Universidade de Aveiro Departamento de Química
2015

Ana Rita
Almeida Pinheiro

Enzimas extracelulares da família *Botryosphaeriaceae*

Extracellular enzymes of *Botryosphaeriaceae* family

Universidade de Aveiro Departamento de Química
2015

Ana Rita
Almeida Pinheiro

Enzimas extracelulares da família *Botryosphaeriaceae*

Extracellular enzymes of *Botryosphaeriaceae* family

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Bioquímica, realizada sob a orientação científica da Doutora Conceição Egas, Investigadora do BIOCANT -Centro de Inovação em Biotecnologia, da Doutora Ana Cristina Esteves, professora Auxiliar Convidada do Departamento de Biologia da Universidade de Aveiro e do Doutor Artur Jorge da Costa Peixoto Alves, Investigador Principal do CESAM e do Departamento de Biologia da Universidade de Aveiro.

Apoio financeiro da FCT e do Fundo Europeu de Desenvolvimento Regional (FEDER) através do COMPETE – Programa Operacional Factores de Competitividade (POFC), no âmbito do projeto PANDORA (PTDC/AGR-FOR/3807/2012 – FCOMP-01-0124-FEDER-027979)

Every accomplishment starts with the decision to try.

o júri

presidente	Prof. Doutor Pedro Miguel Domingues Professor Auxiliar com Agregação do Departamento de Química da Universidade de Aveiro
	Prof. Doutora Cláudia Sofia Oliveira Professora Auxiliar Convidada do Departamento de Biologia da Universidade de Aveiro
	Prof. Doutor Ana Cristina Esteves Professora Auxiliar Convidada do Departamento de Biologia da Universidade de Aveiro

agradecimentos

Ao Professor Doutor António Correia, pela partilhada da sua experiência e saber e por me ter dado a oportunidade de desenvolver este trabalho no MicroLab.
Aos meus orientadores, Dra Conceição Egas, Dra Ana Cristina Esteves e Dr Artur Alves um muito obrigada pelo conhecimentos transmitidos, pelas suas opiniões e pelo apoio no solucionar de questões e contratempos que foram surgindo ao longo da realização deste trabalho. Muito especialmente, desejo agradecer à Dra Ana Cristina Esteves, pela sua disponibilidade, paciência e dedicação.

Aos integrantes do Microlab e aos meus colegas de mestrado por me terem acolhido tão bem e terem sempre uma palavra amiga.

À minha família, em particular à minha mãe e irmã, por serem o meu grande alicerce.

palavras-chave

Familía *Botryosphaeriaceae*, fungos fitopatogénicos, sequências de proteínas, alinhamento múltiplo de sequências.

resumo

As espécies da família *Botryosphaeriaceae* são morfologicamente diversas e descritas como endofíticas, patogénias e saprófitas. Estas são normalmente encontradas numa grande diversidade de hospedeiros. Os fungos patogénicos para plantas *Macrophomina phaseolina, Neofusicoccum parvum* e *Diplodia corticola* secretam uma variedade de enzimas extracelulares, tais como proteases e glicosil hidrolases, algumas das quais envolvidas na interação hospedeiro-patogénio.

A fim de elucidar a correlação entre microrganismo secretoma-hospedeiro, fo comparado entre estes organismos a quantidade de sequências que codificam para enzimas tais como proteases extracelulares e glicosil hidrolases (xilanases e endoglucanases). Através de ferramentas bioinformáticas, tais como, Clustal X2 e T-Coffee, foi realizado o alinhamento múltiplo de sequências dos domínios das proteínas. Além disso, para estudar a relação evolutiva entre as sequências de proteínas foram construídas árvores filogenéticas utilizando a ferramenta MEGA.

Entre *M. phaseolina, N. parvum* e *D. corticola*, o genoma de *D. corticola* contém genes que codificam para uma maior diversidade de famílias glicosi hidrolases sugerindo uma melhor capacidade de adaptação durante sua interação com espécies hospedeiras. A similaridade de sequências observada no alinhamento múltiplo de sequências entre M*. phaseolina, N. parvum* e *D. corticola* é explicado pela sua relação evolutiva e não pelo hospedeiro de cada um. A análise filogenética demonstra que a nível evolutivo, *M. phaseolina e D. corticola* estão mais próximos entre si do que a *N. parvum.*

keywords

Botryosphaeriaceae family, phytopathogenic fungi, protein sequence, multiple sequence alignment

abstract

Species of the *Botryosphaeriaceae* family are morphologically diverse and are described as endophytes, pathogens and saprophytes. They are commonly found in a wide range of hosts. The plant pathogenic fungi *Macrophomina phaseolina, Neofusicoccum parvum* and *Diplodia corticola* secrete a variety of extracellular enzymes, such as proteases and glycoside hydrolases, some of which are involved in host-pathogen interaction.

In order to elucidate the correlation microorganism secretome-host, the amount of sequences encoding extracellular enzymes such as proteases and glycoside hydrolase (xylanases and endoglucanases) was compared between organisms. Through bioinformatics tools, namely Clustal X2 and T-Coffee, multiple sequence alignment of the protein domains was performed. Furthermore, to study the phylogenetic relationship between protein sequences, phylogenetic trees were constructed using MEGA tool.

Between *M. phaseolina, N. parvum* and *D. corticola, D. corticola* genome contains genes that encode a larger diversity of glycoside hydrolase families suggesting a better capacity for adaptability during its interaction with host species. The sequence similarity observed in the multiple sequence alignment between *M. phaseolina, N. parvum* and *D. corticola* is explained by the evolutionary relationship and not by their host type. The phylogenetic analysis shows that at the evolutionary level, *M. phaseolina* and *D. corticola* are closer to each other than to *N. parvum.*

Contents

List of Abbreviations

BaCelLo - Balanced Subcellular Localization Predictor

CAZy - Carbohydrate-Active Enzymes

CBM - Carbohydrate-Binding Modules

CE - Carbohydrate Esterases

CTAB - Cetyltrimethylammonium Bromide

DNA - Deoxyribonucleic Acid

GH - Glycoside Hydrolases

GT - Glycosyltransferases

INSDC - International Nucleotide Sequence Database Collaboration

ISSR **-** Inter Simple or Short Sequence Repeat

ITS - Internal Transcribed Space

JGI - Joint Genome Institute

Mbp **-** Millions of Base Pairs

MEGA - Molecular Evolutionary Genetics Analysis

NCBI - National Center for Biotechnology Information

NLM - United States National Library of Medicine

NJ - Neighbor-Joining

PL - Polysaccharide Lyases

ProMED - Program for Monitoring Emerging Diseases

SMART - Simple Modular Architecture Research Tool

SVM - Support Vector Machines

UniProt - Universal Protein Resource

USA - United States of America

WAG - Whelan And Goldman

List of Figures

List of Tables

1. Introduction

2

1.1. Fungi

The kingdom Fungi comprises a group of organisms with a cell organization clearly eukaryotic.(1) These non-photosynthetic organisms are non-motile and extract their nourishment from the substrate. Their metabolism is characteristically heterotrophic (2) because they are able to create energy for metabolism by absorbing complex organic materials. This is why they have a fundamental role in decomposition of organic matter.(3) The cells, which have rigid cell walls, assimilate and absorb, instead of ingesting, nutrients. The majority of fungi are haploid organisms and they reproduce through spores that can the divided in two types: asexual and sexual spores. Although there are several spore forms, these organisms reproduce either sexually or asexually.(4) Despite most fungi are haploid, some are diploid and a few fungi can alternate between haploid and diploid somatic phases. Some yeasts such as *Candida albicans* are permanently diploid.(1)

Fungi have been used in many fields like in fermentation processes and in the production of antibiotics and other bioactive compounds This group of organisms have a wide industrial and economic significance.(4)

The number of fungal species known in 1983 was about 64,000 and in 1994 the number increased to 70,000.(2) By 2008, more than 100,000 species of fungi had been described. However it is estimated that there are more than 1.5 million fungal species.(3) Between 2006 and 2010, DNA sequencing allowed the discovery of a multitude of new genes for analysis allowing the identification of about 10 phyla belonging to kingdom Fungi. In 2011 a study claims the existence of over 5 million species yet to identify.(5)

1.1.1. Fungal genome sequencing

Comparing to other eukaryotes, the nuclear genomes of fungi are smaller (1) but display genomes with widely differing sizes. The average fungal genome size is 37.7 Mbp.(6) Until 2011, the two largest fungal genomes declared were *Neottiella vivida* (Nyl.) Dennis with 750Mbp (7) and *Scutellospora castanea*

Walker with 795Mbp (8) belonging to the phyla *Ascomycota* and *Glomeromycota*, respectively.

Several factors distinguish fungi from other eukaryotes: they have a shorter life cycle; they create genetically uniform population through asexual reproduction and because they are haploid it's easier to mutate.(1) Because of that, fungi are one of the major targets for genetic research.

The research fields are so vast that can range from descriptive accounts of organisms found in nature to complex laboratory experimentation at the cellular and molecular level. Due to all projects, our knowledge of fungal processes have the potential to expand.(4)

Sequencing the genome of some species brings a new era of information because not only new knowledge of fungi is revealed, but also revolutionized what was already known. Fungi were the first among eukaryotic organisms for which it was possible to apply molecular genetics techniques.(4)

The identification of specific genes and their function in *Saccharomyces cerevisiae* has provided information that has enabled scientists to assign specific functions to human counterparts and provided an invaluable tool to identify genes and their functions in humans and animals.(9)

Using the information obtained from comparing genomes of species from the same family, one can perceive gene organization and function. Using databases with reliable information, the researchers have available several tools to identify specific genes and suggest their function. And even the knowledge about protein-protein interaction of fungi gene products helps to understand the interaction that are also present in humans and animals.(3) Progressively, new databases are launched or re-released having significant importance to genomics. Free database access with information about genome structure and genomic-based comparative biodiversity reports are making a crucial influence in the world of genetics.(7)

1.1.2. Pathogenic fungi

Some fungi can cause diseases in humans, plants, animals.(3) Fungi invaded every conceivable ecological niche and even grow on other fungi.(4)

The majority of fungi have saprotrophic nutrition in which they obtain nutrients from non-living organic matter by absorbing soluble organic compounds. The digestion is external because they secrete enzymes, degrade nutrients from the environment and absorb small molecules.(10) However, some fungi can only complete their life cycle with the help of a host, such as plants or animals.(11) Within this type of fungi, there are fungi able to survive against host defence responses and obtain nutrients from them. When that happens, some can cause diseases that can even lead to the death of the host.

For a particular disease to occur it is necessary the interaction of a susceptible host, a virulent pathogen and favourable environmental factors.(12) Factors such as temperature, precipitation, solar radiation and wind affect fungi and also interfere with pathogen-versus-host relationship.(13)

Fungi affect predominantly plants.(14) When they are associated with plants, fungi can act in one hand as partners in the formation of mycorrhizae and in the other hand as overt pathogens.(4) Among plant pathogens there is a great variety of pathogenesis mechanisms. There are various strategies that fungal pathogens use to invade a plant depending, namely, on host tissue conditions. The fungi can invade immature tissues aggressively and fast often involving extensive destruction of tissues by enzymes or toxins; can stay as symbiotic endophytes until eventually cause disease and even host death; and can invade mature tissues when fungi typically show a high degree of host specialization characterised by tissue invasion and maintenance of host cell viability or cell death.(1) Some pathogens have the capacity to infect the whole plant, while others infect only specific tissues, like leaves, roots, seeds and floral structures, or even propagative organs.(15)

In a study about emerging infectious diseases, major taxonomic groups of pathogens that cause plant diseases or symptoms of the disease were analysed. The database used was ProMED (www.promed-mail.org), Program for Monitoring Emerging Diseases, which is a reporting system dedicated to warning of outbreaks of infectious diseases and critical exposures to toxins that affect human health.(16)

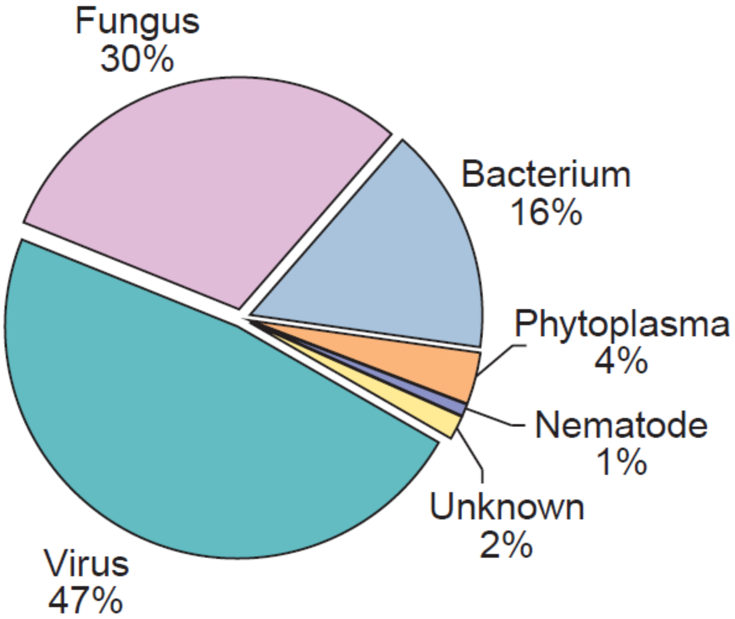


Figure 1. Major taxonomic groups of pathogens causing plant diseases (16)

The mechanism that pathogens use to penetrate the host is considered the most important driver to plant disease. Weather conditions are also an important cause, which might be related to the sensitivity of plants to humidity and moisture levels.(16)

Annually, there are colossal losses worldwide caused by fungi that infect plants, affecting the economy and the ecology. The majority of plant diseases are caused by fungi, making these organisms very relevant for agriculture.(3) With negative impacts to flora and fauna (17) phytopathogenic fungi have a wide range of infection mechanisms.(18) In some situations the injury is very large; this is why the development of disease control methods is needed. Forms of intervention are different depending on the location whether in the city, in the countryside or in forest.(19)

In humans, fungal diseases, also known as mycosis, may possess different degrees of severity and can even put in risk the patient's life. The study about the interaction between human host and fungal pathogens may help to identify why a fungi infect a specific host and if there are more susceptible genomes to different fungi, as well as the consequences behind fungal infection.(3) In contrast to their major significance as parasites of plants, fungi cause relatively few diseases in humans and other animals.(1) Of 1400 human pathogens, merely 20% (about 325) of them are fungal. However, through the years, it has been noted an increase of patients with a type of fungal diseases due to the increase of immunocompromised patients.(17) In patients with immune deficient systems fungal pathogens can be life threatening.(1)

To prevent fungal diseases it is essential a deep study about the genes that are expressed during the infection mechanism. Fungal genome sequencing can reveal the regulation of gene expression that influences the fungus pathogenicity. Furthermore, knowing which genes are turned on or off during infection can help to explain how and when a pathogen's weapons are deployed against a host. Considering the impact of fungal plant parasites in plant production, it is crucial to understand virulence strategies.(11) This kind of information can lead to improvements in the detection, prevention and treatment of fungal diseases. In fact, the information obtained from genomics of the fungal kingdom can open new opportunities in energy production, explain secondary metabolites, unveil enzymes secondary structure and even improve the management of fungal disease.(3)

1.2. The family *Botryosphaeriaceae*

The family *Botryosphaeriaceae* [*Botryosphaeriaceae* Theiss. & P. Syd. (1918)], belongs to the phylum *Ascomycota*, order *Botryosphaeriales* and kingdom *Fungi*.

The fungi of the phylum *Ascomycota* preferentially parasitize fruits and plants, including bananas, apples, strawberries and grapes. For instance, the main host of *Sphaeropsis pyriputrescens* fungus is the apple.(20) These microorganisms are biotrophic, necrotrophic or saprotrophic.(21) Species of the family *Botryosphaeriaceae* are found as endophytes, parasites and saprophytes on a vast number of plants.(22)

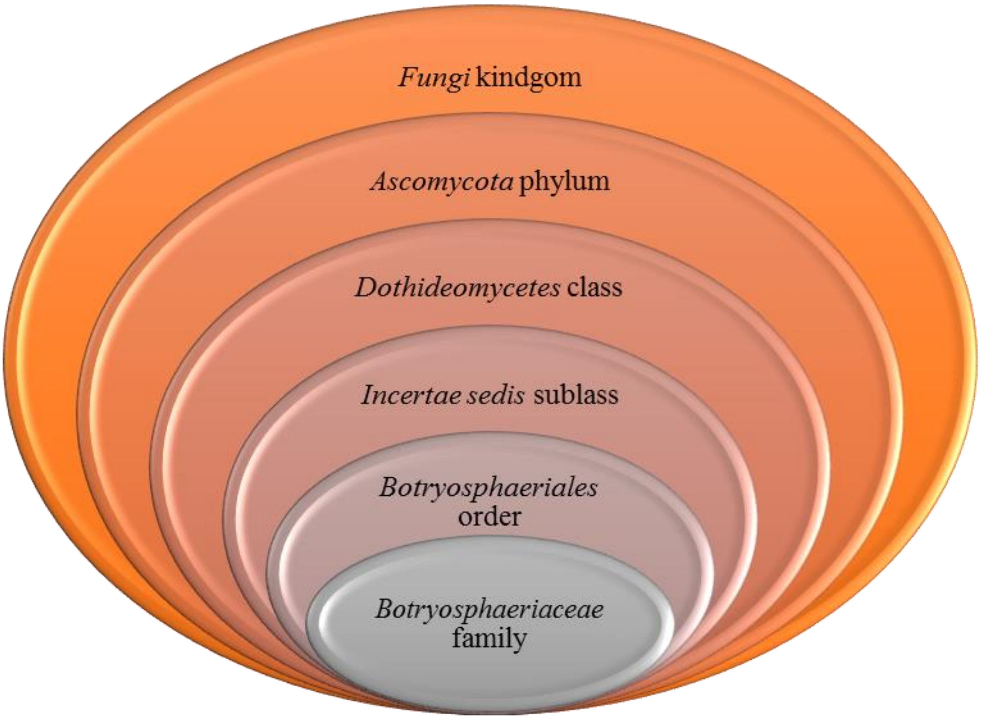


Figure 2. Taxonomy of *Botryosphaeriaceae* family.

In 2007, more than 1500 species were associated with this family. The genera *Diplodia, Botryosphaeria, Fusicoccum, Dothiorella, Lasiodiplodia* and *Sphaeropsis* are the ones that contain a larger number of species.(23)

It's a difficult job to analyse and confirm biological aspects of this family such as their endophytic nature, due to their taxonomic complexities.(24) The interpretation of the taxonomy of genus and species in the *Botryosphaeriaceae* can be frustrating because for a long time many distinct species where treated with the same name.(23)

Applying DNA-based molecular techniques, it was possible to separate the family into several different genera.(24) Recent advances in this area provide efficient tools to classify this group with a reliable identification(23) Molecular techniques, especially the analysis of sequences of variable regions of ribosomal DNA, have been used to clarify the taxonomic genus and have become a powerful tool for species differentiation, allowing to distinguish between closely related species.(25)

It was only in the late 1980's that members of the family *Botryosphaeriaceae* were recognized as endophytes.(26) They are commonly found in a wide range of hosts, ranging from plants to humans. Based on that information, it is now thought that the majority of *Botrysphaeriaceae* members, might have an endophytic phase.(23) Endophytes are microorganisms that preserve themselves in the plant interior. Together, the plant and its endophytic microbiome act as an organizational unit. The non-harmful fungi do not change plant morphology even when interacts with the host as mutualist or commensal at some point of their life cycle.(27)

Several species of the family *Botryosphaeriaceae* are plant pathogens causing leaf spots, fruit rots, dieback, cankers and even death. Furthermore, some species have been classified as human pathogens causing subcutaneous, ocular and internal organ infections.(28) *Botryosphaeria dothidea*, *D. mutila* and *Lasiodiplodia theobromae* were registered in Chile (1986) for causing grapevine cankers and dieback, in Italy (1987) and in California (1990). (23)

It is well known the economic importance of these fungi. However, due to taxonomic difficulties, it's not clear their presence and ecological role in native plant communities. Factors influencing pathogenicity include extreme weather conditions, such as unpredictable rainfall, lower or higher rainfall in different areas and extreme temperatures. Depending on the intensity and extension of factors of stress, disease symptoms can spread rapidly and cause extensive losses over large areas.(29) Diseases caused by *Botryosphaeriaceae* arise due to the onset of stress factors and the biological pressure from the infection itself.

1.2.1. Genus *Neofusicoccum*

Neofusicoccum includes species as *N. parvum, N. vitifusiforme, N. luteum, N. ribis* and *N. australe*.(30)

Neofusicoccum is a vascular pathogen that causes severe decline and dieback symptoms in grapevines worldwide.(30) Species belonging to this genus are considered pathogenic in numerous other plants causing serious economic losses in agriculture.(31)

1.2.1.1. *Neofusicoccum parvum*

A project estimated the size of *N. parvum* genome with a total size of 42.5Mb, in accordance with the genome size of other plant-pathogenic ascomycetes.(32) Lignin peroxidases and cytochrome P450 monooxygenases involved in lignin degradation were described. Also 163 glycoside hydrolases, 22 polysaccharide lyases, and 8 cutinases that might have an important role during the colonization of host tissues were described. In total, 1097 proteins identified as potentially secreted were reported.(33)

Neofusicoccum parvum was reported as the pathogen responsible for causing canker symptoms on *Eucalyptus* species in distant parts of the world as, for instance, Australia, Chile, China, Ethiopia, Indonesia, South Africa, Uganda, Uruguay, Venezuela (34) and Spain.(35)

Neofusicoccum parvum causes dark streaking of the wood (22) and bunch rot, as can be confirmed by the Chardonnay and Shiraz grapes study in Australia.(36) To fully understand the infection mechanism of *N. parvum*, developments on genome sequence data should be made.(24)

The infection mechanism involves cell wall degrading proteins that allow to penetrate in plant cells and phytotoxins that induct cell death. Through pruning wounds, the pathogen colonizes the host tissues causing shoot dieback, cane bleaching, bud necrosis, and graft failure.(37)

Neofusicoccum parvum strain UCRNP2 was obtained from the margin of a grapevine (*Vitis vinifera*) wood canker collected in Riverside County (California, USA) in 2011. The genome sequenced using the Illumina HiSeq 2000 platform has a median total length of 42.59Mb, 10,470 predicted complete protein-coding genes and a median GC content of 56.7%.(33) The whole genome shotgun project was deposited at the INSDC AORE00000000.

1.2.2. Genus *Macrophomina*

The genus *Macrophomina* includes only one species, *M. phaseolina*. Although previously five species were described, due to recent phylogenetic studies currently the genus is monotypic.(38)

1.2.2.1. *Macrophomina phaseolina*

Macrophomina phaseolina is a pathogen (a root inhabitant) that causes diseases in more than 500 species, becoming one of the most destructive necrotrophic fungal pathogens. The fungus has a worldwide distribution, the capacity to develop in extreme climatic conditions (39), great longevity and highly competitive saprophytic ability.(40) Among the many hosts, some stand out for their economic importance such as soybean, common bean, corn, sorghum, cowpea, peanut and cotton.(39) Only in United States of America, loses due to plant diseases caused by this fungus were estimated at about 27 million bags of soybeans per year. And in the zone of West Africa (Niger and Senegal), the cowpea rot disease cause a loss of $US 146 million.(41) Not only are the plants the target of this fungus, some reports reveals *M. phaseolina* acting as an opportunistic human pathogen, in particular of immunosuppressed patients.(42)

When *M. phaseolina* induces diseases symptoms on a plant, it results in decreased stem height, root rot, stem rot of grasses and even death of the affected plant. The disease caused by this particular fungus is known as charcoal rot for the abundant production black sclerotia that cause the rotted tissues.(43) Infection by this fungus is favoured by a set of conditions such as soils poor in organic matter and with low levels of potassium; temperature between 28ºC and 30ºC; stress caused by the attack of other pathogens, environmental or nutritional factors; successive cultivations and soil with low water retention capacity and high heat absorption capacity.(14) This explains the dominances of this pathogen in the warmer months of the year (the late spring/early summer) when temperatures often exceeds 30ºC.(44)

Fungus can remain viable for more than 4 years in the soil. First the hyphae invade the cortical tissue of plants, then begins sclerotia formation and finally it sets stem rot disease. The infected area are invaded by grey black mycelia and sclerotia displaying disease symptoms. (45)

Macrophomina phaseolina strain MS6 was isolated from an infected jute plant at Bangladesh Jute Research Institute (BJRI), Dhaka. Using a whole-genome shotgun approach a total of 6.92 Gb of raw sequence was generated using a combination of 454 and Illumina platforms. This strain has a genome size of 48.88Mb of which 98.53% is non-gapped sequence, median protein count of 13806 and a median GC content of 51.2%. The *M. phaseolina* genome comprises 2.84% repetitive DNA and 3.98% transposable elements. After the draft genome sequencing and assembly, 14,249 protein-coding genes were predicted and 9,934 were validated by the transcriptome. At NCBI 3154 nucleotide sequences are deposited.(45) The whole genome shotgun project was deposited at the International Nucleotide Sequence Database Collaboration (INSDC) AHHD00000000.1.

1.2.3. Genus *Diplodia*

Diplodia is known to enclose species that are pathogens, endophytes and saprophytes on a wide range of mainly woody hosts.(46) *Diplodia* is a large genus with more than 1 000 species recognized in 2012 but the number decreased to 388 species based on a search of Catalogue of Life (last actualization has in 2014-02-19).(47)

There are reports with diverge statements regarding the pathogenicity of some of the species, possibly due to an incomplete knowledge of the taxonomy of the genus leading to unreliable species recognition and identification. The most influential pathogens are *Diplodia sapinea, Diplodia mutila, Diplodia seriata* and *Diplodia corticola.*(48)

1.2.3.1. *Diplodia corticola*

Botryosphaeria corticola (anamorph: *Diplodia corticola*), reported by Alves at al. (59), was initially identified as *Botryosphaeria stevensii* (anamorph: *Diplodia mutilates*).

Recently, it was shown that *B. corticola* and *B. stevensii* are different species that can be distinguished by morphology and by nucleotide sequence from ribosomal DNA.(49)

Although fungi within *Botryosphaeria* can cause disease symptoms in hundreds of plant genus, *D. corticola* was only found on oaks, grapevines and eucalypts. This fungus penetrates into plants through wounds, scars including leaf or stomata open for gas exchange.(3)

Bot canker is the informal name of the disease identified in 2010 as the cause of oak death in California and Florida, and grapevine mortality in California, Texas, and Spain.(50) This name comes from the sexual stage the fungus that cause bleeding trunks and branches cancer.(51) However, information available regarding the pathogenicity of *D. corticola* is very limited. The first evidence was in Greece, Hungary, Italy, Morocco, Portugal and Spain and was recently associated with disease symptoms in *Quercus chrysolepis* and *Quercus agrifolia* in California.(50) The main cause of the cork oak disease are fungi from *Botryosphaeria* family.(52) It was reported in the main producing countries of cork such as Portugal, France, Spain, Italy and Morocco. Overall, the plant diseases (in all sorts of plants) can cause massive economic losses. The damage can occur from cultivation to harvesting or storage.

A species that belongs to the same genus, D. *corticola*, produces several phytotoxins, especially a newly diplopyrone identified as monosubstituted tetrahydropyranpyran-2-one, in toxic concentrations for *Q. suber* (0.01-0.1 mg/ml). The known metabolites produced are diplopirones and sphaeropsidins.(53) In addition, *D. corticola* secretes three classes of hydrolases able to degrade cellulose: exoglucanases, endoglucanases, and β-glycosidases.(54,55) These enzymes have already been described on the secretome of fungi that infect plants and wood.

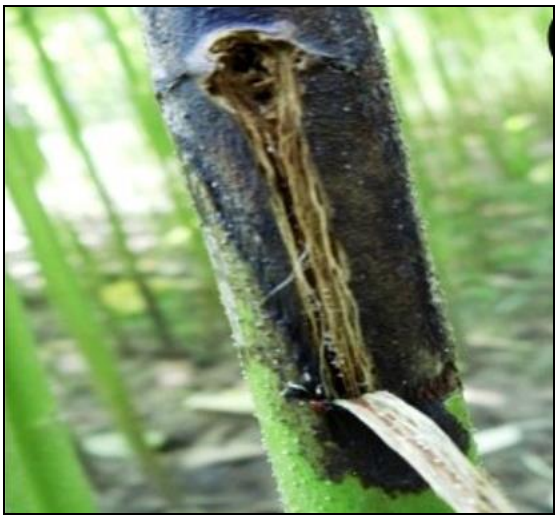


Figure 3. During fungal infection, hyphae penetrate the plant cell wall and produce disease symptoms.(53)

1.2.4. Infection mechanism

It is still unknown how fungi overcome the plant's defence system induced by the microbial attack. And to what extent fungal pathogenicity genes reveal information regarding to fungus's ability to recognize a parasite, attach to the plant surface and penetrate and colonize the host.(56) Mechanism of pathogenesis by a fungus is a very complex process that include numerous steps: arrival of an infectious organism (usually a spore of some kind) in the host or in a nearby area, adhesion and recognition of the host, penetration into the host, invasive growth within the host, lesion development in the host, and finally production of additional infectious particles.(57)

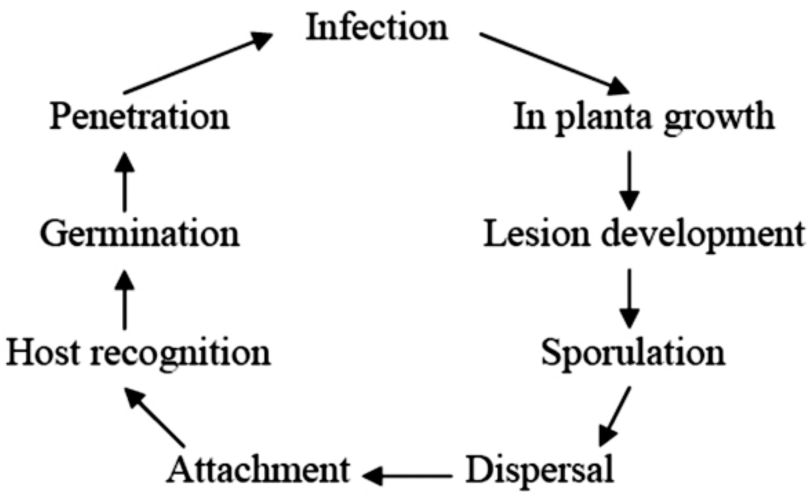


Figure 4. Infection and disease cycle caused by fungi.(64)

To start a new infection process, dispersal of spores is the most common process.(58) Adhesion of spores to the host involves physical and chemical processes. Throughout adhesion molecules, the spore connects to the surface of a host tissue natural openings or injuries (caused by adverse environmental conditions, such as temperature).(59) Primary sites of infection of blueberry species by *B. dothidea* are pruning wounds and sites of winter injury.(60)

When the pathogen evolves, it also feeds itself into the host. A biotrophic fungus may grow, thus acquiring nutrients from the host without killing the cells and maintaining close contact with them. But many facultative parasites secrete enzymes that cause death and the disintegration of the cellular components of the plant.(61) Colonization and therefore the disease process, only develops when the pathogen's mechanisms of action outweigh the host defence mechanisms.(62)

Despite the mechanism of pathogenesis of *D. corticola* being largely unknown, the involvement of this fungus on the decline of cork oak forests is well known.(49) The first project using proteomic to characterise a member of the *Botryosphaeriaceae* family was only in 2010.(63) Since then, proteomics techniques have been use to characterise phytopathogenic fungi secretome, and thereby contributing to elucidate the infection mechanism and provide information for the development of disease management strategies.(55) *De* *novo* sequencing methods allow to identify proteins from organisms with unsequenced genomes, which is the case of many *Botryosphaeriaceae* species, including *D. corticola*.(64) Despite the increased number of researches that use proteomic to study phytopathogenic fungi in the last decade, same factors have been interfering, such as the low concentration of extracellular proteins, high amount of polysaccharides and the presence of low-molecular-weight metabolites also secreted by fungi.(55,65)

It is now known some important data regarding to *D. corticola* extracellular proteins as their identification and characterization. Despite being an asset, it is still required studies that unlock information about *D. corticola* infection mechanism and plant pathogen interactions. Knowing the proteins involved in the interactions and mechanisms is the key to full understanding.

1.3. Extracellular enzymes

The production of enzymes is common in many species of filamentous fungi. Fungi secrete a variety of extracellular enzymes such as amylases, cellulases, proteases, lipases, xylanases, pectinases, laccases, among others.(66). Due to the wide range of enzymes, they can be involved in gathering nutrients and ecological interactions. Some are involved in the host-pathogen interaction such as: polygalacturonases, pectate lyases, xylanases and lipases.(20)

The synthesis of these enzymes is subject to various regulatory mechanisms of inhibition and induction. Signal peptides are part of the protein that are exported and after being recognized by the cell can pass through the pores of the cell membrane.(67) Many extracellular enzymes have been implicated in the infection mechanism of many plant pathogens and therefore, deserve special attention.

In the present study, we provide an overview of the distribution of fungal proteases of 3 different organisms. We performed amino acid sequence alignment of the proteases functional domain, and then a preliminary phylogenetic analysis.

1.3.1. Proteases

Proteases are hydrolytic enzymes that break the peptide bonds in proteins(68) and fragments of proteins forming amine and carboxyl groups, resulting in smaller peptides and/or amino acids.(69,70) Proteases are one of the three largest groups of industrial enzymes representing 60% of international sales of enzymes.(71)

For a long time the exact meaning of the terms proteases, proteinases and peptidases has not been clear. For a more accurate definition of the nomenclature of proteolytic enzymes, it is required to explain each one. Some proteolytic enzymes act best on intact proteins, whereas others show a preference for small peptides as substrates, so in 1928, Grassmann and Dyckerhoff proposed the new term proteinase for proteases that show specificity for intact proteins.(68) Some might expect the word proteinase to apply to exopeptidases but in previous works the enzymes included in sub-subclasses EC 3.4.21-99 carried the same meaning as endopeptidase. To avoid mistakes, the word has been replaced by endopeptidase for consistency.(72) The peptidase terms are undoubtedly the more rational, and probably should eventually become standard. The majority of proteases are specific, meaning that do not hydrolyse protein molecule at any peptide bond, but only links between particular amino acids.(73)

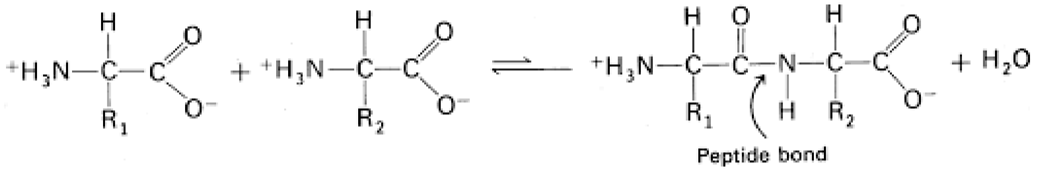


Figure 5. Hydrolysis of a peptide bond. Proteolysis involves the addition of a water molecule to break the carbonyl carbon-nitrogen single bond connecting individual amino acids.(181)

The process is thermodynamically favourable by adding a water molecule, but however, has a very slow kinetics without the catalyst.(74)

Proteases include two groups of enzymes: the endopeptidases, which break peptide bonds within the protein chain, and exopeptidases which reach the terminal amino acids of the polypeptide chain. The implemented classification system is currently based on comparison of active centres, mechanism of action and three-dimensional structure.(75) Proteolytic enzymes are divided in several families according to their homology to the type enzyme.(75) Each family is identified by a letter representing the catalytic mechanism of the proteolytic enzymes: aspartic (A), cysteine (C), glutamic (G), metallo (M), asparagine (N), serine (S) and threonine (T) proteases.(76,77) The metalloproteases have, as the name indicates, in their structure a binding site for a metal ion.

In the presence of plant extract, the proteases secretion by filamentous fungi increase.(78) Plant cell walls and extracellular matrices contain glycoproteins and proteoglycans, so the degradation of these enzymes can help in fungal hyphal penetration.(79) Usually proteases were not seen as key partners on fungal phytopathogenesis. However, proteases have an important role in physiological and pathological processes.(80) Due to the functional diversity (exo and endoproteases), proteases have the ability to interact synergistically and therefore degrade a variety of substrates.

1.3.1.1. Aspartic proteases

Aspartic proteases are a catalytic type of protease enzymes that catalyses peptide substrates using aspartate residues. Aspartic proteases are widely distributed among all the biological kingdoms. They can be found in vertebrate, plants, yeast, fungi, parasites, nematodes, bacteria and even in viruses.

According to the MEROPS database, aspartic proteases can be divided into 16 families and 5 clans, according to their similarity in amino acid sequences and structural relationships, respectively.(81)

Structurally, they have a two-domain structure, arising from ancestral duplication. Each domain contributes with a catalytic Asp residue, with an extended active site cleft localised between the two lobes of the molecule. One lobe has probably evolved from the other through a gene duplication event in the distant past. (82) Typically it has a sequence of Asp-Thr-Gly but Asp-Ser-Gly are also found. It is usually represented as monomeric enzymes with twofold symmetry and has a tertiary structure with an N-terminal and a C-terminal.

The great majority of the aspartic peptidases are members of the pepsin (Al) family, which have been found only in eukaryotic organisms. The viral retropepsins (family A2) have a monomeric structure in which each molecule contains only half of the functional catalytic site and dimerization is needed to form the active enzyme. Both families Al and A2 have a very similar tertiary structure and sensitivity to inhibition by pepstatin.(83,84)

Even though the amino acid sequences between aspartic sequences are different, the catalytic site motif is well conserved. And the three-dimensional structures between aspartic proteases are very similar.(85) The presence and position of disulphide bridges are a conserved characteristic of aspartic proteases, predicted to be important in maintaining the structure of the enzyme.(86)

Catalytic mechanism

The mechanism of action is an acid-base mechanism involving the coordination of a water molecule between the two highly conserved aspartic acid residues in the active site. One aspartate activates the water by abstracting a proton, enabling the water to attack the carbonyl carbon of the substrate scissile bond, generating a tetrahedral oxyanion intermediate. This catalytic cleavage of their peptide substrates is optimally active at acidic pH. Rearrangement of this intermediate leads to protonation of the scissile amide.(87,88)

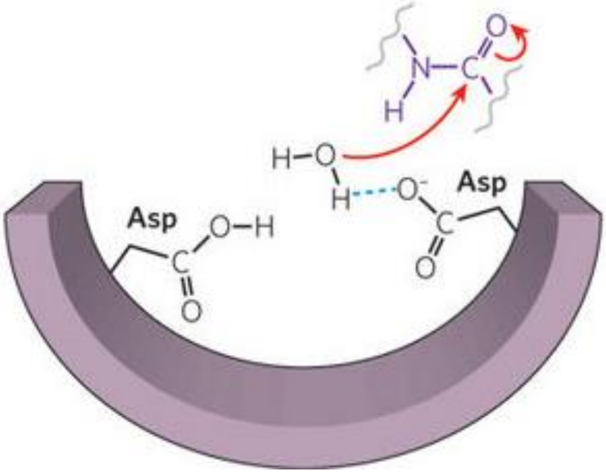


Figure 6. Mechanism of peptide bond hydrolysis by an aspartic protease.(89)

Some examples of the Eukaryotic aspartic proteases are pepsins, cathepsins and renins. Nearly all known aspartyl proteases are inhibited by pepstatin. Members of the aspartic protease family can be found in different organisms, ranging from vertebrates to plants and retroviruses. The best known sources of aspartic proteases are the stomach of mammals, yeast and fungi, with porcine pepsin as the proto type.(88)

1.3.1.2. Serine proteases

Serine proteases are a family of enzymes with a catalytic serine residue and constitute the largest group of peptidases.(90) According to the MEROPS database, serine proteases can be divided into 52 families and 12 clans, according to their similarity in amino acid sequences and structural relationships, respectively.(81)

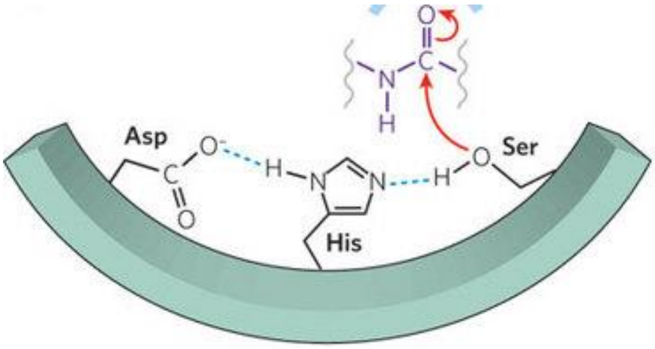


Figure 7. Mechanism of peptide bond hydrolysis by a serine protease.(89)

Commonly serine proteases utilise a uniquely activated serine residue to catalytically hydrolyse peptide bonds in proteins.(91) The functionality of the catalytic Ser is dependent on a catalytic triad of Asp, His and Ser residues. Each amino acid plays an essential role in the cleaving ability of the proteases. Each amino acid in the triad performs a specific task in this process. The serine has an hydroxyl group that donates a pair of electron attacking the carbonyl carbon of the peptide bond of the substrate.(92) A pair of electrons on the histidine nitrogen has the ability to accept the hydrogen from the serine hydroxyl group, thus coordinating the attack of the peptide bond. Aspartic acid has a carboxyl group that forms hydrogen bonds with histidine, making the nitrogen atom much more electronegative.

The triad is located in the active site of the enzyme, where catalysis occurs, and is preserved in all serine protease enzymes. The particular geometry of the triad members is highly characteristic to their specific function. The amino acid members of the triad are located far from one another on the sequence of the protein, but due to folding, they will be very close to one another in the center of the enzyme.(93) Many serine peptidases combine a simple mechanism where Lys or His is paired with the catalytic Ser. Inhibitors such as diisopropylfluorophosphate and phenylmethanesulfonyl fluoride can inactivate this mechanism. The Table 1 summarise the catalytic units in serine peptidases.

