Pathogenesis-Related Proteins: Phylogenetic characterization

Inaugural-Dissertation
zur
Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Nicole de Miranda Scherer
aus Porto Alegre, Brasilien

Juni 2010

Aus dem Institut für Informatik
der Heinrich-Heine Universität Düsseldorf

Gedruckt mit der Genehmigung der Mathematisch-Naturwissenschaftlichen
Fakultät der Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Martin Lercher
Koreferent: Prof. Dr. William Martin

Tag der mündlichen Prüfung: 01.06.2010

Dedicated to my Family

(all branches and leaves)

Danksagung

An dieser Stelle möchte ich all jenen danken, die durch ihre fachliche und persönliche Unterstützung zum Gelingen dieser Arbeit beigetragen haben.

Ich bedanke mich bei Prof. Arndt von Haeseler, der mich herzlich in seiner Arbeitsgruppe aufgenommen hat, und mir die Gelegenheit gegeben hat, mein Promotionsvorhaben in Deutschland zu realisieren. Darüber hinaus danke ich Prof. Martin Lercher für die Bereitschaft die Begutachtung zu übernehmen, und Prof. William Martin für die freundliche Übernahme des Korreferats.

Für die freundliche Hilfe und das Verständnis möchte ich mich ganz herzlich bei Claudia Kiometzis, Anja Walge und Angelika Simons bedanken, ohne die ich alle amtlichen Anforderungen nicht überstanden hätte.

Mein ganz besonderer Dank gilt an die Bioinformatiker für die gemütliche Stimmung und lustige "procrastination" in den richtigen und falschen Momenten. Danke Thomas "Schluffi", Jochen, Ingo P, Tanja, Andrea, Simone, Ricardo, Thomas L, Michael, Vinh, Minh, Markus, Sasha, Maren, Stephan, Dominic, Heiko, Roland, Achim, Ingo E, Lutz, Jutta, Mireille (gehört auch dazu) und Jule! Dieser Dank geht ebenfalls an die neuen Bioinformatiker der Gruppe Sabine und Gabriel, die mich aus der Ferne unterstützt haben.

Insbesondere danke ich Anja und Eike, und auch der Familie Dentz, die mich immer wieder in Ihrem Trautem Heim willkommen heißen.

Vor allem danke ich allen meinen Freunden in Deutschland für die wunderschöne Zeit zusammen. Volleyballer, UniChor, Bioinformatiker und Co, und alle die dazu gekommen sind. Ihr habt meine vier Jahre in Düsseldorf viel angenehmer gemacht.

Ich vermisse Euch!

Agradecimentos

Esta tese é dedicada a minha imensa família, com todos seus ramos e folhas, seus frutos e seus anexos, todas as pessoas que eu amo, que são muitas! O espaço desta página não será suficiente para citar cada nome em particular.

Essa imensa família tem muitos nomes: Miranda, Scherer, Klassmann, Daudt, e anexa também os laços que vão além do sangue e da lei: Beinroth, Berlowitz, Bitar, Blank, Bressel, Dentz, Fagundes, Fröhlich, Gersdorf-Fiedrich, Gilgen, Gondim, Kegler, Loureiro, Reitz, Ribas, Rocha, Schiengold, Schlegel, Schreiner, Schulze-Hofer, Selbach-Peccin, entre tantas outras.

Acima de tudo, agradeço aos meus pais o investimento financeiro e emocional na minha formação acadêmica e humana, a melhor irmã do mundo e, principalmente, o amor mais puro e sincero que pode existir.

Além da família, há aquelas pessoas que no decorrer dessa caminhada contribuíram de diferentes formas. Agradeço aos professores Loreta, Sandro e Salzano, que me apresentaram às PRs há dez anos, quando tudo começou. Agradeço às professoras Maria Luiza Campos e Ana Lucia Bazzan, que abriram as portas de seus laboratórios para mim e ofereceram a oportunidade de lecionar em suas disciplinas de bioinformática. Ao Daniel, meu colaborador preferido, agradeço a criatividade e parceria na criação do DNATagger.

Agradeço à Deya e à Rosvi as "terapias de grupo para doutorandas na Alemanha" e os exemplos de persistência. Agradeço ao meu melhor amigo, Nelson, os sábios conselhos e carinhosos xingamentos. À Karen agradeço o puxão de orelha no final da tese. Ao pessoal da Fiocruz, do NCE e aos biofísicos do IBCCF, o acolhimento e o companheirismo, mesmo que em encontros tão esporádicos. À mulherada (e aos meninos) do Verdinho, os deliciosos momentos de procrastinação e as aulas de cultivo de cáctus. E a todos os amigos que compartilharam dos meus anseios e minhas descobertas, agradeço a inspiração, as palavras de carinho e a torcida organizada.

Não menos importante é o meu agradecimento ao governo brasileiro, que, através da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), financiou a maior parte deste trabalho.

Agradeço enormemente ao meu tio Bob a dedicada correção desta tese. Aprendi, nesses meses de intensa troca de e-mails, muito mais do que qualquer curso de inglês poderia me ensinar. Nossas discussões sobre o uso adequado de termos e expressões foram enriquecedoras. Foram incontáveis mensagens recheadas de carinho, com a paciência que só um padrinho pode ter com sua afilhada cheia de questionamentos. Mais uma vez, thank you very much!!!!

Este trabalho nunca teria encontrado um fim se não fosse a persistência de três pessoas especiais: minha Alma-Gêmea, meu Anjo da Guarda, e o meu melhor Marido. À eles agradeço o apoio incondicional, a presença constante (mesmo que por telefone e internet), os puxões de orelha, as palavras de consolo, a torcida, as lágrimas, as risadas, os abraços, o carinho, a paciência, o companheirismo e o amor. Obrigada Rosvita, Mainá e Diego!

Abstract

Pathogenesis-related proteins (PR) are expressed by plants in response to pathogenic attack, conferring enhanced resistance to subsequent infection. Seventeen protein families have already been associated with this class. They include enzymes like glycoside-hydrolases (glucanases and chitinases), oxidoreductases (peroxidases, oxalate oxidases and superoxide dismutase), proteinases and proteinase inhibitors, antimicrobial peptides (defensins, thionins and lipid-transfer proteins) and other proteins with unknown function. Most of these families are multigene families, presenting from just a few to more than a hundred copies in one single genome. In some families, only a few genes have been demonstrated to be pathogenesis-related.

The main objective of this thesis was to investigate and characterize all PR-families from an evolutionary point of view, using the genetic data available in public databases, to gain insights into the question of how a protein becomes "pathogenesis-related". At the core of this project stands the configuration of a methodological framework for bioinformatic analysis that is adequate for the specific characteristics of these datasets, giving preference to free and open source software. The first goal was to assemble a representative dataset for each PR-family, and further employ it to characterize other members of these families and to search the databases for homologous sequences. To achieve this, a set of bioinformatic resources was employed sequentially, while the intermediate datasets were constantly evaluated. In the next step, the curated datasets were used to estimate the phylogeny of each family, and subsequently to infer the strength of natural selection acting on these proteins. As a result, amino acid sites predicted to be under positive selection were identified in five PR-families.

The present thesis also explores and discusses the diversity of plant chitinases and chitin-binding proteins, since they belong to four distinct PR-families, namely PR-3, PR-4, PR-8 and PR-11. In this survey, the classifications of different chitinase classes are revisited and phylogenetic analysis is employed to clarify the evolutionary relationship between family members.

Zusammenfassung

An der Pathogenese beteiligte Proteine (PR) werden von Pflanzen als Abwehrreaktion gegen angreifende Pathogene exprimiert und verleihen der Pflanze darüber hinaus bei späteren Infektionen eine erhöhte Resistenz. Siebzehn Proteinfamilien werden bereits mit dieser Klasse von Proteinen assoziiert. Dazu gehören Enzyme wie Glycosid-Hydrolasen (Glucanasen und Chitinasen), Oxidoreduktasen (Peroxidasen, Oxalatoxidasen und Superoxiddismutasen), Proteinasen, Proteinase-Inhibitoren, antimikrobiellen Peptide (Defensine, Thionine und Lipid-Transfer-Proteine) und andere Proteine mit unbekannter Funktion. Die meisten PRs gehören zu Multigenfamilien, die in einigen wenigen bis zu mehreren hundert verschiedenen Varianten in einem einzigen Genom vorhanden sein können. In manchen Familien, wurden nur wenige Gene als an der Pathogenese beteiligt identifiziert.

Das Hauptziel dieser Arbeit ist, alle PR-Familien von einem evolutivem Standpunkt aus zu untersuchen und zu charakterisieren. Die verfügbaren genetischen Daten aus öffentlichen Datenbanken sollen Anhaltspunkte liefern, wie ein Protein zu einem an der Pathogenese beteiligtem Protein wird. Im Mittelpunkt dieses Projektes steht die Konfiguration eines, auf die spezifischen Eigenschaften der Daten angepassten, methodischen Arbeitsablaufs für die bioinformatischen Analysen, wobei Freie und Open Source Software bevorzugt wird. Der erste Aspekt der Arbeit ist die Erstellung eines repräsentativen Datensatzes für jede PR-Familie. Das Verfahren wird zur Charakterisierung anderer Mitglieder dieser Familien verwendet. Weiterhin werden die Datenbanken nach homologen Sequenzen durchsucht. Zu diesem Zweck wird eine Reihe von aufeinander aufbauenden bioinformatischen Methoden eingesetzt und Teilergebnisse kontinuierlich evaluiert. Im nächsten Schritt werden ausgewählte Datensätze verwendet, um die Phylogenie jeder Familie abzuleiten und anschließend die Einwirkung der natürlichen Selektion auf diese Proteine zu untersuchen. Als Ergebnis wurden fünf PR-Familien mit positiver Selektion identifiziert.

Die vorliegende Arbeit untersucht und diskutiert im weiteren die Vielfalt der pflanzlichen Chitinasen und Chitin-bindenden Proteine, da diese zu vier verschiedenen PR-Familien, nämlich PR-3, PR-4, PR-8 und PR-11, gehören. In dieser Arbeit wurde die Einteilung der verschiedenen Klassen von Chitinasen überprüft. Phylogenetische Analysen wurde eingesetzt, um die evolutionären Beziehungen zwischen den Familienmitgliedern zu ermitteln.

Publications

Articles:

N. M. Scherer, C. E. Thompson, L. B. Freitas, S. L. Bonatto, and F. M. Salzano. (2005) Patterns of molecular evolution in pathogenesis-related proteins. *Genet. Mol. Biol.*, 28(4):645–653. ISSN 1415-4757. DOI: 10.1590/S1415-47572005000500001

N. M. Scherer and D. M. Basso. (2008) DNATagger, colors for codons. *Genet. Mol. Res.*, 7(3):853–860. ISSN 1676-5680. DOI: 10.4238/vol7-3x-meeting003

Conference proceedings:

N. M. Scherer, C. E. Thompson, L. B. Freitas, S. L. Bonatto, and F. M. Salzano. (2006) Evolutionary Analysis in Pathogenesis-Related Proteins. In: *NIC Workshop: From Computational Biophysics to Systems Biology*, Jülich – Germany. NIC Series. John von Neumann Institute for Computing, v. 34. p. 193–196.

N. M. Scherer and D. M. Basso. (2007) DNATagger, colors for codons. In: *X-meeting 2007, The Third Annual Conference of the Brazilian Association for Bioinformatics and Computational Biology*, São Paulo – Brazil. X-meeting 2007, Proceedings.

N. M. Scherer. (2008) Evolução Molecular em Proteínas Relacionadas à Patogênese em Plantas. In: *I Escola Brasileira de Bioinformática (EBB)*, Santo André – Brazil.

N. M. Scherer. (2009) Phylogenetic characterization of glycoside hydrolase family 18 pathogenesis-related proteins using profile HMM and maximum likelihood methods. In: *Brazilian Simposium on Bioinformatics (BSB)*, Porto Alegre – Brazil.

N. M. Scherer. (2009) Interactive methodological framework for evolutionary analysis of pathogenesis-related proteins. In: *5th International Conference of the Brazilian Association for Bioinformatics and Computational Biology*, Angra dos Reis – Brazil. X-Meeting Eletronic Abstracts Book 2009, p. 17.

Supplementary Material

Supporting information is provided with this thesis in a CD-ROM. The files correspond to the final dataset of all PR-families.

Supplementary_Material/Tables: The tables included in this folder were automatically produced during the filtering procedure. They reproduce information from the databank annotation.

Supplementary_Material/Alignments_phylip: Two alignments are provided for each family, one codon alignment and one amino acid alignment. The amino acid alignments are direct translations of the codon alignments. All alignments provided in this folder are stored in plain phylip – sequential format. A Perl script for complementing the names of the sequences in the alignment is available.

Supplementary_Material/Trees: The trees stored in this folder were constructed with PhyML, whereas the branch lengths were estimated with codeml, model M0 (one rate). All trees are in NEXUS file format. A Perl script for complementing the names at the tips of the trees is also available.

Supplementary_Material/Alignments_DNATagger: Contains the colored alignments of the SEED dataset as well as of the EXTENDED–Final dataset in PDF/A format. The sequences were colored by DNATagger.

Abbreviations

PR	**P**athogenesis-**R**elated Protein
HR	**H**ypersensitive **R**eaction
ROS	**R**eactive **O**xygen **S**pecies
SAR	**S**ystemic **A**cquired **R**esistance
TMV	**T**obacco **M**osaic **V**irus
ABA	**AB**scisic **A**cid
SA	**S**alicylic **A**cid
JA	**J**asmonic **A**cid
ET	**ET**hylene
BSP	**B**asic **S**ecretory **P**rotein
BTH	**B**enzo**TH**iadiazole
GlcNAc	N**-ac**etil-D-**gl**u**c**osamine
MSA	**M**ultiple **S**equence **A**lignment
AOS	**A**verage **O**verlap **S**core
MOS	**M**ultiple **O**verlap **S**core
HMM	**H**idden **M**arkov **M**odel
NJ	**N**eighbor **J**oining
ML	**M**aximum **L**ikelihood
LRT	**L**ikelihood **R**atio **T**est
AIC	**A**kaike **I**nformation **C**riterion
AICc	**c**orrected **A**kaike **I**nformation **C**riterion
BIC	**B**ayesian **I**nformation **C**riterion

Contents

Introduction

Pathogenesis-related (PR) proteins are expressed in plants in response to pathogenic attack, conferring enhanced resistance to subsequent infection. The proteins represented in this class usually belong to multigenic families, showing a variety of forms, functions, tissue specificities and expression patterns. This class encompasses to date 17 protein families, including chitinases, glucanases, proteinase inhibitors, peroxidases, defensins, thionins, lipid-transfer proteins, thaumatin-like proteins, germin and germin-like proteins, among others, with unknown function. What do they have in common? (There must be one thing they have in common!) How do they differ from each other?

Since the discovery of the PRs in 1970, an increasing amount of molecular data from pathogenesis-related proteins has been accumulated in protein and nucleotide databases. With the advance of bioinformatics, the use of the stored sequences to produce knowledge has been facilitated. What can we learn from this data? What do protein and nucleotide sequences from different species tell us? How can bioinformatics help in the understanding of pathogenesis-related protein evolution?

In the present work, computer based analyses are used to unveil the evolutionary history of pathogenesis-related proteins.

The main objective of this work is to investigate and characterize the PR-families from an evolutionary point of view, using the genetic data available in public databases. At the core of this project is the configuration of a methodological framework for bioinformatic analysis that is adequate for the specific characteristics of these datasets, giving preference to free and open source software.

The first goal is to assemble a representative dataset for each PR-family, which can be further employed to characterize other members of these families, and to search the databases for homologous sequences. The next goal is to use these curated datasets to estimate the phylogeny of each family, and subsequently to infer the strength of natural selection acting on these proteins. More specifically, this step aims to identify amino acid sites that are predicted to be under positive selection.

The present work also aims to explore and discuss the diversity of plant chitinases and chitin-binding proteins, since they belong to four distinct PR-families, namely PR-3, PR-4, PR-8 and PR-11. In this survey, the classifications of different chitinase classes are revisited and phylogenetic analysis is employed to clarify the evolutionary relationship between family members.

The full comprehension of this thesis requires background information on three main topics:

- Biology of plant defense and pathogenesis-related proteins
- Process of molecular sequence evolution and natural selection
- Bioinformatics of phylogenetic analysis and inference

These topics are reviewed in Chapter 1 (Background theory). The framework used to perform phylogenetic analysis is described in Chapter 2. A general description of the results obtained for all PR-families is presented in Chapter 3, while detailed descriptions of the results obtained for chitinases and chitin-binding proteins (PR-3, PR-4, PR-8 and PR-11) are given in Chapter 4.

———-

Several times, I was told: "You must decide whether your thesis is methodological or biological". I decided: both! I cannot disconnect the components of bioinformatics. In my work, they grew up together. The methodology I pieced together is the result of long time experimentation with the same protein families. Of course, the results can be considered in separate. I present here one chapter dedicated to the general results, where the importance of each component of the methodological workflow is discussed, and one chapter where specific results of selected protein families are considered in detail.

Chapter 1

Background

In silico studies of molecular evolution require a good background in knowledge obtained from *in vivo* and *in vitro* research. Bioinformatic methods evolve as the biological knowledge becomes more complex.

A brief description of the pathogenesis-related proteins and defense strategies employed by plants is presented in the first section of this chapter, followed by an introduction about the evolution of multigenic families. The last section reviews bioinformatic methods and tools used in phylogenetic analysis.

1.1 Pathogenesis-Related Proteins

Pathogenesis-related (PR) is the denomination for a class of plant-produced proteins expressed in response to pathogenic attack, conferring enhanced resistance to subsequent infection. In the last 30 years, the definition of pathogenesis-related proteins has been constantly reviewed [1–7]. Today, the authors have reached a consensus defining PRs as proteins encoded by the host plant and induced by various types of pathogens such as viruses, bacteria, fungi, and animals including phytophagous insects, nematodes and herbivores, as well as by the exposure to abiotic stress factors (see box below) [3, 4, 8]. In other words, PR-proteins are produced by plants in a context of pathogenesis or physiological danger for the organism. They belong to the ensemble of proteins involved in the mechanism of systemic acquired resistance. In this type of induced resistance, non-infected parts of previously infected plants become resistant to future infections [9].

Biotic stresses

- Viral infection
- Bacterial infection
- Fungal infection
- Nematode infection
- Herbivory

Abiotic stresses

- Wounding
- Osmotic stress
- Drought stress
- Senescence
- Chemical treatment

1.1.1 Plant Defense against Pathogens

Plants are constantly attacked by other organisms. Viral, bacterial and fungal pathogens, as well as herbivorous animals or even parasitic plant species, may infest plants in all organs, below and above ground. Disease, though, is not the rule, since preformed anatomical and biochemical barriers prevent most of the infections. When the pathogen overcomes this first barriers, the mutual perception between plants and their pathogens triggers induced defenses, including the expression of pathogenesis-related proteins (PRs) [5, 10, 11].

Constitutive Defense Preformed barriers, such as the cell wall, cutin, waxes, hairs and thorns, are typical traits of constitutive defense, together with poisonous secondary plant metabolites.

Induced Defense The main defense mechanisms induced by pathogens are i) cell wall stiffening by cross-link of cell wall components, lignification, formation of papillae, callose deposition; ii) production of phytoalexins, which present antibiotic effect against a broad spectrum of fungi and bacteria; and iii) synthesis of pathogenesis-related proteins.

1.1.2 Induced Resistance

Induced resistance is a state of enhanced defensive capacity, naturally triggered by the exposure of plants to virulent, avirulent or non-pathogenic microbes, or artificially induced by various chemical agents [12]. Primary defense responses, induced in cells surrounding the site of infection, prevent the proliferation and expansion of pathogens

[13]. The secondary (induced) defenses often act systemically throughout the plant and are typically effective against a broad spectrum of attackers [14].

In incompatible interactions (where the host is resistant to the pathogen), cellular components that are liberated during the interaction become elicitors, molecules that can be recognized by the defense system of the host. Elicitors can be of pathogen or host origin, either fungal enzymes or hydrolysis products of the pathogen or host cell wall. The binding of a proper elicitor to a host receptor induces the defense responses [13]. In compatible interactions, the host fails to detect the pathogen due to the lack or suppression of proper receptors (i.e., the defense responses are not induced and the infection is successful).

Induced defenses involve phytohormone-mediated signal transduction that links the damage with the phenotypic change in the plant [15]. There are three main signal transduction pathways that underlie induced defenses:

jasmonate pathway characterized by phytohormone jasmonic acid (JA)

shikimate pathway characterized by phytohormone salicylic acid (SA)

ethylene pathway characterized by phytohormone ethylene (ET)

Other plant hormones, including abscisic acid (ABA), auxin, and brassinosteroids, have also been implicated in plant defense, but their significance is less well studied [14].

The jasmonate pathway is mostly involved in insect-induced and wound-induced plant responses, but is not exclusive of this interaction. JA is a typical stress hormone and is known to act synergistically with ABA and ET, while the interaction between JA and SA can have either mutually antagonistic or synergistic effects. Salicylic acid is often related to virus and fungal infection. SA is predominantly effective against biotrophic pathogens (pathogens that establish a long-term feeding relationship with the living cells of their hosts, rather than killing the host cells), whereas necrotrophic pathogens and herbivorous insects induce more JA/ET-mediated defenses (see [13–16] and references therein). Ethylene and auxin are the two most-studied phytohormones with regard to effects on plant-nematode interactions. ET is also involved in senescence, abscission, and fruit ripening [17].

Hypersensitive Response

The front line of the induced resistance is the hypersensitive response (HR), characterized by the induction of rapid localized cell death in the infected area. During HR, the plant liberates reactive oxygen species (ROS) at the site of the infection, leading to an oxidative burst and killing both pathogens and infected cells [13].

Systemic Acquired Resistance

One of the consequences of the HR is the establishment of a systemic acquired resistance (SAR). While vertebrates have an immune system based on clonal selection of antibodies, which provides acquired immunity through "immunological memory", plants develop a "stress memory", which is responsible for acquired resistance [5, 10, 11].

Systemic acquired resistance is characterized by the activation of a cascade of host defense responses (including the synthesis of PR-proteins), locally at the site of the initial pathogen's attack and systemically in non-infected plant organs remote from the infected site [18]. In an analogous manner to vaccine immunization, a new pathogen will fail to infect the plant because the defense mechanisms are already preformed. This protects the plant against further infection. Furthermore, SAR is effective against a broad spectrum of pathogens other than the pathogen of the primary infection [13].

1.1.3 PR-Families

Pathogenesis-related proteins usually belong to multigenic families, showing a variety of forms, functions, tissue specificities and expression patterns.

Seventeen classes of PR-proteins have been described, comprising four families of chitinases (PR-3, PR-4, PR-8 and PR-11), a family with unknown biochemical properties (PR-1), homologous to wasp venom allergen, the 1,3-glucanases (PR-2), the thaumatin-like PR-5 family, proteinase inhibitors (PR-6), a subtilisin-like endoproteinase (PR-7), plant peroxidases (PR-9), the birch allergen Betv1-related PR-10 family, plant defensins (PR-12), thionins (PR-13), nonspecific lipid-transfer proteins (PR-14), germin and germin-like proteins (PR-15 and PR-16), and a basic secretory protein classified as PR-17 [8, 19, 20].

1.2 Multigene Families and Sequence Evolution

Pathogenesis-related proteins are usually members of multigenic families. As an example, PR- 9 proteins belong to the class III peroxidase (POX) multigenic family, for which the recent completion of the rice genome sequencing has allowed the identification of 138 genes and 14 pseudogenes, distributed among the 12 rice chromosomes [21]. In the *Arabidopsis thaliana* genome, Welinder and colleagues (2002) identified 73 full-length POX genes, two pseudo-genes, and six fragments spread rather evenly on the five *Arabidopsis* chromosomes [22]. PR-3, in turn, can be subdivided into several subfamilies, while PR-15 and PR-16 belong to the same superfamily.

Plant genes are particularly affected by gene duplications. Although found across eukaryotic lineages, gene duplication seems to occur at an elevated rate in plants [23]. Gene family members may be tandem duplicates, dispersed duplications, or genome-wide duplications.

1.2.1 Gene Duplication

A hundred and fifty years ago, Charles Darwin already proposed that “parts”, many times repeated, would improve diversity and adaptation [24]. In 1932, J. B. S. Haldane suggested that duplication events might be favorable because they produce genes that can be altered without disadvantage to the organism [cited by 25]. Organisms with multiple copies of genes would be, therefore, less prone to harmful mutations.

Michael Lynch (2002) remarks that duplicate genes could provide the ultimate substrate on which evolution could work. Either one member of the duplicate gene pair could take on a new function (neofunctionalization), or two duplicate genes could divide the multiple functions of the ancestral gene between them (subfunctionalization). The acquisition of copy-specific mutational refinements and complementary degenerative mutations promote the preservation of both copies. In most cases, however, one copy may be silenced by degenerative mutations (nonfunctionalization) and becomes a pseudogene [26–29].

1.2.1.1 Particularities of Plant Genomes

Gene duplication appears to occur at an elevated rate in plants [23]. The *de novo* gene creation from raw DNA (or RNA) sequence must have occurred early in the history of life on earth. The majority of the "new" genes arise from the duplication, rearrangement and divergence of pre-existing genes [30]. Numerous mechanisms exist for increasing gene number by segmental or full genome duplication.

Segmental Duplication Local duplications are tandem duplications that occur through transposable element activity, replication slippage or unequal recombination. Tandem duplications can promote rapid increase in gene copies, and the duplicated genes can be also transferred to different chromosomes [25, 30].

Polyploidy Polyploids arise either by duplication of the genome within a single species (autopolyploidy) or the acquisition of genomes from two closely related species into the same nucleus (allopolyploidy) [30]. Polyploidy duplicates every gene in the genome, providing the raw material for divergence or partitioning of function in *homoeologous* copies [31]. It has been previously proposed that 70% of all angiosperm species and 95% of pteridophytes are of polyploid origin [32]. Polyploidization events must have occurred multiple times in evolutionary history, and probably 100% of flowering plants are current polyploids or have a polyploid history [33, 34]. Therefore, most of today's "diploid" angiosperm species are now considered to carry paleopolyploid genomes [35].

Polyploidization is considered to be the primary source of duplicate loci in plants, with ploidy levels in angiosperms reaching 2n=640 (approx. 80x) in the stonecrop *Sedum suaveolens*, and up to 2n=1260 (approx. 84x) in the fern *Ophioglossum pycnostichum* [23]. The prevalence of polyploids in plants probably reflects the evolutionary and ecological advantage of having extra gene copies [36]. Even though polyploidy alone does not explain all the redundancy (gene copies) found in plants, it is still one of the major processes shaping the evolution of plant genomes.

1.2.2 Homology

Homology is determined by common ancestrality. Two genes are homologous when they are derived from a common ancestor. Traditionally, homologous genes are classified on the basis of their origin.

Orthology Genes are said to be orthologous when they originate from a single ancestral gene in a speciation event. (The same gene in different species)

Paralogy Gene duplication is the origin of paralogous genes. (More than one gene copy in the same organism)

Complex cases Naturally, this binarity cannot explain the universe of possible consecutive events that originate the multigenic families observed in nature. Trying to adjust to it, an intrincate nomenclature has already been proposed, but is not widely employed. Terms such as “xenologs”, “pseudoorthologs”, “co-orthologs”, “inparalogs” (symparalogs), “outparalogs” (alloparalogs) and “pseudoparalogs” are very well discussed by Fitch [37] and Koonin [38].

1.2.3 Natural Selection

> *As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be* naturally selected*.*
>
> Charles Darwin, 1859

Darwin called the preservation of favorable variations and the rejection of injurious variations “natural selection”. He also pointed out that variations which are neither useful nor injurious would not be affected by natural selection, and would be left as fluctuating elements. He was, indeed, convinced that natural selection has been the main, but not exclusive, means of modification [24]. This gave rise afterwards to fruitful discussions, that can be cited as the *selectionism versus neutralism* (and the *mutationism* *versus selectionism*) debate [see 39, and references therein]. Today, evolutionists agree that all these processes are important in phenotypic evolution as well as in molecular evolution. However, the discussion remains on the relative importance of mutation, selection and chance in the whole process of evolution.

Positive Selection Also called positive Darwinian selection, refers to the increase in fitness by advantageous variations. As explained by Darwin, "individuals having any advantage, however slight, over others, would have the best chance of surviving and of procreating their kind".

Negative Selection (or purifying selection) causes the removal of harmful variations. (Darwin: "we may feel sure that any variation in the least degree injurious would be rigidly destroyed.")

Neutral Evolution where the existing variation does not influence the chance of survival and reproduction, so that variants become eventually fixed by random genetic drift (the "fluctuating elements" referred to by Darwin).

Many other names have been introduced to define different subcategories of these terms (e.g. adaptive selection, diversifying selection, directional selection, nearly neutral selection) and will be occasionally discussed later.

1.2.3.1 Natural Selection at Molecular Level

In molecular evolution we can exemplify these evolutionary processes by comparing homologous proteins. The patterns of amino acid substitution reflect the selective forces acting on the proteins. Multigene families offer very didactic examples of selection and neutral evolution.

Structurally and functionally important regions of the protein are expected to be subject to purifying selection. Amino acid substitutions may not be accepted, or only similar amino acids will be allowed. Frequently, cysteine residues implicated in disulfide bond formation are conserved in all members of a family. Substitutions at these sites may inactivate the protein, being deleterious.

On the other hand, there can be sites that will not be at all affected by amino acid changes, and mutations are considered neutral [40]. In other cases, there are preferred amino acids for certain regions, such as leucine, isoleucine, valine, alanine and phenylalanine residues in the signal peptide [41].

There may, however, be cases in which a substitution for a dissimilar amino acid in a functionally important region can confer an adaptive gain. In these cases, positive selection will favor the new variant (directional selection), that may become fixed. And there is also diversifying selection, improving variability among multigene family members, as in the case of the adaptive immune system of vertebrates [42].

Naturally, many other processes can be observed in protein sequences, but they will not be discussed until later.

1.3 Phylogenetic Bioinformatics

1.3.1 Protein Sequence Database

A variety of protein sequence databases have been created to store the continuously increasing amount of data produced. Specialized databases contain information about a particular protein family or groups of proteins, or are related to a specific organism, while universal protein databases cover protein from all species. There is a distinction between sequence repositories and expertly curated databases [43].

UniProtKB

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data [44]. UniProt is maintained by the UniProt Consortium, a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). The UniProt Knowledgebase (UniProtKB), one of UniProt's four components, has been created from Swiss-Prot, TrEMBL and PIR-PSD (see box below) to be a single, stable, high quality, comprehensive and authoritative protein knowledgebase [43, 44].

Swiss-Prot is a curated protein sequence database containing fully manually annotated entries with a minimal level of redundancy and high level of integration with other databases [45–47].

TrEMBL (Translated EMBL Nucleotide Sequence Data Library) consists of computer-annotated entries derived from the translation of all the coding sequences (CDS) deposited in the EMBL/GenBank/DDBJ nucleotide sequence databases, waiting to be manually curated by an annotator. TrEMBL is therefore a complement to Swiss-Prot [45, 46, 48].

PIR-PSD (Protein Information Resource - International Protein Sequence Database), the world's first database of classified and functionally annotated protein sequences, had been discontinued by the end of 2004. PIR-PSD sequences and annotations have been integrated into the UniProt Knowledgebase [49, 50].

UniProtKB consists of two sections, referred to as the Swiss-Prot (UniProtKB/Swiss-Prot) and the TrEMBL (UniProtKB/TrEMBL) sections. The section UniProtKB/Swiss-Prot contains manually annotated high quality records with information extracted from literature and curator-evaluated computational analysis. The annotation is performed by biologists with specific expertise and brings together experimental results, computed features and scientific conclusions. The section UniProtKB/TrEMBL contains computationally analyzed records enriched with automatic annotation and classification [44]. Together, they cover all the proteins characterized or inferred from all publicly available protein-coding nucleotide sequences [48].

It is worthy of note that the complete UniProtKB/TrEMBL section is approximately ten times larger (in file size) than the complete UniProtKB/Swiss-Prot section. In number of entries, this proportion increases to 20 TrEMBL sequences for each manually curated protein. This difference is (clearly) due to the larger amount of information contained in a fully annotated entry file. In the particular case of human data, the UniProtKB/Swiss-Prot dataset is 20% larger than the unannotated dataset. Not surprisingly, *Homo sapiens* (human) is the most represented species in UniProtKB/Swiss-Prot, with 20,330 entries, followed by *Mus musculus* (mouse), represented in 16,140 entries, and *Arabidopsis thaliana* (Mouse-ear cress), a plant model organism with 8,338 entries (UniProtKB/Swiss-Prot Release 57.4 of 16-Jun-2009). At UniProtKB/TrEMBL, on the other hand, there are 274,327 entries for human immunodeficiency virus 1 (HIV1). *Oryza sativa* subsp.

japonica (rice) is the second in this rank, with 94,803 entries, approximately one third of the number of HIV1 entries. *Homo sapiens* appear in the third place, represented in 66,494 of the not yet annotated sequences (UniProtKB/TrEMBL Release 40.4 of 16-Jun-2009) [51, 52].

Annotation

For the manual annotation of the proteins deposited in UniProtKB/Swiss-Prot, biologists use scientific literature and bioinformatic analysis, such as multiple sequence alignments with paralogous or orthologous proteins, to validate the described information [48, 53]. The available evidence for the existence of a protein is classified in five levels ("Evidence at protein level"," Evidence at transcript level", "Inferred from homology", "Predicted", and "Uncertain"). In the absence of explicit experimental evidence, the level of confidence of the information is indicated by the non-experimental qualifiers: "Potential" (predicted by computer analysis), "Probable" (inferred from sources other than literature) or "By similarity" (demonstrated for a homologous protein).

Plant Specific Database Collection

Although two plant model organisms are on the top list of the most represented species in both database sections, the entries of Viridiplantae represent 8% (650,985) of the complete UniProtKB/TrEMBL section (8,594,382 entries) and only 6% (27,824 entries) of the UniProtKB/Swiss-Prot entries (470,369) (same release data). For each fully annotated plant sequence there are another 20 computer-annotated sequences awaiting manual (expertise) annotation [51, 52, 54]. Furthermore, half of all the entries from Viridiplantae originate from the 10 most highly represented species [48].

In the UniProtKB/Swiss-Prot Release 57.4 of 16-Jun-2009, the entries from *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays* (maize), *Nicotiana tabacum* (common tobacco), *Solanum lycopersicum* (or *Lycopersicon esculentum*) (tomato), *Solanum tuberosum* (potato), *Pisum sativum* (garden pea), *Glycine max* (soybean), *Hordeum vulgare* (barley) and *Triticum aestivum* (wheat) add up to 14,222 sequences, 8,338 alone being from *Arabidopsis thaliana*.

PPAP The Plant Proteome Annotation Program (PPAP) of UniProtKB/Swiss-Prot focuses on the manual annotation of plant-specific proteins and protein families. It is currently focused on the annotation of proteins from the fully sequenced model plant organisms *Arabidopsis thaliana* and *Oryza sativa*, without neglecting annotation of proteins from other plant species, with an emphasis on species that are the target of genomic, proteomic or transcriptomic projects (maize, wheat, soybean, etc.). These efforts might also help to identify and reveal the function of proteins originating from other plants [48, 53].

1.3.2 Multiple Sequence Alignments

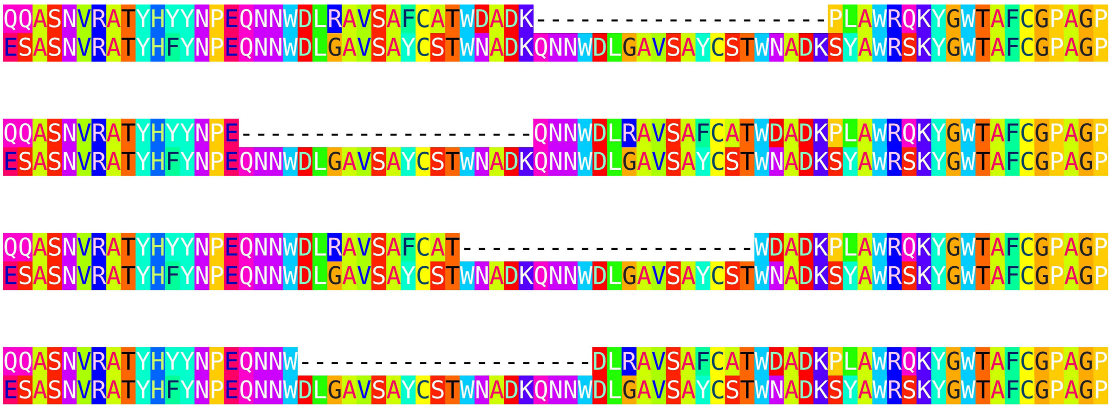
The most widely used representation of the relationship between biological sequences is the sequence alignment. The comparison of two related sequences is defined as pairwise alignment. When more than two sequences are to be compared, they are represented in a multiple sequence alignment (MSA), which is in most cases built up from several pairwise alignments.

An MSA can be represented by a two-dimensional array, in which each row fits one sequence and the columns represent single sites. A gap in an alignment represents an assumed event of insertion or deletion (indel), that may have occurred at some point in the evolutionary history, after the divergence of the aligned sequences. The commonly used symbol for representing an indel is a dash (-). Some alignment representation formats (e.g. Stockholm format) differentiate insertions from deletions using dots (.) for insertions and dashes for deletions.

Gaps are artifices of alignment algorithms to allow sequences of different length or with a history of insertions and deletions to be aligned. The process of alignment involves the introduction of gaps into sequences in order to find maximal matching scores. These gaps illustrate hypotheses about historical indel events. They are not part of the peptide or DNA molecule. A common mistake is to consider a gap to be an informative character in the interpretation of sequences. Of course, gaps are very informative about the evolution of a lineage when compared to others. But they have no biological meaning regarding the function of the gene.

Sequence alignments can be structurally based or evolutionary motivated. In structural alignments, residues are assigned to the same column if they are considered structurally equivalent. An evolutionary alignment, however, assumes that aligned residues have originated from the same residue in a common ancestor. Both kinds of alignments should converge if we consider that structural features represent important constraints in sequence evolution. The approaches diverge when the sequences become very dissimilar. It is even possible to structurally align non evolutionarily related proteins, that share common structural motifs.

Ideally, each column in a multiple sequence alignment must represent homologous sites, i.e. all elements placed in one column derived from a single site in the common ancestral sequence. Unfortunately, this is not always the case. Because of the diversity of processes involved in sequence evolution, not every event is likely to be reliably represented in the form of an alignment. Sequence inversions and translocations, for example, will not fit in an alignment (unless indirectly described by the use of parentheses and arrows). Moreover, segment duplications can be tricky, since the decision of which of the duplicated segments should be aligned to the single segment in the other sequences is arbitrary (see Fig. 1.1).



```
QQASNVRATYHYYNPEQNNWDLRAVSAFCATWDADK--------------------PLAWRQKYGWTAFCGPAGP
ESASNVRATYHFYNPEQNNWDLGAVSAYCSTWNADKQNNWDLGAVSAYCSTWNADKSYAWRSKYGWTAFCGPAGP

QQASNVRATYHYYNPE--------------------QNNWDLRAVSAFCATWDADKPLAWRQKYGWTAFCGPAGP
ESASNVRATYHFYNPEQNNWDLGAVSAYCSTWNADKQNNWDLGAVSAYCSTWNADKSYAWRSKYGWTAFCGPAGP

QQASNVRATYHYYNPEQNNWDLRAVSAFCAT--------------------WDADKPLAWRQKYGWTAFCGPAGP
ESASNVRATYHFYNPEQNNWDLGAVSAYCSTWNADKQNNWDLGAVSAYCSTWNADKSYAWRSKYGWTAFCGPAGP

QQASNVRATYHYYNPEQNNW--------------------DLRAVSAFCATWDADKPLAWRQKYGWTAFCGPAGP
ESASNVRATYHFYNPEQNNWDLGAVSAYCSTWNADKQNNWDLGAVSAYCSTWNADKSYAWRSKYGWTAFCGPAGP
```

Figure 1.1: Four alternative pairwise alignments of an internal duplication. Alignment colored by DNATagger [55]

How are pairwise sequence alignments constructed? Given two small strings of similar length, we could align them by sliding one in respect to the other, inserting gaps in both sequences, until we find the alignment that has the greatest number of matches and the smallest number of gaps.

A scoring system is also needed to assign values for matches, mismatches and gaps (gap penalties). Elaborated scoring schemes assign specific scores to each pair of amino acids or nucleotides (substitution matrices) and also differ between gap opening and gap extension (affine gap penalties) [56]. Traditional pairwise alignment methods use position-independent scoring parameters. Position-specific scores (profile methods) for amino acids or nucleotides and position specific penalties for opening and extending an insertion or deletion can be alternatively applied. For a more detailed explanation on this topic, we refer the interested reader to [56–58].

In practice, it is unfeasible to evaluate every possible way of aligning two sequences, because the number of alternative alignments increases factorially with sequence length. To solve this problem in polynomial time, Needleman and Wunsch (1970) introduced an iterative matrix method of alignment calculation, using dynamic programming [59]. All possible pairs of amino acids (or nucleotides) are represented by a two-dimensional array, and all possible comparisons are represented by pathways through the array. Each element of this array stores a value that represents the maximum score for the partial alignment ending at this point.

Given two sequences X and Y, the score for the partial alignment ending at sites X_i and Y_j is stored at element $d_{i,j}$. The value in $d_{i,j}$ is the maximum value among the three scores: the score of the pathway coming from the diagonal $d_{i-1,j-1}$ (producing a match or mismatch), or from one of the adjacent elements in either the same row $d_{i,j-1}$ (indicating the insertion of a gap in sequence X), or in the same column $d_{i-1,j}$ (introducing a gap in sequence Y). The best alignment is defined by the pathway leading to the maximal score (or minimal cost) in the very last element of the array. The advantage of the Nedelman-Wunsch algorithm is that only a reduced set of the possible pathways needs to be evaluated. Nevertheless, it is an exact algorithm, and it will always give as solution the alignment with the highest score (given the scoring scheme).

There are two basic modes of sequence alignment:

global ⟶ over the entire length of the sequences [59]
local ⟶ only high-scoring (highly similar) regions are considered [60]

The Needleman-Wunsch algorithm calculates a global alignment. A method to calculate local alignments was developed by Smith and Waterman in 1981. It aims to identify common subsequences in a pair of long sequences. The algorithm searches for pairs of maximally similar segments, considering the possibilities for ending at any pair of residues. The innovative strategy was to include a "zero" in the objective function to "reset" negative scores. This allows the finding of high scoring segments, even when they are surrounded by highly variable regions [60].

The extension of the dynamic programming algorithms for more than two sequences increases the magnitude of the problem to "N-dimensional" matrices. The memory usage is proportional to L^N, where L is the length of the sequences and N is the number of sequences. Here again, the optimal alignment is the alignment with the maximum score. The finding of the path that maximizes the score (minimizes the costs) through the "N-dimensional" matrix is an NP-hard problem.

A commonly used objective function to assign a score to an MSA is the "sum-of-pairs" (SP) measure, the sum of the pairwise scores for every pair of rows in the alignment. The alignment score of a pair of sequences is computed as the sum of substitution matrix scores for each aligned pair of residues, plus gap penalties. The calculation of the SP score grows exponentially with the sequence length and the number of sequences to be aligned. The SP function optimized by Carrillo and Lipman (1988) reduces the search space for an optimal alignment. The algorithm is based on the idea that every multiple alignment imposes a pairwise alignment on each pair of sequences. Considering each imposed pairwise alignment as a projected path in a standard two-dimensional path graph, they calculate upper bounds on the cost of pairwise alignments, thus limiting the points through which the projection can possibly pass. The method still guarantees the discovery of an optimal alignment, but is also limited to a few sequences [61, 62]. In addition to being computationally expensive, the SP measure ignores the structure of any associated phylogenetic tree and thus attributes greater weight to some evolutionary events than to others. The SP measure is not appropriate when evolutionary distances between members are not evenly distributed.

The "weighted sum of pairs" (WSP) measure, proposed by Altschul and colleagues (1989) and optimized by Gotoh (1995), adjusts the weights given to individual sequence pairs in order to compensate biased datasets. The WSP score depends, in turn, on the guide tree used [63–65]. Also, using either the SP or the WSP measure, each multiple alignment of N sequences implies the calculation of N(N-1)/2 pairwise scores. Therefore, heuristics must be used to construct multiple sequence alignments.

One interesting approach is to cut the sequences, compute short optimal SP alignments and concatenate them to a full alignment. This is the idea underlying the "divide-and-conquer" alignment algorithm (DCA), that provides near-to-optimal results for sufficiently homologous sequences [66–68]. Similarly, fragment based MSA compares whole segments of the sequences instead of single residues, and searches for high scoring diagonals (DIALIGN). The final alignment is based on a set of high scoring diagonals that are consistent in the overall order of the positions in each sequence [69].

The most widely used heuristic for MSA calculation is the strategy of progressive (hierarchical) alignments proposed by Feng and Doolittle (1987) [70]. Progressive alignment methods iteratively align pairs of sequences or sequence profiles. The preferred order of including sequences in the alignment is determined by a guide tree, generally constructed by applying a clustering method to the pairwise distances between all sequences. The two most recently diverged sequences are aligned first, and each new incorporated sequence is "pairwisely" aligned to the set of already aligned sequences. This strategy reduces the problem to a different type of pairwise alignment, in which partial MSAs are reduced to one-dimensional sequences that can therefore be used in pairwise dynamic programming alignment. Depending on the tree, a set will be aligned to another set. A set of aligned sequences can be represented as a profile, as a consensus sequence, as a set of probabilities, or as a directed acyclic graph, as recently implemented for partial order MSA [71].

As already stated by Feng and Doolittle, "The thrust of the method involves putting more trust in the comparison of recently diverged sequences than in those evolved in the distant past" [70]. One known shortcoming of the method is the rule "once a gap, always a gap", because gaps introduced into the early aligned sequences are not reevaluated after the inclusion of divergent information. This becomes inconvenient when aligning distantly related sequences, in which the placement of gaps is context-dependent. Iterative refinement approaches are used to circumvent this inconvenience.

