8-2016

Understanding Plant Response to Stress Using Gene Model Quality Evaluation and Transcriptome Analysis

Karthik Ramaswamy Padmanabhan
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Karthik Ramaswamy Padmanabhan

Entitled

Understanding Plant Response to Stress Using Gene Model Quality Evaluation and Transcriptome Analysis

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Michael Gribskov
Chair

Daisuke Kihara

Carol Beth Post

Laurie L. Parker

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Michael Gribskov

Approved by: Daoguo Zhou 7/10/2016

Head of the Departmental Graduate Program Date

UNDERSTANDING PLANT RESPONSE TO STRESS USING GENE MODEL QUALITY EVALUATION AND TRANSCRIPTOME ANALYSIS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Karthik Ramaswamy Padmanabhan

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2016

Purdue University

West Lafayette, Indiana

This dissertation is dedicated to my loving and ever-supportive parents.

ACKNOWLEDGMENTS

The work presented in this dissertation was possible thanks to the support and guidance of many people. Firstly, I would like to express my deepest gratitude to my advisor, Prof. Michael Gribskov. I am thankful for his patience, encouragement and participation in various scientific discussions over the course of my graduate work. He gave me the freedom to think independently and the opportunity to grow as a scientist. I would also like to thank the members of my dissertation committee - Dr. Daisuke Kihara, Dr. Carol Post and Dr. Laurie Parker for their thoughtful advice, supervision and guidance.

I would like to extend my thanks to our collaborators Dr. Stephen Weller, Dr. Burkhard Schulz and Kabelo Segobye for providing us with the data required for our analysis, and for their expertise and help during the project.

I would also like to thank the former and current members of the Gribskov lab. I am truly fortunate to have been in the company of people with brilliant scientific minds and I am sure they are destined to achieve great success in their respective fields. I would like to specifically thank Jim and Reazur for helping me settle in when I first joined the lab.

I would like to acknowledge Dr. Nina Robinson for providing me with the opportunity to work with her as a graduate assistant. Thanks are also due to Dr. Barrett Foat and Dr. Dmitry Grapov, scientists at Monsanto, for being my mentors during my summer internship there.

I would like to thank my parents - Mr. Padmanabhan and Mrs. Hemalatha, for their love and encouragement. They always believed in me, and I am blessed to have such wonderful people in my life. I would also like to acknowledge my relatives and family friends who have played important roles at various stages of my life. I would also like to dedicate this dissertation to my aunt Mrs. Parvathy who I consider to be my second mother.

Finally, I would like to thank my friends, both at Purdue and elsewhere for sharing my joy and frustrations. I would especially like to thank my best friend, Prahatha, for being my support system and for always pushing me to go the extra mile. Thank you for inspiring me to be the person I am now. To Phoebe, you are the best dog in the world, and don't let anyone tell you otherwise.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

ABBREVIATIONS

ATP	adenosine triphosphate
BLAST	basic local alignment search tool
DNA	deoxyribonucleic acid
EPK	eukaryotic protein kinase
EPSPS	enolypyruvylshikimate-3-phosphate synthase
HMM	hidden markov model
RNA	ribonucleic acid
mRNA	messenger RNA

ABSTRACT

Padmanabhan, Karthik Ramaswamy PhD, Purdue University, August 2016. Understanding Plant Response to Stress Using Gene Model Quality Evaluation and Transcriptome Analysis. Major Professor: Michael R. Gribskov.

The overall aim of the project was to understand how plants reacted to environmental stress and evolved to overcome it. The land plants that we see today evolved from a green algal ancestor around 510 million years ago. Plants had to make significant changes to their cellular, morphological, regulatory and physiological processes during their adaptation to the terrestrial environment from an aquatic environment. The first part of the project was to find out how these changes were reflected on the protein makeup of the early land plants. The gene model sequence data of two early land plants, *Physcomitrella patens* (moss) and *Chlamydomonas reinhardtii* (green algae). We specifically focused on the protein family expansion of protein kinases due to their roles in various important functions that would affect the transition from water to land. We developed a gene model quality evaluation method to score the gene models of *P. patens* and *C. reinhardtii* using well-studied plants such as *Arabidopsis thaliana* and *Oryza sativa* (rice) to improve the poor quality gene models that currently exist. The resulting corrected gene models were analyzed using functional annotation methods to understand how the proteomics of the early land plants varied from modern land plants.

The second part of the project was to identify the genes responsible for herbicide resistance in *Ambrosia trifida* (giant ragweed). Giant ragweed is one of the most competitive annual weeds in corn and soybean production across the eastern Corn Belt in the United States. The use of glyphosate (commercial name: Roundup)

and glyphosate-ready crop systems managed to keep giant ragweed populations under control. Glyphosate-ready crop systems consist of seeds that are resistant to glyphosate, which enables farmers to use glyphosate to control the population of weeds. But in the last decade, glyphosate-resistant giant ragweed populations have been reported across the world. It is a huge problem to farmers since it results in unusable glyphosate-ready cropping systems and huge yield losses. Glyphosate-resistant and sensitive plants were identified from across the Midwestern United States and a RNA-seq experiment was performed by isolating the total mRNA from leaf material, and obtaining the expressed messenger RNA sequences. The genetic makeup of the sensitive and resistant strains was thus compared based on their transcriptome data, and a list of potential genes that were differentially expressed between them was identified. We also analyzed how much the quality of the transcriptome can be improved by using the transcriptome and genome of sunflower, a closely-related plant.

1. INTRODUCTION

1.1 Background and significance

Environmental stress in plants can be defined as conditions that negatively affect plant growth or development (Buchanan et al., 2015). Stress can affect the gene expression, rate of cellular metabolism and rate of development in plants (Reddy et al., 2004). Changes in plant cellular and metabolic states occur via a process called acclimation, which involves modifications to multiple metabolic pathways (Mittler, 2006). Due to the increasing awareness about climate change, and widespread increase in drought occurrences across the globe, understanding how plants respond to stress is increasingly crucial, especially in the field of agriculture (Petit et al., 1999; Chaves et al., 2003). Stress response also plays a major role in weeds developing resistance to herbicides (Powles and Yu, 2010).

Thanks to the rapid advances in genomics and sequence analysis, many genes that take part in stress response have been identified. Abscisic acid (ABA), a plant hormone, has been recognized as playing an important role in the response to stress conditions such as salinity, physical damage, and water scarcity, by plants (Tuteja, 2007). Plants are known to regulate the levels of ABA to counter environmental stress (Tuteja, 2007).

Plant protein kinases play a major role in plant stress response. A study by Saijo et al. examined the contribution of Calcium-Dependent Protein Kinases (CPKs) to the plant stress responses (Saijo et al., 2000). In the study, the authors found that a single CPK in rice was responsible for conferring tolerance to low temperature, salinity and drought, thus suggesting that this CPK could be a commonly used regulator

in different stress response pathways. Mitogen Activated Protein Kinases (MAPKs) also help in pathogen response in tobacco plants (Yang et al., 2001). The study concluded that a MAPK kinase kinase (MAPKK), NtMEK2, regulates the activation of the hypersensitive response in tobacco, which is a key mechanism involved in plant disease resistance. A protein cascade involving NtMEK2 was also discovered to control the expression of *HMGR* and *PAL*, two key defense genes which express enzymes playing a major role in phytoalexin and salicyclic acid production (Yang et al., 2001). Another study found that several genes that code for protein kinases are upregulated in response to stress conditions such as salinity, cold, and drought (Liu et al., 2000b). In addition, several protein kinases in the MAPK cascade have been identified to play a major role in plant signaling under abiotic stress conditions such as cold stress, salt stress, dehydration, wounding, ozone, and heavy metal stress (Sinha et al., 2011).

Nitric Oxide (NO) is known to play an important role in plant biotic and abiotic stress response through a process called NO burst, a term used to describe rapid NO production (Asai and Yoshioka, 2008). NO is understood to induce the activation of a MAPK during the process of programmed cell death (Clarke et al., 2000).

1.2 Abiotic stress during early land plant evolution

The emergence of land plants, otherwise known as embryophytes, around 400-500 million years ago, and their early diversification and development, delineates a very important phase in the growth of terrestrial life on earth (Karol et al., 2001). Early land plants such as liverworts, hornworts and mosses evolved from charophytes, which are comprised almost entirely of green algae (Rensing et al., 2008). These three early land plant groups are collectively termed bryophytes, or non-vascular plants (Figure 1.1). Plants with a developed vascular system are called tracheophytes. The morphologies and life cycles of charophytes differ from both bryophytes and tracheophytes. While tracheophytes and bryophytes alternate between the diploid sporo-

phyte and the haploid gametophyte generations, the gametophytic haploid generation is generally the only generation of the life cycle for the charophytes (Graham and Wilcox, 2000). However, certain charaophytes such as *Spirogyra* and *Coleochaete scutata* have been shown to have unusual life cycles that include a diploid generation (Haig, 2010).

Plants had to make significant changes to their cellular, morphological, regulatory, and physiological processes, during their adaptation to the terrestrial environment (Rensing et al., 2008). During the evolution of bryophytes from charophytes, the dynein-based transport system present in algae was replaced with a kinesin-based transport system. Early land plants developed signaling systems that used auxin, ABA and cytokinins. The complexity of systems such as the ATP-binding cassette (ABC) transporter family and photoreceptor signaling increased, while an increase in tolerance to desiccation and stress was mandated by the move to dryer environments. Mosses, in particular, elaborated a complex two-component signaling system. Two-component signaling systems, at a basic level, involve a histidine kinase and a corresponding response regulator. The histidine kinase receives a signal which results in the autophosphorylation of a conserved histidine residue. The resulting phosphate is then relayed to the response regulator protein (Stock et al., 2000). Two-component systems are found in both prokaryotes and eukaryotes, although more complex two-component systems have been discovered in plants (Lohrmann and Harter, 2002). These complex systems can include more than two response regulator proteins for signal relay and a hybrid kinase instead of a histidine kinase. Hybrid kinases are known to contain more than one phosphodonor and phosphoacceptor sites in order to use multiple step phosphoryl transfer system. Mosses also elaborateded more efficient homologous recombination DNA repair systems, adaptations to growth in shade, and dehydration-rehydration adaptations (Rensing et al., 2008). When tracheophytes evolved from bryophytes, they lost motile gametes and vegetative desiccation toler-

ance, but gained the ability to signal via gibberellin, jasmonate, ethylene,and brassinosteroids.

1.2.1 Eukaryotic protein kinases

Protein kinases are enzymes that catalyze the phosphorylation of serine, threonine, or tyrosine residues in proteins by transferring the γ-phosphate of ATP to the substrate residue (Lehti-Shiu and Shiu, 2012). Eukaryotic protein kinases (EPKs) are a large family of highly regulated and conserved proteins involved in many cellular processes (Hanks and Hunter, 1995). A characteristic feature of EPKs is the high degree of sequence and structural similarity across different species and families. EPKs possess a highly conserved protein kinase catalytic domain of roughly 250-300 amino acid residues (Stone and Walker, 1995).

EPKs can be broadly classified into eight different groups : AGC (Protein Kinase A, G and C families), CAMK (Calmodulin/Calcium-Mediated Kinases), CMGC (CDK, MAPK, GSK3 and CLK families), RGC (Receptor Guanylate Cyclases), TK (Tyrosine Kinases), TKL (Tyrosine Kinase-Like), STE (homologs of yeast Sterile kinases) and CK1 (Casein Kinase 1 group) protein kinases (Manning et al., 2002; Hanks and Hunter, 1995). The classification is based mainly on sequence similarity and the presence of certain conserved domains in each group.

Previous research has shown that the conserved regions in the protein kinase catalytic domain can be classified into 12 subdomains; each region has been studied to identify the reasons for its conservation (Hanks and Quinn, 1991). The subdomains consist of several sites that have consensus residues (Figure 1.2). Subdomain I has a conserved GxGxxG motif which acts as the ATP-binding loop. The primary function of this motif is to orient the γ-phosphate of ATP for the transfer of the phosphoryl group (Johnson et al., 2001). Subdomain II is centered on a conserved lysine residue

that has been found to be required for maximum enzyme activity. It anchors the α and β phosphates of ATP during catalysis (Johnson et al., 2001). An invariant glutamate residue in subdomain III acts as a salt bridge between the conserved lysine in subdomain II, and the protein kinase active site [25]. Subdomain IV contains a pair of conserved phenylalanine residues which act as anchors for the ATP-binding pocket (Johnson et al., 2001). Subdomain V is a part of both the N and C lobes; it forms a hydrophobic beta strand in the N-lobe and takes the form of an alpha helix in the C-lobe. The N lobe of the protein kinase consists of a β-sheet containing five strands, and the C-lobe comprises of α-helices and loops (McClendon et al., 2014). Residues in subdomain V help to stabilize the ATP-binding pocket, and the binding of peptide with the substrate (Hanks and Hunter, 1995).

Subdomain VIb contains the consensus sequence HRDLKxxN, which is the most important conserved sequence in the protein. It is termed the catalytic loop because the aspartate residue interacts with the three ATP-phosphates, either through magnesium atoms, or through direct contact (Kornev et al., 2006). In Protein Kinase A (PKA), and some other protein kinases, the loop contains a tyrosine residue (Y) in place of the histidine (H). Certain non-protein kinases such as the Phosphatidylinositol phosphate kinases (PIPK) contain a MDYSL motif instead (Schramp et al., 2012). Subdomain VII comprises the highly conserved DFG triplet, which helps position the magnesium ion, and orient the -phosphate of the ATP for transfer. Another conserved triplet is the APE triplet which is found in subdomain VIII. The glutamate residue in this motif forms a salt bridge with the invariant arginine in subdomain XI. This stabilizes the kinase core, and acts as an anchor for the movement of the activation loop (Hanks and Hunter, 1995). In most protein kinases, this subdomain also contains a phosphorylatable amino acid residue about seven to ten residues upstream of the APE triplet motif, which creates an ionic bond with the arginine in the HRDLKxxN motif in subdomain VIb. The region between this residue and the APE motif is termed the P+1 loop [21]. Subdomain IX has a conserved DxWxxG

motif and is involved in hydrophobic interactions that stabilize the structure of the protein kinase. Subdomain XI contains an invariant arginine residue,. This conserved arginine residue improves the stability of the large carboxyl-terminal lobe.