Table 1 Catalytic units of serine peptidases. Adapted from (102). *The MEROPS family S58 has been transferred as family P1

Families	Representative enzymes	Catalytic residues
S1	Trypsin	His, Asp, Ser
S8, S53	Subtilisin, sedolisin	Asp, His, Ser
S9, S10, S15, S28, S33, S37	2Prolyl oligopeptidase	Ser, Asp, His
S11, S12, S13	D-Ala-D-Ala carboxypeptidase	Ser, Lys
S24	LexA peptidase	Ser, Lys/His
S21, S73, S77, S78, S80	Prohead peptidase	His, Ser, His
S16, S50, S69	Lon peptidase	Ser, Lys
S14, S41, S49	Clp peptidase	Ser, His, Asp
S59	Nucleoporin	His, Ser
P1*	Aminopeptidase DmpA	Ser
S60	Lactoferrin	Lys, Ser
S66	L,D-Carboxypeptidase	Ser, Glu, His
S54	Rhomboid	His, Ser

Serine proteases are found in both eukaryotes and prokaryotes. Serine proteases are divided into two categories based on their structure: chymotrypsin-like (trypsin-like) or subtilisin-like. In humans, these enzymes are responsible for a diverse array of physiological functions, the best known being digestion, blood coagulation, fibrinolysis, fertilization, and complement activation during immune responses.(91,94) Peptidases of family S1, trypsin-like, are the most abundant among serine peptidases. These enzymes participate in many important physiological functions including digestive and degradative processes, cellular and humoral immunity and embryonic development.(84,90)

There has been increasing interest in the identification, structural, and functional characterization of all members of the serine protease family of enzymes in humans and other organisms. Structural characterization of all serine proteases and extensive analysis of their location are the first steps towards understanding the control of their gene expression and their involvement in various physiological and pathological conditions.(91) The protease domain in serine proteases is both necessary and sufficient to specify function and evolution.(95)

1.3.1.3. Cysteine peptidases

Cysteine proteases are a family of proteolytic divided into families on the basis of the architecture of their catalytic dyad (cysteine-histidine). The sequences around the three active site residues are well conserved.(96)

Papain (from *Carica papaya*) is the best characterized family of cysteine proteases, characterized by a two-domain structure. The active site (catalytic pocket), where the substrate is bound, is located between the domains. The catalytic residues of papain are Cys25 and His159, and they are evolutionarily preserved.(97)

Catalytic mechanism

Cysteine proteases have a catalytic mechanism that involves a nucleophilic cysteine in a catalytic triad. Cysteine and histidine in the active site are highly conserved, hence have been used to classify proteases. Their catalytic mechanism is similar to serine proteases, however due to the extra shell of electron cysteine proteases are better nucleophiles.(98)

Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. The basic mechanism of action of serine proteases involves transfer of the acyl portion of a substrate to a functional group of the enzyme. Firstly occurs the formation of an ester bond between the oxygen atom of serine and the acyl portion of the substrate. This step produces a tetrahedral intermediate and releases the amino part of the substrate. And then the attack of water on the acyl-enzyme intermediate, which breaks it down and releases the acidic product.(99) This mechanism is shown in Figure 8.

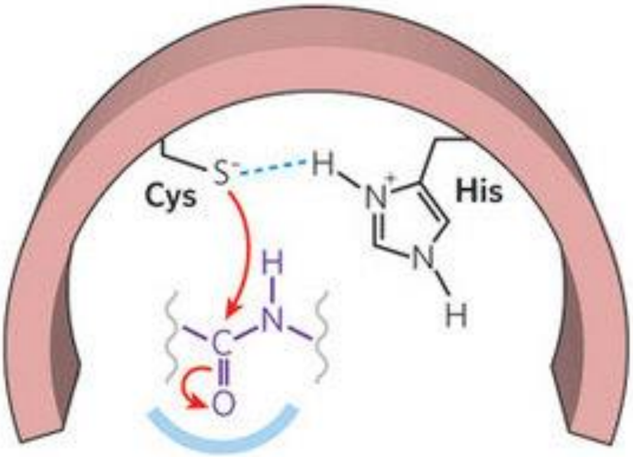


Figure 8. Mechanism of peptide bond hydrolysis by a cysteine protease.(89)

Cysteine proteases are commonly encountered in fruits including the papaya, pineapple, fig and kiwifruit. The proportion of protease tends to be higher when the fruit is unripe. These enzymes are also used as an ingredient in meat tenderizers. These enzymes participate in many roles in microsporogenesis, symbiosis, hypersensitive response, signal transduction and differentiation, senescence, and protein degradation.(99)

1.3.1.4. Metallo peptidases

According to the MEROPS database, metalloproteases are the most diverse of the four main protease types. Metalloproteases are divided into 89 families and 15 clans, according to their similarity in amino acid sequences and structural relationships, respectively.(81) The majority of zinc-dependent metallopeptidases share a common pattern of primary structure in the part of their sequence involved in the binding of zinc.(100)

In these enzymes, a divalent cation, usually zinc, activates the water molecule. The metal ion is held in place by amino acid ligands, usually three in number. The known metal ligands are His, Glu, Asp or Lys and at least one other residue is required for catalysis, which may play an electrophilic role.(101) This mechanism is shown in Figure 9.

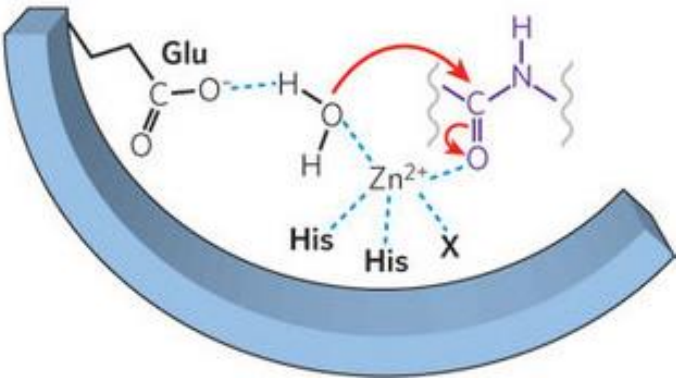


Figure 9. Mechanism of peptide bond hydrolysis by a metallo protease.(89)

The majority of the enzymes contain zinc, and for several of them the residues involved in binding the zinc have been identified by X-ray crystallography. Of the families of metallopeptidases, 13 contain the sequence HEXXH, which is known or suspected to provide two of the three ligands for the zinc atom.(84)

Explaining the domain structure in a more detailed manner, it can be defined as 'abXHEbbHbc'. The “a” is most often valine or threonine; “b” is an uncharged residue, and “c” a hydrophobic residue. Proline is never found in this site, possibly because it would break the helical structure adopted by this motif in metalloproteases.(102)

1.3.2. Glycosidase hydrolases

1.3.2.1. Xylanases

Xylanases are hydrolytic enzymes able to degrade the linear polysaccharide beta-1,4-xylan. Endo-xylanases and β-xylosidases are the main responsible for the hydrolysis of xylan, the major component of hemicellulose.(103)

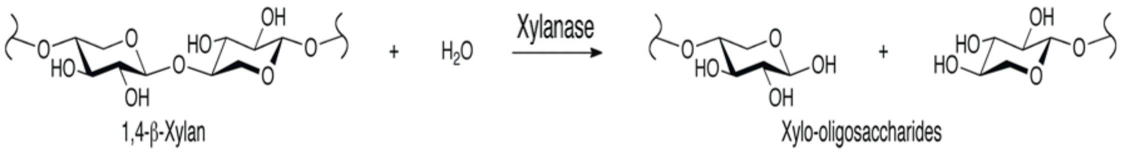


Figure 10. Schematic representation of the overall xylanase reaction.

Hemicellulose typically includes five different sugars: L-arabinose, D-galactose, D-glucose, D-mannose, and D-xylose and even components such as acetic, glucuronic, and ferulic acids. Based on the backbone of the chains, hemicelluloses can be a homopolymer (only a single sugar repeat unit) or a heteropolymer (mixture of different sugars). Hemicellulose has different classifications namely xylans (are polysaccharides made from units of xylose), mannans, glucans, glucuronoxylans, arabinoxylans, glucomannans, galactomannans, galactoglucomannans, β-glucans, and xyloglucans.(104)

In plants, xylans are situated in the overlying layer of lignin and cellulose fibres underneath.(105) Based on their structural chemistry, they seem to be covalently linked to lignin and a non-covalently interacting with cellulose; this is an important factor to maintain the integrity of the cellulose *in situ* and to help to protect the fibres against degradation by cellulases.(106)

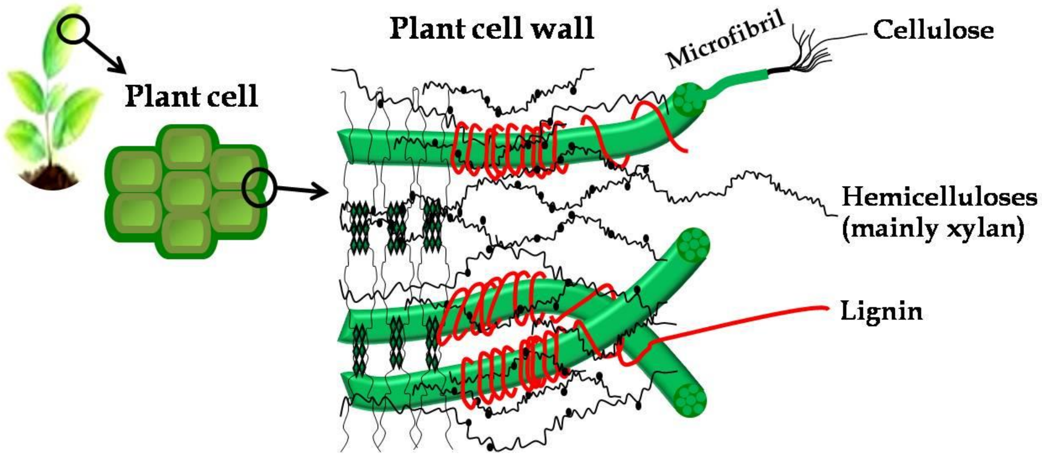


Figure 11. Structure of plant cell wall. Image adapted from (114)

In hardwood from angiosperms is common to find xylan as the major hemicellulose, but it is less abundant in softwood from gymnosperms. Homoxylans are built exclusively of xylosyl residue.(107) Even though they have been isolated from esparto grass, tobacco stalks and guar seed husk, is rare to find homoxylans in nature.(105) The majority occur as heteropolysaccharides, containing different substituent groups in the backbone chain and in the side chain acetyl, arabinosyl, and glucuronysyl residues are used as substituents in the backbone of xylan.(108)

To achieve a complete degradation of xylan, several hydrolytic enzymes with diverse specificity and modes of action are required. Usually, xylan hydrolysis is achieve through a multifunctional xylanolytic enzymatic system that includes: β-1,4-endoxylanase, β-xylosidase, α-L-arabinofuranosidase, α-glucuronidase, acetyl xylan esterase, and phenolic acid.(105)

In 1998, a study verify that in a fungus (*Trichosporon cutaneum*) xylanase is induced by xylose, but is repressed in the presence of glucose.(109) Lignocelluloses such as wheat bran, rice straw, corncobs, and sugarcane bagasse are able to induce the production of xylanases.(110) Moreover, it is possible to suppress xylanase synthesis applying sugars, such as glucose and/or xylose.(111)(105)

1.3.2.2. Endoglucanases

Glucanases are enzymes that hydrolase glucans by breaking the glucosidic bond. Glucans are polysaccharides made of several glucose sub-units. Glucanases can be divided into α-glucanase and β-glucanases: enzymes that break down α-glucans and β-glucans, respectively.

Cellulolytic systems are constituted by enzymes that can act on ends (exoglucanases) or on inside (endoglucanases) of the cellulose chains. Endoglucanases hydrolyse mainly internal links in the polymer cellulose, producing new terminal chains. Exoglucanases start the hydrolysis at the end of the chain, and do not produce a significant amount of new terminals on the surface of the cellulose chains.(112)

A study developed in 2009 confirmed that a β-1,3-glucanase produced by a fungus (*Colletotrichum higginsianum*)was involved in the process of infecting a plant.(113) The results of a recent study show that several hydrolases are potentially involved in pathogenesis of *Diplodia corticola*, including endoglucanases.(55)

Glucans are present in plant cell walls and their degradation is used as medium for invading fungi to penetrate the plant cell wall or to remodel fungal cell-wall during infection.(114)

1.4. OBJECTIVES

Fungi belonging to *Botryosphaeriaceae* family are mostly endophytes able to colonize a wide variety of hosts. Several species, such as *Macrophomina phaseolina, Neofusicoccum parvum* and *Diplodia corticola* are important phytopathogens causing leaf spots, fruit rots, dieback, perennial cankers and death of economically important plants. It is expected that phytopathogenic fungi express high amounts of hydrolytic and oxidative enzymes, such as proteases and glycoside hydrolases.

Although there is some information regarding the extracellular enzymes expressed by *M. phaseolina, N. parvum* and *D. corticola*, a detailed comparison of proteases and glycoside hydrolases is missing.

The main goal of this work is to analyse the extracellular enzymes, namely proteases and glycoside hydrolases, encoded in the genomes of these fungi, contributing for a better understanding of their pathogenesis mechanisms. For that purpose, it will be analysed the enzymes (proteases and glycoside hydrolases) between organisms and correlate with their hosts. Furthermore, the use of bioinformatics tools will allow us to obtain information regarding these fungi proteins sequence similarity and their phylogenetic relationships.

2. Materials and Methods

2.1. Selection of Enzyme Sequences

2.1.1. *Macrophomina phaseolina* and *Neofusicoccum parvum*

The information about *Macrophomina phaseolina* and *Neofusicoccum parvum* genomes are stored at the National Center for Biotechnology Information (NCBI) at database GenBank. The National Center for Biotechnology Information is a United States of America resource for molecular biology information that shares a series of databases relevant to biotechnology and biomedicine.(115) All these databases are available online through the Entrez search engine http://www.ncbi.nlm.nih.gov/.

One of the main databases is GenBank, an open access sequence database of annotated nucleotide sequences and their protein translations. Only original sequences can be submitted to GenBank in order to provide access to the most recent and reliable DNA sequence information.(116)

Due to the improvements in sequencing technologies, the number of nucleotide and protein sequences deposited in databases has increased. However, the scientific community does not have the capacity to annotation these sequences leading to the increase of unannotated sequences.

Proteases

To select the set of sequences of proteases in the organisms under study (*M. phaseolina* and *N. parvum*), NCBI's Protein database was used. Which is a collection of sequences from several sources including translations from annotates coding regions in GenBank. Protein sequences are the fundamental determinants of biological structure and function.

In this database, a search was made crossing the name of each organism with the protease, peptidase and proteinase keywords.

Glycoside hydrolases

To select all known glycoside hydrolases was used the Carbohydrate-Active Enzymes (CAZy).database.(117)

CAZy database displays and describe families of enzymes that degrade, modify or create glycosidic bonds, such as the glycoside hydrolases (GH), the polysaccharide lyases (PL), the carbohydrate esterases (CE), the glycosyltransferases (GT) and their appended non-catalytic carbohydrate-binding modules (CBM).

CAZy defines families based on significant amino acid sequence similarity (118) linking the sequence to the specificity and 3D structure of the enzymes that assemble, modify and breakdown oligo and polysaccharides.(119) However, sequences with low similarity are not included, to ensure a significant alignment, and are stored in the non-classified section of each CAZyme category, awaiting biochemical characterization.(119) Furthermore, to avoid analysing unfinished sequences that may change accession number, this database only analyses protein sequences that are regularly updated in GenBank. As a result of the classification systems of CAZy a single family may exhibit a variety of substrate specificities or chemical reactions they catalyse among its members.(120)

In this work it is intended to make an overall analysis of the glycoside hydrolases families of the species of interest, focusing on the enzymes with xylanolytic and glucanolytic activities.

Using CAZy tool, a search was made crossing the names xylanases and endoglucanases in the “known activities” search field. So, it was possible to know which glycoside hydrolases families have enzymes with xylanolytic and glucanolytic activities.

With the information obtained regarding the glycoside hydrolase family number that includes enzymes with the required activities, a search was made in NCBI tool. In the Protein field at Protein Advanced Search Builder the name of each organism (*D. corticola, N. parvum* and *M. phaseolina*) was crossed with glycoside hydrolase family number.

At Protein Advanced Search Builder:

Organism: *Macrophomina phaseolina*

All fields: Glycoside hydrolase family 5

2.1.2. *Diplodia corticola*

Diplodia corticola strain CBS112549 (= CAA004) genome was recently sequenced by University of Aveiro in collaboration with Ghent University. The nucleotide and the protein sequences as well as their description, length and molecular function were annotated with Blast2Go. The genome size of this fungus is 34,997 Mb (data not shown). The selection of proteases and glycoside hydrolases of *D. corticola* was based on the sequence annotation.

2.2. Selection of Sequences According to their Subcellular Localization

The purpose of this study was the characterisation of the secretome of these fungi involved in the infection mechanisms, so only extracellular enzymes were selected.

BaCelLo (Balanced Subcellular Localization Predictor) is a predictor for the subcellular localization (secretory pathway, cytoplasm, nucleus, mitochondrion and chloroplast) of identified proteins in eukaryotes.(121) BaCelLo is a fungi-specific predictor used in previous works of our laboratory to confirm the extracellular localization of proteins from a fungal plant pathogen belonging to the family *Botryosphaeriaceae.*(55)

BaCelLo analyses the sequence of both the N- and C-termini and does an alignment profile to obtain evolutionary information. The system displays different several support vector machines (SVMs) organized in a decision tree. The training set is curated in order to avoid redundancy. Three kingdom-specific predictors are implemented: animals, plants and fungi, in which the last one is classify to four localizations (secretory pathway, cytoplasm, nucleus and mitochondrion). BaCelLo's predictions are balanced among different classes and all the localizations are considered as equiprobable.(121)

For this work, the field “Fungi” was selected and then the complete amino acid sequence in FASTA format of up to 5 globular proteins (software maximum analysing capacity) was submitted.

2.3. Noise reduction

In order to obtain a reliable selection of proteases and glycoside hydrolases enzymes, it was established additional steps. After the gathering of all known enzyme sequences in NCBI tool, the sequences were thoroughly analysed to be sure we did to we did not have repetitive protein sequences, enzymes without activity or even non-peptidase homologues.

2.3.1. Elimination of sequences

The enzyme description was analysed in detail. Search results were filtered against false positives regarding the enzymes with "peptidase"/"protease"/"proteinase" in their description, excluding enzymes representatives of other classes, such as "peptidase inhibitor".

2.3.2. Elimination of redundant sequences

All sequences with repetitive protein sequence (100% identity) but with different names were summarized at just one exemplary. When performing multiple sequence alignment, we were aware of a few sequences with different names but with the same exact composition.

2.3.3. Elimination of proteases without peptidase activity

SMART

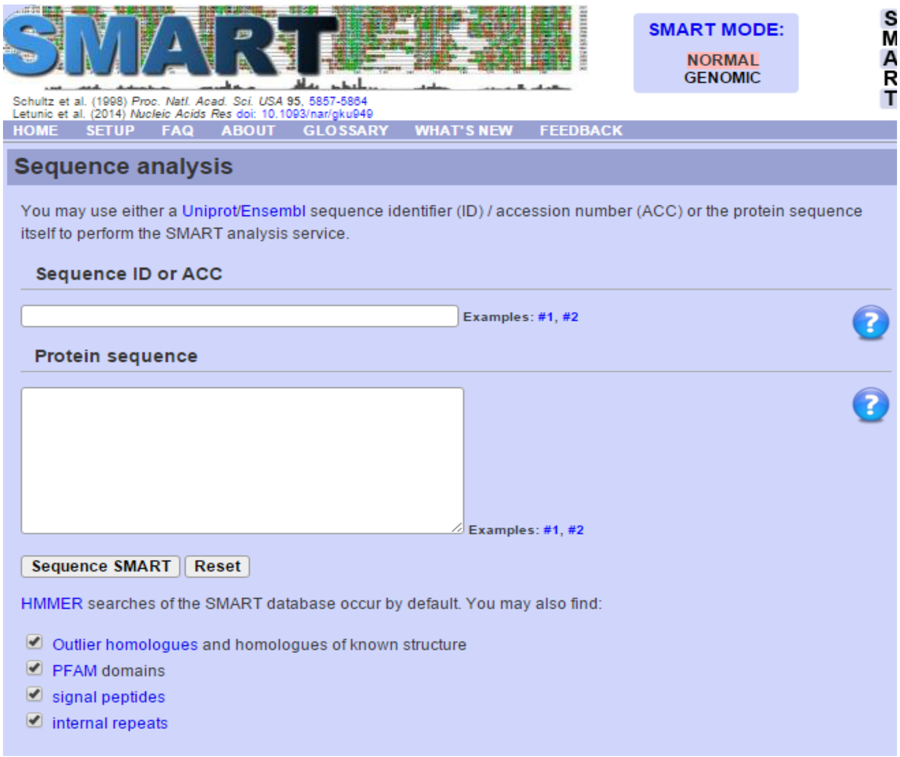
Each sequence was analysed in order to assess their family through SMART (Simple Modular Architecture Research Tool) analysis service.

SMART is an online tool used in the identification and annotation of genetically mobile domains and the analysis of domain architectures.(122)

The SMART database provides the identification and annotation of protein domain in protein sequences. In addition it allows the analysis and visualization of protein domain architectures.(123) The use of computational sequence analysis tools is essential for the annotation of novel genes or genomes, and the prediction of protein structure and function. The most recent release of SMART (2011) contains manually curated models of more than 1204 domains. SMART uses profile-hidden Markov models built from multiple sequence alignments to detect protein domains in protein sequences. This models describe the conservation of residues over entire domains or whole proteins.(124)

The main protein database in SMART consists of the complete Universal Protein Resource protein database combined with predicted proteins from all stable Ensembl genomes.(124) Data from SMART was used to create the Conserved Domain Database collection and it was distributed as part of the InterPro database.(125)

With just a Uniprot or Ensembl protein sequence identifier or accession number, SMART performs sequence analysis. However in this work we submitted the full protein sequences for analysis to identify the regions representative of the conservative domains (Figure 12).



Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* 95, 5857-5864
Letunic et al. (2014) *Nucleic Acids Res* doi: 10.1093/nar/gku949

Sequence analysis

You may use either a Uniprot/Ensembl sequence identifier (ID) / accession number (ACC) or the protein sequence itself to perform the SMART analysis service.

Sequence ID or ACC

Examples: #1, #2

Protein sequence

Examples: #1, #2

HMMER searches of the SMART database occur by default. You may also find:

☑ Outlier homologues and homologues of known structure
☑ PFAM domains
☑ signal peptides
☑ internal repeats

Figure 12 Sequence search using the SMART web interface

<u>InterPro</u>

Despite the reliability of SMART tool, all the protein sequences were also analysed and classified using InterPro version 51.0 to verify the results.

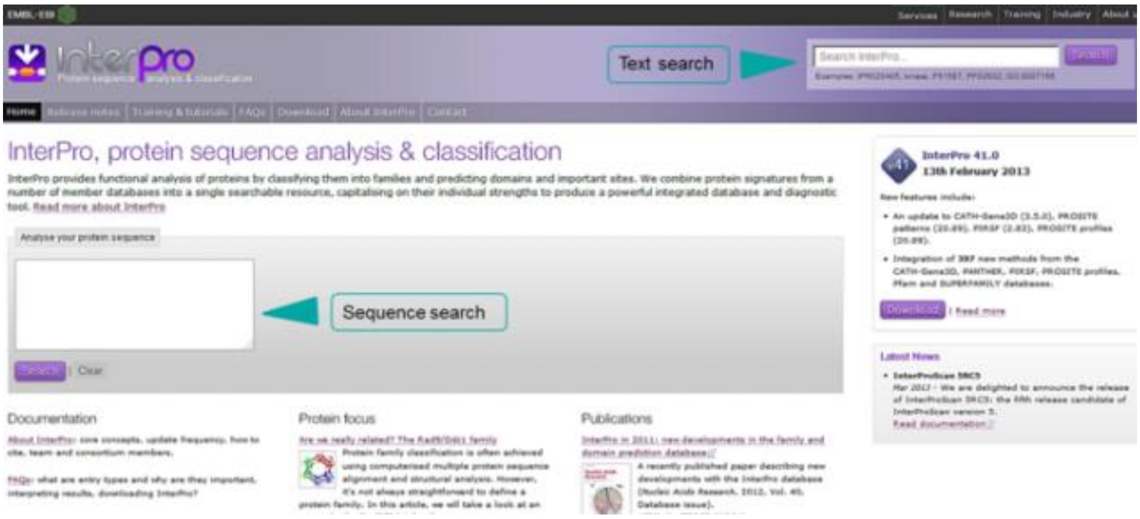
InterPro can be found at www.ebi.ac.uk/interpro

InterPro is an open-resource of protein families that provides functional analysis of protein sequences and is used for the automatic annotation of proteins.(126) Protein signatures databases describe protein families, functional domains and conserved sites within related groups of proteins. The signatures consist of models (simple types, such as regular expressions or more complex ones, such as Hidden Markov models) which describe protein families, domains or sites. Models are built from the amino acid sequences of known families or domains and they are subsequently used to search unknown sequences (such as those arising from novel genome sequencing) in order to classify them.(125)

This homology-based method rely primarily on identifying homologous sequences in other genomes and/or in public databases. A total of 11 databases, CAT-Gene3D, HAMAP, PANTHER, Pfam, PIRSF, ProDom, PROSITE, SMART, SUPERFAMILY and TIGRFAMS were combined in a single into one annotation tool- InterPro.

The InterPro database does not generate diagnostic models itself, but rather, groups one or more related member database signatures, and provides additional overarching functional annotations.(127) Each of the member databases of InterPro contributes towards a different niche, from very high-level, structure-based classifications through to quite specific sub-family classifications. So, with the different areas of specialization of each database, is produced a single resource that provides protein classification on multiple levels.(128)

The search in InterPro is carried out via the InterProScan analysis tool to predict the family and domain information for the amino acid sequence submitted.



InterPro, protein sequence analysis & classification

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. Read more about InterPro

InterPro 41.0
13th February 2013

New features include:

- An update to CATH-Gene3D (3.5.0), PROSITE patterns (20.89), PIRSF (2.83), PROSITE profiles (20.89).
- Integration of 387 new methods from the CATH-Gene3D, PANTHER, PIRSF, PROSITE profiles, Pfam and SUPERFAMILY databases.

Latest News

- InterProScan 5RC5
Mar 2013 - We are delighted to announce the release of InterProScan 5RC5; the fifth release candidate of InterProScan version 5.
Read documentation

Documentation

About InterPro: core concepts, update frequency, how to cite, team and consortium members.

FAQs: what are entry types and why are they important, interpreting results, downloading InterPro?

Protein focus

Are we really related? The Rad9/Ddc1 family
Protein family classification is often achieved using computerised multiple protein sequence alignment and structural analysis. However, it's not always straightforward to define a protein family. In this article, we will take a look at an

Publications

InterPro in 2011: new developments in the family and domain prediction database
A recently published paper describing new developments with the InterPro database (Nucleic Acids Research, 2012, Vol. 40, Database issue).

Figure 13 Sequence search using the InterPro web interface.

To search InterPro with a protein sequence, an amino acid sequence is inserted into the large box on the home page. InterPro will then attempt to assign the sequence to a protein family and identify any domains, repeats and sites. After the search, InterPro output includes protein family membership, sequence length, domains and detailed signature matches and the gene ontology (GO) terms. It is possible to collect essential information about the proteins in a secretome using gene ontology. This tool provides information regarding to biological process, molecular function and cellular localization.

According to the signatures matches, the sequences were filtered by type predicted protein family membership.

2.3.4. Elimination of non-peptidase homologues

To find the non-peptidase homologues MEROPS, the Peptidase Database, was used. (81)

MEROPS can be found at http://merops.sanger.ac.uk

MEROPS database is an information resource for peptidases (also termed proteases, proteinases and proteolytic enzymes) and for proteins that inhibit them. The MEROPS database uses a hierarchical, structure-based classification of the peptidases. In this, each peptidase is assigned to a Family on the basis of statistically significant similarities in amino acid sequence, and families that are thought to be homologous are grouped together in a Clan.(81,129)

According to this database some protease's families contain peptidases but also many homologues known as non-peptidase homologues. In other words, they are homologues that are not peptidases, usually because one of the active site residues has been replaced.(77)

The Summary page describing a given peptidase can be reached by use of an index under its source Organism. Searching for each organism (*Macrophomina phaseolina MS6* and *Neofusicoccum parvum*) there is an extensive list of all putative peptidase and non-peptidase homologue. So, it is easily noticeable which sequences are non-peptidase homologues. Since MEROPS do not associate the protein sequence with the GenBank accession number, to detect has compare the amino acid sequence of the non-peptidase homologues to all the selected protein sequences. Crossing that information was possible to eliminate the non-peptidase homologues.

2.4. Function Prediction/Domain Identification

Proteins have several functionally active regions called domains. To determine the functions of all known enzymes and identify these conservative domains computational sequence analysis tools were used. There are software programs which, given unannotated input sequence, attempt to identify one or more candidate motifs. Using that information, one can identify, shared and distinct sequence patterns, which facilitates an annotation of the uncharacterized sequence. For the annotation of each enzyme it is essential the prediction of domain structure and function.

SMART

SMART allows the identification and annotation of genetically mobile domains and the analysis of domain architectures.(130)

This database provides the identification and an extensive annotation of protein domains as well as the exploration of protein domain architectures. SMART allows finding protein domains within protein sequences among more than 1200 protein domains in the current version (2015). Besides that user interfaces allow searches for proteins containing specific combinations of domains in defined taxa. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues.(123)

SMART allowed us to classify the protein sequence by family for proteases and for glycoside hydrolases. Besides that, give us the protein sequence of the conservative domain of each enzyme.

2.5. Active Site Finding/Domain Analysis

Given a set of biological sequences, it is often a desire to identify the similarities shared between them. This information gives further information about the functionality, originality and the evolution of the species.

In order to locate the active sites in the functional domains a global multiple sequence alignment was performed. A variety of methods for isolating the motifs have been developed: all are based on identifying short highly conserved patterns within the larger alignment and constructing a matrix similar to a substitution matrix which reflects the amino acid or nucleotide composition of each position in the putative motif.

In this work, the motif searching was performed for each protease family. The MSA generated blocks of gapped regions that where thoroughly reviewed.

2.5.1. Multiple Sequence Alignment (MSA)

Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides.(131) Multiple sequence alignments were performed using the amino acid sequences of the domain region selected (2.4-Function prediction/Domain identification. Conserved domains contain conserved sequence patterns or motifs, which allow for their detection in polypeptide sequences.

Clustal X 2.1

The program used to perform multiple alignments was Clustal X 2.1 (ClustalX is a version of ClustalW with a graphical user interface).

Clustal is known to be a fast and light program that can align a large number of sequences and it has been applied for multiple sequence alignment and phylogenetic tree construction.(132)

The method Clustal used to construct the alignment is called pairwise progressive sequence alignment and it is carried out in 3 stages. First, the most similar sequences are aligned and then successively less related sequence pairs or groups are aligned. Pairwise sequence alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid). Then, a dendrogram (guide tree) of the sequences is constructed according to the pairwise similarity of the sequences describing the approximate groupings of the sequences by similarity.(133)

Finally a multiple sequence alignment is constructed by aligning sequences in the order, using the dendrogram as a guide. Progressive alignment results are dependent on the choice of "most related" sequences and thus can be sensitive to inaccuracies in the initial pairwise alignments.(133)

To run ClustalX 2.1 is a simple multiple steps process.

Sequence input; three or more sequences can be entered directly in the program in GCG, FASTA, EMBL, PIR, NBRF or UniProtKB/Swiss-Prot format. There is a limit of 500 sequences or 1MB of data. All non-alphabetic characters (spaces, digits, punctuation marks) are ignored except "-" which is used to indicate a GAP.

In this work a file containing the desired sequences in FASTA format was uploaded and used as input for the multiple sequence alignment. The sequences were in protein format (by default) but this tool also aligns nucleotide (DNA/RNA) sequences.

Multiple Sequence Alignment Options; Slow-accurate alignment parameters do not have any effect on the speed of the alignments. They are used to give initial alignments which are then rescored to give percent identity scores. These scores are the ones which are displayed on the screen (converted to distances for the trees). To perform

- Protein Weight Matrix » The multiple alignment protein sequence comparison matrix series used to score the alignment Blosum (for protein only).
- Gap Opening [0-100] » Multiple alignment penalty for the first residue in a gap.
- Gap Extension [0-100] » Multiple alignment penalty for each additional residue in a gap.

Several parameters combination were used in order to improve the multiple sequence alignments. In Table 2 lists some experimental parameters used to perform different multiple sequence alignment to improve their quality.

Table 2 Experimental parameters used on MSA.

Protein Weight Matrix	**Gap Opening [0-100]**	**Gap Extension [0-100]**
Gonnet (Default value)	10 (Default value)	0.2 (Default value)
BLOSUM	100	10
BLOSUM	100	0
Gonnet.	100	0

The following parameters control the final multiple alignment. Each step in the final multiple alignment consists of aligning two alignments or sequences. This progressive process is based on the branching order in the guide tree. In this step was used all the default parameters.

- Delay Divergent sequences (%) » Switch delays the alignment of the most distantly related sequences until after the most closely related sequences have been aligned. Sequences that are less identical than this level to any other sequences will be aligned later. Default value 30
- Use negative matrix » In the weight matrices, you can use negative as well as positive values if you wish, although the matrix will be automatically adjusted to all positive scores, unless the negative matrix option is selected. Default parameter Off

Alignment display; the alignment is displayed on the screen with the sequence names on the left hand side.

A line above the alignment is used to mark strongly conserved positions

In this program it is possible to:

- Realign selected sequences » it is used to realign badly aligned sequences in the alignment. Sequences can be selected by clicking on the sequence names. The unselected sequences are then 'fixed' and a profile is made including only the unselected sequences. Each of the selected sequences in turn is then realigned to this profile.
- Realign selected sequence range » it is used to realign a small region of the alignment. A residue range can be selected by clicking on the sequence display area. The new alignment of the range is pasted back into the full sequence alignment.

Output format

Alignment » Output format options.

The output format can be chosen among several formats: CLUSTAL, NBRF/PIR, GCG/MSF, PHYLIP, NEXUS, GDE or FASTA. All sequences are written in a single file. Users can also choose to include the residue range numbers by appending them to the sequence names.

T-coffee

T-Coffee (Tree-based Consistency Objective Function or alignment Evaluation) is a multiple sequence alignment software that uses a progressive approach whose algorithm has been constantly improved.

T-Coffee progressive alignment algorithm has 3 main stages: pre-processing a library, pairwise alignments, and refining the library. It generates a library of pairwise alignments to guide the multiple sequence alignment. T-Coffee produces several intermediate alignments that can be used and the algorithm can also stop at that point. T-coffee uses the primary library of both global and local alignments to measure distance between pairwise sequences or combines multiple alignments from other programs. It has advanced features to evaluate the quality of the alignments and some capacity for identifying occurrence of motifs. (134)

This tool provides a considerable improvement in accuracy so that even it is slower than commonly used alternatives but T-Coffee is considered to be the recommendation when you want to produce high accurate MSAs. In the last versions, 3D-Tcofee can use structure information from Protein Data Bank.(134)

To perform the T-Coffee alignment, the Unipro UGENE 1.18, a open-source cross-platform bioinformatics software, was used. The file with all sequences was uploaded to the program in txt format. However, the alignment can also be done using the web tool, a simple multiple steps process. Three or more sequences to be aligned can be entered directly in the program in GCG, FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP or UniProtKB/Swiss-Prot format.

All non-alphabetic characters (spaces, digits, punctuation marks) are ignored except "-" which is used to indicate a GAP. Most multiple sequence alignment methods try to minimize the number of insertions/deletions (gaps) and, as a consequence, produce compact alignments. GAP (-) symbols inserted such that maximizing the similar regions across the sequences. In biological perspective, the gaps represent residue symbols being deleted from the sequences during the course of evolution. In general, when a gap is inserted into a sequence a penalty is applied to the alignment. The weight of the penalty depends on the location of the insertion.(135)

2.6. Phylogenetic Analysis

To represent the evolution of a set of proteins it is required to arrange them in a phylogenetic tree.

In this work maximum likelihood and neighbour joining phylogenetic trees were constructed. Both methods were developed using MEGA (Molecular Evolutionary Genetics Analysis) tool. This software was developed with the goal to provide statistical analyses of DNA and protein sequence data from an evolutionary standpoint. Through the comparison of DNA and protein sequences is construct the molecular evolutionary patterns of genes, genomes, and species over time is determined.(136)

We compared Neighbor-Joining (automatically generated trees) and Maximum Likelihood.

An evolutionary tree by the Neighbor-Joining (NJ) algorithm, can be constructed by this tool, using a matrix of pairwise distances. Firstly, it is necessary to measure all pairwise distances of the given set of sequences and derive a distance matrix from this. The distances can be gathered from a multiple sequence alignment by comparing each pair of aligned sequences in the alignment. From this distance matrix, a phylogenetic tree can be computed and display the evolutionary history of a set of sequences. The sequences are represented by the leaves of the tree and internal nodes are putative ancestor sequences.(137)

In reality, Neighbor Joining is not a phylogenetic method, but a phenetic one. It establish relationships between sequences according to their genetic distance, without taking into account an evolutionary model. Thus, ancestry is never considered. That’s why a true phylogenetic tree - Maximum Likelihood (ML) was additionally construct. ML is a statistical method of estimating unknown parameters of a probability model given data.(138) One of the advantages of this methods is the statistical flexibility by allowing varying many parameters, including rates of evolution, differential transformation costs, and, even, the tree itself.(139) In fact, the method claims that evolution at different sites and along different lineages must be statistically independent. It is reliable for reconstructing sequence histories because it uses a more complex evolution mode. All possible combinations of tree topology and branch length are generate leading this method well suited to the analysis of distantly related sequences.(140)

For this work, phylogenetic trees were constructed using both methods: maximum likelihood and neighbour joining.

To construct the phylogenetic tree using MEGA6, several steps were necessary.

a. Input alignment file

Convert To MEGA Format (Main File Menu)

This item allows the user to choose the file and/or the format that the user would like to use to convert a given sequence data file into a *MEGA* format. It converts the data file and displays the converted data in the editor.

Files written in a number of popular data formats can be converted into *MEGA* format. MEGA supports conversion of CLUSTAL, NEXUS (PAUP, MacClade), PHYLIP, GCG, FASTA, PIR, NBRF, MSF, IG, and XML formats.

b. Find best protein model

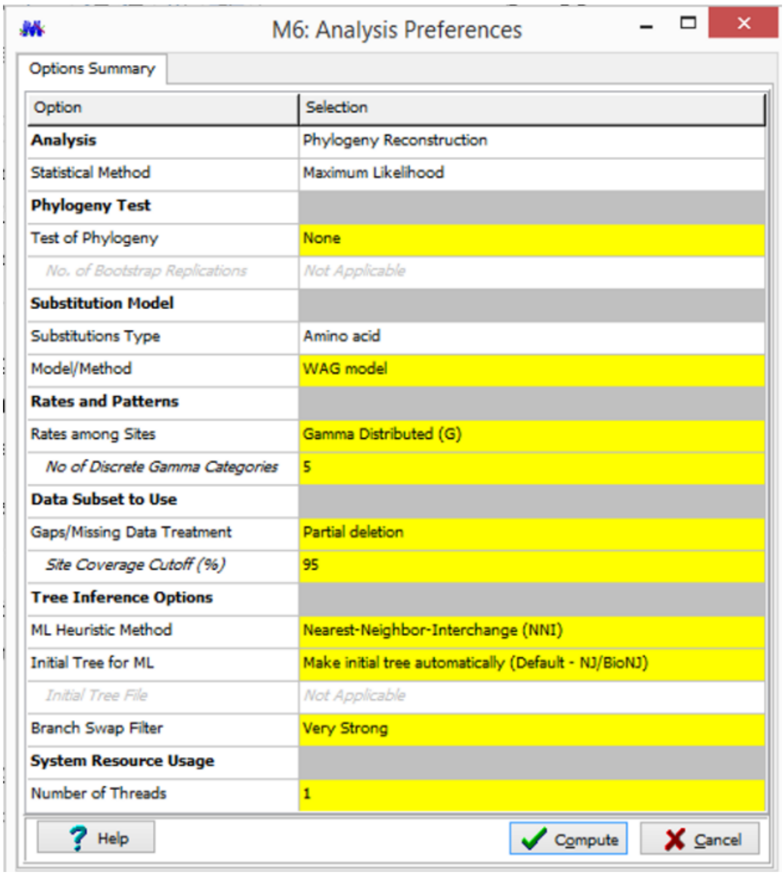
Multiple alignment uses a variety of substitution models to correct for multiple changes at the same site during the evolutionary history of the sequences. MEGA6 provides a feature that chooses the best model for the user. This step is only available for maximum likelihood method.

From the Models menu choose Find Best DNA/Protein Models (ML). Click the Compute button to start the run. *Models* can take quite a while to consider all the available models, but a progress bar shows exactly how things are coming along. The list of evaluated substitution models along with their relative fits, number of parameters and estimates of evolutionary parameters.

When completed, a window appears which lists the models in order of preference. The WAG + G + I model was the preferred model to all datasets. This information allows you to check the robustness of the estimates of evolutionary parameters under different models of substitutions and assumptions about the distribution of evolutionary rates among sites.

The Gamma distance improves upon the Poisson correction distance by taking care of the inequality of the substitution rates among sites.(141) The WAG (Whelan And Goldman) is an empirical model of globular protein evolution. It was estimated from 182 protein families (provided by David Jones) using a maximum likelihood procedure that takes into account the evolutionary relationships within each family. WAG is implemented in many widely used programs for phylogeny.(140)

The Analysis Preferences Dialog is used for specifying the substitution model to use as well as the distribution of rates for ML based ancestral sequence inference.



Option	Selection
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Amino acid
Model/Method	WAG model
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
No of Discrete Gamma Categories	5
Data Subset to Use	
Gaps/Missing Data Treatment	Partial deletion
Site Coverage Cutoff (%)	95
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File	*Not Applicable*
Branch Swap Filter	Very Strong
System Resource Usage	
Number of Threads	1

Figure 14.The Analysis Preferences Dialog

c. Construct phylogenetic tree

The evolutionary history was inferred by using the Maximum Likelihood method based on the WAG matrix- based model and no bootstrap value. Evolutionary analyses were conducted in MEGA 6.06(136). The trust level assigned in the various branches is determined by bootstrap replications of 500. Calculating bootstrap is important to estimate the reliability of a tree. In computational phylogenetics, boostrapping refers to creating multiple pseudo-alignments from randomly chosen columns from the original alignment until the random alignments have the same length as the original alignment. For each random alignment a tree is calculated with the same parameters as the tree calculated for the original alignment. That procedure is commonly used for providing confidence to branches in phylogenetic trees. Bootstrapping values are typically presented from 100 to 2000 repeated calculations, and the number of cycles increases the calculation time. Bootstrap values of >70% is recommended.