Different algorithms have been developed to construct sequence alignments. The aim in all of them is to calculate the "true alignment", if any. They in fact search for the "optimal" alignment given a specific model, although nothing guarantees that the optimal alignment is the most biologically relevant. Any one of them is ideal, and the different programs for MSA calculation have been demonstrated to be more or less adequate depending on the dataset under study [72–77]. To maximize the chance of obtaining the most reliable alignment, it is therefore recommended to apply different algorithms to align the data and use the researche's expertise to extract the meaningful information from each calculated alignment [73, 77]. Of course, a visual evaluation can only be applied to relatively small sets of data. (In the case of large scale studies, this becomes unrealistic). Consistency based methods for alignment calculations turn out to be a practical alternative to incorporate information from different alignments in a single MSA [78].

Assessing alignment quality

The quality of an MSA is difficult to determine. Besides the SP and the WSP scores, another method that uses the sum of pairs is an accuracy metric for a multiple alignment relative to a reference alignment [79]. The "sum-of-pairs score" (SPS) described by Thompson and colleagues (1999) is based on the numbers of residue pairs identically aligned in the test and the reference alignment. The authors also introduced the "column score" (CS), that counts how many columns are identical in the reference and test alignments. More recently, Lassmann and Sonnhammer (2005) introduced the overlap score to assess alignment quality without the need of a reference alignment. The method employs inter-consistency to define both the accuracy of individual alignments by the multiple overlap score (MOS) and the difficulty of an alignment case by the average overlap score (AOS) [73].

1.3.3 Profile Hidden Markov Models

Profiles are statistical descriptions of the consensus of a multiple sequence alignment, with position-specific scoring models based on MSAs of gene or protein families. They capture important information about the degree of conservation at individual positions of the MSA, and the varying degree to which gaps and insertions are permitted. There are heuristic and probabilistic approaches for the application of profiles.

Hidden Markov models (HMM) have a formal probabilistic basis and provide a coherent theory for profile methods [80, 81]. HMMs are probabilistic models generally applicable to time series or linear sequences. In our case, linear amino acid sequences are the targets. An HMM describes a probability distribution over a potentially infinite number of sequences by recursive enumeration of possible sequences from a finite set of rules. These rules are represented by states, state transitions, and symbol emission probabilities.

Here it is important to introduce the terminology associated with HMMs. It is crucial to bear in mind the distinction between state sequences and symbol sequences. A state sequence is generated from the model and emits a symbol sequence with a given probability. In hidden Markov models, the state sequence is a first-order Markov chain where only the symbol sequence is observed, while the state sequence itself is hidden.

HMM Architecture A profile HMM describes a state path in which each consensus column of the multiple alignment is represented by a match state (M), and each match state has associated insert (I) and delete (D) states, forming what is called a "node" (M/D/I) at the same consensus position in the alignment.

States:

M $\longrightarrow$ *Match state – models the distribution of residues allowed in the column*

I $\longrightarrow$ *Insert state – allows insertion of one or more residues between match states (or between match and delete states)*

D $\longrightarrow$ *Delete state – allows deletion of the consensus residue*

In addition to match, insert and delete states, the profile HMM architecture also includes a begin (B) state and an end (E) state (dummy non-emitting states), that are not the same as the start state (S) and the terminal state (T). S and T refer to the query, while B and E refer to the model. Prefix (N) and suffix (C) states account for the flanking sequences, that are not part of the model.

Each match state is represented by a vector of emission probabilities for each nucleotide or amino acid residue (p_x). The states of the path are connected by state transition probabilities (t_{BM}, t_{MM}, t_{MI}, t_{MD}, t_{II}, t_{IM}, t_{DD},t_{DM}, t_{ME}).

Transitions:

t_{MM} $\longrightarrow$ *state transition probability for move from match state to match state*
t_{MI} $\longrightarrow$ *state transition probability for move from match state to insert state*
t_{MD} $\longrightarrow$ *state transition probability for move from match state to delete state (or "jump" a match state)*
t_{II} $\longrightarrow$ *self-transition probability of the insert states*
...
t_{BM} $\longrightarrow$ *"entry" transition*
t_{ME} $\longrightarrow$ *"exit" transition*

These parameters (probabilities) can be estimated by training the HMM from unaligned sequences. This option is, however, much harder than using pre-aligned sequences. Emission and transition probabilities are obtained by converting the observed counts of symbol emissions and state transitions into probabilities (additive log-odds scores) – $log(p_x/f_x)$ [80].

Residue-specific parameters:

p_x $\longrightarrow$ *probability of the match state emitting residue x (position-specific)*
f_x $\longrightarrow$ *expected background frequency of the residue x in the database*
$log(p_x/f_x)$ $\longrightarrow$ *score for residue x in this specific match state*

Alignment Modes Profile HMMs are models of global alignment: the generated sequences usually start at the first match state and end at the last match state. Nevertheless, models can be built to work as local alignment algorithms, which look for a high-scoring alignment between a sub-sequence of the target sequence and a part of the query model. The alignment mode is determined by the HMM configuration, not by the search algorithm. The model architecture differs when the profile is meant to find (perform) global or local alignments. In global alignments (with respect to the model), the begin state can only be followed by a match or delete state. Similarly, the end state can only be preceded by a match or delete state (that is very different than the local option!). Local alignments can further be distinguished as being local with respect to the model (profile) or to the sequence (subject). Local alignments with respect to the model incorporate a non-zero transition probability from a begin state to any internal match state, and from internal match states to the end state. The HMM architecture can also deal with multiple hits. More than one hit to the HMM per sequence is allowed by a cycle of non-zero transitions through a third special insert state (J - Joining segment unaligned sequence state) [80].

The most widely used software package for profile HMM analyses, HMMER, uses the "Plan 7" profile HMM architecture (Fig. 1.2), a fully probabilistic model of both local and global profile alignments [80]. The core section of the Plan 7 model is composed of M, D, and I nodes, flanked by B and E states. The other states (S,N,C,T,J) are "special states" that control algorithm dependent features of the model, e.g. how likely the model is to generate various sorts of local or multihit alignments. The main model has 7 transitions per node (one of the origins of the name Plan 7, according to Eddy [81]), t_{MM}, t_{MI}, t_{MD}, t_{IM}, t_{II}, t_{DM}, t_{DD}. There is no transition from delete to insert state or *vice-versa*. Plan 7 models one or more (possibly incomplete) matches to the domain model, and also models the unaligned sequence in the target.

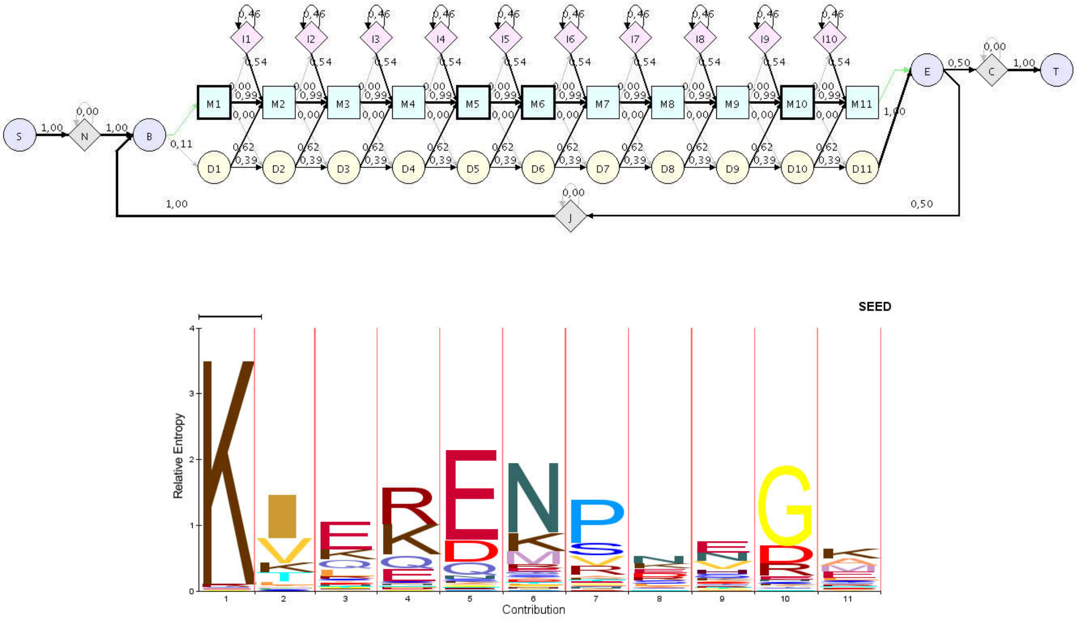


FIGURE 1.2: Illustrations of a profile HMM. Top: HMM architecture (Plan 7), figure created with HMMEditor [82]. Squares, diamonds and circles represent match, insert and delete states, respectively. Arrows represent transitions. Bottom: HMM Logo of the same profile, created using HMM Logo web server [83].

1.3.4 Phylogenetic Trees

The evolutionary relationship among sequences is typically presented as a phylogenetic tree. We may here generalize the terms to explain the structure of such a tree. The objects under study (molecular sequences, species, population, individuals, etc.) can be referred to as *operational taxonomic units* (OTUs). The leaves of the trees, or external nodes, represent the OTUs. The external nodes are connected by branches to internal nodes that represent their common ancestor. Internal nodes can also be referred to as *hypothetical taxonomic units* (HTUs). The branching order of the OTUs defines the tree topology. The length of the branches (in weighted trees) represents the amount of evolutionary change (observed or estimated) accumulated since the divergence from the common ancestor.

Considering the directionality of ancestry and descent, it is natural to think of rooted trees, in which the root is the last common ancestor of all OTUs. As a matter of fact, for almost the totality of the studied objects, the available information refers to contemporary (extant) data. An unrooted tree is, for this reason, the most reasonable representation of the relationship among them. The order of the events may therefore be estimated by introducing an external reference, commonly referred to as an *outgroup*, thus allowing the “rooting” of the tree.

The number of all possible tree topologies $T(s)$ for a set of s sequences is given by $\dfrac{(2s-5)!}{2^{s-3}(s-3)!}$ for unrooted trees and $\dfrac{(2s-3)!}{2^{s-2}(s-2)!}$ for rooted trees.

The easiest way to infer a phylogeny from molecular data is to calculate the pairwise distance between OTUs and compute a tree by hierarchical clustering. Among the hierarchical clustering methods, UPGMA (Unweighted Pair Group Method with Arithmetic Mean) [84] and neighbor-joining (NJ) [85] are the most widely used. The UPGMA method assumes that all OTUs evolved at the same rate, implying that the distance from the root to each external node is the same. The root of the tree is therefore defined as the point where the last two clusters connect. The neighbor-joining method, otherwise, necessarily produces an unrooted tree. The principle of the NJ method is to find pairs of sequences that minimize the total branch length at each stage of clustering, starting with a starlike tree.

The transformation of the qualitative differences in quantities (pairwise distances) causes a significative loss of phylogenetic information. Methods that can directly use qualitative criteria to infer phylogenies are considered to improve the chance of getting the true tree. Maximum parsimony (MP) [86] and maximum likelihood (ML) [87, 88] methods use completely different approaches to search for the best tree. The principle of maximum parsimony considers the best tree to be the one that explains the observed data by the smallest number of changes. ML methods search for the tree that maximises the likelihood of the observed data, given an evolutionary model. The debate about the merits of MP and ML is beyond the scope of this thesis.

It must be noted that MP and ML are NOT methods of tree construction. They are, in actual fact, principles used to decide which tree is the best among a set of trees. MP and ML principles do not determine how to find the topologies that will be evaluated (these must be given). If the number of sequences is relatively short (up to five), it would be possible to evaluate all alternative topologies. As the number of candidate topologies increases factorially with the number of sequences, the use of heuristics to generate trees remains as the reasonable solution. Programs implementing MP and ML methods usually rely on distance-based methods to create initial trees and apply different optimization methods (such as nearest-neighbor interchange, subtree pruning and regrafting, quartet-puzzling) to wander through the tree topology space. To prevent ML algorithms getting stuck at local optima, it is sometimes helpful to perform multiple optimization runs using random topologies as initial trees.

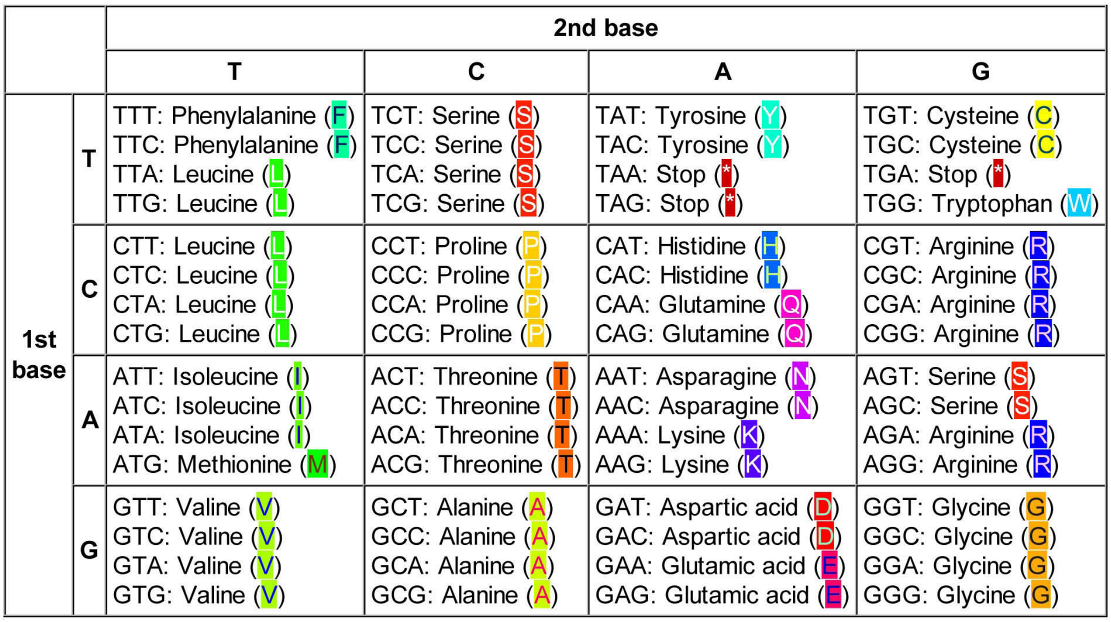
1.3.5 Codon Substitution Models

A classical approach to detect natural selection at molecular level is to compare observed and expected substitutions in a pair (or a group) of sequences [89, 90]. This requires the inference of substitutions accumulated over a phylogenetic tree, and of the expected number of substitutions if the mutations are fixed by chance.

Mutations occur in the DNA, and natural selection acts (theoretically) on the proteins, so that any approach should consider the particularities of the genetic code.

The informational unit of the genetic code is the “codon”. Three consecutive nucleotides in the DNA strand configure a codon, and each codon codifies a single amino acid. The reverse correlation, however, is not so easy. The genetic code is highly redundant. A single amino acid may be codified by up to six different codons in the standard genetic code (Table 1.1).

TABLE 1.1: Codon table for the standard genetic code. Source: [55]



		2nd base			
		T	C	A	G
1st base	T	TTT: Phenylalanine (F) TTC: Phenylalanine (F) TTA: Leucine (L) TTG: Leucine (L)	TCT: Serine (S) TCC: Serine (S) TCA: Serine (S) TCG: Serine (S)	TAT: Tyrosine (Y) TAC: Tyrosine (Y) TAA: Stop (*) TAG: Stop (*)	TGT: Cysteine (C) TGC: Cysteine (C) TGA: Stop (*) TGG: Tryptophan (W)
	C	CTT: Leucine (L) CTC: Leucine (L) CTA: Leucine (L) CTG: Leucine (L)	CCT: Proline (P) CCC: Proline (P) CCA: Proline (P) CCG: Proline (P)	CAT: Histidine (H) CAC: Histidine (H) CAA: Glutamine (Q) CAG: Glutamine (Q)	CGT: Arginine (R) CGC: Arginine (R) CGA: Arginine (R) CGG: Arginine (R)
	A	ATT: Isoleucine (I) ATC: Isoleucine (I) ATA: Isoleucine (I) ATG: Methionine (M)	ACT: Threonine (T) ACC: Threonine (T) ACA: Threonine (T) ACG: Threonine (T)	AAT: Asparagine (N) AAC: Asparagine (N) AAA: Lysine (K) AAG: Lysine (K)	AGT: Serine (S) AGC: Serine (S) AGA: Arginine (R) AGG: Arginine (R)
	G	GTT: Valine (V) GTC: Valine (V) GTA: Valine (V) GTG: Valine (V)	GCT: Alanine (A) GCC: Alanine (A) GCA: Alanine (A) GCG: Alanine (A)	GAT: Aspartic acid (D) GAC: Aspartic acid (D) GAA: Glutamic acid (E) GAG: Glutamic acid (E)	GGT: Glycine (G) GGC: Glycine (G) GGA: Glycine (G) GGG: Glycine (G)

Codons that codify for the same amino acid are said to be "synonymous", while codons codifying different amino acids are called "non-synonymous". If natural selection acts at the protein level, only non-synonymous mutations will be considered as targets for selection, while synonymous mutations will remain invisible.

1.3.5.1 Selection Inference

The ratio of relative fixation rates of non-synonymous and synonymous mutations is a recognized indicator of selection pressure and neutral evolution. Following this concept, equal rates of non-synonymous and synonymous substitutions are expected when mutations are neutral [40, 91]. A strong purifying selection, otherwise, will tend to eliminate every non-synonymous mutation. For the same reason, an excess of non-synonymous substitutions indicates positive selection [92–96].

Maximum likelihood approaches for the estimation of non-synonymous and synonymous rates of substitution were proposed by Muse and Gaut (1994) and Goldman and Yang (1994), and further adapted by several authors [99–108]. The ML methods use a continuous-time Markov chain, where the states are the 61 sense codons (in the standard genetic code), to model substitutions among codons.

The model of Goldman and Yang (1994) is based on an explicit model of codon substitution, described by a substitution rate matrix, $Q = \{q_{ij}\}$, where the elements q_{ij} represent the instantaneous substitution rate from codon i to j $(i \neq j)$. A simplified version of q_{ij} is given by

$$q_{ij} = \begin{cases} 0 & \text{if codons } i \text{ and } j \text{ differ in more than one position} \\ \pi_j & \text{for synonymous transversion} \\ \kappa\pi_j & \text{for synonymous transition} \\ \omega\pi_j & \text{for non-synonymous transversion} \\ \omega\kappa\pi_j & \text{for non-synonymous transition} \end{cases}$$

where κ is the transition/transversion rate ratio, ω is the non-synonymous/synonymous rate ratio ($\omega = d_N/d_S$), and π_j is the equilibrium frequency of codon j [99]. The matrix of transition probabilities over time, $P(t) = e^{Qt}$, is used in the log-likelihood calculation following Felsenstein (1981).

Improved models implement statistical distributions to account for heterogeneous ω ratios among sites (*site models*), and likelihood ratio tests (LRT) are employed to detect variation of ω ratios along lineages (*branch models*), while a Bayesian approach is used to identify positively selected amino acid sites [99–103].

1.3.6 Hypothesis Testing and Model Selection

The mode of gene evolution varies greatly amongst gene families, organisms and even populations, so that real datasets rarely fit perfectly to a model of sequence evolution represented in one of the several proposed substitution matrices. Apart from the matrices, a large number of other parameters can vary amongst the datasets. We thus seek a model that is best supported by the empirical data. For this reason, it is advisable to employ a model selection framework as a way to identify the "best-fit model" from a set of candidate models.

Two basic approaches dominate the scene in molecular evolution. The most traditional method is null hypothesis testing, where an alternative hypothesis is accepted when it is significantly better than the null hypothesis according to a test statistic. In molecular evolution, for instance, the preferred method is the likelihood ratio test. This method requires the models being compared to be nested, when the alternative hypothesis is a more complex version of the null hypothesis. More recently, model selection criteria such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and their derivatives have begun to dominate the field of phylogenetic analysis [109]. The use of model selection criteria enables the simultaneous test of several candidate models, that must not be nested.

The likelihood function (L) measures the fit of a model of sequence evolution (M) to a data set (D), given a tree (T) and branch lengths (B):

$$L = P(D|M, T, B)$$

The statistics of model selection can be applied to select among models, varying one or more of these parameters (M, T and B). Because the maximum likelihood of a tree typically assumes very small values, it is usually presented in the logarithmic form (lnL).

Likelihood Ratio Test The LRT is applied to test whether selection models provide a better fit to the data than does a null model [110]. The likelihood ratio (Δ) is given by the difference between the log-likelihood of the selection model (lnL_S) and the log-likelihood of the null model (lnL_{Null}):

$$\Delta = lnL_S - lnL_{Null}$$

For nested models, twice the log of this likelihood ratio statistic follows a χ^2 distribution, with the degree of freedom (df) being the number of extra parameters in the selection model.

$$\chi^2 = -2\Delta$$

Akaike Information Criterion A convenient way to choose among different models with different numbers of parameters is to make use of “information criteria”. Akaike (1973) defined an information criterion to estimate the expected distance of a given model from the unknown true model [111].

Being K the number of parameters, AIC is calculated by

$$AIC = -2\, lnL + 2\, K$$

The model yielding the smallest value of AIC is considered to better approximate the model that generated the data.

Second-Order Akaike Sugira (1978) included the size of the sample (n) in the denominator of a penalty term for AIC.

$$AICc = AIC + \frac{2K(K+1)}{n-K-1}$$

The second-order AIC or corrected AIC (AICc) is recommended when the sample size is small compared to the number of parameters ($n/K < 40$), but it can be applied equally well in larger sample sizes, because the penalty term tends to zero when the sample increases and AICc converges to AIC.

Bayesian Information Criterion Schwarz (1978) proposed an alternative information criterion that approximates to a Bayes estimator [112]. BIC differs from AIC by multiplying K by $log(n)$ rather than by 2.

$$BIC = -2\, lnL + K\, log(n)$$

(lnL: log-likelihood of the model, K: number of parameters, n: size of the sample)

Chapter 2

Methodological Framework

Many strategies exist to exploit the phylogenetic context of multigenic protein families. There is, however, no “magical formula” capable of dealing with all evolutionary patterns of genes and genomes. The classical examples of protein families present the genes as well defined paralogous and orthologous, in which each subtree reflects a species tree. This is far from what one can observe in the phylogenetic trees constructed for PR gene families.

Pathogenesis-related proteins are members of multigenic families encoded by genes that present an extremely variable number of copies in different species. The collection of sequences that are currently available in the public sequence databases lead to this conclusion. This intrinsic feature coupled with the disparate availability of sequences in the public databases makes the assembly of a concise and reliable dataset for evolutionary analysis a difficult task. Moreover, automatic methods often lead to the inclusion of defective or misclassified sequences in the dataset and, consequently, produce biased or erroneous results [48]. Manual selection of sequences in the databases is otherwise inappropriately time consuming. A strategy is needed to improve the automatic retrieval of sequences, while the quality of the resulting dataset is assured.

This chapter describes a methodological framework developed to generate reliable datasets for a large scale phylogenetic analysis of the 17 PR-families known to date. Aiming to absorb the characteristics of specific datasets and increase the quality of the analysis, this framework incorporates several methods, algorithms and strategies from bioinformatics.

The following sections are divided into two parts. The first part describes the workflow in chronological order. The second part presents the programs and parameters used in the framework in greater detail.

2.1 Workflow

The proposed workflow is divided into two stages, concerning the SEED and the EXTENDED datasets. A small representative alignment (SEED) is defined for each PR-family and is used to build a profile HMM. This profile HMM is able to capture virtually all homologous protein sequences in a database search and is used to assemble an enriched dataset (EXTENDED) that will be used to characterize the phylogenetic context of the protein families (Fig. 2.1).

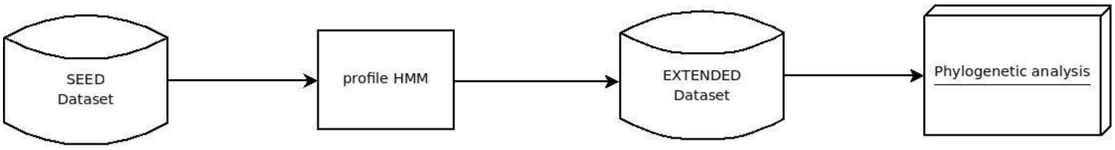


FIGURE 2.1: Workflow outline

2.1.1 SEED Dataset

The first step of the analysis is the construction of a SEED alignment to build a profile HMM for each PR-family. This step encompasses the selection of representative sequences, the definition of the region of interest (mature peptide), multiple sequence alignment, phylogenetic tree reconstruction and building of profile HMMs. i) Representative sequences are to be collected using previous knowledge about the protein families; ii) Sequences are aligned by an ensemble of multiple sequence alignment programs; iii) Alignments are visually and automatically evaluated, and one alignment is chosen as the SEED alignment; iv) A profile HMM is built from the SEED alignment. A tree can be constructed from the chosen alignment to complement/illustrate the data.

2.1.1.1 Data Selection

Representative sequences

Previous knowledge about PR-protein families is used for the collection of representative sequences. Reference sequences indicated in the literature [8, 19, 20] are used as queries in database searches. The annotation of the matches is manually examined for references about the source of the sequence. Preference is given to sequences obtained in experimental studies on plant defense against pathogenic attack. Experiments on other general features of PR-proteins are also considered as relevant for the inclusion of the sequences in the primary dataset. If this search is unsuccessful, sequences that present a reasonable annotation are preferred in the construction of the dataset.

Filtering database entries

The primary dataset must be filtered in order to increase the reliability of the data. The data file containing the collected database entries is processed by a Perl script to extract relevant information, and identify and exclude problematic sequences (i.e. fragments, redundant sequences, sequences with non-expected characters, etc.). Entries are also scanned for annotations about start and end positions of the mature peptide, signal peptide and pro-peptide (information not available for every entry). The presence of hinge regions and spacer sequences, sequence variations and conflicts can also be detected. The sequence regions annotated as mature peptide or catalytic domain are collected and stored for later use in the pre-alignment.

Pre-alignment

The set of full-length sequences and the set of collected catalytic domain and/or mature peptide sequences are aligned by fast alignment algorithms. The resulting alignments are manually revised and the full-length sequences compared to the set of domain regions. Sequence substrings corresponding to signal peptides and pro-peptides are removed by hand, and a new data file containing only sequences corresponding to the mature peptides is created (Fig. 2.2). This step also supports the identification of sequences containing large internal deletions and probable frameshift mutations that must be removed from the SEED dataset (see 2.2.4).

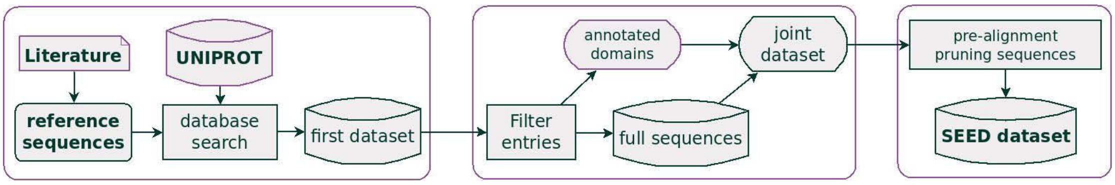


Figure 2.2: Workflow SEED dataset

2.1.1.2 SEED alignment

Alignment of the SEED sequences

The sequences corresponding to the mature peptides are aligned by an ensemble of multiple sequence alignment programs (see Table 2.1 for description).

Evaluation of Alignment Quality

All alternative alignments are evaluated comparatively. Two criteria are used to evaluate the resulting alignments: inter-consistency and visual evaluation. The collection of MSAs is first analyzed automatically by a program that uses the concept of inter-consistency to rank the individual alignments. In addition, the alignments are visually inspected for biologically relevant misalignments. The alignment best ranked by inter-consistency will be chosen as the SEED alignment, unless visual evaluation supports the choice of another MSA (Fig. 2.3).

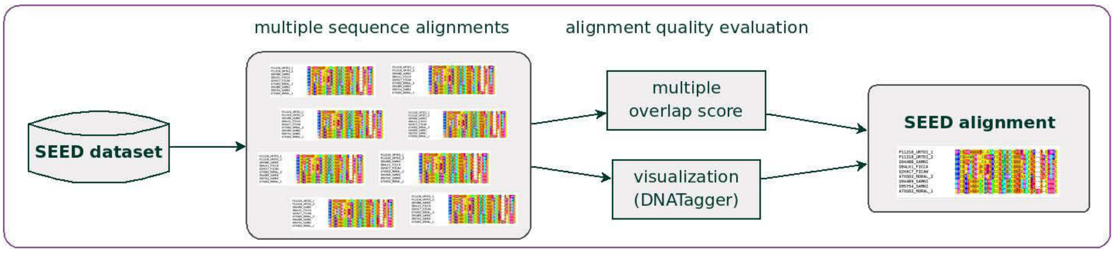

multiple sequence alignments
alignment quality evaluation
SEED dataset
multiple overlap score
visualization (DNATagger)
SEED alignment

Figure 2.3: SEED alignment workflow

2.1.2 Profile HMM

In this step we generate statistical models of multiple sequence alignments, collect sequences for an enriched dataset, select reliable sequences, and define the target region. i) The SEED alignment is used to build profile hidden Markov models, ii) The profile HMMs are used to find homologous sequences in a database; iii) Entries are filtered; iv) Selected sequences are aligned to the profile HMMs to define the location of the domain in the full-length sequences.

Building models

Four models are built for each PR-family:

global-single (gs) alignments are global with respect to the HMM and only one aligned region per target sequence contributes to the target's score, preventing over-scoring of multi-domain sequences.

global-multi (gm) allows multiple hits of the entire profile HMM. Useful to find multi-domain sequences.

local-single (ls) scores only one single local alignment per target sequence.

local-multi (lm) accounts for multiple local matches.

Using profile HMM in the search for homologous sequences

Each profile HMM is used independently to search a local copy of all plant entries from the UniProtKB. The lists containing the matches and the scores are individually analyzed. This manual curation is important to define the limits of the acceptable score values for matches. Matches with scores below these limits are excluded from the list. The entries of the remaining matches are retrieved from the UniProt web server and constitute the enriched dataset (Fig. 2.4).

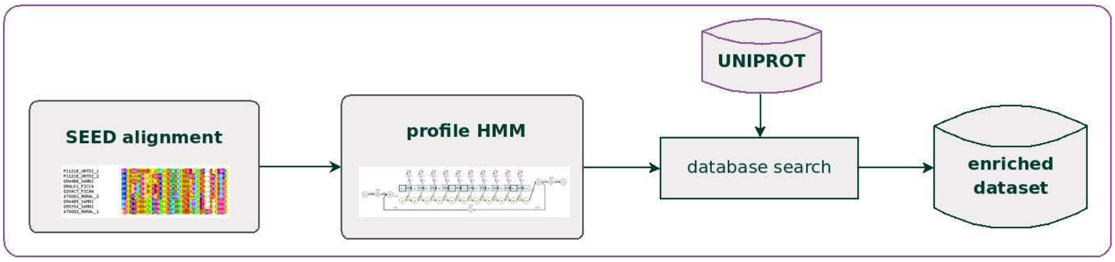


Figure 2.4: Using profile HMM in the database search

Filter entries

All entries are submitted to an annotation-based filter. Sequences matching at least one of the exclusion criteria are eliminated. The exclusion criteria are set according to dataset features. All sequences that pass the filter are considered "approved" and maintained in the dataset. Entries are filtered using a Perl script that reads annotation data and selects sequences to build the datasets. By default, the filter is set to exclude entries matching the following criteria:

- Evidence for the protein's existence is unavailable
- Sequence length is outside a predefined range
- Protein sequence presents ambiguous characters
- Sequence is redundant (the full-length sequence matches exactly with an already stored sequence)

The exclusion criteria can be adjusted according to the dataset. As output, the script produces a fasta file with the sequences, feature reports and Perl scripts to edit the sequence names in alignment and tree files.

Align sequences to profile HMM

The "approved" sequences are aligned to a global and a local profile HMM (*gs* and *lm*). These alignments allow the identification of the sequence regions that match the models. Sequences that do not match the profile HMM over the entire length of the model are excluded from the dataset, as well as sequences that present very large insertions or deletions. The sequence regions that are N-terminal and C-terminal to the model are deleted by a Perl script. The collected set of sequences constitutes the EXTENDED dataset (Fig. 2.5)

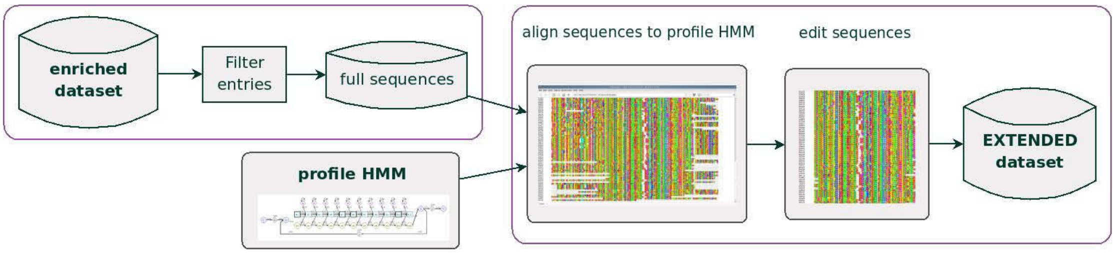


Figure 2.5: EXTENDED dataset workflow

2.1.3 EXTENDED Dataset

The large scale phylogenetic analysis needs an EXTENDED alignment containing a large number of homologous sequences.

In this step we calculate and evaluate multiple sequence alignments and use them to describe the phylogenetic context of each PR-family. i) Sequences are aligned by an ensemble of multiple sequence alignment programs; ii) Alignments are automatically evaluated; iii) One alignment is chosen as the EXTENDED alignment; iv) Coding sequences are retrieved and aligned 'to' the protein sequences; v) Coding sequences are aligned as nucleotide sequences and the MSAs are compared to the protein alignment.

2.1.3.1 Alignment

Multiple sequence alignment

The sequences corresponding to the profile HMM are aligned by an ensemble of multiple sequence alignment programs (see 2.1 for descriptions). Here, the profile HMM of the family is also used to construct an alignment that will be included in the MSA set.

Guide trees In specific cases, phylogenetic trees are inferred from the best intermediate alignment (inter-consistency criteria) using neighbor-joining (NJ) methods (see 2.2.7) and used as guide trees in a second alignment round.

Evaluation of alignment quality

In this step, the visual check was also performed, but only to detect extreme discrepancies in the selected alignment compared to the SEED alignment, considering important features of the family. As for the SEED dataset, the EXTENDED alignment is chosen from the list of those best ranked by inter-consistency, unless visual evaluation supports the choice of another MSA over the first in rank.

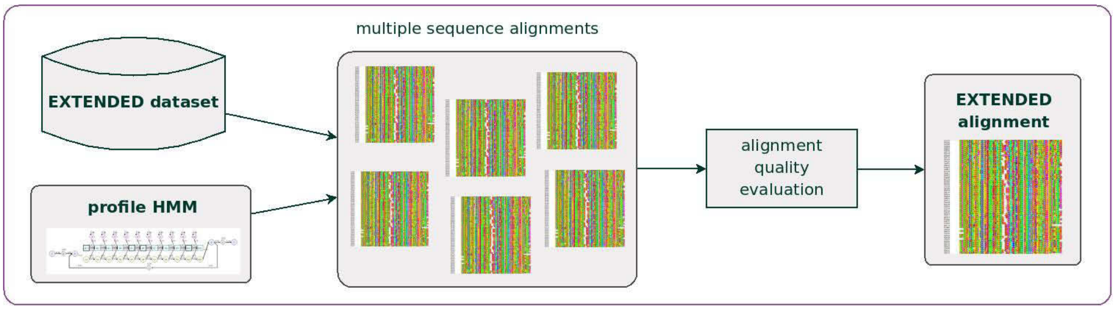


Figure 2.6: EXTENDED alignment workflow

2.1.4 Protein-coding sequences

Protein-coding DNA sequences are necessary to conduct the codon substitution model analysis. The protein-coding sequences of all datasets are retrieved from nucleotide databases and aligned following the amino acid based alignment.

Eliminate redundancy

More than one DNA sequence corresponding to the target amino acid sequences may be found. This is probable due to the redundancy of the nucleotide databases. Multiple matches for single target sequences are eliminated from the dataset with a Perl script.

Multiple sequence alignment

The protein-coding sequences are submitted to the alignment routine and aligned as nucleotide sequences irrespective of coding features. The alignments are manually compared to the coding-based alignment, and sequences with clear indications of frameshift mutations are excluded from the dataset.

2.1.5 Phylogenetic Analysis

The produced alignments are used for phylogenetic inference. i) Best-fit models of protein evolution and nucleotide substitutions are selected for the corresponding alignments; ii) Phylogenetic trees are inferred for both alignments (amino acid sequences and protein-coding DNA sequence alignments); iii) Evolutionary models are applied to analyze the sequences and trees.

Model selection for phylogenetic inference

Model selection is applied to define parameters for use in the calculation of phylogenetic trees. The choice of the most appropriate model among the set of candidate models is based on the ranking given by the second order Akaike information criterion (AICc).

Phylogenetic Tree Inference

Phylogenetic trees are inferred for the EXTENDED alignment of amino acids and nucleotide sequences using maximum likelihood (ML) methods. The parameters are set according to the best-fit models of protein and DNA evolution (Fig. 2.7).

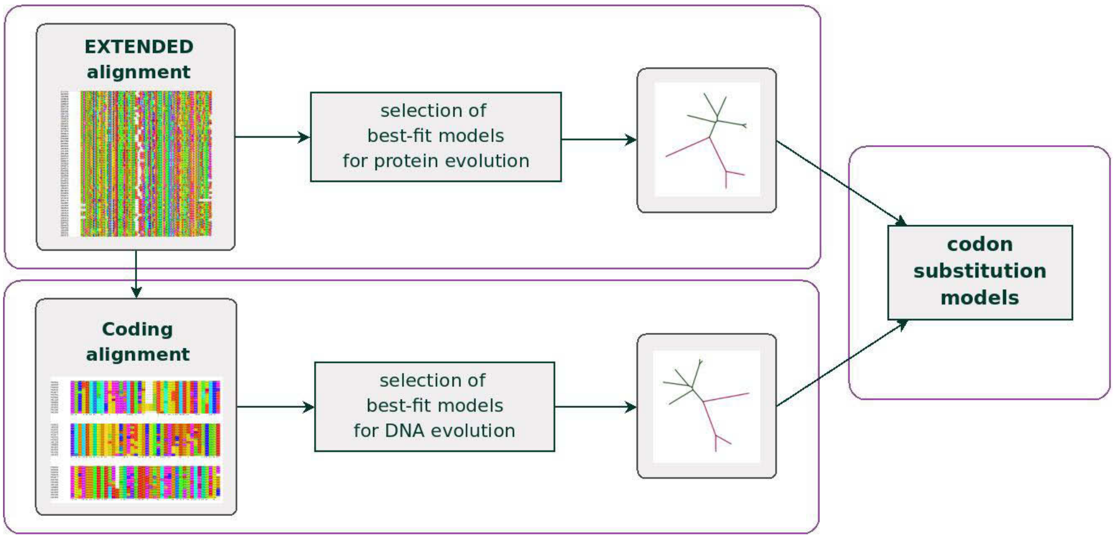

EXTENDED alignment
selection of best-fit models for protein evolution
Coding alignment
selection of best-fit models for DNA evolution
codon substitution models

Figure 2.7: Phylogenetic analysis workflow

2.1.6 Codon substitution models

Codon-based substitution models and maximum likelihood methods are used to identify genes that are likely to be evolving under positive selection pressure.

Selection inference

The alignments of the protein-coding DNA sequences are analyzed with codon substitution models. Two candidate tree topologies are given as user trees, one inferred from the amino acid alignment using models of protein evolution and the other inferred with DNA models (from the protein-coding sequences). Both trees are compared under the model M0 (one ratio) and the tree that best fits the data under this model is chosen for the application of more complex models. Two pairs of nested models of codon substitution are applied to infer positive selection with a likelihood ratio test (LRT). Models M1a (NearlyNeutral) and M2a (PositiveSelection), as well as models M7 (beta) and M8 (beta&ω), are compared using χ^2 with two degrees of freedom. Additionally, the model M3 (Discrete, 3 categories) is employed and applied in an LRT with M0 as a test of variable ω ratios among sites.

Posterior probabilities for site classes, calculated with Bayes empirical Bayes (BEB) method, are used to identify sites under positive selection if the LRT is significant.

2.2 Databases and Programs

2.2.1 Protein database

This study uses the UniProt Knowledgebase (UniProtKB) as the main source of data [44]. UniProtKB consists of two sections, referred to as the Swiss-Prot section and TrEMBL section.

All plant entries from the UniProtKB/Swiss-Prot and UniProtKB/TrEMBL datasets were downloaded in flat file format from the UniProt FTP directory (Copyright 2002-2009 UniProt Consortium), subdirectory /taxonomic_divisions.

The files uniprot_sprot_plants.dat (134 MB) and uniprot_trembl_plants.dat (1.8 GB) were merged into a single file. The complete UniProtKB dataset in flat file format reaches more than 20 gigabytes in size, being 1.9 GB from UniProtKB/Swiss-Prot (Release 57.4 of 16-Jun-2009) and 18.8 GB from UniProtKB/TrEMBL (Release 40.4 of 16-Jun-2009).

Annotation

Each UniProtKB entry contains an entry identifier, an accession number, a recommended name, date of creation and last modification, the taxonomic classification of the organism, bibliographical references, and the protein sequence. Additional annotation includes: names and origin, protein attributes, ontologies, alternative products, general annotation (comments), sequence annotation (features), sequences, references, cross-references, entry information and relevant documents [48].

The section "Protein Attributes" includes the "Protein Existence" (PE) line, which indicates the available evidence for the existence of a protein, classified in five levels:

Evidence at protein level (PE 1)

Supported by clear experimental evidence (e.g. unambiguous identification by mass spectrometry)

Evidence at transcript level (PE 2)

Supported by transcription data (e.g. cDNAs, RT-PCR, microarray, Northern blot)

Inferred from homology (PE 3)

Strong sequence similarity to known proteins in related species

Predicted (PE 4)

Without evidence at protein, transcript, or homology levels

Uncertain (PE 5)

Dubious sequences that may represent the translation of a pseudogene or an erroneously assigned ORF

The PE tag discriminates between proteins whose existence has been experimentally proven and those whose existence has been computationally inferred [44, 48]. It does not constitute *per se* a measure of the correctness of the sequence. The stored sequence itself may contain errors, especially for sequences derived from gene predictions from genomic sequences, but also well characterized sequences, as in an example found during this study (see note on page 90, Chapter 4).

In the present work, the information displayed in the PE line is used as exclusion criteria rather than for validation. The same is true for the DE (DEscription) lines containing non-experimental qualifiers “probable” and “by similarity”. Other exclusion criteria used here are the terms “hypothetical”, “putative”, “fragment”, “partial”, “undetermined”, “homology to unknown gene”, “unknown” and “scaffold”, appearing in the DE line.

2.2.2 Multiple sequence alignment

Assessment of alignment consistency The inter-consistency of a collection of multiple sequence alignments is assessed with MUMSA [73]. MUMSA uses the concept of “pairs of aligned residues” to compare a set of multiple sequence alignments and calculate the average overlap score (AOS) and multiple overlap score (MOS). The AOS is a good prediction of alignment difficulty, while the MOS is an efficient measure to assess the accuracy of individual alignments [73]. The AOS is calculated over all alignments and the individual MSAs are ranked according to the MOS.

Alignment programs MSAs were generated using the following programs:

Kalign (Fast and accurate multiple sequence alignment algorithm)
Muscle (Multiple sequence comparison by log-expectation)
ClustalW (Multiple alignment of nucleic acid and protein sequences)
MAFFT (Multiple sequence alignment based on Fast Fourier Transform)
DIALIGN-T (Improved algorithm for segment-based multiple sequence alignment)
DIALIGN-TX (Greedy and progressive segment-based multiple sequence alignment)
POA (Partial Order Alignment)
PRANK (Probabilistic Alignment Kit)
ProbCons (Probabilistic Consistency-based multiple sequence alignment)
T-Coffee (Tree-based consistency objective function for alignment evaluation)
DCA (Divide-and-Conquer Alignment)

Options and versions of the MSA programs are listed in Table 2.1.

Table 2.1: List of programs for multiple sequence alignment used in the workflow

Program	options	obs	version	references
DCA		max. 42 sequences	1.1	[67, 113]
Kalign			2.03	[114, 115]
Kalign_a2	-gpo 5.0 -gpe 5.0 [a]		2.03	[115]
Kalign_b2	-matrix_bonus 5.2	all-positive matrix	2.03	[115]
Muscle			3.6	[116, 117]
ClustalW2	-quicktree	faster	2.0.10	[118, 119]
ClustalW2	-iteration=tree	slower	2.0.10	[119]
MAFFT	L-INS-i		6.240	[120, 121]
DIALIGN-T		greedy algorithm	0.2.2	[122]
DIALIGN-TX		greedy + progressive	1.0.2	[123]
POA global	-do_global		2.0	[124]
POA progressive	-do_progressive		2.0	[71]
PRANK			080709	[125]
PRANK_F	+F	‘permanent’ insertions	080709	[126]
ProbCons			1.12	[127]
T-Coffee			5.31	[128]

[a] Gap open penalty (-gop); Gap extension penalty (-gep)

The SEED dataset is aligned with Kalign (default, a2 and b2), Muscle, Clustal W2 (-iteration=tree), MAFFT L-INSi, POA progressive, POA global, ProbCons, DIALIGN-TX, DIALIGN-T, DCA, T-COFFEE, PRANK_F and PRANK. The set of alternative alignments are compared by inter-consistency evaluation with MUMSA.