1.2.2 Eukaryotic protein kinases in land plant evolution

As mentioned earlier, early land plants underwent major changes to different fundamental systems during their adaptation to life on land from water. This involved changes in response to both abiotic and biotic conditions. Alteration in water and salt concentrations are the most important factors that affected these early plants during the transition. Salinity and drought stress have major effects on the metabolism and physiology of plants. Many of these changes must have involved protein kinases. The SOS pathway has been recently identified in plants as one of the primary regulatory pathways that is triggered during saline and drought stress (Zhu, 2000). It involves three proteins SOS1, SOS2 and SOS3, of which SOS2 is a serine/threonine protein kinase (McDonald and Linde, 2002). Plant defense responses are mainly mediated by protein kinase families such as calcium-dependent protein kinases, and MAP kinases (Romeis, 2001). Protein kinases have been found to be required for salt tolerance (Liu et al., 2000a), ABA signaling (Hirayama and Shinozaki, 2007), carbon metabolism (Halford and Hardie, 1998), apoptosis (Bialik and Kimchi, 2006), dehydration tolerance (Yoshida et al., 2002), osmotic stress response (Mikołajczyk et al., 2000), regulation of reactive oxygen species (ROS) production (Kobayashi et al., 2007), auxin signaling (Lee et al., 2009), jasmonate signaling (Takahashi et al., 2007), and ethylene signaling (Guo and Ecker, 2004).

MAP kinases, and the MAPK cascade of proteins, are one of the most important regulators of stress in plants, due to their role in signaling (Cristina et al., 2010). It is known that H_2O_2 is a significant molecule used in signaling stress responses, wounding and pathogen resistance (Kovtun et al., 2000). A majority of eukaryotes

use protein phosphorylation mediated by MAPK cascades as responders to oxidative signals. A typical cascade consists of MAP kinase kinase kinases (MAP3Ks), MAP kinase kinases (MAP2Ks) and MAP kinases (MAPKs) (Nakagami et al., 2005). Research has shown that MAPKs play an important role in plant pathogen response in *Arabidopsis*, tobacco, rice and parsley (Nakagami et al., 2005).

Plant receptor-like kinases (RLKs) are a major class of protein kinases, and are similar to animal receptor tyrosine kinases. They typically span the cell membrane, and contain receptor domains esposed on the extracellular side of the cell membrane that receive signals, and a cytoplasmic protein kinase domain that is activated when ligands bind to the extracellular receptor (Becraft, 1998). They are known to take part in a variety of processes such as resistance to disease, regulation of cell growth, symbiosis and brassinosteroid signaling (De Smet et al., 2009). Only about 2% of the total number of RLKs identified so far have been assigned functions (Shiu and Bleecker, 2001). RLKs have been implicated in both normal growth and development of the plant, and in plant stress responses. Various genes in rice, wheat, tomato and *Arabidopsis* have been associated with disease resistance, defense response, and microbial stress response functions (Shiu and Bleecker, 2001). RLK signaling pathways are known to activate defense response genes in various plants (Afzal et al., 2008). For instance, the FLS2 receptor kinase is involved in *Arabidopsis* innate immunity. FLS2 binds bacterial flagellin, which activates downstream signals, causing plant defense response. The expansion of the RLK family of protein kinases in early land plants is therefore assumed to have allowed plants make suitable adjustments to signaling systems during their evolution (Afzal et al., 2008).

Therefore, in order to track the adaptations that bryophytes and land plants had to make, the specific protein kinase families that expanded, contracted or remained constant during the course of evolution from charophytes to tracheophytes were analyzed. To track these changes, functional analysis of bryophytes and charophytes

were done. This helped us understand the various adaptations that the early land plants had to make to their various bio-systems.

When EPKs are compared across bryophytes, charophytes and tracheophytes, the latter have the highest number of protein kinases [20]. *Physcomitrella patens*, a bryophyte model system, has 685 gene models annotated as protein kinases in its genome, while *Chlamydomonas reinhardtii*, the most studied green algae, has only 426 protein kinases. In comparison, *Arabidopsis thaliana* has close to 1000 protein kinases, and *Oryza sativa* has more than 1400 protein kinases. This disparity in the number of protein kinases is mainly due to the extent of protein duplication in each of the species. Lehti-Shiu et al. (2012) compared the percentage sequence identity between the paralogues of proteins in *C. reinhardtii*, *P. patens*, and *A. thaliana* (Lehti-Shiu and Shiu, 2012). The higher the sequence identity, the more recent the duplication event. To explain further, when a species undergoes a whole genome duplication event, the duplicate genes start diverging from each other due to random mutations. Therefore, when the sequence identity between duplicated genes is high, this implies that the duplication event was fairly recent. On the other hand, when the sequence identity is low, it can be said that the duplication event occurred farther in the past. They discovered that while *C. reinhardtii* has an average paralog percentage identity of 56.6, *A. thaliana*, *O. sativa*, and *P. patens* have a much higher percentage identities of 81.1%, 79.0% and 85.3%, respectively. This suggests that these three plants have protein kinases that are more recently duplicated than the proteins in the green algae. Research suggests that the *A. thaliana* lineage underwent at least three whole genome duplication events, with the last one occurring around 25 million years ago (Rizzon et al., 2006; Blanc and Wolfe, 2004). Similarly, *O. sativa* is known to have undergone a whole genome duplication event approximately 21 million years ago, with another duplication event earlier occurring 170-235 million years ago (Yu et al., 2005). In *P. patens*, a whole genome duplication event is assumed to have occurred between 30 and 60 million years ago (Rensing et al., 2007). These studies

generally agree with the results obtained from the percentage sequence identity study that was performed by the group.

Protein family expansion may contribute to the adaptation of the organism to its environment. It has been suggested that the increase in size of the protein kinase family of proteins occurred via lineage-specific expansion (Haig and Wilczek, 2006). Lineage-specific expansion can be termed as the expansion of a particular family of proteins in a specific lineage, when compared to its sister lineage. A study found that up to 80% of proteins in the *Arabidopsis thaliana* genome consisted of lineage-specific expansions of protein families, whose functions were mainly related to pathogen response, stress response, and signaling pathways (Lespinet et al., 2002) This is mainly because it is simpler for the organism to undergo gene duplication to increase their functional diversity than start from scratch (Kondrashov, 2012).

1.2.3 Plant gene modeling

Computational methods for identifying genes in a genome typically fall into two categories genome-guided, and ab initio gene modeling. In genome-guided gene modeling, the genomic sequence of a closely-related organism is used to infer the structure of genes. In plant genomics, only certain model systems such as *A. thaliana* and to a certain extent, *O. sativa* have reasonably well annotated genomes (Kaul et al., 2000; Project, 2005). On the other hand, ab initio gene modeling is used when there is no closely related genome to work with, and uses signal sensors and content sensors to identify genes (Wang et al., 2004). These will be discussed in detail in the next chapter. Ab initio methods are often less accurate, and suffer from various drawbacks and limitations (Li et al., 2005). Therefore, when working with sequences from plants such as *P. patens* and *C. reinhardtii*, which dont have closely related reference genomes, there is a need for a method to evaluate the ab initio gene models.

1.3 Herbicide resistance in plants

Herbicides are chemicals that are used to kill unwanted plants that compete with crop plants for resources. Herbicide application has been the most widely used and effective form of weed control over the past four decades (Powles and Yu, 2010). However, weeds have recently developed resistance to various herbicides due to improper herbicide treatment strategies such as constant treatment with a similar class or family of chemicals (Jasieniuk et al., 1996).

Herbicide-resistant weeds are a growing threat to food crops and agriculture. Due to the immense selective pressure produced by herbicide application, any plant carrying an allele providing resistance to herbicides is strongly selected. There are two types of mechanisms by which weeds can develop resistance to herbicides target-site resistance (TSR) and non-target-site-based resistance (NTSR). TSR occurs when an amino acid change at the target protein occurs thereby preventing the binding of the herbicide or many other effects. It can also be achieved when the target enzyme is overexpressed via gene amplification or duplication. NTSR occurs when the plant prevents the herbicide from reaching the target site through various mechanisms such as reduced herbicide translocation, herbicide degradation, efflux and sequestration (Kemp et al., 1990; Kern and Dyer, 1998; Preston and Wakelin, 2008; Ge et al., 2010).

1.3.1 Glyphosate resistance and weed evolution

N-(phosphonomethyl)glycine, commonly known as glyphosate or RoundUp, is one of the most widely used herbicides in the world (Shaner, 2000). Glyphosate closely resembles the chemical structure of the amino acid glycine, which results in its uptake by plants without causing any stress response. It demonstrates herbicidal activity against a wide variety of weeds (Malik et al., 1989). Glyphosates mode of action involves the inhibition of 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), an

enzyme that is present only in plants, fungi and bacteria (Nandula, 2010). Therefore, it is inherently non-toxic to humans, other animals, and insects. The enzyme plays an important role in the synthesis of aromatic amino acids in the shikimate pathway. The inhibition of EPSPS by glyphosate ultimately results in plant death due to build up of shikimate pathway intermediates (Gomes et al., 2014).

Resistance to herbicides in plants is an evolutionary process due to selection pressure that enables plants to survive a normal dose of herbicide treatment (Bradshaw et al., 1997). The mechanisms behind herbicide resistance can be the EPSPS enzyme target-site modification, degradation of the herbicide, bypassing the toxic activity of the herbicide, or prevention of herbicide-target interaction by utilizing physical barriers such as enhanced cuticles or physiological barriers like active transporters (Sammons and Gaines, 2014). These mechanisms are further explained in detail in the next section. Due to the use of glyphosate-resistant cropping systems, and to general overuse of herbicides, various resistant weeds have been reported (Duke and Powles, 2008).

1.3.2 Mechanisms of glyphosate resistance

In one mechanism, the sequence of the gene encoding EPSPS is altered with amino acid residue 106P typically being replaced by Serine, Alanine or Threonine, which affects the strength of glyphosate binding to the enzyme (Powles and Preston, 2006). Common weeds exhibiting this type of resistance mechanism include Malaysian goosegrass, Italian ryegrass, and Rigid ryegrass (Gomes et al., 2014; Jasieniuk et al., 2008; Preston et al., 2009).

Another major mechanism of glyphosate resistance in various weeds is reduced translocation. Studies done using radioactively-labeled 14C glyphosate have shown that many weeds restrict translocation of glyphosate within the plant (Pratley et al.,

1999). When nuclear magnetic resonance (NMR) was used to track the transport of glyphosate inside cellular compartments, glyphosate was found to be sequestered inside vacuoles in some glyphosate resistant weed biotypes (Ge et al., 2013; Ge et al., 2010). This indicates that these plants are able to recognize and isolate glyphosate before it can cause any adverse effects . However, there has been no evidence, so far, that glyphosate can be catabolized by plants, and therefore, such a mechanism of resistance has not been discovered (Whitaker et al., 2013).

Another mechanism found in glyphosate-resistant weed biotypes is gene duplication. In some weeds, the EPSPS gene is duplicated multiple times to overcome the effect of the herbicide (Boerboom et al., 1990; Jones et al., 1996; Shah et al., 1986; Suh et al., 1993; Widholm et al., 2001). The increase in EPSPS gene expression due to gene duplication leads to increases in protein levels, thus circumventing the effect of the herbicide. However, studies have shown that the multiple copies of the EPSPS gene are not stable, and are not passed to subsequent generations (Sammons and Gaines, 2014; Pline-Srnic, 2006).

1.3.3 Herbicide resistance and plant stress

In giant ragweed, NTSR has been found to be the most common type of mechanism for acquiring resistance (Powles and Yu, 2010). It has been reported that NTSR can be caused by environmental stresses introduced by the application of herbicides (Delye, 2013). In a process called 'gene stacking', weed genotypes that have reduced sensitivity to herbicides are progressively naturally selected after each generation until resistance to herbicide is achieved. Due to the extreme selective pressure, in which up to 99% of sensitive weeds may be eliminated by spraying with herbicides, any genetic change that enables the weed to survive and reproduce is strongly selected. Several gene families involved in NTSR have been shown to play an important role in plant stress response. These include the Cytochrome P450 family, oxidases, perox-

idases, esterases, hydrolases, glutathione-S-transferases, glycosyl-transferases, transporter proteins, transcription factors, and protein kinases (Delye, 2013). Studies of comparing the stress responses due to herbicide application to other abiotic stresses show that they affect similar pathways (Das et al., 2010; Vivancos et al., 2011; Unver et al., 2010). Therefore, glyphosate resistance in weeds can be considered to be a type of rapid evolutionary stress response.

1.4 Organization of Dissertation

The overall aim of this dissertation is to understand how plants react to environmental stress, and evolve to overcome it. I focus on two specific evolutionary scenarios: the adaptation of early land plants to the terrestrial environment, and the evolution of glyphosate resistance in giant ragweed. Early land plants had to withstand an enormous amount of abiotic stress during the move from an aquatic environment to a terrestrial environment. Since eukaryotic protein kinases play an important role in various stress related processes, it is be fair to hypothesize that the EPK family underwent substantial functional elaboration during the transition from charophytes to embryophytes. While some groups of proteins in the family remained unchanged, others underwent expansion or contraction during the process of adaptation. Therefore, if we can isolate these modified groups of proteins, we can estimate the role played by different EPK family proteins during the stress.