From the Phylogeny menu choose Construct/Test Maximum Likelihood Tree/ Construct/Test Neighbour-Joining Tree. A preferences dialog similar to that in Figure will appear.

Below are indicated the tree parameters options used to construct the phylogenetic trees presented in the Results and Discussion.

Substitution Model: Substitutions Type - Amino acid; Model/Method - WAG model

Rates and Patterns: Rates among Sites - Gamma Distributed (G); No of Discrete Gamma Categories - 5

Data Subset to Use: Gaps/Missing Data Treatment - Partial deletion; Site Coverage Cutoff (%) – 95

Tree Inference Options: ML Heuristic Method - Nearest-Neighbor-Interchange (NNI); Initial Tree for ML - Make initial tree automatically (Default - NJ/BioNJ)

Branch Swap Filter - Very Strong

After building the tree, the Tree Explorer displays the evolutionary tree based on the options used to compute or display the phylogeny. The Tree Explorer also allows customization of the tree display.

3. Results and discussion

3.1. Genome analysis

Most of the species of the *Botryosphaeriaceae* cause disease symptoms such as die-back and cankers on numerous woody and non-woody hosts. In this work, the enzymes distribution of available sequenced genomes of *Macrophomina phaseolina, Neofusicoccum parvum* and *Diplodia corticola* were described. During the infection mechanism is expected that phytopathogenic fungi express high amounts of hydrolytic and oxidative enzymes, such as proteases and glycoside hydrolases (namely xylanases and endoglucanases).

In order to determine the sequence of a genome, several complex steps are required, and eventually errors in the sequence can be introduced. The source of such errors be the technique used: the sequencing process available, the produced signal, the fragmentation of the DNA into smaller fragments or even the reassembling of the small DNA fragments into a continuous genome. The nature of the sequence can also obstruct the accuracy of the sequencing. Problems such as sequence repetitions of the same base or tandem repeats, secondary structure formation or sequences which cannot be cloned easily into bacteria for amplification. In order to increase the reliability of the sequenced genome, the whole genome has to be sequenced multiple times. The same sequence has to be covered multiple times, usually at least 3-6x for a draft genome, and at least 7-10x for a full genome.(142,143)

Since the focus of this study is species of the *Botryosphaeriaceae* family, so were sought genomes of all species belonging to this family. Only seven species of *Botryosphaeriaceae* family are covered by the 1000 Fungal Genome project. The aim of this project is to provide genomic information for every specie across the Kingdom Fungi with the objective to significantly advance genome-enabled mycology. This sequencing project is supported by the Joint Genome Institute (JGI). Students and postdocs provide high quality DNA and RNA samples to JGI for sequencing. Once completed the genome assemble and annotation, will have open access for researchers use at JGI Mycocosm site.(144)

There are underway seven genome projects in which 4 of them are still incomplete. In the fungi *Aplosporella prunicola, Lasiodiplodia theobromae,* *Sphaeropsis sapinea* and *Botryosphaeria dothidea* the determination of the DNA genome sequence is still being determined. The *Diplodia pinea* project status is defined as permanent draft. Even though the genome of *D. pinea* had been recently sequenced by Forestry & Agricultural Biotechnology Institute, University of Pretoria, it was not described any gene that encodes proteases or glycoside hydrolases. The two remain species (out of seven), *Neofusicoccum parvum* and *Macrophomina phaseolina,* have their genome completely sequenced and available. Besides that, *Diplodia corticola* recently sequenced genome it will also be analysed, even though is still not available in online databases.

The characteristics of the genome of *M. phaseolina* MS6*, N. parvum* UCRNP2 and *D. corticola* can give us insights into the complex set of proteins factors that these pathogens may use in the phypathogenesis.

In GenBank *Macrophomina phaseolina* genome overview presents 7 different Genome Assembly and Annotation reports but only one of *N. parvum* available. This can suggest that the information (genome size, protein sequences, sequences description, number of proteins) deposited about *M phaseolina* is more complete and with fewer errors. The genome sequencing platform used for *M. phaseolina* genome was the combination of 454 and Illumina platforms (45), while for *N. parvum* was Illumina HiSeq 2000 platform (33) and for *D. corticola* was used Illumina paired-end sequencing technology. The large number of transposons in the *M. phaseolina* genome suggests that they could be the primary mechanism for mutagenesis and gene duplications, which may promote the ability of *M. phaseolina* to infect new plant species. Another interesting fact is that a high percentage of the genes encoded by the genome have significant similarities with genes involved in pathogen-host interactions.(45) In all the 3 organisms, it was sequenced the draft genome and *Diplodia corticola* it has the smaller genome size comparing to *N. parvum* and *M. phaseolina*.

Comparing the amount of putative proteins identified as potentially secreted between these organisms, the results were very similar. According to the literature, *Macrophomina phaseolina* holds the highest amount of proteins (45) than *Neofusicoccum parvum* (35) and *Diplodia corticola* genome.

According to a recent study (2012) *M. phaseolina* has the highest number of genes when compared to other fungi regarding the glycosidase and secondary metabolite backbone genes. That data suggested that this could be a strategy to overcome the host plant defence response by using various secondary metabolites.(45)

3.2. Selected sequences

All the amino acid sequences of proteases and glycoside hydrolases of *Macrophomina phaseolina, Neofusicoccum parvum* and *Diplodia corticola* were selected Analysing each sequence individually, several sequences were eliminated based on their Protein database definition (for example, Peptidase S41 [*Macrophomina phaseolina* MS6])

Besides that, non-peptidase homologues, redundant sequences and proteases without peptidase activity were eliminated.

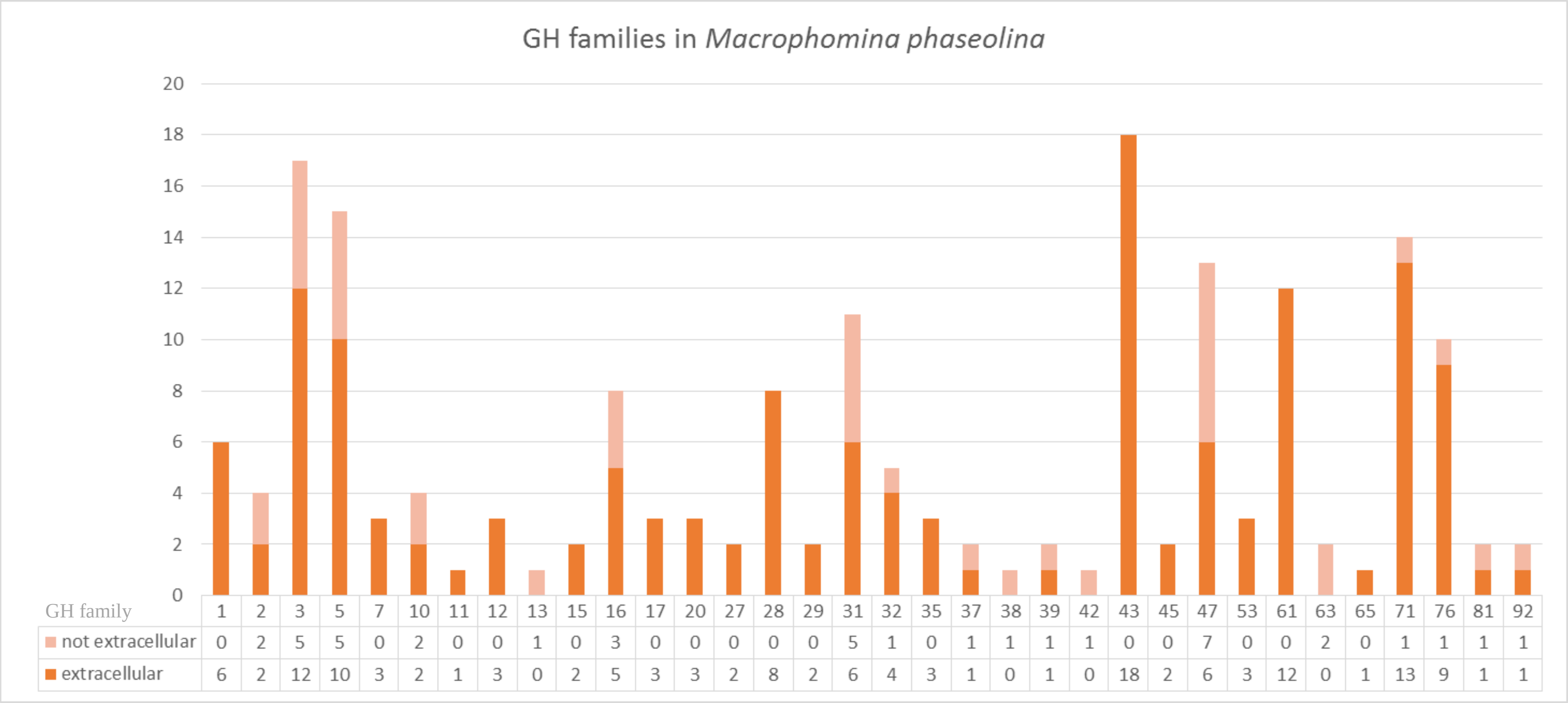
The sequences not considered are listed in Annexe 5.1, organised by organism.

3.3. Assembly protein sequences into families

The cellular localisation of all the protein sequences of the three organisms were predicted using BaCelLo tool. After selecting the sequences corresponding to extracellular enzymes, the enzyme family has confirmed by SMART and InterPro version 51.0.analysis service.

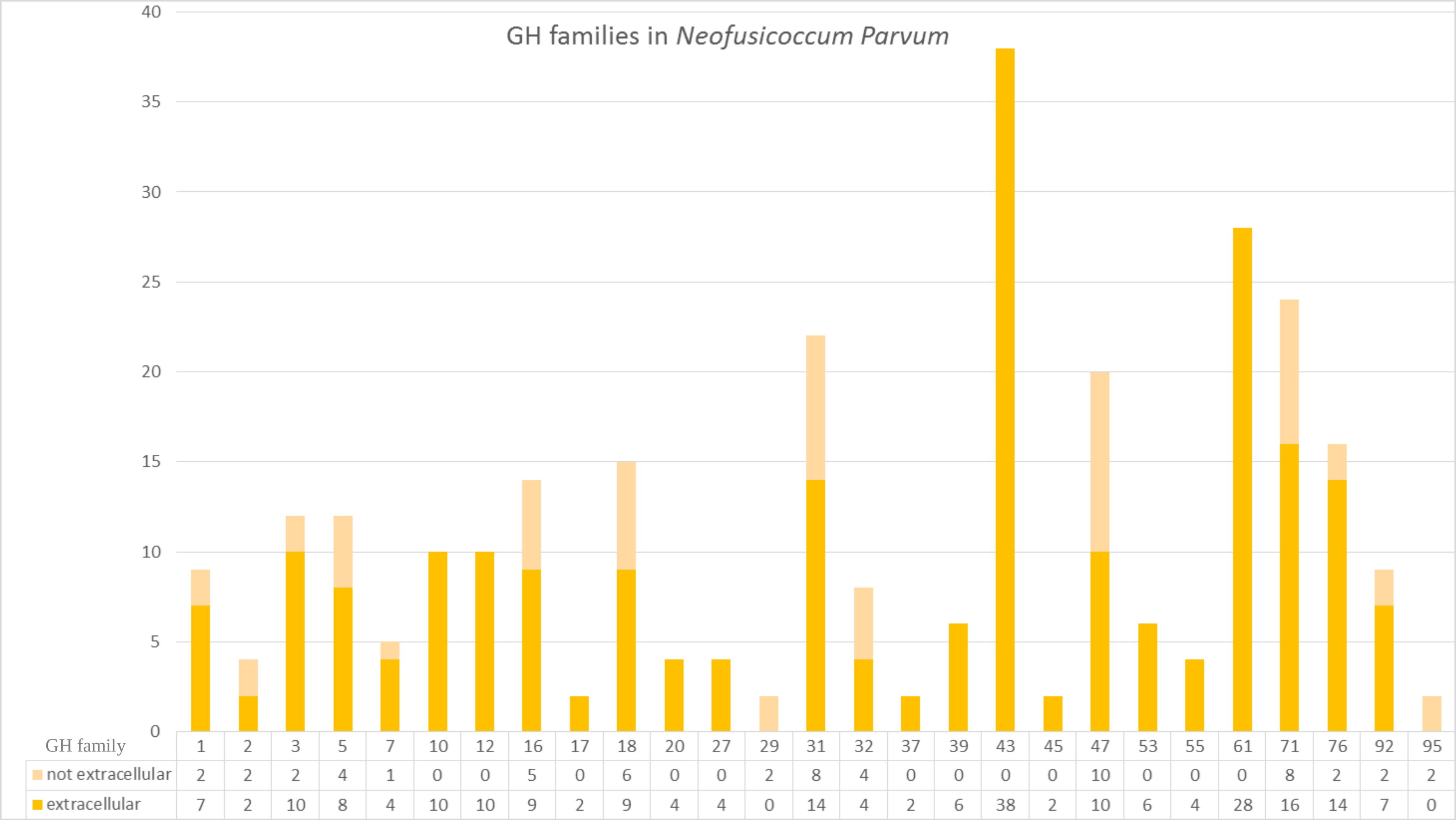
3.3.1. Glycoside hydrolases (GH)

The complete genomes of *Macrophomina phaseolina* and *Neofusicoccum parvum* were released into GenBank database (NCBI). The amino acid sequences of all glycoside hydrolases of both organisms were selected to analysis. Using BaCelLo, a predictor for the subcellular localisation, we selected the extracellular proteins. Then we proceeded to “noise reduction” step as described in *Materials and Methods*. The following figures (Figure 16, Figure 16 and Figure 17) show the number of sequences of each family of GH (according to CAZy classification) (extracellular and not extracellular) from each organism (*M. phaseolina, N. parvum* and *D. corticola)*.



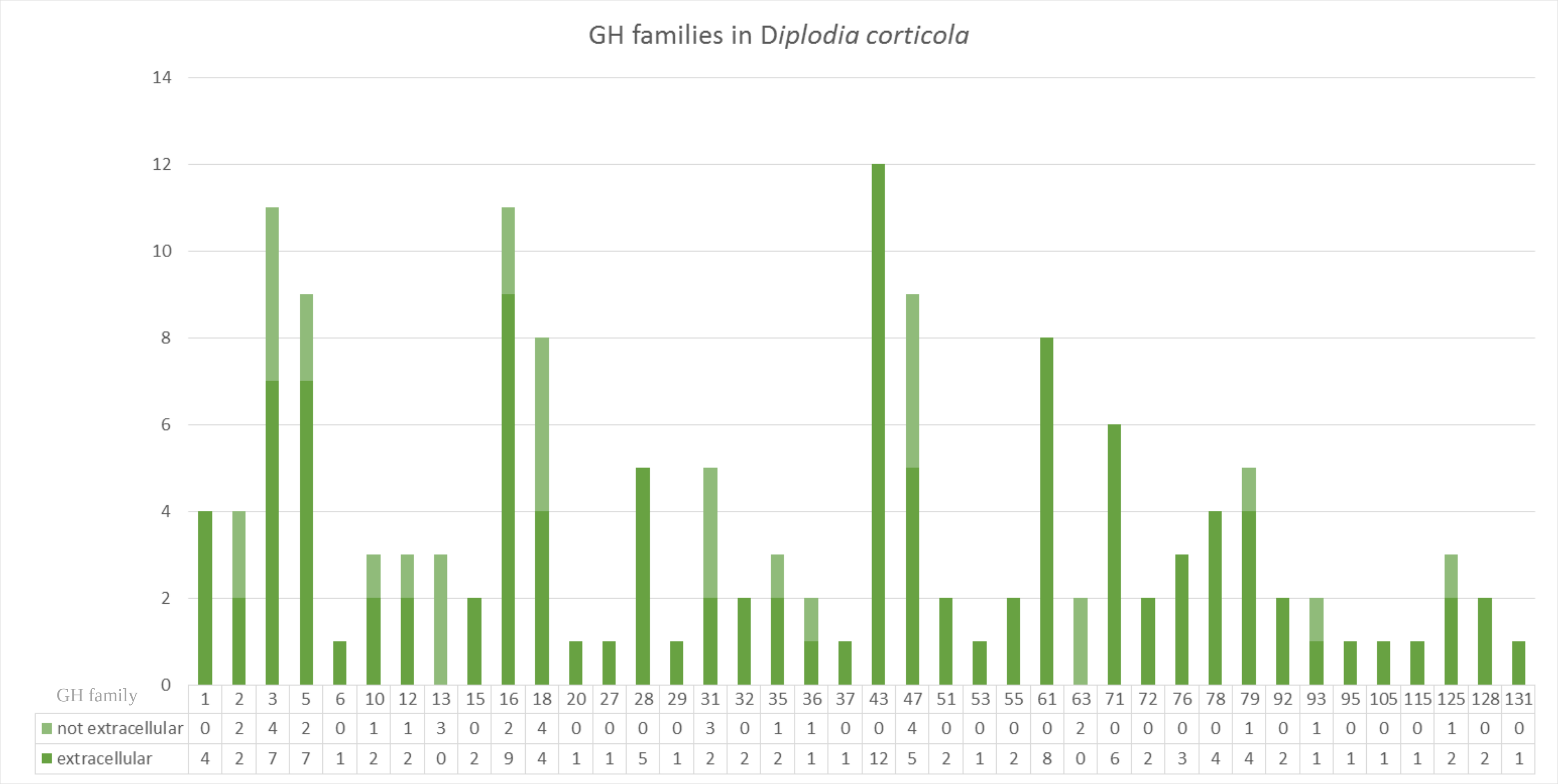
GH family	1	2	3	5	7	10	11	12	13	15	16	17	20	27	28	29	31	32	35	37	38	39	42	43	45	47	53	61	63	65	71	76	81	92
■ not extracellular	0	2	5	5	0	2	0	0	1	0	3	0	0	0	0	0	5	1	0	1	1	1	1	0	0	7	0	0	2	0	1	1	1	1
■ extracellular	6	2	12	10	3	2	1	3	0	2	5	3	3	2	8	2	6	4	3	1	0	1	0	18	2	6	3	12	0	1	13	9	1	1

Figure 15. Number of sequences of each glycoside hydrolases families in *Macrophomina phaseolina*



GH family	1	2	3	5	7	10	12	16	17	18	20	27	29	31	32	37	39	43	45	47	53	55	61	71	76	92	95
▨ not extracellular	2	2	2	4	1	0	0	5	0	6	0	0	2	8	4	0	0	0	0	10	0	0	0	8	2	2	2
■ extracellular	7	2	10	8	4	10	10	9	2	9	4	4	0	14	4	2	6	38	2	10	6	4	28	16	14	7	0

Figure 16. Number of sequences of each glycoside hydrolases families in *Neofusicoccum parvum*



GH family	1	2	3	5	6	10	12	13	15	16	18	20	27	28	29	31	32	35	36	37	43	47	51	53	55	61	63	71	72	76	78	79	92	93	95	105	115	125	128	131
not extracellular	0	2	4	2	0	1	1	3	0	2	4	0	0	0	0	3	0	1	1	0	0	4	0	0	0	0	2	0	0	0	0	1	0	1	0	0	0	1	0	0
extracellular	4	2	7	7	1	2	2	0	2	9	4	1	1	5	1	2	2	2	1	1	12	5	2	1	2	8	0	6	2	3	4	4	2	1	1	1	1	2	2	1

Figure 17. Number of sequences of each glycoside hydrolases families in *Diplodia corticola*.

Fungi are unable to create organic matter from inorganic materials. Thus their nutrition comes, directly or indirectly, from plants. The degree of fungal-plant specificity is, however, very variable. The mycelium release extracellular digestive enzymes that degrade organic molecules. Then, hyphae absorb the digestion products, using them as energy source and raw material for their survival and growth. In fact, the absorptive mode of nutrition of these organisms has resulted in the evolution and secretion of a variety of enzymes able to catabolise environmental complex organic polymers, such as cellulose, chitin and proteins into smaller constituents.(145)

In this study, we consider a set of fungal genes encoding all characterized GH family enzymes. About 14 exclusive glycoside hydrolase families can be found in these fungus: GH6, 18, 36, 51, 73, 78, 79, 93, 95,105, 115, 125, 128 and 131. In *M. phaseolina* were described 4 exclusive GH families: 38, 42, 65 and 81. And finally, in *N. parvum* no glycoside hydrolase family was found exclusively in this fungus.

Analysing the genes that encodes extracellular proteins of these three organisms is clear the diversity of glycoside hydrolase families in *Diplodia corticola.* Such enzymes may function in the penetration of plant tissues and for the acquisition of nutrients from plant constituent. Furthermore, they may be involved as effector molecules that cause host-plant defence responses. The high genetic diversity suggests that *D. corticola* has a great capacity for adaptability during its interaction with host species.

Besides the analysis of all glycoside hydrolases, we focused on the ones with xylanolytic and endoglucanolytic activity. To know the GH families with the activities of interest, the Carbohydrate-Active Enzymes (CAZy) database was used. Then, to select de protein sequences of enzymes with xylanolytic and endoglucanolytic activity in *M. phaseolina, N. parvum* and *D. corticola* genomes, NCBI database was used. CAZy database groups' glycoside hydrolases into families based on their amino acid sequence and, as a result, a single family may exhibit a variety of activities among its members. Therefore, a predicted specific activity is found in several glycoside hydrolase families.

Described below are the glycoside families corresponding to the activities in focus in this work (xylanase and endoglucanase). Xylanases and endoglucanases have been classified into glycoside hydrolase (GH) families (http://www.cazy.org) based on sequence similarities of the catalytic domain.

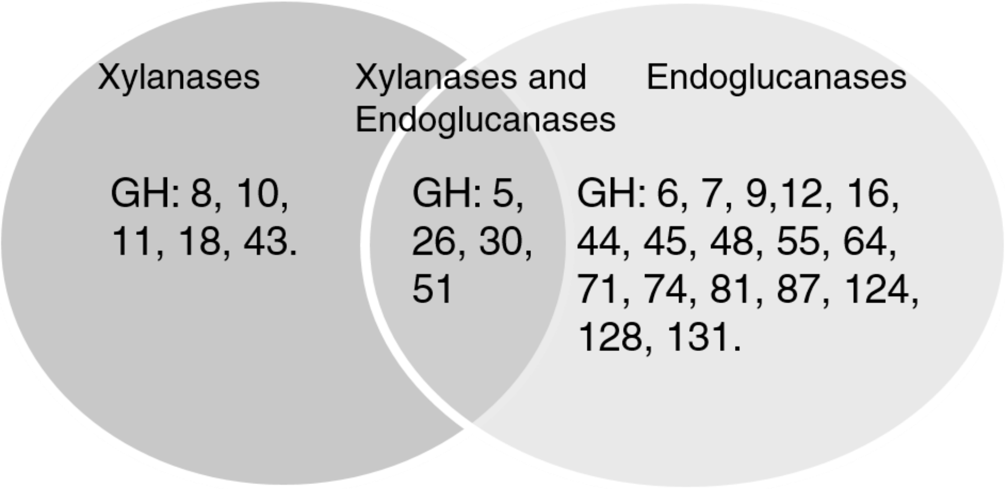
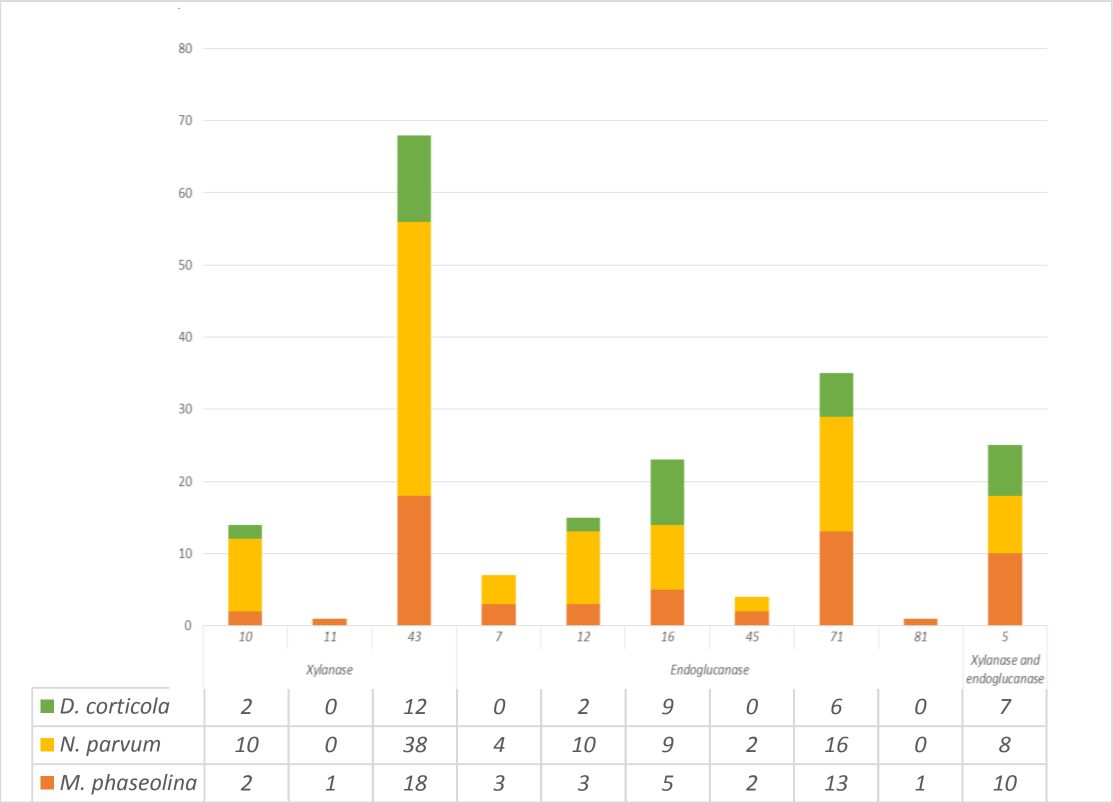

Xylanases
Xylanases and Endoglucanases
Endoglucanases
GH: 8, 10, 11, 18, 43.
GH: 5, 26, 30, 51
GH: 6, 7, 9,12, 16, 44, 45, 48, 55, 64, 71, 74, 81, 87, 124, 128, 131.

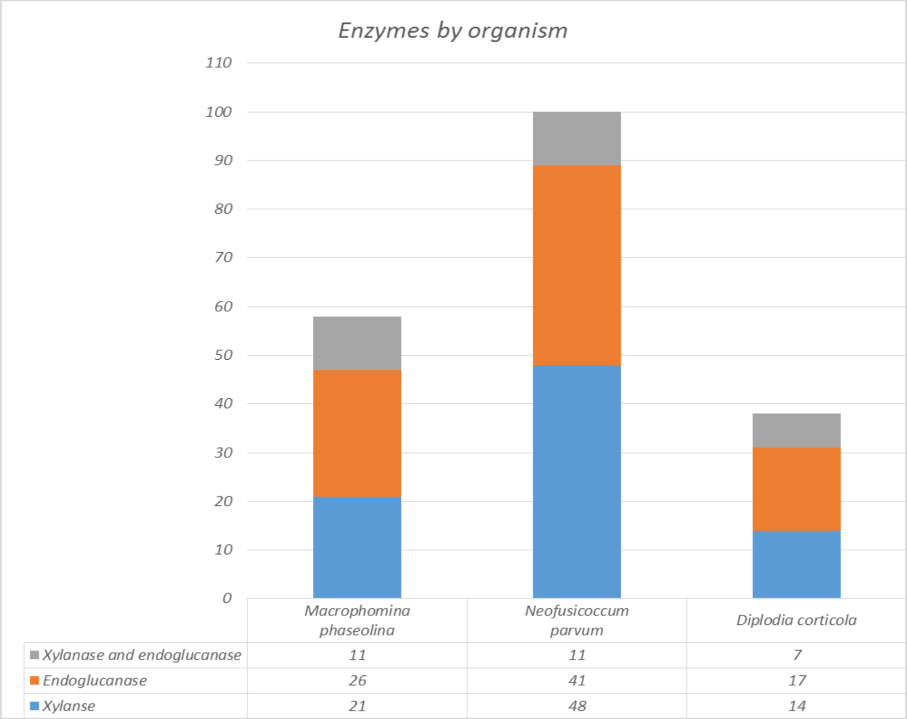
Figure 18. Glycoside hydrolases families that have enzymes with xylanolytic activity, endoglucanolytic activity and xylanolytic and endoglucanolytic activity.

However, not all these families were found among the described proteins. And, in some cases, the enzymes of some GH family were all “not extracellular” (nuclear, cytoplasmic or mitochondrial). In the following graphics are presented the extracellular enzymes of *N. parvum, M. phaseolina* (Figure 16), and *D. corticola* (Figure 17) with xylanolytic and endoglucanolytic activity are presented.



	10	11	43	7	12	16	45	71	81	5
	Xylanase			Endoglucanase						Xylanase and endoglucanase
🟩 *D. corticola*	2	0	12	0	2	9	0	6	0	7
🟨 *N. parvum*	10	0	38	4	10	9	2	16	0	8
🟧 *M. phaseolina*	2	1	18	3	3	5	2	13	1	10

Figure 19. Number of sequences of glycoside hydrolase families grouped by xylanase and endoglucanase activity in *M. phaseolina, N. parvum* and *D. corticola*



	Macrophomina phaseolina	Neofusicoccum parvum	Diplodia corticola
■ Xylanase and endoglucanase	11	11	7
■ Endoglucanase	26	41	17
■ Xylanse	21	48	14

Figure 20. Number of protein sequences of xylanases and endoglucanases grouped by organism (*M. phaseolina, N. parvum* and *D. corticola*)

Phytopathogenic fungi secrete a cocktail of hydrolytic enzymes (including glycoside hydrolases) for degrading the plant cell wall and penetrating into the host tissue.(45) The enzymes involved in the biodegradation of cellulose are: endoglucanases (hydrolyse cellulose to glucooligosaccharides), cellobiohydrolases (release cellobiose from crystalline cellulose) and glucosidases (degrade the oligosaccharides to glucose). And the biodegradation of the xylan backbone (the major hemicellulose portion of cell walls) depends on two classes of enzymes: exoxylanases and endoxylanases that are able to cleave the xylan backbone into smaller oligosaccharides.(146,147) The set of enzymes (such as glycoside hydrolases, polysaccharide lyases and cutinases), identified as potentially secreted by a fungus might have a role during the colonisation of host tissues.(33)

All the ascomycete wood pathogens showed a wider array of enzymes that target cellulose and hemicellulose. Since these are wood-colonizing fungi, we expect their genomes to include a range of genes encoding for wood-degrading enzymes, especially *N. parvum*, which colonizes grapevine wood more rapidly than most trunk pathogens.(148) In fact, when compared to the others fungi, a greater number of proteins, such as xylanases and endoglucanases, were found in *N. parvum* genome. The number of GHs possessed by this plant pathogenic fungi is consistent with the importance of carbohydrate degradation for plant invasion.(45) There are 48 sequences of xylanases, 41 sequences of endoglucanases and 11 sequences that belong to glycoside hydrolase families with endoglucanase and xylanases enzymes. These data suggest that the *N. parvum* genome encodes a large repertoire of pathogenicity- associated genes.

Among the xylanases glycoside families, GH10 and 11 xylanases have distinct three-dimensional structures, mechanisms of action and, therefore distinct, substrate specificity to xylan. Despite these G10 and 11 being the most abundant families, in these organisms just a few sequences were found.(149) A high number of GH10 xylanases are usually found in plant pathogen or saprobic fungi, as Aspergillus species.(150) Xylanases are predominantly found in GH43 family, but this family also contains arabinanase, galactan 1,3-β-galactosidase and exo-α-1,5-L-arabinofuranosidase. This glycosil family is commonly found in fungi and because it is an inverting mechanism, they have a broad substrate specificity. It should be noted that in several plant cell wall degrading organisms there has been an increase amount of GH43 family enzymes described, which may reflect their wide range of specificities.

Even though there are some GH families with xylanases and endoglucanases members (GH: 5, 26, 30, 51), it was only found in the genome of *M. phaseolina, N. parvum* and *D. corticola* genes that encodes for GH5. This family is one of the largest of all GH families and contains enzymes acting on a wide variety of substrates. Most of them endoglucanases and mannanases, but other activities include xylanases, endoglycoceramidase and chitosanase. In an Ascomycota phylum member, *Aspergillus nidulans*, has already been described genes that encode for GH5 β-mannanases.(151) Actually, GH5 family has been found to contain catalytic domain that exhibit xylanase activity but, until recently, not in fungi. It will be interesting for future studies to understand if the protein sequences are mannanases or whether it is xylanases.

3.3.2. Proteases

Proteases with a significant similarity in amino acid sequence are grouped into families. The MEROPS groups sequences into a peptidase family with homologous domains with the active site residues. When comparing sequences to group them into peptidase families, it is important to consider only the domain with the active site residues. No peptidase family exists containing peptidases of different catalytic types. A family can contain a single peptidase if no homologues are known and a single gene product can contain more than one peptidase each assigned to a different family.(70)

The results of grouping sequences into MEROPS families through a screening procedure are illustrated in Table 3, which consists of 5 different protease families.

Table 3. Number of extracellular proteases of each MEROPS protease's family of *M. phaseolina, N. parvum* and *D. corticola.*

	Macrophomina phaseolina	*Neofusicoccum parvum*	*Diplodia corticola*	Total
Serine	38	44	31	113
Aspartic	16	28	23	67
Metallo	21	14	28	63
Cysteine	1	1	0	2
Glutamic	2	0	0	2
Total	78	87	82	

In this work, peptidases sequences were analysed in SMART and InterPro tool and grouped into families according to their functional domains. Considering the three organisms, a total of 69 protein sequences were identified as aspartic peptidases, distributed among A1 and A4 families. A total of 109 protein sequences were identified as serine peptidases, distributed among a number of serine families: S8, S9, S10, S15, S16, S28, S41 and S53. A total of 64 protein sequences were identified as metallo peptidases, distributed among a number of metallo families: M12; M14; M16; M22; M24; M28; M35; M41; M43; M48; M54; M79. (Annexe 5.2) A total of 2 protein sequences were identified as cysteine peptidases, C13, C54. And finally, only two protein sequences was identified as been a glutamic or cysteine peptidase: G1.

3.3.2.1. *Macrophomina phaseolina*

The GenBank accession numbers of the sequences of all secreted proteases of *Macrophomina phaseolina* are listed in Annexe 5.3, organised by MEROPS proteases' families.

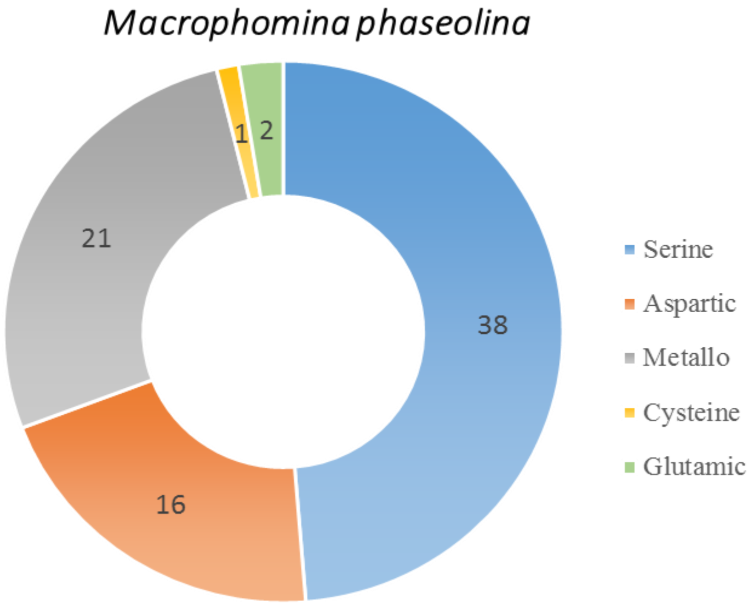


Figure 21. Distribution of extracellular proteases of *Macrophomina phaseolina* by MEROPS families (Serine, Aspartic, Metallo, Cysteine and Glutamic).

The draft genome sequence of *M. phaseolina* contains 13.07% secreted proteins as compared to 7-10% in other saprophytes, such as *Aspergillus niger* and *Aspergillus flavus*.(152)

Between the organisms in study, this is the only one that contains genes that codify for extracellular proteases representative of 5 different protease families (aspartic, serine, cysteine, metallic and glutamic). Serine peptidases are notoriously the majority (about 49% of the total proteases) followed by the aspartic peptidases (about 27%).

In *M. phaseolina*, the only secreted cysteine peptidase with the domain properly identified, belongs to C13 family. The accession number of the amino acid sequence is EKG15700.1. Peptidase family C13 contains asparaginyl endopeptidases/legumains and protein transamidases.(153)

This family is also known as legumain-like because it was first identified and charaterised from a legumen. Legumain is usually associated with cysteine endopeptidase activity and cleaves only carboxyterminally to the amino acid asparagine. Despite the strict specificity of the substrate, is associated with a general digestion function in the lysosomes. More recently, legumain has been suggested to have a role in cleaving/activating other proteases, such as matrix metalloproteinase-2.(154) It is interesting that the only cysteinase secreted is a member of a MEROPS C13 family that is mainly found in plants(155) and humans (in immunity and tumor progression).(156) So far the literature did not associated legumains with fungal secretome. Since proteins of this family are well described in plants genome, the presence the enzymes in *M. phaseolina* can be associated with an adaptive process, but further analysis is required.

3.3.2.2. *Neofusicoccum parvum*

The GenBank accession numbers of the sequences of all secreted proteases are listed in Annexe 5.4, organised by MEROPS protease's families.

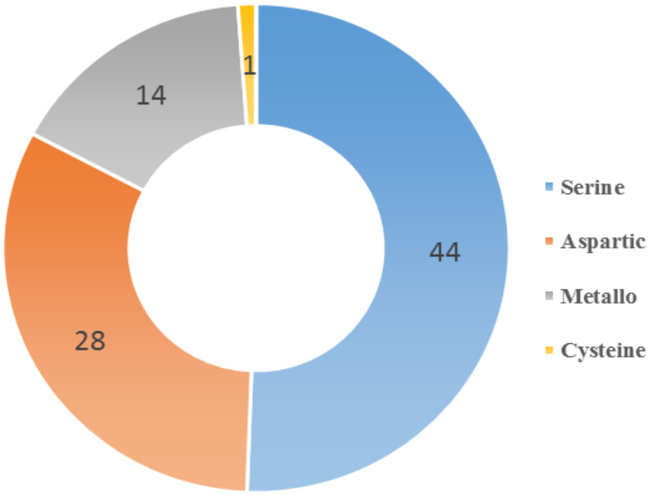


Figure 22 Distribution of extracellular proteases of Neofusicoccum parvum by MEROPS families (Serine, Aspartic, Metallo and Cysteine).

Neofusiccoum parvum expresses extracellular proteases of 4 different families: serine, aspartic, metallic and cysteinic. The only secreted cysteine peptidase with the domain properly identified, belongs to C54 family. The accession number of the amino acid sequence is EOD48149.1. Peptidase family C54 contains endopeptidases with specificity for glycyl bonds. Family C54 endopeptidases activate and remove lipids of proteins important in a number of intracellular signalling pathways that involve transfer of proteins across membranes, including autophagy, intra-Golgi transport and receptor sorting. The active site residues specific of MEROPS peptidase family C54 cysteine proteases are Y54 C74 D278 H280: tyrosine, cysteine, aspartic acid and histidine, respectively. The family C54 is distributed among Protozoa, Fungi, Plants and Animals. According to MEROPs the only fungus belonging to *Botryosphaeriaceae* family that has described a C54 peptidase homologues is *Macrophomina phaseolina*, with one cyoplasmatic cysteine peptidase. In this work, the only sequence described and with its domain properly identified as C54, belong to *N. parvum.*

3.3.2.3. *Diplodia corticola*

The GenBank accession numbers of the sequences of all secreted proteases are listed in Annexe 5.5, organised by MEROPS protease's families.

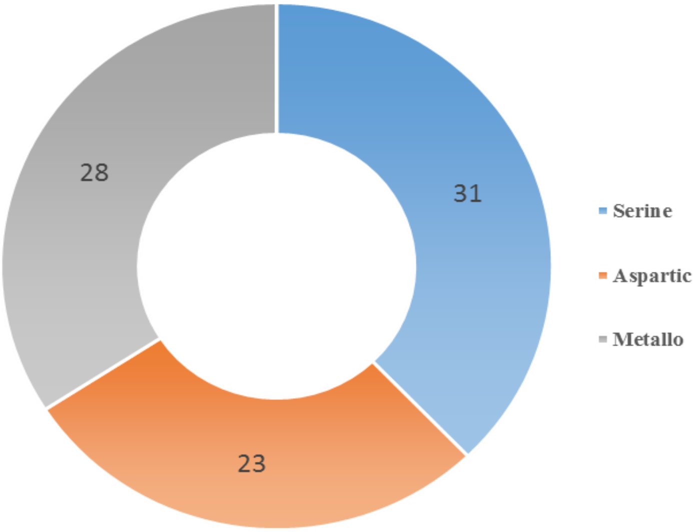


Figure 23. Distribution of extracellular proteases of *Diplodia corticola* by MEROPS families (Serine, Aspartic, Metallo).

Based on Interpro and Smart tool for sequence analysis, it was not possible to predict the protein family membership of any extracellular cysteine or glutamic proteases sequences, due the lack of functional domain identification.