The alignment routine of the EXTENDED dataset is achieved in six steps. i) alignment with Kalign (default, a2 and b2), Muscle, Clustal W2 (-quicktree), MAFFT L-INSi, POA progressive, POA global, ProbCons, DIALIGN-TX, DIALIGN-T, and PRANK_F; ii) automatic selection of one representative alignment by inter-consistency assessment with MUMSA; iii) calculation of an NJ tree with PHYLIP (*protdist* and *neighbor*); iv) alignment with PRANK, PRANK_F and Clustal W2 (-ITERATION=alignment) using the NJ tree as guide tree; v) alignment to profile HMM (lm); vi) inter-consistency assessment with MUMSA using all generated alignments.

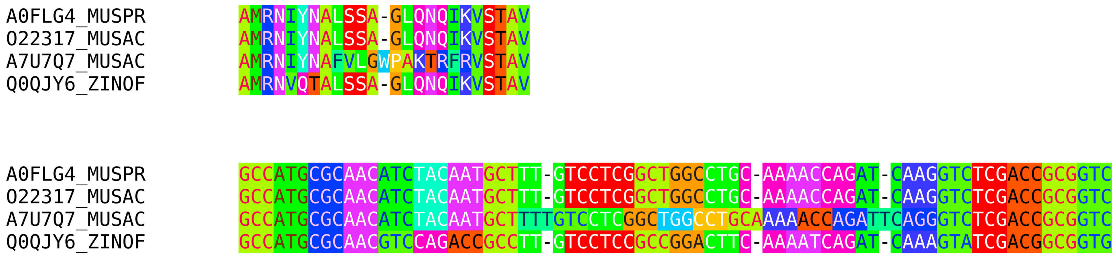
2.2.3 Protein-coding sequences

The protein-coding sequences of all datasets are retrieved using PROTOGENE (PROtein TO GENE) at the T-Coffee server (`www.tcoffee.org`) [129, 130]. PROTOGENE searches nucleotide databases in order to identify the transcript or genomic sequences most likely to be associated with the original protein sequences, and turns the protein multiple sequence alignment into the associated CDS (CoDing Sequence) MSA [129].

2.2.4 Finding Frameshift Mutations

The observed reading-frame alterations can be either natural mutational events registered in the DNA, or (unfortunately frequently) mere products of technical artifacts arising in the reading of sequenced DNA/RNA. Naturally-occurring frameshift mutations are normally short and occur in regions that do not strongly affect the protein's stability and function. If this occurs, the newly formed codons may code for similar amino acids. Coding sequences with frameshifts are not indicated for analysis by codon substitution models, because the models suppose only point mutations. The reason is that the nucleotides forming the codon triplets on the "reverse-translated" amino acid-based alignment are no longer really homologous. The inclusion of these "artificially aligned" codons will not represent the actual history of the sites, and it is not compatible with the codon models used here. Sequencing errors may in turn produce "noisy" results in sequence alignments, and can easily be identified by a trained researcher's eyes. In both cases, coding sequences containing differences in the reading frame should be excluded from the dataset. The technique used here to identify reading-frame alterations is simple, but requires manual and visual interference on the part of the researcher.

DNATagger The coding sequences are aligned as nucleotide sequences and visualized with DNATagger [55]. This program colors coding sequences relative to the coded amino acids, irrespective of the positions of gaps. Frameshift mutations and reading-frame errors can be easily recognized as demonstrated in Figure 2.8.



```
A0FLG4_MUSPR          AMRNIYNALSSA-GLQNQIKVSTAV
O22317_MUSAC          AMRNIYNALSSA-GLQNQIKVSTAV
A7U7Q7_MUSAC          AMRNIYNAFVLGWPAKTRFRVSTAV
Q0QJY6_ZINOF          AMRNVQTALSSA-GLQNQIKVSTAV


A0FLG4_MUSPR          GCCATGCGCAACATCTACAATGCTTT-GTCCTCGGCTGGCCTGC-AAAACCAGAT-CAAGGTCTCGACCGCGGTC
O22317_MUSAC          GCCATGCGCAACATCTACAATGCTTT-GTCCTCGGCTGGCCTGC-AAAACCAGAT-CAAGGTCTCGACCGCGGTC
A7U7Q7_MUSAC          GCCATGCGCAACATCTACAATGCTTTTGTCCTCGGCTGGCCTGCAAAAACCAGATTCAGGGTCTCGACCGCGGTC
Q0QJY6_ZINOF          GCCATGCGCAACGTCCAGACCGCCTT-GTCCTCCGCCGGACTTC-AAAATCAGAT-CAAAGTATCGACGGCGGTG
```

FIGURE 2.8: Effects of a frameshift mutation in a protein and a nucleotide alignment.

2.2.5 Profile Hidden Markov Models

The software package HMMER (profile hidden Markov models for biological sequence analysis), Version 2.3.2 (Oct 2003), is used to build profile HMMs and to perform database searches and sequence alignments [80, 81]. HMMER uses the "Plan 7" profile HMM architecture, a fully probabilistic model of both local and global profile alignments.

Programs from the HMMER package used here:

hmmbuild builds a profile HMM from a multiple sequence alignment. By default, *hmmbuild* builds models for the finding of multiple non-overlapping alignments to the complete model (global-multi). In this analysis, preference is given to the model that finds a single global alignment to a target sequence (-g option, global-single). Alternatively, local models can be built, accounting for a single local alignment per sequence (option -s, local-single), or even multiple domains per target sequence (-f option, local-multi). Local profile HMMs are useful to find conserved domains in unrelated sequences.

hmmcalibrate calibrates HMM search statistics. *hmmcalibrate* scores a large number of synthesized random sequences with a profile HMM and uses the extreme value distribution (EVD) to calibrate the model.

hmmsearch searches a sequence database with a profile HMM. The program reads a profile HMM and searches a database for significantly similar sequence matches. The results are presented in a ranked list of the best scoring sequences. The scores are relative to the size of the database.

hmmalign aligns sequences to a profile HMM. The alignments are saved in Stockholm format.

Box: Considerations about Stockholm format output

The Stockholm format, used as default output by HMMER, was developed by the Pfam Consortium to support extensible markup of multiple sequence alignments. The syntax adopted by HMMER is useful to highlight the model states in the alignment. The structure of the alignment in reference to the model is represented in markup lines. The columns which were assigned to match states will be marked with x's, and columns containing insert states will be marked with dots in the reference line (#=GC RF) annotation. Other types of Stockholm markup annotation are described in the HMMER User's Guide [81]. In the alignment lines, uppercase (capital) letters are used to indicate emissions on match states, while lowercase symbols are used for insert states. Dashes (-) are used for deletions inside match state columns, and dots (.) fill the insert state columns (or non-emission of character states).

Match states – uppercase (capital) letters

Insert states – lowercase letters or dots (.)

Delete states – dashes (-) for deletions inside match state columns

It is worth noting that characters within insert states are not aligned. Alignments produced by hmmalign must be used with care, with eventual excision of the insert states. This property is actually useful to retrieve the sequences corresponding to the domain of interest from the entire sequences.

2.2.6 Model Selection

The program ProtTest (version 2.1) [131] is used to select among models of protein evolution and jModelTest (version 0.1.1) [132] is for best-fit models of nucleotide substitution. Both programs use PhyML [133] to calculate the models.

Model selection for amino acid sequences ProtTest computes the likelihood of each candidate model of protein evolution and estimates the fit of all candidate models using Akaike and Bayesian information criteria (AIC, AICc and BIC). The general matrices of amino acid substitution included in ProtTest are: Dayhoff, JTT, WAG, VT, DCMut, LG, Blosum62, MtREV, MtMam, MtArt, RtREV, and CpREV. Each model can also be combined with parameters of proportion of invariable sites (I), gamma rate distribution (G) and amino acid frequencies observed from data (F). ProtTest uses the

PhyML program for the computation of likelihoods and the estimation of parameters, given an alignment and a tree (user-provided or calculated with BIONJ algorithm [134]).

Model selection for nucleotide sequences jModelTest is similar to ProtTest, given that it also takes advantage of the PhyML program for likelihood calculations, including model parameters and tree estimates. The program implements 11 nucleotide substitution schemes (JC, HKY, TN, TPM1, TPM2, TPM3, TIM1, TIM2, TIM3, TVM and GTR), with equal/unequal base frequencies (+F), proportion of invariable sites (+I), and gamma rate variation among sites (+G) (4 rate categories by default).

2.2.7 Phylogenetic Trees

The calculation of phylogenetic trees has been incorporated in different steps of the described workflow. In the first part of the analysis it is optional to include a subroutine for tree reconstruction in the multiple sequence alignment step. The use of a guide tree can reduce the runtime of several MSA programs, but it may introduce a bias in the resulting alignment. A user-provided guide tree can be useful when running the same program multiple times with different parameters. A tree is also constructed with the SEED alignment to provide an illustration of sequence diversity and distribution.

In the EXTENDED analysis, phylogenetic trees play roles as results (*per se*) and as intermediate data for evolutionary models of sequence evolution. Here again, intermediate guide trees can reduce the computational time of multiple sequence alignments of large datasets. They can also be given as initial trees in ML phylogenetic programs for further topology optimization.

A reliable tree topology is needed to estimate the parameters of codon-based models. A codon-based phylogeny would be desirable in this case. For large datasets there is, however, no feasible way to infer a phylogeny under a codon model yet. To execute codon model analysis, it is advisable to compare several candidate trees inferred under DNA or AA models [135].

Programs implementing neighbor-joining and maximum likelihood algorithms have been incorporated in the workflow. All programs are freely available and run under Linux.

Neighbor-joining

PHYLIP (PHYLogeny Inference Package) Perhaps the most widely-distributed phylogeny package, PHYLIP consists of 35 programs, implementing a variety of methods for phylogenetic analysis [136]. PHYLIP is used here to construct guide trees in the alignment routine. A distance matrix is calculated from aligned protein sequences by *protdist*, and is used by the program *neighbor* to construct a tree with the neighbor-joining method of Saitou and Nei (1987) [85, 136].

Maximum Likelihood

PhyML A very fast and accurate maximum likelihood program for nucleotide or protein sequence data [133]. The latest version (v.3.0) offers three choices of tree topology improvement: "nearest neighbor interchange" (NNI), "subtree pruning and regrafting" (SPR), and BEST, where both SPR and NNI methods are used and the best tree is chosen. The first method is the fastest and the last is computationally very intensive. In this study, the BEST option is preferred.

IQPNNI An efficient tree reconstruction method that implements the "important quartet puzzling" (IQP) algorithm and NNI optimization to reconstruct phylogenetic trees based on DNA or amino acid sequence data [137]. It also implements codon models for tree evaluation and is able to detect sites evolving under positive selection. IQPNNI is well suitable for small and medium size datasets, but becomes unfeasibly time-consuming for large datasets.

TREE-PUZZLE More than a tree reconstruction method, TREE-PUZZLE is a program package for quartet-based maximum likelihood phylogenetic analysis able to compare and test trees and models on DNA and protein sequences [138]. The tree reconstruction takes, however, much longer than other methods, and the resulting tree has often unresolved branches (multifurcation)[1]. Therefore, TREE-PUZZLE is used here only to compare trees constructed by the other methods.

[1] In my opinion, multifurcations on a phylogenetic tree are not as bad as they are claimed. They draw our attention to the uncertainty of tree topology with the very short internal branches, probably due to small divergence times. These branches show typically low support in bootstrap analysis.

2.2.8 Codon Substitution Models

The alignments of the protein-coding DNA sequences are analyzed with the PAML 4 software package [139], using the *site models* implemented in *codeml* [100, 102, 140]. The analysis encompasses a test for the presence of sites under positive selection using LRT to compare nested models, the estimation of the proportion of positively selected sites and the strength of selection by ML estimation of parameters in the ω distribution, and subsequently identification of sites under positive selection by Bayes prediction [141].

2.2.8.1 Number of free parameters in the ω distribution

The simplest site model, M0 (one ratio), has only ω as a free parameter. M1a (NearlyNeutral) has two categories in the ω distribution, and two free parameters. ω_1 is fixed in 1, while $\omega_0 < 1$, and p_0 is the proportion of sites in ω_0 (and $p_1 = 1 - p_0$). M2a (PositiveSelection) has one extra category and, therefore, two additional parameters, p_2 and $\omega_2 > 1$. The number of free parameters in M3 (Discrete) depends on the number of categories of ω ratios defined for the analysis. The default option (3 categories) has three free parameters for ω values (ω_0, ω_1 and ω_2), and two free parameters for the proportion of sites in each category (p_0, p_1, $p_2 = 1 - p_0 - p1$). For models M7 (beta) and M8 (beta&ω) the number of site classes for the beta distribution does not change the number of free parameters. The beta distribution is described by p and q. In M8, an extra class of ω accounts for sites outside the beta distribution ($\omega_s > 1$, $p_1 = 1 - p_0$). M8a fixes $\omega_s = 1$, and is used as the null model for M8 in a special LRT [142].

2.2.8.2 Further variable parameters

In addition to the ω distribution parameters, some other parameters can be estimated or fixed to a predefined value. Here, κ (transition/transversion rate ratio) and branch lengths are usually optimized from initial values[2]. Because of the computational burden, analyses with models M7 and M8 are performed with fixed branch lengths, obtained with less complex models.

[2] **obs:** Some analyses with free branch lengths can take days, or even weeks, to complete. In such cases, it is very advisable to set the control variable "noisy" to a higher value (here, 3) in the file `codeml.ctl` to force *codeml* to print the current parameter values in the file `rub` during the iteration process. These parameters can be given as initial values in a new run. The values recorded after the symbol "`x:`" in the last line of the `rub` file must be pasted into a new file named `in.codeml` [142].

Chapter 3

General Results

3.1 SEED Dataset

3.1.1 Selection of representative sequences

The homepage `www.bio.uu.nl/~fytopath/PR-families.htm`, maintained by the Phytopathology group of Utrecht University, stores the general and specific references for all recognized families of pathogenesis-related proteins to date. Based on this information and the reference literature indicated by Van Loon and Van Strien (1999), Van Loon, Rep and Pieterse (2006), representative members for each PR-family have been defined [8, 19, 20].

Each representative member has been used as a query to search the UniProtKB database. Among the matched entries, sequences representing PRs that had been experimentally studied in a context of induced defense response were revised and collected to complete the first dataset.

The representativity of the different PR-families in UniProtKB is largely variable. Families PR-2, PR-3 and PR-9 are abundantly represented in the database and many sequences are directly related to experimental studies. PR-7 has many homologous sequences, but most of them originated from whole genome sequencing and few sequences are properly annotated; the only related experimental studies are derived from the same organism. PR-17, otherwise, was recently described as PR and the available annotation has not been updated by the authors, but the related literature supported the choice of the sequences. The criteria applied in the selection of representatives for each PR-family takes this variability into account.

3.1.2 Finding the boundaries of the mature peptides

The retrieved sequences correspond to the precursor peptides, which frequently contain sequence regions that are removed by post-translational processing. The annotation provided with the sequences about the position of signal peptides, pro-peptides and hinge regions, as well as the coordinates of the chain itself, were used to define the location of the mature peptide in the precursor sequences.

Except for PR-10, all PR-families presented signal peptides at the N-termini and eight families had pro-peptides described for some members (Table 3.1). Sequence features for PR-7 and PR-11 were not available in the annotation. From the literature, it is known that P69 (a PR-7) precursor has an N-terminal signal peptide (22 residues) followed by a 92-amino acid pro-peptide and the 631-amino acid mature peptide [143]. The signal peptides found in PR-11 sequences could be identified by visual inspection of the alignment and their identification was confirmed by SignalP prediction [144]. C-terminal pro-peptides of variable length were found in PR-1, PR-2, PR-3, PR-4 and PR-5 sequences, while in some PR-6 members, a potential pro-peptide was located right after the N-terminal signal peptide. Some members of the chitinase families PR-3 and PR-4 present one or more chitin-binding domains (CBD), and these domains were also removed from the sequences. On the other hand, the two domains that characterize PR-13 were not separated in the analysis.

Because not all sequences in the datasets were annotated for these features, the collected chains were aligned to all sequences and the flanking regions were manually removed from alignment. The set of edited sequences will hereafter be called the SEED dataset.

3.1.3 Generation of alternative multiple sequence alignments

Each SEED dataset was submitted to a set of alignment programs. Kalign was the fastest among all programs and provided reliable alignments for most of the datasets. It was useful to test alternative parameters for gap penalties and substitution matrices. DCA, on the other hand, could not be applied to all datasets. The largest number

Table 3.1: SEED Dataset

PR-family	number of sequences	length (precursor)	signal peptide	pro-peptide	length (chain)
PR-1	20	157 - 179	21 - 30	20	135 - 155
PR-2	19	331 - 372	22 - 32	20 - 23	306 - 340
PR-3	37	243 - 415	16 - 33	7 - 15	229 - 261
PR-4	17	125 - 231	17 - 25	6	120 - 138
PR-5	15	169 - 251	20 - 26	6	148 - 230
PR-6	16	65 - 111	12 - 23	14 - 17	65 - 92
PR-7	10	666 - 754	22 - 28	87 - 92	631
PR-8	9	291 - 302	22 - 30		267 - 273
PR-9	18	312 - 364	22 - 45		290 - 332
PR-10	11	155 - 160	∅		155
PR-11	10	327 - 398	16 - 35		321 - 366
PR-12	9	72 - 105	25 - 29	27 - 33	45 - 51
PR-13	5	133 - 137	24 - 28		109 - 110 [a]
PR-14	7	114 - 132	24 - 27		90 - 105
PR-15	13	201 - 229	22 - 31		198 - 201
PR-16	9	217 - 229	20 - 27		193 - 206
PR-17	11	223 - 641	20 - 26		203 - 209

[a] Two domains

of sequences DCA was able to align was 42, but even smaller datasets caused DCA to run out of memory if relatively divergent sequences were included in the dataset. Comparatively, the most time-consuming programs were PRANK, ProbCons and T-Coffee. Even so, the worst mark was a well feasible five minutes, required by PRANK to align the PR-7 dataset.

With the exception of PR-13, all SEED datasets were reasonably well aligned. For PR-13 the best results were obtained aligning each domain separately and concatenating the selected alignments.

3.1.4 Choosing the SEED alignment

Assessment of alignment inter-consistency The inter-consistency among the alternative alignments was assessed with MUMSA. The average overlap score (AOS) ranged from 0.76 to 0.92, indicating that the sequences are relatively well "alignable". ProbCons and MAFFT L-INS-i most frequently reached the highest multiple overlap score (MOS). POA, DIALIGN-T and DIALIGN-TX were most frequently among the lowest ranked.

Visual evaluation of alignment quality All alternative alignments were visualized with DNATagger. Particular features of each family, such as the position of cysteine residues involved in disulfide bridges, were considered when comparing the alignments. Preferentially, the alignment with highest MOS was chosen to be the SEED alignment. In a few cases, different algorithms produced the same systematic errors, resulting in higher MOS values for incorrectly aligned sequences. In these cases, the alternative alignment that best solved the question was selected as the SEED alignment. If necessary, the selected alignment was manually edited to correct eventually misplaced gaps.

Table 3.2 summarizes the alignment quality evaluation of the SEED dataset and the choice of the SEED alignment.

Table 3.2: Inter-consistency scores for SEED alignments

		first ranked MSA		SEED alignment	
PR-family	AOS	algorithm	MOS	algorithm	MOS
PR-1	0.91	Kalign_b2	0.96		
PR-2	0.89	MAFFT L-INS-i	0.97	ProbCons	0.96
PR-3	0.78	MAFFT L-INS-i	0.86		
PR-4	0.92	ProbCons	0.99	(edited)	0.97
PR-5	0.76	ProbCons	0.85		
PR-6	0.83	ProbCons	0.91	MAFFT L-INS-i	0.91
PR-7	0.92	Kalign	0.99	PRANK_F (edited)	0.98
PR-8	0.90	ProbCons	0.97		
PR-9	0.86	MAFFT L-INS-i	0.94	(edited)	
PR-10	0.89	MAFFT L-INS-i	0.97	PRANK	0.96
PR-11	0.84	ProbCons	0.92	PRANK	0.89
PR-12	0.85	ProbCons	0.92	DIALIGN-TX	0.92
PR-13	0.76	ProbCons	0.85	Kalign + DIALIGN-TX	0.75 [a]
PR-14	0.85	Muscle	0.94		
PR-15	0.93	ProbCons	1.00		
PR-16	0.88	Muscle	0.95	(edited)	0.95
PR-17	0.90	MAFFT L-INS-i	0.98	DCA	0.97

[a] Each chain was aligned independently, joined and compared to the set of full alignments.

3.2 Profile HMM

3.2.1 Building models

The SEED alignments were used to build statistical models of multiple sequence alignments (profile HMM) with HMMER [80]. Profile HMMs were built with *hmmbuild* and empirically calibrated with *hmmcalibrate* to increase the sensitivity of the database search.

3.2.2 Database search

The profile HMM built from the SEED alignment is used to search the UniProtKB database for homologous proteins with the *hmmsearch* program from the HMMER package [80]. The first search is performed on a local copy of the plant dataset from UniProtKB [48, 53]. The per-sequence E-value cutoff is set to 10.0, while the bit score cutoff is set to zero.

An alternative search for homologous sequences in other phyla (complete UniProtKB) is performed through the web server Mobyle Portal [145]. The threshold is defined with E-value cutoff set at 0.1, while the bit score is allowed to negative infinity (default settings). (By default, the threshold is controlled by E-value and not by bit score.)

Filtering search results

In four PR-families, sequences derived from large sequencing for characterization of protein isoforms in crop plants account for excessive redundancy in the extended dataset. A convenient solution prevents this data from inflating the set. The dataset was restricted to entries labeled PE1 and PE2 (“Evidence at protein level” and “Evidence at transcript level”, respectively). By default, the required PE level was 3 (“Inferred from homology”).

In the extreme opposite case, PR-families that were poorly represented in the UniProtKB database were submitted to a less stringent filter, that did not exclude entries based on their description or PE labels. It only restricted entries by sequence size (lower and upper values defined according to previous knowledge on the protein size distribution).

Table 3.3 lists the quantity of whole sequence hits matched by different profile HMMs in the search for homologous sequences in UniProtKB – plant dataset, and the numbers of sequences that remained after the application of the filter.

Table 3.3: Number of matches found by different profile HMMs in UniProtKB

	number of whole sequence hits				selected	
PR-family	global-single	global-multi	local-single	local-multi	filter	set
PR-1	368	368	396	398	PE3	65
PR-2	899	898	1012	1012	PE2	158
PR-3	702	702	813	813	PE3	181
PR-4	86	86	98	98	PE3	26
PR-5	652	652	887	888	PE3	142
PR-6	237	237	240	241	PE3	26
PR-7	599	599	824	825	PE3	52
PR-8	221	221	268	268	PE3	89
PR-9	1412	1411	1971	1973	PE2	330
PR-10	665	665	754	754	PE2	231
PR-11	57	57	95	96	length	31
PR-12	280	280	281	281	PE3	79
PR-13	91	90	114	114	PE3	62
PR-14	481	481	591	592	PE2	102
PR-15	501	500	573	573	PE3	160
PR-16	513	513	576	578	PE3	160
PR-17	51	51	60	60	length	33

3.2.3 Aligning sequences to the profile HMM

The sequences approved by the filter were aligned to each one of the four profile HMMs. The alignments indicated the regions of the sequences that matched the profiles. The visualization of the alignments in DNATagger permitted a qualitative evaluation of the matches and the elimination of poorly related sequences of the dataset.

Pruning sequences The syntax adopted by HMMER with the Stockholm format is useful to highlight the model states in the alignment[1]. This property was used in a script to eliminate the sequence regions flanking the profile HMM with regular expressions. The alignments produced with the *global-single* profile HMM were used for the sequence edition.

[1] In the alignment lines, uppercase (capital) letters are used to indicate emissions on match states, while lowercase symbols are used for insert states. Dashes (-) are used for deletions inside match state columns, and dots (.) fill the insert state columns (or non-emission of character states).

3.3 EXTENDED Dataset

The EXTENDED datasets were assembled with the sequences acquired and edited using profile HMM. PR-4 and PR-6 families ended with the smallest datasets, with only 26 sequences each. PR-9 produced the largest dataset, with 330 sequences, followed by PR-10, with 231 sequences. Most important, the EXTENDED dataset of PR-15 corresponds exactly to the same sequences composing the EXTENDED dataset of PR-16. They will therefore be treated hereafter as a single dataset (PR-15/PR-16).

Table 3.4 summarizes the number of sequences included in the EXTENDED datasets, the original size of the precursor sequences and annotated chains, and the size of the aligned sequences.

Table 3.4: EXTENDED dataset

PR-family	number of sequences	sequence length (precursor)	sequence length (chain)	alignment length
PR-1	65	136 – 418	135 – 155	168
PR-2	158	310 – 544	305 – 340	397
PR-3	181	208 – 459	229 – 261	306
PR-4	26	125 – 231	119 – 138	127
PR-5	142	169 – 665	148 – 234	297
PR-6	26	65 – 128	65 – 75	69
PR-7	52	421 – 840	505 – 651	875
PR-8	89	274 – 530	263 – 308	347
PR-9	330	294 – 367	281 – 343	494
PR-10	231	150 – 178	153 – 160	184
PR-11	31	327 – 479	259 – 366	404
PR-12	79	45 – 132	32 – 108	56
PR-13	62	103 – 142	31 – 110	123
PR-14	102	91 – 237	89 – 105	112
PR-15 / PR16	160	144 – 263	185 – 227	271
PR-17	33	206 – 641	203 – 209	224

3.3.1 Generating alternative multiple sequence alignments

Following the same strategy employed in the SEED dataset, all EXTENDED datasets were submitted to different multiple sequence alignment programs.

Providing PRANK with a guide tree reduced computation time by half, because the pairwise alignments for distance calculations were computationally expensive.

Assessing alignment consistency The alignment representing the EXTENDED dataset was chosen from among the alignment set using inter-consistency criteria. The visual evaluation did not detected major problems in the first ranked alignments. Consequently, the best ranked were always selected for the EXTENDED alignments.

Table 3.5 summarizes the inter-consistency scores calculated by MUMSA, and indicates the chosen EXTENDED alignment for each PR-family.

Table 3.5: Inter-consistency scores for EXTENDED alignments

		first ranked MSA	
PR-family	AOS	algorithm	MOS
PR-1	0.86	MAFFT L-INS-i	0.94
PR-2	0.82	MAFFT L-INS-i	0.91
PR-3	0.81	MAFFT L-INS-i	0.91
PR-4	0.92	MAFFT L-INS-i	0.98
PR-5	0.76	ProbCons	0.85
PR-6	0.89	Kalign	0.98
PR-7	0.69	MAFFT L-INS-i	0.79
PR-8	0.84	MAFFT L-INS-i	0.93
PR-9	0.80	MAFFT L-INS-i	0.90
PR-10	0.90	MAFFT L-INS-i	0.97
PR-11	0.80	MAFFT L-INS-i	0.87
PR-12	0.84	ProbCons	0.92
PR-13	0.74	MAFFT L-INS-i	0.83
PR-14	0.85	MAFFT L-INS-i	0.93
PR-15 / PR16	0.80	MAFFT L-INS-i	0.89
PR-17	0.90	MAFFT L-INS-i	0.97

3.3.2 Retrieving coding sequences

Protein-coding sequence alignments of all datasets were obtained by submitting the EXTENDED alignments to PROTOGENE at the T-Coffee server [129, 130]. Not every protein sequence included in the EXTENDED dataset had a corresponding DNA sequence stored in the nucleotide databases (from where PROTOGENE retrieves the sequences). This is because several PR sequences were obtained from direct protein sequencing, and the corresponding DNA sequence had not yet been described.

Another issue concerning the retrieval of protein-coding sequences is the redundancy of matches coding for the same amino acid sequence. Several times, they contain synonymous nucleotide substitutions, but frequently, they are identical sequences originating from different database entries. The problem is more notable here, because the given amino acid alignment contains only a region of the full sequences. The redundant sequences were excluded from the dataset using Perl scripts and manual curation.

In order to identify probable sequencing errors, large deletions, and frameshift mutations, the DNA sequences were submitted to the same alignment procedure as used for the protein sequences. The produced alignments were carefully assessed using DNATagger for visualization. Ninety-one sequences were excluded from the whole dataset using these criteria.

3.4 Phylogenetic Analysis

3.4.1 Model selection

To make a proper use of the tree reconstruction programs, it is advisable to choose a model of sequence evolution that best describes the data. Two model programs for selection of best-fit models of both protein and nucleotide evolution have been employed here. The best-fit models for each dataset are listed in Table 3.6.

ProtTest [131] was used to compare 56 candidate models of protein evolution. The models are the product of seven general matrices of amino acid substitution (Dayhoff, JTT, WAG , VT, DCMut, LG, and Blosum62) and eight combinations of alternative parameters (+F, +I, +I+F, +G, +G+F, +I+G, +I+G+F and none). Since PR-proteins are encoded by nuclear genes, the other matrices implemented in ProtTest (MtREV – for mitochondrial DNA; MtMam – eutherian mitochondrial genes; MtArt – Arthropoda mitochondrial genes; RtREV – retrovirus and reverse transcriptase and CpREV – for proteins encoded by chloroplast DNA) can be disregarded.

jModelTest [132] compared the fitting of 88 candidate models to the aligned nucleotide sequences for all but one of the EXTENDED datasets. It was not possible to test gamma rate variation among sites for PR-9, because the program crashed during the test. The 88 candidate models are the combination of 11 nucleotide substitution schemes (JC, HKY, TN, TPM1, TPM2, TPM3, TIM1, TIM2, TIM3, TVM and GTR) with +F, +I, and +G.

TABLE 3.6: Best-fit models for amino acid and nucleotide EXTENDED alignments

	ProtTest Model		jModelTest Model	
PR-family	subst. model	parameters	subst. model	parameters
PR-1	WAG	+G	TPM1uf	(+F) +I +G
PR-2	LG	+G	GTR	(+F) +I +G
PR-3	WAG	+G	TVM	(+F) +I +G
PR-4	WAG	+G	TIM1	(+F) +I +G
PR-5	WAG	+F	TVM	(+F) +I +G
PR-6	WAG	+G	TPM2uf	(+F) +I
PR-7	WAG	+I +G +F	TVM	(+F) +I +G
PR-8	WAG	+G	TVM	(+F) +I +G
PR-9	WAG	+I +G +F	GTR	(+F) +I +G
PR-10	LG	+I +G +F	SYM	+I +G
PR-11	WAG	+I +G +F	TVMef	+I +G
PR-12	WAG	+I +G	HKY	(+F) +I +G
PR-13	JTT	+G	TPM2	+I +G
PR-14	LG	+I +G	TPM1uf	(+F) +I +G
PR-15 / PR-16	WAG	+G	GTR	(+F) +I +G
PR-17	WAG	+I +G	GTR	(+F) +I +G

+I: proportion of invariable sites; +G: gamma rate variation among sites (4 rate categories); +F: observed amino acid frequencies (protein), and equal/unequal base frequencies (DNA).

(+F) indicates that the unequal base frequencies are intrinsic to the model.

3.4.2 Tree reconstruction

Two trees were calculated for each family EXTENDED alignment. The models selected by ProtTest and jModelTest were given as parameters for PhyML to calculate the trees for the amino acid and nucleotide sequences. Trees calculated with the amino acid sequences are presented in Appendix B. The figures were created with cTree [146].

3.4.3 Testing trees with codon models

Each pair of trees was assessed through codon substitution models using the one-rate model (M0) in *codeml*. The tree which best fits the codon sequences is further used in the multi-model analysis. Table 3.7 summarizes the results obtained with model M0 for both trees (amino acid and nucleotide based) in each PR-family. The best-fitting trees are indicated by asterisks.

Table 3.7: Likelihood estimates for AA and DNA trees under codon substitution models

PR	ns	tree	*lnL*	ΔlnL	pRELL	tree length
PR-1	63					
		AA	-12276.997	-28.021	0.246	45.472
		NT*	-12248.976	0.000	0.754	37.981
PR-2	141					
		AA	-62489.495	-88.906	0.081	101.274
		NT*	-62400.589	0.000	0.919	94.768
PR-3	157					
		AA	-30343.999	-121.015	0.006	81.644
		NT*	-30222.984	0.000	0.994	77.248
PR-4	24					
		AA	-4529.841	-64.853	0.003	14.822
		NT*	-4464.988	0.000	0.997	12.610
PR-5	130					
		AA	-22278.638	-67.304	0.118	86.982
		NT*	-22211.334	0.000	0.882	77.398
PR-6	15					
		AA	-1724.497	-0.745	0.418	9.883
		NT*	-1723.752	0.000	0.581	9.839
PR-7	41					
		AA*	-53243.549	0.000	0.996	57.649
		NT	-53387.482	-143.933	0.004	48.973
PR-8	77					
		AA	-28425.403	-24.198	0.320	48.119
		NT*	-28401.204	0.000	0.680	44.859
PR-9	308					
		AA	-151153.762	-162.601	0.119	280.754
		NT*	-150991.161	0.000	0.881	269.025
PR-10	225					
		AA	-27278.129	-128.746	0.001	80.858
		NT*	-27149.383	0.000	0.999	80.049
PR-11	19					
		AA*	-14657.678	0.000	0.699	23.162
		NT	-14667.118	-9.440	0.301	23.215
PR-12	47					
		AA	-3999.260	-70.370	0.032	62.072
		NT*	-3928.891	0.000	0.968	46.476
PR-13	46					
		AA	-3359.937	-42.795	0.015	15.741
		NT*	-3317.142	0.000	0.985	15.195
PR-14	82					
		AA	-12582.270	-92.875	0.019	71.909
		NT*	-12489.394	0.000	0.981	60.947
PR-15 /	147					
PR-16		AA	-33623.376	-69.861	0.071	103.529
		NT*	-33553.515	0.000	0.929	97.695
PR-17	26					
		AA	-8718.216	-42.164	0.123	13.547
		NT*	-8676.052	0.000	0.877	13.056

ns: number of sequences; *lnL*: log-likelihood of the tree under M0; ΔlnL : log-likelihood difference from the best-fitting tree; pRELL: bootstrap proportions using the RELL method.

3.4.4 Using codon substitution models to predict positive selection

The best-fitting trees from M0 were given as *user tree* in the multi-model (M0, M1a, M2a, M3, M7, M8 and M8a) analysis. Initially, the branch lengths estimated by M0 were set as initial values for optimization. Comparative tests indicated that branch length optimization was computationally very expensive. The very first attempt to perform all models in a row demonstrated that, for large datasets, M7, M8 and M8a (models with beta distribution) tend to enter a virtually endless iteration or fall into non-numeric ("NaN") parameter values. Nonetheless, models M0, M1a, M2a and M3 were processed in one run with branch lengths as free parameters. The results were tested with AIC, AICc and BIC, and the branch length estimates from the best-fit model were thus set as fixed values to perform M7, M8 and M8a analysis. For PR-9, the M0 branch lengths were given as fixed values already for M1a, M2a and M3 analysis, because the branch length optimization needed weeks to complete (and crashed in the meantime).

AIC, AICc and BIC Model selection with *information criteria* allows the comparison of several non-nested models simultaneously. Akaike information criterion (AIC), corrected AIC (AICc) and Bayesian information criterion (BIC) were applied to compare M0, M1a, M2a, and M3 maximum likelihood estimates. For AICc and BIC, the sample size was defined as the number of sequences times the alignment length (number of codon sites). The number of free parameters (np) includes model specific free parameters, κ, and branch length estimates (if not fixed). Table 3.8 presents the calculated values and the difference (ΔAIC, ΔAICc, ΔBIC) in relation to the best-fit model. A single asterisk is given for one best-fit model in one criterion. Here, M3 was unanimously chosen as the best-fit model in all three tests, for all PR-families.

Table 3.8: Model selection for codon models

Family	Model	np	*LnL*	AIC	ΔAIC	AICc	ΔAICc	BIC	ΔBIC
PR-1	M0	125	-12248.98	24747.95	1178.05	24752.02	1177.78	24984.98	1170.46
	M1a	126	-11878.83	24009.65	439.75	24013.78	439.55	24248.58	434.06
	M2a	128	-11878.83	24013.65	443.75	24017.91	443.68	24256.37	441.85
	M3 ***	129	-11655.95	23569.91	0.00	23574.24	0.00	23814.52	0.00
PR-2	M0	281	-62400.59	125363.18	4343.33	125367.60	4343.20	126081.82	4333.10
	M1a	282	-61182.52	122929.04	1909.19	122933.49	1909.09	123650.24	1901.52
	M2a	284	-61182.52	122933.04	1913.19	122937.56	1913.16	123659.35	1910.63
	M3 ***	285	-60224.92	121019.85	0.00	121024.40	0.00	121748.72	0.00
PR-3	M0	313	-30222.98	61071.97	1779.78	61080.79	1779.55	61808.85	1770.37
	M1a	314	-29973.80	60575.59	1283.41	60584.47	1283.23	61314.83	1276.34
	M2a	316	-29973.80	60579.59	1287.41	60588.58	1287.35	61323.54	1285.05
	M3 ***	317	-29329.09	59292.19	0.00	59301.23	0.00	60038.49	0.00
PR-4	M0	47	-4464.99	9023.98	417.22	9025.60	416.93	9092.22	411.41
	M1a	48	-4314.72	8725.44	118.68	8727.13	118.47	8795.14	114.33
	M2a	50	-4306.75	8713.50	106.74	8715.34	106.67	8786.11	105.29
	M3 ***	51	-4252.38	8606.76	0.00	8608.67	0.00	8680.81	0.00
PR-5	M0	259	-22211.33	44940.67	1798.83	44949.45	1798.55	45508.69	1790.05
	M1a	260	-21933.20	44386.40	1244.55	44395.24	1244.35	44956.61	1237.98
	M2a	262	-21933.20	44390.40	1248.55	44399.38	1248.49	44964.99	1246.36
	M3 ***	263	-21307.92	43141.84	0.00	43150.90	0.00	43718.63	0.00
PR-6	M0	29	-1723.75	3505.50	45.87	3507.44	45.30	3533.59	42.00
	M1a	30	-1705.78	3471.56	11.92	3473.63	11.49	3500.61	9.02
	M2a	32	-1705.57	3475.15	15.52	3477.50	15.37	3506.14	14.55
	M3 ***	33	-1696.82	3459.63	0.00	3462.14	0.00	3491.59	0.00
PR-7	M0	81	-53243.55	106649.10	4173.93	106649.72	4173.87	106838.07	4164.60
	M1a	82	-52044.18	104252.37	1777.20	104253.00	1777.15	104443.67	1770.20
	M2a	84	-52044.18	104256.37	1781.20	104257.03	1781.18	104452.33	1778.87
	M3 ***	85	-51152.58	102475.17	0.00	102475.85	0.00	102673.47	0.00
PR-8	M0	153	-28401.20	57108.41	2196.76	57111.10	2196.62	57452.10	2187.78
	M1a	154	-27795.07	55898.15	986.50	55900.88	986.39	56244.08	979.76
	M2a	156	-27795.07	55902.15	990.50	55904.95	990.46	56252.57	988.25
	M3 ***	157	-27298.82	54911.65	0.00	54914.48	0.00	55264.32	0.00
PR-9 [a]	M0	2	-150925.37	301854.74	12817.21	301854.74	12817.21	301860.51	12805.68
	M1a	3	-147725.51	295457.01	6419.48	295457.01	6419.48	295465.66	6410.83
	M2a	5	-147725.51	295461.01	6423.48	295461.01	6423.48	295475.42	6420.60
	M3 ***	6	-144512.77	289037.54	0.00	289037.54	0.00	289054.82	0.00
PR-10	M0	449	-27149.38	55196.77	819.95	55212.52	819.66	56281.84	810.28
	M1a	450	-27010.27	54920.54	543.72	54936.36	543.51	56008.03	536.47
	M2a	452	-27010.18	54924.37	547.55	54940.34	547.48	56016.69	545.13
	M3 ***	453	-26735.41	54376.82	0.00	54392.86	0.00	55471.56	0.00
PR-11	M0	37	-14657.68	29389.36	769.91	29389.83	769.80	29455.06	762.80
	M1a	38	-14418.11	28912.22	292.77	28912.72	292.69	28979.70	287.45
	M2a	40	-14418.11	28916.22	296.77	28916.77	296.74	28987.25	295.00
	M3 ***	41	-14268.72	28619.45	0.00	28620.03	0.00	28692.25	0.00
PR-12	M0	93	-3928.89	8043.78	426.29	8053.32	425.44	8163.28	421.16
	M1a	94	-3816.24	7820.49	203.00	7830.24	202.36	7941.27	199.15
	M2a	96	-3807.19	7806.38	188.89	7816.56	188.68	7929.73	187.61
	M3 ***	97	-3711.74	7617.49	0.00	7627.88	0.00	7742.12	0.00
PR-13	M0	91	-3317.14	6816.28	240.17	6822.45	239.61	6948.06	234.38
	M1a	92	-3217.47	6618.94	42.83	6625.25	42.41	6752.17	38.49
	M2a	94	-3196.44	6580.89	4.78	6587.47	4.63	6717.01	3.33
	M3 ***	95	-3193.05	6576.11	0.00	6582.84	0.00	6713.68	0.00
PR-14	M0	163	-12489.39	25304.79	973.81	25312.55	973.42	25606.06	966.41
	M1a	164	-12139.47	24606.94	275.95	24614.80	275.66	24910.06	270.41
	M2a	166	-12134.29	24600.59	269.60	24608.64	269.51	24907.41	267.76
	M3 ***	167	-11998.49	24330.98	0.00	24339.13	0.00	24639.65	0.00
PR-15/ PR-16	M0	293	-33553.52	67693.03	1820.75	67701.17	1820.53	68376.21	1811.43
	M1a	294	-33296.17	67180.34	1308.06	67188.53	1307.89	67865.85	1301.07
	M2a	296	-33296.17	67184.34	1312.06	67192.64	1312.00	67874.51	1309.73
	M3 ***	297	-32639.14	65872.28	0.00	65880.64	0.00	66564.78	0.00
PR-17	M0	51	-8676.05	17454.10	519.11	17455.12	518.95	17541.95	512.22
	M1a	52	-8509.02	17122.05	187.06	17123.10	186.93	17211.62	181.89
	M2a	54	-8509.02	17126.05	191.06	17127.19	191.01	17219.06	189.33
	M3 ***	55	-8412.50	16934.99	0.00	16936.17	0.00	17029.73	0.00

[a] PR-9: Branch lengths fixed at M0 estimates

LRT The likelihood ratio test can be employed in pairs of nested models (when one is a more complex case of the other). Here, LRT was applied to M1a *vs.* M2a, M0 *vs.* M3, and M7 *vs.* M8, with *df* (degree of freedom) equal to 2, 4 and 2, respectively. M8a *vs.* M8 uses the critical values 2.71 at 5%, 5.41 at 1% and 9.55 at 0.1% [142]. In the site models, M1a *vs.* M2a was significant for PR-4, PR-12, PR-13 and PR-14, while M3 was significatively better than M0 in all comparisons. M8 was significantly higher than M7 and M8a for PR-4 and PR-3.

Although Yang [142] suggests that the M0 *vs.* M3 comparison should be used as a test of variable ω among sites rather than a test of positive selection, the presence of a category with $\omega_2 > 1$ and p_2 not zero accounts for M3 as a model capable of identifying positive selection in specific cases. This could be observed for PR-6 and PR-13, where M3 detected sites under positive selection, that were not detected by M2a or M8.

NEB, BEB Posterior probabilities for site classes are automatically calculated by *codeml* using the naïve empirical Bayes (NEB) and the Bayes empirical Bayes (BEB) approaches [147]. Sites under positive selection were identified in PR-4, PR-6, PR-12, PR-13 and PR-14. PR-4 has one site identified with NEB and BEB for models M2a and M8. Two sites were detected by M3 with NEB for PR-6. Model M2a detected two sites in PR-12 and one in PR-14 with BEB. For PR-13, five sites were identified by BEB under M2a and M8, while 12 sites were detected using NEB under M3.

Average branch length The sequence divergence is quantified by the tree length, which is defined as the expected number of nucleotide substitutions per codon along the tree. As this value is directly affected by the number of sequences, Anisimova and colleagues (2001) introduced a measure of relative sequence divergence to enable a qualitative comparison between datasets of different sizes. They used the average number of nucleotide changes per codon per branch (referred to as A in Table 3.9), which is calculated as the tree length (S) divided by the number of branches of an unrooted tree of T taxa[2] ($2T - 3$) [148].

The average branch length calculated for the each dataset ranged from 0.17 (PR-13, M0) to 0.78 (PR-7, M3) nucleotide substitutions per codon, which is within the values considered as a medium level of sequence divergence by the authors.

[2] Here, T is the number of sequences (ns) in the alignment.

Table 3.9: Results obtained with codon substitution models.