Similarly, weeds encounter severe abiotic stress when treated with herbicides, and resistant weeds are results of quick evolution in action. Resistance to glyphosate in particular, has been a huge problem for farmers due to its widespread availability and simple application. Here I focus on a weed, giant ragweed (GR) (*Ambrosia trifida*), for which glyphosate-resistant biotypes have been observed, but the mechanism of resistance has, so far, been unknown. Using a time-course study to compare the

gene expression patterns of resistant and sensitive GR, I identify genes and pathways responsible for conferring resistance to glyphosate.

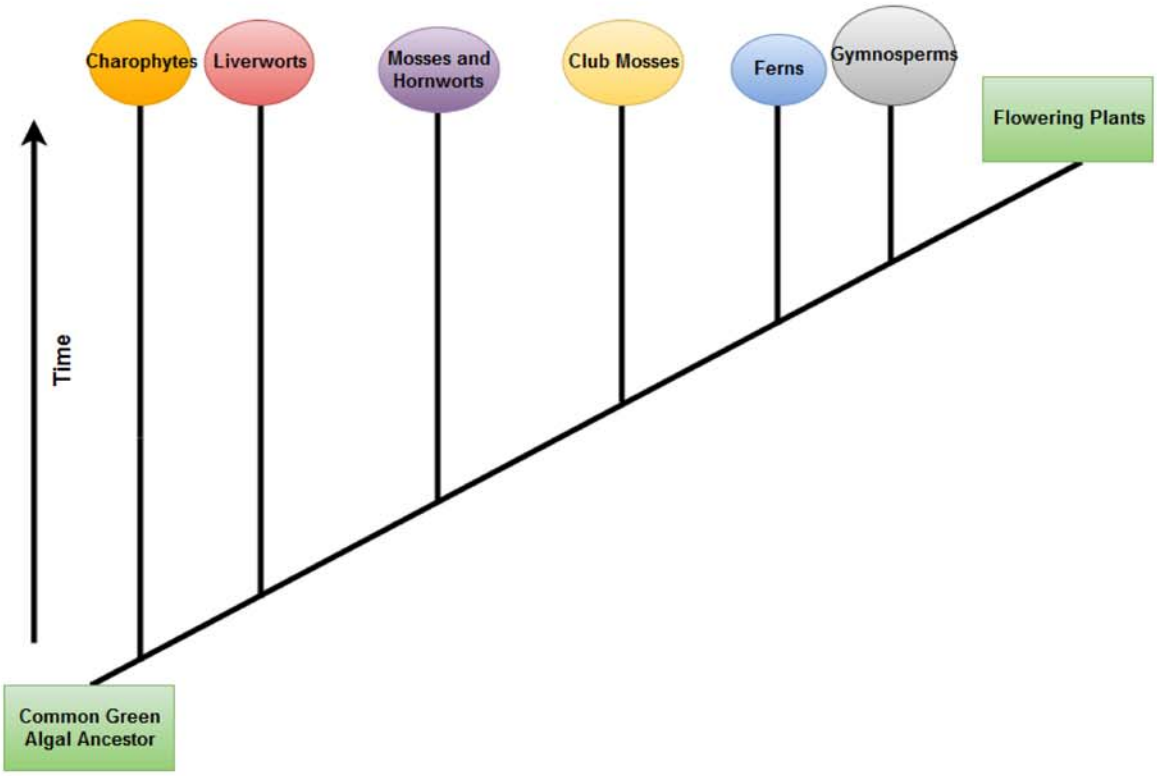

Fig. 1.1.: Cladogram showing plant evolution. All land plants originate from a common green algal ancestor. Liverworts were the first plants to move from an aquatic environment to a terrestrial environment. Liverworts, hornworts and mosses were the three earliest land plant groups.

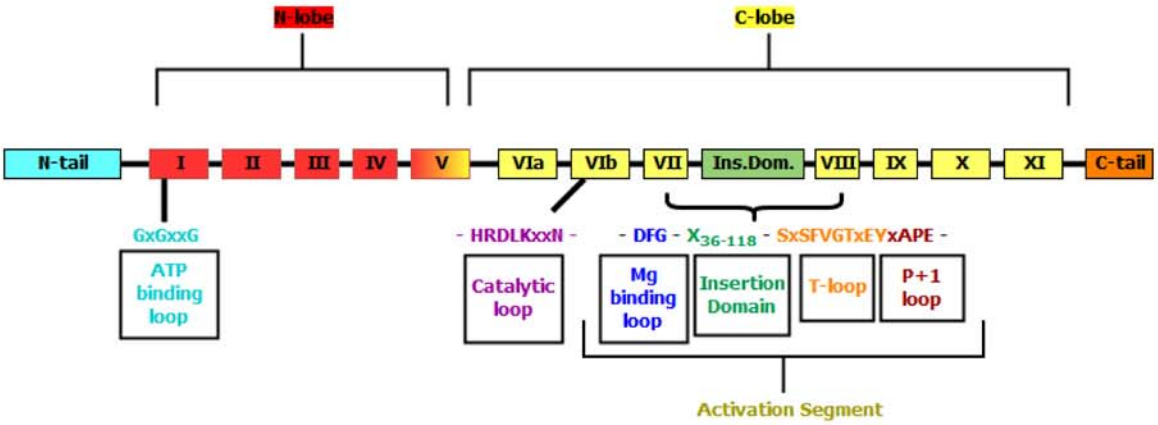

Fig. 1.2.: Eukaryotic Protein Kinase (EPK) domain and its subdomains. The EPK domain consists of 2 lobes the N lobe (end containing the free amine group) and the C lobe (end containing the free carbocyl group). The N and C tails flank 12 conserved subdomains. Motifs that are highly conserved include the GxGxxG loop in subdomain I, which is involved in ATP binding, the HRDLKxxN motif in subdomain VIb, which is a part of the catalytic site and involved in conformational change due to phosphorylation, the DFG motif in subdomain VII, which helps bind Mg2+, and the APE motif in subdomain VIII, which forms a salt bridge with an invariant arginine in subdomain XI.

2. DEVELOPING A GENE MODEL QUALITY EVALUATION METHOD TO SCORE GENE MODELS FROM *PHYSCOMITRELLA PATENS* AND *CHLAMYDOMONAS REINHARDTII*

2.1 Introduction

The advent of next-generation sequencing along with reducing costs and high throughput of sequencing technologies have led to characterization of a wide variety of organisms by genome and transcriptome sequencing (Alkan et al., 2011). An important stage of any genome sequencing experiment is the assembly of reads to form contiguous sequences (contigs) that represent the DNA of the organism, and predicting the structure and position of genes (Zerbino and Birney, 2008). This assembly can be done using a closely-related genome as a template, which is termed a genome-guided assembly method. In the absence of closely-related genomes, an *ab initio* genome annotation is performed where inherent intron and exon signals from the sequence are used to predict the gene characteristics. The accuracy of *ab initio* genome annotation thus vary depending on the nature and evidence for this prediction.

Accuracy of gene models is important for the study of an organism's genetic features (Testa et al., 2015). When performing comparative genomic analyses, incorrect gene models might result in arriving at faulty functional annotations. Besides, when incorrect functional annotations are submitted to public databases such as NCBI, the errors are disseminated further. Gene model accuracy is also very important in designing drugs and treatment mechanisms for various diseases.

The moss *Physcomitrella patens* is a popular model system in the field of genetics in order to study plant development and evolution (Schaefer and Zrÿd, 1997). This is due to its straight-forward developmental configuration and the dominance of the haploid phase in its life cycle (Nishiyama et al., 2003). The genome of *P. patens* was sequenced in 2008, being the first of the mosses to have its genome published (Rensing et al., 2008). Similarly, the unicellular eukaryotic alga *Chlamydomonas reinhardtii* is a popular model system which is used to study photosynthetic processes, biogenesis of chloroplasts, eukaryotic cilial functions, and systems biology, due to the fact that its genetics are well-understood (Rochaix, 1995; Rupprecht, 2009). It was the first algal model organism to be sequenced and its genome was published in 2007 (Merchant et al., 2007). The *P. patens* genome contains 35,938 genes with 84% of the protein sequences predicted to be complete. The closest relative of *P. patens* that has a well-studied genome is *A. thaliana*, which is 400 million years apart in evolutionary distance (Haas et al., 2005). Thus, 63% of the gene modeling in *P. patens* was done by *ab initio* methods. Similarly, the genome of *C. reinhardtii* was published in 2007 and is thought to be 95% complete, with 15,143 predicted genes (Merchant et al., 2007). About 44% of the genes modeled were based on *ab initio* predictions.

As mentioned earlier, *ab initio* gene predictions use a combination of different signals embedded in the genome to construct a statistical model that can predict genes and their exon-intron structures. They depend on signal information and content information of sequences to make predictions of the location and structure of genes (Wang et al., 2004). They do not require any prior experimental knowledge or information about specific genes (Picardi and Pesole, 2010). Statistical methods such as Hidden Markov Models, Neural Networks, and Dynamic Programming are generally used to make gene predictions. Computational gene finders typically look at various sequence elements including splicing regions, transcription promoter and terminator regions, start and stop codons, binding sites for transcription factors, polyadenylation sites, ribosome and topoisomerase II binding sites, and topoisomerase I cleavage

sites (Haussler, 1998). More complicated gene finders include homology searches and predictions of gene structure along with the above listed signal and content sensors to make gene predictions.

The accuracy of *ab initio* gene predictors is dependent on a multitude of factors. GC content is understood to affect the accuracy of most *ab initio* predictors (Nasiri et al., 2011). GC content is defined as the percentage of guanine or cytosine bases in DNA. It is known to be associateed with variations in the gene density and patterns of methylation in genes (Jabbari and Bernardi, 1998; Mouchiroud et al., 1991; Duret et al., 1995). Another major factor that affects accuracy is the frequency and number of introns in the genome (Tenney et al., 2004). Generally, the accuracy of *ab initio* predictors decreases with increased intron size and frequency, as this makes it harder to detect intron-exon boundaries. Most *ab initio* prediction programs also require training data sets to help set their initial parameters, which could bias the output (Hoff and Stanke, 2015).

The performance of *ab initio* gene predictors with real data is questionable to say the least. In a study done with the maize genome, the predictive accuracies of five popular *ab initio* gene prediction software were compared (Yao et al., 2005). It was found that even the best *ab initio* prediction got only 50% of the gene models right. In a more recent study, the authors compared the accuracy of *ab initio* gene finders using the *Toxoplasma gondii* genome (Goodswen et al., 2012). They concluded that in the absence of experimental evidence, the accuracy of such predictors is very low. In a study done with mouse genomic DNA, it was found that *ab initio* gene predictors had a low predictive accuracy (Nasiri et al., 2011). In a study comparing the accuracy of computational gene finders for large DNA sequences, the authors concluded that while the algorithms gave satisfactory results while analyzing single genes with no introns, they had difficulty with genomic sequences with large number of introns and complex gene structures (Guigó et al., 2000). With more and more genomes being

sequenced each day, it is often impossible to experimentally verify the structure and function of each gene. There are possibly many such incorrect gene models present in the *C. reinhardtii* and *P. patens* genomes, since many genes were predicted using *ab initio* predictors. Therefore, a method to score the predicted gene models would help identify the incorrect gene models.

To evaluate a gene model effectively, the most convincing approach would be to integrate multiple sources of evidence. In the current study, two types of evidence were used to create a gene model evaluation score that make use of the characteristics of protein kinases : consensus regions in the primary sequence, and domain relationships. A scoring function was devised to integrate the results of the various approaches.

2.2 Materials and Methods

2.2.1 Data collection

Amino acid sequences of *P. patens* were downloaded from PlantGDB, an online resource for comparative plant genomics database (Duvick et al., 2008). Similarly, sequences for *C. reinhardtii* were downloaded from Joint Genome Institute's (JGI) online genome portal (Nordberg et al., 2014). As references, *A. thaliana* and *O. sativa* sequences were obtained from The Arabidopsis Information Resource (TAIR) and the Rice Genome Annotation Project (Michigan State University) respectively (Lamesch et al., 2012; Ouyang et al., 2007).

2.2.2 Evaluation using consensus catalytic regions

As mentioned in the previous chapter, all eukaryotic protein kinases (EPKs) have a protein kinase domain, with varying levels of conservation. The twelve subdomains of the protein kinase catalytic domain are conserved depending on the family of protein kinases; therefore, any gene model representing protein kinases can be evaluated

based on the presence of consensus catalytic regions at specific locations in the sequence. This was done using three different methods: using Hidden Markov Models (HMMs), regular expression searches, and by comparison to orthologous proteins.

For performing the evaluation based on HMMs, a profile HMM representing the eukaryotic protein kinase (PF00069.20) was obtained from Pfam (Bateman et al., 2000). Pfam is a database of profiles and functional units that can be used to identify protein domains and families. The eukayotic protein kinase HMM contains information depicting the conservation of the protein kinase active site and ATP binding site. The active site and the ATP binding site are the most important functional parts of the protein kinase, and therefore, it is expected that a true protein kinase gene model will contain the two sites at an appropriate spacing. The gene models downloaded for *P. patens*, *C. reinhardtii*, *A. thaliana* and *O. sativa* were then compared to this HMM using a program called HMMER (Finn et al., 2011). The search function hmmscan is a part of HMMER, and it compares protein sequences against a HMM and returns their respective similarities. The E-value obtained from the comparison for each of the protein sequences against the HMM was extracted from the results and used in the scoring function. All sequences that did not match the protein kinase HMM were excluded from further analysis, and the matching sequences were used as the working set of data for subsequent processing and analysis.

To make the technique more robust, a regular expression (regex) search was also performed. A regex is defined as a specific pattern that can be used to search for particular characters, words or patterns of characters. Regex have frequently been used to model protein motifs. In this study, regex pattern depicting the protein kinase domain (PS50011) was obtained from Prosite (Bairoch, 1991). A program called ScanProSite, which looks for the specified protein patterns in the submitted database of proteins, was used for the comparison (De Castro et al., 2006). Each protein is given a score by ScanProSite that correlates with the similarity of the protein to the

regular expression pattern depicting the eukaryotic protein kinase. The hits for each set of protein kinases were extracted and the score for the hit was used in the scoring function.