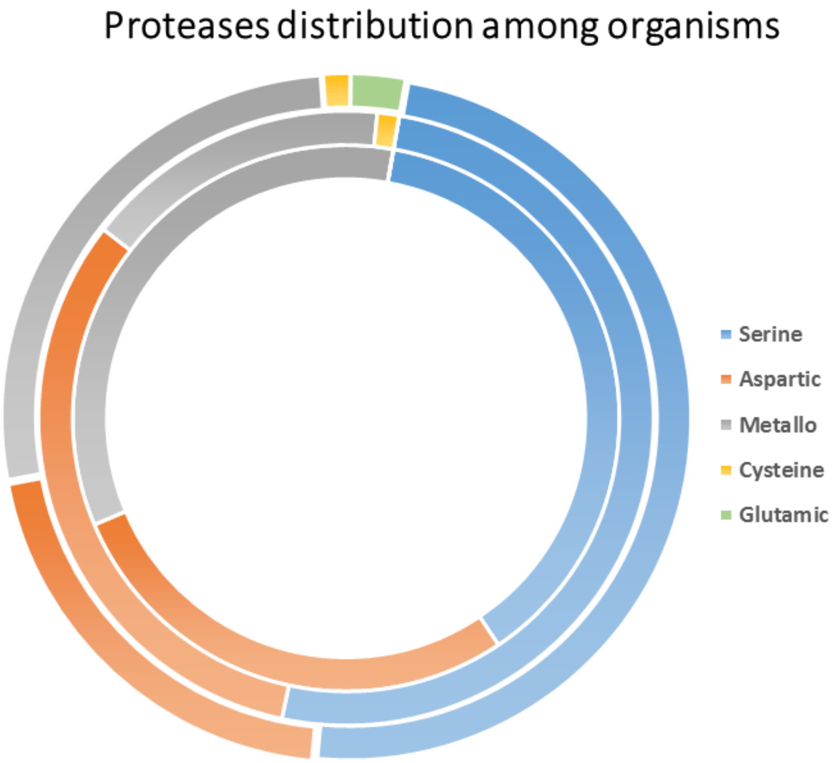


Figure 24 Comparison of number of proteases sequences per megabases of genome (millions of base pairs, Mbp) between *M. phaseolina, N. parvum* and *D. corticola.*

Plant pathogenic fungi secrete a large amount of proteins, among them degrading enzymes. In the presence of plant extract, the proteases secretion by filamentous fungi increases.(78) In order to protect themselves from invasive parasites, the plant defence mechanism involves the secretion of proteases inhibitors. This complex process involves the activation or repression of different signaling pathways leading to the overexpression of target genes with defence properties.(99) Plant proteinase inhibitors have been well established to play a potent defensive role against predators and pathogens. The presence of protease inhibitors in plant protection against fungi is an indicator of the importance of proteases in phytopatogenesis.

However, it is remarkable the distinctive secretome profiles between the microorganisms. The amount of secreted proteases associated with *Neofusicoccum parvum* (87 sequences) is higher than *Diplodia corticola* (82 sequences) and even *Macrophomina phaseolina* (78 sequences, Table 3). However, to a fair comparison is necessary take into account the size of their genomes (millions of base pairs, Mbp) (Figure 24)

It was identified a bigger amount of serine and aspartic proteases in *N. parvum* genome, however, regarding to metallo proteases *D. corticola* high number of protein coding sequences.

It is very important to accentuate the diversity of proteases' genes present in the genome of *M. phaseolina* genome. Contrary to the others organisms studied, this fungus presentes proteases belonging to the 5 different families. The diverse secretome profile might be associated with their hosts' types, their geographical location or their living conditions. Over the years, this worldwide distributed fungus has been isolated from increasingly diverse hosts, such as strawberry(40), soybean(157) and even humans(158)*. M. phaseolina* can cause diseases in more than 500 plant species, being one of the most destructive necrotrophic fungal pathogens. In fact, this fungus has the capacity to develop in extreme climatic conditions, making it extremely adaptable and versatile.(45)

Neofusicoccum parvum has a broad localisation spectrum and can be found in numerous countries such as Australia, Chile, China, Ethiopia, Indonesia, South Africa, Uganda, Uruguay, Venezuela, Spain (34) and Portugal.(159) But this pathogen host species are more limited than *M. phaseolina* hosts. It was reported as the pathogen responsible for causing canker symptoms on *Eucalyptus* species in distant parts of the world and in grapevine.(35)

Diplodia corticola presents genes codifying for only 3 families of proteases. Despite fungi of *Botryosphaeria* genera causing disease in hundreds of plant species, *D. corticola* was found only in oaks, grape vines and eucalypts. It was reported in the main producing countries of cork such as Portugal, France, Spain, Italy and Morocco.(50) And was also associated with oak disease symptoms in pathogens in Greece, Hungary, California and Florida. It is interesting that the fungus with a spare number of hosts, *Diplodia corticola*, also has low proteases families' diversity present in the secretome.

Even though proteolytic enzymes knowledge is increasing, many family members still have to be studied in detail. We are becoming aware of the physiological role that proteases play in proteolysis networks. However, there are still many open questions regarding their actions in pathophysiology. Very few protein are resistant to proteolysis so these enzymes are involved in key decisions that determine life and death of an organism. Proteolysis is required to inhibit the signals that proteins initiate by degrading either them or the proteins they bind to. Besides that, pathogens use proteolytic enzymes to invade their hosts and to destroy/inactivate potential lethal or toxic proteins express by the host that could affect them or interfere with their reproduction. (160)

3.4. Multiple sequences alignment/Motif finding

Every protein sequence has an evolutionary history. Due to that history, sequences similarity might imply a functional or structural conservation. Biologically important residues/nucleotides are assumed to be less-likely to mutate than unimportant ones, so the sequence conservation might be a hint to a functionally important region. Multiple Sequence Alignment (MSA) allows grouping the conserved residues/nucleotides according to the highest sequence similarity. A sequence motif is a short nucleotide or amino acid sequence pattern (usually between 5 and 30 nucleotides, respectively amino acids) which is conjectured or known to have an important biological role.

The MSA output is influenced by: the number of sequences, the length of the sequences, the alphabet of the sequences (DNA, RNA or amino acids) and the divergence of the sequence. Therefore, the choice of program should take into account the types of sequences being aligned (nucleic acid or protein), how related the sequences are, the sequence lengths or how many sequences are being compared.

In this work, an experimental trial was performed using some of the more popular and available MSA programs: Clustal X2.1, T-Coffee and Muscle. Several multiple sequences alignments with various parameters' combinations were made for different sets of sequences. A successful alignment is crucial because it is used in a number of applications, such as the identification of conserved motifs and domains within related families of proteins that then may be inferred to play a role in structure. Besides, the MSA is the first step of a phylogenetic analysis that can lead to the prediction of evolutionary relationships.

From the several algorithms used, the one with the most accurate alignment profile was T-Coffee method. This result was expected because in previous studies T-Coffee had a higher score over MUSCLE and Clustal in accuracy.(161) However, if we consider running time, MUSCLE outperforms all of the programs reviewed here. MUSCLE was optimised for speed so its performance speed is expected to be high. Contrasting is T-Coffee, the slowest of the three programs.(162) Clustal is the oldest of the programs so, nowadays, there are definitely programs that are faster and more accurate.(132) Despite speed and memory being important criteria in MSA methods, the main one is their accuracy. What excels T-coffee is the fact that it uses a combination of local and global pairwise alignments to generate the sequence library, increasing the accuracy.(163) In a short note, despite the low speed of T-Coffee method, its accuracy comparing to CLUSTAL X2.1 and MUSCLE overcomes its overall performance. For that reason, the following results were obtained using T-Coffee method.

To align with T-Coffee the advanced options used to all the alignment were: Gap opening penalty -50; Gap extension penalty 0; and Number of iterations 0.

3.4.1. Aspartic proteases – family A1

Almost all the proteases found in the 3 organisms belong to A1 family (MEROPS classification). In 2 sequences belonging to *N. parvum* an A4 domain was identified.

The aspartic proteases secreted by fungi are generally similar to pepsin (A1 family) or belong to the A4 family which contains only fungal secreted enzymes.(164)

Peptidase family A1 contains endopeptidases, most of which are most active at acidic pH values. In this family, peptidases were formerly known as "acid proteinases" or "carboxyl proteinases". The amino acid sequences for the catalytic site motif are much conserved between organisms.

Their general structural pattern contains 3 highly conserved regions and 4 cysteine residues. Two of the conserved regions with the motifs DSG or DTG contain the two reactive aspartic residues (D) of the active site of the pepsin-like proteases. The core motif is Xaa-Xaa-Asp-Xbb-Gly-Xbb, where Xaa is a hydrophobic residue and Xbb is either Ser or Thr. The hydrophobic amino acid is normally buried inside the protein core: alanine (A), isoleucine (I), leucine (L), phenylalanine (F), valine (V), proline (P) and glycine (G) One lobe may be evolved from the other through ancient gene-duplication event.(84) The third conserved region is found at the C-terminus of the protein. The cysteine residues form two disulphide bridges implicated in the maintenance of the three-dimensional structure. The presence and position of disulphide bridges are other conserved features of aspartic peptidases.

In plants, plant aspartic proteinases are grouped in three classes depending on their putative domain organisations and their active site sequence motifs.(165)

- Typical; Hydrophobic-hydrophobic-DTG-serine-serine
- Atypical; Hydrophobic-hydrophobic-DTG-serine-acidic
- Nucellin-like; Acidic-hydrophobic-DTG-serine-acidic

The evolutionary or biological significance of this variation observed in aspartic proteases of different kingdoms has not been established.

3.4.1.1. *Macrophomina phaseolina*

The multiple sequence alignment covered just a part of the protein sequence: the functional domain (Figure 25). For domain identification and selection, SMART tool was used. Analysing only the functional domain of each enzyme, allow to standardise the sequence length, which allows a better alignment. The range of the functional domain of each sequence is shown in Annexe 5.6.1

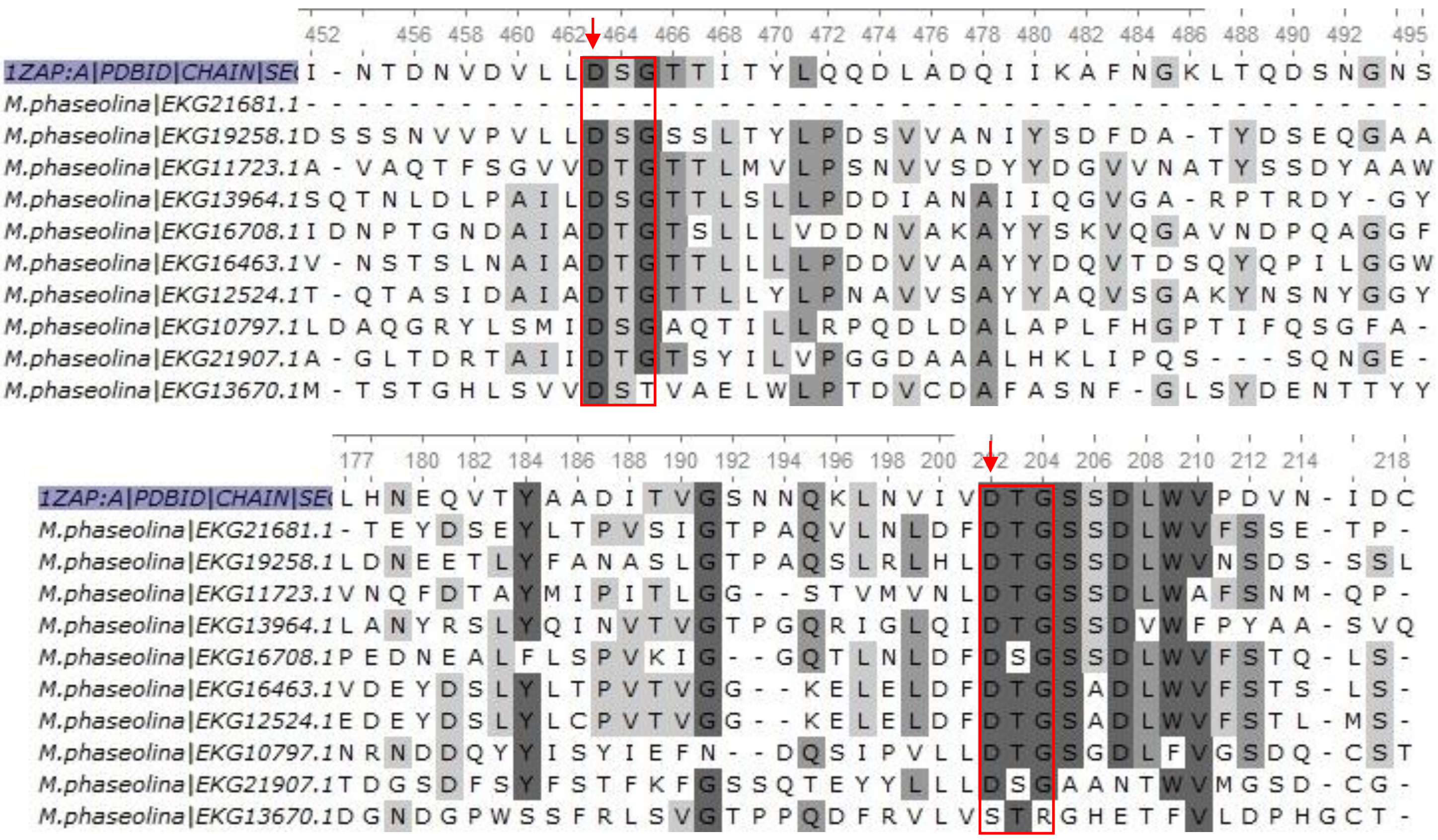
The active site is the region responsible for the enzyme function. In aspartic proteases the two acid aspartic at the active site play key catalytic roles in the pepsin family and are conserved for all family members.(166) The catalytic motifs in both N- and C-terminal lobes of this family are conserved regions with the motifs DSG or DTG.

Secreted aspartic proteinases (SAP) domains are found on diverse nuclear proteins and they are found in a wide range of organisms, including yeast, plants, animals, fungi and even bacteriophages.(167) The production of SAP like proteins, has already been described in several fungi of the phylum Ascomycota(168) as *Neofusicoccum parvum*. Although it has not yet been identified SAP domains in any of aspartic proteases of *N. parvum,* previous work confirm their presence.(33)

They are secreted from the pathogen to degrade host proteins. SAP is one of the most significant extracellular hydrolytic enzymes produced by *C. albicans*.(169) In *C. albicans* was observed that SAP deficient strains are less virulent or avirulent in infected hosts. In this specific scenario it is difficult to confirm the virulence contribution of the SAPs, however, the secretion of proteases in general correlates strongly with pathogenicity. (170)

This motif has been found in a number of chromatin associating proteins, such as scaffold attachment factors, DNA repair proteins, RNA processing complexes and proto-oncogene proteins.(167) In plants was associated with responses to mild and severe oxidative stresses, by mediating DNA repair and programmed cell death processes.(171)

A multiple sequence alignment was performed with the functional domain sequence of aspartic proteases and the reference sequence (pepsin A (Homo sapiens), MEROPS Accession MER000885)



```
                             452   456 458 460 462 464 466 468 470 472 474 476 478 480 482 484 486 488 490 492  495
1ZAP:A|PDBID|CHAIN|SE(       I - N T D N V D V L L D S G T T I T Y L Q Q D L A D Q I I K A F N G K L T Q D S N G N S
M.phaseolina|EKG21681.1      - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
M.phaseolina|EKG19258.1      D S S S N V V P V L L D S G S S L T Y L P D S V V A N I Y S D F D A - T Y D S E Q G A A
M.phaseolina|EKG11723.1      A - V A Q T F S G V V D T G T T L M V L P S N V V S D Y Y D G V V N A T Y S S D Y A A W
M.phaseolina|EKG13964.1      S Q T N L D L P A I L D S G T T L S L L P D D I A N A I I Q G V G A - R P T R D Y - G Y
M.phaseolina|EKG16708.1      I D N P T G N D A I A D T G T S L L L V D D N V A K A Y Y S K V Q G A V N D P Q A G G F
M.phaseolina|EKG16463.1      V - N S T S L N A I A D T G T T L L L L P D D V V A A Y Y D Q V T D S Q Y Q P I L G G W
M.phaseolina|EKG12524.1      T - Q T A S I D A I A D T G T T L L Y L P N A V V S A Y Y A Q V S G A K Y N S N Y G G Y
M.phaseolina|EKG10797.1      L D A Q G R Y L S M I D S G A Q T I L L R P Q D L D A L A P L F H G P T I F Q S G F A -
M.phaseolina|EKG21907.1      A - G L T D R T A I I D T G T S Y I L V P G G D A A A L H K L I P Q S - - - S Q N G E -
M.phaseolina|EKG13670.1      M - T S T G H L S V V D S T V A E L W L P T D V C D A F A S N F - G L S Y D E N T T Y Y

                             177  180 182 184 186 188 190 192 194 196 198 200 202 204 206 208 210 212 214  218
1ZAP:A|PDBID|CHAIN|SE(       L H N E Q V T Y A A D I T V G S N N Q K L N V I V D T G S S D L W V P D V N - I D C
M.phaseolina|EKG21681.1      - T E Y D S E Y L T P V S I G T P A Q V L N L D F D T G S S D L W V F S S E - T P -
M.phaseolina|EKG19258.1      L D N E E T L Y F A N A S L G T P A Q S L R L H L D T G S S D L W V N S D S - S S L
M.phaseolina|EKG11723.1      V N Q F D T A Y M I P I T L G G - - S T V M V N L D T G S S D L W A F S N M - Q P -
M.phaseolina|EKG13964.1      L A N Y R S L Y Q I N V T V G T P G Q R I G L Q I D T G S S D V W F P Y A A - S V Q
M.phaseolina|EKG16708.1      P E D N E A L F L S P V K I G - - G Q T L N L D F D S G S S D L W V F S T Q - L S -
M.phaseolina|EKG16463.1      V D E Y D S L Y L T P V T V G G - - K E L E L D F D T G S A D L W V F S T S - L S -
M.phaseolina|EKG12524.1      E D E Y D S L Y L C P V T V G G - - K E L E L D F D T G S A D L W V F S T L - M S -
M.phaseolina|EKG10797.1      N R N D D Q Y Y I S Y I E F N - - D Q S I P V L L D T G S G D L F V G S D Q - C S T
M.phaseolina|EKG21907.1      T D G S D F S Y F S T F K F G S S Q T E Y Y L L L D S G A A N T W V M G S D - C G -
M.phaseolina|EKG13670.1      D G N D G P W S S F R L S V G T P P Q D F R V L V S T R G H E T F V L D P H G C T -
```

Figure 25. Protein alignment of aspartic protein sequences of *Macrophomina phaseolina* aligned using T-Coffee software with default settings. The conserved regions are indicated by boxes.

Not surprisingly, the catalytic domain is the region displaying the greatest sequence similarity among the peptidase A1 members. This domain also best aligns with *C. albicans* secreted aspartic protease, the prototypical enzyme for the A1 family of aspartic endopeptidases (MEROPS the Peptidase Database; http://merops.sanger.ac.uk/)

Analysing the peptidase family A1, the active site as not found in all sequences. Despite all the sequences had been thoroughly analysed before been classify into families (using SMART and InterPro tools), the conserved regions were not found in some sequences. Therefore, the following protein sequences were eliminated: EKG11095.1; EKG22569.1; EKG19996.1. The catalytic motif of the sequence *M.phaseolina*|EKG21681.1| was only found in the C- terminal. It was not possible to localise the DTG/DTG section in the C-terminal.

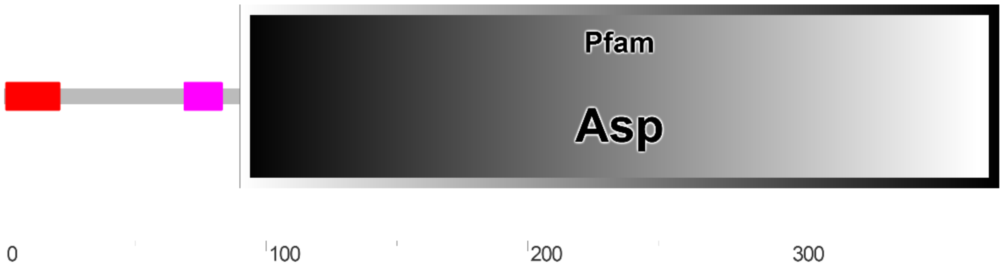
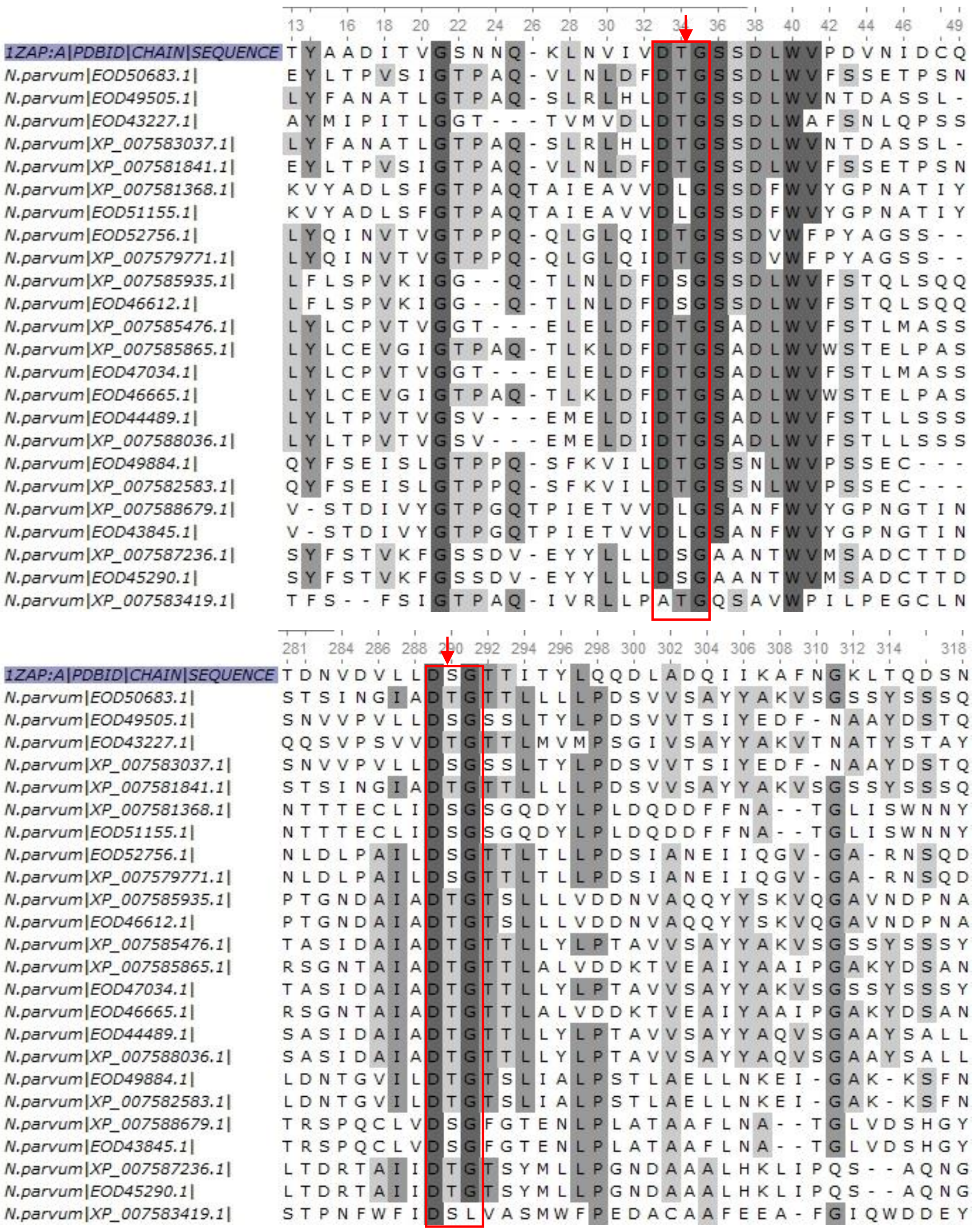


Figure 26 Sequence organisation with domain architecture representation of the sequence EKG21681.1

3.4.1.2. *Neofusicoccum parvum*

In *N. parvum* genome a total of 28 protein sequences of secreted aspartic proteases belonging to A1 family were identified. A multiple sequence alignment covered just a part of the protein sequence: the functional domain (Figure 27). The range of the functional domain in each sequence is shown in Annexe 5.6.2.1.

Despite all the sequences had been thoroughly analysed before been classify into families, the conserved regions of some sequences were not found. Therefore, those protein sequences were eliminated: EOD50588; EOD44319; XP_007587137 and XP_007588209.



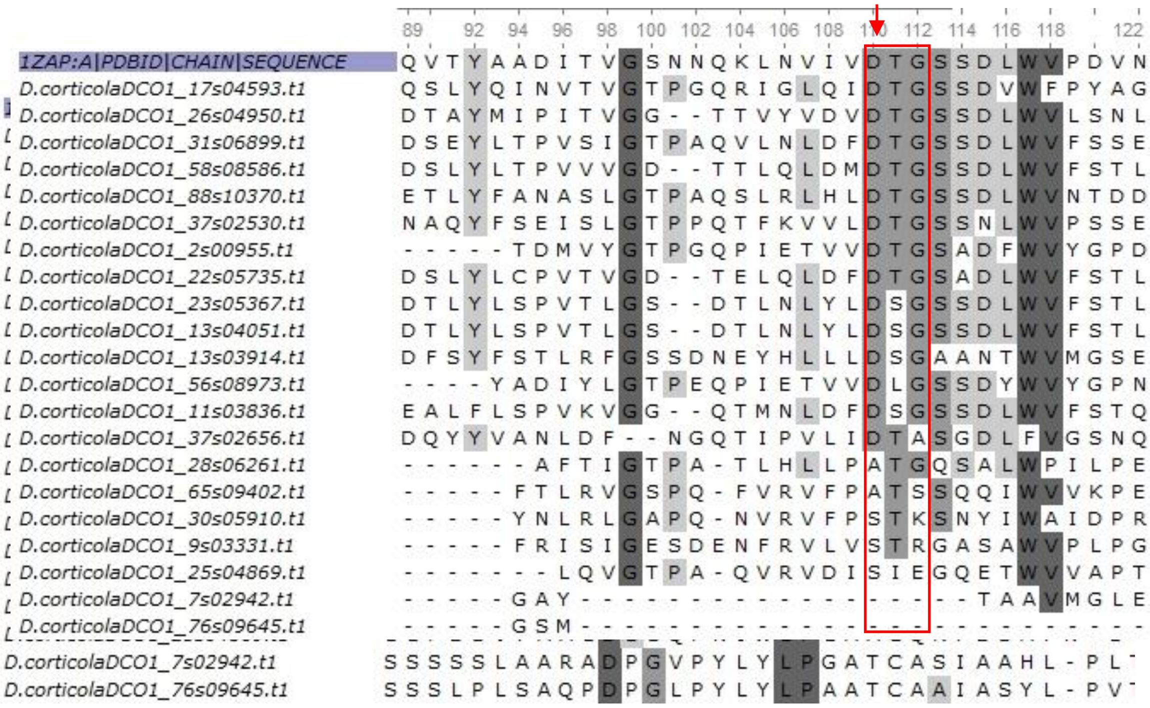
```
                                13    16  18  20  22  24  26  28  30  32  34  36  38  40  42  44  46   49
1ZAP:A|PDBID|CHAIN|SEQUENCE     T Y A A D I T V G S N N Q - K L N V I V D T G S S D L W V P D V N I D C Q
N.parvum|EOD50683.1|            E Y L T P V S I G T P A Q - V L N L D F D T G S S D L W V F S S E T P S N
N.parvum|EOD49505.1|            L Y F A N A T L G T P A Q - S L R L H L D T G S S D L W V N T D A S S L -
N.parvum|EOD43227.1|            A Y M I P I T L G G T - - - T V M V D L D T G S S D L W A F S N L Q P S S
N.parvum|XP_007583037.1|        L Y F A N A T L G T P A Q - S L R L H L D T G S S D L W V N T D A S S L -
N.parvum|XP_007581841.1|        E Y L T P V S I G T P A Q - V L N L D F D T G S S D L W V F S S E T P S N
N.parvum|XP_007581368.1|        K V Y A D L S F G T P A Q T A I E A V V D L G S S D F W V Y G P N A T I Y
N.parvum|EOD51155.1|            K V Y A D L S F G T P A Q T A I E A V V D L G S S D F W V Y G P N A T I Y
N.parvum|EOD52756.1|            L Y Q I N V T V G T P P Q - Q L G L Q I D T G S S D V W F P Y A G S S - -
N.parvum|XP_007579771.1|        L Y Q I N V T V G T P P Q - Q L G L Q I D T G S S D V W F P Y A G S S - -
N.parvum|XP_007585935.1|        L F L S P V K I G G - - Q - T L N L D F D S G S S D L W V F S T Q L S Q Q
N.parvum|EOD46612.1|            L F L S P V K I G G - - Q - T L N L D F D S G S S D L W V F S T Q L S Q Q
N.parvum|XP_007585476.1|        L Y L C P V T V G G T - - - E L E L D F D T G S A D L W V F S T L M A S S
N.parvum|XP_007585865.1|        L Y L C E V G I G T P A Q - T L K L D F D T G S A D L W V W S T E L P A S
N.parvum|EOD47034.1|            L Y L C P V T V G G T - - - E L E L D F D T G S A D L W V F S T L M A S S
N.parvum|EOD46665.1|            L Y L C E V G I G T P A Q - T L K L D F D T G S A D L W V W S T E L P A S
N.parvum|EOD44489.1|            L Y L T P V T V G S V - - - E M E L D I D T G S A D L W V F S T L L S S S
N.parvum|XP_007588036.1|        L Y L T P V T V G S V - - - E M E L D I D T G S A D L W V F S T L L S S S
N.parvum|EOD49884.1|            Q Y F S E I S L G T P P Q - S F K V I L D T G S S N L W V P S S E C - - -
N.parvum|XP_007582583.1|        Q Y F S E I S L G T P P Q - S F K V I L D T G S S N L W V P S S E C - - -
N.parvum|XP_007588679.1|        V - S T D I V Y G T P G Q T P I E T V V D L G S A N F W V Y G P N G T I N
N.parvum|EOD43845.1|            V - S T D I V Y G T P G Q T P I E T V V D L G S A N F W V Y G P N G T I N
N.parvum|XP_007587236.1|        S Y F S T V K F G S S D V - E Y Y L L L D S G A A N T W V M S A D C T T D
N.parvum|EOD45290.1|            S Y F S T V K F G S S D V - E Y Y L L L D S G A A N T W V M S A D C T T D
N.parvum|XP_007583419.1|        T F S - - F S I G T P A Q - I V R L L P A T G Q S A V W P I L P E G C L N

                                281   284 286 288 290 292 294 296 298 300 302 304 306 308 310 312 314   318
1ZAP:A|PDBID|CHAIN|SEQUENCE     T D N V D V L L D S G T T I T Y L Q Q D L A D Q I I K A F N G K L T Q D S N
N.parvum|EOD50683.1|            S T S I N G I A D T G T T L L L L P D S V V S A Y Y A K V S G S S Y S S S Q
N.parvum|EOD49505.1|            S N V V P V L L D S G S S L T Y L P D S V V T S I Y E D F - N A A Y D S T Q
N.parvum|EOD43227.1|            Q Q S V P S V V D T G T T L M V M P S G I V S A Y Y A K V T N A T Y S T A Y
N.parvum|XP_007583037.1|        S N V V P V L L D S G S S L T Y L P D S V V T S I Y E D F - N A A Y D S T Q
N.parvum|XP_007581841.1|        S T S I N G I A D T G T T L L L L P D S V V S A Y Y A K V S G S S Y S S S Q
N.parvum|XP_007581368.1|        N T T T E C L I D S G S G Q D Y L P L D Q D D F F N A - - T G L I S W N N Y
N.parvum|EOD51155.1|            N T T T E C L I D S G S G Q D Y L P L D Q D D F F N A - - T G L I S W N N Y
N.parvum|EOD52756.1|            N L D L P A I L D S G T T L T L L P D S I A N E I I Q G V - G A - R N S Q D
N.parvum|XP_007579771.1|        N L D L P A I L D S G T T L T L L P D S I A N E I I Q G V - G A - R N S Q D
N.parvum|XP_007585935.1|        P T G N D A I A D T G T S L L L V D D N V A Q Q Y Y S K V Q G A V N D P N A
N.parvum|EOD46612.1|            P T G N D A I A D T G T S L L L V D D N V A Q Q Y Y S K V Q G A V N D P N A
N.parvum|XP_007585476.1|        T A S I D A I A D T G T T L L Y L P T A V V S A Y Y A K V S G S S Y S S S Y
N.parvum|XP_007585865.1|        R S G N T A I A D T G T T L A L V D D K T V E A I Y A A I P G A K Y D S A N
N.parvum|EOD47034.1|            T A S I D A I A D T G T T L L Y L P T A V V S A Y Y A K V S G S S Y S S S Y
N.parvum|EOD46665.1|            R S G N T A I A D T G T T L A L V D D K T V E A I Y A A I P G A K Y D S A N
N.parvum|EOD44489.1|            S A S I D A I A D T G T T L L Y L P T A V V S A Y Y A Q V S G A A Y S A L L
N.parvum|XP_007588036.1|        S A S I D A I A D T G T T L L Y L P T A V V S A Y Y A Q V S G A A Y S A L L
N.parvum|EOD49884.1|            L D N T G V I L D T G T S L I A L P S T L A E L L N K E I - G A K - K S F N
N.parvum|XP_007582583.1|        L D N T G V I L D T G T S L I A L P S T L A E L L N K E I - G A K - K S F N
N.parvum|XP_007588679.1|        T R S P Q C L V D S G F G T E N L P L A T A A F L N A - - T G L V D S H G Y
N.parvum|EOD43845.1|            T R S P Q C L V D S G F G T E N L P L A T A A F L N A - - T G L V D S H G Y
N.parvum|XP_007587236.1|        L T D R T A I I D T G T S Y M L L P G N D A A A L H K L I P Q S - - A Q N G
N.parvum|EOD45290.1|            L T D R T A I I D T G T S Y M L L P G N D A A A L H K L I P Q S - - A Q N G
N.parvum|XP_007583419.1|        S T P N F W F I D S L V A S M W F P E D A C A A F E E A - F G I Q W D D E Y
```

Figure 27. Protein alignment of aspartic protein sequences of *Neofusicoccum parvum* aligned using T-Coffee software with default settings. The conserved regions are indicated by boxes.

In the alignment it is possible to observe aligned regions, except for the XP_007588209 sequence. The active-site centre of these endopeptidases is composed of two Asp residues that generally occur in a highly conserved sequence motif. Since that region is key for a protein function this sequence was eliminated.

3.4.1.3. *Diplodia corticola*

In *D. corticola* genome a total of 23 protein sequences of secreted aspartic proteases belonging to A1 family were identified. A multiple sequence alignment covered just a part of the protein sequence: the functional domain (Figure 28). The range of the functional domain in each sequence is shown in Annexe 5.6.3.1.



```
                                 89   92 94 96 98 100 102 104 106 108 110 112 114 116 118   122
1ZAP:A|PDBID|CHAIN|SEQUENCE      Q V T Y A A D I T V G S N N Q K L N V I V D T G S S D L W V P D V N
D.corticolaDCO1_17s04593.t1      Q S L Y Q I N V T V G T P G Q R I G L Q I D T G S S D V W F P Y A G
D.corticolaDCO1_26s04950.t1      D T A Y M I P I T V G G - - T T V Y V D V D T G S S D L W V L S N L
D.corticolaDCO1_31s06899.t1      D S E Y L T P V S I G T P A Q V L N L D F D T G S S D L W V F S S E
D.corticolaDCO1_58s08586.t1      D S L Y L T P V V V G D - - T T L Q L D M D T G S S D L W V F S T L
D.corticolaDCO1_88s10370.t1      E T L Y F A N A S L G T P A Q S L R L H L D T G S S D L W V N T D D
D.corticolaDCO1_37s02530.t1      N A Q Y F S E I S L G T P P Q T F K V V L D T G S S N L W V P S S E
D.corticolaDCO1_2s00955.t1       - - - - - T D M V Y G T P G Q P I E T V V D T G S A D F W V Y G P D
D.corticolaDCO1_22s05735.t1      D S L Y L C P V T V G D - - T E L Q L D F D T G S A D L W V F S T L
D.corticolaDCO1_23s05367.t1      D T L Y L S P V T L G S - - D T L N L Y L D S G S S D L W V F S T L
D.corticolaDCO1_13s04051.t1      D T L Y L S P V T L G S - - D T L N L Y L D S G S S D L W V F S T L
D.corticolaDCO1_13s03914.t1      D F S Y F S T L R F G S S D N E Y H L L L D S G A A N T W V M G S E
D.corticolaDCO1_56s08973.t1      - - - - Y A D I Y L G T P E Q P I E T V V D L G S S D Y W V Y G P N
D.corticolaDCO1_11s03836.t1      E A L F L S P V K V G G - - Q T M N L D F D S G S S D L W V F S T Q
D.corticolaDCO1_37s02656.t1      D Q Y Y V A N L D F - - N G Q T I P V L I D T A S G D L F V G S N Q
D.corticolaDCO1_28s06261.t1      - - - - - - - A F T I G T P A - T L H L L P A T G Q S A L W P I L P E
D.corticolaDCO1_65s09402.t1      - - - - - - F T L R V G S P Q - F V R V F P A T S S Q Q I W V V K P E
D.corticolaDCO1_30s05910.t1      - - - - - - Y N L R L G A P Q - N V R V F P S T K S N Y I W A I D P R
D.corticolaDCO1_9s03331.t1       - - - - - - F R I S I G E S D E N F R V L V S T R G A S A W V P L P G
D.corticolaDCO1_25s04869.t1      - - - - - - - - L Q V G T P A - Q V R V D I S I E G Q E T W V V A P T
D.corticolaDCO1_7s02942.t1       - - - - - - G A Y - - - - - - - - - - - - - - - - - - T A A V M G L E
D.corticolaDCO1_76s09645.t1      - - - - - - G S M - - - - - - - - - - - - - - - - - - - - - - - - - -
D.corticolaDCO1_7s02942.t1     S S S S S L A A R A D P G V P Y L Y L P G A T C A S I A A H L - P L
D.corticolaDCO1_76s09645.t1    S S S L P L S A Q P D P G L P Y L Y L P A A T C A A I A S Y L - P V
```

Figure 28. Protein alignment of aspartic protein sequences of *Diplodia corticola* aligned using T-Coffee software with default settings. The conserved regions are indicated by boxes

The grey areas highlight the regions with >40% conservation level.

For a more reliable motif finding the *C. albicans* secreted aspartic protease reference sequence was included.

Unexpectedly, the active site was not identified in all sequences. Despite all the sequences had been thoroughly analysed before been classify into families (using SMART and InterPro tools), in some sequences the conserved regions were not detected after the MSA. Therefore, the following protein sequences were eliminated: DCO1_34s06771.t1; DCO1_16s03028.t1; DCO1_34s06771.t1

In addition, the catalytic motif in the N-terminal in several protein sequences could not be found. In only the following sequences, it as possible to localise the DTG/DTG section in the C-terminus conserved regions:

DCO1_28s06261.t1; DCO1_65s09402.t1; DCO1_30s05910.t1; DCO1_9s03331.t1; DCO1_25s04869.t1; DCO1_7s02942.t1; DCO1_76s09645.t1

3.4.2. Metallo proteases – family M35

Even though the majority of metallo peptidase sequences did not belong to family M35, we chose this family to analyse. Peptidase family M35 contains fungal metalloendopeptidases which has been previous described been as potentially involved in pathogenesis.(55)

For better a comprehensive view of fungal M35 family (deuterolysin), we conducted multiple sequences alignment and phylogenetic analyses of genes in this family from 3 sequenced Ascomycota fungi with different life style, *M. phaseolina, N. parvum* and *D. corticola.*

Metalloproteinases have been identified to play important roles in the pathogenicity of pathogenic fungi including phytopathogens. The metalloproteinases identified as secreted by pathogenic fungi mainly belong to two different families: the deuterolysin (M35) and the fungalysin (M36).(164,172)

Family M35 members contain two zinc binding histidines and a catalytic glutamate in an HEXXH motif. There is a third zinc ligand, an Asp, found in a GTXDXXYG motif C-terminal to the His zinc ligands.(70)

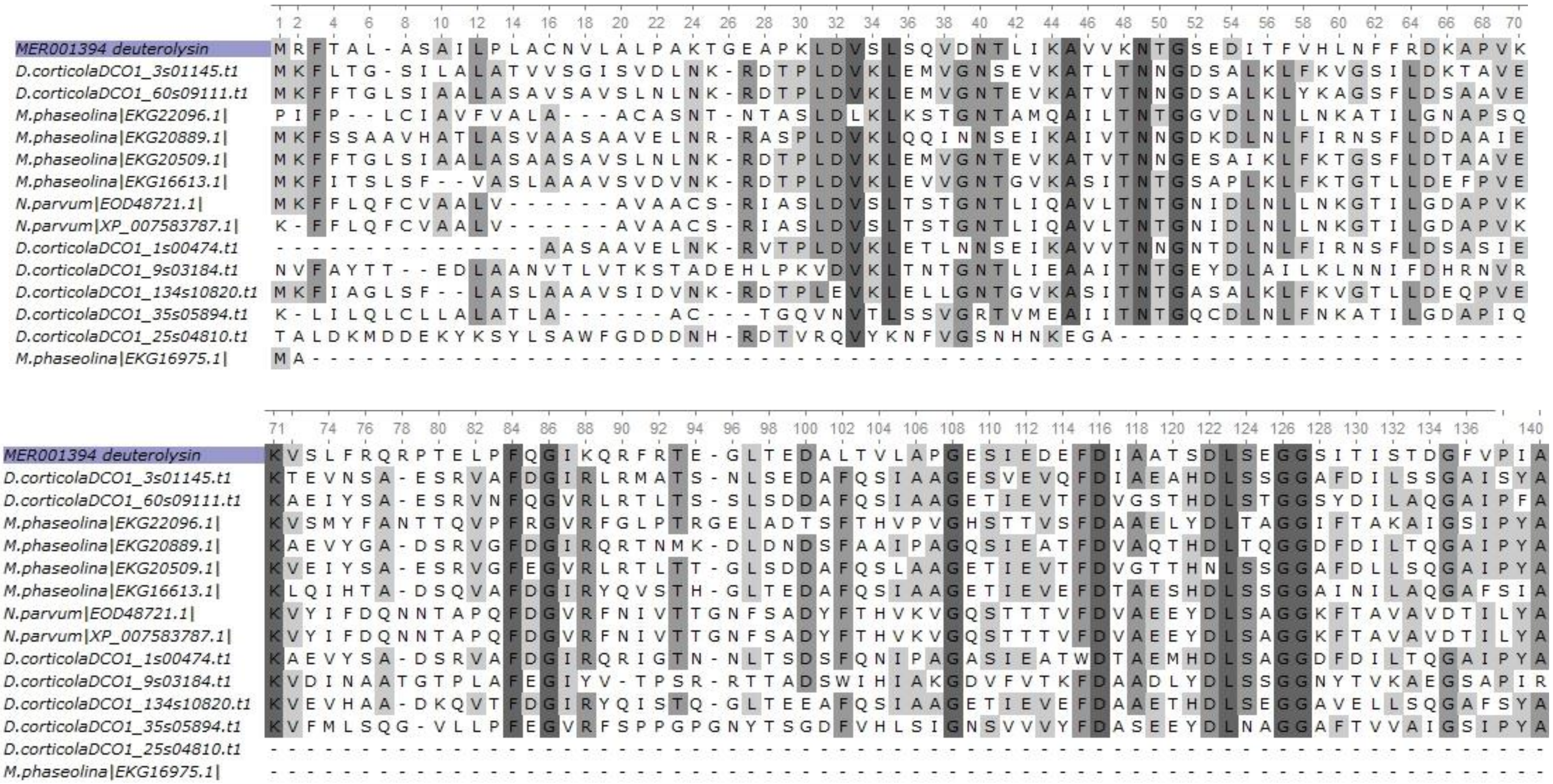
The amino acid sequences have two conserved features that contain three zinc ligands, known as 'aspzincin', defined as the "HEXXH + D" domain: Finding this consensus sequence is the most reliable way to detect a metallopeptidase. The N-terminus conserved region is composed by two histidine residues ligands of the metal atom; and the glutamate residue with a catalytic role.