Family	ns	ls	Model	np	S	A	κ	lnL	estimates	LRT (P-value)	pos. sites
PR-1	63	125	M0	125	37.98	0.31	1.32	-12248.98	ω: 0.20761		
			M1a	126	42.73	0.35	1.49	-11878.83	p: 0.67225 0.32775 ω: 0.13049 1.00000		
			M2a	128	42.73	0.35	1.49	-11878.83	p: 0.67225 0.25679 0.07096 ω: 0.13049 1.00000 1.00000	M2 vs M1 0.000	
			M3	129	45.42	0.37	1.35	-11655.95	p: 0.28387 0.40956 0.30657 ω: 0.01480 0.17686 0.57949	M3 vs M0 **1186.046** (<0.001)	
			M7	3	45.42	0.37	1.35	-11645.35	p: 0.12500 ... 0.12500 ω: 0.00266 ... 0.75818		
			M8	5	45.42	0.37	1.35	-11645.35	p: 0.12500 ... 0.12500 0.00001 ω: 0.00266 ... 0.75816 1.00000	M8 vs M7 -0.000	
			M8a	4	45.42	0.37	1.35	-11645.35	p: 0.12500 ... 0.12500 0.00001 ω: 0.00266 ... 0.75816 1.00000	M8 vs M81 0.000	
PR-2	141	256	M0	281	94.77	0.34	1.24	-62400.59	ω: 0.18553		
			M1a	282	97.56	0.35	1.40	-61182.52	p: 0.75342 0.24658 ω: 0.15406 1.00000		
			M2a	284	97.56	0.35	1.40	-61182.52	p: 0.75342 0.24658 0.00000 ω: 0.15406 1.00000 28.44398	M2 vs M1 -0.000	
			M3	285	101.10	0.36	1.26	-60224.92	p: 0.36847 0.38073 0.25080 ω: 0.04826 0.18785 0.50781	M3 vs M0 **4351.331** (<0.001)	
			M7	3	101.10	0.36	1.26	-60078.52	p: 0.12500 ... 0.12500 ω: 0.00947 ... 0.59680		
			M8	5	101.10	0.36	1.26	-60078.52	p: 0.12500 ... 0.12500 0.00001 ω: 0.00947 ... 0.59678 1.00000	M8 vs M7 -0.002	
			M8a	4	101.10	0.36	1.26	-60078.52	p: 0.12500 ... 0.12500 0.00001 ω: 0.00947 ... 0.59678 1.00000	M8 vs M81 0.000	
PR-3	157	144	M0	313	77.25	0.25	1.48	-30222.98	ω: 0.13313		
			M1a	314	78.03	0.25	1.65	-29973.80	p: 0.86700 0.13300 ω: 0.12559 1.00000		
			M2a	316	78.03	0.25	1.65	-29973.80	p: 0.86698 0.13302 0.00000 ω: 0.12559 1.00000 19.01704	M2 vs M1 0.000	
			M3	317	80.77	0.26	1.46	-29329.09	p: 0.28627 0.47313 0.24060 ω: 0.02068 0.11585 0.34552	M3 vs M0 **1787.783** (<0.001)	
			M7	3	80.77	0.26	1.46	-29269.77	p: 0.12500 ... 0.12500 ω: 0.00603 ... 0.42827		
			M8	5	80.77	0.26	1.46	-29269.77	p: 0.12500 ... 0.12500 0.00001 ω: 0.00603 ... 0.42826 1.00000	M8 vs M7 -0.005	
			M8a	4	80.77	0.26	1.46	-29269.77	p: 0.12500 ... 0.12500 0.00001 ω: 0.00603 ... 0.42826 1.00000	M8 vs M81 0.000	

Family	ns	ls	Model	np	S	A	κ	lnL	estimates	LRT (P-value)	pos. sites
PR-4	24	118	M0	47	12.61	0.28	1.51	-4464.99	ω: 0.13647		
			M1a	48	13.33	0.30	1.74	-4314.72	p: 0.78936 0.21064 ω: 0.07786 1.00000		
			M2a	50	13.93	0.31	1.78	-4306.75	p: 0.78402 0.20750 0.00848 ω: 0.07896 1.00000 5.54825	M2 vs M1 **15.939** (<0.001)	**102N**
			M3	51	13.63	0.30	1.60	-4252.38	p: 0.45129 0.36598 0.18273 ω: 0.00816 0.15750 0.65334	M3 vs M0 **425.219** (<0.001)	
			M7	3	13.63	0.30	1.59	-4252.60	p: 0.12500 ... 0.12500 ω: 0.00005 ... 0.70643		
			M8	5	13.63	0.30	1.60	-4242.75	p: 0.12416 ... 0.12416 0.00670 ω: 0.00009 ... 0.62157 3.92046	M8 vs M7 **19.683** (<0.001)	**102N**
			M8a	4	13.63	0.30	1.58	-4249.65	p: 0.12210 ... 0.12210 0.02324 ω: 0.00010 ... 0.57981 1.00000	M8 vs M81 **13.801** (<0.001)	
PR-5	130	120	M0	259	77.40	0.30	1.29	-22211.33	ω: 0.10963		
			M1a	260	75.76	0.29	1.53	-21933.20	p: 0.77766 0.22234 ω: 0.10060 1.00000		
			M2a	262	75.76	0.29	1.53	-21933.20	p: 0.77767 0.22233 0.00000 ω: 0.10060 1.00000 18.55511	M2 vs M1 -0.000	
			M3	263	83.62	0.33	1.32	-21307.92	p: 0.28448 0.45165 0.26387 ω: 0.00882 0.09424 0.32251	M3 vs M0 **1806.827** (<0.001)	
			M7	3	83.62	0.33	1.31	-21256.63	p: 0.12500 ... 0.12500 ω: 0.00151 ... 0.43688		
			M8	5	83.62	0.33	1.31	-21256.63	p: 0.12500 ... 0.12500 0.00001 ω: 0.00151 ... 0.43688 1.25088	M8 vs M7 -0.003	
			M8a	4	83.62	0.33	1.31	-21256.63	p: 0.12500 ... 0.12500 0.00001 ω: 0.00151 ... 0.43688 1.00000	M8 vs M81 -0.001	
PR-6	15	62	M0	29	9.84	0.36	1.90	-1723.75	ω: 0.30699		
			M1a	30	10.84	0.40	2.09	-1705.78	p: 0.72086 0.27914 ω: 0.20809 1.00000		
			M2a	32	11.02	0.41	2.11	-1705.57	p: 0.71460 0.26781 0.01759 ω: 0.21282 1.00000 2.52957	M2 vs M1 0.409	
			M3	33	11.30	0.42	2.04	-1696.82	p: 0.26209 0.68662 0.05128 ω: 0.05049 0.38220 1.86563	M3 vs M0 **53.870** (<0.001)	**6T** 28K
			M7	3	11.30	0.42	2.06	-1699.36	p: 0.12500 ... 0.12500 ω: 0.02551 ... 0.81078		
			M8	5	11.30	0.42	2.06	-1697.14	p: 0.12063 ... 0.12063 0.03494 ω: 0.03387 ... 0.71240 2.20961	M8 vs M7 4.424	
			M8a	4	11.30	0.42	2.06	-1698.70	p: 0.11432 ... 0.11432 0.08545 ω: 0.03281 ... 0.64786 1.00000	M8 vs M81 3.102 (<0.05)	

Family	ns	ls	Model	np	S	A	κ	lnL	estimates	LRT (P-value)	pos. sites
PR-7	41	525	M0	81	57.65	0.73	1.34	-53243.55	ω: 0.17156		
			M1a	82	52.05	0.66	1.50	-52044.18	p: 0.64103 0.35897 ω: 0.14079 1.00000		
			M2a	84	52.05	0.66	1.50	-52044.18	p: 0.64104 0.35896 0.00000 ω: 0.14079 1.00000 26.77653	M2 vs M1 0.000	
			M3	85	61.83	0.78	1.37	-51152.58	p: 0.28111 0.36983 0.34906 ω: 0.02334 0.14028 0.42355	M3 vs M0 **4181.931** (<0.001)	
			M7	3	61.83	0.78	1.38	-51048.33	p: 0.12500 ... 0.12500 ω: 0.00521 ... 0.61226		
			M8	5	61.83	0.78	1.38	-51048.33	p: 0.12500 ... 0.12500 0.00001 ω: 0.00521 ... 0.61224 1.00000	M8 vs M7 -0.002	
			M8a	4	61.83	0.78	1.38	-51048.33	p: 0.12500 ... 0.12500 0.00001 ω: 0.00521 ... 0.61224 1.00000	M8 vs M81 0.000	
PR-8	77	229	M0	153	44.86	0.30	1.28	-28401.20	ω: 0.18611		
			M1a	154	45.13	0.30	1.47	-27795.07	p: 0.69534 0.30466 ω: 0.13354 1.00000		
			M2a	156	45.13	0.30	1.47	-27795.07	p: 0.69535 0.30465 0.00000 ω: 0.13354 1.00000 30.06301	M2 vs M1 -0.000	
			M3	157	47.48	0.31	1.28	-27298.82	p: 0.38440 0.33716 0.27845 ω: 0.03368 0.18126 0.49488	M3 vs M0 **2204.762** (<0.001)	
			M7	3	47.48	0.31	1.29	-27254.49	p: 0.12500 ... 0.12500 ω: 0.00583 ... 0.61269		
			M8	5	47.48	0.31	1.29	-27254.49	p: 0.12500 ... 0.12500 0.00001 ω: 0.00583 ... 0.61268 1.38521	M8 vs M7 -0.005	
			M8a	4	47.48	0.31	1.29	-27254.49	p: 0.12500 ... 0.12500 0.00001 ω: 0.00583 ... 0.61268 1.00000	M8 vs M81 -0.002	
PR-9	308	247	M0	2	255.52	0.42	1.29	-150925.37	ω: 0.17466		
			M1a	3	255.52	0.42	1.48	-147725.51	p: 0.67807 0.32193 ω: 0.13796 1.00000		
			M2a	5	255.52	0.42	1.48	-147725.51	p: 0.67807 0.16986 0.15207 ω: 0.13796 1.00000 1.00000	M2 vs M1 0.000	
			M3	6	255.52	0.42	1.30	-144512.77	p: 0.25013 0.45864 0.29122 ω: 0.01582 0.12829 0.42799	M3 vs M0 **12825.207** (<0.001)	
			M7	3	255.52	0.42	1.30	-143872.06	p: 0.12500 ... 0.12500 ω: 0.00377 ... 0.54979		
			M8	5	255.52	0.42	1.30	-143871.95	p: 0.12489 ... 0.12489 0.00084 ω: 0.00367 ... 0.54971 2.72874	M8 vs M7 0.229	
			M8a	4	255.52	0.42	1.30	-143872.29	p: 0.12486 ... 0.12486 0.00111 ω: 0.00352 ... 0.55216 1.00000	M8 vs M81 0.685	

Family	ns	ls	Model	np	S	A	κ	lnL	estimates	LRT (P-value)	pos. sites
PR-10	225	116	M0	449	80.05	0.18	1.47	-27149.38	ω: 0.27192		
			M1a	450	81.99	0.18	1.61	-27010.27	p: 0.75632 0.24368 ω: 0.26670 1.00000		
			M2a	452	82.02	0.18	1.61	-27010.18	p: 0.75456 0.24197 0.00347 ω: 0.26738 1.00000 1.49354	M2 vs M1 0.169	
			M3	453	85.05	0.19	1.46	-26735.41	p: 0.13625 0.42465 0.43910 ω: 0.06393 0.19484 0.45968	M3 vs M0 **827.948** (<0.001)	
			M7	3	85.05	0.19	1.46	-26701.85	p: 0.12500 ... 0.12500 ω: 0.03693 ... 0.66292		
			M8	5	85.05	0.19	1.47	-26697.04	p: 0.12165 ... 0.12165 0.02679 ω: 0.04419 ... 0.59537 1.00000	M8 vs M7 **9.635** (<0.01)	
			M8a	4	85.05	0.19	1.47	-26697.04	p: 0.12165 ... 0.12165 0.02679 ω: 0.04419 ... 0.59537 1.00000	M8 vs M81 0.000	
PR-11	19	314	M0	37	23.16	0.66	1.62	-14657.68	ω: 0.18761		
			M1a	38	24.38	0.70	1.84	-14418.11	p: 0.74921 0.25079 ω: 0.15084 1.00000		
			M2a	40	24.38	0.70	1.84	-14418.11	p: 0.74921 0.19699 0.05380 ω: 0.15084 1.00000 1.00000	M2 vs M1 0.000	
			M3	41	25.84	0.74	1.70	-14268.72	p: 0.23250 0.47205 0.29545 ω: 0.02411 0.15169 0.49611	M3 vs M0 **777.907** (<0.001)	
			M7	3	25.84	0.74	1.69	-14264.31	p: 0.12500 ... 0.12500 ω: 0.01144 ... 0.60195		
			M8	5	25.84	0.74	1.69	-14264.32	p: 0.12500 ... 0.12500 0.00001 ω: 0.01143 ... 0.60201 2.84888	M8 vs M7 -0.004	
			M8a	4	25.84	0.74	1.71	-14263.14	p: 0.12086 ... 0.12086 0.03316 ω: 0.01278 ... 0.54727 1.00000	M8 vs M81 -2.358	
PR-12	47	41	M0	93	46.48	0.51	1.26	-3928.89	ω: 0.15474		
			M1a	94	40.41	0.44	1.74	-3816.24	p: 0.26829 0.73171 ω: 0.01668 1.00000		
			M2a	96	42.30	0.46	1.80	-3807.19	p: 0.26829 0.68332 0.04839 ω: 0.01747 1.00000 2.97373	M2 vs M1 **18.110** (<0.001)	**11V** 26R
			M3	97	51.31	0.56	1.29	-3711.74	p: 0.21951 0.27179 0.50870 ω: 0.00000 0.12918 0.32499	M3 vs M0 **434.295** (<0.001)	
			M7	3	51.31	0.56	1.30	-3714.23	p: 0.12500 ... 0.12500 ω: 0.00115 ... 0.67925		
			M8	5	51.31	0.56	1.30	-3714.22	p: 0.12462 ... 0.12462 0.00306 ω: 0.00118 ... 0.67193 1.02446	M8 vs M7 0.003	
			M8a	4	51.31	0.56	1.30	-3714.22	p: 0.12469 ... 0.12469 0.00251 ω: 0.00118 ... 0.67316 1.00000	M8 vs M81 0.001	

Family	ns	ls	Model	np	S	A	κ	lnL	estimates	LRT (P-value)	pos. sites
PR-13	46	61	M0	91	15.20	0.17	1.63	-3317.14	ω: 0.68134		
			M1a	92	16.97	0.19	1.52	-3217.47	p: 0.38180 0.61820 ω: 0.15361 1.00000		
			M2a	94	18.13	0.20	1.66	-3196.44	p: 0.32722 0.53697 0.13581 ω: 0.15140 1.00000 2.90460	M2 vs M1 **42.058** (<0.001)	6L 16F **20V 23G 24A**
			M3	95	18.10	0.20	1.60	-3193.05	p: 0.20476 0.49359 0.30165 ω: 0.05997 0.60535 1.91057	M3 vs M0 **248.175** (<0.001)	**6L** 13T **16F 17A 19G 20V 23G 24A** 28S 36R **60I** 61Q
			M7	3	18.10	0.20	1.45	-3206.04	p: 0.12500 ... 0.12500 ω: 0.01367 ... 0.99967		
			M8	5	18.10	0.20	1.59	-3187.94	p: 0.10563 ... 0.10563 0.15493 ω: 0.01255 ... 0.99855 2.39911	M8 vs M7 **36.197** (<0.001)	6L 16F **20V 23G 24A**
			M8a	4	18.10	0.20	1.46	-3204.48	p: 0.07087 ... 0.07087 0.43306 ω: 0.01288 ... 0.77784 1.00000	M8 vs M81 **33.073** (<0.001)	
PR-14	82	86	M0	163	60.95	0.38	1.30	-12489.39	ω: 0.27603		
			M1a	164	66.87	0.42	1.37	-12139.47	p: 0.43792 0.56208 ω: 0.13763 1.00000		
			M2a	166	68.47	0.43	1.38	-12134.29	p: 0.43264 0.52004 0.04732 ω: 0.14002 1.00000 1.94955	M2 vs M1 **10.350** (<0.01)	24A
			M3	167	69.34	0.43	1.26	-11998.49	p: 0.26254 0.35927 0.37819 ω: 0.04368 0.27030 0.67348	M3 vs M0 **981.806** (<0.001)	
			M7	3	69.34	0.43	1.26	-11962.13	p: 0.12500 ... 0.12500 ω: 0.00691 ... 0.89297		
			M8	5	69.34	0.43	1.27	-11960.98	p: 0.11957 ... 0.11957 0.04341 ω: 0.00551 ... 0.85268 1.27691	M8 vs M7 2.291	
			M8a	4	69.34	0.43	1.26	-11961.79	p: 0.11938 ... 0.11938 0.04492 ω: 0.00525 ... 0.85818 1.00000	M8 vs M81 1.611	
PR-15/	147	146	M0	293	97.69	0.34	1.45	-33553.52	ω: 0.14696		
PR-16			M1a	294	93.51	0.32	1.55	-33296.17	p: 0.83249 0.16751 ω: 0.14733 1.00000		
			M2a	296	93.51	0.32	1.55	-33296.17	p: 0.83249 0.16751 0.00000 ω: 0.14733 1.00000 14.10641	M2 vs M1 -0.000	
			M3	297	101.61	0.35	1.42	-32639.14	p: 0.24337 0.40647 0.35017 ω: 0.02575 0.11376 0.32309	M3 vs M0 **1828.75** (<0.001)	
			M7	3	101.61	0.35	1.42	-32561.96	p: 0.12500 ... 0.12500 ω: 0.00863 ... 0.47947		
			M8	5	101.61	0.35	1.43	-32556.94	p: 0.12416 ... 0.12416 0.00672 ω: 0.00950 ... 0.44618 1.00000	M8 vs M7 **10.040** (<0.01)	
			M8a	4	101.61	0.35	1.43	-32556.94	p: 0.12416 ... 0.12416 0.00672 ω: 0.00950 ... 0.44618 1.00000	M8 vs M81 0.000	

Family	ns	ls	Model	np	S	A	κ	lnL	estimates	LRT (P-value)	pos. sites
PR-17	26	203	M0	51	13.06	0.27	1.43	-8676.05	ω: 0.20223		
			M1a	52	14.18	0.29	1.50	-8509.02	p: 0.78372 0.21628 ω: 0.14159 1.00000		
			M2a	54	14.18	0.29	1.50	-8509.02	p: 0.78372 0.21628 0.00000 ω: 0.14159 1.00000 30.81466	M2 vs M1 0.000	
			M3	55	14.25	0.29	1.40	-8412.50	p: 0.38987 0.44973 0.16039 ω: 0.03025 0.24453 0.74413	M3 vs M0 **527.112** (<0.001)	
			M7	3	14.25	0.29	1.39	-8416.58	p: 0.12500 ... 0.12500 ω: 0.00365 ... 0.69741		
			M8	5	14.25	0.29	1.40	-8415.21	p: 0.11884 ... 0.11884 0.04926 ω: 0.00509 ... 0.59406 1.00000	M8 vs M7 2.745	
			M8a	4	14.25	0.29	1.40	-8415.21	p: 0.11884 ... 0.11884 0.04926 ω: 0.00509 ... 0.59406 1.00000	M8 vs M81 0.000	

ns, number of sequences; ls, length (in codon sites); np, number of free parameters; S, sequence divergence (tree length); A, average branch length; κ, transition/transversion rate ratio (ts/tv); lnL, log-likelihood; p, proportion of codon sites for site classes; ω, non-synonymous/synonymous rate ratio (d_N/d_S); LRT, likelihood ratio test. Sites inferred to be under positive selection at $P > 95\%$ ($P > 99\%$ shown in bold).

3.5 Discussion

The analysis presented in this chapter was motivated by the observation that most of the publications on PR-proteins concentrate on their expression patterns and agricultural applications, but dedicate less effort to the evolution of these extraordinary proteins. Moreover, in recent years, the attention previously given to the PR-proteins has turned the focus to other aspects of the plant defense system, directing the research to signaling pathways and triggering factors, addressing the economic benefits of genetic engineering. This shift of interests may be attributed to the undesirable side effects produced by the enhanced expression of certain PR genes. We cannot lose sight of the fact that many PR-proteins are on the list of important food allergens [149, 150], and that the most resistant cultivar may also produce the most unpalatable fruit.

Studies concerning the evolution of specific PR-proteins were conducted by a few authors [21, 151–158], yielding very interesting insights. Other PR-families, however, remain neglected by the evolutionists. The present work aims to investigate and characterize all PR-families in the context of sequence diversity, evolutionary patterns and phylogenetic relationships, using the genetic data available in the public databases.

3.5.1 Dataset

It is indeed not surprising that even the molecular biology studies are unevenly distributed among these groups, so that publications describing members of less known families, such as PR-11 and PR-17, are rarely available. The same is true for the annotation of the sequences in the protein databases. Sequences identified as pathogenesis-related were abundantly found for PR-1, PR-2, PR-3, PR-4, PR-5 and PR-10, while almost no references about defensive features were found in the annotation of PR-7, PR-11 and PR-17. Moreover, as discussed by Tuzun and Somanchi (2006), homologies at the nucleotide sequence level may be encountered without information on the expression or characteristics of the encoded protein, leading to a complexity in comparative analysis of PR-proteins from different species. For that reason, the assemblage of the SEED datasets required a different approach for each case, based on a close inspection of the related literature.

During this investigation, it became clear that the availability and organism diversity of the sequences were largely heterogeneous among the PR-families. The number of homologous sequences found in the UniProtKB ranged from 51, for PR-17, to 1412, for PR-9 (using the *global-single* profile HMM). A large number of stored sequences does not necessarily imply a great diversity of represented organisms. Among the 1412 PR-9 sequences for example, 465 alone belonged to rice (*Oryza sativa*), 132 sequences originated from maize (*Zea mays*), 113 from grapevine (*Vitis vinifera*), and 104 from *Arabidopsis thaliana*. Paralogous genes do not represent a problem for the analysis, on the contrary, they are an important piece of the whole puzzle. On the other hand, the large number of isoforms and cultivar samples stored for the same species is very difficult to trace indeed. These data inflate the datasets with an excess of low variable clusters, introducing very little phylogenetic information.

To overcome this inconvenience without resorting to manual selection of database entries, a variety of filtering strategies were developed to identify undesirable redundancies. These filters benefit from the structured annotation system of UniProtKB, and are also able to identify and eliminate entries with other unwanted features (fragments, hypothetical proteins, putative genes) according to criteria defined for each specific dataset. The use of these filtering strategies provided a significant quality increase for the automatic selection of entries, although it did not solve the problem of uneven diversity of represented species.

It must be borne in mind that most of the plant sequences stored in the public databases originated from model organisms and crop species. Therefore, these species will be naturally overrepresented in most datasets. Rarely, sequences from singular unusual species become the most famous members of a specific PR-family, as in the case of the well known pollen allergen Betv1 from *Betula verrucosa* (white birch) in PR-10, and the intensely sweet-tasting thaumatin (PR-5) from *Thaumatococcus daniellii*, a West African rain forest plant.

3.5.2 Selection

Bishop, Dean and Mitchell-Olds (2000) were the first to provide empirical evidence that plant defense proteins not involved in pathogen recognition may also be subject to pathogen-imposed selection. They tested positive selection on class I chitinase genes

(PR-3) from the genus *Arabis* (Cruciferae) using codon substitution models, finding an excess of amino acid replacements in the active site and substrate binding cleft [160].

Further on, Bishop and colleagues (2005) detected strong evidence of positive selection on the soybean endo-1,3-glucanase A (EGaseA), a PR-2 family member. EGaseA is the major elicitor-releasing isozyme, and a high-affinity ligand for a glucanase inhibitor protein (GIP1) produced by the pathogen *Phytophthora sojae.* The elevated number of positively selected sites in the proximity of the binding site to GIP1 suggests a repeated adaptation to pathogen attack and inhibition [161].

Unlike the results found by Bishop and colleagues in both surveys, the selection analysis conducted here does not reproduce the findings. In the present study no evidence of positive selection could be found for PR-2 and PR-3. This outcome is not necessarily unexpected, since the structure of either dataset differs fundamentally. The datasets used in the two cited works encompass mostly sequences of the same genus or even of the same species, and consequently present very low sequence divergence. In this case, codon substitution models can easily detect rapidly evolving sites occurring in a short period of time.

The present work, otherwise, attempts to assess positive selection on a larger timescale. The datasets of PR-2 and PR-3, as well as most of the datasets analyzed here, attempt to encompass the maximal diversity available for these proteins in the publicly available sources of sequences. This may include sequences from different plant families and, even more deeply, different plant phyla (sequences from different kingdoms were not included in the selection analysis). When analyzing very large datasets with a reasonable sequence divergence using site models (not branch-site models), selection occurring at codon sites restricted to specific branches is overshadowed by the excess of neutrally evolving codons at the same position in other branches.

As asserted by Wong and colleagues (2004), a reasonable amount of synonymous and non-synonymous substitutions are necessary for the success of the method, as low divergent sequences provide too little information, while at high levels of divergence synonymous substitutions are often saturated [140].

The dataset compiled for PR-4 was ca. six times smaller then the datasets of PR-2 and PR-3, although it still comprises the variability expected for the proposed analysis, presenting sequences from several species of Liliopsida and Eudicotyledons, and average branch length of 0.3 (medium level of sequence divergence). Here, the codon-models analysis was able to detect positive selection and identify one codon site with an excess of amino acid substitutions. The site Asn 102 (or Ile 107 in the Barwin (PDB: 1bw3) mature peptide) accommodates the most diverse types of amino acids (positively and negatively charged, uncharged polar, hydrophobic) and therefore could be undergoing diversifying selection in an arms race interaction.

Diversifying selection can be a source of new tools to expand the possibilities for protein function, a repository of variability to explore new environments, or even a strategy to escape the attack of enemies. In the specific context of pathogenesis-related proteins, it can be a way to circumvent the loss of functional efficiency in response to a pathogen, avoiding, for example, the binding of enzyme inhibitors. Response to inhibitory proteins is, in fact, one of the suggestions made by Bishop and colleagues to explain the substitutions observed within the active site cleft in class I chitinase from *Arabis* species.

Another PR-protein estimated to undergo positive selection was PR-6, that is itself a proteinase inhibitor. This finding is consistent with the suggestion of compensatory replacements to recover the lost affinity for a binding site. Proteinase inhibitors may act by reducing the ability of the pathogen to use its lytic enzymes (fungi), to complete its replication cycles (viruses), or to obtain nutrition through digestion of host proteins (nematodes, insects) [162]

In the case of the antimicrobial peptides PR-12 (plant defensins), PR-13 (α- and β-thionins) and PR-14 (non-specific lipid-transfer proteins – nsLTPs) however, the positive selection could be more easily attributed to diversifying selection. Although they are better known as membrane-permeabilizing peptides, their toxicity can be related to different properties, including protease inhibition.

Most plant defensins (PR-12) exhibit antimicrobial properties, including antifungal activity against a broad range of fungi and antibiotic effects against Gram-positive and Gram-negative bacteria. Antifungal plant defensins can interact with fungal membrane components causing an alteration in permeability and consequently increased Ca2+ influx and K+ efflux [163]. It has been shown that Rs-AFPs (from radish, *Raphanus sativus*) specifically interact with the sphingolipid glucosylceramide, component of the cell wall of the yeasts *Pichia pastoris* and *Candida albicans*, and that Dm-AMP1 (from *Dahlia merckii*) targets the sphingolipid mannosyldiinositolphosphorylceramide (M(IP)2C) from *Saccharomyces cerevisiae* [164–166]. These interactions might facilitate the insertion of the defensin into the fungal plasma membrane, leading to membrane destabilization and arrest of fungal growth [167]. Other defensins may stay outside the cell and induce fungal cell death via modulation of intercellular signaling cascades, such as the induction of reactive oxygen species (ROS), but they can also bind to sodium channels or inhibit protein translation [168–170].

Besides the antimicrobial potential, plant defensins can be engaged in the defense against feeding insects, acting as enzyme inhibitors. Some defensins have been shown to inhibit alpha-amylase, enzymes present in insect gut and involved in digestion of plant material. Similarly, protease inhibitors, like trypsin inhibitors, may inhibit protease activity during insect predation [171–175]. It is worth noting that defensins exibiting alpha-amylase activity are not involved in antifungal activities [174, 176]. The same is valid for other properties: individual defensins can exert one or two of of these defense related attributes, but none of them can act as “do all” proteins [166].

Thionins (PR-13) have been shown to be toxic to a broad range of biological systems, including bacteria, fungi, insect larvae, and cultured mammalian cells as well as to small laboratory animals [177, 178]. It has been proposed that the primary mode of action for thionin toxicity is their ability to form ion channels in cell membranes by electrostatic interaction with negatively charged membrane phospholipids. There are also indications that thionins may function as regulatory proteins [178]. Interestingly, the toxicity of thionins was first observed in connection with the inhibition of bread and beer fermentation. Already in the 19th century, it was proposed that unknown grain compound, present in some batches of wheat flour, could be lethal to bread yeast [cited by 177–180]. This is very unwanted side effect for increasing the resistance of cereals by genetic engineering.

Finally, the proposed biological roles for nsLTPs (PR-14) include direct antimicrobial defense, defensive signaling, production of cuticular waxes in epidermal cells, beta-oxidation, somatic embryogenesis and modulation of plant growth and development [181, 182]. They have also been reported as important fruit allergens [149, 150].

All these possibilities of action of antimicrobial peptides are consistent with the observation of positive (diversifying) selection acting on these PR-families.

Chapter 4

Chitinases

4.1 Introduction

Chitinases were the first pathogen-related proteins whose function was identified [183, 184]. Three chitinases were among the second group of characterized PR-proteins, isolated from hypersensitively reacting virus-infected tobacco leaves [185]. The proteins designated by van Loon and colleagues (1987) as Tobacco P (27 kDa) and Tobacco Q (28 kDa) fall into the PR-3 family, while the two components of Tobacco R (13 and 15 kDa) belong to the PR-4 family. Later, a cucumber chitinase presenting lysozyme activity was designated as PR-8. The last identified PR-chitinase (PR-11) is less studied, and presents a distant homology to a bacterial chitinase [186].

Besides their extreme importance in self-defense against pathogens, plant chitinases are also constitutively expressed, and are induced by environmental stresses, such as wounding, frost, osmotic pressure (variation in salinity, drought), and chemical treatments [187].

4.1.1 Definition

The term "chitinase" designates enzymes that catalyze the hydrolysis of the β-1,4-N-acetyl-D-glucosamine (GlcNAc) linkages in chitin polymers [186]. This natural biopolymer is a key structural component of fungal cell walls and exoskeletons of invertebrates, such as insects and nematodes. GlcNAc is also a prevalent building block of bacterial peptidoglycan [188, 189]. Plant chitinases are endo-chitinases and have been shown to hydrolyze fully acetylated chitin, chitosan (partially deacetylated chitin), chitin oligomers (GlcNAc) of variable length and bacterial cell wall [190].

4.1.2 Antifungal activity

Although the primary identification of chitinases as PR-proteins was in virus-infected plants, their most studied feature in defense response is related to fungal pathogens.

The fungal cell wall is basically composed of carbohydrate polymers (chitin, β-1,3-glucan, β-1,6-glucan), glycosylated cell wall proteins, glycosylphosphatidylinositol, and a protein with internal repeats. The incorporation of chitin, β-1,3-glucan and β-1,6-glucan into the cell walls occurs in the tips of the growing hyphae [191]. Consequently the fungal cell growth is affected by chitin-binding proteins, chitinases and glucanases. The chitinases are able to inhibit fungal growth due to their ability to degrade recently incorporated chitin from hyphae that penetrate the plant tissues.

The antifungal capacity of chitinases is marked by two main functions. When the plant is invaded by a fungus, apoplastic acidic chitinases bind the hyphal cell wall and release GlcNAc oligomers from larger chitin molecules. These oligomers play a role in cellular signaling, working as elicitors, molecules that can bind to specific receptors and then trigger a signaling cascade for the activation of defense mechanisms. Amongst the activated defense mechanisms is the synthesis of new apoplastic and vacuolar chitinases. Vacuolar chitinases can then play a direct chitinolytic role in the invader organism, leading to the degradation of chitin molecules from the hyphal cell wall which prevents the fungal spread on subjacent tissues [187, 192, 193].

Due to the nature of the fungal cell wall, the antifungal effect of chitinases is synergistically potentiated by β-1,3-glucanases (PR-2) [194]. These proteins catalyze the hydrolytic cleavage of the 1,3-β-D-glucosidic linkages in β-1,3-glucans [195]. Simultaneous expression of a tobacco class I chitinase and a class I β-1,3-glucanase gene in tomato resulted in increased fungal resistance, whereas transgenic tomato plants expressing either one of these genes, but not the other, were not protected against fungal infection [196]. The synergistic effect occurs because the parallel degradation of two major structural components of the cell wall propitiates a more efficient inhibition of fungal growth.

The PR enzymes that are able to inhibit fungal growth through the degradation of structural carbohydrates can be thought of as an orchestra, where every single form of these proteins has a specific feature and they work together to perform their functions.

4.1.3 Substrate specificities

Brunner and colleagues (1998) studied the substrate specificities of ten tobacco chitinases, five of which were PR-3 members, three were PR-8, one was PR-4, and one was PR-11. Differential kinetics of chitin oligomer accumulation and degradation indicate that distinct chitinases have different cleavage specificities toward chitin and are capable of further processing the released oligomers. PR-3 had the most active isoforms on chitin and chitin oligomers. PR-4 and PR-11 showed lower activity against chitin, but performed rapid hydrolysis of chitin oligomers. Basic PR-8 isoforms were particularly efficient in inducing the lysis of bacterial cells, and the acidic isoforms were most active against larger oligomers (with 5 and 6 GlcNAc units) and 4-methyl-umbelliferyl-chitotriose, another typical lysozyme substrate. The authors suggested that the different chitinases act synergistically against their substrates, by distinct mechanisms, and play an important role in the concerted biochemical defense response against pathogenic attack [190].

In 2006, Sasaki and colleagues examined the sugar recognition specificities of rice PR-3 and PR-8 by nuclear magnetic resonance spectroscopy. Their experiments confirm that PR-3 is likely to hydrolyze GlcNAc polymers (chitin) of the chitinous components of the fungal cell wall, while PR-8 may not be specific to consecutive GlcNAc sequences, but rather act toward GlcNAc-containing glycolipid or glycoprotein, producing or degrading signal molecules important to other biological processes not directly involved in pathogenesis. However, PR-8 could recognize and hydrolyze chitosan (partially deacetylated chitin) in the cell wall of mature hyphae, producing elicitor compounds and participating in the defensive action against pathogens [197].

4.1.4 Multi-domain structure

A remarkable feature of many chitinases is a characteristic domain architecture. Members of the PR-3 and PR-4 family are categorized by the absence or presence of one or two chitin-binding domains preceding the catalytic domain. This particular chitin-binding domain is named after hevein, a 43 amino acid polypeptide present in the latex of *Hevea brasiliensis* (rubber tree) [198, 199].

Chitin-binding domains (CBD) have a common structural motif of 30 to 43 residues that are rich in glycines and cysteines organized around a highly conserved four-disulfide core [188, 198]. They possess the ability to reversibly bind to chitin, but are devoid of chitinolytic activity. Plant proteins that possess at least one non-catalytic domain that reversibly binds to a specific mono- or oligosaccharide can be defined as lectins [200]. The hevein domain found in chitinases is structurally homologous to the repetitive domain that composes plant agglutinins, and to the single chitin-binding domain of Ac-AMP antimicrobial peptides. Hololectins (lectins comprising exclusively hevein domains) are able to promote the agglutination of glycoconjugates (carbohydrates covalently linked with other chemical species) on bacterial and fungal surfaces, and they can interact with chitin in the peritrophic membrane that lines the intestinal tracts of herbivorous insects, possibly inhibiting absorption of nutrients [201, 202]. It has been demonstrated that the agglutinating properties are conferred by the duplicated chitin-binding domains [202]. Therefore, even lacking agglutination properties and without any known enzymatic activity, the single domain Ac-AMP also presents antifungal, antibacterial and antinutrient activity in insects [203].

The "chitin-binding chitinases" are chimerolectins, which consist of one or more N-terminal hevein domains fused to a C-terminal chitinase domain [200]. The latex hevein of the rubber tree is responsible for the coagulation of isoprene monomers by cross-linking the rubber particles via sugar linkage. Latex coagulation seals wounds and it is a vital defense mechanism whenever the plant is exposed to physical injury [149]. Apart from coagulating rubber particles, hevein can protect the plant using its affinity to chitin. The rubber tree produces hevein as a larger pre-pro-hevein of 204 amino acid residues, containing an N-terminal signal peptide and a small C-terminal targeting signal, which are both cleaved off during protein maturation. The mature pro-hevein is a PR-4 chitinase, formed by the hevein domain and the chitinase catalytic domain (barwin domain). In the formation of latex, in the lutoids (vacuole of the lactifers, cells which produce the latex), both domains are separated and the barwin domain is probably degraded by proteolysis, given the large molar excess of hevein (30:1) in the lutoid-body fraction of rubber latex [204, 205].

The hevein domain alone has demonstrated antifungal activity in vitro, but to a lesser extent than the conjugated pro-hevein [206]. Hevein may interfere with fungal growth by binding or cross-linking newly synthesized chitin chains. Van Parijs and colleagues (1991) speculated that the antifungal properties of hevein are somehow related to its particularly small size. The CBD monomer is small enough to cross the fungal cell wall and reach the plasma membrane, where it may have an effect on the active sites that are involved in cell wall morphogenesis.

Taira and colleagues (2002) studied the differences between PR-3 class I (RSC-a) and class II (RSC-c) rye seed chitinases (respectively with and without hevein domain) in binding to the fungal cell walls, to understand why the basic class I chitinase inhibited fungal growth more effectively than basic class II chitinase did. They reported that the C-terminal catalytic domain of RSC-a acted more effectively on the old hyphae than its class II homologous, and proposed that this was due to differences in charge. The conservation of hydrophobic residues in the chitin-binding domain could indicate that the binding ability of CBD resides in a hydrophobic interaction. Following these observations they suggested that basic class I chitinase binds to hyphal tips, lateral walls and septa, mainly by ionic interaction of the C-terminal catalytic domain, but also by hydrophobic interaction of the chitin-binding domain, being able to degrade mature chitin fibers as well as nascent chitin by its hydrolytic action. Basic class II chitinase, which lacks the CBD, binds only to the hyphal tip (by ionic interaction), and it is only able to hydrolyze newly incorporated chitin [207].

Hevein domains show an extended binding site that is perfectly pre-organized for the exclusive recognition of linear β-1,4-linked GlcNAc polymers when they adopt a conformation close to the native structure of the peptide. Five to seven GlcNAc units in the ligand are required to express all the possible protein-sugar interactions simultaneously. The different GlcNAc units of the polysaccharide exchange their positions at the binding site by interacting with different protein subsites in a dynamic process [188].

The chitin-binding domain found in members of the PR-3 family is homologous to the hevein of PR-4 proteins. The C-terminal catalytic domain of PR-3 presents, on the other hand, no amino acid sequence similarities with the PR-4 C-terminal domain. It has been suggested that the hevein domain was incorporated by some form of genetic transposition in the structure of a common ancestral gene [208].

4.1.5 Classification and nomenclature

Different classifications and nomenclatures are simultaneously used for plant chitinases.

4.1.5.1 Glycoside hydrolases

Carbohydrate-active enzymes that hydrolyze the glycosidic bonds are classified in glycoside hydrolase families [209]. To date there are 114 families of glycoside hydrolases classified in the CAZy database (Carbohydrate Active Enzymes database, accessed on February 2nd 2009, *http://www.cazy.org/fam/acc_GH.html*) [210]. Three glycoside hydrolase families are among the PR-families. Glycoside hydrolase family 17 (GH17) is represented by the PR-2 family (β-1,3-glucanases). In turn, the glycoside hydrolases of families 18 (GH18) and 19 (GH19) are chitinases (EC 3.2.1.14) [195]. The PR-3 family is a member of GH19, whereas PR-8 and PR-11 are classified as GH18. Interestingly, PR-4 has not been assigned to any of these families [186].

Glycoside hydrolases present two basic mechanisms of action, which result in either an overall retention or inversion of the configuration at the anomeric carbon of the sugar ring undergoing hydrolysis. The enzymes from families GH17 and GH18 operate through retention of configuration, while GH19 members are inverting enzymes [209, 211]. There are no sequence similarities between the three families and the three-dimensional structure of GH19 chitinases is also completely different from the others. GH17 and GH18 proteins have an 8-fold β/α-barrel structure, while GH19 proteins contain predominantly α-helices [211–213].

4.1.5.2 Protein classification

Following the classification presented by Neuhaus (1999), plant chitinases are divided into classes assigned by roman numerals. Primordial attempts to classify the chitinases distributed them in classes on the basis of sequence patterns [192, 208]. Class I chitinases possess an N-terminal chitin-binding domain and a highly conserved catalytic domain. Some members also have a short carboxy-terminal extension. Class II shares the catalytic domain with class I, but does not have the chitin-binding domain. Class III chitinases are completely unrelated to the former classes: they have higher homology to fungal chitinases than to other plant chitinase classes and they lack the CBD [214]. Classes IV, VI and VII are distantly homologous to classes I and II, and are characterized by particular deletions and duplications. The numeral V has been used simultaneously for different protein classes (namely class V PR-3 and class V PR-11) and this may cause some confusion [186]. In 1994, Neuhaus classified a stinging nettle (*Urtica dioica*) lectin, which had two CBDs and a catalytic domain homologous to class I, as a class V chitinase [186]. At the same time, Melchers (1994) assigned a new class of tobacco chitinase with sequence similarities to bacterial exochitinases as class V [215, 216]. Classes I, II, IV, V[1], VI and VII belong to family 19 of glycoside hydrolases and are PR-3 members, whereas classes III and V[2] belong to family 18 and are respectively PR-8 and PR-11 [186, 187, 195]. PR-4 has its own classes I and II. Class I PR-4 has a hevein domain, whereas the PR-4 class II lacks the CBD [186].

4.1.5.3 Gene names

PR-3 genes can be named using the acronym *Chia* plus a number relating to the class it belongs to. Class I PR-3 genes are named *Chia1*, class II, *Chia2*, class IV, *Chia4*, and so on. *Chib1* designates PR-8 genes, *Chic1* is for PR-11, and PR-4 genes are named *Chid1* and *Chid2* [186].

4.1.5.4 Isoelectric properties

To complement this complex nomenclature scheme, chitinases from all classes can be characterized by their isoelectric properties. Acidic chitinases are found among classes I, II, III, IV and VI and are mainly secreted to the apoplast. Classes I, III and VI also contain basic chitinases, which are located in vacuoles [187].

[1] Neuhaus, 1994

[2] Melchers, 1994

4.2 "Chitin-binding chitinases"

4.2.1 PR-3

Pathogenesis-related proteins of family PR-3 are present in the entire plant kingdom and are characterized by a conserved catalytic domain. The proteins are formed basically by α-helices and assume a globular structure with a catalytic groove [186].

They typically present an N-terminal signal sequence, followed by a chitin-binding domain linked to the catalytic domain by a hinge region of variable length, rich in glycine, proline and threonine (Gly/Pro/Thr-rich). The catalytic domain is ca. 240 amino acids long, in the class I members. Vacuolar forms of these chitinases also present a short (7 to 15 AA) C-terminal pro-peptide, required for targeting the protein to the vacuole.

Initially, class I and II chitinases were differentiated according to the presence or absence of the chitin-binding domain. Class II chitinases lack the chitin-binding domain and the N-terminal signal peptide is directly followed by the catalytic domain. However, two distinct groups are commonly classified as class II chitinases, and they are probably paraphyletic. The original class II chitinases, prototyped by tobacco chitinases P and Q, have a unique catalytic domain, presenting an important deletion relative to the second loop of class I chitinases. The barley *CHI2* gene, previously classified as class II because of the lack the hevein domain, does not present the characteristic deletion of the loop 2 region, and its catalytic domain is much more similar to those of class I sequences. Sequences presenting these characteristics will hereafter be referred to as Ib.

Class IV chitinases are characterized by four large deletions, all in loop regions, one being in the CBD and three being in the catalytic domain (Fig. 4.1). The complete deletion of loop 1 in the catalytic domain probably causes the loss of a sugar-binding subsite, whereas the loss of loops 3 and 4 reduces the volume of the protein [186]. The class V chitinases of PR-3 were first described in stinging nettle lectin. This protein has two hevein domains followed by a catalytic C-terminal domain homologous to class I and class II chitinases. The protein isolated from sugar beet (*Beta vulgaris*) classified as class VI presents a truncated CBD, which lacks four out of eight cysteines, and possesses a proline-rich spacer sequence of 131 amino acids [217]. Finally, class VII denominates a class IV homolog found in rice that lacks the hevein domain [186].

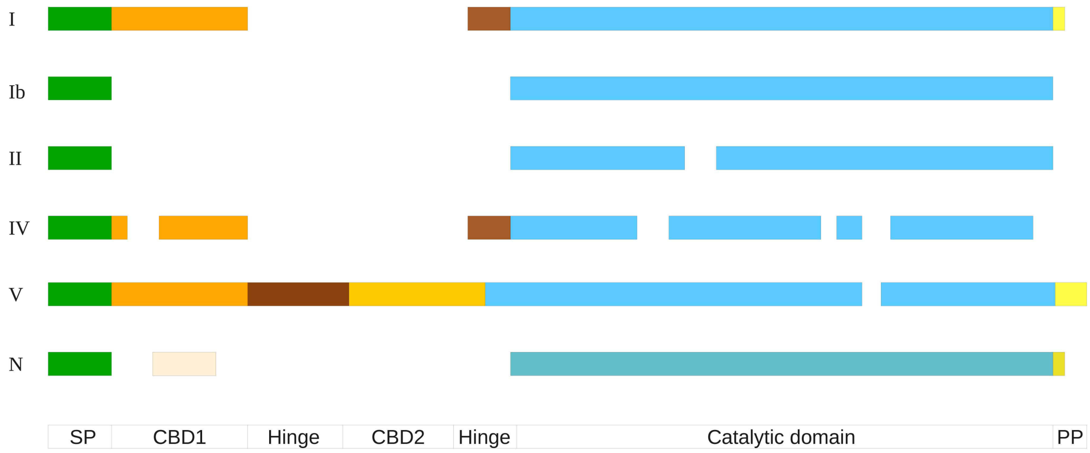


FIGURE 4.1: Scheme of chitinase classes I, Ib, II, IV and V. Signal peptide (green), chitin-binding domain (orange), hinge region (brown), catalytic domain (blue), and C-terminal pro-peptide (yellow). N refers to the non-hydrolytic chitinase-like genes.