In addition to the HMM and regex based searches, another approach based on the comparison to protein homologs of the gene model was also performed. Homologs of protein kinases in *P. patens* and *C. reinhardtii* were extracted from *A. thaliana* and *O. sativa* proteomes using BLASTP comparisons, and the similarity was compared (Altschul et al., 1990). Since protein kinases are well-conserved across species, the similarity between proteins and their homologous sequences is bound to give a reasonable idea about gene model reliability. This similarity was measured using BLASTP (Altschul et al., 1990). Protein orthologs in *A. thaliana* and *O. sativa* were functionally identified using the best BLAST hit approach (Altschul et al., 1990). To perform this, gene models from *P. patens* and *C. reinhardtii* were used as the query against a protein database containing the protein kinases from both *A. thaliana* and *O. sativa* in the BLASTP search. The best BLAST hit for each gene model against this combined set of reference protein kinases was then used as the scoring function.

2.2.3 Evaluating gene models using protein domain co-occurrence

A protein domain is a functional, structural and evolutionary unit of a protein. Domains that have similar sequence and structure often have similar functions. As a corollary, proteins that have similar functions tend to have the same domain compositions. For instance, in *A. thaliana*, Calcium-dependent protein kinase-1 (CPK1) and Calcium-dependent protein kinase-6 (CPK6) have similar functions. Looking at their domain organizations, they share the same domains one protein kinase catalytic domain, and four EF-hand calcium-binding domains. Therefore, this was used as the basis of an approach based on protein domain co-occurrence to evaluate gene models.

Since we are analyzing plant protein kinases, protein sequences of *A. thaliana* and *O. sativa* were used to construct a protein domain co-occurrence matrix. All sequences annotated as protein kinases were extracted into a separate file. Thus, a file containing all protein kinases, and another containing all proteins that were not protein kinases were obtained.

An R script was then used to calculate the protein domain co-occurrence matrix for each case. The script analyzes each protein sample and looks through the domains it contains. Each time two domains occur in the same proteion, the pairwise counts for the co-occurring domains are increased. Protein kinase sequences from *P. patens*, *C. reinhardtii*, *A. thaliana* and *O. sativa* were analyzed using InterProScan to identify domains for each protein kinase (Quevillon et al., 2005). The domains in each protein sequence were then analyzed to get the probability that they occur together, given that it is a protein kinase. This was done using Bayes rule. For multiple variables, Bayes rule states that for a given independent variable K and dependent variables $D_1, D_2, \ldots, D_n$, we can say that:

$$P(K|D_1, D_2, \ldots, D_n) = \frac{P(K)\prod_{i=1}^{n} P(D_i|K)}{P(K)\prod_{i=1}^{n} P(D_i|K) + P(\neg K)\prod_{i=1}^{n} P(D_i|\neg K)} \tag{2.1}$$

where,

K denotes the probability that the protein is a protein kinase,

$\neg K$ denotes the probability that the protein is not a protein kinase,

D_i denotes the domain i.

Using Bayes rule, we can successfully predict the probability of a protein kinase having certain domains given that we know the probability of the domain in all protein kinases, the probability of the domain in all non-protein kinases, and the probability of a kinase in a given protein space. In equation (2.1), $P(K)$ and $P(\neg K)$ were values

obtained from the HMM search based on how many of the reference proteins had matches to the protein kinase HMM from the total set of proteins used.

2.2.4 Designing a scoring system

Since the score is based on four different methods, it is important to have a good scoring function that can distinguish between the high quality and the low quality gene models when the results of the different methods are combined. The total score, S_T, for the gene model can be given as,

$$S_T = S'_H + S'_R + S'_O + S'_D \tag{2.2}$$

where S'_H, S'_R, S'_O, and S'_D are the scores from each method used.

S_H is the score from the HMM-based consensus search, and can be calculated as,

$$S_H = -log(hmmscanEvalueofquery) \tag{2.3}$$

S_R is the score from the regular expression pattern search using Prosite. Each gene model is assigned a score by Prosite based on the extent of similarity to the regular expression. That score was used here. The score is unitless and is directly proportional to the similarity to the protein kinase domain.

S_O is the score from the ortholog comparison method, which can be calculated as,

$$S_O = -log(EvaluefromBLAST) \tag{2.4}$$

where the E-value is from the best match of the gene model in the *A. thaliana* and *O. sativa* protein sequences.

S_D is the score from the domain co-occurrence method. This score is obtained from the probability calculation discussed in the previous section. Finally, each score was normalized to a value range of (0,1) using the following:

$$S_{i'} = \frac{S_i - S_{min}}{S_{max} - S_{min}} \tag{2.5}$$

where S_i is the score for each gene model i, S_{min} and S_{max} are the minimum and maximum scores among gene all gene models compared, and $S_{i'}$ is the normalized score.

2.3 Results

2.3.1 Hidden Markov Model search results

The program *hmmsearch* was used to compare the protein sequences of *P. patens*, *C. reinhardtii*, *A. thaliana* and *O. sativa* against the HMM representing the region containing the protein kinase active site and ATP binding site (Finn et al., 2011). The results are tabulated in Table 2.1. Overall, *P. patens* was found to contain 950 matches to the protein kinase HMM, while *C. reinhardtii*, *A. thaliana* and *O. sativa* contained 581, 1361 and 2058 matches respectively. The search was done with a very liberal E-value cutoff of 10. While the cutoff resulted in some false positives, the final score was determined by the results obtained from the other comparisons as well, which minimized the impact of false positives.

2.3.2 Regular expression search results

The next step in the analysis was to do the regular expression analysis using the protein kinase domain pattern obtained from Prosite (Bairoch, 1991). Using the program called ScanProSite, we searched the set of protein kinases obtained for each plant against the regular expression pattern PS50011 which represents the protein kinase active site and the ATP binding site (Fig 2.1 and 2.2) (De Castro et al., 2006).

The results are tabulated in Table 2.2. Overall, close to 95% of all sequences that matched the protein kinase HMM have matches using the regular expression search as well. To find out the reason why a small set of sequences did not have hits for the regular expression search, we performed a BLASTP comparison using the TAIR set of proteins as the database (Lamesch et al., 2012). We found that while most sequences are annotated as protein kinases, the E-value for the hit was high (between 10^{-5} and 0.5). This could be a possible reason why they don't show up as protein kinases in the regular expression search.

2.3.3 Protein domain co-occurrence analysis

Gene models for *P. patens*, *C. reinhardtii*, *A. sativa* and *O. sativa* were each divided into kinases and non-kinases depending on the matches to the protein kinase HMM in the first step (Table 2.3). The program InterProScan was then run for each set of protein kinases and non-protein kinases for each plant (Quevillon et al., 2005). Each sequence was annotated with the set of domains it contains using sequence similarity. Domain annotations include Pfam domains, PANTHER domain annotations, and InterPro architectures. (Bateman et al., 2000; Mi et al., 2016). On analyzing the results file, we found that almost all proteins had annotations for InterPro signatures. Therefore, InterPro annotations were used for constructing the domain co-occurrence matrices. Overall, the co-occurrence matrix constructed using the protein kinases had 221 unique domains, and the matrix constructed using non-protein kinases had 6997 unique domains.

Next, an R script was used to extract the domain annotations from the InterPro results file for both the protein kinases and the non-kinases (Apweiler et al., 2001). Since we are using *A. thaliana* and *O. sativa* sequences as references, we used the sequences from these two plants in order to construct the reference domain co-occurrence matrix. The R script was used to analyze each protein sequence and count

the domains that co-occur in the same sequence. Using this script, we constructed two matrices - one for the protein kinases and one for the non-protein kinases (Figures 2.3 and 2.4). These matrices were used as the reference to calculate the conditional probabilities of protein domain co-occurrences. The next step was to analyze the sequences of the two reference plants using the domain co-occurrence matrix to verify how well the method works. A function was written to calculate the probability that a protein sequence encodes a kinase given that it contains certain protein domains. This was done using the Bayes rule as mentioned in the equation (2.1).

The results for the domain co-occurrence for protein kinases and non-protein kinases for *A. thaliana* and *O. sativa* are shown in Figures 2.5, 2.6, 2.7 and 2.8. Among the groups of protein kinases, sequences from *Arabidopsis* performed very well in the test, with all of the sequences scoring having a probability of 0.8 for being a protein kinase. Similarly for the protein kinase sequences from *O. sativa*, all protein kinases scored above 0.8 for the probability.

For the non-protein kinases, in *A. thaliana*, close to 13,000 proteins scored 0 for the probability of being a kinase, while approximately 1500 had a probability of more than 0.9. In *O. sativa*, there was a similar trend, as close to 14000 having a probability of less than 0.2, while around 2000 had a probability of greater than 0.9. We decided to investigate the sequences that were in the non-kinase group, but still had a high probability of being a kinase. We used UniProt to retrieve the functional annotations for the sequences that had greater than 0.9 probability of being a kinase in the non-kinase sequence group. We compared the results from this study to the set of known protein kinases and non-protein kinases from *A. thaliana*. For the protein kinases, all the proteins had a probability of 1 (Figure 2.9). For the non-protein kinases, a majority of the protein kinases had a probability score of less than 0.2, but just like the HMM-based non-protein kinases, there was a small set of proteins that had a high probability (Figure 2.10). Therefore, these probability scores have

the same distribution as the protein kinases and the non-protein kinases from the HMM-based classification.

In *A. thaliana*, out of a total of 1656 sequences having a probability greater than 0.9, 797 sequences had unreviewed functions with no experimental verification, 12 were annotated as protein kinases, 52 were annotated as other kinases, 154 were phosphatases, 117 were proteins involved in phospho-transfer reactions, and 183 were annotated with a probable but unconfimed function. Similarly, in *O. sativa*, out of the 2193 sequences that had a high probability score, 1003 were annotated as protein kinases, 114 were expressed proteins with no function, 95 were phosphatases, and 140 proteins were phosphor-transfer proteins. Therefore, it seems like the domain co-occurrence is picking up protein kinases that were not detected by a simple HMM search alone. It also seems to pick up sequences involved with phosphoryl group transfer that are very similar to protein kinases. Thus, the use of protein domain co-occurrence alone may not be sufficient since it seems to contain false positives. So it is important to use the results from the protein domain co-occurrence with other homology-based methods for accurate scoring and gene model evaluation.

The next step was to compute the protein kinase probabilities for *P. patens* and *C. reinhardtii* sequences. Using a procedure similar to that used for the two reference plants, we used the domain co-occurrence matrices computed for protein kinases and non-protein kinases to determine the probability of the protein being a protein kinase given the list of domains it contains. We calculated the probabilities separately for the protein kinases and non-protein kinases determined using the HMM search.

In *P. patens*, for the protein kinase group, we found that all the sequences had a probability greater than 0.8 (Figure 2.11). In the non-kinase group, more than 10000 sequences had a probability less than 0.2, while close to 2000 sequences had a probability greater than 0.9 (Figure 2.12). In *C. reinhardtii*, almost all sequences

except a handful had a probability greater than 0.8 in the kinase group (Figure 2.13). In the non-kinase group, more than 3500 sequences had a probability less than 0.2, and more than 4000 had a probability less than 0.6 (Figure 2.14). Approximately 1000 sequences had a probability of more than 0.9. These results mirror the results we obtained for *Arabidopsis* and *O. sativa*, and it is possible that most of these sequences are in fact protein kinases which did not get picked up in the HMM search.

In order to further probe the set of sequences that are predicted as protein kinases even in the non-protein kinase group, we combined the sequences that had a protein kinase probability greater than 0.9 from the non-protein kinase group from *P. patens* and *C. reinhardtii* and combined them with the protein kinase sequences that we had originally classified using the HMM search. Thus, we had a total of 2221 sequences in *P. patens* and 1422 sequences in *C. reinhardtii*. We ran hmmsearch again using this new set of sequences in order to verify if we were able to get any new matches to the protein kinase domain HMM. Unfortunately, there was no change in the number of matches to the protein kinase domain in either of the plants. Similarly, we compared the new set of sequences against the protein kinase Prosite profile using ScanProSite. Once again, we found no new hits to the protein kinase domain among the sets of sequences for both the plants. This suggests that there may be inherent changes in the protein kinase domain region either due to sequence deletions or due to fusion of two proteins leading to reduced similarity against the protein kinase domain HMM and Prosite profile. We used BLASTP to calculate the E-value for the comparison against the orthologs in *A. thaliana* and *O. sativa* as done previously. The overall distribution of E-value scores for *P. patens* and *C. reinhardtii* can be found in Figure 2.15 and 2.16 respectively.

2.3.4 Scoring the protein kinase gene models

We constructed a data table for storing the results from the different analyses using an R script. The score for each sequence was calculated by combining the results from the HMM search, the Prosite motif search, the BLAST homology search, and the protein domain co-occurrence study. Specifically, we used the E-values from the hmmscan results and the BLASTP search, the ScanProSite scores, and the domain co-occurrence probability scores for each sequence to calculate the score.

Since not all sequences had scores in all four categories, we imputed the missing values by using the "na.roughfix()" function in the randomForest package in R. The function imputes missing values by using the column medians for each column. Thus, we were able to obtain the scores for each sequence and each method.

As mentioned previously, we calculated the negative logarithm (base 10) of the E-value to obtain scores for the HMM-based search and the BLASTP search. The scores for three methods - HMM, Prosite, and BLASTP, were then normalized to a value between 0 and 1 using equation (2.5). This step was skipped for the protein domain co-occurrence based score because the score is already a value between 0 and 1. The sum of the normalized scores is the final score for each gene model.