Typically is found in a GTXDXXYG motif containing glycine, threonine, acid aspartic and tyrosine. The aspartic acid is the third zinc-binding residue. The conserved tyrosine residue acts as a proton donor during catalysis. This amino acid is usually located three residues C-terminal to the acid aspartic. Besides, the molecule of water becomes a ligand when activated and mediates the nucleophilic attack on the scissile peptide bond.(173,174)

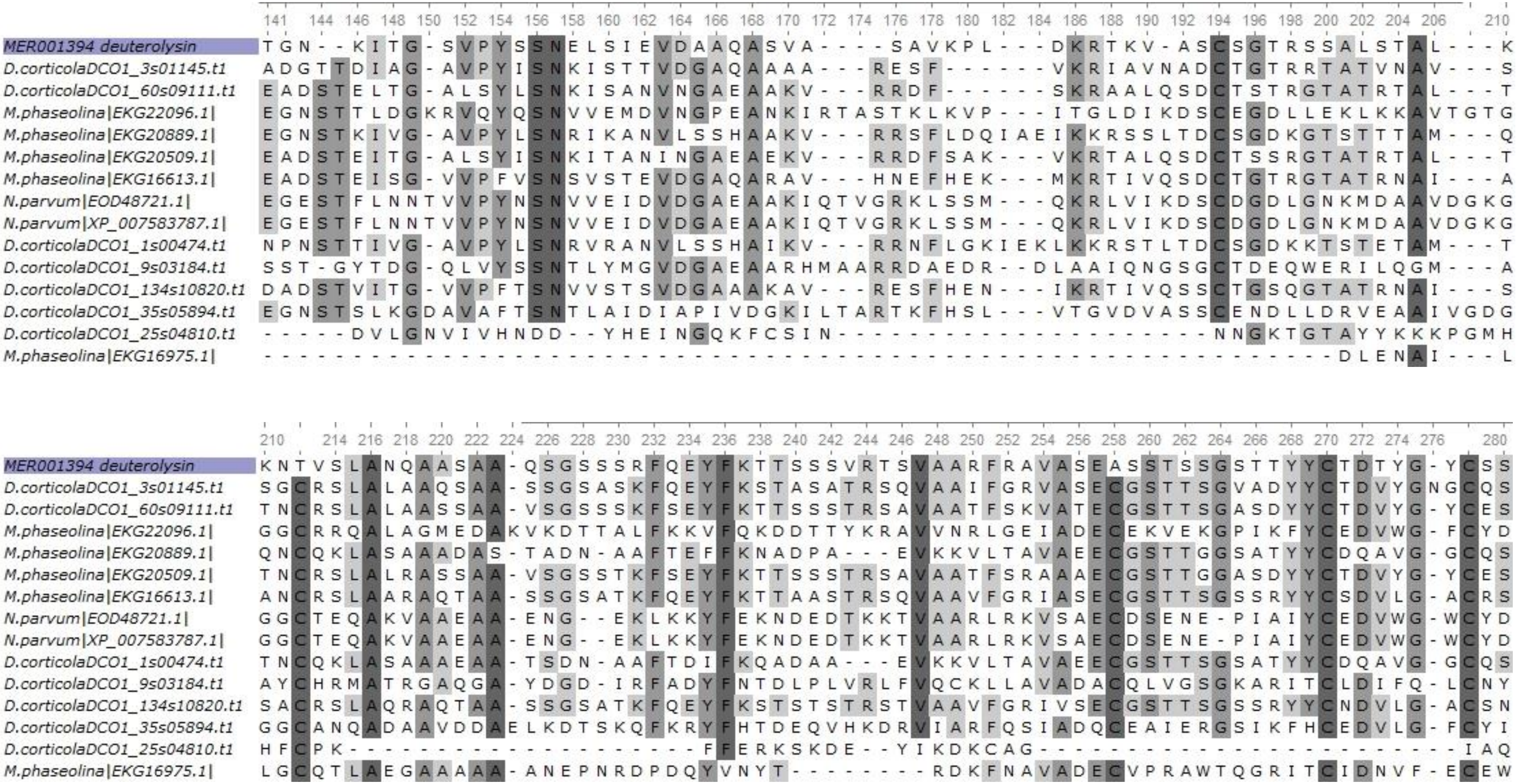
Deuterolysins are highly active towards basic nuclear proteins such as histones and protamines, with a preference for a Lys or Arg residues. Many members of the M35 peptidases display unusual thermostabilities.(175)

In total, only 16 protein sequences of secreted metallo proteases belonging to the M35 family were identified: nine sequences of *D. corticola,* six of *M. phaseolina* and only two of *N. parvum.*

Since the number of M35 peptidases is fairly low, we decided to analyse the multiple sequence alignment of all the sequences simultaneously. The multiple sequence alignment was performed with the protein sequences of metallo proteases and the reference sequence (deuterolysin (Aspergillus flavus), MEROPS Accession MER001394

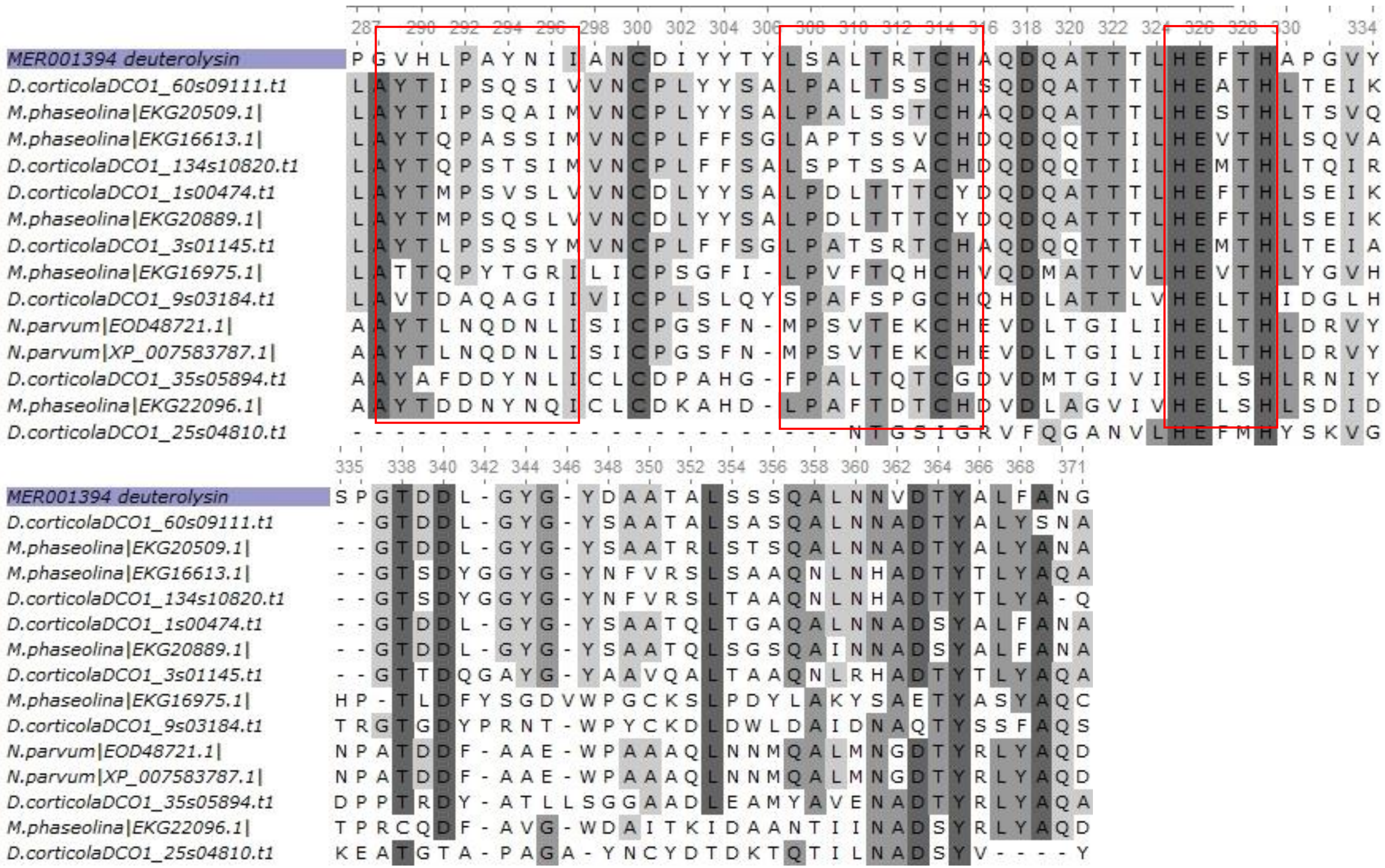


```
                              1 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70
MER001394 deuterolysin        MRFTAL-ASAILPLACNVLALPAKTGEAPKLDVSLSQVDNTLIKAVVKNTGSEDITFVHLNFFRDKAPVK
D.corticolaDCO1_3s01145.t1    MKFLTG-SILALATVVSGISVDLNK-RDTPLDVKLEMVGNSEVKATLTNNGDSALKLFKVGSILDKTAVE
D.corticolaDCO1_60s09111.t1   MKFFTGLSIAALASAVSAVSLNLNK-RDTPLDVKLEMVGNTEVKATVTNNGDSALKLYKAGSFLDSAAVE
M.phaseolina|EKG22096.1|      PIFP--LCIAVFVALA---ACASNT-NTASLDLKLKSTGNTAMQAILTNTGGVDLNLLNKATILGNAPSQ
M.phaseolina|EKG20889.1|      MKFSSAAVHATLASVAASAAVELNR-RASPLDVKLQQINNSEIKAIVTNNGDKDLNLFIRNSFLDDAAIE
M.phaseolina|EKG20509.1|      MKFFTGLSIAALASAASAVSLNLNK-RDTPLDVKLEMVGNTEVKATVTNNGESAIKLFKTGSFLDTAAVE
M.phaseolina|EKG16613.1|      MKFITSLSF--VASLAAAVSVDVNK-RDTPLDVKLEVVGNTGVKASITNTGSAPLKLFKTGTLLDEFPVE
N.parvum|EOD48721.1|          MKFFLQFCVAALV-------AVAACS-RIASLDVSLTSTGNTLIQAVLTNTGNIDLNLLNKGTILGDAPVK
N.parvum|XP_007583787.1|      K-FFLQFCVAALV-------AVAACS-RIASLDVSLTSTGNTLIQAVLTNTGNIDLNLLNKGTILGDAPVK
D.corticolaDCO1_1s00474.t1    ---------------AASAAVELNK-RVTPLDVKLETLNNSEIKAVVTNNGNTDLNLFIRNSFLDSASIE
D.corticolaDCO1_9s03184.t1    NVFAYTT--EDLAANVTLVTKSTADEHLPKVDVKLTNTGNTLIEAAITNTGEYDLAILKLNNIFDHRNVR
D.corticolaDCO1_134s10820.t1  MKFIAGLSF--LASLAAAVSIDVNK-RDTPLEVKLELLGNTGVKASITNTGASALKLFKVGTLLDEQPVE
D.corticolaDCO1_35s05894.t1   K-LILQLCLLALATLA------AC---TGQVNVTLSSVGRTVMEAIITNTGQCDLNLFNKATILGDAPIQ
D.corticolaDCO1_25s04810.t1   TALDKMDDEKYKSYLSAWFGDDDNH-RDTVRQVYKNFVGSNHNKEGA-----------------------
M.phaseolina|EKG16975.1|      MA--------------------------------------------------------------------

                              71 74 76 78 80 82 84 86 88 90 92 94 96 98 100 102 104 106 108 110 112 114 116 118 120 122 124 126 128 130 132 134 136 140
MER001394 deuterolysin        KVSLFRQRPTELPFQGIKQRFRTE-GLTEDALTVLAPGESIEDEFDIAATSDLSEGGSITISTDGFVPIA
D.corticolaDCO1_3s01145.t1    KTEVNSA-ESRVAFDGIRLRMATS-NLSEDAFQSIAAGESVEVQFDIAEAHDLSSGGAFDILSSGAISYA
D.corticolaDCO1_60s09111.t1   KAEIYSA-ESRVNFQGVRLRTLTS-SLSDDAFQSIAAGETIEVTFDVGSTHDLSTGGSYDILAQGAIPFA
M.phaseolina|EKG22096.1|      KVSMYFANTTQVPFRGVRFGLPTRGELADTSFTHVPVGHSTTVSFDAAELYDLTAGGIFTAKAIGSIPYA
M.phaseolina|EKG20889.1|      KAEVYGA-DSRVGFDGIRQRTNMK-DLDNDSFAAIPAGQSIEATFDVAQTHDLTQGGDFDILTQGAIPYA
M.phaseolina|EKG20509.1|      KVEIYSA-ESRVGFEGVRLRTLTT-GLSDDAFQSLAAGETIEVTFDVGTTHNLSSGGAFDLLSQGAIPYA
M.phaseolina|EKG16613.1|      KLQIHTA-DSQVAFDGIRYQVSTH-GLTEDAFQSIAAGETIEVEFDTAESHDLSSGGAINILAQGAFSIA
N.parvum|EOD48721.1|          KVYIFDQNNTAPQFDGVRFNIVTTGNFSADYFTHVKVGQSTTTVFDVAEEYDLSAGGKFTAVAVDTILYA
N.parvum|XP_007583787.1|      KVYIFDQNNTAPQFDGVRFNIVTTGNFSADYFTHVKVGQSTTTVFDVAEEYDLSAGGKFTAVAVDTILYA
D.corticolaDCO1_1s00474.t1    KAEVYSA-DSRVAFDGIRQRIGTN-NLTSDSFQNIPAGASIEATWDTAEMHDLSAGGDFDILTQGAIPYA
D.corticolaDCO1_9s03184.t1    KVDINAATGTPLAFEGIYV-TPSR-RTTADSWIHIAKGDVFVTKFDAADLYDLSSGGNYTVKAEGSAPIR
D.corticolaDCO1_134s10820.t1  KVEVHAA-DKQVTFDGIRYQISTQ-GLTEEAFQSIAAGETIEVEFDAAETHDLSEGGAVELLSQGAFSYA
D.corticolaDCO1_35s05894.t1   KVFMLSQG-VLLPFEGVRFSPPGPGNYTSGDFVHLSIGNSVVVYFDASEEYDLNAGGAFTVVAIGSIPYA
D.corticolaDCO1_25s04810.t1   ----------------------------------------------------------------------
M.phaseolina|EKG16975.1|      ----------------------------------------------------------------------
```



```
                              141  144 146 148 150 152 154 156 158 160 162 164 166 168 170 172 174 176 178 180 182 184 186 188 190 192 194 196 198 200 202 204 206  210
MER001394 deuterolysin        TGN--KITG-SVPYSSNELSIEVDAAQASVA----SAVKPL---DKRTKV-ASCSGTRSSALSTAL---K
D.corticolaDCO1_3s01145.t1    ADGTTDIAG-AVPYISNKISTTVDGAQAAAA---RESF------VKRIAVNADCTGTRRTATVNAV---S
D.corticolaDCO1_60s09111.t1   EADSTELTG-ALSYLSNKISANVNGAEAAKV---RRDF------SKRAALQSDCTSTRGTATRTAL---T
M.phaseolina|EKG22096.1|      EGNSTTLDGKRVQYQSNVVEMDVNGPEANKIRTASTKLKVP---ITGLDIKDSCEGDLLEKLKKAVTGTG
M.phaseolina|EKG20889.1|      EGNSTKIVG-AVPYLSNRIKANVLSSHAAKV---RRSFLDQIAEIKKRSSLTDCSGDKGTSTTTAM---Q
M.phaseolina|EKG20509.1|      EADSTEITG-ALSYISNKITANINGAEAEKV---RRDFSAK---VKRTALQSDCTSSRGTATRTAL---T
M.phaseolina|EKG16613.1|      EADSTEISG-VVPFVSNSVSTEVDGAQARAV---HNEFHEK---MKRTIVQSDCTGTRGTATRNAI---A
N.parvum|EOD48721.1|          EGESTFLNNTVVPYNSNVVEIDVDGAEAAKIQTVGRKLSSM---QKRLVIKDSCDGDLGNKMDAAVDGKG
N.parvum|XP_007583787.1|      EGESTFLNNTVVPYNSNVVEIDVDGAEAAKIQTVGRKLSSM---QKRLVIKDSCDGDLGNKMDAAVDGKG
D.corticolaDCO1_1s00474.t1    NPNSTTIVG-AVPYLSNRVRANVLSSHAIKV---RRNFLGKIEKLKKRSTLTDCSGDKKTSTETAM---T
D.corticolaDCO1_9s03184.t1    SST-GYTDG-QLVYSSNTLYMGVDGAEAARHMAARRDAEDR--DLAAIQNGSGCTDEQWERILQGM---A
D.corticolaDCO1_134s10820.t1  DADSTVITG-VVPFTSNVVSTSVDGAAAKAV---RESFHEN---IKRTIVQSSCTGSQGTATRNAI---S
D.corticolaDCO1_35s05894.t1   EGNSTSLKGDAVAFTSNTLAIDIAPIVDGKILTARTKFHSL---VTGVDVASSCENDLLDRVEAAIVGDG
D.corticolaDCO1_25s04810.t1   -----DVLGNVIVHNDD--YHEINGQKFCSIN---------------------NNGKTGTAYYKKKPGMH
M.phaseolina|EKG16975.1|      ------------------------------------------------------------DLENAI---L
```

```
                              210  214 216 218 220 222 224 226 228 230 232 234 236 238 240 242 244 246 248 250 252 254 256 258 260 262 264 266 268 270 272 274 276  280
MER001394 deuterolysin        KNTVSLANQAASAA-QSGSSSRFQEYFKTTSSSVRTSVAARFRAVASEASSTSSGSTTYYCTDTYG-YCSS
D.corticolaDCO1_3s01145.t1    SGCRSLALAAQSAA-SSGSASKFQEYFKSTASATRSQVAAIFGRVASECGSTTSGVADYYCTDVYGNGCQS
D.corticolaDCO1_60s09111.t1   TNCRSLALAASSAA-VSGSSSKFSEYFKTTSSSTRSAVAATFSKVATECGSTTSGASDYYCTDVYG-YCES
M.phaseolina|EKG22096.1|      GGCRRQALAGMEDAKVKDTTALFKKVFQKDDTTYKRAVVNRLGEIADECEKVEKGPIKFYCEDVWG-FCYD
M.phaseolina|EKG20889.1|      QNCQKLASAAADAS-TADN-AAFTEFFKNADPA---EVKKVLTAVAEECGSTTGGSATYYCDQAVG-GCQS
M.phaseolina|EKG20509.1|      TNCRSLALRASSAA-VSGSSTKFSEYFKTTSSSTRSAVAATFSRAAAECGSTTGGASDYYCTDVYG-YCES
M.phaseolina|EKG16613.1|      ANCRSLAARAQTAA-SSGSATKFQEYFKTTAASTRSQVAAVFGRIASECGSTTSGSSRYYCSDVLG-ACRS
N.parvum|EOD48721.1|          GGCTEQAKVAAEAA-ENG--EKLKKYFEKNDEDTKKTVAARLRKVSAECDSENE-PIAIYCEDVWG-WCYD
N.parvum|XP_007583787.1|      GGCTEQAKVAAEAA-ENG--EKLKKYFEKNDEDTKKTVAARLRKVSAECDSENE-PIAIYCEDVWG-WCYD
D.corticolaDCO1_1s00474.t1    TNCQKLASAAAEAA-TSDN-AAFTDIFKQADAA---EVKKVLTAVAEECGSTTSGSATYYCDQAVG-GCQS
D.corticolaDCO1_9s03184.t1    AYCHRMATRGAQGA-YDGD-IRFADYFNTDLPLVRLFVQCKLLAVADACQLVGSGKARITCLDIFQ-LCNY
D.corticolaDCO1_134s10820.t1  SACRSLAQRAQTAA-SSGSATKFQEYFKSTSTSTRSTVAAVFGRIVSECGSTTSGSSRYYCNDVLG-ACSN
D.corticolaDCO1_35s05894.t1   GGCANQADAAVDDAELKDTSKQFKRYFHTDEQVHKDRVIARFQSIADQCEAIERGSIKFHCEDVLG-FCYI
D.corticolaDCO1_25s04810.t1   HFCPK--------------------FFERKSKDE--YIKDKCAG------------------------IAQ
M.phaseolina|EKG16975.1|      LGCQTLAEGAAAAA-ANEPNRDPDQYVNYT--------RDKFNAVADECVPRAWTQGRITCIDNVF-ECEW
```



```
                              287  290 292 294 296 298 300 302 304 306 308 310 312 314 316 318 320 322 324 326 328 330  334
MER001394 deuterolysin        P G V H L P A Y N I I A N C D I Y Y T Y L S A L T R T C H A Q D Q A T T T L H E F T H A P G V Y
D.corticolaDCO1_60s09111.t1   L A Y T I P S Q S I V V N C P L Y Y S A L P A L T S S C H S Q D Q A T T T L H E A T H L T E I K
M.phaseolina|EKG20509.1|      L A Y T I P S Q A I M V N C P L Y Y S A L P A L S S T C H A Q D Q A T T T L H E S T H L T S V Q
M.phaseolina|EKG16613.1|      L A Y T Q P A S S I M V N C P L F F S G L A P T S S V C H D Q D Q Q T T I L H E V T H L S Q V A
D.corticolaDCO1_134s10820.t1  L A Y T Q P S T S I M V N C P L F F S A L S P T S S A C H D Q D Q Q T T I L H E M T H L T Q I R
D.corticolaDCO1_1s00474.t1    L A Y T M P S V S L V V N C D L Y Y S A L P D L T T T C Y D Q D Q A T T T L H E F T H L S E I K
M.phaseolina|EKG20889.1|      L A Y T M P S Q S L V V N C D L Y Y S A L P D L T T T C Y D Q D Q A T T T L H E F T H L S E I K
D.corticolaDCO1_3s01145.t1    L A Y T L P S S S Y M V N C P L F F S G L P A T S R T C H A Q D Q Q T T T L H E M T H L T E I A
M.phaseolina|EKG16975.1|      L A T T Q P Y T G R I L I C P S G F I - L P V F T Q H C H V Q D M A T T V L H E V T H L Y G V H
D.corticolaDCO1_9s03184.t1    L A V T D A Q A G I I V I C P L S L Q Y S P A F S P G C H Q H D L A T T L V H E L T H I D G L H
N.parvum|EOD48721.1|          A A Y T L N Q D N L I S I C P G S F N - M P S V T E K C H E V D L T G I L I H E L T H L D R V Y
N.parvum|XP_007583787.1|      A A Y T L N Q D N L I S I C P G S F N - M P S V T E K C H E V D L T G I L I H E L T H L D R V Y
D.corticolaDCO1_35s05894.t1   A A Y A F D D Y N L I C L C D P A H G - F P A L T Q T C G D V D M T G I V I H E L S H L R N I Y
M.phaseolina|EKG22096.1|      A A Y T D D N Y N Q I C L C D K A H D - L P A F T D T C H D V D L A G V I V H E L S H L S D I D
D.corticolaDCO1_25s04810.t1   - - - - - - - - - - - - - - - - - - - - - - - - N T G S I G R V F Q G A N V L H E F M H Y S K V G

                              335  338 340 342 344 346 348 350 352 354 356 358 360 362 364 366 368  371
MER001394 deuterolysin        S P G T D D L - G Y G - Y D A A T A L S S S Q A L N N V D T Y A L F A N G
D.corticolaDCO1_60s09111.t1   - - G T D D L - G Y G - Y S A A T A L S A S Q A L N N A D T Y A L Y S N A
M.phaseolina|EKG20509.1|      - - G T D D L - G Y G - Y S A A T R L S T S Q A L N N A D T Y A L Y A N A
M.phaseolina|EKG16613.1|      - - G T S D Y G G Y G - Y N F V R S L S A A Q N L N H A D T Y T L Y A Q A
D.corticolaDCO1_134s10820.t1  - - G T S D Y G G Y G - Y N F V R S L T A A Q N L N H A D T Y T L Y A - Q
D.corticolaDCO1_1s00474.t1    - - G T D D L - G Y G - Y S A A T Q L T G A Q A L N N A D S Y A L F A N A
M.phaseolina|EKG20889.1|      - - G T D D L - G Y G - Y S A A T Q L S G S Q A I N N A D S Y A L F A N A
D.corticolaDCO1_3s01145.t1    - - G T T D Q G A Y G - Y A A V Q A L T A A Q N L R H A D T Y T L Y A Q A
M.phaseolina|EKG16975.1|      H P - T L D F Y S G D V W P G C K S L P D Y L A K Y S A E T Y A S Y A Q C
D.corticolaDCO1_9s03184.t1    T R G T G D Y P R N T - W P Y C K D L D W L D A I D N A Q T Y S S F A Q S
N.parvum|EOD48721.1|          N P A T D D F - A A E - W P A A A Q L N N M Q A L M N G D T Y R L Y A Q D
N.parvum|XP_007583787.1|      N P A T D D F - A A E - W P A A A Q L N N M Q A L M N G D T Y R L Y A Q D
D.corticolaDCO1_35s05894.t1   D P P T R D Y - A T L L S G G A A D L E A M Y A V E N A D T Y R L Y A Q A
M.phaseolina|EKG22096.1|      T P R C Q D F - A V G - W D A I T K I D A A N T I I N A D S Y R L Y A Q D
D.corticolaDCO1_25s04810.t1   K E A T G T A - P A G A - Y N C Y D T D K T Q T I L N A D S Y V - - - - Y
```

Figure 29. Protein alignment of M35 protein sequences of *M. phaseolina, N. parvum* and *D. corticola* aligned using T-Coffee software with default settings. The motifs are

indicated by boxes.

Despite all the sequences had been thoroughly analysed before been classify in families, in some sequences the conserved regions were not found. Therefore, those protein sequences were eliminated: *M.phaseolina*|EKG10265.1| and *D.corticola*|DCO1_106s10656.t1

For M35 family proteases, the active zinc ligands are composed of 2 histidines in the HEXXH motif and the Aspartic acid in motif GTXDXXYG. In it is visible the HEXXH motif, highlighted by a box. All the 14 protein sequences align perfectly in this conserved region. However, substitutions of the original amino acid can occur due to duplications phenomenon for instance. It is generally thought that side-chain of residues can influence the flexibility of the ligand-binding site of a protein. Therefore, these substitutions may exert influences on the binding of the zinc ion.(176)

In the sequence *M.phaseolina*|EKG22096.1|, at site 338, T changed to C and that substitution can influence the coordination of the zinc ligands and protein flexibilities. This amino acid change is considered important for M35 family gene because the hydroxyl group of T338 can interact with the second zinc binding histidine H329 playing an important role in sustaining the coordination of the catalytic zinc ligands.(173)

In a more complex way, the motif is composed by: Xaa-Xbb-Xcc-His-Glu-Xbb-Xbb-His-Xbb-Xdd. Xaa is hydrophobic or Thr; Xbb is an uncharged residue, Xcc is any amino acid except Pro, and Xdd is hydrophobic. The sequences analysed in this work do not follow those restricted parameters. We can deduce that even though the conserved regions aligned with the consensus sequence, the remaining sequence are very heterogeneous.

3.4.3. Serine proteases – family S8

A majority of the serine peptidases belong to the S8 family (44 sequences). Also known as the subtilase family, family S8 is the second largest family of serine peptidases, both in terms of number of sequences and of characterised peptidases. Subtilases are widespread, being found in eubacteria, archaebacteria, eukaryotes and viruses.(90)

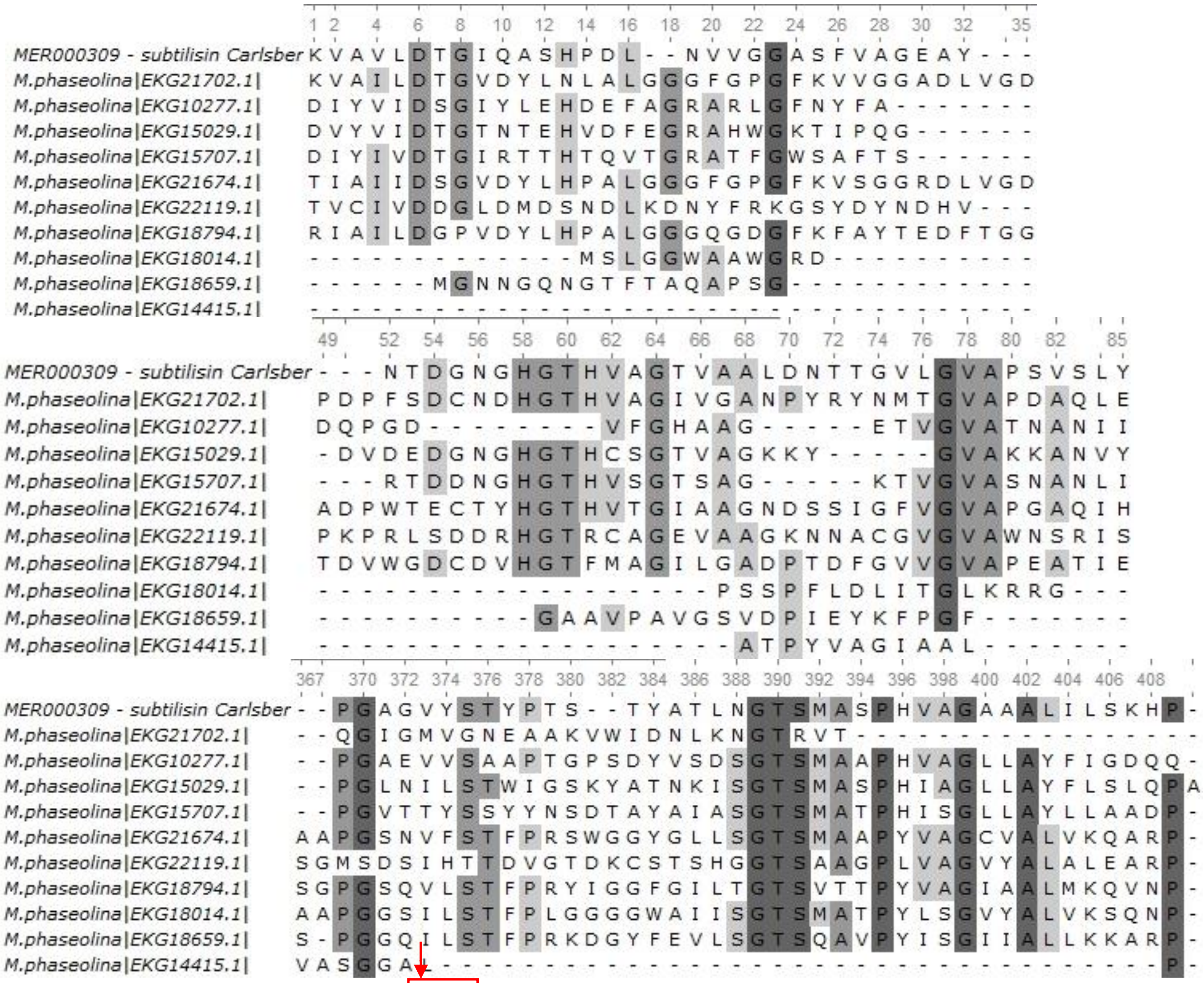
Members of family S8 are divided into two subfamilies: subtilisin the type-example for subfamily S8A and kexin the type-example for subfamily S8B. Several previous studies indicate that subfamily S8A peptidases members secreted from bacteria can been implicated in the pathogenesis. In subfamily S8B are grouped enzymes such as kexin and furin, typically found in yeasts.(87)

In the family S8 the catalytic triad order is always Asp, His and Ser, however, depending to what subfamily the enzyme belongs, the residues surrounding this motif differ. In subfamily S8A, the active site residues frequently occurs in the motifs Asp-Thr/Ser-Gly (which is similar to the sequence motif in families of aspartic endopeptidases), His-Gly-Thr-His and Gly-Thr-Ser-Met-Ala-Xaa-Pro. In subfamily S8B, the catalytic residues frequently occur in the motifs Asp-Asp-Gly, His-Gly-Thr-Arg and Gly-Thr-Ser-Ala/Val-Ala/Ser-Pro.(90)

A multiple sequence alignment was performed with the protein sequences of serine proteases and the reference sequence (subtilisin Carlsberg (Bacillus licheniformis), MEROPS Accession MER000309)

3.4.3.1. *Macrophomina phaseolina*

In *M. phaseolina* genome a total of 10 protein sequences of secreted serine proteases belonging to S8 family were identified. A multiple sequence alignment covered just a part of the protein sequence: the functional domain. The range of the functional domain in each sequence is shown in Annexe 5.6.1.3.



```
                                   1 2   4   6   8   10  12  14  16  18  20  22  24  26  28  30  32   35
MER000309 - subtilisin Carlsber    K V A V L D T G I Q A S H P D L - - N V V G G A S F V A G E A Y - - -
M.phaseolina|EKG21702.1|           K V A I L D T G V D Y L N L A L G G G F G P G F K V V G G A D L V G D
M.phaseolina|EKG10277.1|           D I Y V I D S G I Y L E H D E F A G R A R L G F N Y F A - - - - - - -
M.phaseolina|EKG15029.1|           D V Y V I D T G T N T E H V D F E G R A H W G K T I P Q G - - - - - -
M.phaseolina|EKG15707.1|           D I Y I V D T G I R T T H T Q V T G R A T F G W S A F T S - - - - - -
M.phaseolina|EKG21674.1|           T I A I I D S G V D Y L H P A L G G G F G P G F K V S G G R D L V G D
M.phaseolina|EKG22119.1|           T V C I V D D G L D M D S N D L K D N Y F R K G S Y D Y N D H V - - -
M.phaseolina|EKG18794.1|           R I A I L D G P V D Y L H P A L G G G Q G D G F K F A Y T E D F T G G
M.phaseolina|EKG18014.1|           - - - - - - - - - - - - - M S L G G W A A W G R D - - - - - - - - - -
M.phaseolina|EKG18659.1|           - - - - - - M G N N G Q N G T F T A Q A P S G - - - - - - - - - - - -
M.phaseolina|EKG14415.1|           - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

                                   49    52  54  56  58  60  62  64  66  68  70  72  74  76  78  80  82   85
MER000309 - subtilisin Carlsber    - - - N T D G N G H G T H V A G T V A A L D N T T G V L G V A P S V S L Y
M.phaseolina|EKG21702.1|           P D P F S D C N D H G T H V A G I V G A N P Y R Y N M T G V A P D A Q L E
M.phaseolina|EKG10277.1|           D Q P G D - - - - - - - - V F G H A A G - - - - - E T V G V A T N A N I I
M.phaseolina|EKG15029.1|           - D V D E D G N G H G T H C S G T V A G K K Y - - - - - G V A K K A N V Y
M.phaseolina|EKG15707.1|           - - - R T D D N G H G T H V S G T S A G - - - - - K T V G V A S N A N L I
M.phaseolina|EKG21674.1|           A D P W T E C T Y H G T H V T G I A A G N D S S I G F V G V A P G A Q I H
M.phaseolina|EKG22119.1|           P K P R L S D D R H G T R C A G E V A A G K N N A C G V G V A W N S R I S
M.phaseolina|EKG18794.1|           T D V W G D C D V H G T F M A G I L G A D P T D F G V V G V A P E A T I E
M.phaseolina|EKG18014.1|           - - - - - - - - - - - - - - - - - - P S S P F L D L I T G L K R R G - - -
M.phaseolina|EKG18659.1|           - - - - - - - - - - G A A V P A V G S V D P I E Y K F P G F - - - - - - -
M.phaseolina|EKG14415.1|           - - - - - - - - - - - - - - - - - - - A T P Y V A G I A A L - - - - - - -

                                   367   370 372 374 376 378 380 382 384 386 388 390 392 394 396 398 400 402 404 406 408
MER000309 - subtilisin Carlsber    - - P G A G V Y S T Y P T S - - T Y A T L N G T S M A S P H V A G A A A L I L S K H P -
M.phaseolina|EKG21702.1|           - - Q G I G M V G N E A A K V W I D N L K N G T R V T - - - - - - - - - - - - - - - - -
M.phaseolina|EKG10277.1|           - - P G A E V V S A A P T G P S D Y V S D S G T S M A A P H V A G L L A Y F I G D Q Q -
M.phaseolina|EKG15029.1|           - - P G L N I L S T W I G S K Y A T N K I S G T S M A S P H I A G L L A Y F L S L Q P A
M.phaseolina|EKG15707.1|           - - P G V T T Y S S Y Y N S D T A Y A I A S G T S M A T P H I S G L L A Y L L A A D P -
M.phaseolina|EKG21674.1|           A A P G S N V F S T F P R S W G G Y G L L S G T S M A A P Y V A G C V A L V K Q A R P -
M.phaseolina|EKG22119.1|           S G M S D S I H T T D V G T D K C S T S H G G T S A A G P L V A G V Y A L A L E A R P -
M.phaseolina|EKG18794.1|           S G P G S Q V L S T F P R Y I G G F G I L T G T S V T T P Y V A G I A A L M K Q V N P -
M.phaseolina|EKG18014.1|           A A P G G S I L S T F P L G G G G W A I I S G T S M A T P Y L S G V Y A L V K S Q N P -
M.phaseolina|EKG18659.1|           S - P G G Q I L S T F P R K D G Y F E V L S G T S Q A V P Y I S G I I A L L K K A R P -
M.phaseolina|EKG14415.1|           V A S G G A - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - P -
```

Figure 30. Protein alignment of S8 protein sequences of *M. phaseolina* aligned using T-Coffee software with default settings. The conserved regions are indicated by boxes.

Analysing the MSA of the peptidase family S8, the functional domain was not found in all sequences. Despite all the sequences had been thoroughly analysed before been classify into families (using SMART and InterPro tools), the conserved regions of some sequences were not identified. The protein EKG14415.1 did not align in any conserved regions with MEROPS S8 representative sequence. Besides, the sequences with the accession numbers EKG18014.1 and EKG18659.1 did not have the amino acid region representative of the first and second sequence motifs. The domain organisations of both of these sequences can is shown in Figure 31 and Figure 32. This domain architecture representation was constructed by SMART analysis service. Despite the manual analysis of the MSA reveal the lack of two active site residues, these representation confirms that both of these this sequences include peptidase S8 domain.