Recently, Nakazaki and colleagues (2006) proposed a classification of the PR-3 genes based on the structure of the catalytic domain. Chitinases with no deletion, one deletion and four deletions are respectively classified as class I, II and IV. Class I and class IV are further subdivided according to the presence (+) or absence (−) of the hevein domain [218]. Following this criterion, class Ib chitinases would be included in class I(−), which is coherent with their evolutionary history. In the same way, class VII members would be shrewdly classified as class IV(−) chitinases. The authors did not mention the class V chitinases, since they had not yet been found in rice, the organism used in their studies.

SEED dataset

In a first approach, 37 sequences of PR-3 family members were selected from the UniProtKB/Swiss-Prot database [44] to characterize this family. The dataset has representatives of 24 plant species, five of which are monocotyledons, with seven sequences, and 19 eudicotyledons, with 30 sequences. Among the eudicotyledons, 10 sequences are from Solanaceae.

The SEED dataset is composed of 19 class I (three without CBD), six class II, and five class IV members. Since only one class V member is represented in Swiss-Prot, six class V sequences from UniProtKB/TrEMBL were added to the dataset. Figure 4.2 shows the SEED alignment and two representations of the corresponding tree.

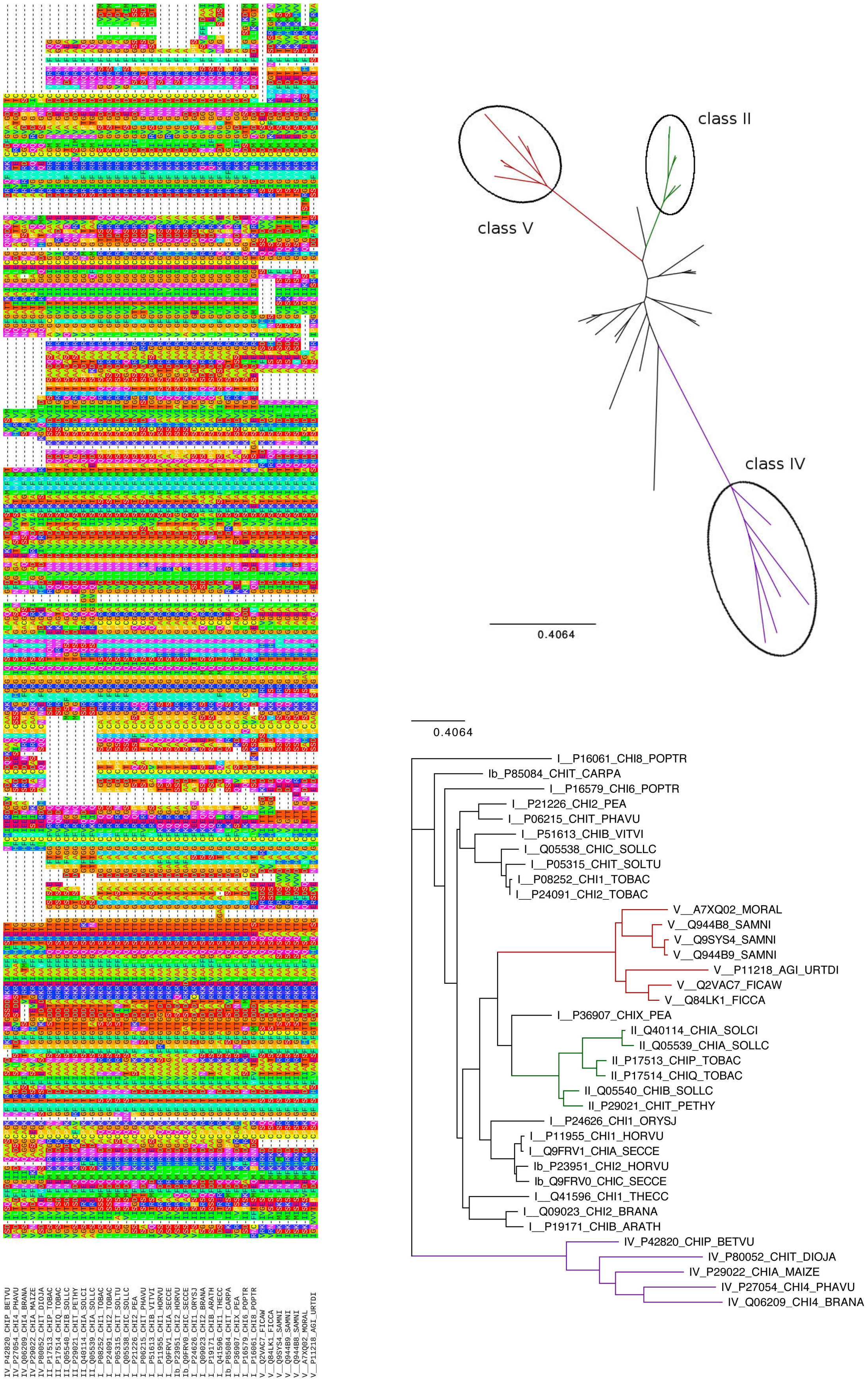


FIGURE 4.2: PR-3 SEED dataset. Left: Alignment of the catalytic domains. Right: Trees representing the SEED alignment. Classes I and Ib have black branches; classes II, IV and V have branches colored in green, magenta and red, respectively.

EXTENDED dataset

The alignment of the catalytic domains of the first set of sequences was used as SEED to build a profile HMM with HMMER [80]. This profile HMM was used to search the plant section of the UniProtKB database [44] for PR-3 homologs. Around 700 sequences were matched, respecting the E-value cutoff (10.0).

After filtering the retrieved entries, an EXTENDED dataset of 181 sequences was created. The dataset contained representatives of 83 plant species: 59 eudicotyledons, 15 monocotyledons, two magnoliids, five coniferales, and two bryophyta. Interestingly, one sequence from *Phytophthora infestans* (potato late blight fungus), an oomycete, was also included in the EXTENDED dataset.

The alignment revealed that 102 of the 181 sequences belong to class I chitinases, 87 presenting CBD and 15 sequences, classified as class Ib, lacking the chitin-binding domain. Seventeen true class II sequences are identified by the deletion in the loop 2 region. Class IV is represented by 46 sequences, seven of which lack the CBD. The *P. infestans* sequence is included in the second group. Nine sequences belong to class V.

Seven sequences included in the EXTENDED dataset could not be assigned to any of these classes. These sequences, labeled with an “N” in the alignment, have a truncated putative CBD, with four cysteines, that slightly resembles the hevein domain and is barely alignable to these. The catalytic domain also differs substantially from the other classes. Although the description coincides to that of class VI, the alignment to the mentioned sugar beet sequence does not support this classification.

An extensive research conducted by Zhang and colleagues (2004) describes the unclassified sequences as a new group of (non-hydrolytic) chitinase-like genes (CTL) [219]. They suggest that a substitution of a glutamic acid catalytic residue by a lysine is responsive to the lack of hydrolytic activity. Substitutions of non-similar amino acids (glutamine to proline and serine to tyrosine) at two functional sites may contribute to this phenotype. The ability to bind chitin appears to be preserved in the CTL proteins though the lack of four cysteine residues, compared to the hevein domain.

All class I sequences are from flowering plants (angiosperms, Magnoliophyta), and are represented in virtually all species in this group. Sequences without CBD are not restricted to cereals, since there are examples in Apiaceae, Ericaceae, Solanaceae, Cucurbitaceae, Fabaceae, Caricaceae, Rutaceae and Rosaceae. The interesting observation is that they appear to have originated independently. The signal peptides of class Ib chitinases, as well as their catalytic domains, are strikingly similar to their class I paralogs in the same species. This can be illustrated by the cereal chitinases HORVU_CHI2 and SECCE_CHIC, respectively paralogous to the class I chitinases HORVU_CHI1 and SECCE_CHIA (Fig. 4.2). They were probably originated by gene duplication in the common cereal ancestor, where one copy had lost the chitin-binding domain.

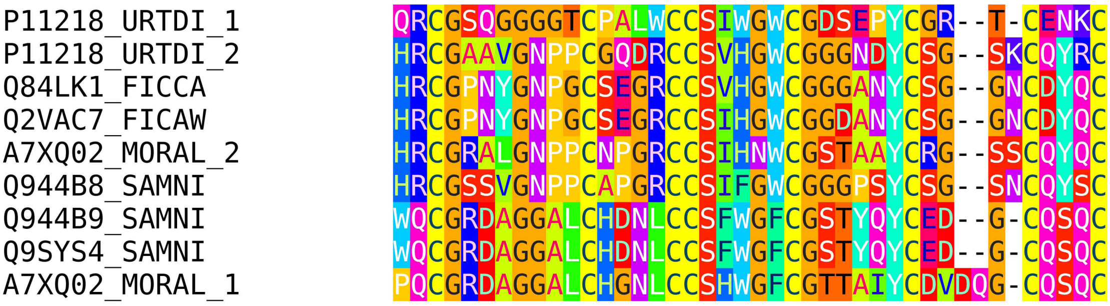
In the database annotation from UniProtKB [44], the class II prototype members, tobacco P and Q, are indicated as *Chitinase class I subfamily* in field *Protein family*, while their homologs in *Solanum* are correctly classified as *Chitinase class II subfamily.* The class Ib members can also be found under the class II designation. Class IV members are classified together with class I.

Class II sequences were found not only in angiosperms, but also in gymnosperms (other seed plants). Two sequences of Coniferales (Q6E6M9_PICAB from *Picea abies*, the Norway spruce, and O04276_PINST from *Pinus strobus*, eastern white pine) lack the loop 2 region. These observations suggest that either loop 2 was independently lost in the course of evolution, or class II also originated before the radiation of flowering plants.

Analogously, the presence of class IV chitinases in Bryophyta and Coniferopsida supports the hypothesis of an ancient origin of this class of chitinases. The location of the deletions is conserved throughout the whole plant kingdom. The chitinase sequence in oomycetes, however, most probably originated via horizontal gene transfer (HGT) (see [220]).

Class V, on the other hand, has only been isolated from eudicotyledons. Sequences from *Ficus awkeotsang* (Q2VAC7_FICAW) and *Ficus carica* (Q84LK1_FICCA), previously classified as class I, are homologous to the stinging nettle lectin (P11218, AGI_URTDI). They possess however only one hevein domain, with higher similarity to the second CBD from AGI_URTDI. A *Sambucus nigra* hevein-like fruit protein (SN-HLPf) was described by Van Damme and colleagues (1999) as a chimeric protein consisting of a hevein domain closely related to that of PR-4 and the C-terminal domain of class V chitinases. Two sequences of SN-HLPf are included in this dataset (Q9SYS4_SAMNI and Q944B9_SAMNI). Another class V chitinase from *Sambucus nigra* (Q944B8_SAMNI), submitted by the same author three years later, presents a hevein domain with 72% identity (81% similarity) to the second CBD from stinging nettle lectin.

The last sequence identified as class V (A7XQ02_MORAL) is a 415 amino acid-long mulberry (*Morus alba*) latex protein, named MLX56 [221]. The inclusion of this sequence in this dataset throws light on the question of the different CBDs found in this group. The mulberry latex protein has two CBDs, which are separated from each other by a 53 amino acid-long proline-rich sequence. The first CBD of this protein (Fig. 4.3, A7XQ02_MORAL_1) is undoubtedly homologous to the hevein domain of the presumed chimeric proteins, with 77% identity (82% similarity) to Q9SYS4_SAMNI and Q944B9_SAMNI. The second CBD of the mulberry latex protein (Fig. 4.3, A7XQ02_MORAL_2) shows indeed 68% identity (78% similarity) to Q944B8_SAMNI, and 65% identity (73% similarity) to the second CBD of the stinging nettle lectin (Fig. 4.3, P11218_URTDI_2). This observation suggests that the ancestor of the *Sambucus nigra* hevein-like fruit proteins could have carried both domains, similar to the mulberry latex protein, and differentially lost one CBD after duplication.



```
P11218_URTDI_1     QRCGSQGGGGTCPALWCCSIWGWCGDSEPYCGR--T-CENKC
P11218_URTDI_2     HRCGAAVGNPPCGQDRCCSVHGWCGGGNDYCSG--SKCQYRC
Q84LK1_FICCA       HRCGPNYGNPGCSEGRCCSVHGWCGGGANYCSG--GNCDYQC
Q2VAC7_FICAW       HRCGPNYGNPGCSEGRCCSIHGWCGGDANYCSG--GNCDYQC
A7XQ02_MORAL_2     HRCGRALGNPPCNPGRCCSIHNWCGSTAAYCRG--SSCQYQC
Q944B8_SAMNI       HRCGSSVGNPPCAPGRCCSIFGWCGGGPSYCSG--SNCQYSC
Q944B9_SAMNI       WQCGRDAGGALCHDNLCCSFWGFCGSTYQYCED--G-CQSQC
Q9SYS4_SAMNI       WQCGRDAGGALCHDNLCCSFWGFCGSTYQYCED--G-CQSQC
A7XQ02_MORAL_1     PQCGRDAGGALCHGNLCCSHWGFCGTTAIYCDVDQG-CQSQC
```

FIGURE 4.3: Alignment of the chitin-binding domains of class V chitinases. Alignment performed with MAFFT-L-INS-i [121] and colored by DNATagger [55].

Like the stinging nettle lectin, the BjCHI1 (Q9SQF7_BRAJU) from *Brassica juncea* is a chitinase with two chitin-binding domains. This protein is however not related to class V, but to class I chitinases. The high identity (97%) of both hevein domains from this protein suggests a recent internal duplication as the origin of the second domain.

The alignment of the full-length sequences of the PR-3 EXTENDED dataset is presented in Figure 4.4. It illustrates the homology between the different classes, and includes the chitin-binding domains (CBD and CBD 2). This alignment was not used in the phylogenetic analysis.

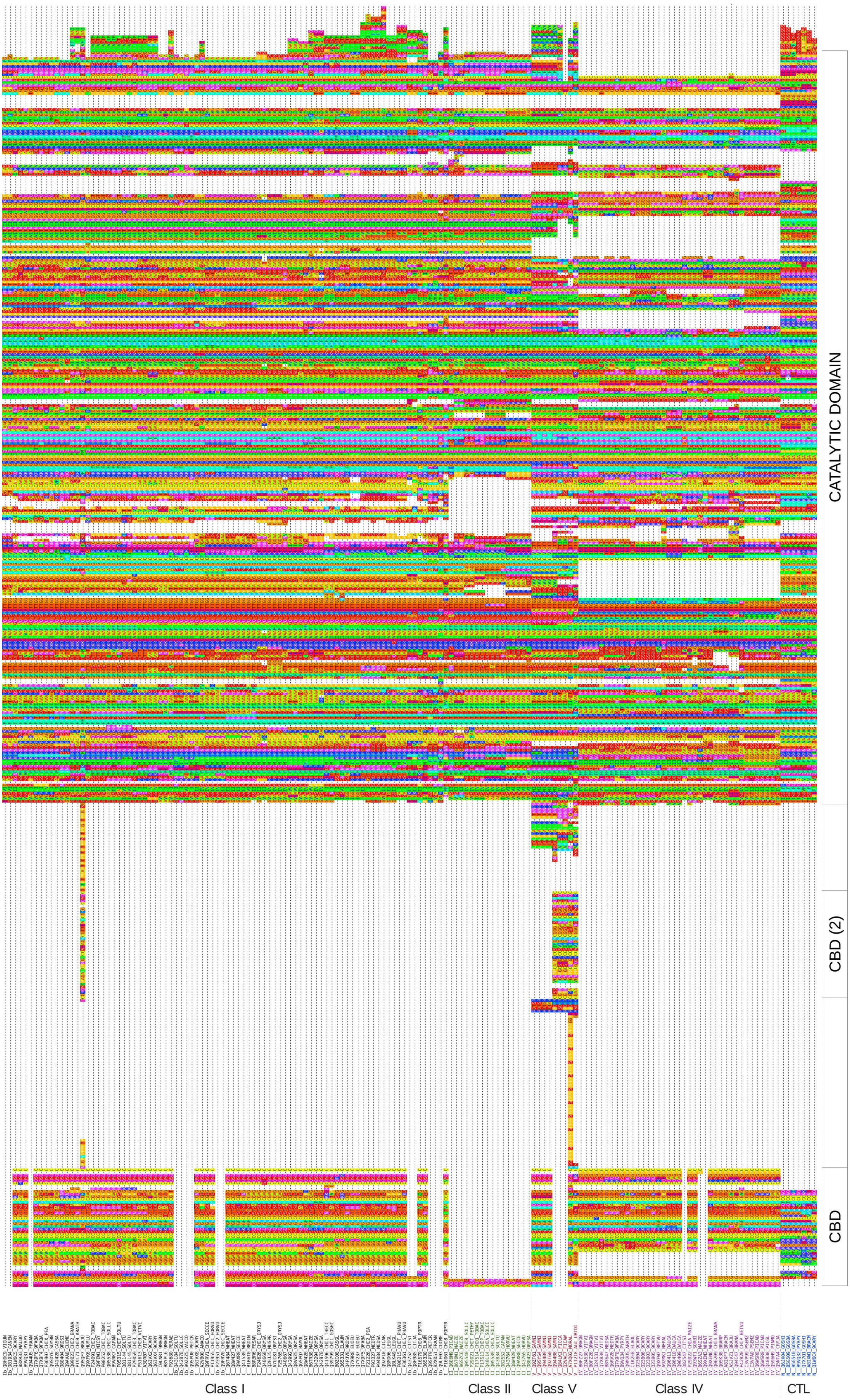


Figure 4.4: Alignment of the full-length sequences of PR-3 EXTENDED dataset. (Note: This is not the EXTENDED alignment.)

Phylogenetic analysis

The EXTENDED dataset was aligned (see chapter 3 for details) and the corresponding coding alignment was retrieved. Four sequences had no corresponding nucleotide sequences. An alternative alignment was performed using nucleotide models, and the results compared to the amino acid based alignment. Sequences presenting clear indication of frameshift mutations or sequencing errors were excluded from the dataset.

The final dataset was composed of 157 sequences, distributed in 85 class I (14 class Ib), 16 class II, 39 class IV (3 without CBD), 9 class V and 7 CTL sequences. Figure 4.5 shows two representations of a tree generated with the full-length amino acid sequences. It can be observed that the CTL members form a tight cluster, quite divergent from the other classes, and can be certainly used as an outgroup.

As expected, the members of class IV chitinases form a clearly separate branch. Similarly, all proteins identified here as class V chitinases cluster together when the catalytic domain is in the analysis. Later in this section we will see that the same is not valid for their chitin-binding domains. Class II, in turn, arrives from a neighboring subtree. The class II sequence from *Picea abies*, however, does not cluster in this group, but is located near the rooting point. Class I has the shortest branches and does not form an independent cluster; on the contrary, class II and class V clusters arrive between class I members. Class Ib members do not diverge from those containing CBDs in class I. Rather, they are spread between class I members.

Selection

The selection analysis was performed with the alignment region corresponding to the catalytic domain. The coding alignment, as well as the amino acid alignment, were used to construct phylogenetic trees under the evolutionary models determined by model selection. Both trees were given as the usertree for evaluation under codon models. Branch lengths were re-estimated and the tree with maximum likelihood was chosen as the initial tree for selection analysis.

Under site models, the discrete model (M0 with 3 categories of ω rates) and the neutral models with β-distribution were indicated as best-fit models. No indication of positive selection could be detected using site models for the entire tree in the complete dataset.

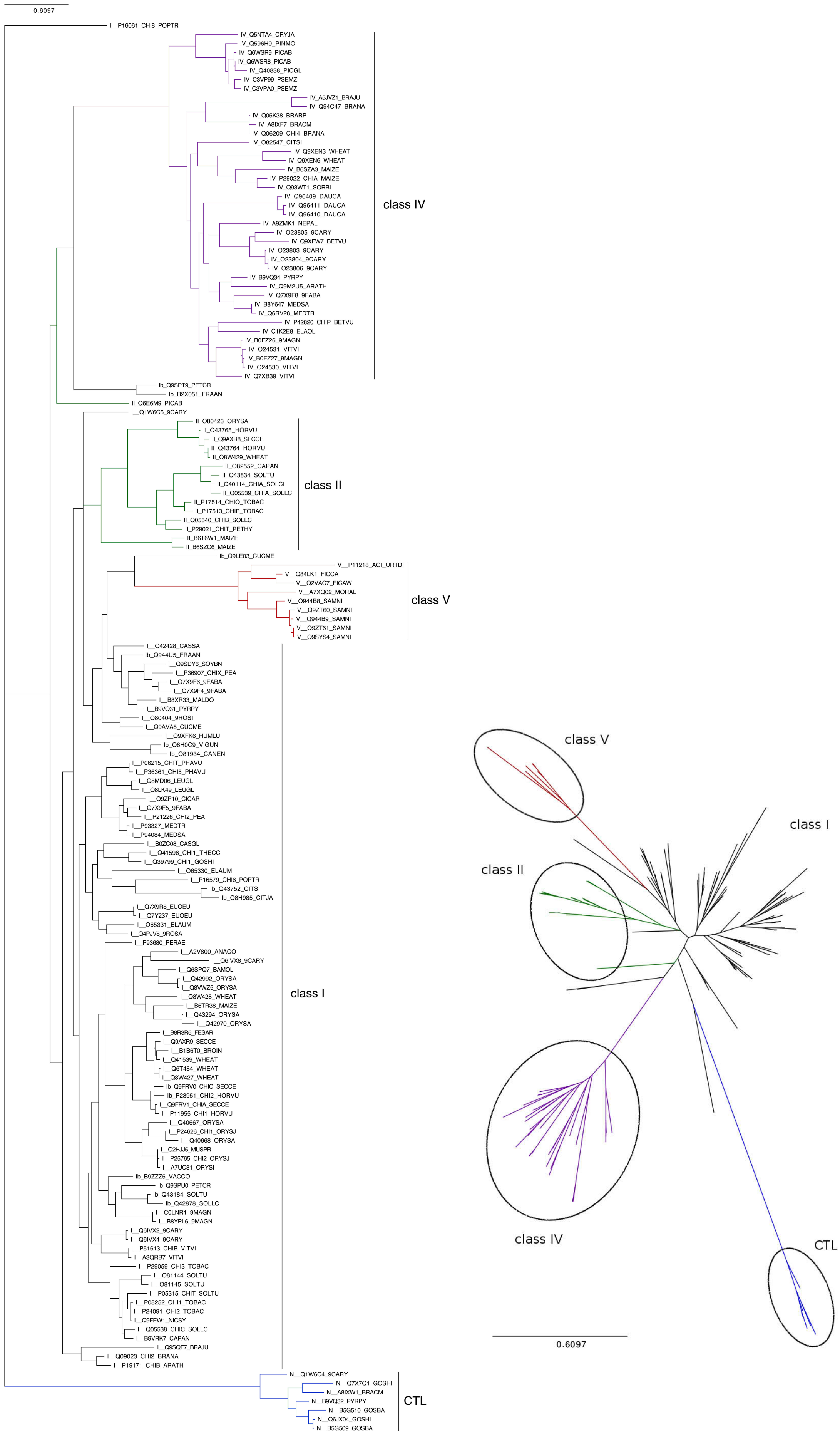


Figure 4.5: Trees of 157 chitinases from GH19 family. Chitinase classes I and Ib (black branches), classes II (green), IV (magenta) and V (red), and CTL (blue).

4.2.2 PR-4

Another chimerolectin among the PR-proteins is PR-4.

The first characterized PR-4 member was Tobacco R, isolated by van Loon and colleagues in 1987 [185]. Years later, CBP20, a tobacco protein with low endochitinase activity was also isolated from leaves inoculated with tobacco mosaic virus [222]. Comparing to Tobacco R, CBP20 possesses a CBD preceding the homologous C-terminal domain. The C-terminal domain of PR-4 proteins has no homology to other known chitinases, nor has it been classified as a glycoside hydrolase. The lack of an obvious catalytic cleft or catalytic site suggested that the chitinase activity of CBP20 was dependent on the chitin-binding domain [186].

The functions and possible catalytic activities of the C-terminal domain remained unknown until Caporale and colleagues (2004) demonstrated that wheatwin1, a wheat PR-4 without chitin-binding domain, has ribonuclease activity. That was the first report of a PR-4 protein with RNase activity. Wheatwin1 showed enzymatic and antifungal activities, degrading wheat coleoptil RNA and inhibiting the hyphal growth and spore germination of *Fusarium culmorum.* In their opinion, PR-4 proteins should not be considered as belonging to the chitinase superfamily, but as possible constitutive members of a distinct protein group. In fact, wheatwin1 has been shown to possess ribonuclease activity [even though its mechanism of action is different from that of PR-10 proteins]. It is likely that PR-4 proteins operate on the invading pathogen by a translation-inhibitory process that could be ascribed to their ribonuclease activity. Therefore, this is probably not the only activity related to its biological function [223].

CBP20, as well as the wound-induced proteins from potato (Win1 and Win2), is homologous to the pro-hevein of *Hevea brasiliensis.* The rubber tree synthesizes a pre-pro-hevein of 204 amino acids containing an N-terminal signal peptide, a CBD domain, a catalytic domain and a C-terminal targeting pro-peptide. The 17-residue signal peptide is removed by post-translational processing, producing a hevein precursor of 187 residues. The C-terminal pro-peptide, which is necessary for vacuolar targeting, is cleaved off during protein maturation. The mature hevein precursor has 173 amino acids and a molecular mass of 20 kDa. This precursor can be further cleaved between residues 49 and 50, where the hevein (4.7 kDa) is separated from the catalytic domain

(14 kDa). At the end, four to six residues are removed from the C-terminus of hevein [205]. Both forms, the processed and the unprocessed precursor, are observed.

One protein found in barley was named Barwin due to its relationship to the C-terminal region of the wound-induced proteins in potato plants [224]. This protein has indeed no chitin-binding domain. After the determination of its three-dimensional structure, the domain that characterizes the PR-4 was then referred to as the barwin domain. The structure of Barwin consists of a well-defined four-stranded antiparallel β-sheet, two parallel β-sheets packed antiparallel to each other, and four short α-helices [225, 226].

In analogy to the classification of PR-3, PR-4 proteins are defined as class II in the absence of the chitin-binding domain, while PR-4 proteins with CBD are referred to as class I [186]. These classes have, however, no relationship to class I and II chitinases.

SEED dataset

Seventeen sequences were selected from both UniProtKB databases [44] to represent the PR-4 protein family in the SEED dataset. The dataset has representatives of 12 plant species, of which five are monocotyledons with six sequences, and seven of which are eudicotyledons with 11 sequences. All sequences from monocotyledons belong to class II, and seven sequences from eudicotyledons present a hevein domain.

Note: A sequence from *Eutrema wasabi* (Q8H950_EUTWA) presented an unexpected internal duplication of 20 amino acids. It is worth noting that the duplicated sequence is identical to the presumable template, which could only be explained by a very recent duplication, in a region characterized as a helix-turn-helix. The article referenced in this entry shows nucleotide and deduced amino acid sequences without the duplication [227]. A careful observation of the database entry for the coding sequence reveals an interesting "line duplication" in the text file of the submitted sequence, and it is thus likely that it is due to an annotation error. We assumed that the sequence of Q8H950_EUTWA was erroneously registered in the database, and deleted this region of the sequence before further analysis.

Alignment

After the excision of the "duplicated" region of *E. wasabi*, the alignment of the selected sequences was 229 sites long. The alignment indicates a high overall conservation of the barwin domain for all PR-4 sequences, as well as of the hevein domains in class I proteins. In the barwin domain, 52 residues are identical in all sequences, 21 sites present conserved substitutions, and 13 show semi-conserved substitutions in the total of 126 aligned sites. In the hevein domain, 21 residues are identical in the seven class I sequences, eight sites present conserved substitutions, and four sites show semi-conserved substitutions. No gap had to be inserted to align the hevein domain, and only five gaps (four single gaps and one extended in one position) were opened in the barwin domain. In the SEED alignment, the signal peptide, hevein domain, and C-terminal pro-peptide were not included.

hmmsearch

In order to check the phylogenetic distribution of PR-4 homologous sequences stored in the public databases, a profile HMM was built with the SEED alignment (barwin domain). The *hmmsearch* revealed that this family is represented by only 15 sequences in UniProtKB/Swiss-Prot, with E-values ranging from e^{-89} to e^{-78} for the first 12 sequences (between 276 and 312) and E-values from e^{-68} to e^{-9} for three fragments (scores, 46 to 245). In the UniProtKB/TrEMBL plant dataset, 71 sequences matched the profile HMM. Excluding fragments, 58 hits present E-values ranging from e^{-89} to e^{-67} (scores from 241 to 312). A database search on the collection *nrprot* (NCBI non-redundant GenBank CDS translations + PDB + Swiss-Prot + PIR) [228] with E-value cutoff of 10.0 (default) matched 92 sequences, from which five were viral sequences (E-value ranging between 6.4 e^{-11} and 7.4 e^{-9}, score around 50), one sequence from Ascomycota fungi *Phaeosphaeria nodorum* SN15 (E-value 5.5 and score 1.0), two high scored sequences from gymnosperm *Picea sitchensis* (Coniferopsida) (E-value ca. e^{-74}, score 268), and 84 sequences from angiosperms. Eudicotyledons were represented in sequences from Solanales, Malpighiales, Vitales, Brassicales, Fabales, Rosales, Dipsacales, and monocotyledons had representatives in Poales, Dioscoreales and Asparagales. Most of the sequences that are not already represented in the SEED dataset are putative uncharacterized proteins or fragments of larger sequences.

EXTENDED dataset

After filtering, 26 sequences of PR-4 homologs were included in the EXTENDED dataset. This number was further reduced to 24 because one sequence was redundant (in the barwin domain region), while the search for the protein-coding DNA sequences retrieved no matches in nucleotide sequence for the entry P28814 (BARW_HORVU).

Phylogenetic analysis

In the trees reconstructed using the whole sequences as well the barwin domain, there is a clear separation between representatives of Liliopsida and eudicotyledons. The class I subtree seems to reflect the corresponding class II cluster in eudicotyledons, and Poaceae class II forms a tight branch. On the other hand, the location of the class II sequences from *Dioscorea bulbifera* (O48880_DIOBU), *Lycoris radiata* (B6EB12_LYCRD) and *Arabidopsis thaliana* (HEVL_ARATH) is uncertain, varying with the dataset and algorithm used (data not shown). Nevertheless, the analysis of the sequences and the trees suggests that the hevein domain has been introduced into the PR-4 in one insertion event, during the radiation of the eudicotyledons.

Selection

All likelihood ratio tests for positive selection were highly significative for PR-4. Model M2a (PositiveSelection) indicates a class with proportion 0.008 for $\omega = 5.548$, and the LRT rejected the null model M1a (NearlyNeutral) with $\alpha < 0.001$. Model M8 (beta&ω) presents a class with proportion 0.007 for $\omega = 3.920$. Both LRT, against M7 (beta) and M8a (beta&$\omega = 1$), were significant with $\alpha < 0.001$. Therefore, using information criteria (AIC, AICc and BIC), the model that best fits the data is the discrete model (M3), which has all categories of ω with values below one.

Posterior probabilities for site classes calculated with Bayes empirical Bayes (BEB), as well as naïve empirical Bayes (NEB) indicated site 102 (in the catalytic domain) as positively selected, with probability 0.999 for $\omega = 6.415 + -1.938$ (M2a BEB), $\omega = 5.548$ (M2a NEB), $\omega = 5.014 + -1.875$ (M8 BEB), and $\omega = 3.920$ (M8 NEB). In figure 4.6, site 102 is highlighted (black box) in the complete alignment of the PR-4 final dataset.

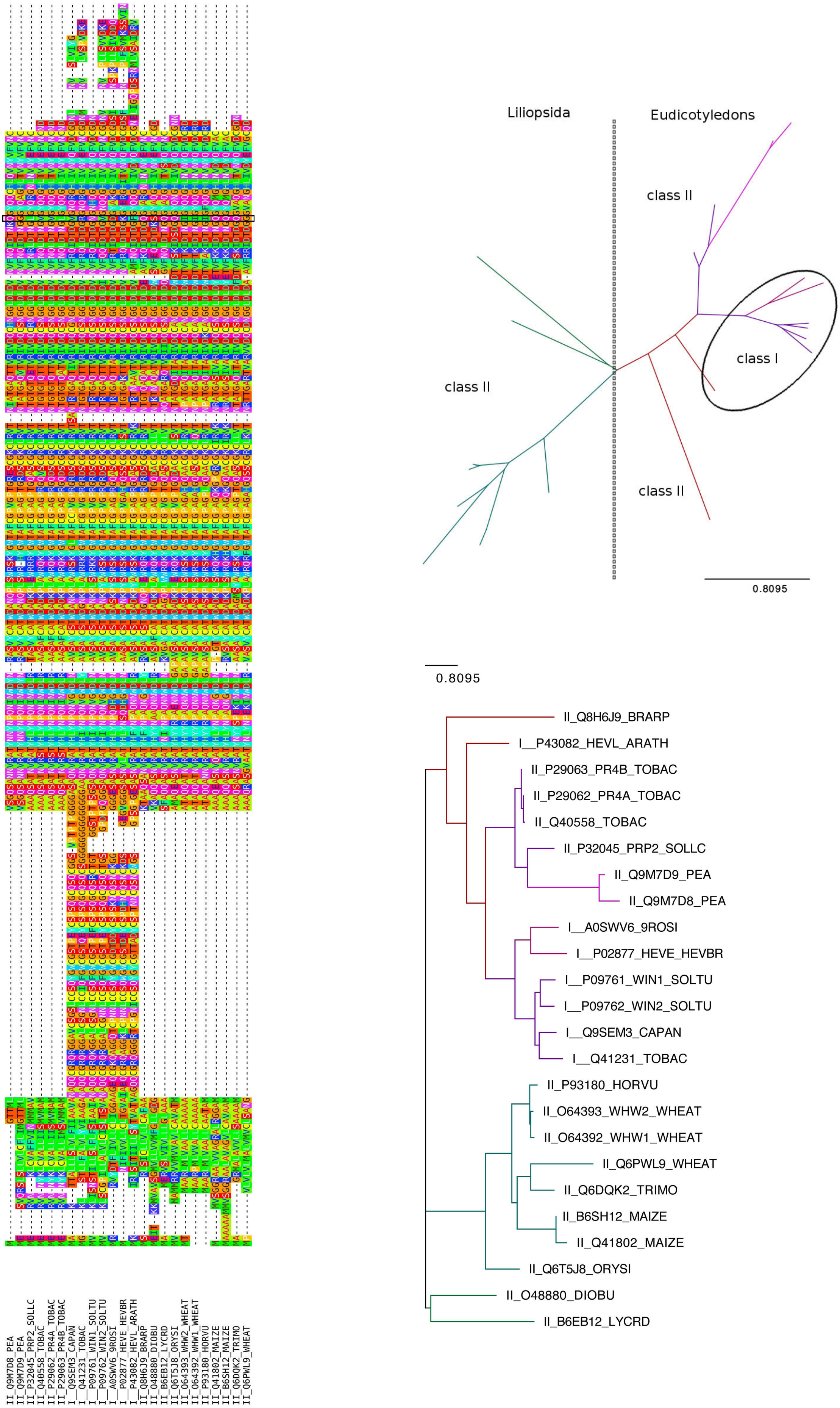


FIGURE 4.6: Final dataset of PR-4. Left: Alignment of the complete sequences. Right: Unrooted and rooted trees from the barwin domain.

4.2.3 CBD

Evolution of CBD

Chitin-binding proteins, or more specifically N-acetylglucosamine binding proteins containing the disulfide-rich domain, have been described by Raikhel and colleagues (1993) as proteins capable of reversibly binding to chitin. The authors observed that all chitin-binding proteins for which the amino acid sequence was known contain a common structural motif of 30 to 43 amino acids with several cysteines and glycines in conserved positions [198].

```
                +-------------+
           +----|------+      |
           |    |      |      |
xxCgxxxxxxxCxxxxCCsxxgxCgxxxxxCxxxCxxxxC
  |              |                |    |
  +--------------+                +----+
```

Figure 4.7: Schematic representation of the four disulfide bonds in chitin-binding domains. Adapted from PROSITE: PDOC00025

The best characterized chitin-binding domain is the wheat germ agglutinin (WGA), a wheat lectin which is a dimer of identical 171-residue polypeptides, each composed of four repetitive CBDs. WGA is therefore a hololectin, originated by tandem duplicative process [229]. Furthermore, many chitin-binding proteins are chimerolectins, consisting of one or more CBDs fused to different protein domains [199, 200]. The presence of highly conserved chitin-binding domains in different classes of proteins suggests that they arose from a common ancestral gene [198]. PR-3 classes I, IV and V, and PR-4 class I members, are chimerolectins [189, 230]. Shinshi and colleagues (1990) proposed that the chitin-binding domain was introduced into the coding region of an ancestral chitinase gene by a transposition event [208].

Dataset

The phylogenetic history of the chitin-binding domains found in PR-3 and PR-4 members was studied here. The aligned chitin-binding domains of 92 PR-3 and seven PR-4 sequences compose the dataset along with 15 lectin sequences: the antimicrobial peptide PN-AMP from *Ipomoea nil* (Japanese morning glory), two lectin sequences of *Phytolacca americana* (American pokeweed), the four domains of WGA from wheat, and their orthologs from barley and rice. The putative CBD domains of the CTL proteins were not considered here. After the exclusion of identical sequences, the dataset ended with 109 sequences, consisting of 56 class I, 24 class IV, 8 class V, 7 PR-4, and 14 lectin sequences.

Alignment

The length of the chitin-binding domains varies from 33 residues in PR-3 class IV members to 41 amino acids of the PR-4 hevein domains. The region of the CBD defined for this analysis begins in the first and ends in the last of the eight conserved cysteins, so that the two "glutamine-rich" columns at the beginning of the alignment were deleted. The reason is in the origin of these sites. In PR-3 class IV sequences, a large gap extended over seven positions. The alignment has 41 sites, 11 of which present gaps. Over the 30 positions without gaps, 15 sites present the same amino acid in at least 50% of the sequences, eight sites are invariant (7 cysteines and one glycine), one site presents only conserved substitutions (Y/F/W), and 2 sites show only semi-conserved substitutions (G/N). Nevertheless, the "variable" cysteine presents a single substitution (C/R) in a class IV grape (*Vitis vinifera*) chitinase transcript.

Phylogenetic analysis

The location of the different clusters in the tree is sensitive to alterations in the dataset composition. As expected, PR-3 class IV members form a distinct cluster in all trees calculated. The same is observed for PR-4 members. PR-3 class V sequences are dispersed in the tree. One cluster, composed by the first CBD of the mulberry latex protein (V_A7XQ02_MORAL_2) and a *Sambucus nigra* hevein-like CBD (V_Q944B9_SAMNI), is located between two PR-4 subtrees. The first CBD of the stinging nettle lectin (P11218_URTDI_1) has a very long branch, arising between class I members. Another cluster, which has the longest branches, contains all sequences related to the second CBDs of mulberry and stinging nettle.

hmmsearch

A profile HMM was built with the alignment of 129 chitin-binding domains from PR-3 and PR-4. The score reached by the sequences present in the seed alignment ranges from 48.2 (E-value 4.1 e^{-13}) to 84.8 (E-value 3.9 e^{-24}). An hmmsearch against the UniProtKB/SwissProt plant dataset indicates that the antimicrobial peptide AMP_IPONI (score 72.6, E-value 3.7 e^{-18}) is the next closest hololectin, followed by the second hevein domain from barley lectin AGI_HORVU (score 65.5, E-value 4.9 e^{-16}) and the other lectins used in the analysis. The databank search against the *nrprot* collection reveals that the next closest related non-plant CBD is from fungal proteins. Two sequences from *Aspergillus*, two from *Talaromyces* and one from *Penicillium* obtained an E-value in the order of e^{-11} (scores between 56.2 and 59.1) when compared to the profile HMM. When a cutoff of 10.0 is set, other fungal groups are included in the hits, along with seven matches in diatoms (Thalassiosiraceae and Naviculales) and one in Entamoeba.

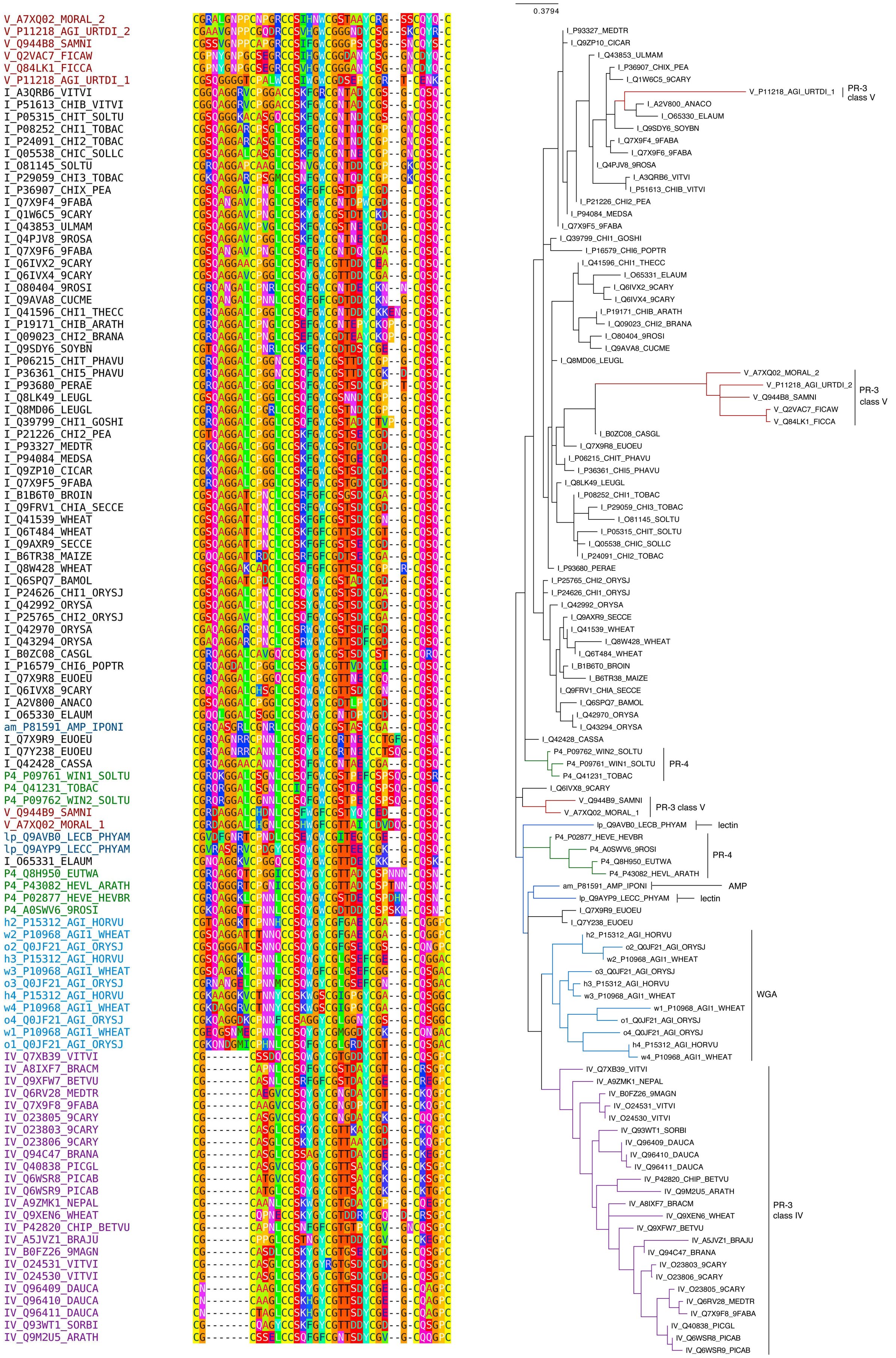


FIGURE 4.8: Chitin binding domains. Left: Alignment of 109 CBD from PR-3, PR-4, AMP, lectin and WGA. Right: Maximum likelihood tree.

4.3 Glycoside Hydrolase Family 18

The glycoside hydrolases family 18 is ubiquitous in the whole tree of life, present in archaea, bacteria, eukaryota and even viruses [210, 231]. A recent study on the GH18 family in animals relates them with innate and adaptive immunity in vertebrates and demonstrates that the phylogeny of animal GH18 genes is consistent with evolution of the family by a birth-and-death process as described for multigene families of the vertebrate immune system [158]. In animals, the GH18 family is divided into three major phylogenetic groups: chitobiases, chitinases/chitolectins, and stabilin-1 interacting chitolectins. In plants, besides PR-8 and PR-11, the GH18 family comprises non-catalytic proteins such as xylanase inhibitor, concanavalin B and narbonin [231].

Phylogeny of Plant Glycoside Hydrolase Family 18

Three hundred and ten sequences of glycoside hydrolases of family 18 from green plants were retrieved from the dataset collection of InterPro: IPR001223 (Glycoside hydrolase, family 18, catalytic domain). After the exclusion of extremely large sequences, small fragments and redundant sequences, 163 sequences were aligned to known members of PR-8 and PR-11 chitinases to identify the members of these families in the dataset. The dataset was further reduced to 52 sequences.

The tree constructed with IQPNNI reached the highest likelihood in comparison to the other methods used. The phylogenetic tree (in all methods used) revealed four long branches with well defined subtrees (Fig. 4.9). One branch leads to the subtree of PR-11 chitinases. A second branch bifurcates into two branches leading to a subtree with narbonins and a subtree with a different class of chitinases, which includes tulip bulb chitinase, referred to as the chitinase class II subfamily of the GH18 family. A third long branch connects PR-8 chitinases, xylanase inhibitor proteins (XIP) and a few diversely classified sequences identified as concavalin B (*Canavalia ensiformis*), yeldin (*Vigna unguiculata*), root vegetative storage protein, Indian jujube *Ziz m1* allergen (*Ziziphus mauritiana*), chitinase homolog and chitinase class III-like. A fourth long branch arises near the branch leading to PR-11. This branch contains two predicted chitinase sequences from the smallest free-living eukaryotes, the unicellular green algae *Ostreococcus tauri* and *Ostreococcus lucimarinus.* They belong to the Prasinophyceae, an early diverging class within the green plant lineage [232]. This supports *Ostreococcus* as a solid alternative to root the tree.