The statistics for the final score for each plant is given in Table 2.4. The scoring function was tested using protein kinases from *A. thaliana* (Figure 2.17) and *O. sativa* (Figure 2.18). The scores for *A. thaliana* ranged from 1.838 to 3.481, while for *O. sativa*, the scores ranged from 1.626 to 3.606. On analyzing known protein kinases in this set, it was found that all protein kinases were present in the scores higher than the first quartile score. In *P. patens*, the gene model scores ranged between 0.9073 and 3.7140 with a mean score of 2.5250 and a median of 2.5230 (Figure 2.19). In *C. reinhardtii*, the scores ranged from a minimum of 0.6811 and a maximum of 3.8480

with a mean and median score of 2.0930 and 2.0020 respectively (Figure 2.20). Protein sequences with a score higher than the first quartile score (2.1730 for *P. patens* and 1.8740 for *C. reinhardtii*) were used for the functional analysis in the next chapter.

2.4 Discussion

We devised a scoring method to evaluate and score the protein kinase gene models in *P. patens* and *C. reinhardtii*. Using the *A. thaliana* and *O. sativa* gene models as references, the gene models of the early plants were compared in order to verify the integrity of the gene models using consensus catalytic regions comparisons and protein domain co-occurrence studies.

2.4.1 Evaluation using consensus catalytic regions

A combination of hidden markov model based search, regular expression search and orthologous proteins based search was used to evaluate the protein kinase gene models. We were able to shortlist a set of protein kinases to be used as the reference by running hmmscan against the two reference proteomes and using it for guiding the analysis. Even though there was a slight reduction in the number of protein kinases detected using Prosite regular expression search when compared to the HMM-based search, we found that they had very weak matches to the reference protein kinases we searched against. Therefore, it is probable that the Prosite based search was more sensitive to minor sequence changes than the HMM based search, thus neglecting any sequence that deviated from the protein kinase domain regular expression. This means that the HMM based method allows for the flexibility of having a slightly modified protein kinase domain due to the method having prior probabilities assigned to a variety of positions in the protein kinase domain HMM. The E-value cutoff of the HMMER search was kept at a very liberal value which could also explain the difference.

2.4.2 Protein domain co-occurrence analysis

We used the sequences from the reference proteins that had hits against the protein kinase HMM for the hmmscan search as the base for constructing the protein domain co-occurrence matrix. Each sequence had an annotation that contained the set of functional protein domains that it contained. This information was obtained using InterProScan, which compares the protein sequence against the set of known protein domains in Pfam, Prosite, PANTHER and InterPro signatures. The domain information for the reference sequences was used to construct a pairwise matrix which contained the number of times each domain occurred with the another domain. A total of 221 unique domains for the protein kinases, and 6997 unique domains for the non-protein kinases were used to construct the pairwise matrix.

When the domain co-occurrence study was first tested using the reference sequences as a benchmark, we found that while the set of protein kinases had high probability scores as expected, there were a significant number of sequences from the non-protein kinase set of sequences that had greater than 0.9 probability of being a protein kinase. On analyzing the functions of these outliers, we found that many of them were protein kinases that did not match either the HMM of the protein kinase domain, nor the regular expression of the protein kinase domain. This leads us to speculate that these sequences may have insertions, deletions, fusions or other mutations in the protein kinase domain which prevents them from matching the HMM and the regular expression pattern of the domain. A large number of sequences also had unknown functions, which could signify the presence of fusion proteins that have domain arrangements similar to protein kinases, but do not have a functional protein kinase domain. Other proteins that had a high probability were mostly involved in phosphorus group transfer functions which suggests that the method may produce a small number of false positives from closely related proteins.

2.4.3 Protein kinase gene model scoring

Finally, the scores from the different methods were combined after normalization and imputation of missing values. While imputation using the column medians may have affected the scores, it was mandatory to avoid the presence of missing values which impeded the calculation of final scores. The scores were designed to eliminate the presence of proteins annotated as protein kinases but lacking the required stuctural and functional domains for protein kinase activity. With that in mind, we chose proteins that scored among the top 75% of the scoring system for functional annotation since this set could have sets of protein kinases that had previously unknown functions.

2.4.4 Future directions

We have shown that a gene model scoring system utilizing the presence of conserved regions, and domain co-occurrence performs reasonably well. Therefore, this can be easily expanded to other protein families in the future. That could lead to having a generalized gene model scoring system that can be designed based on conservation of specific gene families. Such a system would drastically reduce the need for manual curation, and will make genome annotation significantly more reliable and faster. The model can also be expanded to species other than plants.

Table 2.1.: The number of protein kinases that match the protein kinase HMM domain across the four species of plants

Species	**No. of protein kinases**
Physcomitrella patens	950
Chlamydomonas reinhardtii	581
Arabidopsis thaliana	1361
Oryza sativa	2058

Table 2.2.: The number of matches to the protein kinase domain found using ScanProSite among the four species

Species	**No. of matches using PROSITE**
P. patens	910
C. reinhardtii	536
A. thaliana	1339
O. sativa	2005

Table 2.3.: The number of protein kinases and non-protein kinases used for the protein domain co-occurrence matrix across the four species of plants

Species	**No. of protein kinases**	**No. of non-protein kinases**
Physcomitrella patens	950	37404
Chlamydomonas reinhardtii	581	16128
Arabidopsis thaliana	1361	34025
Oryza sativa	2058	64280

Table 2.4.: The statistics for the final score that represents the strength of the gene model being a protein kinase for *P. patens* and *C. reinhardtii*.

Species	**Minimum value**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum value**
P. patens	0.9073	2.1730	2.5230	2.5250	2.9370	3.7140
C. reinhardtii	0.6811	1.8740	2.0020	2.0930	2.2530	3.8480
A. thaliana	1.838	2.939	2.949	2.944	2.949	3.481
P. patens	1.626	2.890	2.911	2.895	2.924	3.606

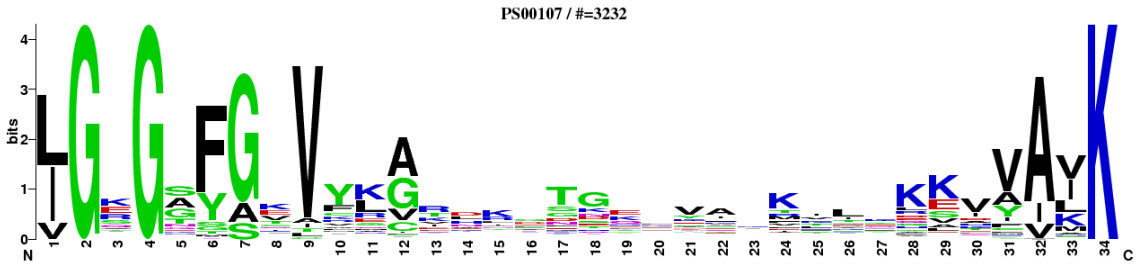

Fig. 2.1.: A representation of the protein kinase active site domain used in ScanProSite. The X axis denotes the position, and the Y axis represents the bit score obtained from BLAST and HMMER log-odds scores.

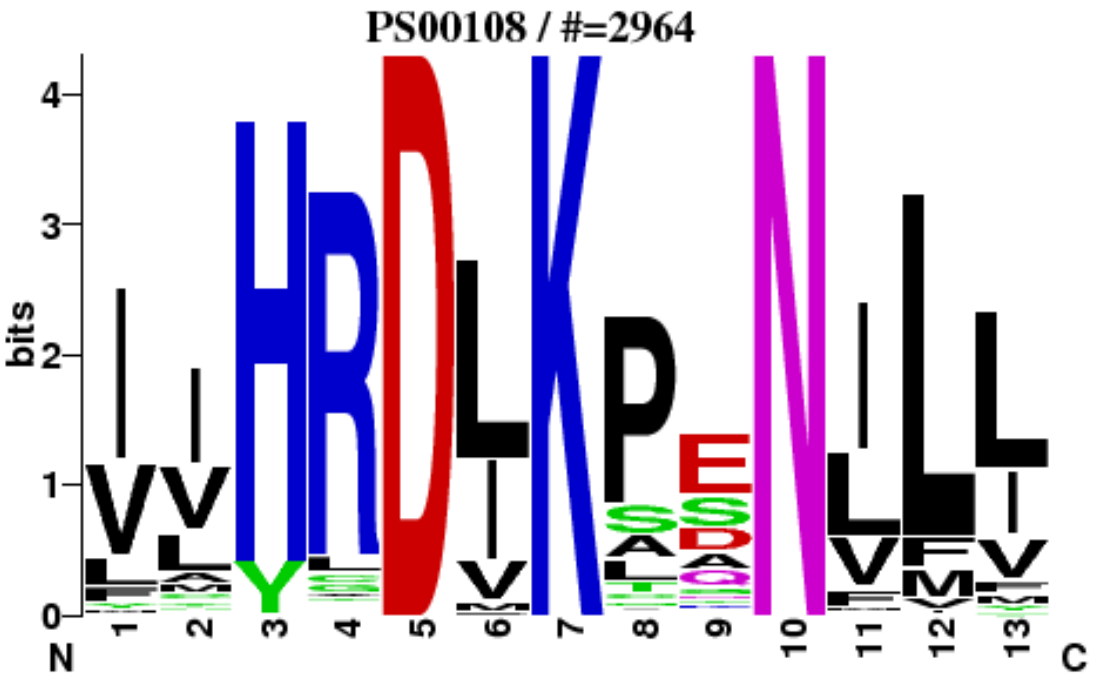

Fig. 2.2.: A representation of the protein kinase ATP-binding domain used in ScanProSite.The X axis denotes the position, and the Y axis represents the bit score obtained from BLAST and HMMER log-odds scores.

row.names	IPR020846	IPR011701	IPR002048	IPR011992	IPR013130	IPR017927	IPR017938	IPR029654	IPR013623	IPR013121
IPR020846	3777	278	1	1	1	1	1	1	1	1
IPR011701	278	147	1	1	1	1	1	1	1	1
IPR002048	1	1	30701	9718	106	91	133	17	100	100
IPR011992	1	1	9718	4435	105	97	130	15	105	99
IPR013130	1	1	106	105	73	69	82	9	52	68
IPR017927	1	1	91	97	69	131	134	9	49	64
IPR017938	1	1	133	130	82	134	189	13	66	78
IPR029654	1	1	17	15	9	9	13	9	8	7
IPR013623	1	1	100	105	52	49	66	8	55	51
IPR013121	1	1	100	99	68	64	78	7	51	65
IPR000778	1	1	363	367	208	192	274	29	186	199
IPR013112	1	1	103	105	76	73	84	8	53	71
IPR004776	1	1	1	1	1	1	1	1	1	1
IPR024571	1	1	1	1	1	1	1	1	1	1
IPR014782	1	1	1	1	1	1	1	1	1	1
IPR001930	1	1	1	1	1	1	1	1	1	1
IPR027417	1	1	225	237	1	1	1	1	1	1
IPR000261	1	1	550	263	1	1	1	1	1	1
IPR022812	1	1	49	26	1	1	1	1	1	1
IPR031692	1	1	45	24	1	1	1	1	1	1
IPR030381	1	1	49	26	1	1	1	1	1	1

Fig. 2.3.: Figure showing a part of the reference domain co-occurrence matrix constructed for protein kinases

row.names	IPR002902	IPR000719	IPR008271	IPR011009	IPR013320	IPR013103	IPR017441	IPR003527	IPR001220	IPR001245
IPR002902	3153	1296	680	772	785	9	661	1	1	559
IPR000719	1296	26981	9944	13171	5596	15	8184	227	539	3924
IPR008271	680	9944	4669	5065	2618	9	3481	38	232	2244
IPR011009	772	13171	5065	7573	3079	9	4184	95	245	3026
IPR013320	785	5596	2618	3079	3823	8	2393	1	510	1314
IPR013103	9	15	9	9	8	7	6	1	3	3
IPR017441	661	8184	3481	4184	2393	6	3903	95	220	1559
IPR003527	1	227	38	95	1	1	95	95	1	1
IPR001220	1	539	232	245	510	3	220	1	255	27
IPR001245	559	3924	2244	3026	1314	3	1559	1	27	8081
IPR032675	1	7352	2924	4075	2749	1	2707	1	1	1494
IPR013210	1	1569	542	856	532	1	562	1	1	262
IPR001611	1	9623	3727	5154	3483	1	3820	1	1	1555
IPR003591	1	8700	3612	4648	3137	1	3393	1	1	1266
IPR001480	1	3513	1792	1947	1796	1	1433	1	1	741
IPR003609	1	1414	763	816	768	1	546	1	1	399
IPR024171	1	490	257	265	263	1	195	1	1	94
IPR021820	1	112	58	61	59	1	18	1	1	60
IPR000858	1	544	282	302	286	1	213	1	1	125
IPR024788	1	427	226	247	230	1	221	1	1	188
IPR002048	1	4070	1584	1619	1	1	1432	1	1	1
IPR018247	1	1486	577	588	1	1	524	1	1	1

Fig. 2.4.: Figure showing a part of the reference domain co-occurrence matrix constructed for non-protein kinases