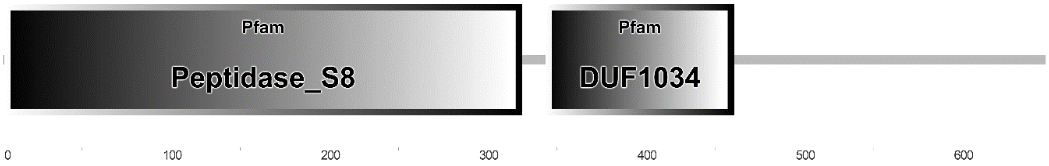


Figure 31. Sequence organisation with domain architecture representation of the sequence EKG18014.1

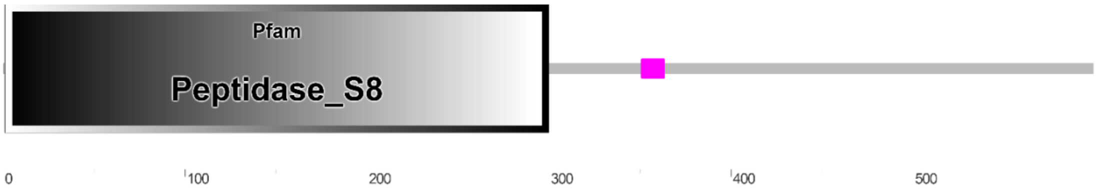
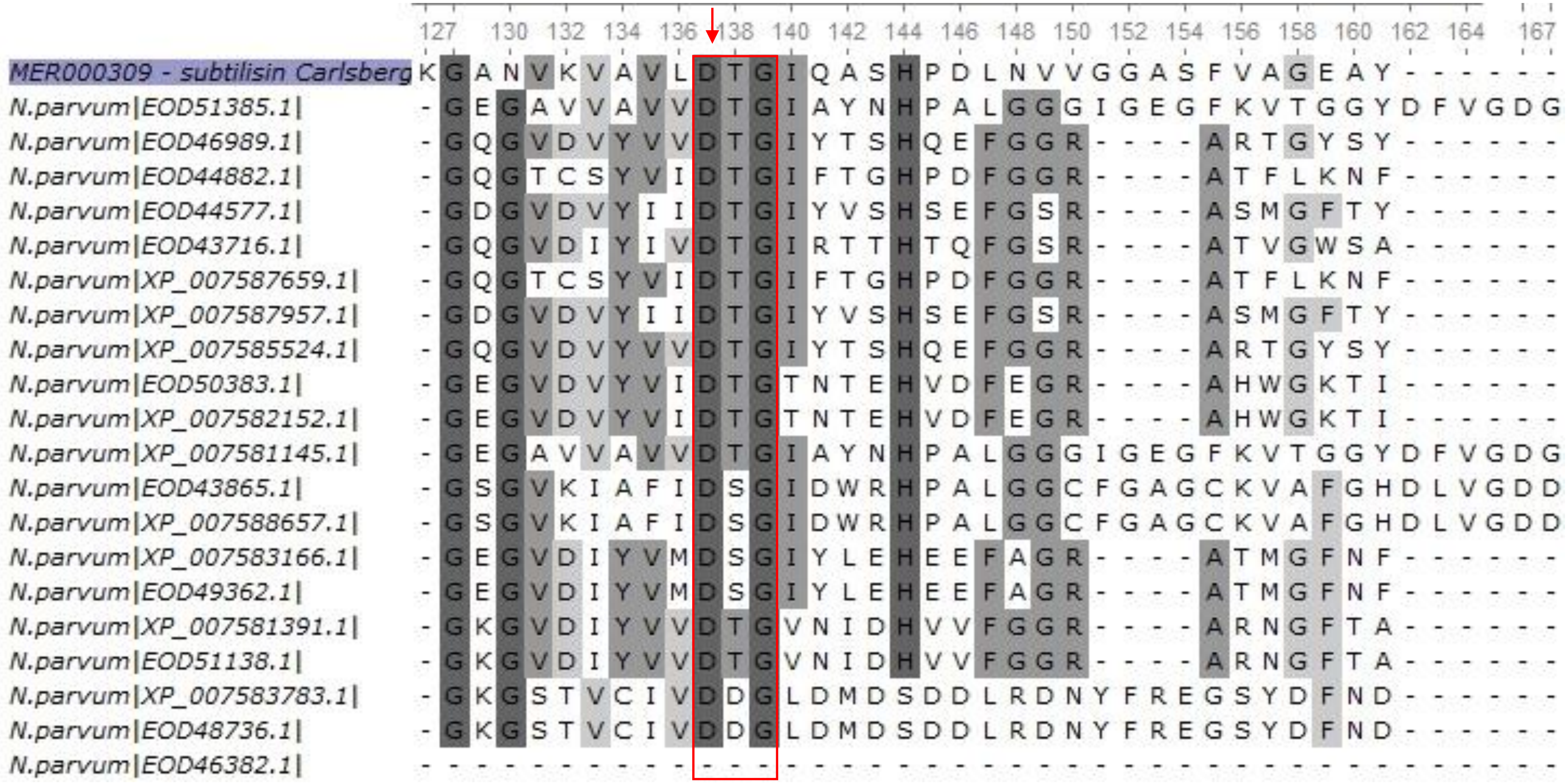


Figure 32. Sequence organisation with domain architecture representation of the sequence EKG18659.1

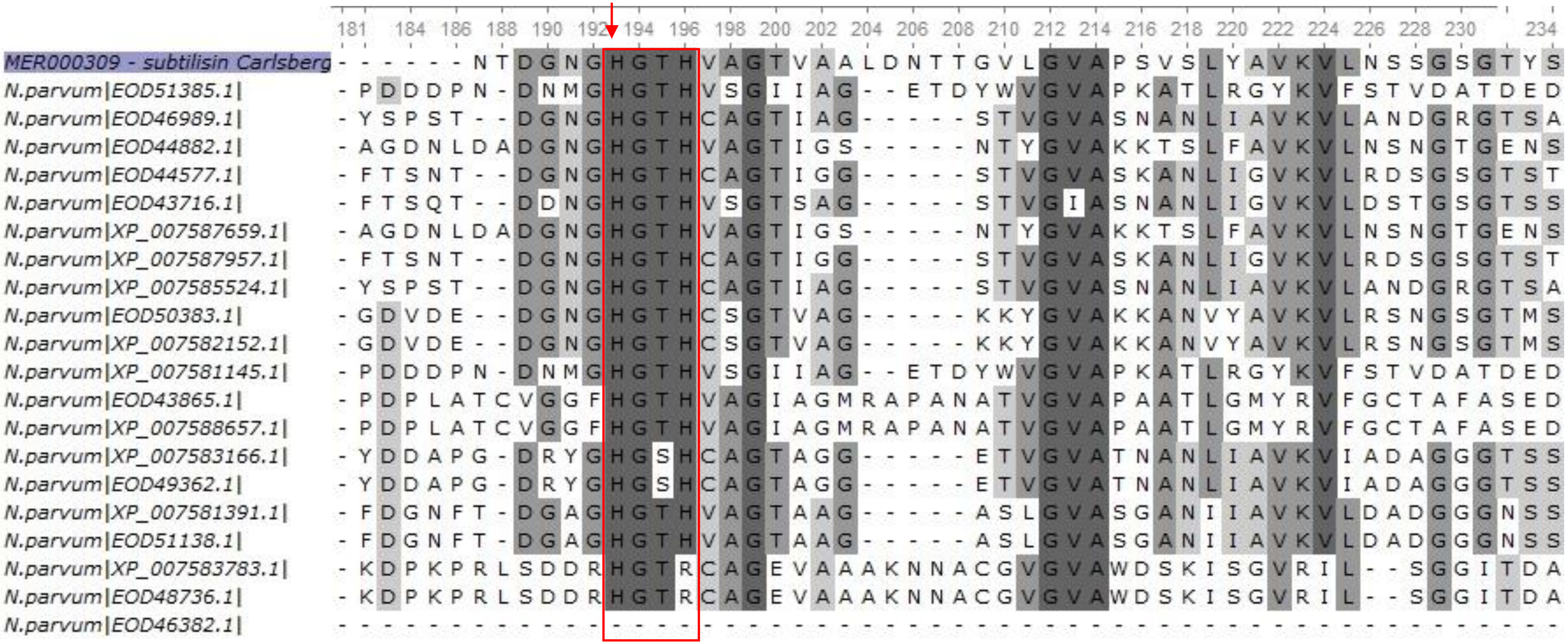
The apparent discrepancies in the alignment can also be related to the different catalytic residues order in both clans. The used tools to found the functional domain in each sequence can only classify the sequence motif in families, however, MEROPS subdivided the families. So members of these subfamilies can be identified by subtle different motifs around the active site. With that knowledge, the protein sequence EKG22119.1 can be grouped as a member of subfamily S8B because the catalytic residues frequently occur in the motifs Asp-Asp-Gly, His-Gly-Arg and Gly-Thr-Ser-Ala-Ala-Pro. All the others enzymes certainly belong to subfamily S8A. To best classify the protein sequence of an enzyme is important to have a detailed description of the family characteristics. Until now, there is no tool able to classify the enzymes into subfamilies based on the functional domain. The only alternative is to manually analyse the amino acid sequence which may be subjective and time consuming.

3.4.3.2. *Neofusicoccum parvum*

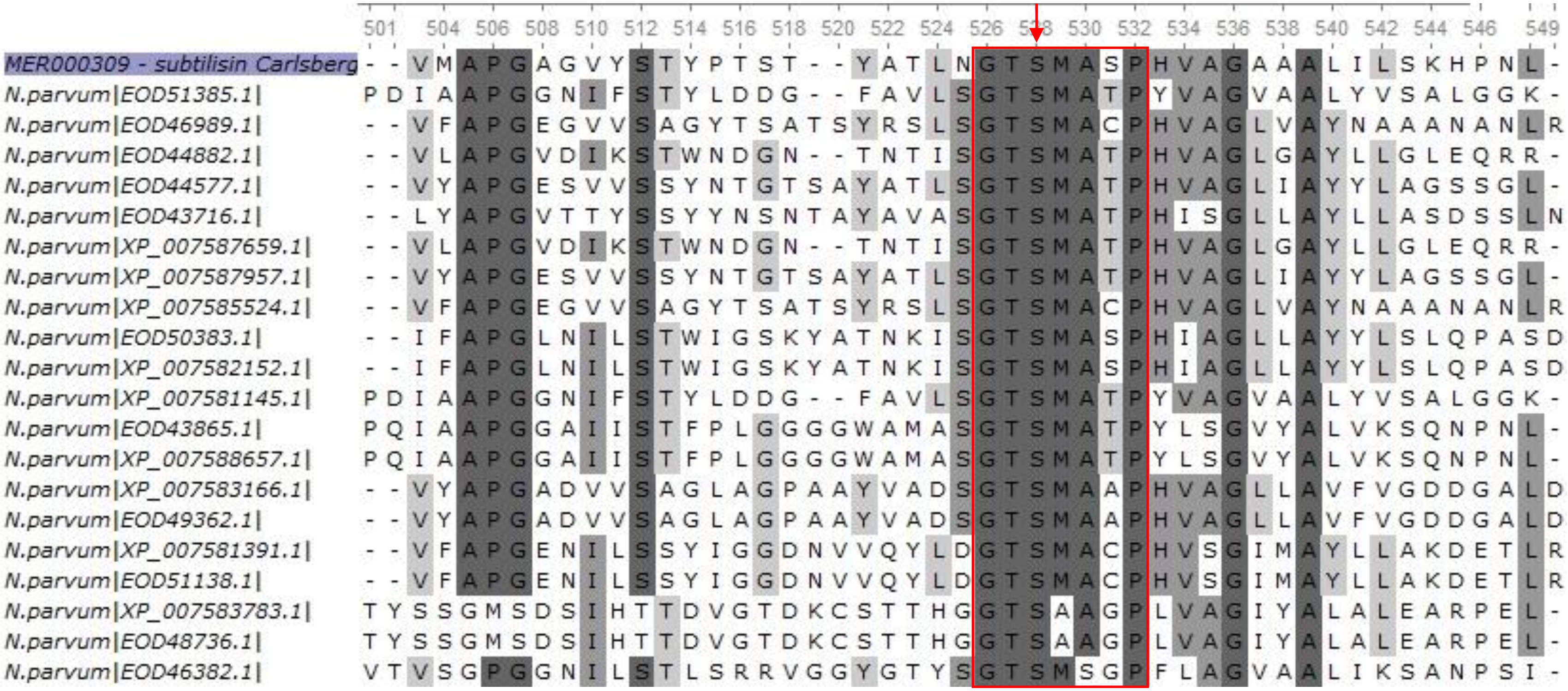
In *N. parvum* genome a total of 21 protein sequences of secreted serine proteases belonging to S8 family were identified. The multiple sequence alignment covered just a part of the protein sequence: the functional domain. The range of the functional domain in each sequence is shown in Annexe 5.6.2.3. A multiple sequence alignment was performed between those 21 protein sequences and a reference sequence (deuterolysin from *Aspergillus flavus*).



```
                              127   130 132 134 136 138 140 142 144 146 148 150 152 154 156 158 160 162 164   167
MER000309 - subtilisin Carlsberg K G A N V K V A V L D T G I Q A S H P D L N V V G G A S F V A G E A Y - - - - - -
N.parvum|EOD51385.1|             - G E G A V V A V V D T G I A Y N H P A L G G G I G E G F K V T G G Y D F V G D G
N.parvum|EOD46989.1|             - G Q G V D V Y V V D T G I Y T S H Q E F G G R - - - - A R T G Y S Y - - - - - -
N.parvum|EOD44882.1|             - G Q G T C S Y V I D T G I F T G H P D F G G R - - - - A T F L K N F - - - - - -
N.parvum|EOD44577.1|             - G D G V D V Y I I D T G I Y V S H S E F G S R - - - - A S M G F T Y - - - - - -
N.parvum|EOD43716.1|             - G Q G V D I Y I V D T G I R T T H T Q F G S R - - - - A T V G W S A - - - - - -
N.parvum|XP_007587659.1|         - G Q G T C S Y V I D T G I F T G H P D F G G R - - - - A T F L K N F - - - - - -
N.parvum|XP_007587957.1|         - G D G V D V Y I I D T G I Y V S H S E F G S R - - - - A S M G F T Y - - - - - -
N.parvum|XP_007585524.1|         - G Q G V D V Y V V D T G I Y T S H Q E F G G R - - - - A R T G Y S Y - - - - - -
N.parvum|EOD50383.1|             - G E G V D V Y V I D T G T N T E H V D F E G R - - - - A H W G K T I - - - - - -
N.parvum|XP_007582152.1|         - G E G V D V Y V I D T G T N T E H V D F E G R - - - - A H W G K T I - - - - - -
N.parvum|XP_007581145.1|         - G E G A V V A V V D T G I A Y N H P A L G G G I G E G F K V T G G Y D F V G D G
N.parvum|EOD43865.1|             - G S G V K I A F I D S G I D W R H P A L G G C F G A G C K V A F G H D L V G D D
N.parvum|XP_007588657.1|         - G S G V K I A F I D S G I D W R H P A L G G C F G A G C K V A F G H D L V G D D
N.parvum|XP_007583166.1|         - G E G V D I Y V M D S G I Y L E H E E F A G R - - - - A T M G F N F - - - - - -
N.parvum|EOD49362.1|             - G E G V D I Y V M D S G I Y L E H E E F A G R - - - - A T M G F N F - - - - - -
N.parvum|XP_007581391.1|         - G K G V D I Y V V D T G V N I D H V V F G G R - - - - A R N G F T A - - - - - -
N.parvum|EOD51138.1|             - G K G V D I Y V V D T G V N I D H V V F G G R - - - - A R N G F T A - - - - - -
N.parvum|XP_007583783.1|         - G K G S T V C I V D D G L D M D S D D L R D N Y F R E G S Y D F N D - - - - - -
N.parvum|EOD48736.1|             - G K G S T V C I V D D G L D M D S D D L R D N Y F R E G S Y D F N D - - - - - -
N.parvum|EOD46382.1|             - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```



```
                                 181   184 186 188 190 192 194 196 198 200 202 204 206 208 210 212 214 216 218 220 222 224 226 228 230     234
MER000309 - subtilisin Carlsberg - - - - - - N T D G N G H G T H V A G T V A A L D N T T G V L G V A P S V S L Y A V K V L N S S G S G T Y S
N.parvum|EOD51385.1|             - P D D D P N - D N M G H G T H V S G I I A G - - E T D Y W V G V A P K A T L R G Y K V F S T V D A T D E D
N.parvum|EOD46989.1|             - Y S P S T - - D G N G H G T H C A G T I A G - - - - - S T V G V A S N A N L I A V K V L A N D G R G T S A
N.parvum|EOD44882.1|             - A G D N L D A D G N G H G T H V A G T I G S - - - - - N T Y G V A K K T S L F A V K V L N S N G T G E N S
N.parvum|EOD44577.1|             - F T S N T - - D G N G H G T H C A G T I G G - - - - - S T V G V A S K A N L I G V K V L R D S G S G T S T
N.parvum|EOD43716.1|             - F T S Q T - - D D N G H G T H V S G T S A G - - - - - S T V G I A S N A N L I G V K V L D S T G S G T S S
N.parvum|XP_007587659.1|         - A G D N L D A D G N G H G T H V A G T I G S - - - - - N T Y G V A K K T S L F A V K V L N S N G T G E N S
N.parvum|XP_007587957.1|         - F T S N T - - D G N G H G T H C A G T I G G - - - - - S T V G V A S K A N L I G V K V L R D S G S G T S T
N.parvum|XP_007585524.1|         - Y S P S T - - D G N G H G T H C A G T I A G - - - - - S T V G V A S N A N L I A V K V L A N D G R G T S A
N.parvum|EOD50383.1|             - G D V D E - - D G N G H G T H C S G T V A G - - - - - K K Y G V A K K A N V Y A V K V L R S N G S G T M S
N.parvum|XP_007582152.1|         - G D V D E - - D G N G H G T H C S G T V A G - - - - - K K Y G V A K K A N V Y A V K V L R S N G S G T M S
N.parvum|XP_007581145.1|         - P D D D P N - D N M G H G T H V S G I I A G - - E T D Y W V G V A P K A T L R G Y K V F S T V D A T D E D
N.parvum|EOD43865.1|             - P D P L A T C V G G F H G T H V A G I A G M R A P A N A T V G V A P A A T L G M Y R V F G C T A F A S E D
N.parvum|XP_007588657.1|         - P D P L A T C V G G F H G T H V A G I A G M R A P A N A T V G V A P A A T L G M Y R V F G C T A F A S E D
N.parvum|XP_007583166.1|         - Y D D A P G - D R Y G H G S H C A G T A G G - - - - - E T V G V A T N A N L I A V K V I A D A G G G T S S
N.parvum|EOD49362.1|             - Y D D A P G - D R Y G H G S H C A G T A G G - - - - - E T V G V A T N A N L I A V K V I A D A G G G T S S
N.parvum|XP_007581391.1|         - F D G N F T - D G A G H G T H V A G T A A G - - - - - A S L G V A S G A N I I A V K V L D A D G G G N S S
N.parvum|EOD51138.1|             - F D G N F T - D G A G H G T H V A G T A A G - - - - - A S L G V A S G A N I I A V K V L D A D G G G N S S
N.parvum|XP_007583783.1|         - K D P K P R L S D D R H G T R C A G E V A A A K N N A C G V G V A W D S K I S G V R I L - - S G G I T D A
N.parvum|EOD48736.1|             - K D P K P R L S D D R H G T R C A G E V A A A K N N A C G V G V A W D S K I S G V R I L - - S G G I T D A
N.parvum|EOD46382.1|             - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```
]



```
                                 501   504 506 508 510 512 514 516 518 520 522 524 526 528 530 532 534 536 538 540 542 544 546   549
MER000309 - subtilisin Carlsberg - - V M A P G A G V Y S T Y P T S T - - Y A T L N G T S M A S P H V A G A A A L I L S K H P N L -
N.parvum|EOD51385.1|             P D I A A P G G N I F S T Y L D D G - - F A V L S G T S M A T P Y V A G V A A L Y V S A L G G K -
N.parvum|EOD46989.1|             - - V F A P G E G V V S A G Y T S A T S Y R S L S G T S M A C P H V A G L V A Y N A A A N A N L R
N.parvum|EOD44882.1|             - - V L A P G V D I K S T W N D G N - - T N T I S G T S M A T P H V A G L G A Y L L G L E Q R R -
N.parvum|EOD44577.1|             - - V Y A P G E S V V S S Y N T G T S A Y A T L S G T S M A T P H V A G L I A Y Y L A G S S G L -
N.parvum|EOD43716.1|             - - L Y A P G V T T Y S S Y Y N S N T A Y A V A S G T S M A T P H I S G L L A Y L L A S D S S L N
N.parvum|XP_007587659.1|         - - V L A P G V D I K S T W N D G N - - T N T I S G T S M A T P H V A G L G A Y L L G L E Q R R -
N.parvum|XP_007587957.1|         - - V Y A P G E S V V S S Y N T G T S A Y A T L S G T S M A T P H V A G L I A Y Y L A G S S G L -
N.parvum|XP_007585524.1|         - - V F A P G E G V V S A G Y T S A T S Y R S L S G T S M A C P H V A G L V A Y N A A A N A N L R
N.parvum|EOD50383.1|             - - I F A P G L N I L S T W I G S K Y A T N K I S G T S M A S P H I A G L L A Y Y L S L Q P A S D
N.parvum|XP_007582152.1|         - - I F A P G L N I L S T W I G S K Y A T N K I S G T S M A S P H I A G L L A Y Y L S L Q P A S D
N.parvum|XP_007581145.1|         P D I A A P G G N I F S T Y L D D G - - F A V L S G T S M A T P Y V A G V A A L Y V S A L G G K -
N.parvum|EOD43865.1|             P Q I A A P G G A I I S T F P L G G G G W A M A S G T S M A T P Y L S G V Y A L V K S Q N P N L -
N.parvum|XP_007588657.1|         P Q I A A P G G A I I S T F P L G G G G W A M A S G T S M A T P Y L S G V Y A L V K S Q N P N L -
N.parvum|XP_007583166.1|         - - V Y A P G A D V V S A G L A G P A A Y V A D S G T S M A A P H V A G L L A V F V G D D G A L D
N.parvum|EOD49362.1|             - - V Y A P G A D V V S A G L A G P A A Y V A D S G T S M A A P H V A G L L A V F V G D D G A L D
N.parvum|XP_007581391.1|         - - V F A P G E N I L S S Y I G G D N V V Q Y L D G T S M A C P H V S G I M A Y L L A K D E T L R
N.parvum|EOD51138.1|             - - V F A P G E N I L S S Y I G G D N V V Q Y L D G T S M A C P H V S G I M A Y L L A K D E T L R
N.parvum|XP_007583783.1|         T Y S S G M S D S I H T T D V G T D K C S T T H G G T S A A G P L V A G I Y A L A L E A R P E L -
N.parvum|EOD48736.1|             T Y S S G M S D S I H T T D V G T D K C S T T H G G T S A A G P L V A G I Y A L A L E A R P E L -
N.parvum|EOD46382.1|             V T V S G P G G N I L S T L S R R V G G Y G T Y S G T S M S G P F L A G V A A L I K S A N P S I -
```

Figure 33. Protein alignment of S8 protein sequences of *N. parvum* aligned using T-Coffee software with default settings. The conserved regions are indicated by boxes.

The sequence with the accession number EOD46382.1 did not have the amino acid region representative of the first and second functional motifs. Despite the manual analysis of the MSA reveal the lack of two active site residues, Pfam tool indicates that this sequence includes peptidase S8 domain.

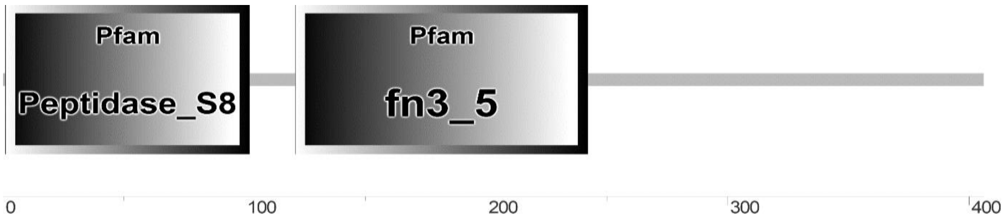


Figure 34. Sequence organisation with domain architecture representation of the sequence EOD46382.1

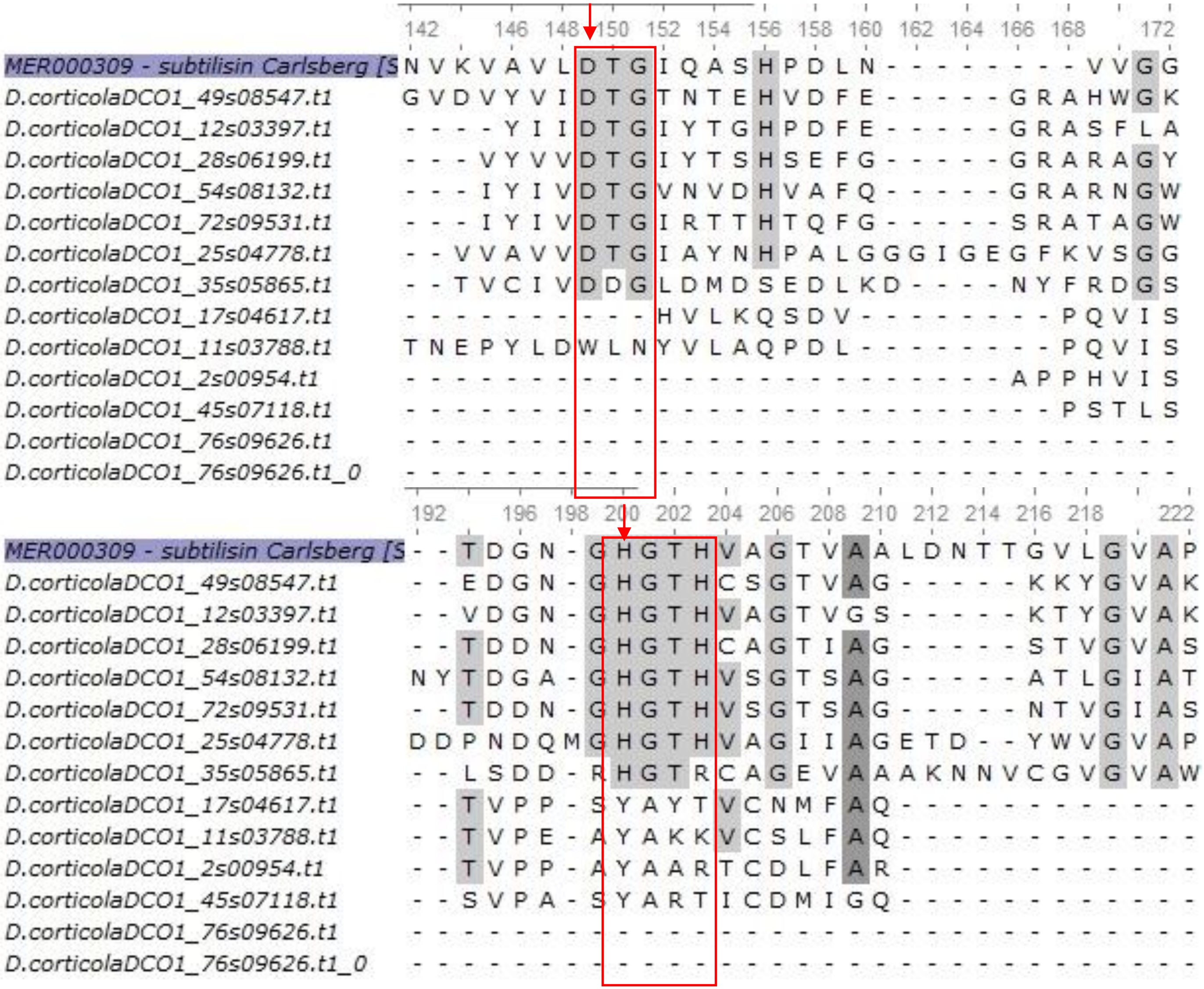
The protein sequence EOD48736.1 and XP_007583783.1 can be grouped as a member of subfamily S8B because the catalytic residues frequently occur in the motifs Asp-Asp-Gly, His-Gly-Thr-Arg and Gly-Thr-Ser-Ala-Ala -Pro. All the others enzymes certainly belong to subfamily S8A.

In general, the multiple sequence alignment of S8 enzymes shows several conserved regions between the sequences. Although the number of protein sequences is higher than in *M. phaseolina* or *D. corticola*, the sequences from *N. parvum* show a much better alignment. The term family is used to describe a group of peptidases in which each member shows evolutionary relationship to at least one other, either throughout the whole sequence or at least in the part of the sequence responsible for catalytic activity. The similarity in these sequences seems to reflect a closer evolutionary relationships.

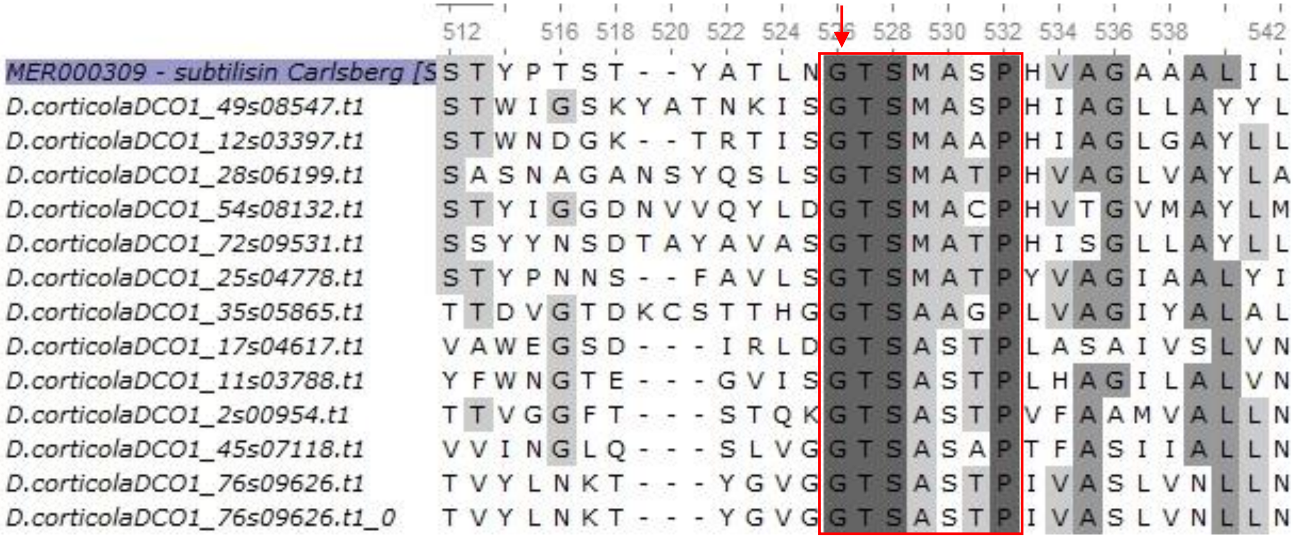
3.4.3.3. *Diplodia corticola*

In *D. corticola* genome a total of 13 protein sequences of secreted serine proteases belonging to S8 family were identified. A multiple sequence alignment covered just a part of the protein sequence: the function domain. (Figure 35).

The range of the functional domain in each sequence is shown in Annexe 5.6.3.3.



```
                                    142     146 148 150 152 154 156 158 160 162 164 166 168     172
MER000309 - subtilisin Carlsberg [S N V K V A V L D T G I Q A S H P D L N - - - - - - - - V V G G
D.corticolaDCO1_49s08547.t1         G V D V Y V I D T G T N T E H V D F E - - - - - G R A H W G K
D.corticolaDCO1_12s03397.t1         - - - - Y I I D T G I Y T G H P D F E - - - - - G R A S F L A
D.corticolaDCO1_28s06199.t1         - - - V Y V V D T G I Y T S H S E F G - - - - - G R A R A G Y
D.corticolaDCO1_54s08132.t1         - - - I Y I V D T G V N V D H V A F Q - - - - - G R A R N G W
D.corticolaDCO1_72s09531.t1         - - - I Y I V D T G I R T T H T Q F G - - - - - S R A T A G W
D.corticolaDCO1_25s04778.t1         - - V V A V V D T G I A Y N H P A L G G G I G E G F K V S G G
D.corticolaDCO1_35s05865.t1         - - T V C I V D D G L D M D S E D L K D - - - - N Y F R D G S
D.corticolaDCO1_17s04617.t1         - - - - - - - - - - H V L K Q S D V - - - - - - - - P Q V I S
D.corticolaDCO1_11s03788.t1         T N E P Y L D W L N Y V L A Q P D L - - - - - - - - P Q V I S
D.corticolaDCO1_2s00954.t1          - - - - - - - - - - - - - - - - - - - - - - - - A P P H V I S
D.corticolaDCO1_45s07118.t1         - - - - - - - - - - - - - - - - - - - - - - - - - - P S T L S
D.corticolaDCO1_76s09626.t1         - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
D.corticolaDCO1_76s09626.t1_0       - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

                                    192     196 198 200 202 204 206 208 210 212 214 216 218     222
MER000309 - subtilisin Carlsberg [S - - T D G N - G H G T H V A G T V A A L D N T T G V L G V A P
D.corticolaDCO1_49s08547.t1         - - E D G N - G H G T H C S G T V A G - - - - - K K Y G V A K
D.corticolaDCO1_12s03397.t1         - - V D G N - G H G T H V A G T V G S - - - - - K T Y G V A K
D.corticolaDCO1_28s06199.t1         - - T D D N - G H G T H C A G T I A G - - - - - S T V G V A S
D.corticolaDCO1_54s08132.t1         N Y T D G A - G H G T H V S G T S A G - - - - - A T L G I A T
D.corticolaDCO1_72s09531.t1         - - T D D N - G H G T H V S G T S A G - - - - - N T V G I A S
D.corticolaDCO1_25s04778.t1         D D P N D Q M G H G T H V A G I I A G E T D - - Y W V G V A P
D.corticolaDCO1_35s05865.t1         - - L S D D - R H G T R C A G E V A A A K N N V C G V G V A W
D.corticolaDCO1_17s04617.t1         - - T V P P - S Y A Y T V C N M F A Q - - - - - - - - - - - -
D.corticolaDCO1_11s03788.t1         - - T V P E - A Y A K K V C S L F A Q - - - - - - - - - - - -
D.corticolaDCO1_2s00954.t1          - - T V P P - A Y A A R T C D L F A R - - - - - - - - - - - -
D.corticolaDCO1_45s07118.t1         - - S V P A - S Y A R T I C D M I G Q - - - - - - - - - - - -
D.corticolaDCO1_76s09626.t1         - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
D.corticolaDCO1_76s09626.t1_0       - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```