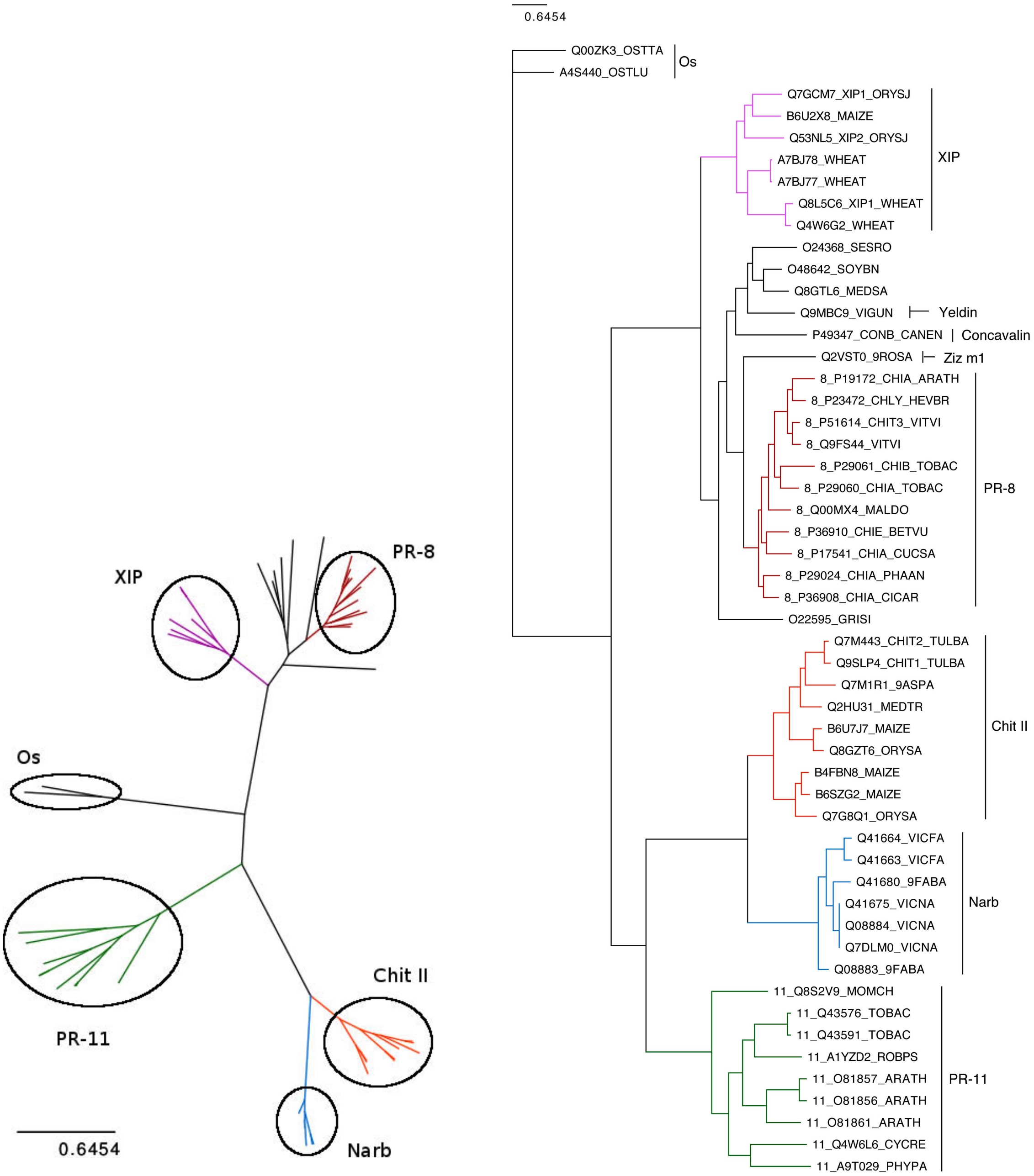


Figure 4.9: Unrooted and rooted trees of 52 GH18 family from plants. Clusters represent xylanase inhibitor proteins (XIP), PR-8, the chitinase class II subfamily of GH18 (Chit II), narbonins (Narb) and PR-11. The tree is rooted with *Ostreococcus tauri* and *Ostreococcus lucimarinus* GH18 chitinase sequences (Os).

4.3.1 PR-8

Bifunctional enzymes with both chitinase and lysozyme activity have been isolated from the latex of rubber tree, papaya and fig, homologous to a cucumber pathogenesis-related protein [233]. They have been assigned as class III chitinases and classified as the PR-8 family of pathogenesis-related proteins [186, 234]. Hevamine, a class III chitinase from the rubber tree, possesses chitinase and lysozyme activity and is important for plant defense against pathogenic bacteria and fungi [212, 213, 216, 233], while PR-8 is the most abundant chitinase upon viral infection in cucumber [235, 236]. Moreover, class III chitinases have higher homology to fungal chitinases than to other plant chitinase classes [214].

SEED dataset

Nine class III chitinases were collected to build the SEED dataset. All sequences belong to eudicotyledon species. Very similar in size, the sequences presented a signal peptide of 22 to 30 residues preceding a catalytic domain of ca. 270 amino acids. No C-terminal pro-peptide was observed. The aligned catalytic domain (SEED alignment) was used to construct four profile HMMs.

hmmsearch **and EXTENDED dataset**

In the UniProtKB plant dataset, the search for PR-8 homologs returned 221 whole sequence hits with global profiles and 268 hits with local profiles (including fragments). After filtering the retrieved entries, 89 sequences remained in the EXTENDED dataset. All sequences in the EXTENDED dataset are from seed plants, 67 being from eudicotyledon species and 22 from Liliopsida (11 alone from rice).

The EXTENDED dataset was aligned to the profile HMM (global-single) and the sequence regions relative to the signal peptide were deleted. The edited sequences were aligned with several programs, and the alignment produced by MAFFT L-INS-i was selected as the EXTENDED alignment. The aligned coding sequences were retrieved for all but one amino acid sequence. A further 11 sequences were excluded due to redundancy and reading frame alterations.

Phylogenetic analysis

The "final" alignment thus contained 77 sequences, with 327 aligned sites, of which 97 had gaps, 34 presented conservative substitutions, 16 presented semi-conservative substitutions, and only 15 sites were invariable (Fig. 4.10). This dataset, however, was not only composed by chitinase class III sequences, but also by xylanase inhibitor proteins, yeldin and yeldin-like sequences, and the allergen *Ziz m1*. Sequence and phylogenetic analysis showed that several sequences annotated as class III chitinase would also require a revision of nomenclature. Further on, there are still some very redundant sequences in this dataset due to the inclusion of isoforms. This can be solved by restricting the dataset to sequences annotated with "Evidence at protein / transcript level" (PE 1 and PE2).

A reduced dataset with 31 class III chitinase sequences was created and separately analyzed. The alignment of the PR-8 sequences had 294 sites, 58 being invariable, 48 with conservative substitutions, and 18 with semi-conservative substitutions. Only 40 positions presented gaps.

The trees calculated with 77 sequences (Fig. 4.11) present a cluster of xylanase inhibitor proteins emerging from the branch PR-8 of monocotyledon species (all XIPs belong to Poaceae species). This supports the hypothesis of very recent divergence after gene duplication [237]. Among the sequences found in these clusters, there are three rice sequences classified as "Class III chitinase homologue". XIP-type proteins present, indeed, no chitinase activity.

Similarly, another well defined cluster emerging from dicotyledon PR-8 sequences groups Yeldin and Yeldin-like protein with sequences classified as "Class III acidic endochitinase", "Class III chitinase-like" and "Chitinase homologue". These sequences, however, are described in connection with plant development, indicating that they can have thus the same function as Yeldin. Yieldin participates in the regulation of the acid-induced wall extension, but the exact mechanism has not yet been elucidated [238].

Attention is also drawn to the solitary branch leading to the Indian jujube allergen *Ziz m1*. It is a long branch that emerges from another subtree of dicotyledon PR-8 sequences, suggesting a further case of enhanced accumulation of substitutions. Whether these differences result in change in functionality remains to be clarified, although recombinant *Ziz m1* showed chitinase activity [239]. Indian jujube *Ziz m1* has been reported to be implicated in the latex-fruit syndrome. It is therefore not surprising that this branch neighbors Hevamine-A (CHLY_HEVBR) from rubber tree.

Selection

The aligned protein-coding sequences were used to infer site specific selection with codon substitution models. None of the datasets indicate positive selection acting on PR-8 proteins. Branch site models, however, were used to detect positive selection acting on the branches leading to the newly emerged proteins, derived from class III chitinases. Indeed, in the three cases, branch site models detected positive selection operating at specific sites. In the test using the branch leading to XIP sequences as foreground, 17 sites were inferred under positive selection (BEB, $\alpha > 0.95$). For Yeldin, as well as Ziz m1, BEB was significant for a single site in each case.

hmmsearch nrprot

A database search with the global-single profile HMM of PR-8 against the *nrprot* collection (NCBI non-redundant GenBank CDS translations + PDB + Swiss-Prot + PIR) by Mobyle Portal returned 695 hits satisfying an E-cutoff of 10.0, or 483 after the exclusion of negative scored hits. Half of the sequences belong to plants, but 167 hits represent fungal sequences, 14 are from animals (hydrozoans and anthozoans) and 30 are from bacteria. The next most related sequences are from fungi (zygomycetes and basidiomycetes). They are, however, distantly related to the plant class III chitinases. Using the E-cutoff of e^{-63}, only plant sequences are represented in the dataset, but it includes xylanase inhibitor proteins, Yeldin, *Ziz m1*, and concavalin B. The E-cutoff of e^{-122} restricts the dataset to 163 class III chitinases from flowering plants. The 125 eudicotyledon sequences are from Brassicales (25), Vitales (23),Fabales (22), Malpighiales (20), Caryophyllales (11), Cucurbitales (8), Solanales (5), Rosales (5), Gentianales (1), Lamiales (1), Asterales (1), Apiales (1), Malvales (1), and Fagales (1). The 38 sequences from monocotyledons are from Poales (36), Zingiberales (1) and Dioscoreales (1).

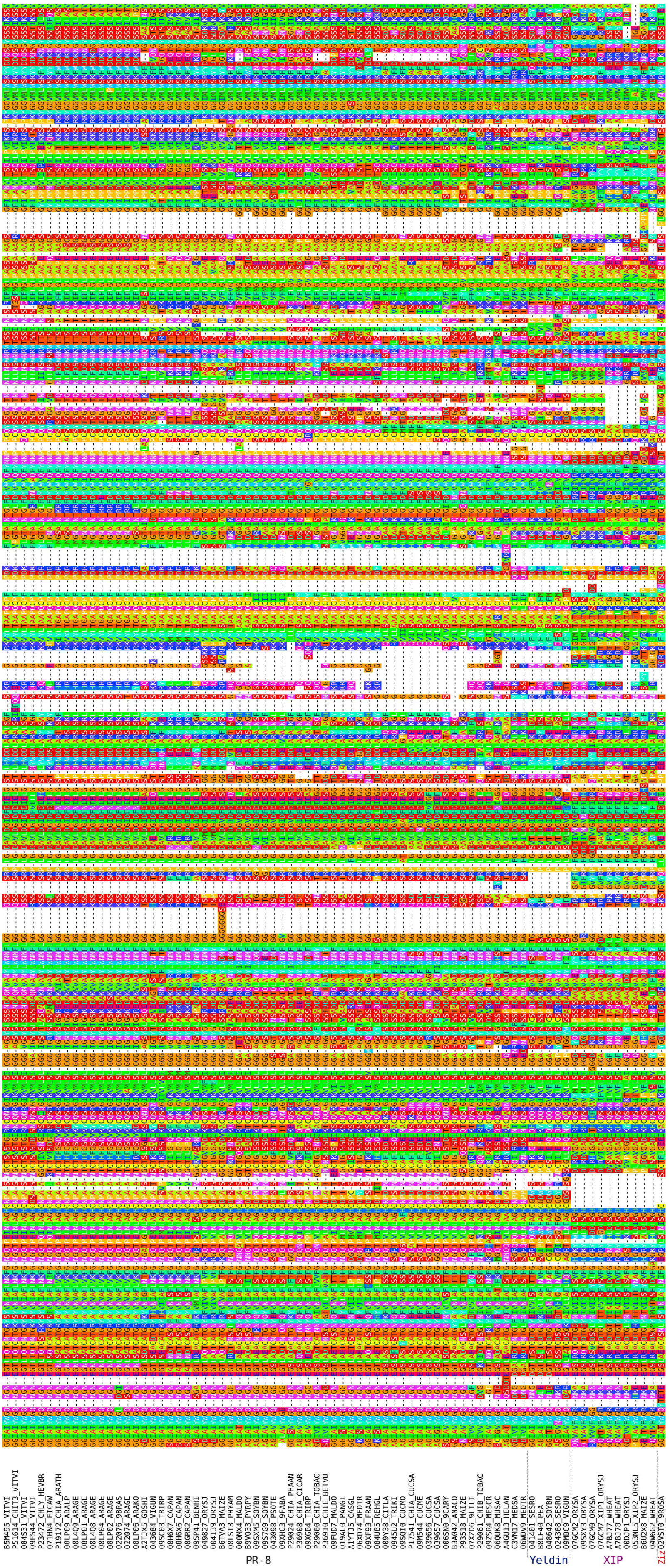


FIGURE 4.10: Alignment of the 77 sequences of PR-8 and class III chitinase homologues

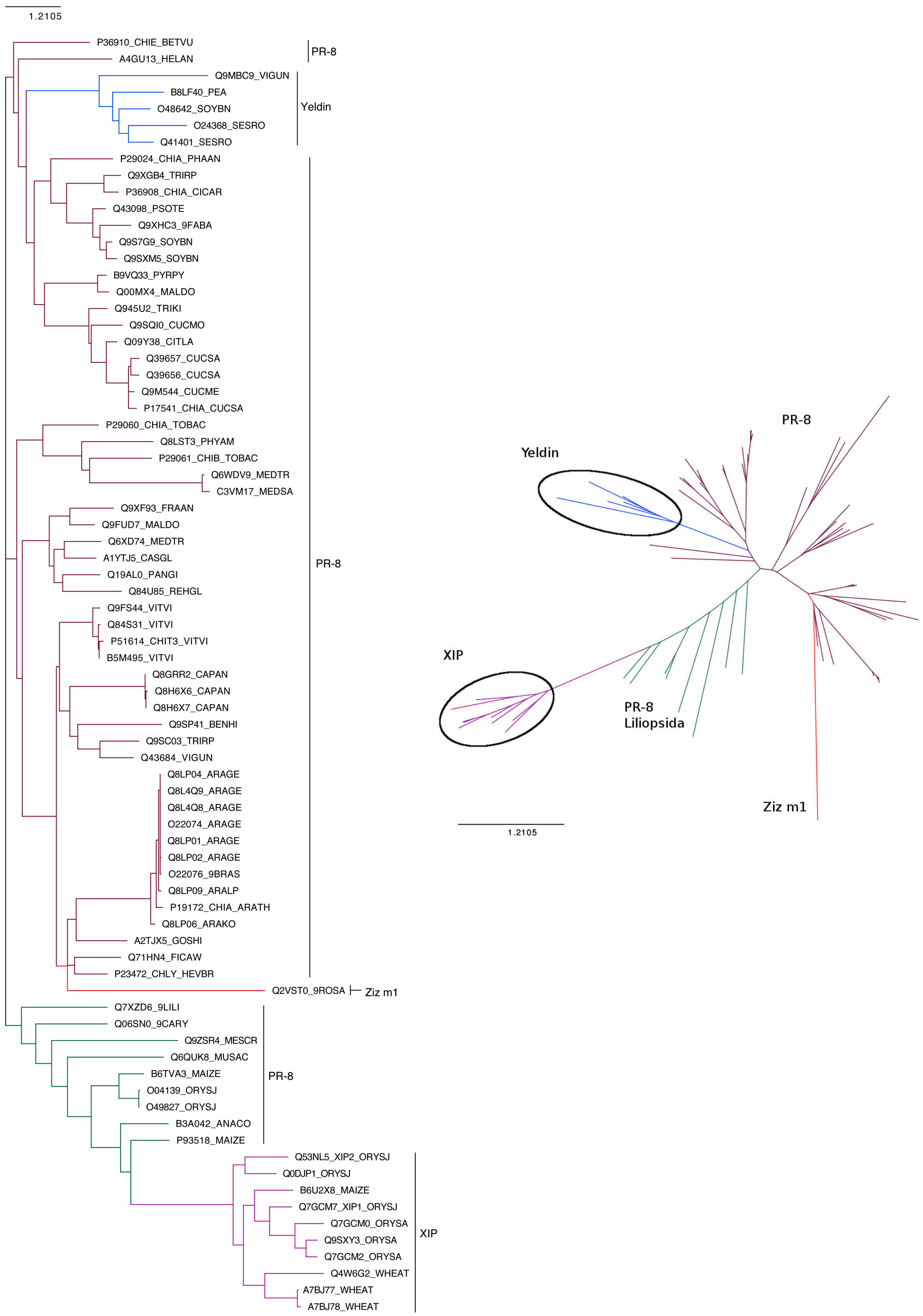


FIGURE 4.11: Trees representing the 77-sequence dataset of PR-8 and class III chitinase homologs

4.3.2 PR-11

The PR-11 family of pathogenesis-related proteins is the chitinase class with least representatives in the protein and nucleotide databases. PR-11 members are distantly related to PR-8 chitinases and share some similarity with bacterial exochitinases, but they present endochitinase activity and participate in the plant defense response [186, 214, 215, 234, 240]. By the time of their discovery, the next most related sequences belonged to chitinases from *Bacillus circulans*, *Serratia marcescens* and *S. plicatus* [186, 215]. They were described almost simultaneously by two unrelated research groups in 1994, isolated from tobacco leaves reacting hypersensitively to tobacco mosaic virus (TMV), and became designated as class V chitinases [215, 240].

The chitinase isolated by Melchers and colleagues, named Chi-V, caused lysis and complete growth inhibition of *Trichoderma viride*, and significantly inhibited growth of the phytopathogenic fungus *Altemaria radicina* Chi-V at low concentrations. On the other hand, even at high concentrations, Chi-V alone had no effect on *Fusarium solani*. In synergy with the class I β-1,3-glucanases, however, Chi-V exhibited a potent antifungal effect on *F. solani*, causing both lysis of germlings and growth inhibition of *F. solani* [215]. The authors observed intracellular accumulation and suggested that Chi-V was likely to contain a carboxy-terminal pro-peptide which directs the protein to the vacuole. This hypothesis was supported by the findings of Heitz and colleagues regarding the chitinase/lysozyme named Pz. Comparison between peptide and cDNA sequences indicated that Pz protein is synthesized as a pre-pro-protein, containing a seven-amino acid C-terminal peptide, which is probably involved in the vacuolar targeting of the protein. Pz mRNA and protein strongly accumulate in TMV-infected tobacco leaves, and Pz transcripts were also found in various tissues of healthy plants, indicating that Pz gene expression is developmentally regulated [240].

Class V homologs were described in balsam pear (*Momordica charantia*, McChi5 protein), sago palm (*Cycas revoluta*, Chitinase A) and barrel medic (*Medicago truncatula*, Mtchit 5 protein) [241–243]. A recently identified lectin from black locust bark (*Robinia pseudoacacia*) shares approximately 50% sequence identity with plant class V chitinases, but it is essentially devoid of chitinase activity [244]. Other sequences with homology to tobacco class V chitinases can be detected among genome and transcriptome products in *Vitis vinifera*, *Arabidopsis thaliana* and *Physcomitrella patens*.

SEED dataset

Ten PR-11 sequences were selected from the GH18 dataset. Among them are the two tobacco class V chitinases, Chi-V (Q43576_TOBAC) and Pz (Q43591_TOBAC), balsam pear McChi5 (Q8S2V9_MOMCH), sago palm Chitinase A (Q4W6L6_CYCRE), barrel medic Mtchit 5 (Q84N00_MEDTR) and black locust lectin (A1YZD2_ROBPS). Three chitinase-like proteins from *Arabidopsis thaliana* proteome analysis (O81856_ARATH, O81857_ARATH, O81861_ARATH), and a predicted protein from *Physcomitrella patens* (A9T029_PHYPA) complement the dataset. Sequence length varies from 327 (A1YZD2_ROBPS, fragment) to 398 (O81861_ARATH).

There was no annotation about domain features for these sequences. Signal peptides could be identified by visual inspection and confirmed by SignalP [144] prediction. A probable C-terminal extension was excluded from some sequences. The predicted catalytic domain was used to produce a SEED alignment to build profile HMMs for PR-11. The SEED alignment shows 43 residues that are identical in all sequences, 52 conserved substitutions, and 24 semi-conserved substitutions over 376 aligned sites. However, 105 alignment columns present at least one gap.

hmmsearch

The search for PR-11 homologs in the UniProtKB plant dataset returned 57 whole sequence hits with global profiles and 95 hits with local profiles. When the default filter was applied, only eight sequences were accepted, because most of the sequences were annotated as “predicted” or “putative”. Considering the scarcity of well annotated PR-11 sequences, an alternative filter was applied. The alternative filter ignores most of the annotation terms and uses a predefined sequence length range as the restriction (limiting) parameter. Sequences with less than 300, or more than 500 amino acid residues were excluded from the dataset.

EXTENDED dataset

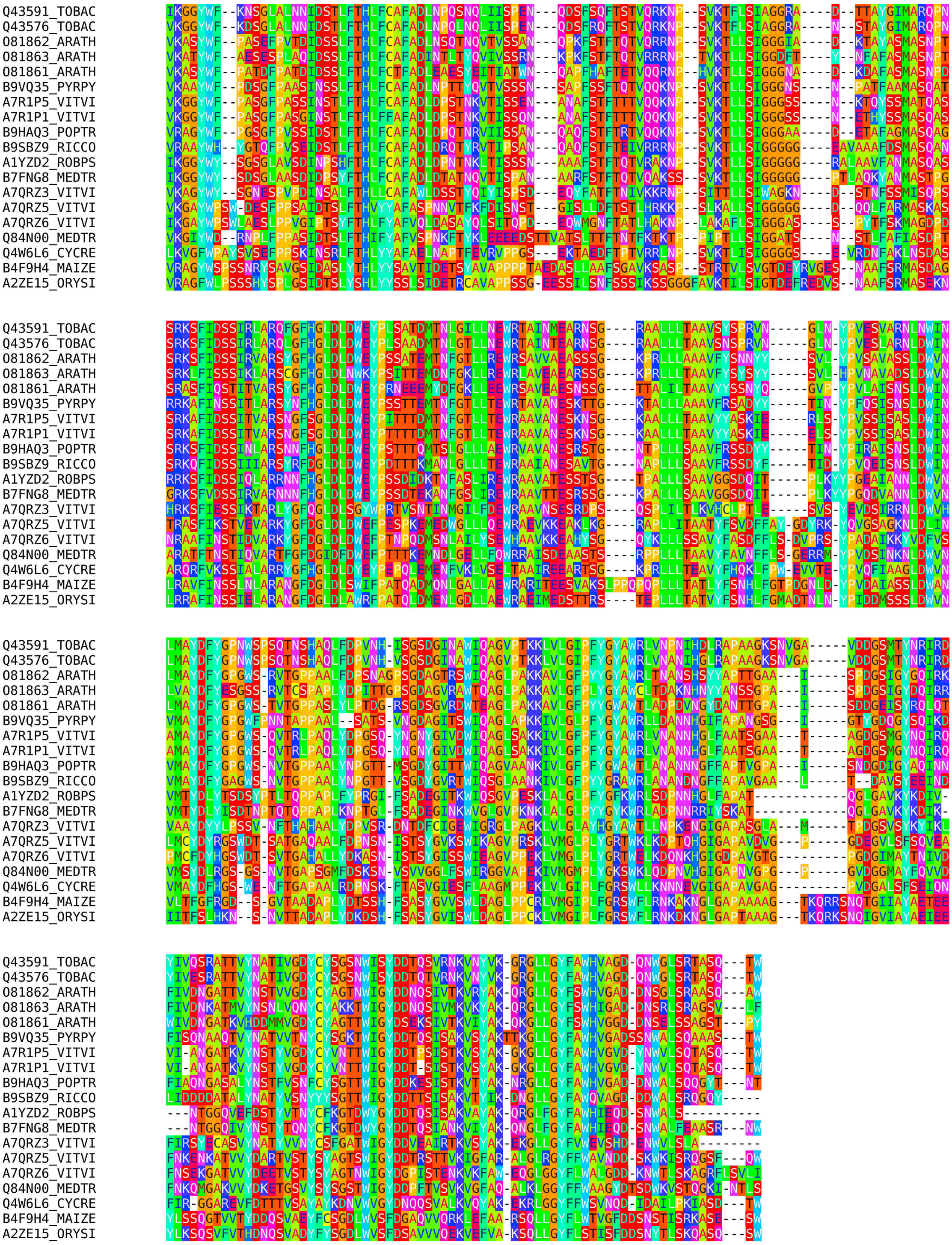
The alternative filter allowed the formation of an EXTENDED dataset with 37 sequences, 29 being from eudicotyledons, seven from Liliopsida and one from Cycadaceae. The profile HMM (global-single) was used to identify end delete the sequence regions relative to the signal peptide and C-terminal extensions. Six sequences containing large gaps in reference to the profile HMM were excluded. The edited sequences were aligned with several programs, and the alignment produced by MAFFT L-INS-i was selected as the EXTENDED alignment. The aligned coding sequences were retrieved, and the final dataset was constructed with 19 sequences (Fig. 4.12).

Selection

The aligned protein-coding sequences were analyzed with codon substitution models. No positive selection could be detected for the PR-11 dataset.

hmmsearch nrprot

The *hmmsearch* with the global-single profile HMM of PR-11 in the *nrprot* collection returned 722 matches with positive scores (E-value less than e^{-18}). Classified PR-11 sequences reached scores between 715.3 and 882.4. The non-plant sequence that reached the highest score (215.7) was a chitinase (CHIT1_HUMAN (NP_003456), chitotriosidase) from *Homo sapiens*, followed by chitotriosidases from horse (*Equus caballus*) and rat (*Rattus norvegicus*), eosinophil chemotactic cytokine from zebrafish (*Danio rerio*), chitinase 8 from red flour beetle (*Tribolium castaneum*), acidic chitinase and chitotriosidase from mouse (*Mus musculus*), a chitinase from garden spider (*Araneus ventricosus*), and chitinases from other metazoan phyla. The bacterial sequence with highest score (82.9) for this profile HMM was from the beta proteobacterium *Collimonas fungivorans*. A chitinase from *Bacillus circulans* reached a score of 67.4, worse than the score of the endochitinase class V precursor from the Ascomycota fungus *Hypocrea virens* (76.1). The bacterial chitinases are nevertheless much more related to PR-11 than the other plant glycoside hydrolases from family 18. Members of class III (PR-8) chitinases, narbonins and xylanase inhibitors present negative score values.



```
Q43591_TOBAC    IKGGYWF--KNSGLALNNIDSTLFTHLFCAFADLNPQSNQLIISPEN---QDSFSQFTSTVQRKNP--SVKTFLSIAGGRA----D--TTAYGIMARQPN
Q43576_TOBAC    VKGGYWF--KDSGLALNNIDSTLFTHLFCAFADLNPQLNQLIISPEN---QDSFRQFTSTVQRKNP--SVKTFLSIAGGRA----N--STAYGIMARQPN
O81862_ARATH    VKASYWF--PASEFPVTDIDSSLFTHLFCAFADLNSQTNQVTVSSAN---QPKFSTFTQTVQRRNP--SVKTLLSIGGGIA----D--KTAYASMASNPT
O81863_ARATH    VKATYWF--AESESPLAQIDSSLFTHLFCAFADINTLTYQVIVSSRN---KPKFSTFTQTVRRRNP--TVKTLLSIGGDFT----Y--NFAFASMASNPT
O81861_ARATH    VKASYWF--PATDFPATDIDSSLFTHLFCTFADLEAESYEITIATWN---QAPFHAFTETVQQRNP--HVKTLLSIGGGNA----D--KDAFASMASNPD
B9VQ35_PYRPY    VKAAYWF--PDSGFPAASINSSLFTHLFCAFADLNPTTYQVTVSSSN---SAPFSSFTQTVQQKNP--SVKTLLSIGGGNS----N--PATFAAMASQAT
A7R1P5_VITVI    VKGGYWF--PASGFPASSINSTLFTHLFCAFADLDPSTNKVTISSEN---ANAFSTFTTTVQQKNP--SVKTLLSIGGGSS----N--KTQYSSMATQAS
A7R1P1_VITVI    VKGGYWF--PASGFPASGINSTLFTHLFFAFADLDPSTNKVTISSQN---ANAFSTFTTTVQQKNP--SVKTLLSIGGGSS----N--ETHYSSMASQAS
B9HAQ3_POPTR    VRAGYWF--PGSGFPVSSIDSTLFTHLFCAFADLDPQTNRVIISSAN---QAQFSTFTRTVQQKNP--SVKTLLSIGGGAA----D--ETAFAGMASQAG
B9SBZ9_RICCO    VRAAYWH--YGTQFPVSEIDSTLFTHLFCAFADLDRQTYRVTIPSAN---QAQFSTFTEIVRRRNP--SVKTLLSIGGGGG----EAVAAAFDSMASQAN
A1YZD2_ROBPS    IKGGYWY--SGSGLAVSDINPSHFTHLFCAFADLDPNTNKLTISSSN---AAAFSTFTQTVRAKNP--SVKTLLSIGGGGG----RALAAVFANMASQAS
B7FNG8_MEDTR    IKGGYWY--SDSGLAASDIDPSYFTHLFCAFADLDTATNQVTISPAN---AARFSTFTQTVQAKSS--SVKTLLSIGGGGG----PTLAQKYANMASTPG
A7QRZ3_VITVI    VKAGYWY--SGNESPVPDINSALFTHLLCAFAWLDSSTYQIYISPSD---EQYFATFTNIVKKRNP--SITTLLSIWAGKN----D--STNFSSMISQPS
A7QRZ5_VITVI    VKGAYWPSW-DESFPPSAIDTSLFTHVYYAFASPNNVTFKFDISNST---GISLLDFTSTLHRKKP--SLKALLSIGGGGG----D--QQLFARMASKAS
A7QRZ6_VITVI    IKGAYWPSWLAESLPPVGIPTSYFTHLFYAFVQLDASAYQLSITQPD---EQWMGNFTATLHAKNP--LAKAFLSIGGGAS----S--PYTFSKMAGDPD
Q84N00_MEDTR    VKGIYWD--RNPLFPPASIDTSLFTHIFYAFVSPNKFTYKLEEEEDSTTVATSLTTFTNTFKTKTP--PIPTLLSIGGATS----N--STLFAFIASDPT
Q4W6L6_CYCRE    LKVGFWPAYSVSEFPPSKINSRLFTHLYYAFAELNAPTFEVRVPPGS---EKTAEDFTPTVRRLNP--SVKTLISIGGGGS----E-VRDNFAKLNSDAS
B4F9H4_MAIZE    VRAGYWSPSSNRYSAVGSIDASLYTHLYYSAVTIDETSYAVAPPPPTAEDASLLAAFSGAVKSASP--STRTVLSVGTDEYRVGES--NAAFSRMASDAG
A2ZE15_ORYSI    VRAGFWLPSSSHYSPLGSIDTSLYSHLYYSSLSIDETRCAVAPPSSG-EESSILSNFSSSIKSSGGGFAVKTILSIGTDEFREDVS--NAAFSRMASEKN

Q43591_TOBAC    SRKSFIDSSIRLARQFGFHGLDLDWEYPLSATDMTNLGILLNEWRTAINMEARNSG----RAALLLTAAVSYSPRVN-----GLN-YPVESVARNLNWIN
Q43576_TOBAC    SRKSFIDSSIRLARQLGFHGLDLDWEYPLSAADMTNLGTLLNEWRTAINTEARNSG----RAALLLTAAVSNSPRVN-----GLN-YPVESLARNLDWIN
O81862_ARATH    SRKSFIDSSIRVARSYGFHGLDLDWEYPSSATEMTNFGTLLREWRSAVVAEASSSG----KPRLLLAAAVFYSNNYY-----SVL-YPVSAVASSLDWVN
O81863_ARATH    SRKLFISSSIKLARSCGFHGLDLNWKYPSITTEMDNFGKLLREWRLAVEAEARSSG----KPRLLLTAAVFYSYSYY-----SVL-HPVNAVADSLDWVN
O81861_ARATH    SRASFIQSTITVARSYGFHGLDLDWEYPRNEEEMYDFGKLLEEWRSAVEAESNSSG----TTALILTAAVYYSSNYQ-----GVP-YPVLAISNSLDWIN
B9VQ35_PYRPY    RRKAFINSSITLARSYNFHGLDLDWEYPSSTTEMTNFGTLLTEWRTAVANESKTTG----KTALLLAAAVFRSADYY-----TIN-YPFQSISNSLDWIN
A7R1P5_VITVI    SRKAFIDSSITVARSNGFSGLDLDWEYPITTTDMTNFGTLLTEWRAAVANESKNSG----KAALLLTAAVYYASKIE-----RLS-YPVSSISASLDWIN
A7R1P1_VITVI    SRKAFIDSSITVARSNGFSGLDLDWEYPTTTTDMTNFGTLLTEWRAAVANESKNSG----KAALLLTAAVYYASKIE-----ELS-YPVSSISASLDWIN
B9HAQ3_POPTR    SRKSFIDSSINLARSNNFHGLDLDWEYPSTTTQMTSLGLLLAEWRVAVANESRSTG----NTPLLLSAAVFRSSDYY-----TIN-YPIRAISNSLDWIN
B9SBZ9_RICCO    SRKQFIDSSIIIARSYRFDGLDLDWEYPDTTTKMANLGLLLTEWRAAIANESAVTG----NAPLLLSAAVFRSSDYF-----TID-YPVQEISNSLDWIN
A1YZD2_ROBPS    RRKSFIDSSIQLARRNNFHGLDLDWEYPSSDIDKTNFASLIREWRAAVATESSTSG----TPALLLSAAVGGSDQIT-----PLKYYPGEAIANNLDWVN
B7FNG8_MEDTR    GRKSFVDSSIRVARNNNFHGLDLDWEYPSSDTEKANFGSLIREWRAAVTTESRSSG----KPALLLSAAVGGSDQIT-----PLKYYPGQDVANNLDWVN
A7QRZ3_VITVI    HRKSFIESSIKTARLYGFQGLDLSGYWPRTVSNTINMGILFDEWRAAVNSESRDPS----QSPLILTLKVHCLPTLE-----SVS-YEVDSIRRNLDWVH
A7QRZ5_VITVI    TRASFIKSTVEVARKYGFDGLDLDWEFPESPKEMEDWGLLLQEWRAEVKKEAKLKG----RAPLLITAATYFSVDFFAY-GDYRK-YQVGSAGKNLDLIN
A7QRZ6_VITVI    NRAAFINSTIDVARKYGFDGLDLDWEFPTNPQDMSNLAILYSEWHAAVKKEAHYSG----QYKLLLSSAVYFASDFFLS-DVPRS-YPADAIKKYVDFVS
Q84N00_MEDTR    ARATFTNSTIQVARTFGFDGIDFDWEFPTTTKEMNDLGELLFQWRRAISDEAASTS----RPPLLLTAAVYFAVNFFLS-GERRM-YPVDSINKNLDWVN
Q4W6L6_CYCRE    ARQRFVKSSIALARRYGFHGLDLDYEYPEPQLEMENFVKLVSELTAAIREEARTSG----KPRLLLTEAVYFHQKLFPW-EVVTE-YPVQFIAAGLDWVN
B4F9H4_MAIZE    LRAVFINSSLNLARANGFDGLDLSWIFPATQADMQNLGALLAEWRARITEESVAKSLPPQPQPLLLTATLYFSNHLFGTPDGNLD-YPVDAIASSLDWAN
A2ZE15_ORYSI    LRRAFINSSIELARANGFDGLDLAWRFPATQLDMENLGDLLAEWRAEIMEDSTTRS----TEPLLLTATVYFSNHLFGMADTNLN-YPIDDMSSSLDWVN

Q43591_TOBAC    LMAYDFYGPNWSPSQTNSHAQLFDPVNH-ISGSDGINAWIQAGVPTKKLVLGIPFYGYAWRLVNPNIHDLRAPAAGKSNVGA-----VDDGSMTYNRIRD
Q43576_TOBAC    LMAYDFYGPNWSPSQTNSHAQLFDPVNH-VSGSDGINAWIQAGVPTKKLVLGIPFYGYAWRLVNANIHGLRAPAAGKSNVGA-----VDDGSMTYNRIRD
O81862_ARATH    LMAYDFYGPGWS-RVTGPPAALFDPSNAGPSGDAGTRSWIQAGLPAKKAVLGFPYYGYAWRLTNANSHSYYAPTTGAA---I-----SPDGSIGYGQIRK
O81863_ARATH    LVAYDFYESGSS-RVTCSPAPLYDPITTGPSGDAGVRAWTQAGLPAKKAVLGFPLYGYAWCLTDAKNHNYYANSSGPA---I-----SPDGSIGYDQIRR
O81861_ARATH    LMAYDFYGPGWS-TVTGPPASLYLPTDG-RSGDSGVRDWTEAGLPAKKAVLGFPYYGWAWTLADPDVNGYDANTTGPA---I-----SDDGEISYRQLQT
B9VQ35_PYRPY    VMAYDFYGPGWFPNNTAPPAAL--SATS-VNGDAGITSWIQAGLAPKKIVLGLPFYGYAWRLLDANNHGIFAPANGSG---I-----GTYGDQGYSQIKD
A7R1P5_VITVI    AMAYDFYGPGWS-QVTRLPAQLYDPGSQ-YNGNYGIVDWIQAGLSAKKIVLGFPFYGYAWRLVNANNHGLFAATSGAA---T-----AGDGSMGYNQIRQ
A7R1P1_VITVI    AMAYDFYGPGWS-QVTRLPAQLYDPGSQ-YNGNYGIVDWIQAGLSAKKIVLGFPFYGYAWRLVNANNHGLFAATSGAA---T-----AGDGSMGYNQIRQ
B9HAQ3_POPTR    VMAYDFYGPGWS-NVTGPPAALYNPGTT-MSGDYGITTWIQAGVAANKIVLGFPFYGWAWTLANANNNGFFAPTVGPA---I-----SNDGDIGYAQINN
B9SBZ9_RICCO    VMAYDFYGAGWS-NVTGPPAALYNPGTT-VSGDYGVRTWIQSGLAANKIVLGFPYYGRAWRLANANDNGFFAPAVGAA---L-----T--DAVSYEEIND
A1YZD2_ROBPS    VMTYDLYTSDSYPTLTQPPAPLFYPRGI-FSADEGITKWIQSGVPESKLALGLPFYGFKWRLSDPNNHGLFAPAT------------QGLGAVKYKDIV-
B7FNG8_MEDTR    VMTYDLYISDTNPTQTQPPAPLKNPTGL-FSADEGINKWVGLGVPKNKLALGLPYYGYKWRLADPNNRRIYSKAT------------QGLGAVKYKDIK-
A7QRZ3_VITVI    VAAYDYYLPSSV-NFTHAHAALYDPVSR-DNTDFCIGEWIGRGLPAGKLVLGLAYHGYAWTLLNPKENGIGAPASGLA---M-----TPDGSVSYKYIKL
A7QRZ5_VITVI    LMCYDYRGSWDT-SATGAQAALFDPNSN-ISTSYGVKSWIKAGVPRSKLVMGLPLYGRTWKLKDPTQHGIGAPAVDVG---P-----GDEGVLSFSQVEA
A7QRZ6_VITVI    PMCFDYHGSWDT-SVTGAHALLYDKASN-ISTSYGISSWIEAGVPPEKLVMGLPLYGRTWELKDQNKHGIGDPAVGTG---------PGDGIMAYTNIVD
Q84N00_MEDTR    VMSYDLRGS-GS-NVTGAPSGMFDSKSN-VSVVGGLFSWIRGGVAPEKIVMGMPLYGKSWKLQDPNVHGIGAPNVGPG---P-----GVDGGMAYFQVVD
Q4W6L6_CYCRE    VMAYDFHGS-WE-NFTGAPAALRDPNSK-FTASVGIESFLAAGMPPEKLVLGIPLFGRSWLLKNNNEVGIGAPAVGAG---------PVDGALSFSEIQN
B4F9H4_MAIZE    VLTFGFRGD--S-GVTAADAPLYDTSSH-FSASYGVVSWLDAGLPPCRLVMGIPLFGRSWFLRNKAKNGLGAPAAAAG---TKQRRSNQTGIIAYAETEE
A2ZE15_ORYSI    IITFSLHKN--S-NVTTADAPLYDKDSH-FSASYGVISWLDAGLPPCKLVMGIPLFGRSWFLRNKDKNGLGAPTAAAG---TKQRKSNQIGVIAYAEIEE

Q43591_TOBAC    YIVQSRATTVYNATIVGDYCYSGSNWISYDDTQSVRNKVNYVK-GRGLLGYFAWHVAGD-QNWGLSRTASQ---TW
Q43576_TOBAC    YIVESRATTVYNATIVGDYCYSGSNWISYDDTQTVRNKVNYVK-GRGLLGYFAWHVAGD-QNWGLSRTASQ---TW
O81862_ARATH    FIVDNGATTVYNSTVVGDYCYAGTNWIGYDDNQSIVTKVRYAK-QRGLLGYFSWHVGAD-DNSGLSRAASQ---AW
O81863_ARATH    FIVDNKATMVYNSNLVQNYCYAKKTWIGYDDNQSIVMKVKYAK-QRGLLGYFSWHIGAD-DNSRLSRAGSV---LF
O81861_ARATH    WIVDNGATKVHDDMMVGDYCYAGTTWIGYDSEKSIVTKVIYAK-QKGLLGYFSWHVGGD-DNSELSSAGST---PY
B9VQ35_PYRPY    FISQNAAQTVYNATVVTNYCYSGKTWIGYDDTQSISAKVSYAKTTKGLLGYFAWHVGADSSNWALSQAAAS---TW
A7R1P5_VITVI    VI-ANGATKVYNSTYVGDYCYVNTTWIGYDDTPSISTKVSYAK-GKGLLGYFAWHVGVD-YNWVLSQTASQ---TW
A7R1P1_VITVI    VI-ANGATKVYNSTYVGDYCYVNTTWIGYDDT-SISTKVSYAK-GKGLLGYFAWHVGVD-YNWVLSQTASQ---TW
B9HAQ3_POPTR    FIAQNGASALYNSTFVSNFCYSGTTWIGYDDKESISTKVTYAK-NRGLLGYFAWHVGAD-DNWALSQQGYT---NT
B9SBZ9_RICCO    LIDDDDATALYNATYVSNYYYSGTTWIGYDDTTSISAKVTYIK-DNGLLGYFAWQVAGD-DDWALSRQGQY-----
A1YZD2_ROBPS    ---NTGGQVEFDSTYVTNYCFKGTDWYGYDDTQSISAKVAYAK-QRGLFGYFAWHIEQD-SNWALS----------
B7FNG8_MEDTR    ---NTGGQIVYNSTYTQNYCFKGTDWYGYDDTQSIANKVIYAK-QNGLFGYFAWHIEQD-SNWALFEAASR---NW
A7QRZ3_VITVI    FIRSYECASVYNATYVVNYCSFGATWIGYDDVEAIRTKVSYAK-EKGLLGYFVWEVSHD-ENWVLSLA--------
A7QRZ5_VITVI    FNKENKATVVYDARTVSTYSYAGTSWIGYDDTRSTTVKIGFAR-ALGLRGYFFWAVNDD-SKWKISRQGSF---QW
A7QRZ6_VITVI    FNSEKGATVVYDEETVSTYSYAGTNWIGYDGPISTENKVKFAW-EQGLGGYFLWALGDD-KNWTLSKAGRFLSVLI
Q84N00_MEDTR    FNKQMGAKVVYDKETGSVYSYSGSTWIGYDDPFTVSVKVGFAQ-ALKLGGYFFWAAGYDTSDWKVSTQGKI-NTLS
Q4W6L6_CYCRE    FIR-GGAREVFDTTTVSAYAYKDNVWVGYDNQQSVALKVQYAK-EKRLGGYFFWSVNQD-IDAILPKIASD---TW
B4F9H4_MAIZE    YLSSQGTVVTYDDQSVAEYFCSGDLWVSFDGAQVVQRKLEFAA-RSQLLGYFLWTVGFDDSNSTISRKASE---SW
A2ZE15_ORYSI    YLKSQSVFVTHDNQSVADYFYSGDLWVSFDSAVVVQEKVEFVA-KSQLLGYFLSTISFDDSNYTLSKQASQ---SW
```

FIGURE 4.12: Alignment of PR-11 EXTENDED dataset

4.4 Discussion

PR-3

Previous phylogenetic analysis of chitinases from classes I, II and IV already revealed a greater evolutionary distance between chitinases of class IV and those of classes I and II, suggesting a remote divergence between these classes [153]. According to Wiweger and colleagues (2003), class IV chitinases probably evolved from class I or II chitinases more than 300 million years ago, i.e. before the separation of angiosperms and gymnosperms [245].

Some bacterial chitinases from *Streptomyces* and *Burkholderia*, as well as a chitinase from an oomycete, *Phytophthora infestans*, are evolutionarily related to the class IV chitinases of plants, supporting the hypothesis that some class IV chitinases in bacteria have evolved from eukaryotic chitinases via horizontal gene transfer [220]. According to Lohtander and colleagues (2008), HGT of a chitinase IV gene from eukaryotes to bacteria has presumably occurred only once. The CAZy database lists members of the GH19 family that have been found in viruses that infect bacteria (bacteriophages) such as *Burkholderia ambifaria* phage, that is in line with the hypothesis held by Lohtander and colleagues. In addition to the plant proteins, the database lists a few non-plant eukaryotic members of the GH19 family belonging to amoeba, *Toxoplasma*, unicellular fungi, and nematodes [210]. Whether these genes are xenologs (originated via HGT) or not remains to be studied.