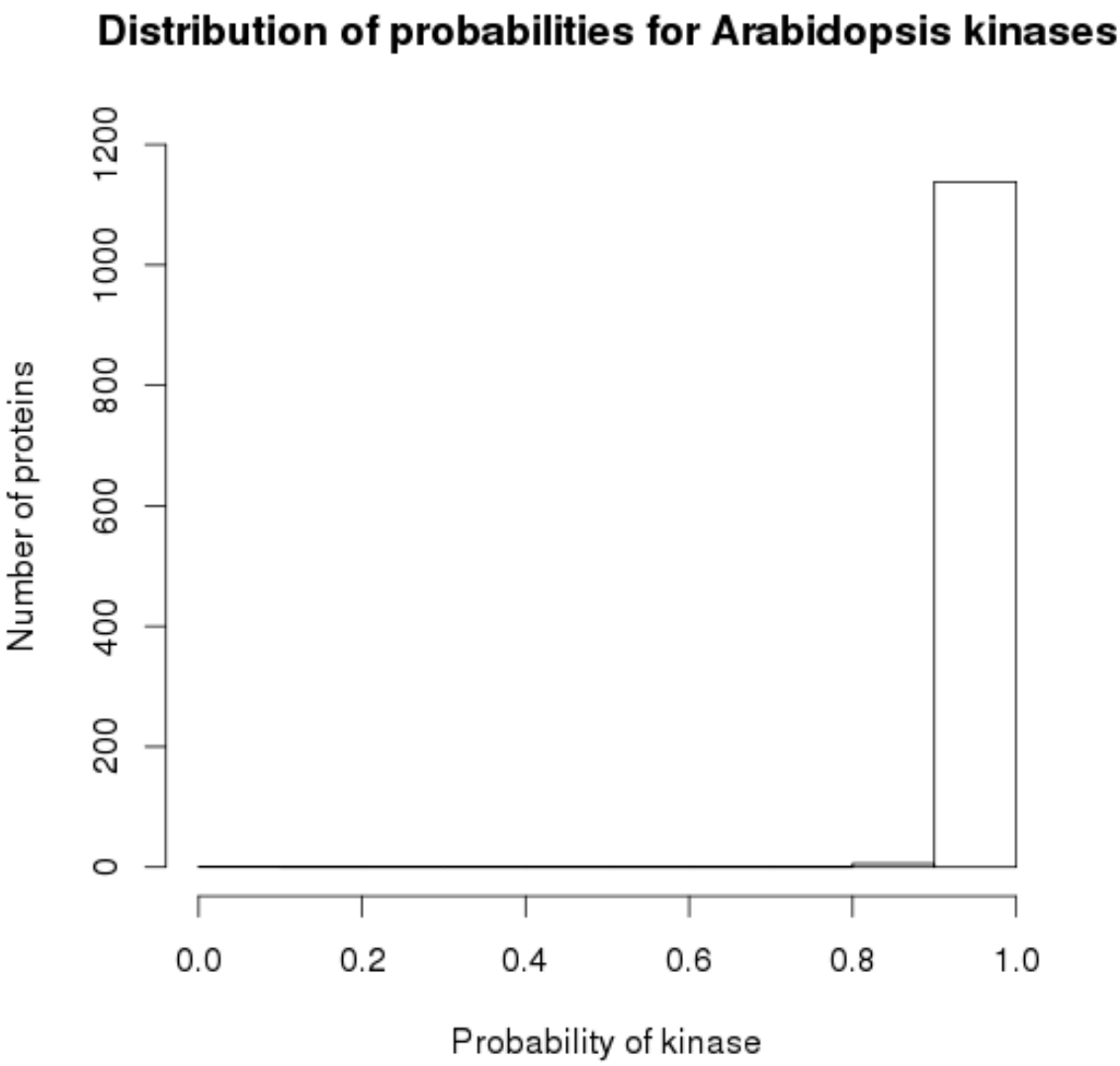

Fig. 2.5.: Histogram showing the distribution of probabilities in the set of protein kinases from *A. thaliana*

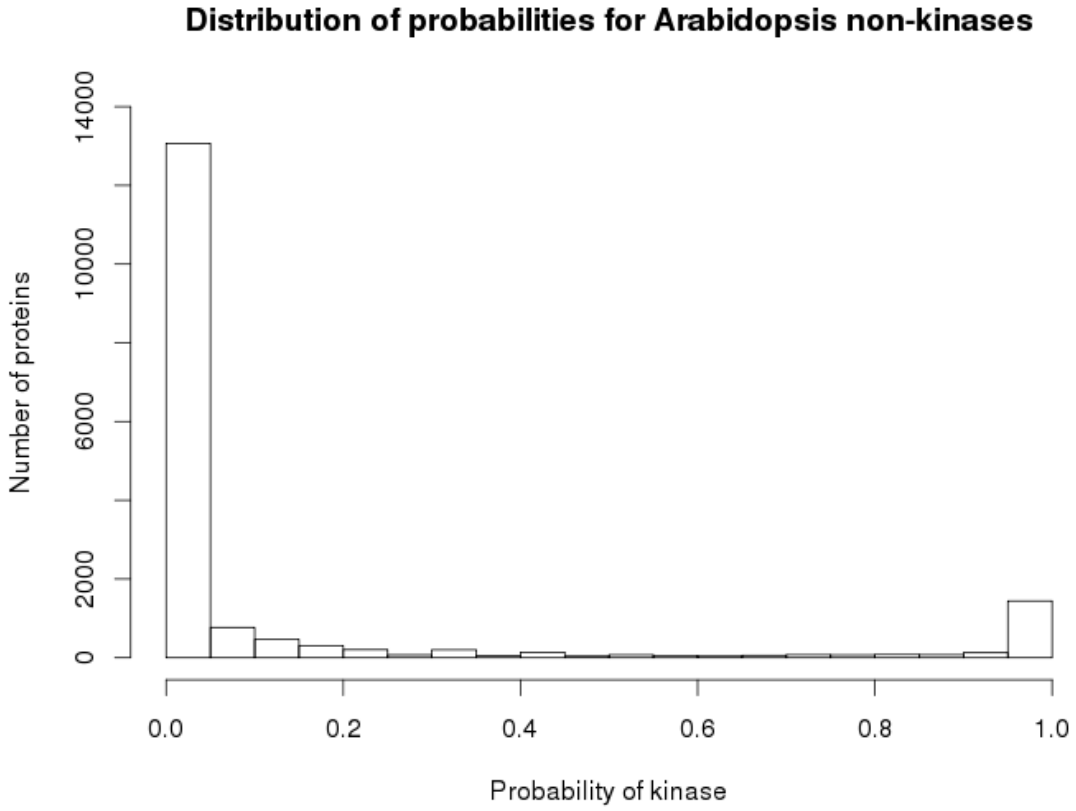

Fig. 2.6.: Histogram showing the distribution of probabilities in the set of non-protein kinases from *A. thaliana*

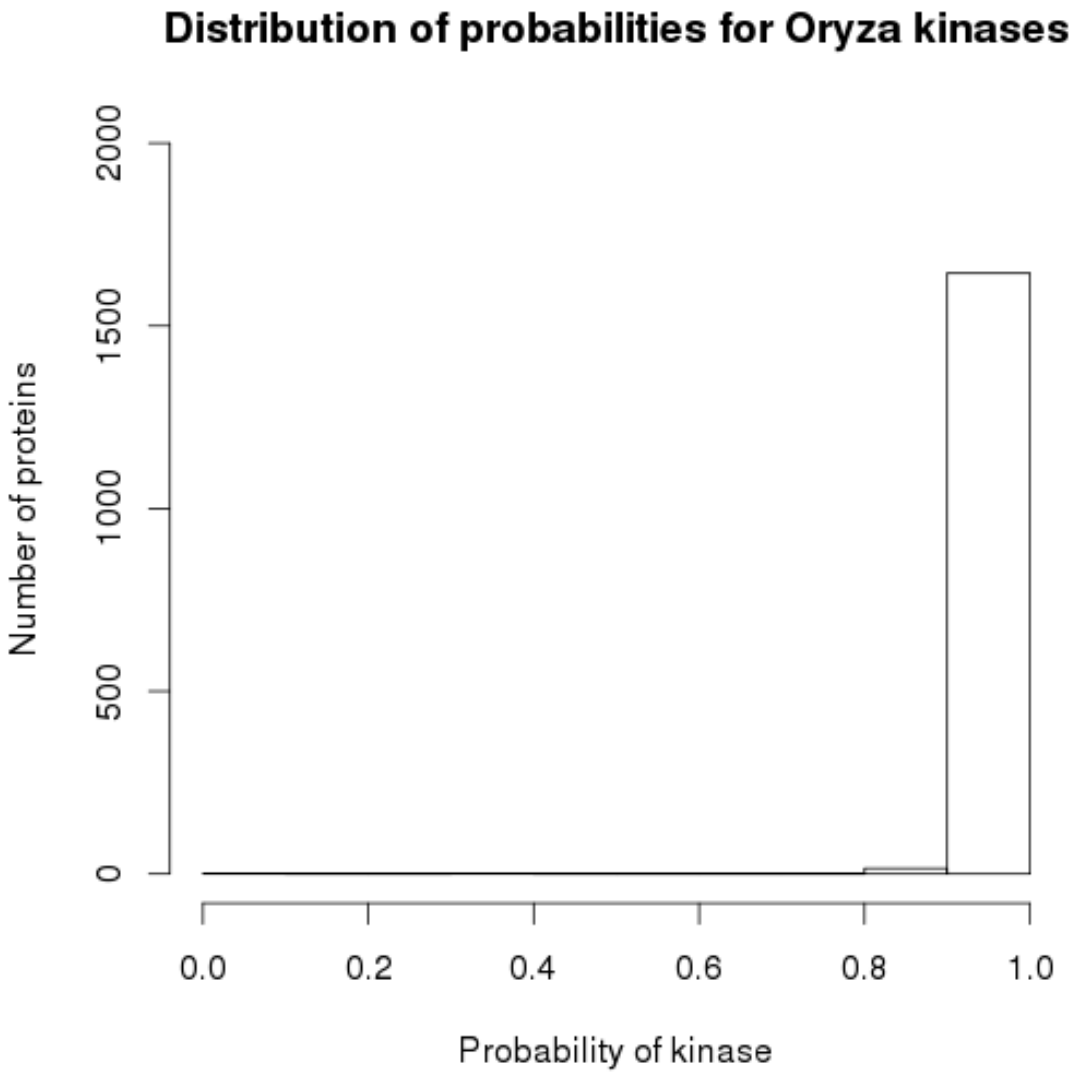

Fig. 2.7.: Histogram showing the distribution of probabilities in the set of protein kinases from *O. sativa*

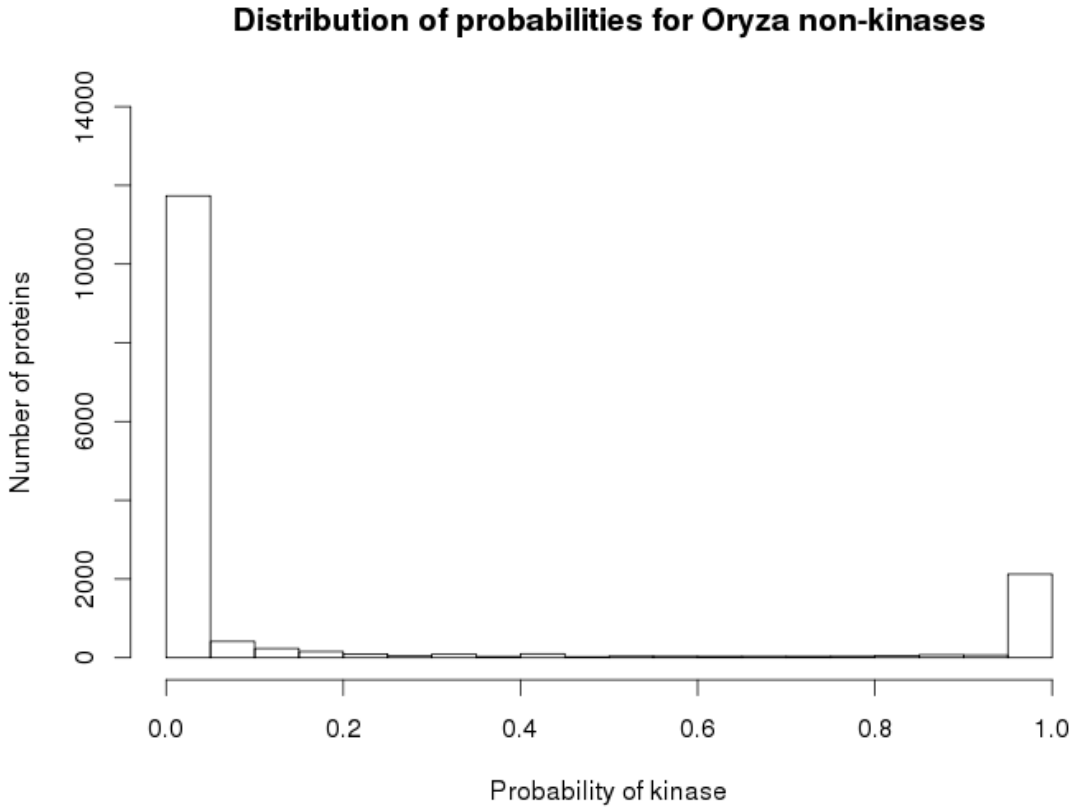

Fig. 2.8.: Histogram showing the distribution of probabilities in the set of non-protein kinases from *O. sativa*