```
                                   512   516 518 520 522 524 526 528 530 532 534 536 538   542
MER000309 - subtilisin Carlsberg [S  STYPTST--YATLNGTSMASPHVAGAAALIL
D.corticolaDCO1_49s08547.t1          STWIGSKYATNKISGTSMASPHIAGLLAYYL
D.corticolaDCO1_12s03397.t1          STWNDGK--TRTISGTSMAAPHIAGLGAYLL
D.corticolaDCO1_28s06199.t1          SASNAGANSYQSLSGTSMATPHVAGLVAYLA
D.corticolaDCO1_54s08132.t1          STYIGGDNVVQYLDGTSMACPHVTGVMAYLM
D.corticolaDCO1_72s09531.t1          SSYYNSDTAYAVASGTSMATPHISGLLAYLL
D.corticolaDCO1_25s04778.t1          STYPNNS--FAVLSGTSMATPYVAGIAALYI
D.corticolaDCO1_35s05865.t1          TTDVGTDKCSTTHGGTSAAGPLVAGIYALAL
D.corticolaDCO1_17s04617.t1          VAWEGSD---IRLDGTSASTPLASAIVSLVN
D.corticolaDCO1_11s03788.t1          YFWNGTE---GVISGTSASTPLHAGILALVN
D.corticolaDCO1_2s00954.t1           TTVGGFT---STQKGTSASTPVFAAMVALLN
D.corticolaDCO1_45s07118.t1          VVINGLQ---SLVGGTSASAPTFASIIALLN
D.corticolaDCO1_76s09626.t1          TVYLNKT---YGVGGTSASTPIVASLVNLLN
D.corticolaDCO1_76s09626.t1_0        TVYLNKT---YGVGGTSASTPIVASLVNLLN
```

Figure 35. Protein alignment of S8 protein sequences of *D. corticola* aligned using T-Coffee software with default settings. The conserved regions are indicated by boxes.

The sequences DCO1_17s04617.t1, DCO1_11s03788.t1, DCO1_2s00954, DCO1_45s07118.t1, DCO1_76s09626.t1 did not have the amino acid region representative of the first and second functional motifs.

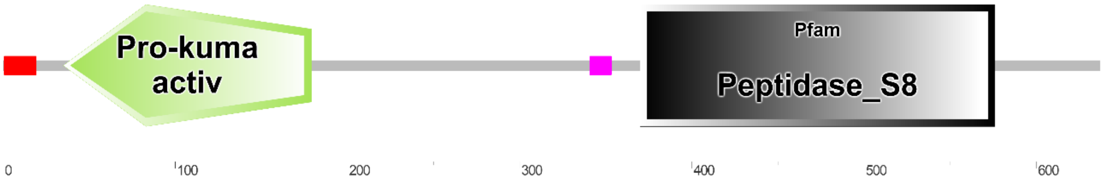


Figure 36. Sequence organisation with domain architecture representation of the sequence DCO1_76s09626.t1

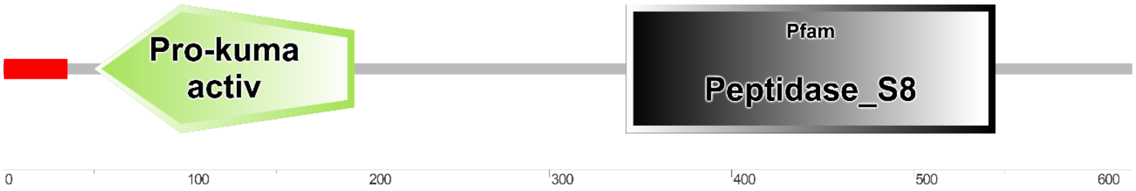


Figure 37. Sequence organisation with domain architecture representation of the sequence DCO1_45s07118.t1

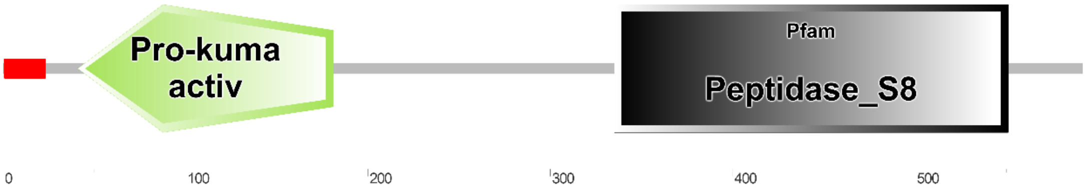


Figure 38. Sequence organisation with domain architecture representation of the sequence DCO1_2s00954

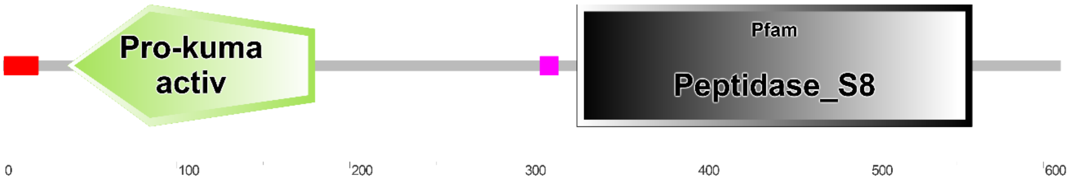


Figure 39. Sequence organisation with domain architecture representation of the sequence DCO1_11s03788.t1

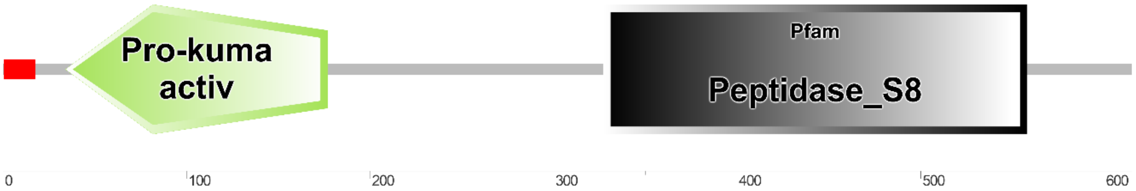


Figure 40. Sequence organisation with domain architecture representation of the sequence DCO1_17s04617.t1

Despite the manual analysis of the MSA that reveals the lack of two active site residues, Pfam representation confirms that both of these this sequences include the peptidase S8 domain. Interestingly, in all the enzymes lacking active sites residues in the first and second functional motifs, a Pro-kumamolisin activation domain was identified. This domain is usually found at the

N-terminus of peptidases belonging to MEROPS peptidase family S53. The predicted cellular role is metabolism and this domain is commonly found in fungi.

The protein sequence DCO1_35s05865.t1 can be grouped as a member of subfamily S8B because the catalytic residues frequently occur in the motifs Asp-Asp-Gly, His-Gly-Thr-Arg and Gly-Thr-Ser-Ala-Ala -Pro. All the others enzymes certainly belong to subfamily S8A.

Family S8 occurs in more than one super kingdom. In fact, according to Pfam database (http://pfam.xfam.org/) this family exists in 11248 species. This peptidase domain is more commonly found in animals with a crucial role in the regulation of plasma cholesterol homeostasis in mouse and human.(177) Fungal kexins (family S8B) have diversified manly from animals, so is expected the amount of kexin genes in that kingdom to be higher. In the human genome, 13 different kexin genes were identified, while when searched for kexin-like sequences in fungi, only identified up to three genes.(178)

Peptidase family S8 contains the serine endopeptidase subtilisin and its homologues. Researchers believe that the subtilase proteases occurred early in ascomycete history with subsequent loss in saprophytic lineages. The fact that several S8 proteases were identified in *Magnaporthe grisea, Metarhizium anisopliae* and *Fusarium graminearium*, confirmed the presence of subtilase genes in these pathogens. The subtilase superfamily is ubiquitous in bacteria and fungi, so they are unlikely to be specifically developed to implement pathogenicity. However, genes directly involved in ecological attributes are hard to identify so the evolutionary process is almost unknown.(179)

The production of a wide range of proteases is key for fungi survival, since it will affected in their ability to gather nutrients from living or dead plant and animal material.(178) However, there remain many gaps in our knowledge concerning the relevance of these molecules for the pathogenicity of fungi towards plants.

In this work is remarkable the families' proteases diversity found in these organisms: *M. phaseolina, N. parvum* and *D. corticola.* The number of proteases in a given family sizes in fungi could correlate with differences in function that are indicative of adaptations to environment and life strategies.(180)

Analysing sequence similarity is key to infer structural or functional characteristics of protein sequences, specially unknown sequences. Since the *D. corticola* genome was only recently sequenced (data not shown), this kind of comparison can be very useful.

The comparison of the complete sequences of these proteases showed that these sequences are very distinct from each other. In order to align properly, the program had to include numerous operations such as inserts, deletions and substitutions. In fact, the only highly conserved regions were the active site residues. Since that is the part of an enzyme that interacts with the substrate during catalysis, it indicates the enzyme function. Even though the active site reveals to be the most conserved region, in some sequences those regions were not found. This inconsistent factor might infer problems in the sequences annotation. Some sequences can be only a protein fragment, which explains the sizes variability. Moreover, in most multicellular eukaryotes, genes are often separated by large intergenic regions, and the genes themselves contain numerous introns, many of them long. The protein length, of the sequences in study, can go from 260 to over 1000 amino acids. The multiple alignment programs are very sensitive to sequence length variations, therefore, that may be a problem.

In *Diplodia corticola* genome annotation project were identified proteins with an unidentified protein domain.(data not shown) These short sequence length proteins suppress plant defence responses, facilitating the colonisation of this pathogenic fungi into and around host cells. However, in several protein sequences of *M. phaseolina, N. parvum* and *D. corticola*, has only recognize a partial functional domain or, in some protein, no domain at all. What can justify such event is if we consider the presence of pseudogenes. A pseudogene is a nucleotide sequence similar to a normal gene but which does not result in a functional product. They are relatives of genes, however, have lost their gene expression in the cell or their ability to code protein. All pseudogenes have distinct mechanisms of origin. Can be caused by retrotransposition of a mature mRNA product or, the most common, gene duplication that is an important process in the evolution of genomes. This event has already been described in fungi species, as *Aspergillus brasiliens,* albeit no conclusive information about the biological impact can be obtain.(181)

3.5. Phylogeny analysis

Phylogenetic analyses of the alignments were performed using MEGA 6.02 (136) for Maximum Likelihood (ML) analysis using WAG + G + I model. The reliability of these topologies was evaluated using bootstrap support with 500 replicates, as described in materials and methods.

It was constructed several phylogenetic trees using Neighbor-joining with different distance matrix models and Maximum Likelihood method. It is not surprising to recovered different trees using different approaches, however some problems arose using Neighbor-joining method which can be seen clearly in the tree in Annexe 5.7. Based on the theoretical knowledge, is was expected this result because ML method is more reliable for reconstructing sequence histories since it uses a more complex evolution mode. While NJ only uses pairwise distance and ignores the detailed alignment. It does not fully utilize the information in a multi-alignment. So, practically speaking, ML is the best option for building the phylogenetic tree.

3.5.1. Protease A1 family

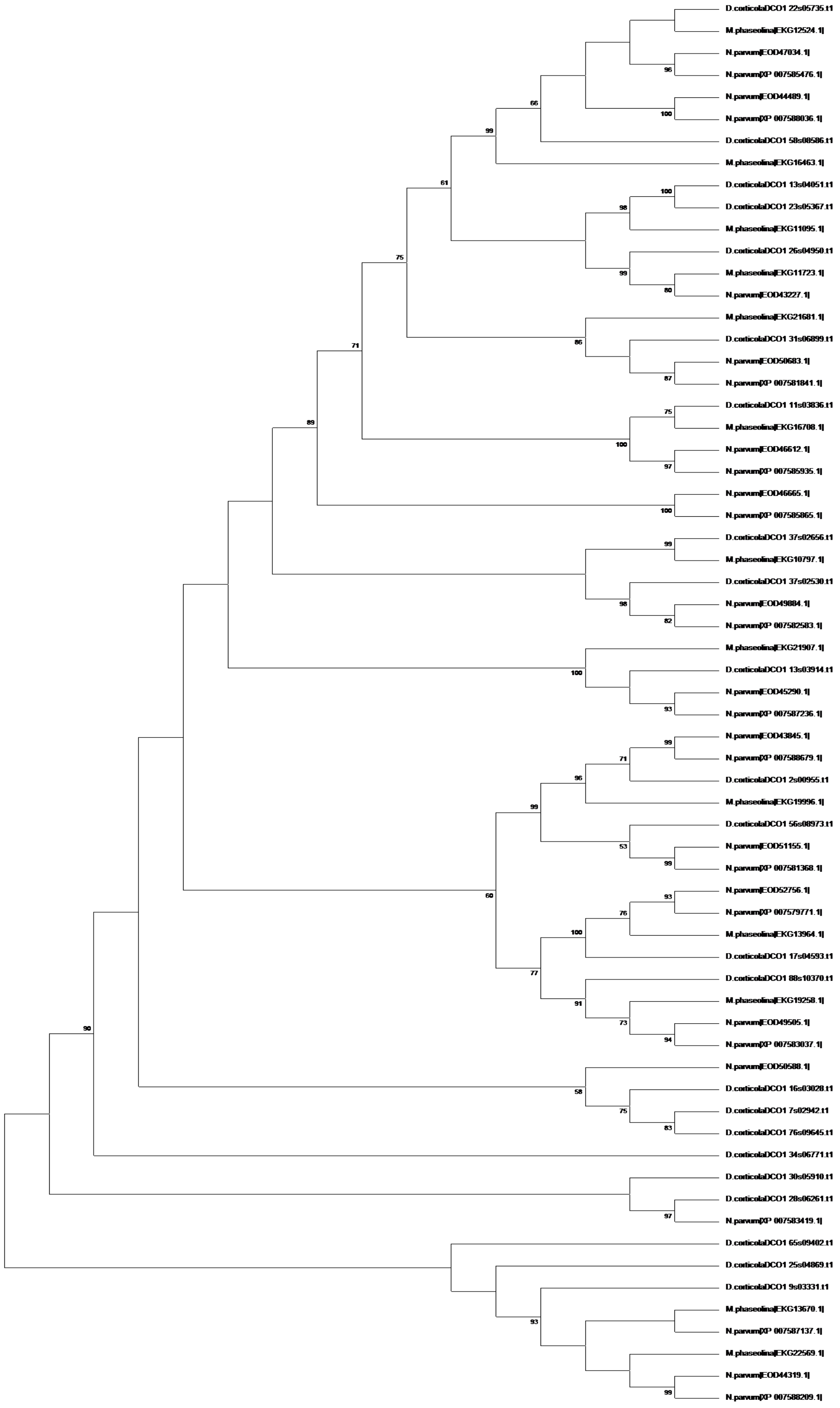


Figure 41. The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan And Goldman model.(140) The tree with the highest log likelihood (-7376,2673) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 3,0375)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 64 amino acid sequences. All positions with less than 95% site coverage were eliminated. Evolutionary analyses were conducted in MEGA6.(136)

The amplification of phylogenetic tree in Figure 41 is stored are Annexe 5.8.

3.5.2. Protease M35 family

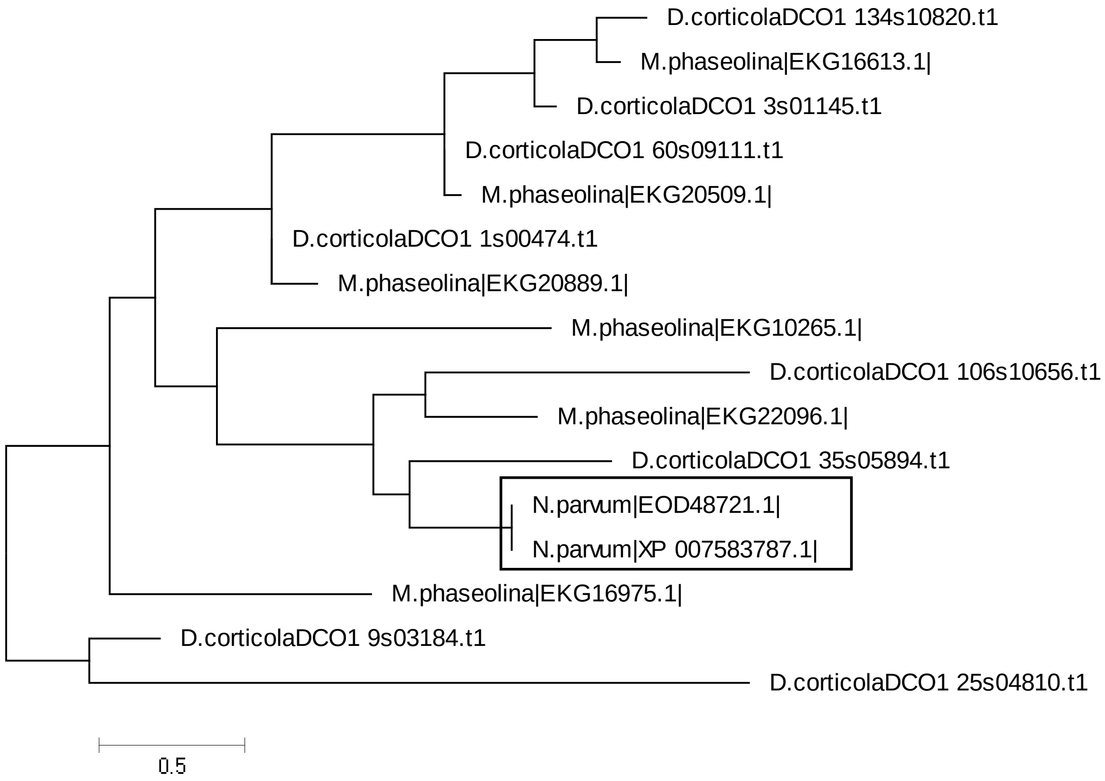

D.corticolaDCO1 134s10820.t1
M.phaseolina|EKG16613.1|
D.corticolaDCO1 3s01145.t1
D.corticolaDCO1 60s09111.t1
M.phaseolina|EKG20509.1|
D.corticolaDCO1 1s00474.t1
M.phaseolina|EKG20889.1|
M.phaseolina|EKG10265.1|
D.corticolaDCO1 106s10656.t1
M.phaseolina|EKG22096.1|
D.corticolaDCO1 35s05894.t1
N.parvum|EOD48721.1|
N.parvum|XP 007583787.1|
M.phaseolina|EKG16975.1|
D.corticolaDCO1 9s03184.t1
D.corticolaDCO1 25s04810.t1
0.5

Figure 42. Molecular Phylogenetic analysis by Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan And Goldman model. The tree with the highest log likelihood (-524.2683) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 2.7866)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 16 amino acid sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 20 positions in the final dataset. Evolutionary analyses were conducted in MEGA6.

3.5.3. Protease S8 family

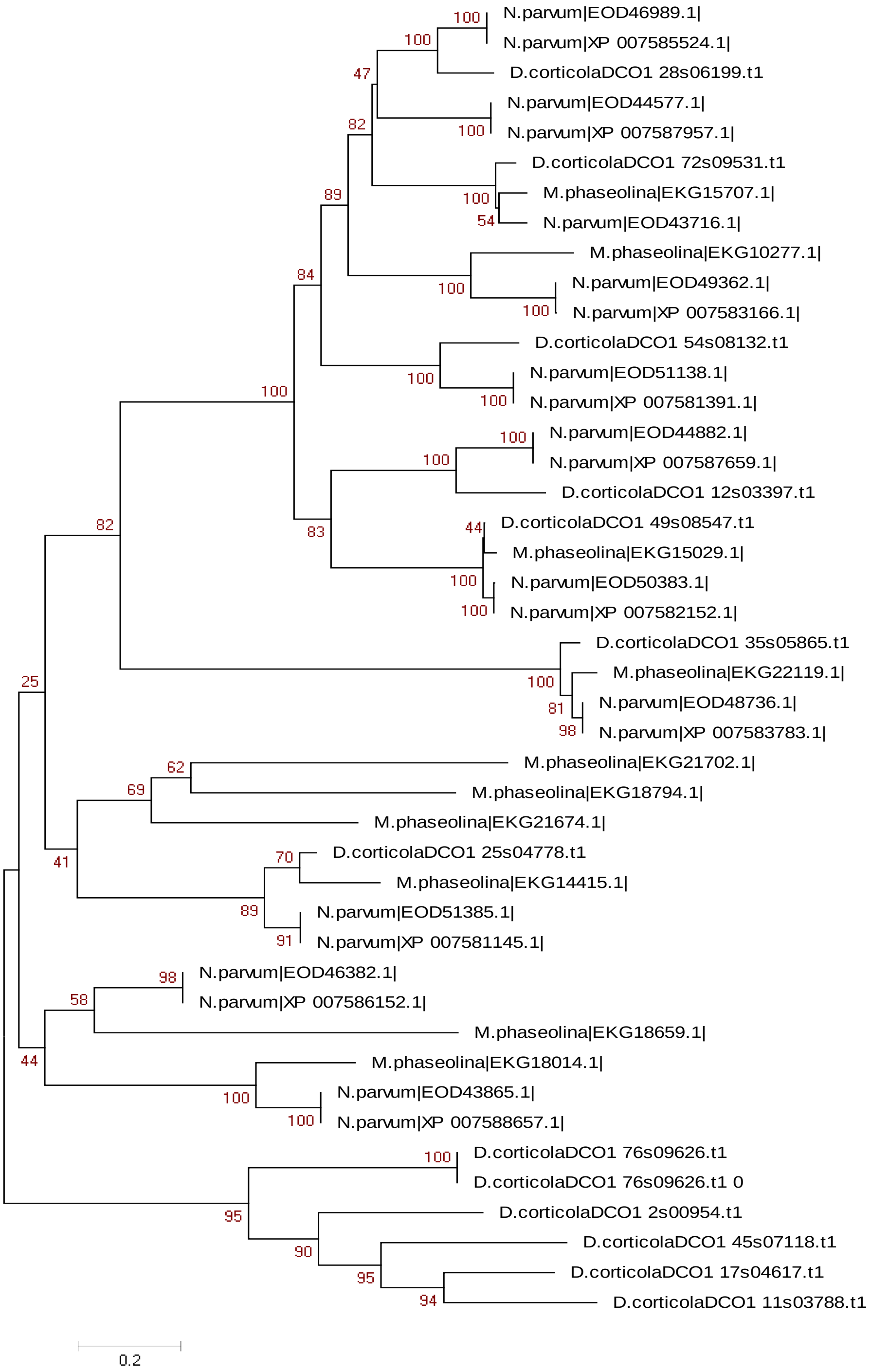


Figure 43. Molecular Phylogenetic analysis by Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan And Goldman model. The tree with the highest log likelihood (-765.2683) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 2.7866)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 44 amino acid sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 632 positions in the final dataset. Evolutionary analyses were conducted in MEGA6.

Only the bootstrap values higher than 50% are visible. Due to the high amount of sequences and additionally the lack of similarity it has a challenge to construct conclusive phylogenetic trees including all the metallic or serine proteases. For that reason, the phylogenetic analysis was based on trees of a single protease family (A1, M35 and S8).

M. phaseolina and *Diplodia corticola* are often associated in the same branch revealing a close evolutionary relationship. Visually analysing the phylogenetic tree, the organism *N. parvum* tends to create individual branches. On an evolutive level, *M. phaseolina* and *D. corticola* are closer to each other than to *N. parvum*. That results is coherent with the multiple sequence alignment where it was possible to verify a greater identity between the protein domains in *N. parvum* proteases. In fact, a study about phylogenetic lineages in the *Botryosphaeriaceae,* suggests that the fungus *N. parvum* has a more distant phylogenetic relationship among the *Botryosphaeriaceae* members*.*

Considering the similar *N. parvum* and *D. corticola* hosts' types, their geographical location or their living conditions, it would be expected that this relationship would extend the evolutionary level. But these work sugest that the evolutionary relationship is not related to the organism's hosts, but rather in the amount and the diversity of the enzymes described in the genome.

4. Conclusion

Analysing the genes that encodes extracellular proteins of these three organisms*,* *Diplodia corticola* genome contains genes that encode a large diversity of glycoside hydrolase families. When compared to *M. phaseolina* and to *N. parvum*, we were able to identify several GH families that are only found in this fungus, suggesting that this could be a strategy to overcome the host plant defence responses. Glycoside hydrolase family 43 that particularly expanded in plant pathogenic species is clearly the most abundant in all the organisms in study. *Neofusicoccum parvum* have just a few genes that encodes enzymes with extracellular xylanolytic and endoglucanolytic activities and a lot with proteolytic activities, indicative of a symbiotic lifestyle.

Furthermore, using bioinformatics' tools we detected some erros in protein annotation. In some cases, the functional domain representative of the family to which the enzyme was described was not present.

The protein sequences of the proteases of *N. parvum* were shown to be very similar which was confirmed by the phylogenetic tree. On an evolutive level, *M. phaseolina* and *D. corticola* are closer to each other than to *N. parvum*. These results reveals that the amount and the diversity of the enzymes described in the genome of these organisms is related on an evolutionary level rather than to types of hosts.

5. Annexes

Erro!

5.1. Sequences not considered

5.1.1. *Macrophomina phaseolina*

Transglutaminas-like proteins; Proteins inhibitors; Intradiol ring-cleavage; Hypothetical proteins; Transcripton factor; Ubiquitin protease associated domain-PA; Retrotransposon; Peptidase A1 EKG17065.1; Peptidase A1 EKG22100.1; Peptidase M20 EKG11719.1; Peptidase EKG18854.1; EKG18756.1; EKG18854.1; EKG20246.1; EKG14974.1; EKG16359.1; EKG15743.1; EKG13531.1; EKG10439.1

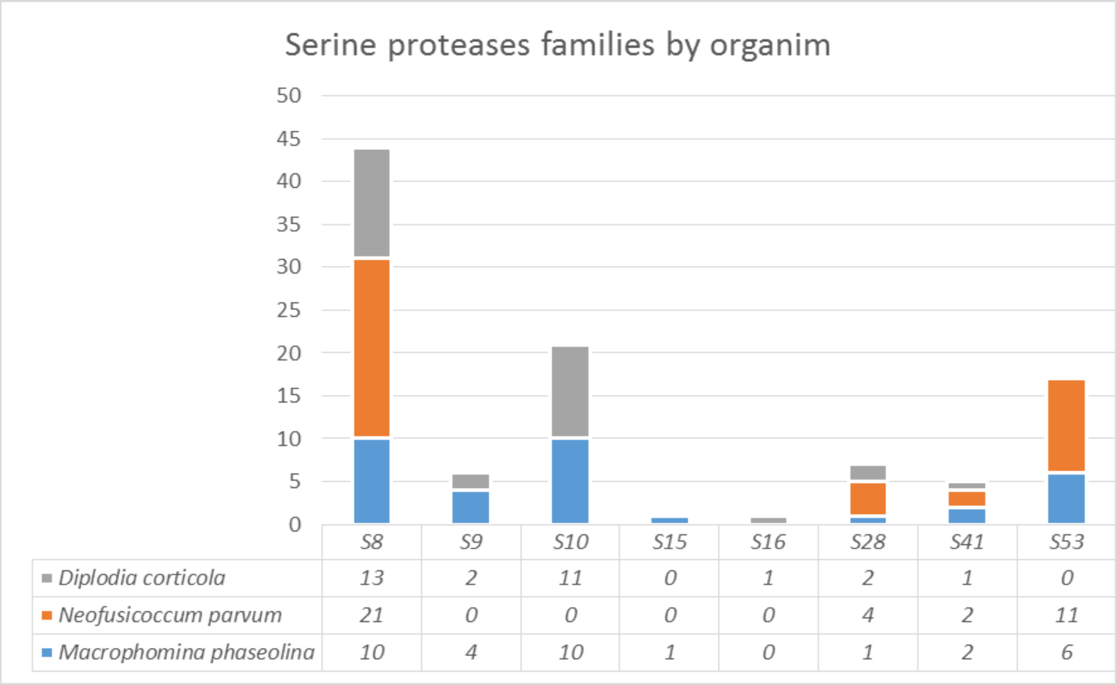
5.1.2. *Neofusicoccum parvum*

Putative integral membrane protease of the rhomboid; Putative family involved in different forms of regulated; Putative intramembrane proteolysis protein; Putative pa and ring finger domain protein; Putative otu domain containing protein GB protein; Putative ubiquitin thioesterase out protein; Putative dj-1 family protein; Putative chaperone protein hsp31 protein; Putative rhomboid family protein; Putative nucleoside diphosphate sugar epimerase protein; Putative math and uch domain protein; Putative hypothetical protein UCRNP2-467; Putative hemolysin-type calcium binding region protein; Putative aminopeptidase y protein; EOD45409.1; EOD49109.1; XP_007580059.1; EOD51531.1; XP_007588044.1; XP_007587466.1; XP_007584266.1; EOD45038.1; EOD48253.1; XP_007589290.1 (identical to EOD43227.1); XP_007581274.1 (identical to EOD51288.1); XP_007588310.1 (identical to EOD44213); XP_007587473.1 (identical to EOD45045.1); XP_007581977.1 (identical to EOD50543.1); XP_007579414.1 (identical to 485929834); XP_007584398.1 (identical to EOD48149.1); XP_007588678.1 (identical to EOD43848.1); XP_007588798.1 (identical to EOD43716.1)

5.1.3. *Diplodia corticola*

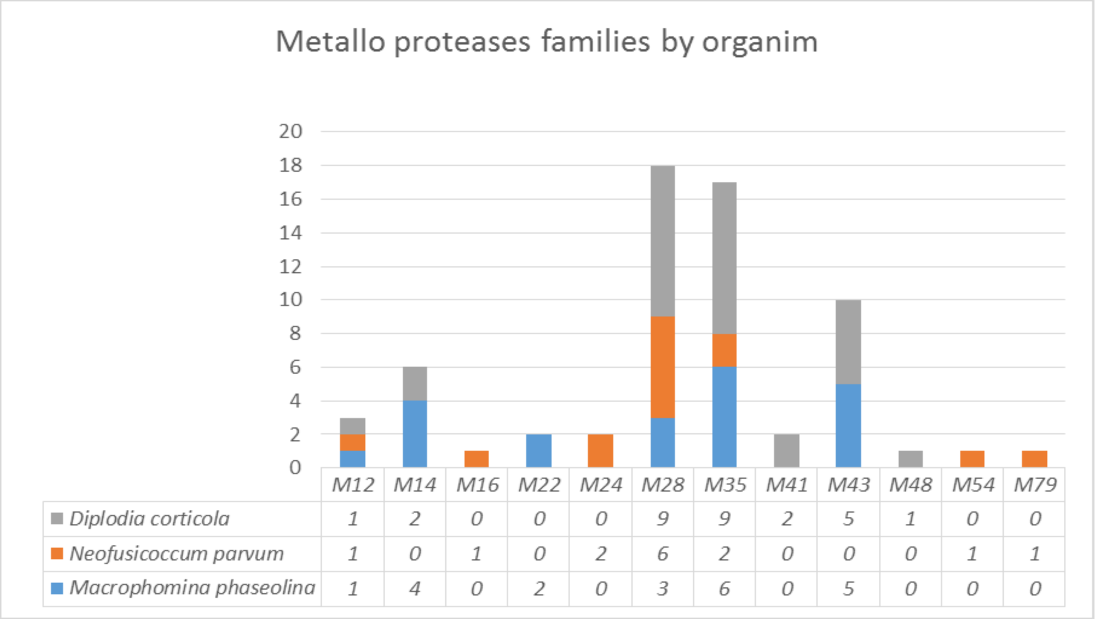
DCO1_38s06614.t1; DCO1_1s00656.t1; DCO1_143s10802.t1; DCO1_25s04759.t1; DCO1_27s06133.t1; DCO1_34s06808.t1; DCO1_38s06577.t1; DCO1_46s07737.t1; DCO1_8s01352.t1; DCO1_27s06131.t1; DCO1_1s00220.t1; DCO1_26s04975.t1; DCO1_26s04890.t1; DCO1_48s08287.t1

5.2. Distribution of proteases families between the organisms



	S8	S9	S10	S15	S16	S28	S41	S53
■ *Diplodia corticola*	13	2	11	0	1	2	1	0
■ *Neofusicoccum parvum*	21	0	0	0	0	4	2	11
■ *Macrophomina phaseolina*	10	4	10	1	0	1	2	6

Figure 44. Distribution of serine proteases families between the *Diplodia corticola, Neofusicoccum parvum* and *Macrophomina phaseolina.*



	M12	M14	M16	M22	M24	M28	M35	M41	M43	M48	M54	M79
■ *Diplodia corticola*	1	2	0	0	0	9	9	2	5	1	0	0
■ *Neofusicoccum parvum*	1	0	1	0	2	6	2	0	0	0	1	1
■ *Macrophomina phaseolina*	1	4	0	2	0	3	6	0	5	0	0	0

Figure 45. Distribution of metallo proteases families between the *Diplodia corticola, Neofusicoccum parvum* and *Macrophomina phaseolina.*

5.3. Protein Sequences of *Macrophomina Phaseolina* Enzymes

5.3.1. Aspartic proteases

EKG22569.1; EKG21907.1; EKG21681.1; EKG19996.1; EKG19258.1; EKG18903.1; EKG17972.1; EKG16708.1; EKG16463.1; EKG13964.1; EKG13670.1; EKG12524.1; EKG12388.1; EKG11723.1; EKG11095.1; EKG10797.1

5.3.2. Metallo proteases

EKG22096.1; EKG20889.1; EKG20509.1; EKG19704.1; EKG19652.1; EKG18003.1; EKG17296.1; EKG16975.1; EKG16613.1; EKG16156.1; EKG14371.1; EKG13918.1; EKG13569.1; EKG13395.1; EKG11890.1; EKG11369.1; EKG11140.1; EKG10934.1; EKG10265.1; EKG10139.1; EKG09239.1

5.3.3. Serine proteases

EKG22119.1; EKG21702.1; EKG21674.1; EKG19995.1; EKG19690.1; EKG18794.1; EKG18659.1; EKG18014.1; EKG16906.1; EKG16459.1; EKG16302.1; EKG16239.1; EKG16094.1; EKG16054.1; EKG15902.1; EKG15707.1; EKG15685.1; EKG15518.1; EKG15029.1; EKG14843.1; EKG14415.1; EKG14081.1; EKG13692.1; EKG13511.1; EKG13315.1; EKG13057.1; EKG12885.1; EKG12582.1; EKG11840.1; EKG11180.1; EKG10314.1; EKG10313.1; EKG10277.1; EKG10175.1

5.3.4. Cysteine proteases

EKG15700.1

5.3.5. Glutamic proteases

EKG19611.1; EKG15813.1

5.4. Protein sequences of *Neofusicoccum parvum* enzymes

5.4.1. Aspartic proteases

EOD52756.1; EOD51155.1; EOD50683.1; EOD49884.1; EOD49505.1; EOD47034.1; EOD46665.1; EOD46612.1; EOD45290.1; EOD44489.1; EOD44319.1; EOD43845.1; EOD43227.1; XP_007588679.1; XP_007588209.1; XP_007588036.1; XP_007587236.1; XP_007587137.1; XP_007585935.1; XP_007585865.1; XP_007585476.1; XP_007583419.1; XP_007583037.1; XP_007582583.1; XP_007581841.1; XP_007581368.1; XP_007579771.1

5.4.2. Metallo proteases

EOD53113.1; EOD52468.1; EOD52192.1; EOD50543.1; EOD49625.1; EOD48721.1; EOD45045.1; EOD44692.1; EOD44213.1; EOD43785.1; XP_007587835.1; XP_007583787.1; XP_007582881.1; XP_007580302.1

5.4.3. Serine proteases

EOD53025.1; EOD52773.1; EOD51385.1; EOD51138.1; EOD50464.1; EOD50383.1; EOD49362.1; EOD49357.1; EOD49349.1; EOD48736.1; EOD48584.1; EOD47983.1; EOD46989.1; EOD46382.1; EOD45322.1; EOD44935.1; EOD44892.1; EOD44882.1; EOD44577.1; EOD43873.1; EOD43865.1; EOD43848.1; EOD43716.1; XP_007588665.1; XP_007588657.1; XP_007587957.1; XP_007587659.1; XP_007587598.1; XP_007587586.1; XP_007587204.1; XP_007586152.1; XP_007585524.1; XP_007584545.1; XP_007583944.1; XP_007583783.1; XP_007583181.1; XP_007583166.1; XP_007583160.1; XP_007582152.1; XP_007582064.1; XP_007581391.1; XP_007581145.1; XP_007579756.1; XP_007579485.1

5.4.4. Cysteine proteases

EOD48149.1

5.4.5. Glutamic proteases

EOD51288.1; EOD50588.1; XP_007581929.1

5.5. Protein sequences of *Diplodia corticola* enzymes

5.5.1. Aspartic proteases

DCO1_31s06899.t1; DCO1_11s03836.t1; DCO1_13s04051.t1; DCO1_23s05367.t1; DCO1_25s04869.t1; DCO1_26s04950.t1; DCO1_28s06261.t1; DCO1_30s05910.t1; DCO1_37s02530.t1; DCO1_37s02656.t1; DCO1_58s08586.t1; DCO1_65s09402.t1; DCO1_9s03331.t1; DCO1_13s03914.t1; DCO1_16s03028.t1; DCO1_17s04593.t1; DCO1_2s00955.t1; DCO1_22s05735.t1; DCO1_34s06771.t1; DCO1_56s08973.t1; DCO1_7s02942.t1; DCO1_76s09645.t1; DCO1_88s10370.t1

5.5.2. Metallo proteases

DCO1_97s10432.t1; DCO1_1s00050.t1; DCO1_1s00474.t1; DCO1_106s10656.t1; DCO1_134s10820.t1; DCO1_44s07437.t1; DCO1_48s08308.t1; DCO1_6s02104.t1; DCO1_78s09893.t1; DCO1_9s03184.t1; DCO1_9s03328.t1; DCO1_10s03645.t1; DCO1_106s10660.t1; DCO1_119s10787.t1; DCO1_12s03468.t1; DCO1_13s03960.t1; DCO1_23s05448.t1; DCO1_25s04810.t1; DCO1_26s04960.t1; DCO1_27s06063.t1;

DCO1_3s01145.t1; DCO1_32s07045.t1; DCO1_35s05894.t1; DCO1_47s08334.t1; DCO1_47s08359.t1; DCO1_60s09111.t1; DCO1_7s02929.t1; DCO1_96s10304.t1

5.5.3. Serine proteases

DCO1_12s03397.t1; DCO1_16s03131.t1; DCO1_35s05865.t1; DCO1_49s08547.t1; DCO1_76s09626.t1; DCO1_1s00288.t1; DCO1_13s03906.t1; DCO1_17s04617.t1; DCO1_19s02442.t1; DCO1_2s00954.t1; DCO1_28s06199.t1; DCO1_28s06180.t1; DCO1_48s08289.t1; DCO1_5s01554.t1; DCO1_54s08132.t1; DCO1_56s08938.t1; DCO1_57s08427.t1; DCO1_58s08590.t1; DCO1_61s09057.t1; DCO1_62s08858.t1; DCO1_45s07117.t1; DCO1_45s07118.t1; DCO1_45s07097.t1; DCO1_71s09421.t1; DCO1_72s09531.t1; DCO1_76s09626.t1; DCO1_78s09882.t1; DCO1_87s10156.t1; DCO1_93s10277.t1; DCO1_11s03788.t1; DCO1_25s04778.t1

5.6. Protein sequence information

5.6.1. *Macrophomina phaseolina*

5.6.1.1. Aspartic proteases

All aspartic proteases belong to MEROPS A1 family

Accession Number	**Sequence Length (Amino Acid)**	**Functional Domain**		
		Start	End	Length
EKG22569.1	538	31	387	356
EKG21907.1	509	88	400	312
EKG21681.1	254	97	252	155
EKG19996.1	407	63	381	318
EKG19258.1	477	71	402	331
EKG18903.1	1024	141	382	241
EKG17972.1	521	179	411	232
EKG16708.1	403	98	397	299
EKG16463.1	399	88	398	310
EKG13964.1	512	65	403	338

EKG13670.1	551	31	387	356
EKG12524.1	380	90	380	290
EKG12388.1	551	1	297	296
EKG11723.1	475	161	469	308
EKG11095.1	290	2	284	282
EKG10797.1	379	44	377	333

5.6.1.2. Metallo proteases

Accession Number	Protease Family	Sequence Length (Amino Acid)	Functional Domain		
			Start	End	Length
EKG11140.1	M12		284	496	
EKG18003.1	M14		134	422	
EKG11369.1			1	118	
EKG10934.1			196	508	
EKG09239.1			128	413	
EKG16156.1	M22;		28	318	
EKG13918.1			185	413	
EKG17296.1	M28; ;		149	333	
EKG11890.1			160	370	
EKG10139.1			174	384	
EKG22096.1	M35		3	373	
EKG20889.1			1	360	
EKG20509.1			1	352	
EKG16975.1			17	203	
EKG16613.1			1	350	
EKG10265.1			146	353	
EKG19704.1	M43		131	277	
EKG19652.1			136	288	
EKG14371.1			177	297	
EKG13569.1			118	223	
EKG13395.1			125	273	

5.6.1.3. Serine proteases

Accession	Protease Family	Sequence Length (Amino Acid)	Functional Domain		
Number			Start	End	Length
EKG22119.1	S8	845	190	472	282
EKG21702.1		389	18	381	363
EKG21674.1		773	18	464	446
EKG18014.1		659	1	328	327
EKG15707.1		426	141	399	258
EKG15029.1		533	180	459	279
EKG14415.1		238	1	77	76
EKG10277.1		401	145	379	234
EKG19690.1	S10	350	6	273	267
EKG15902.1		413	1	372	371
EKG15518.1		475	57	469	412
EKG14843.1		571	56	513	457
EKG13692.1		337	77	273	196
EKG13315.1		577	1	421	420
EKG12885.1		420	1	382	381
EKG12582.1		544	131	539	408
EKG11840.1		554	57	500	43
EKG10314.1		557	75	512	437
EKG16239.1	S15	423	68	257	189
EKG13057.1	S28	563	67	512	445
EKG16459.1	S41	774	374	603	229
EKG16094.1		851	361	585	224
EKG15685.1		281	135	249	114
EKG11180.1		800	411	671	260
EKG19995.1	S53	494	344	479	135
EKG16054.1		433	43	374	331
EKG14081.1		617	268	563	295
EKG13511.1		656	380	605	225
EKG10313.1		606	334	539	205
EKG10175.1		641	373	583	210
EKG16906.1	GlpG	267	108	266	158
EKG18794.1	S8_S53	844	105	540	435
EKG18659.1		600	1	300	299

Erro!

5.6.1.4. Cysteine and glutamic proteases

Accession number	Protease Family	Sequence Length (Amino Acid)	Functional Domain Range		
			Start	End	Length
EKG15700.1	C13	1136	41	306	265
EKG19611.1	G1	934	67	269	202
EKG15813.1		136	2	135	133

5.6.2. *Neofusicoccum parvum*

5.6.2.1. Aspartic proteases

All aspartic proteases belong to MEROPS A1 family.

Accession Number	Sequence Length (Amino Acid)	Functional Domain		
		Start	End	Length
EOD52756.1	512	65	403	338
EOD51155.1	391	12	365	353
EOD50683.1	409	96	407	311
EOD49884.1	400	86	399	313
EOD49505.1	476	70	401	331
EOD47034.1	401	89	398	309
EOD46665.1	441	93	427	334
EOD46612.1	411	97	410	313
EOD45290.1	736	89	401	312
EOD44489.1	388	84	387	303
EOD44319.1	553	28	197	169
EOD43845.1	417	38	391	353
EOD43227.1	462	147	455	308
XP_007588679.1	417	77	430	353
XP_007588209.1	553	28	197	169
XP_007588036.1	388	84	387	303
XP_007587236.1	736	160	472	312
XP_007587137.1	549	34	391	357
XP_007585935.1	411	97	410	313
XP_007585865.1	441	93	427	334
XP_007585476.1	401	89	398	309
XP_007583419.1	597	50	397	347
XP_007583037.1	476	70	401	331
XP_007582583.1	400	86	399	313

XP_007581841.1	409	96	407	311
XP_007581368.1	391	12	365	353
XP_007579771.1	512	65	403	338

5.6.2.2. Metallo proteases

Accession Number	**Protease Family**	**Sequence Length (Amino Acid)**	**Functional Domain**		
			Start	End	Length
EOD52468.1	M12	764	284	477	193
EOD50543.1	M16	855	261	448	187
EOD44692.1	M24	785	504	763	259
XP_007587835.1		785	504	763	259
EOD52192.1	M28	521	252	467	215
EOD49625.1		843	149	333	184
EOD44213.1		1086	326	526	200
EOD43785.1		810	420	613	193
XP_007582881.1		834	149	333	184
XP_007580302.1		521	252	467	215
EOD48721.1	M35	384	11	373	362
XP_007583787.1		384	2	363	361
EOD45045.1	M54	521	1	356	355
EOD53113.1	M79	171	1	87	86

5.6.2.3. Serine proteases

Accession Number	Protease Family	Sequence Length (Amino Acid)	Functional Domain		
			Start	End	Length
EOD53025.1	S8	569	301	520	219
EOD52773.1		619	340	562	222
EOD51385.1		874	115	521	406
EOD51138.1		520	192	458	266
EOD50383.1		533	176	450	274
EOD49362.1		390	140	383	243
EOD49349.1		601	323	535	212
EOD48736.1		844	190	467	277
EOD47983.1		619	274	559	285
EOD46989.1		385	135	377	242
EOD46382.1		406	1	102	101
EOD44892.1		610	327	558	231
EOD44882.1		304	48	287	239
EOD44577.1		402	138	391	253
EOD43865.1		1054	167	612	445
EOD43848.1		584	336	514	178
EOD43716.1		425	137	385	248
XP_007588657.1		1054	167	612	445
XP_007587957.1		402	138	391	253
XP_007587659.1		304	48	287	239
XP_007587598.1		610	327	558	231
XP_007586152.1		406	1	102	101
XP_007585524.1		385	135	377	242
XP_007584545.1		619	274	559	285
XP_007583783.1		844	190	467	277
XP_007583181.1		601	323	535	212
XP_007583166.1		390	140	383	243
XP_007582152.1		553	176	450	274
XP_007581391.1		520	192	458	266
XP_007581145.1		874	115	521	406
XP_007579756.1		619	340	562	222
XP_007579485.1		569	362	580	218
EOD50464.1		737	498	708	210

5.6.2.4. Cysteine and glutamic proteases

Accession number	Protease Family	Sequence Length (Amino Acid)	Functional Domain		
			Start	End	Length
EOD48149.1	C54	239	102	167	65
EOD51288.1*	G1	850	84	296	212
			342	554	212
			580	796	216
EOD50588.1		273	66	273	207
XP_007581929.1		273	67	272	205

*SMART tool predicted three peptidase G1 domains.

5.6.3. *Diplodia corticola*

5.6.3.1. Aspartic proteases

All aspartic proteases belong to MEROPS A1 family

Sequence Name	Sequence Length (Amino Acid)	Functional Domain		
		Start	End	Length
DCO1_31S06899.T1	412	89	409	320
DCO1_11S03836.T1	415	97	413	316
DCO1_13S04051.T1	414	88	407	319
DCO1_23S05367.T1	556	32	387	355
DCO1_25S04869.T1	379	39	349	310
DCO1_26S04950.T1	494	169	486	317
DCO1_28S06261.T1	603	45	398	353
DCO1_30S05910.T1	821	57	410	353
DCO1_37S02530.T1	400	71	399	328
DCO1_37S02656.T1	420	41	386	345
DCO1_58S08586.T1	394	78	392	314
DCO1_65S09402.T1	660	61	406	345
DCO1_9S03331.T1	563	28	386	358
DCO1_13S03914.T1	496	50	399	349
DCO1_16S03028.T1	620	110	415	305
DCO1_17S04593.T1	412	35	414	379
DCO1_2S00955.T1	421	43	395	352
DCO1_22S05735.T1	420	84	398	314
DCO1_34S06771.T1	485	40	383	343
DCO1_56S08973.T1	423	41	398	357

DCO1_7S02942.T1	677	124	410	286
DCO1_76S09645.T1	603	124	425	301
DCO1_88S10370.T1	475	54	423	369

5.6.3.2. Metallo proteases

Accession Number	Protease Family	Sequence Length (Amino Acid)	Functional Domain		
			Start	End	Length
DCO1_44S07437.T1	M12	820	281	493	212
DCO1_106S10660.T1	M14	433	132	420	288
DCO1_13S03960.T1		591	201	513	312
DCO1_10S03645.T1	M28	805	430	610	180
DCO1_12S03468.T1		1036	156	323	167
DCO1_23S05448.T1		480	150	358	208
DCO1_26S04960.T1		499	259	436	177
DCO1_27S06063.T1		377	170	358	188
DCO1_32S07045.T1		721	360	546	186
DCO1_47S08334.T1		522	265	458	193
DCO1_7S02929.T1		391	183	372	189
DCO1_96S10304.T1		259	14	207	193
DCO1_1S00474.T1	M35	541	16	417	401
DCO1_106S10656.T1		531	183	391	208
DCO1_134S10820.T1		348	1	348	347
DCO1_9S03184.T1		393	30	391	361
DCO1_25S04810.T1		279	76	258	182
DCO1_3S01145.T1		352	1	350	349
DCO1_35S05894.T1		383	2	361	359
DCO1_60S09111.T1		350	1	349	348
DCO1_47S08359.T1	M41	804	452	709	257
DCO1_1S00050.T1		820	566	765	199
DCO1_48S08308.T1	M43	292	139	285	146
DCO1_6S02104.T1		276	120	272	152
DCO1_78S09893.T1		281	133	277	144
DCO1_9S03328.T1		284	134	273	139
DCO1_119S10787.T1		288	148	281	133
DCO1_97S10432.T1	M48	364	146	259	113

5.6.3.3. Serine proteases

Accession Number	Protease Family	Sequence Length (Amino Acid)	Functional Domain		
			Start	End	Length
DCO1_12S03397.T1	S8	385	135	366	231
DCO1_35S05865.T1		850	194	475	281
DCO1_49S08547.T1		535	39	432	393
DCO1_76S09626.T1		637	401	569	168
DCO1_17S04617.T1		615	358	549	191
DCO1_2S00954.T1		592	349	520	171
DCO1_28S06199.T1		402	143	376	233
DCO1_54S08132.T1		515	204	451	247
DCO1_45S07118.T1		605	346	517	171
DCO1_72S09531.T1		428	142	373	231
DCO1_76S09626.T1		637	401	569	168
DCO1_11S03788.T1		610	217	610	393
DCO1_25S04778.T1		906	152	594	442
DCO1_45S07097.T1	S9	776	549	742	193
DCO1_71S09421.T1		735	496	703	207
DCO1_13S03906.T1	S10	1178	58	473	415
DCO1_19S02442.T1		573	63	511	448
DCO1_28S06180.T1		556	67	516	449
DCO1_5S01554.T1		517	52	506	454
DCO1_56S08938.T1		544	138	538	400
DCO1_57S08427.T1		576	64	469	405
DCO1_61S09057.T1		553	63	516	453
DCO1_62S08858.T1		572	83	512	429
DCO1_45S07117.T1		557	77	506	429
DCO1_78S09882.T1		580	52	493	441
DCO1_93S10277.T1		568	89	521	432
DCO1_16S03131.T1	S16	942	8	255	247
DCO1_1S00288.T1	S28	551	68	505	437
DCO1_87S10156.T1		545	59	524	465
DCO1_48S08289.T1	S41	712	347	417	70
DCO1_58S08590.T1		790	372	586	214

Erro!

5.7. Phylogenetic tree constructed using Neighbor-Joining method

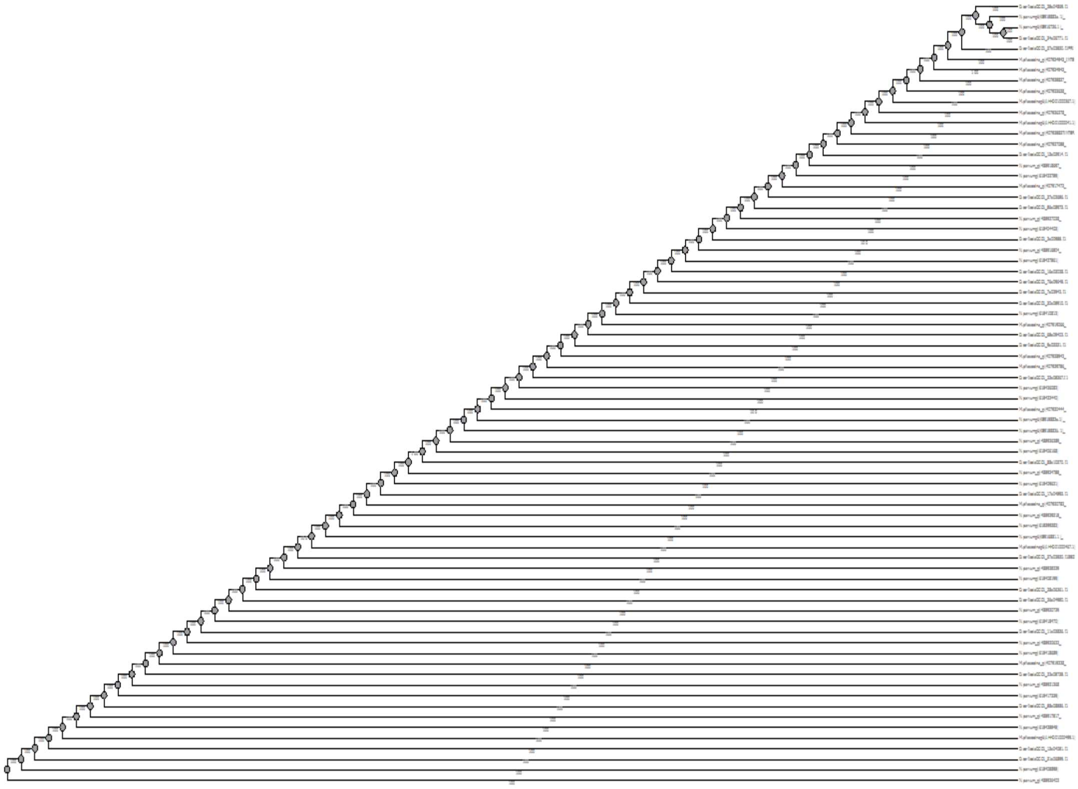