PR-4

PR-4 proteins without the hevein domain, such as Tobacco R and Barwin, can be found in monocotyledons and dicotyledons. Two sequences of the Coniferopsida *Picea sitchensis* (Sitka spruce) with 71% identity to Barwin have been found in transcriptome analysis, but remain uncharacterized. On the other hand, CBD-containing PR-4 proteins have only been described in dicotyledons. Besides a fragment of 9 amino acids isolated from the Gnetophyta *Ephedra distachya* (Joint-fir) identical to a region in the barwin domain of tomato, no other sequences of PR-4 protein have been reported in "lower plants". Two sequences from fungi and five sequences from viruses have been related to barwin domains in the Integrative Protein Signature Database (InterPro: IPR001153) [246] and in the Protein Families Database (Pfam: PF00967) [247].

Following the idea that Barwin is not a chitinase, and based on the observation that PR-4 and PR-3 catalytic domains are "not alignable", it is intriguing that some PR-4 entries in UniProtKB are cross-referenced with InterPro entries for glycoside hydrolase family 19 (InterPro: IPR000726). The "mystery" is explained by the "equivocal" (wrong) inclusion of six PR-4 sequences in the training dataset for the chitinase-related family in the PANTHER library (PTHR22595, Glyco_hydro_19_cat, April 2009) [248].

GH18

Family 18 of glycoside hydrolases has been shown to have a complex evolutionary history. The diversity of the GH18 classes found in plants originates early, before the separation of the higher eukatiotic phyla. This became clear by the observation that the class V chitinases of the GH18 family (PR-11) were much more similar to mammal proteins than to other plant proteins.

Profile HMMs demonstrated efficiency in the identification of distant homologs of PR-8 and PR-11 in other phyla. Among the plant GH18 members, class III chitinases appear to be a great source for the emergence of new functions, as in the case of their nearest homologs, the xylanase inhibitors. Xylanase inhibitor proteins are similar to class III chitinases, but they do not exhibit chitinase activity. They possess competitive inhibiting activity against several fungal endo-1,4-β-D-xylanases, secreted during pathogenic infestation. Durand and colleagues (2005) suggested that XIP-type proteins evolved from chitinases as part of the plant defense pathway [237]. XIP-type proteins are an amazing example of neofunctionalization after gene duplication. Yieldin, a chitinase-like wall protein isolated from cowpea hypocotyls, was found to enhance wall extension [238], and is also derived from class III chitinases.

Twenty-five years after the discovery of class V chitinases, the abundance of molecular data stored in the databases increased exponentially. The use of profile HMMs to search a collection of protein and nucleotide databases led to the conclusion that the next closest related sequences to PR-11 are not bacterial, but indeed belong to animal chitinases. Even more interestingly, the sequence that retrieved the highest score is a human chitinase. It is noteworthy that both the plant class V chitinases and the human protein are related to defense response against invaders. The human chitotriosidase is a vesicular and secreted protein produced by activated macrophages, while the class V chitinase is induced by different types of pathogens.

Furthermore, an investigation of a possible role of Class II GH18 chitinases in defense response should be considered.

Conclusions

When plants are in danger, they defend themselves by producing different sets of defense proteins and chemical compounds. Pathogenesis-related proteins (PRs) form a class of defense proteins induced in plants in a context of pathogenicity or physiological danger for the organism, and confer enhanced resistance to subsequent infections. They are members of multigenic families encoded by genes with an extremely variable number of copies in different species. This intrinsic feature, coupled with the disparate availability of sequences in public databases, makes the assembly of a concise and reliable dataset for evolutionary analysis a delicate quest.

In this thesis I proposed a methodological framework to generate reliable datasets and perform phylogenetic analysis of the PR-proteins. It incorporates several methods, algorithms and strategies from bioinformatics, carefully selected to accommodate the characteristics of each specific dataset and increase the quality of the analysis.

This framework was applied to the 17 PR-families known to date, and the results confirm a great improvement in the quality of the analysis when compared to previous surveys. Approximately four thousand sequences were examined during the analysis. Profile hidden Markov models proved to be an outstanding resource for the assembly of representative datasets. The efficient retrieval of sequence entries, combined with a good filtering strategy, represents a great improvement to the methodology. The produced datasets contained a relatively good diversity of sequences and species, considering the restrictions of the databases.

Another great improvement can be attributed to the incorporation of two different approaches to alignment quality evaluation in the framework. The result is a reliable alignment for the construction of phylogenetic trees and analysis of evolutionary models. The phylogenetic analysis was also carefully conducted, and different trees were tested with codon substitution models. Nevertheless, this method is not indicated for unsupervised analysis. The most significant aspect of the proposed framework is the intensive “neatness” with which the data is treated, and this requires the constant intervention of the researcher.

Apart from the advanced automated bioinformatic frameworks, the visual inspection of sequence data is still a powerful tool for the researcher. DNATagger, a sequence visualization tool, uses the codons as informative units for coloring DNA and RNA sequences. It was developed during this study to facilitate the comprehension of evolutionary processes in sequence analysis. In the present work, the coding nucleotide sequences were aligned as non-coding, and the results were compared to the coding-based alignment. The visual analysis with DNATagger resulted in the exclusion of 90 sequences ($\sim$5%) from the final dataset.

Positive selection analysis of the constructed datasets indicated site-specific selection in five PR-families. The sites predicted to be positively selected can be used as targets for mutational selection analysis *in vitro*, or even *in silico*. Today, sophisticated computational biology employs molecular modeling and molecular docking utilities to simulate and evaluate the effects of amino acid substitutions in the structure of the proteins.

Unfortunately, not every biologically relevant result could be considered in detail in this thesis. At least three further chapters were proposed initially to explore the specific results obtained for i) antimicrobial peptides (PR-12, PR-13 and PR-14), ii) germin and germin-like proteins (PR-15 and PR-16), and iii) PR-17, the last protein included in the PR-classification, since less is known about its function and evolution. Nevertheless, chapter four gives an in-depth description of an important group of PR-families: the chitinases (PR-3, PR-4, PR-8 and PR-11).

The next step is to investigate the molecular differences that may account for the specific actions of PR-15 and PR-16 using branch site models. The phylogenetic analysis demonstrated that PR-15, which has oxalate oxidase activity, groups as a subtree of PR-16, which has superoxide dismutase activity. A preliminary analysis with branch site models (data not shown) indicated 10 sites subject to positive selection, that may be involved with the change in function.

It must be mentioned that all programs used here are freely available, and most of them are open source software. Thanks to the scientific community, the bioinformatic resources are becoming more and more accessible. Also due to the increasing availability of bioinformatic web-portals, researchers of all countries can reach powerful resources. Today, if you have a good internet access, a desktop computer and, most important, a good idea, you can produce quality research from anywhere in the world.

Appendix A

DNATagger, colors for codons

Software

DNATagger is designed to colorize codons in DNA/RNA sequence alignments, and amino acids in protein sequence alignments. Therefore, all DNA/RNA sequences and sequence alignments are interpreted as being from protein coding genes. The coding frame is determined by the first base letter recognized in each sequence, irrespective of its position in the alignment. Gaps occurring in the sequence string are interpreted as non-informative characters, and are not included as part of the codon. For example, the sequence string `ACAT--GG-C-T-A` will be displayed with the colors representing the amino acids coded by the codons `ACA`, `TGG`, and `CTA`. The gaps will continue to be displaced in the original position, but they will not be colorized.

Usage

Sequences and alignments should be introduced as plain text on the "Edit" tab by pasting or typing. There is no need for file uploads. The user can work with his/her alignments in raw text. The simple text editing function provides freedom over the standard formats required by most applications. The sequences are interpreted with respect to the genetic code chosen by the user. There are 17 genetic codes available at the selection box on the "Translation Table" tab. By default, the standard code is used.

Amino acid and codon sequences are colored on the “View” tab following the color table selected by the user on the “Colors” tab. Several color combinations are already available. A monochromatic table was also created, motivated by the publication costs of color illustrations. The user is invited to create his/her own color table using the color bar option, or to write his/her own code. This table can be saved for future use in a text file. For reuse of this color table, the user should place this code in the code box and click on “interpret this color table code”.

Availability and requirements

DNATagger is a JavaScript application, following the W3C guidelines, designed to work on standards-compliant web browsers. It therefore requires no installation and is platform independent. The web-based DNATagger is available as free and open source software at `http://www.inf.ufrgs.br/~dmbasso/dnatagger/`.

Appendix B

Datasets – UniProtKB Entries

The sequences used in the first and in the last step of the analysis are identified here by their UniProtKB accession numbers.

B.1 SEED

PR01

Q04108, P04284, P33154, Q05968, Q41359, Q08697, P08299, P07053, P09042, Q00008, A0N0C1, A0N0C2, A0N0C3, B5QTD3, O65157, O82086, Q0KIX9, Q43392, Q00MX6, Q96344.

PR02

P52401, P33157, P49237, Q03467, Q01413, P23547, P07979, A0FLG4, P93519, Q42890, Q43778, Q56AP0, Q6S9W0, Q84JM2, Q9ATR3, Q9M3U4, Q9M563, Q9XFW9, Q9ZP12.

PR03

P42820, P27054, Q06209, P29022, P80052, P17513, P17514, Q05540, P29021, Q40114, Q05539, P08252, P24091, P05315, Q05538, P21226, P06215, P51613, P11955, Q9FRV1, P23951, Q9FRV0, P24626, Q09023, P19171, Q41596, P85084, P36907, P16579, P16061, Q2VAC7, Q84LK1, Q9SYS4, Q944B9, Q944B8, A7XQ02, P11218.

PR04

O64392, O64393, P28814, P93180, Q6T5J8, P43082, Q8H950, O48880, P02877, A0SWV6, P09761, P09762, Q41231, P29062, P29063, P32045, Q9M7D9.

PR05

P12670, P50701, P25871, P13046, P81370, P13867, P81295, P02884, P50696, P50698, P32937, P83335, Q9FSG7, Q9SMH2, P28493.

PR06

P80211, P82381, P24076, Q6XNP7, P81712, P16064, P05118, P08454, Q03198, Q03199, Q02214, P01053, P82977, P16062, P16063, P19873.

PR07

O65835, O65834, O82007, Q96478, O04678, Q9LWA3, Q9LWA4, O65836, Q9SAN2, Q9ZR46.

PR08

P29024, P36910, P29060, P36908, P17541, Q00MX4, P19172, Q9FS44, P29061.

PR09

Q96520, Q43032, O49866, Q5U1S8, Q9XIV8, Q8W4V8, Q5I3F2, Q9XGV6, Q43212, Q40949, Q43099, Q43101, Q42905, Q43102, Q8RVP3, Q8S3U4, Q43790, Q42580.

PR10

B5KVN9, Q941P7, Q9LEP0, Q9FS42, Q94IM3, O24256, Q9LLQ3, Q9SXX8, P25985, Q9M500, Q9LKJ9.

PR11

Q4W6L6, A9T029, Q84N00, Q43576, Q43591, O81856, O81857, O81861, A1YZD2, Q8S2V9.

PR12

Q01783, Q01784, P30224, P69241, Q43413, O65740, Q8GTM0, Q8H6Q1, Q8H6Q0.

PR13

Q9SBK8, Q42596, P01543, P09618, Q8LSZ9.

PR14

Q41258, Q43767, Q9SDS2, Q9SDS3, Q52RN7, Q850K5, Q9SMM1.

PR15

P15290, P45851, Q70PK0, O24004, Q0GR11, P45850, P26759, Q8L696, Q8L697, Q9FEW6, Q851K0, Q851K1, Q851J8.

PR16

O65358, Q5DT23, O64439, Q9S8P4, P92995, Q43487, Q9SM35, Q6YZY5, Q6JVN1.

PR17

Q9XIY9, Q84XQ4, Q94I89, Q84ZU9, Q84ZV0, Q41523, A7YA66, Q9SWZ5, A7YA60, O24003, O24002.

B.2 EXTENDED

PR01

O65157, B2LW68, P11670, Q08697, Q00MX6, Q39188, B2CZ52, P08299, Q40557, P07053, P09042, Q40397, O82086, B3TLW3, C3UZE5, C3UZE4, O82714, Q3S4I3, Q43489, P35793, Q05968, P35792, Q94F73, C3UZE2, C3UZE3, A0N0C3, Q00008, O82715, Q41359, B5QTD3, A0N0C2, A0N0C1, Q7FP72, Q8LLU7, Q3S4I4, Q0KIX9, P33154, Q2HZ50, Q6IUZ8, B1N8M3, Q8L688, Q9SC15, Q941G6, Q8L687, P04284, Q04108, O24026, Q2VT59, Q6WHB9, O81889, Q9LJM5, Q9XFB4, Q40374, B6UG61, Q9M0C8, Q9SV22, Q6ID87, Q2QLQ1, Q39186, Q0DWY7, Q0DFE6, A6GVD5, Q8S9B6.

PR02

P07979, Q01413, P52401, P27666, P23546, Q9M563, Q00NV3, Q9ZP99, Q5RLY0, A9YYK4, Q40314, Q9XFW9, Q03467, P23535, Q8VZJ2, B2ZP01, A9CSM2, Q9M3U4, Q9M5I9, O23783, Q1X7Q1, Q03773, Q43778, P36401, Q42890, Q68V46, Q6TQD8, A0FLG4, O22317, Q0QJY6, B3TLW8, B3TLW9, Q1W6B9, Q69D51, P93153, Q8LG04, Q9M2M0, P33157, Q2HZ53, Q2VT22, P49236, A4USG1, Q2HZ52, Q84RT6, Q84JM2, Q56AP0, Q6S9W0, Q9ZP12, B2NK62, Q9XFW8, P93519, P49237, Q9XEN7, Q5DM81, Q9SXY6, Q1ERG1, Q1ERF7, Q1ERG2, Q84V65, Q1ERF9, Q1ERF8, Q1ERG0, Q8W4V0, Q4JH28, O82716, Q1EMA4, Q1EMA2, Q8S3U1, P15737, Q9SXY7, Q5JMU8, Q02438, Q02437, Q9XEN5, Q9ATR3, Q5UAW3, Q7DLM1, Q42518, B6T391, B6TH79, B6TNS8, Q1EMA6, B6U015, B6TDK5, B6TDV9, Q8LEU0, P52399, P52398, P23433, P23432, P23547, Q70C53, O82063, Q01412, Q944C0, O82673, Q588B8, Q42944, Q9FXL4, Q8H0I0, Q0V7P5, Q8S9I6, Q8L9D9, Q45X99, Q4ZGK1, B6TRI0, Q10P58, B6TJF7, Q0JDD4, B6SUM3, Q0DEW3, Q8L837, Q1WAK1, B6T289, B6TU78, Q94G86, Q9M2T6, Q8L935, Q9M4A9, B9VQ36, Q9AVE5, B5M9E5, B6TY63, P52409, B6THD2, Q08A62, Q94EN5, B6TLN1, B6TCY1, B7F9R0, B9DGU2, B6UBU1, Q9LK41, Q2ACE6, B6TIF7, B6SWD3, B6STZ9, B6SXV6, Q6TQD7, Q9LK11, B6U2N5.

PR03

P24091, Q9FEW1, P08252, P05315, B9VRK7, Q05538, O81144, O81145, P29059, Q09023, P19171, C0LNR1, B8YPL6, P51613, A3QRB7, Q6IVX2, Q6IVX4, P93680, B9ZZZ5, Q9SPU0, Q43184, Q42878, B8XR33, B9VQ31, Q944U5, Q7X9F6, Q7X9F4, P36907, Q9SDY6, Q42428, Q1W6C5, O80404, Q9AVA8, Q6IVX8, Q9FRV1, P11955, P23951, Q9FRV0, P24626, Q40667, Q2HJJ5, P25765, A7UC81, Q6T484, Q8W427, Q9AXR9, Q41539, B1B6T0, B8R3R6, Q42992, Q8VWZ5, Q6SPQ7, A2V800, Q8W428, B6TR38, Q43294, Q42970, Q8H0C9, O81934, Q9XFK6, O65330, Q7X9F5, P21226, P93327, P94084, Q9ZP10, Q8MD06, Q8LK49, P06215, P36361, Q41596, Q39799, B0ZC08, O65331, Q4PJV8, Q7X9R8, Q7Y237, Q9SQF7, Q43752, Q8H985, P16579, Q6E6M9, Q9SPT9, B2X051, Q05540, P29021, P17514, P17513, Q40114, Q43834, Q05539, O82552, Q43765, Q43764, Q8W429, Q9AXR8, O80423, B6T6W1, B6SZC6, Q9LE03, Q40668, Q9ZT61, Q9SYS4, Q944B9, Q9ZT60, Q944B8, Q84LK1, Q2VAC7, A7XQ02, P11218, P16061, Q6JX04, B5G509, B5G510, B9VQ32, Q7X7Q1, A8IXW1, Q1W6C4, P42820, B0FZ27, O24530, O24531, B0FZ26, Q7XB39, B8Y647, Q6RV28, B9VQ34, Q9M2U5, Q7X9F8, P29022, Q93WT1, O23803, O23804, O23806, O23805, Q9XFW7, O82547, Q9XEN3, Q9XEN6, B6SZA3, A9ZMK1, C1K2E8, Q96411, Q96410, Q96409, C3VP99, C3VPA0, Q6WSR8, Q6WSR9, Q40838, Q596H9, Q5NTA4, Q06209, Q05K38, A8IXF7, A5JVZ1, Q94C47.

PR04

P09761, P09762, Q41231, A0SWV6, P02877, P43082, O48880, P29062, Q40558, P29063, P32045, Q8H6J9, Q6DQK2, B6EB12, Q9M7D9, Q9M7D8, Q6PWL9, O64393, O64392, P93180, Q6T5J8, B6SH12, Q41802, Q9SEM3.

PR05

Q9ARG0, Q8SA57, P50701, Q5XPJ5, Q9FT35, Q8LPU1, P12670, B2CZJ9, P50702, P14170, Q94JN9, P25871, Q75W82, Q7Y1P9, P50703, Q6X1B8, Q9M3X2, Q9SPE0, A9QVJ4, Q8LSM9, P81370, Q93XD4, Q93XD2, B6ZHC0, Q2VAD0, P50700, Q7XA89, Q8LEV9, P13046, P07052, O04708, A2T1L9, Q8GUQ2, Q7XAU7, A3QRB4, A3QRB5, P93621, A9ZMG0, A9ZMG2, Q0QJL7, Q2VAC9, P33679, Q946Y8, O04364, A1IIJ1, P02883, P02884, A1IGE2, Q69CS6, Q69CS5, P81295, Q69CS3, Q69CS2, A4PBQ1, Q5RZ93, Q4W6C7, A2TEG9, A2TEG5, A2TEG7, A2TEH0, A2TEG8, Q8H995, Q8H996, A5HIY3, Q9ZRS9, Q5QJU2, Q5DWG2, Q946Y9, Q8S4P7, Q946Z0, O23997, Q5MBN2, Q38769, Q5SFH8, O81927, P32937, P32938, P27357, Q94F70, Q9S776, B6TA80, P50695, P50696, P50697, O82087, P50698, P31110, B7SDH2, Q946Z1, Q3S4I2, Q41584, Q8LKF7, P83335, P50694, Q3BCT4, Q00MX5, Q9FSG7, B6E2B4, P83332, O80327, Q7X9V4, Q9SMH2, Q53MB8, Q9FLD4, Q8GTA0, Q2VAC8, Q41350, B6SKP5, B5G511, Q9SNY0, P28493, Q9C9P9, O24468, Q1PFD2, Q9LN66, Q5DWG1, A8IXE7, Q8LE09, O49965, P50699, Q9LQT4, A8IXG8, B9DGE1, O65638, B6TEB6, Q10P77, Q9STX6, B3LFC8, Q38925, Q7F267.

PR06

Q02214, Q40416, Q03199, Q03198, P20076, P08454, P16231, P05118, Q8GT64, P80211, P01053, Q40059, P82977, P16062, P16063.

PR07

O04678, Q96478, C3PTS6, B6SWF4, B6SYH0, Q9LRF2, Q01H99, Q0D3H9, Q0DR00, Q39547, Q9ZTT3, Q8L7D2, Q9FGU3, Q9LLL8, A9XG41, A9XG40, O65351, P93204, B6U0R8, P93205, B6SZ82, Q8LGA0, Q9LPD1, C3VDI0, B6U4E9, Q93WQ0, Q40764, O82777, O82006, Q38708, Q0WVJ9, Q1PDX5, Q9SVT4, Q9SPA0, P93221, Q0JBB7, Q68Q08, B7ETP3, B6SWW5, Q948Q4, O48798.

PR08

B5M495, P51614, Q84S31, Q9FS44, P23472, Q71HN4, P19172, Q8LP09, Q8L4Q9, Q8LP01, Q8L4Q8, Q8LP04, Q8LP02, O22076, O22074, Q8LP06, A2TJX5, Q43684, Q9SC03, Q8H6X7, Q8H6X6, Q8GRR2, Q9SP41, O49827, O04139, B6TVA3, Q8LST3, Q00MX4, B9VQ33, Q9SXM5, Q9S7G9, Q43098, Q9XHC3, P29024, P36908, Q9XGB4, P29060, P36910, Q9FUD7, Q19AL0, A1YTJ5, Q6XD74, Q9XF93, Q84U85, Q09Y38, Q945U2, Q9SQI0, P17541, Q9M544, Q39656, Q39657, Q06SN0, B3A042, P93518, Q7XZD6, P29061, Q9ZSR4, Q6QUK8, A4GU13, C3VM17, Q6WDV9, Q41401, B8LF40, O48642, O24368, Q9MBC9, Q7GCM2, Q9SXY3, Q7GCM0, Q7GCM7, A7BJ77, A7BJ78, Q0DJP1, Q53NL5, B6U2X8, Q4W6G2, Q2VST0.

PR09

Q43102, B9GYK0, Q9LEH3, Q5JBR5, Q0ZA88, B5U1R2, Q43099, Q40949, Q43101, Q8RVP3, Q0ZA67, Q42905, Q8S3U4, Q42578, Q9FG34, Q9XFL6, P11965, Q8W174, Q43774, Q43790, Q43791, Q40365, Q18PQ8, Q18PQ7, O23961, Q8GZS1, O24081, Q18PQ9, Q40366, O22443, Q18PR0, A4UN77, Q9SMU8, P24101, P00433, Q4PJU0, A9XEK4, Q9LHB9, Q9LDA4, Q9LDN9, O80912, P24102, B6T7B1, P59120, O65773, Q6UBM4, Q39652, Q40559, Q6UNK7, Q39034, Q42517, Q9XGV6, Q58GF4, Q18PR1, B3V2Z3, B6SNF9, P93548, B6SMR2, Q5I3F2, Q9FLC0, B9VRK9, Q9XIV9, Q5JBR1, B3SHI1, Q7XYR7, Q84ZT5, B6U6W0, Q6T1D0, Q8RVP7, Q84ZT7, C0KXH4, Q9LVL2, Q9LVL1, Q9XFL4, Q43212, Q5I3F7, P27337, Q40068, Q5I3F6, Q5I3F5, Q43218, B4F6F1, Q05855, B4F6E7, B4F6F0, B4F6F2, B4F6E5, B4F6E6, Q0D3N0, Q7F1U0, O22438, Q43006, Q9LKY9, Q7F1U1, O22439, O22440, Q5I3F4, Q43417, Q43220, A5H452, Q41577, A5H454, A5H453, Q4W2V2, Q4W2V3, Q4W2V4, Q4W2V5, Q0JW34, P22195, Q7XMP4, A0SWU6, Q9SSZ7, B5U1R3, B6SRR3, B6TU39, B6U2M7, Q02200, A4ZCI6, Q1AJZ4, O49192, Q9SSZ8, O49193, Q9M9Q9, Q9LE15, Q4W1I8, Q0JW35, Q0JW36, Q5W5I3, Q9SC55, Q94IQ0, Q96512, Q9SI16, Q9SI17, O23237, B6E1W9, Q0ZA68, B0ZC10, Q5JBR3, Q5JBR2, B3SHI2, Q9FJZ9, Q9M4Z4, Q9MAX9, Q0JM38, Q9M4Z2, Q9M4Z3, A7J0U4, Q8H285, Q7F936, Q9SD46, A8W7V9, O23474, Q9SLH7, Q9FX85, Q8RVP6, Q9SJZ2, B6T750, Q5I3F1, Q96519, Q682W9, Q50LG4, Q50LG5, P12437, Q5JBR4, O04796, O04795, Q5JBR6, Q9SQ62, Q5U1S8, B6U0D4, Q5I3F0, A5H8G4, B6T173, Q9M4Z5, Q9SSZ9, Q10SI9, Q5U1M4, Q5U1P7, Q5U1P8, Q9FEQ8, Q0J459, B6U0T8, Q96518, Q96522, Q52QY2, Q84UA9, A0S5Z4, B6E500, O24080, Q9ZP15, Q9SZE7, Q43731, Q43873, Q96510, C0KKH7, Q96509, B6UI45, B6UBB5, B6T5R9, B6TYF5, Q9FL16, Q24JM5, Q9LHA7, O23609, Q9FJR1, O48677, Q7X8H7, B6SI04, O22959, Q8W4V8, Q9XIV8, C1KA97, Q4ADU9, Q8RVP4, Q8RVW0, P22196, Q43387, Q9FKA4, O80822, Q4A3Z3, Q4A3Z2, Q9FMI7, Q96511, Q5QEB4, Q40486, Q40487, Q0IMX5, Q9ZV04, B6SRH9, Q1H8N1, P93546, Q9FMR0, Q9SK52, O81772, Q9SY33, Q43729, Q9SS67, Q93V93, Q9LNL0, Q1PGA3, B9GQQ9, Q43872, Q9LT91, Q58A85, C0KKH9, Q9SZB9, B3SHI0, O23044, Q56V16, Q9LSY7, Q9SUT2, Q07446, Q07445, Q66RM0, Q43158, P37834, B6TMY7, B6SIA9, Q43735, Q9LXG3, Q96506, B6UB27, B6TWB1, Q8RVP5, Q9SZH2, Q5K4K5, Q5U1G1, O49293, Q43032, C0KKH8, B2G335, B2G334, Q0VYC8, Q94IQ1, Q43854, B7UCP4, Q43782, Q401B7, P93547, P93545, Q5W5I4, B9VSG0, Q96520, Q6PQF2, Q84U03, Q40069, B6THG0, Q5W5I2, Q9FYS6, O49866, Q5GMP4, B1A9R4, O22510, Q42580, C1KA92, Q9FXL6, Q9XFL2, O64970, Q9XFI6, Q08671, A3FPF7, Q9FT05, A9XN55, Q9SB81, A0S7R2, Q8H958, Q9LSP0, B6SU07.

PR10

Q2I6V8, B5KVN9, B5KVP1, O24248, Q40280, Q9SYV4, Q9SYV9, Q9SYV3, Q9SYV2, Q9SYW3, P43211, Q9SYV6, Q9SYV7, Q9SYV5, Q9SYV8, O65200, Q5VJR0, Q5VJR1, Q5VJQ9, Q5VJQ8, Q5VJR2, Q5VJR3, Q5VJR5, Q5VJR4, Q6QHU2, Q6QHU1, Q6QHU3, O50001, Q5VJQ7, P43184, P43176, P45431, P43186, Q9ZS38, Q9ZS39, Q9SCH6, O23753, O23751, P43177, Q9SCH5, P43180, Q96365, Q9SCI3, Q9SYW0, Q9SCI0, Q96367, Q96368, Q9AYS2, Q9AYS3, P43185, Q9AYS4, Q96370, O23752, O24642, Q96366, P15494, Q96371, Q42499, Q9SCI2, Q9SCH9, Q9SCH8, Q9SYW1, P43183, P43179, P43178, Q9SYW2, Q39417, P38948, Q96382, Q96381, Q96379, Q96377, Q96378, B6RQR9, B6RQR7, Q96503, B6RQR6, B6RQR8, B6RQS0, P38949, Q96501, Q96380, Q08407, P38950, B7TWE6, B7TWE8, B7TWE7, Q9ZRU8, Q9SWR4, Q9FPK2, Q9FPK3, Q9FPK4, Q39415, Q7Y083, B6RQS2, B6RQS3, B7TWE3, B7TWE5, B7TWE4, B6RQS1, Q2I305, B7SL50, A9CSL9, Q9MB25, Q9M500, Q2VT55, Q5DUH6, Q4KYL1, P17642, P17641, Q53U35, Q9FE19, Q6Q4B3, Q9FUI5, B9VRH3, Q9FUI6, Q6Q4B4, Q6XC94, Q945E7, P27538, P92918, Q8SAE7, P19417, P19418, Q40795, O04298, Q75T31, O81640, P49372, Q8S903, Q75T32, O49065, Q9LWB1, Q8W2B4, Q9LLQ3, Q9AXK1, Q93XI0, Q7Y1W5, Q9AXK2, Q9LLQ2, O24010, Q39450, Q06930, P14710, P13239, P27047, P93333, Q43560, Q40320, P26987, Q43453, P25986, Q9SXX8, A7LNN6, Q41711, Q2VU97, P25985, B2ZGS2, Q6VT83, Q0PKR4, P52779, Q9SPB2, P52778, Q8W1M7, Q6YNP8, Q06931, Q8L6K8, B0YIU5, Q5USC6, Q5USC5, Q5USC4, B5M1X5, B5M1X6, B7U9Z2, Q8H1L1, Q40154, Q9ZTP6, A0FIJ6, Q7X9W3, Q7X9W8, Q7X9W6, Q7X9W5, Q7X9W2, Q7X9W7, Q7X9W4, Q7X9W1, O81134, Q7X9W0, Q7X9V8, Q7X9V9, Q7X9V7, Q5YBE7, Q5YBE8, Q5YBE9, Q5YBF0, B3TM18, B3TLX2, Q96233, B5B3P8, Q945E9, B6SXF5, Q41298, Q9SEY7, O24453, Q9ZPP9, Q9SEY9, Q9ZPP8, Q9SB87, Q05736, Q9ZRR2, Q9ZRR3, Q9ZRR1, Q5YJR3, B1Q190, P93330, Q40707, Q2QNT0, Q5VJQ6, Q84QC7, Q93VR4, A8IXG5, Q9FYU3, Q9ZWP8, Q19VG6, Q5XLE0.

PR11

Q43591, Q43576, O81862, O81863, O81861, B9VQ35, A7R1P5, A7R1P1, B9HAQ3, B9SBZ9, A1YZD2, B7FNG8, A7QRZ3, A7QRZ5, A7QRZ6, Q84N00, Q4W6L6, B4F9H4, A2ZE15.

PR12

P69241, Q94IN7, P30230, P30224, Q39313, Q9FS38, Q5KU48, O24331, Q9FI23, O80995, Q8H6Q0, Q8GTM0, P32026, O24115, Q43413, Q01784, P81929, Q8H6Q1, Q8W4V6, B2CM18, O65740, Q40901, Q39182, Q9ZUL7, Q41914, C1K3M7, P82659, A3FPF2, Q9C947, Q01783, Q8W434, P18646, Q40128, B3F051, P82784, Q19JA1, Q9ZUL8, Q670N7, P82782, P82787, Q9FFP8, B2CNV2, Q8L698, Q9FZ31, Q84ZX5, B6SQK6, Q9SEM1.

PR13

P09618, P08772, P09617, Q42838, Q8LSZ9, Q8LT00, Q8LT04, Q5Z4K0, Q0DBX2, B7F9C4, Q8LT03, Q8LT02, Q8LT01, Q43225, Q43226, Q43227, Q43224, Q41609, Q9SBK8, Q42597, A1Z1S5, Q42596, B5M1X2, Q38L62, B8YLY8, B8YM12, P01543, B8YLZ1, B8YM08, B8YM20, B8YLZ3, B8YM05, B8YLZ6, B8YM21, B8YM03, Q9T0P2, B8YM15, P01544, P32032, Q43205, Q9ZNY5, P21742, P01545, Q05806, Q38770, Q41585

PR14

Q9S9G0, Q9S9F9, Q9S9G1, Q42589, Q9S7I3, Q9XFS7, P93224, P27056, P10976, Q9M5X8, Q9M5X7, Q9M5X6, Q8S4Y3, A0AT29, A0AT32, Q153I9, Q9LLR7, Q9LLR6, Q9SW93, Q43129, Q43019, Q8H2B3, C0L0I5, P27631, B6SGP7, P19656, O24583, B6SY96, B6T089, Q2QYL2, A2ZHF1, B6SJ07, B8A3E0, B6SKH5, A2ZAS9, B4FB54, Q8W533, Q2PCB6, Q84N29, Q5NE29, Q42848, Q93Z88, Q42842, Q43875, Q43767, Q43871, A9NUI4, A9NLY0, A9NP77, A9NY87, A9NKX7, B8LRP3, A9NKD5, Q9AXZ6, Q41073, A9NLQ3, A9NJW4, A9NY14, A9NJW5, A9NKV1, B6SIF2, C0P9Y4, Q9SES6, Q2QKE7, Q2V3C1, Q9LDB4, Q9ZUK6, B6T3G3, B6U968, B6U964, Q9ZPW9, B6TRH6, B6TLQ7, Q0JPJ4, Q0WYX3, Q0WYX5, P10975, Q07A25, O04004, O24485, Q5NE32, Q6AWW0.

PR15 and PR16

P15290, C3UZE7, P45851, Q70PK0, C3UZE6, C3UZE9, C3UZE8, P93600, Q0GR11, O24004, P45850, P26759, Q8L697, Q8L696, Q9FEW6, Q851K1, Q851K0, Q851J8, Q6YZZ6, Q6YZZ7, Q6ZCR3, B6TVW2, Q6YZ97, Q6YZA1, Q6YZZ2, Q6YZA4, Q6YZ99, Q6YZA6, Q6YZY5, Q0GR07, Q43487, Q0GR08, Q6DQK3, Q9SM35, Q9SM34, B6UEL1, Q6YZA9, B6UGC5, B7U512, Q2QXJ2, Q2QXJ4, B3TLX7, Q9SFF9, Q9M8X6, A8QK90, A8QK89, Q6JVN1, Q9LEA7, P92999, P92996, Q9FIC8, Q9FIC6, Q9FIC9, Q9FL89, Q9FID0, P92997, Q9FMA8, Q9FMA9, Q7XSN6, B6TTY1, Q9ST00, P45852, Q93WX8, Q9M8X5, Q9M8X4, Q9M8X3, Q942A7, Q10BU2, Q75HJ4, Q8H2A6, Q94JF3, P92995, Q9LMC9, Q9S772, Q9S8P4, A7LIS6, Q9M263, Q9FZ27, Q6I544, B6TWC3, Q942A8, B6TWH9, B6TKA5, B6TKE1, B6TU44, Q7F731, Q0GR06, Q9SPV5, Q94EG3, O64439, Q5DT23, Q5VJG4, O65358, B5U961, B5U962, P93000, Q8LEQ3, Q5KT31, Q8GSQ5, Q5KT25, Q5KSB5, O65010, Q49SH2, O49135, Q2R352, Q8H021, P92998, Q6ESF0, B6TF80, Q9SR72, Q4A3V0, Q5KSC2, Q5KSC1, Q5KSC4, Q5KSC3, Q652P9, Q84XR7, Q652Q1, Q7Y255, B5G500, Q9ZRA4, O04012, O04011, A7Y2G2, B9A6I8, Q9AR81, P94072, P45854, Q7XZV3, P94040, Q8VZ99, P46271, Q0GR10, O49871, Q8L686, Q6ZBZ2, Q6Z964, B6T6K1, B6U6C5, A7LIS5, Q9M3Y4, Q1H8M7, Q84RC0, Q9FPQ1, Q84V63, Q9FPQ0, Q9FPP9.

PR17

Q84XQ4, Q9XIY9, Q94I89, B9HS69, B9GHT3, B9RY80, A5BS35, A7QW51, Q9SKL6, Q9ZUJ8, A5BS34, Q84ZV0, Q84ZU9, B9HS70, Q41523, A7YA66, A7YA60, Q9SWZ5, B6TDW7, O24003, Q7XD55, A2Z8T9, A2Z8U1, Q9FWU4, O24002, B8BHL9.

Bibliography

[1] L. C. Van Loon and A. Van Kammen. Polyacrylamide disc electrophoresis of the soluble leaf proteins from *Nicotiana tabacum* var. "Samsun" and "Samsun NN". II. Changes in protein constitution after infection with tobacco mosaic virus. *Virology*, 40(2):190–211, February 1970. ISSN 0042-6822.

[2] J. F. Antoniw, C. E. Ritter, W. S. Pierpoint, and L. C. Van Loon. Comparison of three pathogenesis-related proteins from plants of two cultivars of tobacco infected with tmv. *J Gen Virol*, 47(1):79–87, March 1980.

[3] J. F. Bol, H. J. M. Linthorst, and B. J. C. Cornelissen. Plant pathogenesis-related proteins induced by virus infection. *Annual Review of Phytopathology*, 28(1):113–138, 1990. doi: 10.1146/annurev.py.28.090190.000553.

[4] L. C. Van Loon, W. S. Pierpoint, T. Boller, and V. Conejero. Recommendations for naming plant pathogenesis-related proteins. *Plant Molecular Biology Reporter*, 12(3):245–264, 1994. doi: 10.1007/BF02668748.

[5] S. Kitajima and F. Sato. Plant pathogenesis-related proteins: Molecular mechanisms of gene expression and protein function. *Journal of Biochemistry*, 125(1):1–8, January 1999.

[6] C. P. Selitrennikoff. Antifungal proteins. *Appl. Environ. Microbiol.*, 67(7):2883–2894, July 2001. doi: 10.1128/AEM.67.7.2883-2894.2001.

[7] A. Edreva. Pathogenesis-related proteins: research progress in the last 15 years. *General and Applied Plant Physiology*, 31(1-2):105–124, 2005.

[8] L. C. Van Loon, M. Rep, and C. M. J. Pieterse. Significance of inducible defense-related proteins in infected plants. *Annual Review of Phytopathology*, March 2006. ISSN 0066-4286. doi: 10.1146/annurev.phyto.44.070505.143425.

[9] J. A. Ryals, U. H. Neuenschwander, M. G. Willits, A. Molina, H. Y. Steiner, and M. D. Hunt. Systemic acquired resistance. *Plant Cell*, 8(10):1809–1819, October 1996. ISSN 1040-4651. doi: 10.1105/tpc.8.10.1809.

[10] P. Reignault and M. Sancholle. Plant-pathogen interactions: will the understanding of common mechanisms lead to the unification of concepts? *Comptes Rendus Biologies*, 328(9):821–833, September 2005. doi: 10.1016/j.crvi.2005.07.002.

[11] L. Király, B. Barna, and Z. Király. Plant resistance to pathogen infection: Forms and mechanisms of innate and acquired resistance. *Journal of Phytopathology*, 155(7-8):385–396, August 2007. ISSN 0931-1785. doi: 10.1111/j.1439-0434.2007.01264.x.

[12] N. Kavroulakis, K. K. Papadopoulou, S. Ntougias, G. I. Zervakis, and C. Ehaliotis. Cytological and other aspects of pathogenesis-related gene expression in tomato plants grown on a suppressive compost. *Annals of botany*, 98(3):555–564, September 2006. ISSN 0305-7364. doi: 10.1093/aob/mcl149.

[13] E. D. Schulze, E. Beck, and K. Müller-Hohenstein. *Plant ecology.* Springer, 2005.

[14] A. Koornneef and C. M. Pieterse. Cross talk in defense signaling. *Plant physiology*, 146(3):839–844, March 2008. ISSN 0032-0889. doi: 10.1104/pp.107.112029.

[15] S. J. Zheng and M. Dicke. Ecological genomics of plant-insect interactions: From gene to community. *Plant Physiol.*, 146(3):812–817, March 2008. doi: 10.1104/pp.107.111542.

[16] J. Browse and G. A. Howe. New weapons and a rapid response against insect attack. *Plant physiology*, 146(3):832–838, March 2008. ISSN 0032-0889. doi: 10.1104/pp.107.115683.

[17] O. Gutierrez, M. Wubben, M. Howard, B. Roberts, E. Hanlon, and J. Wilkinson. The role of phytohormones ethylene and auxin in plant-nematode interactions. *Russian Journal of Plant Physiology*, 56(1):1–5, January 2009. doi: 10.1134/S1021443709010014.

[18] M. Ghosh. Antifungal properties of haem peroxidase from *Acorus calamus*. *Ann Bot*, 98(6): 1145–1153, December 2006. ISSN 0305-7364. doi: 10.1093/aob/mcl205.

[19] Recognized families of pathogenesis-related proteins. Website. URL `http://www.bio.uu.nl/~{}fytopath/PR-families.htm`.

[20] L. C. Van Loon and E. A. Van Strien. The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins. *Physiological and Molecular Plant Pathology*, 55 (2):85–97, August 1999. doi: 10.1006/pmpp.1999.0213.

[21] F. Passardi, D. Longet, C. Penel, and C. Dunand. The class III peroxidase multigenic family in rice and its evolution in land plants. *Phytochemistry*, 65(13):1879–1893, July 2004. ISSN 00319422. doi: 10.1016/j.phytochem.2004.06.023.

[22] K. G. Welinder, A. F. Justesen, I. V. Kjaersgård, R. B. Jensen, S. K. Rasmussen, H. M. Jespersen, and L. Duroux. Structural diversity and transcription of class iii peroxidases from arabidopsis thaliana. *European journal of biochemistry / FEBS*, 269(24):6063–6081, December 2002. ISSN 0014-2956.

[23] A. Lawton-Rauh. Evolutionary dynamics of duplicated genes in plants. *Molecular Phylogenetics and Evolution*, 29(3):396–409, December 2003. doi: 10.1016/j.ympev.2003.07.004.

[24] C. R. Darwin. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* John Murray, London, first edition, 1859.

[25] J. S. Taylor and J. Raes. Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics*, 38(1):615–643, 2004. ISSN 0066-4197. doi: 10.1146/annurev.genet.38.072902.092831.

[26] M. Lynch. Genomics. gene duplication and evolution. *Science (New York, N.Y.)*, 297(5583): 945–947, August 2002. ISSN 1095-9203. doi: 10.1126/science.1075472.

[27] A. L. Hughes. The evolution of functionally novel proteins after gene duplication. *Proceedings. Biological sciences / The Royal Society*, 256(1346):119–124, May 1994. ISSN 0962-8452. doi: 10.1098/rspb.1994.0058.

[28] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, April 1999. ISSN 0016-6731.

[29] M. Lynch, M. O'Hely, B. Walsh, and A. Force. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4):1789–1804, December 2001. ISSN 0016-6731.

[30] J. L. Bennetzen. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica*, 115(1):29–36, May 2002. doi: 10.1023/A:1016015913350.

[31] J. J. Doyle, L. E. Flagel, A. H. Paterson, R. A. Rapp, D. E. Soltis, P. S. Soltis, and J. F. Wendel. Evolutionary genetics of genome merger and doubling in plants. *Annual Review of Genetics*, 42 (1):443–461, 2008. doi: 10.1146/annurev.genet.42.110807.091524.

[32] J. Masterson. Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science*, 264(5157):421–424, April 1994. ISSN 1095-9203. doi: 10.1126/science.264.5157.421.

[33] D. E. Soltis, P. S. Soltis, and J. A. Tate. Advances in the study of polyploidy since *Plant speciation*. *New Phytologist*, 161(1):173–191, 2004. ISSN 1469-8137. doi: 10.1046/j.1469-8137.2003.00948.x.

[34] E. A. Kellogg and J. L. Bennetzen. The evolution of nuclear genome structure in seed plants. *Am. J. Bot.*, 91(10):1709–1725, October 2004.

[35] J. F. Wendel. Genome evolution in polyploids. *Plant Molecular Biology*, 42(1):225–249, January 2000. doi: 10.1023/A:1006392424384.

[36] S. Anssour, T. Krugel, T. F. Sharbel, H. P. Saluz, G. Bonaventure, and I. T. Baldwin. Phenotypic, genetic and genomic consequences of natural and synthetic polyploidization of nicotiana attenuata and nicotiana obtusifolia. *Ann Bot*, 103(8):1207–1217, June 2009. doi: 10.1093/aob/mcp058.

[37] W. Fitch. Homology a personal view on some of the problems. *Trends in Genetics*, 16(5):227–231, May 2000. ISSN 01689525. doi: 10.1016/S0168-9525(00)02005-9.

[38] E. V. Koonin. Orthologs, paralogs, and evolutionary genomics1. *Annual Review of Genetics*, 39 (1):309–338, 2005. ISSN 0066-4197. doi: 10.1146/annurev.genet.39.073003.114725.

[39] M. Nei. Selectionism and neutralism in molecular evolution. *Mol Biol Evol*, 22(12):2318–2342, December 2005. doi: 10.1093/molbev/msi242.

[40] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, February 1968. ISSN 0028-0836.

[41] L. Patthy. *Protein evolution*. Blackwell Publishing, Oxford, 1999. ISBN 0-632-04774-7.

[42] T. Ohta. Role of diversifying selection and gene conversion in evolution of major histocompatibility complex loci. *Proceedings of the National Academy of Sciences of the United States of America*, 88(15):6716–6720, August 1991.

[43] R. Apweiler, A. Bairoch, and C. H. Wu. Protein sequence databases. *Curr Opin Chem Biol*, 8(1): 76–80, February 2004. ISSN 1367-5931. doi: 10.1016/j.cbpa.2003.12.004.

[44] The-Uniprot-Consortium. The universal protein resource (uniprot). *Nucl. Acids Res.*, 36(suppl_1): D190–195, January 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm895.

[45] C. O'Donovan, M. J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, and R. Apweiler. High-quality protein knowledge resource: Swiss-prot and trembl. *Briefings in bioinformatics*, 3(3): 275–284, September 2002. ISSN 1467-5463.

[46] A. Bairoch, B. Boeckmann, S. Ferro, and E. Gasteiger. Swiss-prot: Juggling between evolution and stability. *Brief Bioinform*, 5(1):39–55, January 2004. ISSN 1467-5463. doi: 10.1093/bib/5.1.39.