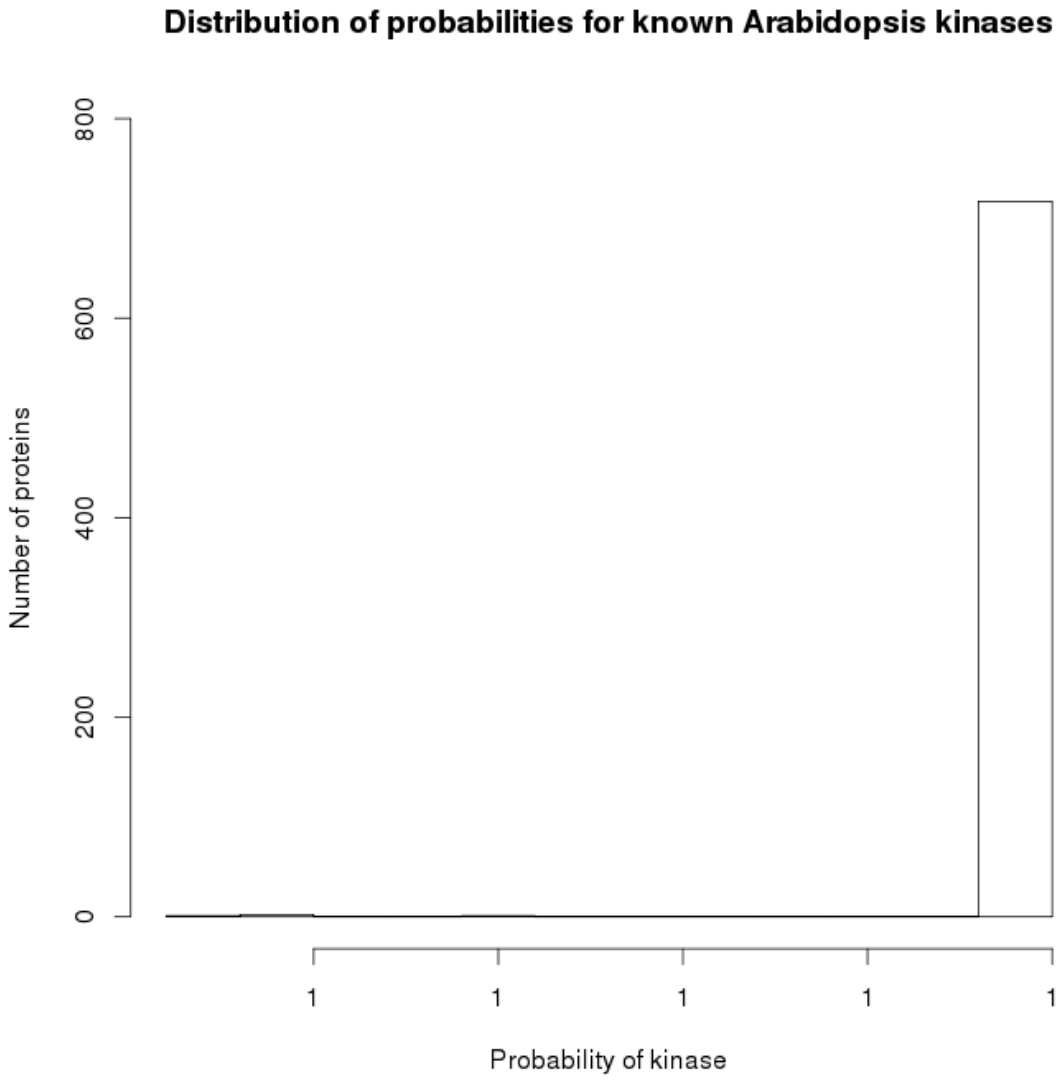

Fig. 2.9.: Histogram showing the distribution of probabilities in the set of known protein kinases from *A. thaliana*

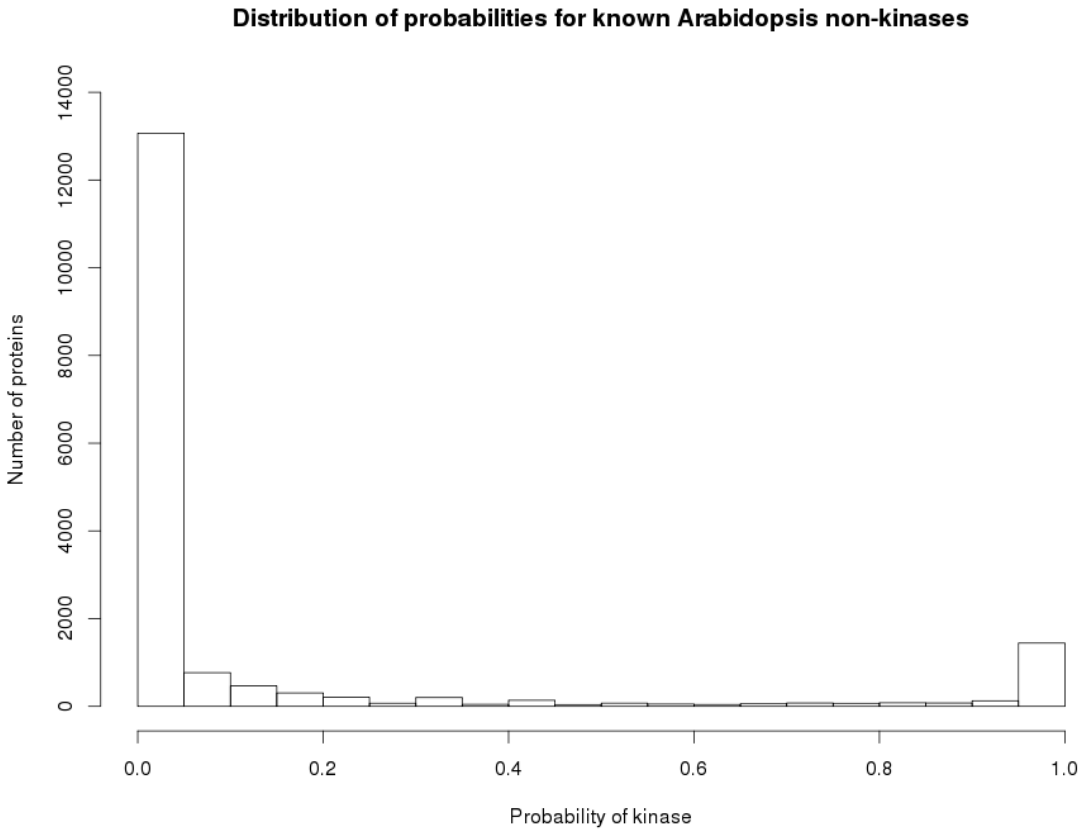

Fig. 2.10.: Histogram showing the distribution of probabilities in the set of known non-protein kinases from *A. thaliana*

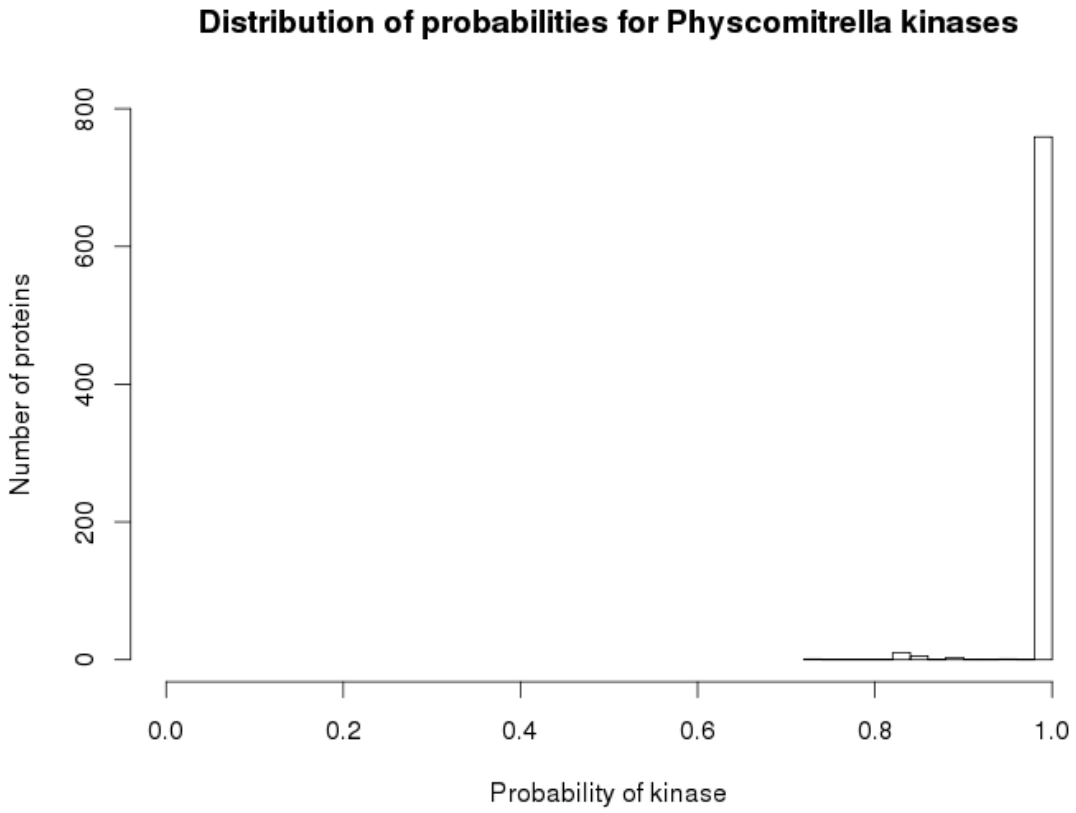

Fig. 2.11.: Histogram showing the distribution of probabilities in the set of protein kinases from *P. patens*

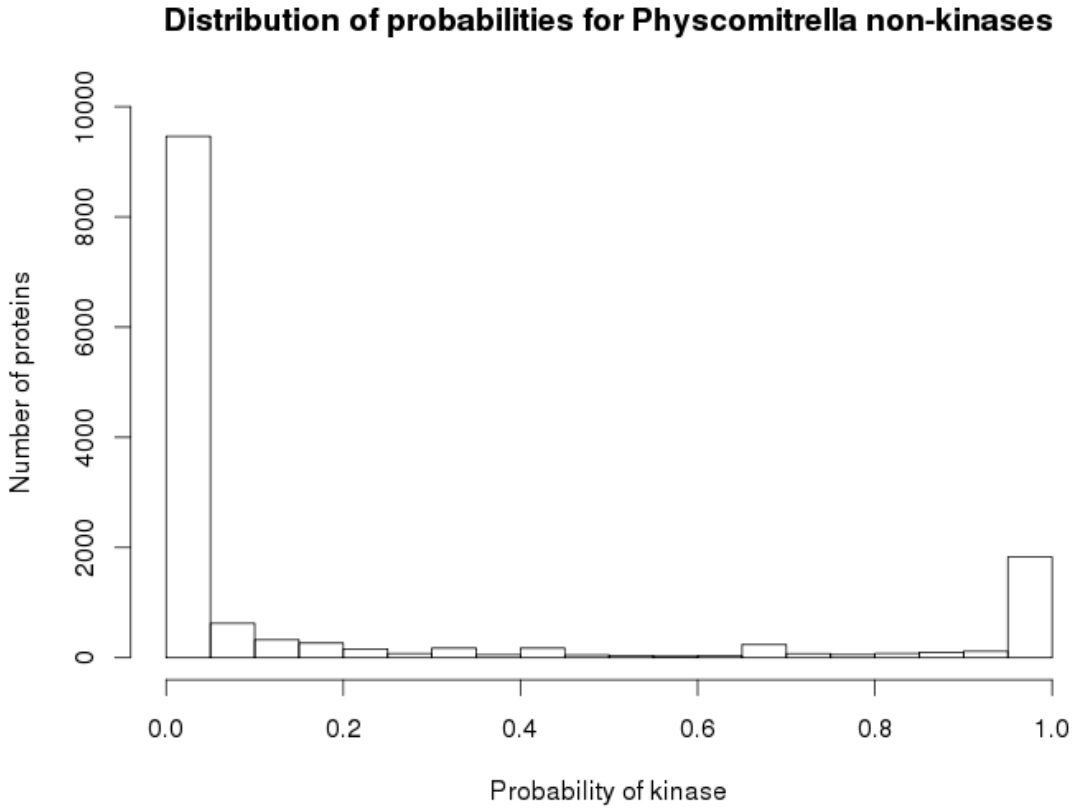

Fig. 2.12.: Histogram showing the distribution of probabilities in the set of non-protein kinases from *P. patens*

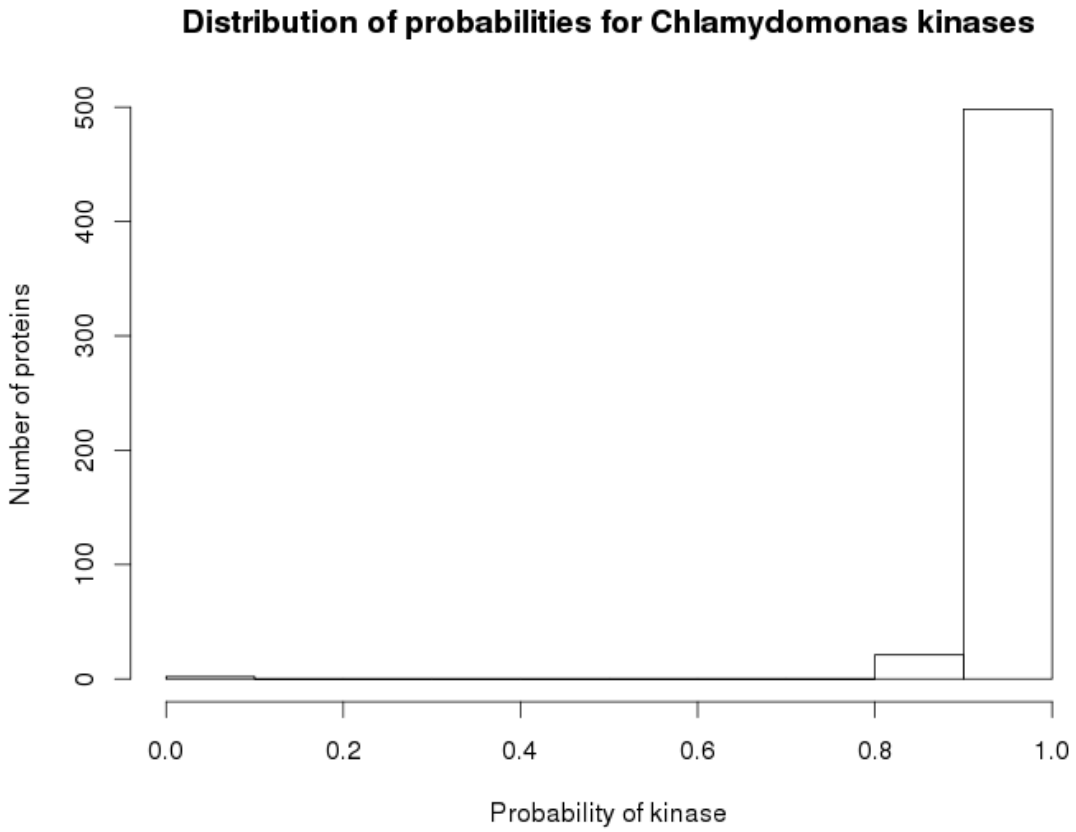

Fig. 2.13.: Histogram showing the distribution of probabilities in the set of protein kinases from *C. reinhardtii*