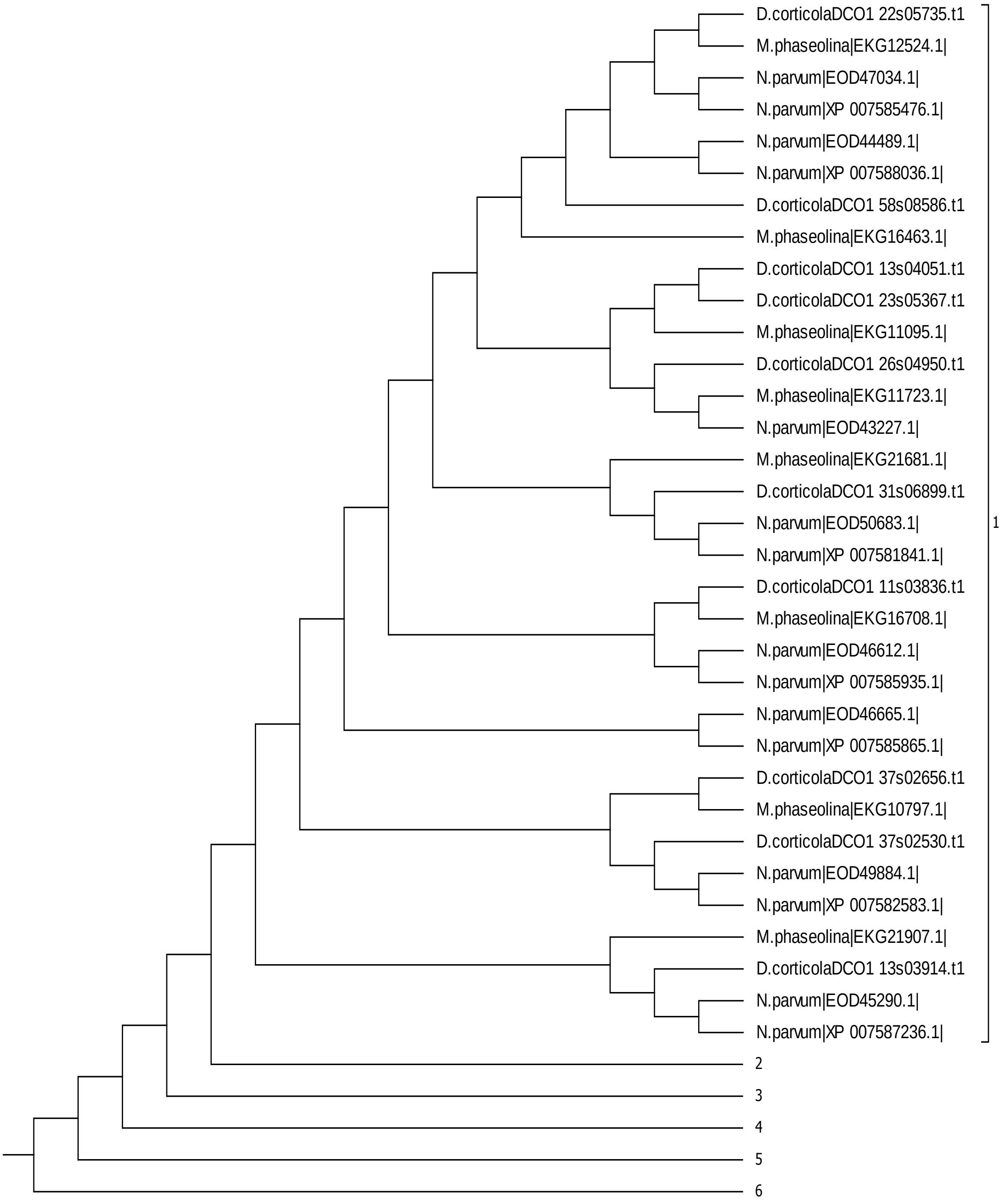


Figure 46. Phylogenetic tree of aspartic proteases of *M. phaseolina, N. parvum* and *D. corticola* using Neighbor-joining method.

5.8. Phylogenetic tree of A1 family



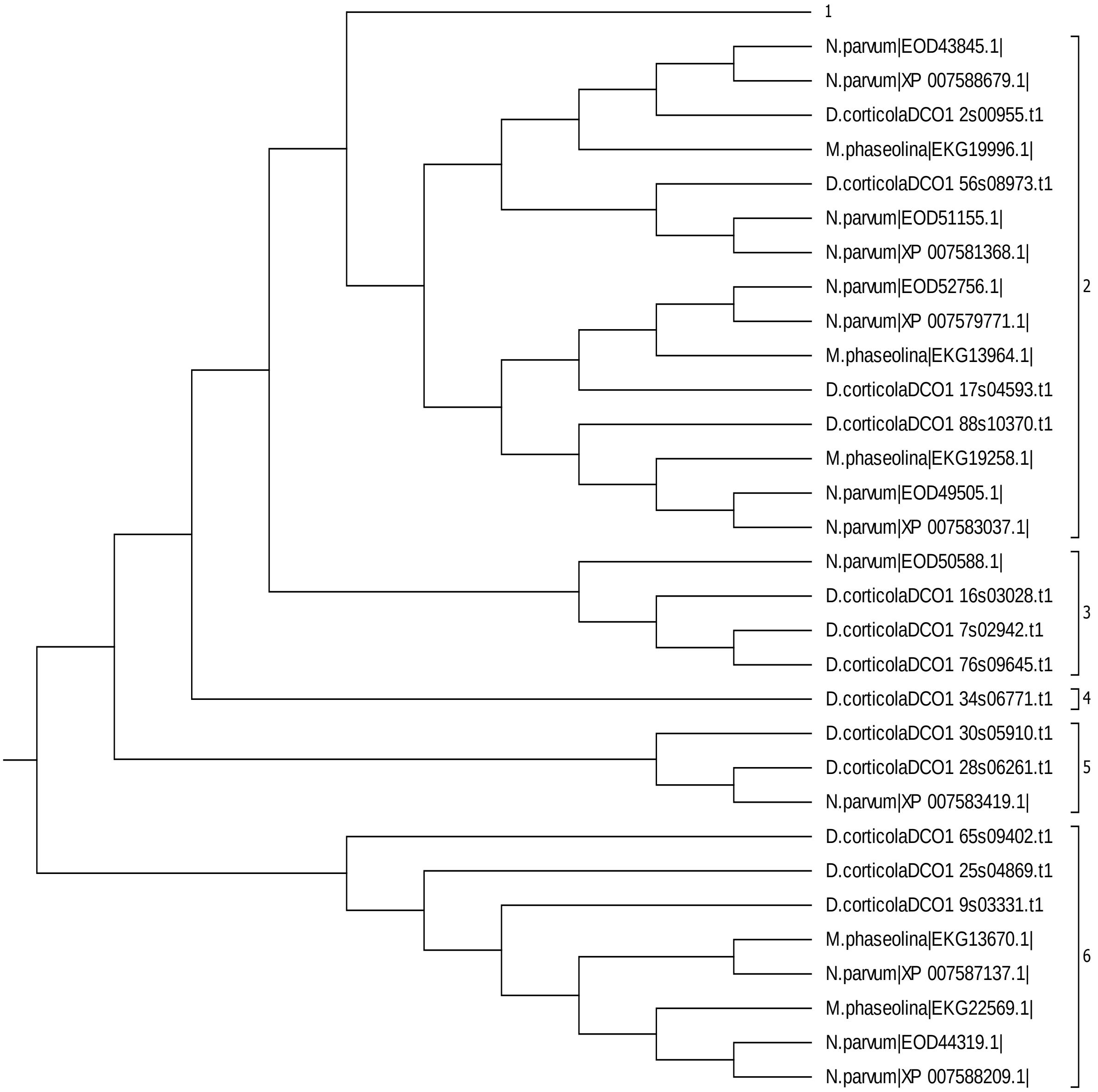


Figure 47. The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan And Goldman model.(140) The tree with the highest log likelihood (-7376,2673) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 3,0375)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 64 amino acid sequences. All positions with less than 95% site coverage were eliminated. Evolutionary analyses were conducted in MEGA6.(136)

6. References

1. Deacon JW. Modern Mycology. 3rd ed. Oxford: Blackwell Science; 1998.

2. Carlile M, Watkinson S, Gooday G. The fungi. Socgenmicrobiol.Org.Uk. 1994.

3. Buckley M. The fungal kingdom diverse and essential roles in earth's ecosystem. American Acedemy of Microbiology. Washington; 2008.

4. Esser K, Lemke PA. The Mycota : a comprehensive treatise on fungi as experimental systems for basic and applied research. Berlin: Springer; 1995.

5. Blackwell M. The fungi: 1, 2, 3 ... 5.1 million species? Am J Bot. 2011;98(3):426–38.

6. Tavares Sã-, Ramos AP, Pires AS, Azinheira HG, Caldeirinha P, Link T, et al. Genome size analyses of Pucciniales reveal the largest fungal genomes. Front Plant Sci. 2014;5(422):11.

7. Gregory TR, Nicol J a., Tamm H, Kullman B, Kullman K, Leitch IJ, et al. Eukaryotic genome size databases. Nucleic Acids Res. 2007;35(SUPPL. 1):332–8.

8. Hijri M, Sanders IR. Low gene copy number shows that arbuscular mycorrhizal fungi inherit genetically different nuclei. Nature. 2005 Jan 13;433(7022):160–3.

9. Albertin W, Marullo P. Polyploidy in fungi: evolution after whole-genome duplication. Proc R Soc B Biol Sci. 2012;279(1738):2497–509.

10. Clegg CJ, MacKean DG. Advanced biology: Principles and applications. 2nd ed. International Student Edition; 2000.

11. Gonzalez-Fernandez R, Jorrin-Novo J V. Contribution of proteomics to the study of plant pathogenic fungi. J Proteome Res. 2012;11(1):3–16.

12. Agrios GN. Plant pathology. 4th ed. San Diego: Elsevier Academic Press; 1997.

13. Harris RH, Clark JR, Matheny NP. Arboriculture: Integrated Management of Landscape Trees. 4th ed. New Jersey: Prentice Hall; 2004.

14. Michereff SJ, Barros R. Proteção de plantas na agricultura sustentável. 1st ed. Recife: University Publisher; 2001.

15. Gibson DM, King BC, Hayes ML, Bergstrom GC. Plant pathogens as a source of diverse enzymes for lignocellulose digestion. Curr Opin Microbiol. 2011;14(3):264–70.

16. Anderson PK, Cunningham A a., Patel NG, Morales FJ, Epstein PR, Daszak P. Emerging infectious diseases of plants: Pathogen pollution, climate change and agrotechnology drivers. Trends Ecol Evol. 2004;19(10):535–44.

17. Olsen L, Choffnes ER, Relman D a, Pray L. Fungal diseases: An emerging threat to human, animal and plant health. New York. 2011. 1-99 p.

18. Hawksworth DL, Kirk PM, Sutton BC, Pegler DN. Ainsworth and Bisby's Dictionary of the Fungi. 9th ed. Trowbridge, UK: International Mycological Institute; 2001.

19. Tello M-L, Tomalak M, Siwecki R, Gáper J, Motta E, Mateo-Sagasta E. Biotic Urban Growing Conditions — Threats, Pests and Diseases. In: Konijnendijk C, Nilsson K, Randrup T, Schipperijn J, editors. Urban Forests and Trees SE - 13. Springer Berlin Heidelberg; 2005. p. 325–65.

20. Gonz R, Prats E. Proteomics of Plant Pathogenic Fungi. 2010;2010.

21. Hawksworth DL, Kirk PM, Sutton BC, Pegler DN. Ainsworth and Bisby's Dictionary of the fungi. 8th ed. Oxon: CAB International; 1995.

22. Úrbez-Torres JR. The status of *Botryosphaeriaceae* species infecting grapevines. Phytopathol Mediterr. 2011;50(SUPPL.):5–45.

23. Slippers B, Wingfield MJ. *Botryosphaeriaceae* as endophytes and latent pathogens of woody plants: diversity, ecology and impact. Fungal Biol Rev. 2007;21(2-3):90–106.

24. Van Niekerk JM, Crous PW, Groenewald JZE, Fourie PH, Halleen F. DNA phylogeny, morphology and pathogenicity of Botryosphaeria species on grapevines. Mycologia. 2004;96(4):781–98.

25. Phillips AJL. *Botryosphaeria* species associated with diseases of grapevines in Portugal. Phytopathol Mediterr. 2002;41(1):3–18.

26. Petrini O, Fisher PJ. A comparative study of fungal endophytes in xylem and whole stem of Pinus sylvestris and Fagus sylvatica. Trans Br Mycol Soc. 1988;91(2):233–8.

27. Podolich O, Ardanov P, Zaets I, Pirttilä AM, Kozyrovska N. Reviving of the endophytic bacterial community as a putative mechanism of plant resistance. Plant Soil. 2014;

28. G. R. *Lasiodiplodia theobromae* as a cause of keratomycoses. Oxford J. 1976;14(2):155–70.

29. Coakley SM, Scherm H, Chakraborty S. Climate Change and Plant Disease Management. Annu Rev Phytopathol. 1999;37:399–426.

30. Vaz AT de A. Doenças causadas por fungos *Botryosphaeriaceae* em videira: Caracterização fenotípica e molecular de isolados e sensibilidade a fungicidas. Universidade Técnica de Lisboa; 2008.

31. Pavlic D, Slippers B, Coutinho T a., Wingfield MJ. Multiple gene genealogies and phenotypic data reveal cryptic species of the Botryosphaeriaceae: A case study on the *Neofusicoccum parvum/N. ribis* complex. Mol Phylogenet Evol. Elsevier Inc.; 2009;51(2):259–68.

32. Gao Q, Jin K, Ying SH, Zhang Y, Xiao G, Shang Y, et al. Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi *Metarhizium anisopliae* and *M. acridum.* PLoS Genet. 2011;7(1).

33. Blanco-Ulate B, Rolshausen P, Cantu D. Draft Genome Sequence of *Neofusicoccum parvum* Isolate UCR-NP2, a Fungal Vascular Pathogen Associated with Grapevine Cankers. Genome Announc. 2013;1(3):5–6.

34. Rodas C a., Slippers B, Gryzenhout M, Wingfield MJ. *Botryosphaeriaceae* associated with Eucalyptus canker diseases in Colombia. For Pathol. 2009;39(2):110–23.

35. Iturritxa E, Slippers B, Mesanza N, Wingfield MJ. First report of *Neofusicoccum parvum* causing canker and die-back of Eucalyptus in Spain. Australas Plant Dis Notes. 2011;6(1):57–9.

36. Wunderlich N, Ash G, Steel C, Raman H, Savocchia S. Trunk disease pathogens within the *Botryosphaeriaceae* are associated with bunch rot disease in the Hunter Valley. Aust New Zeal Grapegrow Winemak. 2009;(548):35–8.

37. Eskalen a, Feliciano a J, Gubler W a. Susceptibility of grapevine pruning wounds and symptom development in response to infection by Phaeoacremonium aleophilum and Phaeomoniella chlamydospora. Plant Dis. 2007;91(9):1100–4.

38. Phillips a JL, Alves a, Abdollahzadeh J, Slippers B, Wingfield MJ, Groenewald JZ, et al. The *Botryosphaeriaceae*: genera and species known from culture. Stud Mycol. 2013 Sep;76(1):51–167.

39. Ndiaye M, Termorshuizen a. J, van Bruggen a. HC. Effects of compost amendment and the biocontrol agent Clonostachys rosea on the development of charcoal rot (Macrophomina phaseolina) on cowpea. J Plant Pathol. 2010;92(1):173–80.

40. Chamorro M, Miranda L, Domínguez P, Medina JJ, Soria C, Romero F, et al. Evaluation of biosolarization for the control of charcoal rot disease (*Macrophomina phaseolina*) in strawberry. Crop Prot. 2015;67:279–86.

41. M. N. Ecology and Management of Charcoal Rot (*Macrophomina phaseolina*) on Cowpea in the Sahel. University of Wageningen; 2007.

42. Tan DHS, Sigler L, Gibas CFC, Fong IW. Disseminated fungal infection in a renal transplant recipient involving *Macrophomina phaseolina* and Scytalidium dimidiatum: case report and review of taxonomic changes among medically important members of the *Botryosphaeriaceae*. Med Mycol. Informa Life Sci; 2008 Jan 1;46(3):285–92.

43. JG R. Transmission of seed borne *Macrophomina phaseolina* in seed. Raut JG. 1983;11:807–17.

44. Fang X, Phillips D, Li H, Sivasithamparam K, Barbetti MJ. Comparisons of virulence of pathogens associated with crown and root diseases of strawberry in Western Australia with special reference to the effect of temperature. Sci Hortic (Amsterdam). Elsevier B.V.; 2011;131(1):39–48.

45. Islam M, Haque M, Islam M, Emdad E, Halim A, Hossen QM, et al. Tools to kill: Genome of one of the most destructive plant pathogenic fungi *Macrophomina phaseolina.* BMC Genomics. BMC Genomics; 2012;13(1):493.

46. Crous PW, Slippers B, Wingfield MJ, Rheeder J, Marasas WFO, Philips AJL, et al. Phylogenetic lineages in the *Botryosphaeriaceae*. Stud Mycol. 2006;55(1915):235–53.

47. Roskov Y, Kunze T, Orrell T, Abucay L, Paglinawan L, Culham A, et al. Species 2000 & ITIS Catalogue of Life. 2014.

48. Phillips a. JL, Lopes J, Abdollahzadeh J, Bobev S, Alves a. Resolving the Diplodia complex on apple and other Rosaceae hosts. Persoonia Mol Phylogeny Evol Fungi. 2012;29:29–38.

49. Alves A, Correia A, Luque J, Phillips A. Botryosphaeria corticola, sp. nov. on Quercus species, with notes and description of Botryosphaeria stevensii and its anamorph, *Diplodia mutila.* Mycologia. 2004;96(3):598–613.

50. Dreaden TJ, Shin K, Smith JA. First Report of *Diplodia corticola* Causing Branch Cankers on Live Oak (*Quercus virginiana*) in Florida. Plant Dis. Scientific Societies; 2011 Jun 2;95(8):1027.

51. Lynch SC, Zambino PJ, Mayorquin JS, Wang DH, Eskalen A. Identification of new fungal pathogens of coast live oak in California. Plant Disease. 2013. p. 1025–36.

52. Sánchez ME, Venegas J, Romero M a., Phillips a. JL, Trapero A. Botryosphaeria and Related Taxa Causing Oak Canker in Southwestern Spain. Am Phytopathol Soc. 2003;87(12):1515–21.

53. Djoukeng JD, Polli S, Larignon P, Abou-Mansour E. Identification of phytotoxins from *Botryosphaeria obtusa*, a pathogen of black dead arm disease of grapevine. Eur J Plant Pathol. 2009;124(2):303–8.

54. Horn SJ, Vaaje-Kolstad G, Westereng B, Eijsink VG. Novel enzymes for the degradation of cellulose. Biotechnol Biofuels. 2012;5(1):45.

55. Fernandes I, Alves A, Correia A, Devreese B, Esteves AC. Secretome analysis identifies potential virulence factors of *Diplodia corticola*, a fungal pathogen involved in cork oak (Quercus suber) decline. Fungal Biol. 2014;118(5-6):516–23.

56. Schafer W. Molecular Mechanisms of Fungal Pathogenicity to Plants. Annu Rev Phytopathol. Annual Reviews; 1994 Sep 1;32(1):461–77.

57. Meng S, Torto-Alalibo T, Chibucos MC, Tyler BM, Dean R a. Common processes in pathogenesis by fungal and oomycete plant pathogens, described with Gene Ontology terms. BMC Microbiol. 2009;9 Suppl 1:S7.

58. Agrios GN. Introdution to plant pathology. In: Agrios GNBT-PP (Third E, editor. Plant Pathology. Academic Press; 1988. p. 3–39.

59. Epstein L, Nicholson RL. Adhesion and adhesives of fungi and oomycetes. Biol Adhes. 2006;41–62.

60. Polashock JJ. Screening for resistance to *Botryosphaeria* stem blight and phomopsis twig blight in blueberry. Acta Hortic. 2006;715(October):493–5.

61. Michereff SJ. Ciclo das relações patógeno-hospedeiro. 2000.

62. Aline A, Romão S. Fungos Patogênicos : Mecanismos Moleculares de Infecção e Estabelecimento na Planta.

63. Cobos R, Barreiro C, Mateos RM, Coque J-JR. Cytoplasmic- and extracellular-proteome analysis of *Diplodia seriata*: a phytopathogenic fungus involved in grapevine decline. Proteome Sci. 2010;8:46.

64. Tannu NS, Hemby SE. De novo protein sequence analysis of Macaca mulatta. BMC Genomics. 2007;8:270.

65. Chevallet M, Diemer H, Van Dorssealer A, Villiers C, Rabilloud T. Toward a better analysis of secreted proteins: The example of the myeloid cells secretome. Proteomics. 2007;7(11):1757–70.

66. Saraiva MRM. Enzimas extracelulares de fungos da família *Botryosphaeriaceae* Márcia Raquel Maia Saraiva. Universidade de Aveiro; 2009.

67. Maheshwari R, Bharadwaj G, Bhat K. M. Thermophilic Fungi: Their Physiology and Enzymes. Microbiol Mol Biol Rev. 2000;64(3):461–88.

68. Grassmann W, Dyckerhoff H. Über die Proteinase und die Polypeptidase der Hefe. 13. Abhandlung über Pflanzenproteasen in der von R. Willstätter und Mitarbeitern begonnenen Untersuchungsreihe. Hoppe-Seyler's Zeitschrift Physiol Chemie. 1928;179(1-3):41–78.

69. Barrett AJ. Classification of peptidases. In: Enzymology AJBBT-M in, editor. Proteolytic Enzymes: Serine and Cysteine Peptidases. Academic Press; 1994. p. 1–15.

70. Barrett AJ, Rawlings ND, O'Brien EA. The MEROPS Database as a Protease Information System. J Struct Biol. 2001 May;134(2–3):95–102.

71. Outtrup H, Boyce COL. Microbial Proteinases and Biotechnology. In: Fogarty W, Kelly C, editors. Microbial Enzymes and Biotechnology. Bagsvaerd, Denmark: Springer Netherlands; 1990. p. 227–54.

72. Lilley DMJ, Clegg RM, Diekmann S, Seeman NC, von Kitzing E, Hagerman P. A nomenclature of junctions and branchpoints in nucleic acids. Eur J Biochem. 1995;230(1):1–2.

73. Barrett a J, McDonald JK. Nomenclature: protease, proteinase and peptidase. Biochem J. 1986;237(3):935.

74. Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th ed. New York: W. H. Freeman and company; 2002.

75. Neurath H. The diversity of proteolytic enzymes, in proteolytic enzymes: A practical approach. Beynon, R. Oxford: IRL University Press; 1990.

76. Barrett a J, Rawlings ND, O'Brien E a. The MEROPS database as a protease information system. J Struct Biol. 2001;134(2-3):95–102.

77. Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2013;42(D1):1–7.

78. Zorn H, Peters T, Nimtz M, Berger RG. The secretome of *Pleurotus sapidus.* Proteomics. 2005 Dec;5(18):4832–8.

79. Espino JJ, Gutiérrez-Sánchez G, Brito N, Shah P, Orlando R, González C. The *Botrytis cinerea* early secretome. Proteomics. 2010 Aug;10(16):3020–34.

80. Sarmento AC, Lopes H, Oliveira CS, Vitorino R, Samyn B, Sergeant K, et al. Multiplicity of aspartic proteinases from *Cynara cardunculus* L. Planta. 2009;230(2):429–39.

81. Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2012;40(Databse issue):343–50.

82. Ghosh AK. Overview of Aspartic Acid Proteases. In: Mannhold R, Kubinyi H, Folkers G, editors. Aspartic Acid Proteases as Therapeutic Targets. 2010.

83. Miller M, Jaskolski M, Rao JKM, Leis J, Wlodawer A. Crystal structure of a retroviral protease proves relationship to aspartic protease family. Nature. 1989 Feb 9;337(6207):576–9.

84. Rawlings ND, Barrett AJ. Evolutionary families of peptidases. Biochem J. 1993;(290):205–28.

85. Gagnon-Arsenault I, Tremblay J, Bourbonnais Y. Fungal yapsins and cell wall: a unique family of aspartic peptidases for a distinctive cellular function. FEMS Yeast Res. 2006;6(7):966–78.

86. Tang J, Wong RN. Evolution in the structure and function of aspartic proteases. J Cell Biochem. 1987 Jan;33(1):53–63.

87. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the peptidase database. Nucleic Acids Res. 2010;38(Database):D227–33.

88. PB S. The aspartic proteases. Scand J Clin Lab Invest Suppl. 1992;210:5–22.

89. Erez E, Fass D, Bibi E. How intramembrane proteases bury hydrolytic reactions in the membrane. Nature. Nature Publishing Group; 2009 May 21;459(7245):371–8.

90. Polgár L. The catalytic triad of serine peptidases. Cell Mol Life Sci. 2005;62(19-20):2161–72.

91. Yousef GM, Kopolovic AD, Elliott MB, Diamandis EP. Genomic overview of serine proteases. Biochem Biophys Res Commun. 2003;305(1):28–36.

92. Hedstrom L. Serine Protease Mechanism and Specificity. Chem Rev. American Chemical Society; 2002 Dec 1;102(12):4501–24.

93. Iván G, Szabadka Z, Ordög R, Grolmusz V, Náray-Szabó G. Four spatial points that define enzyme families. Biochem Biophys Res Commun. Elsevier Inc.; 2009;383(4):417–20.

94. Hörl WH, Heidland A. Design of Enzyme Inhibitors as Drugs. In: Sandler M, Smith HJ, editors. Proteinases: Potential Role in Health and Disease. Oxford: Oxford University Press; 1989. p. 573–81.

95. Krem MM, Rose T, Di Cera E. Sequence determinants of function and evolution in serine proteases. Trends Cardiovasc Med. 2000;10(4):171–6.

96. Barrett AJ, Rawlings ND. Evolutionary lines of cysteine peptidases. Biol Chem. MRC Molecular Enzymology Laboratory, The Babraham Institute, Cambridgeshire, UK.; 2001;382(5):727–33.

97. Grzonka Z, Jankowska E, Kasprzykowski F, Kasprzykowska R, Łankiewicz L, Wiczk W, et al. Structural studies of cysteine proteases and their inhibitors. Acta Biochim Pol. 2001;48(1):1–20.

98. Sajid M, McKerrow JH. Cysteine proteases of parasitic organisms☆. Mol Biochem Parasitol. 2002;120(1):1–21.

99. Domsalla A, Melzig MF. Occurrence and properties of proteases in plant latices. Planta Med. 2008;74(7):699–711.

100. Jongeneel C V., Bouvier J, Bairoch a. A unique signature identifies a family of zinc-dependent metallopeptidases. FEBS Lett. 1989;242(2):211–4.

101. Rawlings ND, Barrett AJ. Evolutionary families of metallopeptidases. Methods Enzymol. Department of Biochemistry, Strangeways Research Laboratory, Cambridge, United Kingdom.; 1995;248:183–228.

102. Minde DP, Maurice MM, Rüdiger SGD. Determining Biophysical Protein Stability in Lysates by a Fast Proteolysis Assay, FASTpp. PLoS One. Public Library of Science; 2012 Oct 3;7(10):e46147.

103. Subramaniyan S, Sandhia GS, Prema P. Control of xylanase production without protease activity in Bacillus sp . by selection of nitrogen source. Biotechnol Lett. 2001;23:369–71.

104. Mussatto S, Teixeira J. Lignocellulose as raw material in fermentation processes. Appl Microbiol an Microb Biotechnol. 2010;2:897–907.

105. Beg QK, Kapoor M, Mahajan L, Hoondal GS. Microbial xylanases and their industrial applications: A review. Appl Microbiol Biotechnol. 2001;56(3-4):326–38.

106. Uffen RL. Xylan degradation: a glimpse at microbial diversity. J Ind Microbiol Biotechnol. 1997;19(1):1–6.

107. Chanda SK, Hirst EL, Jones JKN, Percival EG V. The Constitution of Xylan from Esparto Grass, etc. J Chem Soc. 1950;1289–97.

108. Biely P, Puls J, Schneider H. Acetyl xylan esterases in fungal cellulolytic systems. FEBS Lett. 1985;186(1):80–4.

109. López C, Blanco A, Pastor FIJ. Xylanase production by a new alkali-tolerant isolate of Bacillus. Biotechnol Lett. Kluwer Academic Publishers; 1998;20(3):243–6.

110. Beg QK, Bhushan B, Kapoor M, Hoondal GS. Production and characterization of thermostable xylanase and pectinase from Streptomyces sp. QG-11-3. J Ind Microbiol Biotechnol. Nature Publishing Group; 2000;24(6):396–402.

111. Liu W, Lu Y, Ma G. Induction and glucose repression of endo-*B*-xylanase in the yeast Trichosporon cutaneum SL409. Process Biochem. 1999;34:67–72.

112. Heikinheimo L. *Trichoderma reesei* cellulases in processing of cotton. VTT Publl.

2002;(483):3–77.

113. Huser A, Takahara H, Schmalenbach W, O'Connell R. Discovery of pathogenicity genes in the crucifer anthracnose fungus *Colletotrichum higginsianum*, using random insertional mutagenesis. Mol Plant Microbe Interact. 2009;22(2):143–56.

114. Martin K, McDougall BM, McIlroy S, Chen J, Seviour RJ. Biochemistry and molecular biology of exocellular fungal beta-(1,3)- and beta-(1,6)-glucanases. FEMS Microbiol Rev. 2007 Mar;31(2):168–92.

115. (MD) B. NCBI Help Manual. National Center for Biotechnology Information (US); 2005. 1 p.

116. Mizrachi I. Chapter 1 : GenBank : The Nucleotide Sequence Database. The NCBI Handbook. 1996. p. 1–14.

117. Cantarel BI, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. Nucleic Acids Res. 2009;37(SUPPL. 1).

118. Henrissat B, Vegetales M, Grenoble F. A classification of glycosyl hydrolases based sequence similarities amino acid. Biochem J. 1991;280((Pt 2)):309–16.

119. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42(D1):1–6.

120. Jacobs J. Characterization of BT3299: A Family GH31 Enzyme from a Prominent Gut Symbiont, Bacteroides thetaiotaomicron. 2011;108.

121. Pierleoni A, Martelli PL, Fariselli P, Casadio R. BaCelLo: A balanced subcellular localization predictor. Bioinformatics. 2006;22(14):408–16.

122. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A. 1998;95(11):5857–64.

123. Letunic I, Doerks T, Bork P. SMART: recent updates, new developments and status in 2015. Nucleic Acids Res. 2014;43(D1):D257–60.

124. Letunic I, Doerks T, Bork P. SMART : recent updates , new developments and status in 2015. 2015;43(October 2014):257–60.

125. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: an integrated documentation resource for protein families, domains and functional sites. Brief Bioinform. 2002;3(3):225–35.

126. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, et al. InterPro in

2011: New developments in the family and domain prediction database. Nucleic Acids Res. 2012;40(D1):306–12.

127. Mitchell a., Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 2014;43(D1):D213–21.

128. McDowall J, Hunter S. InterPro Protein Classification. In: Wu CH, Chen C, editors. Bioinformatics for Comparative Proteomics SE - 3. Humana Press; 2011. p. 37–47.

129. Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2014;42(D1):503–9.

130. Letunic I, Doerks T, Bork P. SMART 7: Recent updates to the protein domain annotation resource. Nucleic Acids Res. 2012;40(D1):302–5.

131. Lin SC, Yen E. Use Cases and Successful Applications of Distributed Computing Infrastructures. Data Driven e-Science. Taiwan: Springer New York; 2011. p. 540.

132. Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, Mcwilliam H, et al. Clustal W and Clustal X version 2 . 0. Bioinformatics. 2007;23(21):2947–8.

133. Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene. 1988 Dec;73(1):237–44.

134. Hang NT, Pedersen CNS, Adviser T. Comparison of multiple sequence title : comparison of multiple sequence alignment. 2008;

135. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, et al. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. 2013;41(Web Server issue):W597–600.

136. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol. 2013;30(12):2725–9.

137. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012 May;13(5):303–14.

138. Nguyen KD. Multiple Biolgical Sequence Alignment : Scoring Functions , Algorithms , and Evaluations. Georgia Stat University; 2011.

139. Will K. Principals of Phylogenetics. 2012;1–4.

140. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 2001;18(5):691–9.

141. Nei M, Kumar S. Molecular evolution and phylogenetics. New York: Oxford University Press; 2000. 333 p.

142. Hahn MW, Zhang S V., Moyle LC. Sequencing, Assembling, and Correcting Draft Genomes Using Recombinant Populations. Genes|Genomes|Genetics. 2014;4(4):669–79.

143. Ayday E, Cristofaro E De, Hubaux J-P, Tsudik G. The Chills and Thrills of Whole Genome Sequencing. Computer (Long Beach Calif). 2013;1:1.

144. Grigoriev I V., Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, et al. MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 2014;42(D1):D699–704.

145. Suryanarayanan TS. Fungal Endosymbionts of Seaweeds. In: Raghukumar C, editor. Biology of Marine Fungi SE - 3. Springer Berlin Heidelberg; 2012. p. 53–69.

146. Caffall KH, Mohnen D. The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. Carbohydr Res. Elsevier Ltd; 2009;344(14):1879–900.

147. Yernool D a., McCarthy JK, Eveleigh DE, Bok JD. Cloning and characterization of the glucooligosaccharide catabolic pathway B-glucan glucohydrolase and cellobiose phosphorylase in the marine hyperthermophile thermotoga neapolitana. J Bacteriol. 2000;182(18):5172–9.

148. Travadon R, Rolshausen PE, Gubler WD, Cadle-Davidson L, Baumgartner K. Susceptibility of Cultivated and Wild Vitis spp. to Wood Infection by Fungal Trunk Pathogens. Plant Dis. Scientific Societies; 2013 Jul 16;97(12):1529–36.

149. Wang G, Huang X, Ng TB, Lin J, Ye XY. High Phylogenetic Diversity of Glycosyl Hydrolase Family 10 and 11 Xylanases in the Sediment of Lake Dabusu in China. 2014;9(11).

150. de Vries RP, Visser J, Ronald P, de Vries, R., P. *Aspergillus* Enzymes Involved in Degradation of Plant Cell Wall Polysaccharides. Microbiol Mol Biol Rev. 2001;65(4):497–522.

151. Rosengren A, Reddy SK, Sjöberg JS, Aurelius O, Logan DT, Kolenová K, et al. An *Aspergillus nidulans* β-mannanase with high transglycosylation capacity revealed through comparative studies within glycosidase family 5. Appl Microbiol Biotechnol. Berlin/Heidelberg: Springer Berlin Heidelberg; 2014 Jun 21;98(24):10091–104.

152. Cuomo CA, Guldener U, Xu J-R, Trail F, Turgeon BG, Di Pietro A, et al. The *Fusarium graminearum* Genome Reveals a Link Between Localized Polymorphism and Pathogen Specialization. Science (80-). 2007;317(5843):1400–2.

153. Atkinson HJ, Babbitt PC, Sajid M. The global cysteine peptidase landscape in parasites. Trends Parasitol. 2009;25(12):573–81.

154. Karamanos N. Extracellular Matrix: Pathobiology and Signaling. Greece: Biochem. Lab, Department of Chemistry; 2012. 888 p.

155. Martinez M, Diaz I. The origin and evolution of plant cystatins and their target cysteine proteinases indicate a complex functional relationship. BMC Evol Biol. 2008;8:198.

156. Dall E, Fegg JC, Briza P, Brandstetter H. Structure and mechanism of an aspartimide-dependent peptide ligase in human legumain. Angew Chem Int Ed Engl. 2015;54(10):2917–21.

157. Bressano M, Lorena Giachero M, Luna CM, Ducasse D a. An in vitro method for examining infection of soybean roots by Macrophomina phaseolina. Physiol Mol Plant Pathol. Elsevier Ltd; 2010;74(3-4):201–4.

158. Premamalini T, Ambujavalli BT, Vijayakumar R, Rajyoganandh SV, Kalpana S, Kindo AJ. Fungal keratitis caused by *Macrophomina phaseolina* – A case report. Med Mycol Case Rep. Elsevier; 2012;1(1):123–6.

159. Abou-Mansour E, Débieux J-L, Ramírez-Suero M, Bénard-Gellon M, Magnin-Robert M, Spagnolo A, et al. Phytotoxic metabolites from *Neofusicoccum parvum*, a pathogen of Botryosphaeria dieback of grapevine. Phytochemistry. Elsevier Ltd; 2015;115:1–9.

160. Rawlings ND. Protease Families, Evolution and Mechanism of Action. In: Brix K, Stöcker W, editors. Proteases: Structure and Function. Springer Vienna; 2013. p. 1–36.

161. Sutton S, Project BF. Multiple Sequence Alignment : A Critical Comparison of Four Popular Programs. 2008;1–18.

162. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acid Res. 2004;32(5):1792–7.

163. Ridgeway T, Nw L. T-Coffee : A Novel Method for Fast and Accurate Multiple Sequence Alignment. 2000;

164. Zaugg C, Holdom M, Jousson O, Monod M, Capoccia S, Le B. Secreted proteases from pathogenic fungi. 2002;419.

165. Simões I, Faro C. Structure and function of plant aspartic proteinases. Eur J Biochem. 2004;271(11):2067–75.

166. Dunn BM. Introduction to the Aspartic Proteinase Family. Aspartic Acid Proteases as Therapeutic Targets. Wiley-VCH Verlag GmbH & Co. KGaA; 2010. p. 1–21.

167. Kelsey JS, Blumberg DD. A SAP domain-containing protein shuttles between the nucleus and cell membranes and plays a role in adhesion and migration in D . discoideum. 2013;

168. Sharpton TJ, Stajich JE, Rounsley SD, Gardner MJ, Wortman JR, Jordar VS, et al. Comparative genomic analyses of the human fungal pathogens Coccidioides and their relatives. 2009;1722–31.

169. Naglik JR, Challacombe SJ, Hube B. Candida albicans secreted aspartyl proteinases in virulence and pathogenesis. Microbiol Mol Biol Rev. 2003;67(3):400–28.

170. Ghadjari A, Matthews RC, Burnie JP. Epitope mapping *Candida albicans* proteinase (SAP 2). 1997;19.

171. Hofmann KAY. SAP – a putative DNA-binding motif involved in chromosomal organization. 2000;0004(99):112–4.

172. Li J, Zhang KQ. Independent expansion of zincin metalloproteinases in onygenales fungi may be associated with their pathogenicity. PLoS One. 2014;9(2).

173. Li J, Yu L, Tian Y, Zhang KQ. Molecular evolution of the deuterolysin (M35) family genes in coccidioides. PLoS One. 2012;7(2).

174. Arnadottir H, Hvanndal I, Andresdottir V, Burr SE, Frey J, Gudmundsdottir BK. The AsaP1 Peptidase of *Aeromonas salmonicida* subsp. achromogenes Is a Highly Conserved Deuterolysin Metalloprotease (Family M35) and a Major Virulence Factor. J Bacteriol. 2009;191(1):403–10.

175. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD: NCBI’s conserved domain database. Nucleic Acids Res. 2014;43(D1):D222–6.

176. Zavodszky MI, Kuhn L a. Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. Protein Sci. 2005;14:1104–14.

177. Poirier S, Mayer G, Benjannet S, Bergeron E, Marcinkiewicz J, Nassoury N, et al. The proprotein convertase PCSK9 induces the degradation of low density lipoprotein receptor (LDLR) and its closest family members VLDLR and ApoER2. J Biol Chem. 2008;283(4):2363–72.

178. Hu G, St. Leger RJ. A phylogenomic approach to reconstructing the diversification of serine proteases in fungi. J Evol Biol. 2004;17(6):1204–14.

179. Duda TF, Palumbi SR. Developmental shifts and species selection in gastropods. Proc Natl Acad Sci U S A. 1999;96(18):10272–7.

180. Upsaliensis AU. Approaches to Species Delineation in Anamorphic (mitosporic) Fungi : A Study on Two Extreme Cases BY. Sci Technol. 2004;

181. Krypotou E, Scazzocchio C, Diallinas G. Functional characterization of NAT/NCS2 proteins of *Aspergillus brasiliensis* reveals a genuine xanthine–uric acid transporter and an intrinsically misfolded polypeptide. Fungal Genet Biol. 2015 Feb;75:56–63.