[47] B. Boeckmann, M. Blatter, L. Famiglietti, U. Hinz, L. Lane, B. Roechert, and A. Bairoch. Protein variety and functional diversity: Swiss-prot annotation in its biological context. *Comptes Rendus Biologies*, 328(10-11):882–899, October 2005. ISSN 16310691. doi: 10.1016/j.crvi.2005.06.001.

[48] M. Schneider, L. Lane, E. Boutet, D. Lieberherr, M. Tognolli, L. Bougueleret, and A. Bairoch. The uniprotkb/swiss-prot knowledgebase and its plant proteome annotation program. *Journal of proteomics*, 72(3):567–573, April 2009. ISSN 1876-7737. doi: 10.1016/j.jprot.2008.11.010.

[49] W. C. Barker, D. G. George, H. W. Mewes, F. Pfeiffer, and A. Tsugita. The pir-international databases. *Nucleic acids research*, 21(13):3089–3092, July 1993. ISSN 0305-1048.

[50] PIR-PSD Database [PIR - Protein Information Resource], 2009. URL `http://pir.georgetown.edu/pirwww/dbinfo/pir_psd.shtml`.

[51] UniProtKB/TrEMBL Release 40.4 Statistics, 2009. URL `http://www.ebi.ac.uk/uniprot/TrEMBLstats/`.

[52] UniProtKB/Swiss-Prot Release 57.4 Statistics, 2009. URL `http://www.expasy.org/sprot/relnotes/relstat.html`.

[53] M. Schneider, M. Tognolli, and A. Bairoch. The swiss-prot protein knowledgebase and expasy: providing the plant community with high quality proteomic data and tools. *Plant physiology and biochemistry : PPB / Société française de physiologie végétale*, 42(12):1013–1021, December 2004. ISSN 0981-9428. doi: 10.1016/j.plaphy.2004.10.009.

[54] Plant Proteome Annotation Program (PPAP), 2009. URL `http://www.expasy.org/sprot/ppap/`.

[55] N. M. Scherer and D. M. Basso. DNATagger, colors for codons. *Genetics and Molecular Research : GMR*, 7(3):853–860, 2008. ISSN 1676-5680. doi: 10.4238/vol7-3X-Meeting003.

[56] P. G. Higgs and T. K. Attwood. *Bioinformatics and Molecular Evolution*. Blackwell Science Ltd., Malden, 2005. ISBN 1-4051-0683-2.

[57] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, April 1998. doi: 10.2277/0521629713.

[58] W. J. Ewens and G. Grant. *Statistical Methods in Bioinformatics : An Introduction*. Springer, April 2001.

[59] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, March 1970. ISSN 0022-2836.

[60] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981. ISSN 00222836. doi: 10.1016/0022-2836(81)90087-5.

[61] H. Carrillo and D. Lipman. The multiple sequence alignment problem in biology. *SIAM Journal on Applied Mathematics*, 48(5):1073–1082, 1988. ISSN 00361399. doi: 10.2307/2101469.

[62] D. J. Lipman, S. F. Altschul, and J. D. Kececioglu. A tool for multiple sequence alignment. *PNAS*, 86(12):4412–4415, June 1989.

[63] S. F. Altschul, R. J. Carroll, and D. J. Lipman. Weights for data related by a tree. *J Mol Biol*, 207(4):647–653, June 1989. ISSN 0022-2836.

[64] O. Gotoh. A weighting system and algorithm for aligning many phylogenetically related sequences. *Computer applications in the biosciences : CABIOS*, 11(5):543–551, October 1995. ISSN 0266-7061.

[65] G. H. Gonnet, C. Korostensky, and S. Benner. Evaluation measures of multiple sequence alignments. *Journal of computational biology : a journal of computational molecular cell biology*, 7(1-2):261–276, 2000. ISSN 1066-5277. doi: 10.1089/10665270050081513.

[66] U. Tönges, S. W. Perrey, J. Stoye, and A. W. Dress. A general method for fast multiple sequence alignment. *Gene*, 172(1), June 1996. ISSN 0378-1119.

[67] J. Stoye. *Divide-and-Conquer Multiple Sequence Alignment*. PhD thesis, Technische Fakultät der Universität Bielefeld, Abteilung Informationstechnik, 1997.

[68] J. Stoye. Multiple sequence alignment with the divide-and-conquer method. *Gene*, 211(2):GC45–GC56, May 1998. ISSN 03781119. doi: 10.1016/S0378-1119(98)00097-3.

[69] B. Morgenstern, A. Dress, and T. Werner. Multiple dna and protein sequence alignment based on segment-to-segment comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 93(22):12098–12103, October 1996. doi: 10.1073/pnas.93.22.12098.

[70] D. F. Feng and R. F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 25(4):351–360, 1987. ISSN 0022-2844.

[71] C. Grasso and C. Lee. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics*, 20(10):1546–1556, July 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth126.

[72] T. Lassmann and E. L. Sonnhammer. Quality assessment of multiple alignment programs. *FEBS Lett*, 529(1):126–130, October 2002. ISSN 0014-5793.

[73] T. Lassmann and E. L. L. Sonnhammer. Automatic assessment of alignment quality. *Nucleic acids research*, 33(22):7120–7128, 2005. ISSN 1362-4962. doi: 10.1093/nar/gki1020.

[74] P. P. Gardner, A. Wilm, and S. Washietl. A benchmark of multiple sequence alignment programs upon structural rnas. *Nucleic Acids Research*, 33(8):2433–2439, April 2005. ISSN 0305-1048. doi: 10.1093/nar/gki541.

[75] G. Blackshields, I. M. Wallace, M. Larkin, and D. G. Higgins. Analysis and comparison of benchmarks for multiple sequence alignment. *In silico biology*, 6(4):321–339, 2006. ISSN 1386-6338.

[76] P. Nuin, Z. Wang, and E. Tillier. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, 7(1):471+, October 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-471.

[77] T. Golubchik, M. J. Wise, S. Easteal, and L. S. Jermiin. Mind the gaps: Evidence of bias in estimates of multiple sequence alignments. *Mol Biol Evol*, August 2007. ISSN 0737-4038. doi: 10.1093/molbev/msm176.

[78] C. Notredame. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol*, 3(8):e123+, August 2007. doi: 10.1371/journal.pcbi.0030123.

[79] J. D. Thompson, F. Plewniak, and O. Poch. Balibase: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics (Oxford, England)*, 15(1):87–88, January 1999. ISSN 1367-4803. doi: 10.1093/bioinformatics/15.1.87.

[80] S. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, October 1998. ISSN 1367-4803. doi: 10.1093/bioinformatics/14.9.755.

[81] S. Eddy. *The HMMER User's Guide*, October 2003. URL `http://hmmer.janelia.org/`.

[82] J. Dai and J. Cheng. Hmmeditor: a visual editing tool for profile hidden markov model. *BMC genomics*, 9 Suppl 1, 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-S1-S8.

[83] B. Schuster-Böckler, J. Schultz, and S. Rahmann. Hmm logos for visualization of protein families. *BMC Bioinformatics*, 5(1):7+, January 2004. ISSN 1471-2105. doi: 10.1186/1471-2105-5-7.

[84] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438, 1958.

[85] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, July 1987. ISSN 0737-4038.

[86] W. M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971. ISSN 00397989. doi: 10.2307/2412116.

[87] L. L. Cavalli-Sforza and A. W. Edwards. Phylogenetic analysis. models and estimation procedures. *American journal of human genetics*, 19(3 Pt 1):233–257, May 1967. ISSN 0002-9297.

[88] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981. ISSN 0022-2844.

[89] W. Messier and C. B. Stewart. Episodic adaptive evolution of primate lysozymes. *Nature*, 385 (6612):151–154, January 1997. ISSN 0028-0836. doi: 10.1038/385151a0.

[90] A. Hughes. Adaptive evolution after gene duplication. *Trends in Genetics*, 18(9):433–434, September 2002. ISSN 01689525. doi: 10.1016/S0168-9525(02)02755-5.

[91] M. Kimura. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608):275–276, May 1977. doi: 10.1038/267275a0.

[92] T. Miyata and T. Yasunaga. Molecular evolution of mrna: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *Journal of Molecular Evolution*, 16(1):23–36, March 1980. ISSN 0022-2844. doi: 10.1007/BF01732067.

[93] W. H. Li, C. I. Wu, and C. C. Luo. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*, 2(2):150–174, March 1985. ISSN 0737-4038.

[94] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3(5):418–426, September 1986. ISSN 0737-4038.

[95] Y. Suzuki and T. Gojobori. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol*, 16(10):1315–1328, October 1999.

[96] Z. Yang and J. P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12):496–503, December 2000. ISSN 01695347. doi: 10.1016/S0169-5347(00)01994-7.

[97] S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, 11(5):715–724, September 1994.

[98] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol Biol Evol*, 11(5):725–736, September 1994.

[99] Z. Yang. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*, 15(5):568–573, May 1998. ISSN 0737-4038.

[100] R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics*, 148(3):929–936, March 1998.

[101] Z. Yang and R. Nielsen. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of molecular evolution*, 46(4):409–418, April 1998. ISSN 0022-2844.

[102] Z. Yang, R. Nielsen, N. Goldman, and A-M K. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449, May 2000.

[103] Z. Yang and R. Nielsen. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19(6):908–917, June 2002. ISSN 0737-4038.

[104] Z. Yang and R. Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, 17(1):32–43, January 2000. ISSN 0737-4038.

[105] J. Zhang, R. Nielsen, and Z. Yang. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, 22(12):2472–2479, December 2005. doi: 10.1093/molbev/msi237.

[106] S. L. K. Pond and S. D. W. Frost. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Molecular Biology and Evolution*, 22(3):478–485, March 2005. ISSN 0737-4038. doi: 10.1093/molbev/msi031.

[107] S. L. K. Pond and S. D. W. Frost. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*, 22(5):1208–1222, May 2005. ISSN 0737-4038. doi: 10.1093/molbev/msi105.

[108] Z. Yang and R. Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*, 25(3):568–579, March 2008. ISSN 0737-4038. doi: 10.1093/molbev/msm284.

[109] J. B. Johnson and K. S. Omland. Model selection in ecology and evolution. *Trends in Ecology & Evolution*, 19(2):101–108, February 2004. ISSN 01695347. doi: 10.1016/j.tree.2003.10.013.

[110] J. P. Huelsenbeck and B. Rannala. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science*, 276(5310):227–232, April 1997. ISSN 0036-8075. doi: 10.1126/science.276.5310.227.

[111] H. Aikake. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Proceedings of 2nd International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, 1973.

[112] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, March 1978. ISSN 0090-5364. doi: 10.1214/aos/1176344136.

[113] J. Stoye, V. Moulton, and A. W. Dress. Dca: an efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. *Computer applications in the biosciences : CABIOS*, 13(6):625–626, December 1997. ISSN 0266-7061.

[114] T. Lassmann and E. L. Sonnhammer. Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6(1):298+, 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-298.

[115] T. Lassmann, O. Frings, and E. L. L. Sonnhammer. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic acids research*, 37(3):858–865, February 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn1006.

[116] R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, March 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh340.

[117] R. C. Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1):113+, August 2004. ISSN 1471-2105. doi: 10.1186/1471-2105-5-113.

[118] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, November 1994. ISSN 0305-1048. doi: 10.1093/nar/22.22.4673.

[119] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics (Oxford, England)*, 23(21):2947–2948, November 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm404.

[120] K. Katoh, K. Kuma, H. Toh, and T. Miyata. Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research*, 33(2):511–518, 2005. ISSN 1362-4962. doi: 10.1093/nar/gki198.

[121] K. Katoh and H. Toh. Recent developments in the mafft multiple sequence alignment program. *Briefings in bioinformatics*, 9(4):286–298, July 2008. ISSN 1477-4054. doi: 10.1093/bib/bbn013.

[122] A. R. Subramanian, J. Weyer-Menkhoff, M. Kaufmann, and B. Morgenstern. Dialign-t: An improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, 6(1): 66+, 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-66.

[123] A. R. Subramanian, M. Kaufmann, and B. Morgenstern. Dialign-tx: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for molecular biology : AMB*, 3(1):6+, May 2008. ISSN 1748-7188. doi: 10.1186/1748-7188-3-6.

[124] C. Lee, C. Grasso, and M. F. Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464, March 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.3.452.

[125] A. Loytynoja and N. Goldman. From the cover: An algorithm for progressive multiple alignment of sequences with insertions. *PNAS*, 102(30):10557–10562, July 2005. doi: 10.1073/pnas.0409137102.

[126] A. Loytynoja and N. Goldman. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320(5883):1632–1635, June 2008. ISSN 1095-9203. doi: 10.1126/science.1158395.

[127] Chuong B. Do, Mahathi S. P. Mahabhashyam, Michael Brudno, and Serafim Batzoglou. Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2):330–340, February 2005. ISSN 1088-9051. doi: 10.1101/gr.2821705.

[128] C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217, September 2000. ISSN 0022-2836. doi: 10.1006/jmbi.2000.4042.

[129] S. Moretti, F. Reinier, O. Poirot, F. Armougom, S. Audic, V. Keduas, and C. Notredame. Protogene: turning amino acid alignments into bona fide cds nucleotide alignments. *Nucleic acids research*, 34(Web Server issue), July 2006. ISSN 1362-4962. doi: 10.1093/nar/gkl170.

[130] O. Poirot, E. O'Toole, and C. Notredame. Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments. *Nucl. Acids Res.*, 31(13):3503–3506, July 2003. doi: 10.1093/nar/gkg522.

[131] F. Abascal, R. Zardoya, and D. Posada. Prottest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104–2105, May 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti263.

[132] D. Posada. jmodeltest: Phylogenetic model averaging. *Mol Biol Evol*, 25(7):1253–1256, July 2008. ISSN 1537-1719. doi: 10.1093/molbev/msn083.

[133] Stéphane Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5):696–704, October 2003. ISSN 1063-5157.

[134] O. Gascuel. Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Mol Biol Evol*, 14(7):685–695, July 1997. ISSN 0737-4038.

[135] M. Anisimova and C. Kosiol. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol*, 26(2):255–271, February 2009. ISSN 1537-1719. doi: 10.1093/molbev/msn232.

[136] J. Felsenstein. Phylip (phylogeny inference package) version 3.6. Distributed by the author, 2005.

[137] L. S. Vinh and A. von Haeseler. Iqpnni: moving fast through tree space and stopping in time. *Molecular biology and evolution*, 21(8):1565–1571, August 2004. ISSN 0737-4038. doi: 10.1093/molbev/msh176.

[138] H. A. Schmidt, K. Strimmer, M. Vingron, and A. von Haeseler. Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics (Oxford, England)*, 18(3):502–504, March 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.3.502.

[139] Z. Yang. Paml 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8):1586–1591, August 2007. ISSN 0737-4038. doi: 10.1093/molbev/msm088.

[140] W. S. W. Wong, Z. Yang, N. Goldman, and R. Nielsen. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, 168(2):1041–1051, October 2004. ISSN 0016-6731. doi: 10.1534/genetics.104.031153.

[141] M. Anisimova, J. P. Bielawski, and Z. Yang. Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol*, 19(6):950–958, June 2002.

[142] Z. Yang. *User Guide, PAML: Phylogenetic Analysis by Maximum Likelihood*, 4.3 edition, September 2009.

[143] P. Tornero, V. Conejero, and P. Vera. Primary structure and expression of a pathogen-induced protease (pr-p69) in tomato plants: Similarity of functional domains to subtilisin-like endoproteases. *Proceedings of the National Academy of Sciences of the United States of America*, 93(13):6332–6337, June 1996.

[144] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak. Improved prediction of signal peptides: Signalp 3.0. *Journal of molecular biology*, 340(4):783–795, July 2004. ISSN 0022-2836. doi: 10.1016/j.jmb.2004.05.028.

[145] B. Neron, H. Menager, C. Maufrais, N. Joly, J. Maupetit, S. Letort, S. Carrere, P. Tuffery, and C. Letondal. Mobyle: a new full web bioinformatics framework. *Bioinformatics*, 25(22):3005–3011, November 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp493.

[146] J. Archer and D. L. Robertson. CTree: comparison of clusters between phylogenetic trees made easy. *Bioinformatics*, 23(21):2952–2953, November 2007. doi: 10.1093/bioinformatics/btm410.

[147] Z. Yang, W. S. W. Wong, and R. Nielsen. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol*, 22(4):1107–1118, April 2005. ISSN 0737-4038. doi: 10.1093/molbev/msi097.

[148] M. Anisimova, J. P. Bielawski, and Z. Yang. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*, 18(8):1585–1592, August 2001. ISSN 0737-4038.

[149] K. Hoffmann-Sommergruber. Plant allergens and pathogenesis-related proteins. what do they have in common? *Int Arch Allergy Immunol*, 122(3):155–166, July 2000. ISSN 1018-2438.

[150] K. Hoffmann-Sommergruber. Pathogenesis-related (PR)-proteins identified as allergens. *Biochemical Society transactions*, 30(Pt 6):930–935, November 2002. ISSN 0300-5127. doi: 10.1042/.

[151] T. Szyperski, C. Fernández, C. Mumenthaler, and K. Wüthrich. Structure comparison of human glioma pathogenesis-related protein glipr and the plant pathogenesis-related protein p14a indicates a functional link between the human immune system and a plant defense system. *Proceedings of the National Academy of Sciences of the United States of America*, 95(5):2262–2266, March 1998. ISSN 0027-8424.

[152] P. B. Høj and G. B. Fincher. Molecular evolution of plant β-glucan endohydrolases. *The Plant Journal*, 7(3):367–379, 1995. ISSN 1365-313X. doi: 10.1046/j.1365-313X.1995.7030367.x.

[153] F. Hamel, R. Boivin, C. Tremblay, and G. Bellemare. Structural and evolutionary relationships among chitinases of flowering plants. *Journal of Molecular Evolution*, 44(6):614–624, June 1997. doi: 10.1007/PL00006184.

[154] P. Tiffin. Comparative evolutionary histories of chitinase genes in the genus Zea and family Poaceae. *Genetics*, 167(3):1331–1340, July 2004. ISSN 0016-6731. doi: 10.1534/genetics.104.026856.

[155] J. T. Christeller. Evolutionary mechanisms acting on proteinase inhibitor variability. *FEBS Journal*, 272(22):5710–5722, 2005. ISSN 1742-4658. doi: 10.1111/j.1742-4658.2005.04975.x.

[156] J. J. Liu and A. K. M. Ekramoddoullah. Characterization, expression and evolution of two novel subfamilies of *Pinus monticola* cDNAs encoding pathogenesis-related (PR)-10 proteins. *Tree Physiol*, 24(12):1377–1385, December 2004. doi: 10.1093/treephys/24.12.1377.

[157] L. B. Freitas, S. L. Bonatto, and F. M. Salzano. Evolutionary implications of intra- and interspecific molecular variability of pathogenesis-related proteins. *Brazilian journal of biology = Revista brasleira de biologia*, 63(3):437–448, August 2003. ISSN 1519-6984.

[158] J. D. Funkhouser and N. N. Aronson. Chitinase family GH18: evolutionary insights from the genomic history of a diverse protein family. *BMC Evol Biol*, 7, 2007. doi: 10.1186/1471-2148-7-96.

[159] S. Tuzun and A. Somanchi. The possible role of PR proteins in multigenic and induced systemic resistance. In S. Tuzun and E. Bent, editors, *Multigenic and Induced Systemic Resistance in Plants*, chapter 6, pages 112–142. Springer US, 2006. ISBN 978-0-387-23265-2. doi: 10.1007/0-387-23266-4_6.

[160] J. G. Bishop, A. M. Dean, and T. Mitchell-Olds. Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution. *PNAS*, 97(10):5322–5327, May 2000.

[161] J. G. Bishop, D. R. Ripoll, S. Bashir, C. M. B. Damasceno, J. D. Seeds, and J. K. C. Rose. Selection on glycine beta-1,3-endoglucanase genes differentially inhibited by a phytophthora glucanase inhibitor protein. *Genetics*, pages genetics.103.025098+, February 2005. doi: 10.1534/genetics.103.025098.

[162] J. Sels, J. Mathys, B. Deconinck, B. Cammue, and M. Debolle. Plant pathogenesis-related (pr) proteins: A focus on pr peptides. *Plant Physiology and Biochemistry*, 46(11):941–950, November 2008. ISSN 09819428. doi: 10.1016/j.plaphy.2008.06.011.

[163] K. Thevissen. Interactions of antifungal plant defensins with fungal membrane components. *Peptides*, 24(11):1705–1712, November 2003. ISSN 01969781. doi: 10.1016/j.peptides.2003.09.014.

[164] K. Thevissen, R. W. Osborn, D. P. Acland, and W. F. Broekaert. Specific binding sites for an antifungal plant defensin from dahlia (*Dahlia merckii*) on fungal cells are required for antifungal activity. *Molecular plant-microbe interactions : MPMI*, 13(1):54–61, January 2000. ISSN 0894-0282. doi: 10.1094/MPMI.2000.13.1.54.

[165] P. B. Pelegrini and O. L. Franco. Plant gamma-thionins: novel insights on the mechanism of action of a multi-functional class of defense proteins. *The international journal of biochemistry & cell biology*, 37(11):2239–2253, November 2005. ISSN 1357-2725. doi: 10.1016/j.biocel.2005.06.011.

[166] F. T. Lay and M. A. Anderson. Defensins–components of the innate immune system in plants. *Current protein & peptide science*, 6(1):85–101, February 2005. ISSN 1389-2037.

[167] K. Thevissen, D. C. Warnecke, I. E. J. A. Francois, M. Leipelt, E. Heinz, C. Ott, U. Zahringer, B. P. H. J. Thomma, K. K. A. Ferket, and B. P. A. Cammue. Defensins from insects and plants interact with fungal glucosylceramides. *J. Biol. Chem.*, 279(6):3900–3905, February 2004. doi: 10.1074/jbc.M311165200.

[168] E. Mendez, A. Moreno, F. Colilla, F. Pelaez, G. G. Limas, R. Mendez, F. Soriano, M. Salinas, and C. de Haro. Primary structure and inhibition of protein synthesis in eukaryotic cell-free system of a novel thionin, gamma-hordothionin, from barley endosperm. *European journal of biochemistry / FEBS*, 194(2):533–539, December 1990. ISSN 0014-2956.

[169] C. Kushmerick, M. de Souza Castro, J. Santos Cruz, C. Bloch, and P. S. Beirão. Functional and structural features of gamma-zeathionins, a new class of sodium channel blockers. *FEBS letters*, 440(3):302–306, December 1998. ISSN 0014-5793.

[170] A. M. Aerts, I. E. J. A. Francois, B. P. A. Cammue, and K. Thevissen. The mode of antifungal action of plant, insect and human defensins. *Cellular and Molecular Life Sciences CMLS*, 65(13): 2069–2079, July 2008. ISSN 1420-682X. doi: 10.1007/s00018-008-8035-0.

[171] C. Bloch. A new family of small (5 kda) protein inhibitors of insect alpha-amylases from seeds or sorghum (sorghum bicolor (l) moench) have sequence homologies with wheat gamma-purothionins. *FEBS Letters*, 279(1):101–104, February 1991. ISSN 00145793. doi: 10.1016/0014-5793(91) 80261-Z.

[172] R. E. Shade, H. E. Schroeder, J. J. Pueyo, L. M. Tabe, L. L. Murdock, T. J. V. Higgins, and M. J. Chrispeels. Transgenic pea seeds expressing the alpha-amylase inhibitor of the common bean are resistant to bruchid beetles. *Nat Biotech*, 12(8):793–796, 1994. doi: 10.1038/nbt0894-793.

[173] R. Wijaya, G. M. Neumann, R. Condron, A. B. Hughes, and G. M. Polya. Defense proteins from seed of cassia fistula include a lipid transfer protein homologue and a protease inhibitory plant defensin. *Plant science (Shannon, Ireland)*, 159(2):243–255, November 2000. ISSN 0168-9452.

[174] B. H. J. Thomma, B. P. A. Cammue, and K. Thevissen. Plant defensins. *Planta*, 216(2):193–202, December 2002. doi: 10.1007/s00425-002-0902-6.

[175] M. A. Graham, K. A. T. Silverstein, and K. A. VandenBosch. Defensin-like genes: Genomic perspectives on a diverse superfamily in plants. *Crop Sci*, 48(Supplement_1):S–3–11, February 2008. doi: 10.2135/cropsci2007.04.0236tpg.

[176] R. W. Osborn, G. W. De Samblanx, K. Thevissen, I. Goderis, S. Torrekens, F. Van Leuven, S. Attenborough, S. B. Rees, and W. F. Broekaert. Isolation and characterisation of plant defensins from seeds of asteraceae, fabaceae, hippocastanaceae and saxifragaceae. *FEBS letters*, 368(2):257–262, July 1995. ISSN 0014-5793.

[177] H. Bohlmann. The role of thionins in the resistance of plants. In S. K. Datta and S. Muthukrishnan, editors, *Pathogenesis-Related Proteins in Plants*, pages 207–234. CRC Press, Boca Raton, 1999.

[178] B. Stec. Plant thionins - the structural perspective. *Cellular and Molecular Life Sciences (CMLS)*, 63(12):1370–1385, June 2006. doi: 10.1007/s00018-005-5574-5.

[179] C. Hernandez-Lucas, R. F. De Caleya, and P. Carbonero. Inhibition of brewer's yeasts by wheat purothionins. *Appl. Environ. Microbiol.*, 28(2):165–168, August 1974.

[180] H. Bohlmann and K. Apel. Thionins. *Annual Review of Plant Physiology and Plant Molecular Biology*, 42(1):227–240, 1991. doi: 10.1146/annurev.pp.42.060191.001303.

[181] A. O. Carvalho and V. M. Gomes. Role of plant lipid transfer proteins in plant cell physiology – a concise review. *Peptides*, 28(5):1144–1153, May 2007. ISSN 01969781. doi: 10.1016/j.peptides.2007.03.004.

[182] T. H. Yeats and J. K. C. Rose. The biochemistry and biology of extracellular plant lipid-transfer proteins (ltps). *Protein Sci*, 17(2):191–198, February 2008. doi: 10.1110/ps.073300108.

[183] M. Legrand, S. Kauffmann, P. Geoffroy, and B. Fritig. Biological function of pathogenesis-related proteins: Four tobacco pathogenesis-related proteins are chitinases. *Proceedings of the National Academy of Sciences of the United States of America*, 84(19):6750–6754, October 1987.

[184] E. Kombrink, M. Schröder, and K. Hahlbrock. Several "pathogenesis-related" proteins in potato are 1,3-beta-glucanases and chitinases. *Proceedings of the National Academy of Sciences of the United States of America*, 85(3):782–786, February 1988.

[185] L. C. Van Loon, Y. A. M. Gerritsen, and C. E. Ritter. Identification, purification, and characterization of pathogenesis-related proteins from virus-infected samsun nn tobacco leaves. *Plant Molecular Biology*, 9(6):593–609, November 1987. doi: 10.1007/BF00020536.

[186] J. M. Neuhaus. Plant chitinases (pr-3, pr-4, pr-8, pr-11). In S. K. Datta and S. Muthukrishnan, editors, *Pathogenesis-Related Proteins in Plants*. CRC Press, Boca Raton, 1999.

[187] A. Kasprzewska. Plant chitinases–regulation and function. *Cellular & molecular biology letters*, 8 (3):809–824, 2003. ISSN 1425-8153.

[188] J. Asensio, F. Canada, H. Siebert, J. Laynez, A. Poveda, P. Nieto, U. Soedjanaamadja, H. Gabius, and J. Jimenez-Barbero. Structural basis for chitin recognition by defense proteins: Glcnac residues are bound in a multivalent fashion by extended binding sites in hevein domains. *Chemistry & Biology*, 7(7):529–543, July 2000. ISSN 10745521. doi: 10.1016/S1074-5521(00)00136-8.

[189] K. P. B. Van den Bergh, P. Rougé, P. Proost, J. Coosemans, T. Krouglova, Y. Engelborghs, W. J. Peumans, and E. J. M. Van Damme. Synergistic antifungal activity of two chitin-binding proteins from spindle tree (*Euonymus europaeus* l.). *Planta*, 219(2):221–232, June 2004. ISSN 0032-0935. doi: 10.1007/s00425-004-1238-1.

[190] F. Brunner, A. Stintzi, B. Fritig, and M. Legrand. Substrate specificities of tobacco chitinases. *The Plant journal : for cell and molecular biology*, 14(2):225–234, April 1998. ISSN 0960-7412.

[191] T. Theis and U. Stahl. Antifungal proteins: targets, mechanisms and prospective applications. *Cellular and Molecular Life Sciences (CMLS)*, 61(4):437–455, February 2004. doi: 10.1007/s00018-003-3231-4.

[192] D. B. Collinge, K. M. Kragh, J. D. Mikkelsen, K. K. Nielsen, U. Rasmussen, and K. Vad. Plant chitinases. *The Plant journal : for cell and molecular biology*, 3(1):31–40, January 1993. ISSN 0960-7412.

[193] G. W. Gooday. Aggressive and defensive roles for chitinases. *EXS*, 87:157–169, 1999. ISSN 1023-294X.

[194] F. Mauch, B. Mauch-Mani, and T. Boller. Antifungal hydrolases in pea tissue : Ii. inhibition of fungal growth by combinations of chitinase and beta-1,3-glucanase. *Plant physiology*, 88(3): 936–942, November 1988. ISSN 0032-0889.

[195] Z. Minic. Physiological roles of plant glycoside hydrolases. *Planta*, 227(4):723–740, March 2008. doi: 10.1007/s00425-007-0668-y.

[196] E. Jongedijk, H. Tigelaar, J. van Roekel, S. Bres-Vloemans, I. Dekker, P. van den Elzen, B. Cornelissen, and L. Melchers. Synergistic activity of chitinases and beta-1,3-glucanases enhances fungal resistance in transgenic tomato plants. *Euphytica*, 85(1):173–180, February 1995. doi: 10.1007/BF00023946.

[197] C. Sasaki, K. M. Vårum, Y. Itoh, M. Tamoi, and T. Fukamizo. Rice chitinases: sugar recognition specificities of the individual subsites. *Glycobiology*, 16(12):1242–1250, December 2006. ISSN 0959-6658.

[198] N. V. Raikhel, H. I. Lee, and W. F. Broekaert. Structure and function of chitin-binding proteins. *Annual Review of Plant Physiology and Plant Molecular Biology*, 44(1):591–615, 1993. doi: 10.1146/annurev.pp.44.060193.003111.

[199] E. J. M. Van Damme, N. Lannoo, and W. J. Peumans. Plant lectins. In J. C. Kader and M. Delseny, editors, *Advances in Botanical Research*, volume 48, chapter 3, pages 107–209. Academic Press, 2008. doi: 10.1016/S0065-2296(08)00403-5.

[200] W. J. Peumans and E. J. M. Van Damme. Lectins as plant defense proteins. *Plant physiology*, 109(2):347–352, October 1995. ISSN 0032-0889.

[201] M. J. Chrispeels and N. V. Raikhel. Lectins, lectin genes, and their role in plant defense. *The Plant cell*, 3(1):1–9, January 1991. ISSN 1040-4651. doi: 10.1105/tpc.3.1.1.

[202] C. M. Tang, M. L. Chye, S. Ramalingam, S. W. Ouyang, K. J. Zhao, W. Ubhayasekera, and S. L. Mowbray. Functional analyses of the chitin-binding domains and the catalytic domain of brassica juncea chitinase bjchi1. *Plant Molecular Biology*, 56(2):285–298, 2004. doi: 10.1007/s11103-004-3382-1.

[203] W. F. Broekaert, W. Marien, F. R. G. Terras, M. F. C. De Bolle, P. Proost, J. Van Damme, L. Dillen, M. Claeys, and S. B. Rees. Antimicrobial peptides from *Amaranthus caudatus* seeds with sequence homology to the cysteine/glycine-rich domain of chitin-binding proteins. *Biochemistry*, 31(17):4308–4314, May 1992. doi: 10.1021/bi00132a023.

[204] I. Broekaert, H. I. Lee, A. Kush, N. H. Chua, and N. Raikhel. Wound-induced accumulation of mrna containing a hevein sequence in laticifers of rubber tree (hevea brasiliensis). *Proceedings of the National Academy of Sciences of the United States of America*, 87(19):7633–7637, October 1990. ISSN 0027-8424.

[205] U. M. Soedjanaatmadja, T. Subroto, and J. J. Beintema. Processed products of the hevein precursor in the latex of the rubber tree (hevea brasiliensis). *FEBS letters*, 363(3):211–213, April 1995. ISSN 0014-5793.

[206] J. Van Parijs, W. F. Broekaert, I. J. Goldstein, and W. J. Peumans. Hevein: an antifungal protein from rubber-tree (*Hevea brasiliensis*) latex. *Planta*, 183(2):258–264, January 1991. doi: 10.1007/BF00197797.

[207] T. Taira, T. Ohnuma, T. Yamagami, Y. Aso, M. Ishiguro, and M. Ishihara. Antifungal activity of rye (secale cereale) seed chitinases: the different binding manner of class i and class ii chitinases to the fungal cell walls. *Bioscience, biotechnology, and biochemistry*, 66(5):970–977, May 2002. ISSN 0916-8451.

[208] H. Shinshi, J. M. Neuhas, J. Ryals, and F. Meins. Structure of a tobacco endochitinase gene: evidence that different chitinase genes can arise by transposition of sequences encoding a cysteine-rich domain. *Plant molecular biology*, 14(3):357–368, March 1990. ISSN 0167-4412.

[209] B. Henrissat. Structural and sequence-based classification of glycoside hydrolases. *Current Opinion in Structural Biology*, 7(5):637–644, October 1997. ISSN 0959440X. doi: 10.1016/S0959-440X(97)80072-3.

[210] B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat. The carbohydrate-active enzymes database (cazy): an expert resource for glycogenomics. *Nucl. Acids Res.*, 37(suppl_1):D233–238, January 2009. doi: 10.1093/nar/gkn663.

[211] G. Davies and B. Henrissat. Structures and mechanisms of glycosyl hydrolases. *Structure*, 3(9):853–859, September 1995. ISSN 09692126. doi: 10.1016/S0969-2126(01)00220-9.

[212] A. C. Terwisscha_van_scheltinga, K. H. Kalk, J. J. Beintema, and B. W. Dijkstra. Crystal structures of hevamine, a plant defence protein with chitinase and lysozyme activity, and its complex with an inhibitor. *Structure*, 2(12):1181–1189, December 1994. ISSN 09692126. doi: 10.1016/S0969-2126(94)00120-0.

[213] E. Bokma, M. Spiering, K. S. Chow, Mulder, T. Subroto, and J. J. Beintema. Determination of cdna and genomic dna sequences of hevamine, a chitinase from the rubber tree hevea brasiliensis. *Plant Physiology and Biochemistry*, 39(5):367–376, May 2001. ISSN 09819428. doi: 10.1016/S0981-9428(01)01247-5.

[214] L. Duo-Chuan. Review of fungal chitinases. *Mycopathologia*, 161(6):345–360, June 2006. ISSN 0301-486X. doi: 10.1007/s11046-006-0024-y.

[215] L. S. Melchers, M. Apotheker-de Groot, J. A. van der Knaap, A. S. Ponstein, M. B. Sela-Buurlage, J. F. Bol, B. J. Cornelissen, P. J. van den Elzen, and H. J. Linthorst. A new class of tobacco chitinases homologous to bacterial exo-chitinases displays antifungal activity. *The Plant journal : for cell and molecular biology*, 5(4):469–480, April 1994. ISSN 0960-7412.

[216] J. Beintema. Structural features of plant chitinases and chitin-binding proteins. *FEBS Letters*, 350(2-3):159–163, August 1994. ISSN 00145793. doi: 10.1016/0014-5793(94)00753-5.

[217] L. Berglund, J. Brunstedt, K. K. Nielsen, Z. Chen, J. D. Mikkelsen, and K. A. Marcker. A proline-rich chitinase from beta vulgaris. *Plant Molecular Biology*, 27(1):211–216, January 1995. ISSN 0167-4412. doi: 10.1007/BF00019193.

[218] T. Nakazaki, T. Tsukiyama, Y. Okumoto, D. Kageyama, K. Naito, K. Inouye, and T. Tanisaka. Distribution, structure, organ-specific expression, and phylogenic analysis of the pathogenesis-related protein-3 chitinase gene family in rice (*Oryza sativa* l.). *Genome / National Research Council Canada = Génome / Conseil national de recherches Canada*, 49(6):619–630, June 2006. ISSN 0831-2796. doi: 10.1139/g06-020.

[219] D. Zhang, M. Hrmova, C. H. Wan, C. Wu, J. Balzen, W. Cai, J. Wang, L. D. Densmore, G. B. Fincher, H. Zhang, and C. H. Haigler. Members of a new group of chitinase-like genes are expressed preferentially in cotton cells with secondary walls. *Plant Molecular Biology*, 54(3): 353–372, February 2004. ISSN 0167-4412. doi: 10.1023/B:PLAN.0000036369.55253.dd.

[220] K. Lohtander, H. L. Pasonen, M. K. Aalto, T. Palva, A. Pappinen, and J. Rikkinen. Phylogeny of chitinases and its implications for estimating horizontal gene transfer from chitinase-transgenic silver birch (*Betula pendula*). *Environmental Biosafety Research*, 7(4):227–239, 2008. ISSN 1635-7922. doi: 10.1051/ebr:2008019.

[221] N. Wasano, K. Konno, M. Nakamura, C. Hirayama, M. Hattori, and K. Tateishi. A unique latex protein, mlx56, defends mulberry trees from insects. *Phytochemistry*, 70(7):880–888, May 2009. ISSN 1873-3700. doi: 10.1016/j.phytochem.2009.04.014.

[222] A. S. Ponstein, S. A. Bres-Vloemans, M. B. Sela-Buurlage, Pjm van den Elzen, L. S. Melchers, and Bjc Cornelissen. A novel pathogen- and wound-inducible tobacco (nicotiana tabacum) protein with antifungal activity. *Plant Physiol.*, 104(1):109–118, January 1994. doi: 10.1104/pp.104.1.109.

[223] C. Caporale, I. Berardino, Leonardi Di, L. Bertini, A. Cascone, V. Buonocore, and C. Caruso. Wheat pathogenesis-related proteins of class 4 have ribonuclease activity. *FEBS Letters*, 575(1-3): 71–76, September 2004. ISSN 00145793. doi: 10.1016/j.febslet.2004.07.091.

[224] B. Svensson, I. Svendsen, P. Højrup, P. Roepstorff, S. Ludvigsen, and F. M. Poulsen. Primary structure of barwin: a barley seed protein closely related to the c-terminal domain of proteins encoded by wound-induced plant genes. *Biochemistry*, 31(37):8767–8770, September 1992. ISSN 0006-2960.

[225] S. Ludvigsen and F. M. Poulsen. Secondary structure in solution of barwin from barley seed using 1h nuclear magnetic resonance spectroscopy. *Biochemistry*, 31(37):8771–8782, September 1992. ISSN 0006-2960.

[226] S. Ludvigsen and F. M. Poulsen. Three-dimensional structure in solution of barwin, a protein from barley seed. *Biochemistry*, 31(37):8783–8789, September 1992. ISSN 0006-2960.

[227] A. Kiba, H. Saitoh, M. Nishihara, K. Omiya, and S. Yamamura. C-terminal domain of a hevein-like protein from *Wasabia japonica* has potent antimicrobial activity. *Plant Cell Physiol.*, 44(3): 296–303, March 2003. doi: 10.1093/pcp/pcg035.

[228] B. Neron, P. Tuffery, and C. Letondal. Mobyle: a web portal framework for bioinformatics analyses. In *NETTAB*, 2005.

[229] H. T. Wright, D. M. Brooks, and C. S. Wright. Evolution of the multidomain protein wheat germ agglutinin. *Journal of Molecular Evolution*, 21(2):133–138, February 1985. doi: 10.1007/BF02100087.

[230] E. J. M. Van Damme, D. Charels, S. Roy, K. Tierens, A. Barre, J. C. Martins, P. Rouge, F. Van Leuven, M. Does, and W. J. Peumans. A gene encoding a hevein-like protein from elderberry fruits is homologous to pr-4 and class v chitinase genes. *Plant Physiol.*, 119(4):1547–1556, April 1999. doi: 10.1104/pp.119.4.1547.

[231] Carbohydrate active enzymes database. URL `http://www.cazy.org/fam/gh18.html`.

[232] B. Palenik, J. Grimwood, A. Aerts, P. Rouze, A. Salamov, N. Putnam, C. Dupont, R. Jorgensen, E. Derelle, S. Rombauts, K. Zhou, R. Otillar, S. S. Merchant, S. Podell, T. Gaasterland, C. Napoli, K. Gendler, A. Manuell, V. Tai, O. Vallon, G. Piganeau, S. Jancek, M. Heijde, K. Jabbari, C. Bowler, M. Lohr, S. Robbens, G. Werner, I. Dubchak, G. J. Pazour, Q. Ren, I. Paulsen, C. Delwiche, J. Schmutz, D. Rokhsar, Y. Van de Peer, H. Moreau, and I. V. Grigoriev. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences*, 104(18):7705–7710, May 2007. doi: 10.1073/pnas.0611046104.

[233] P. A. Jekel, B. H. Hartmann, and J. J. Beintema. The primary structure of hevamine, an enzyme with lysozyme/chitinase activity from hevea brasiliensis latex. *European journal of biochemistry / FEBS*, 200(1):123–130, August 1991. ISSN 0014-2956.

[234] R. Cohen-Kupiec and I. Chet. The molecular biology of chitin digestion. *Current opinion in biotechnology*, 9(3):270–277, June 1998. ISSN 0958-1669.

[235] J. P. Metraux, W. Burkhart, M. Moyer, S. Dincher, W. Middlesteadt, S. Williams, G. Payne, M. Carnes, and J. Ryals. Isolation of a complementary dna encoding a chitinase with structural homology to a bifunctional lysozyme/chitinase. *Proceedings of the National Academy of Sciences of the United States of America*, 86(3):896–900, February 1989. ISSN 0027-8424.

[236] K. Lawton, E. Ward, G. Payne, M. Moyer, and J. Ryals. Acidic and basic class iii chitinase mrna accumulation in response to tmv infection of tobacco. *Plant molecular biology*, 19(5):735–743, August 1992. ISSN 0167-4412.

[237] A. Durand, R. Hughes, A. Roussel, R. Flatman, B. Henrissat, and N. Juge. Emergence of a subfamily of xylanase inhibitors within glycoside hydrolase family 18. *The FEBS journal*, 272(7):1745–1755, April 2005. ISSN 1742-464X. doi: 10.1111/j.1742-4658.2005.04606.x.

[238] K. Takahashi, S. Hirata, N. Kido, and K. Katou. Wall-yielding properties of cell walls from elongating cucumber hypocotyls in relation to the action of expansin. *Plant Cell Physiol.*, 47(11):1520–1529, November 2006. ISSN 0032-0781. doi: 10.1093/pcp/pcl017.

[239] M. Lee, G. Hwang, Y. Chen, H. Lin, and C. Wu. Molecular cloning of indian jujube (*Zizyphus mauritiana*) allergen ziz m 1 with sequence similarity to plant class iii chitinases. *Molecular Immunology*, 43(8):1144–1151, March 2006. ISSN 01615890. doi: 10.1016/j.molimm.2005.07.021.

[240] T. Heitz, S. Segond, S. Kauffmann, P. Geoffroy, V. Prasad, F. Brunner, B. Fritig, and M. Legrand. Molecular characterization of a novel tobacco pathogenesis-related (pr) protein: a new plant chitinase/lysozyme. *Molecular and General Genetics MGG*, 245(2):246–254, March 1994. doi: 10.1007/BF00283273.

[241] Y. H. Xiao, L. Hou, X. H. Yuan, X. Y. Yang, Y. Pei, X. Y. Luo, and Y. Pei. [cloning and characterization of a homologous gene of plant class v chitinase from balsampear, *Momordica charantia* linn.]. *Yi chuan xue bao = Acta genetica Sinica*, 29(11):1028–1033, 2002. ISSN 0379-4172.

[242] P. Salzer, N. Feddermann, A. Wiemken, T. Boller, and C. Staehelin. *Sinorhizobium meliloti*-induced chitinase gene expression in *Medicago truncatula* ecotype r108-1: a comparison between symbiosis-specific classv and defence-related classiv chitinases. *Planta*, 219(4):626–638, August 2004. doi: 10.1007/s00425-004-1268-8.

[243] Y. Yamauchi, M. Nishimura, H. Hayshi, T. Taira, M. Ishihara, and C. Fukamizo. Characterization of the reaction catalyzed by class v chitinase from cycas revoluta. *Chitin and Chitosan Research*, 12(2):203+, 2006.

[244] E. J. M. Van Damme, R. Culerrier, A. Barre, R. Alvarez, P. Rougé, and W. J. Peumans. A novel family of lectins evolutionarily related to class v chitinases: an example of neofunctionalization in legumes. *Plant physiology*, 144(2):662–672, June 2007. ISSN 0032-0889. doi: 10.1104/pp.106.087981.

[245] M. Wiweger, I. Farbos, M. Ingouff, U. Lagercrantz, and S. von Arnold. Expression of chia4-pa chitinase genes during somatic and zygotic embryo development in norway spruce (*Picea abies*): similarities and differences between gymnosperm and angiosperm class iv chitinases. *J. Exp. Bot.*, 54(393):2691–2699, December 2003. doi: 10.1093/jxb/erg299.

[246] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. Interpro: the integrative protein signature database. *Nucleic acids research*, 37(Database issue), January 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn785.

[247] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. The pfam protein families database. *Nucl. Acids Res.*, 36(suppl_1):D281–288, January 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm960.

[248] P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129–2141, September 2003. ISSN 1088-9051. doi: 10.1101/gr.772403.

Die hier vorgelegte Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 1. Juni 2010

(Nicole de Miranda Scherer)