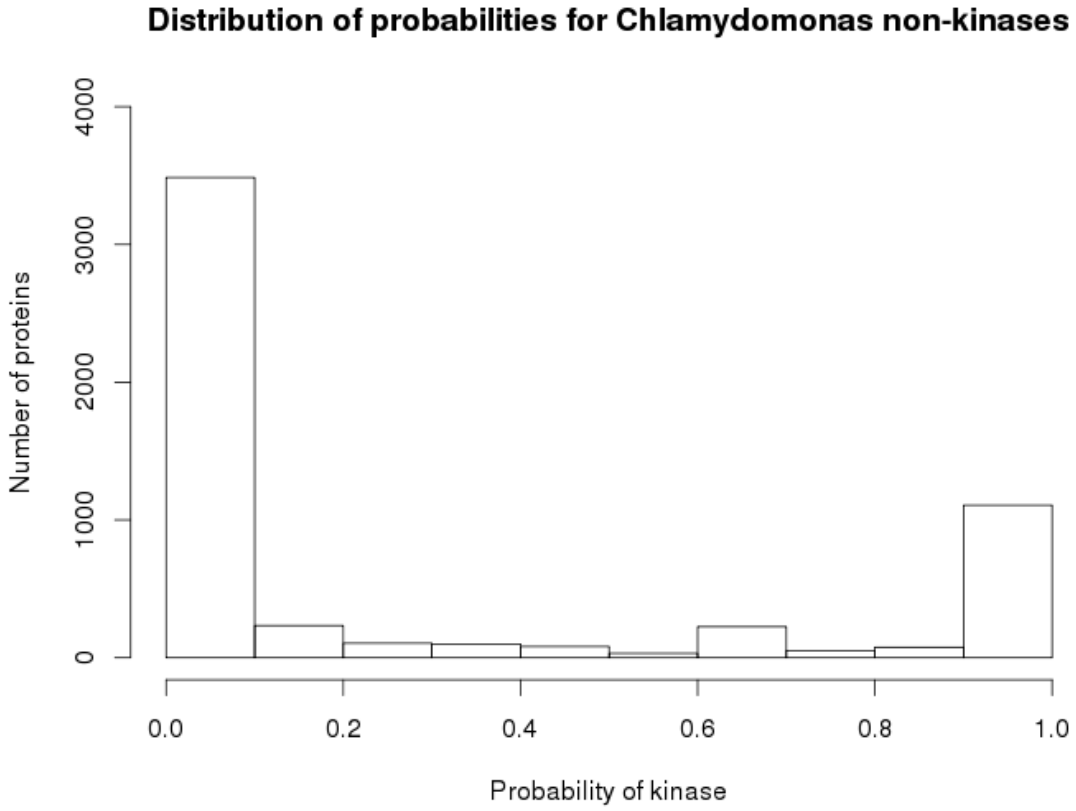

Fig. 2.14.: Histogram showing the distribution of probabilities in the set of non-protein kinases from *C. reinhardtii*

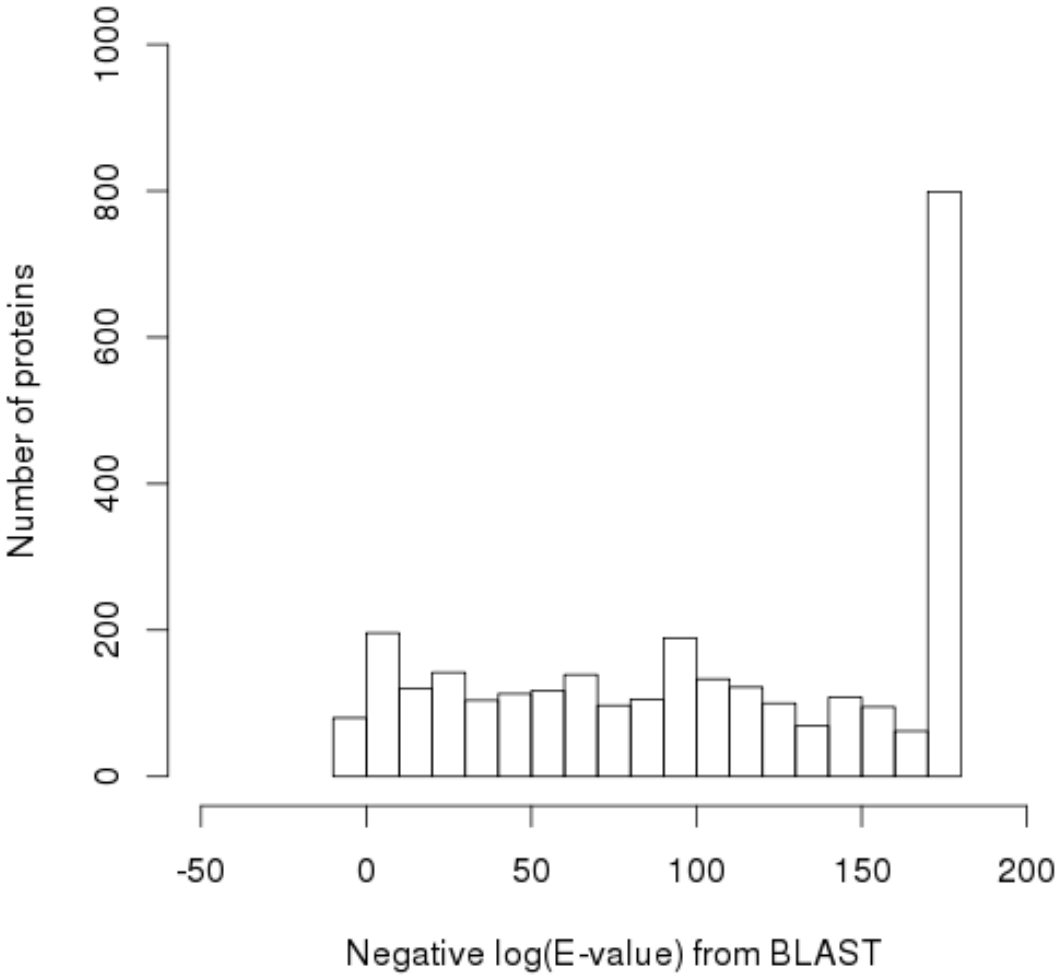

Fig. 2.15.: Histogram showing the distribution of E-value based BLASTP scores for the potential protein kinases from *P. patens*

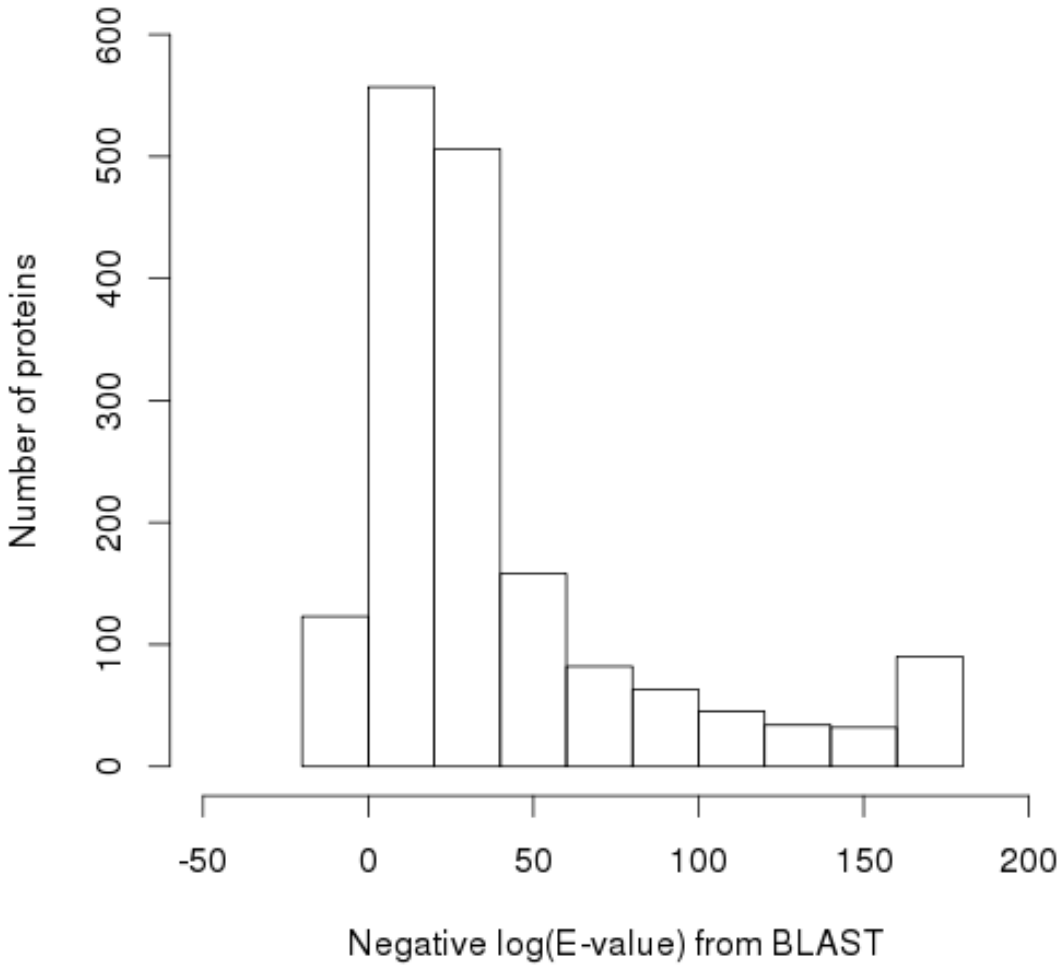

Fig. 2.16.: Histogram showing the distribution of E-value based BLASTP scores for the potential protein kinases from *C. reinhardtii*

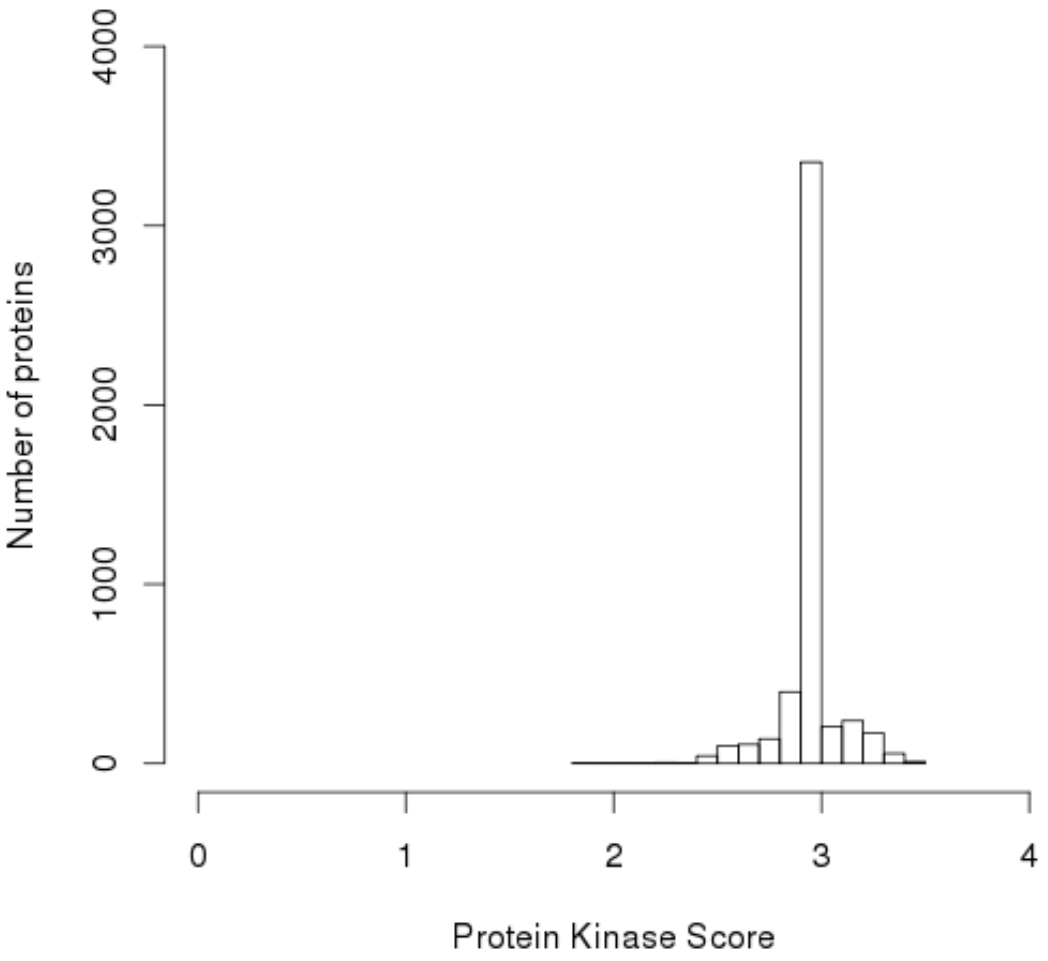

Fig. 2.17.: Histogram showing the distribution of final scores in the set of protein kinases from *A. thaliana*

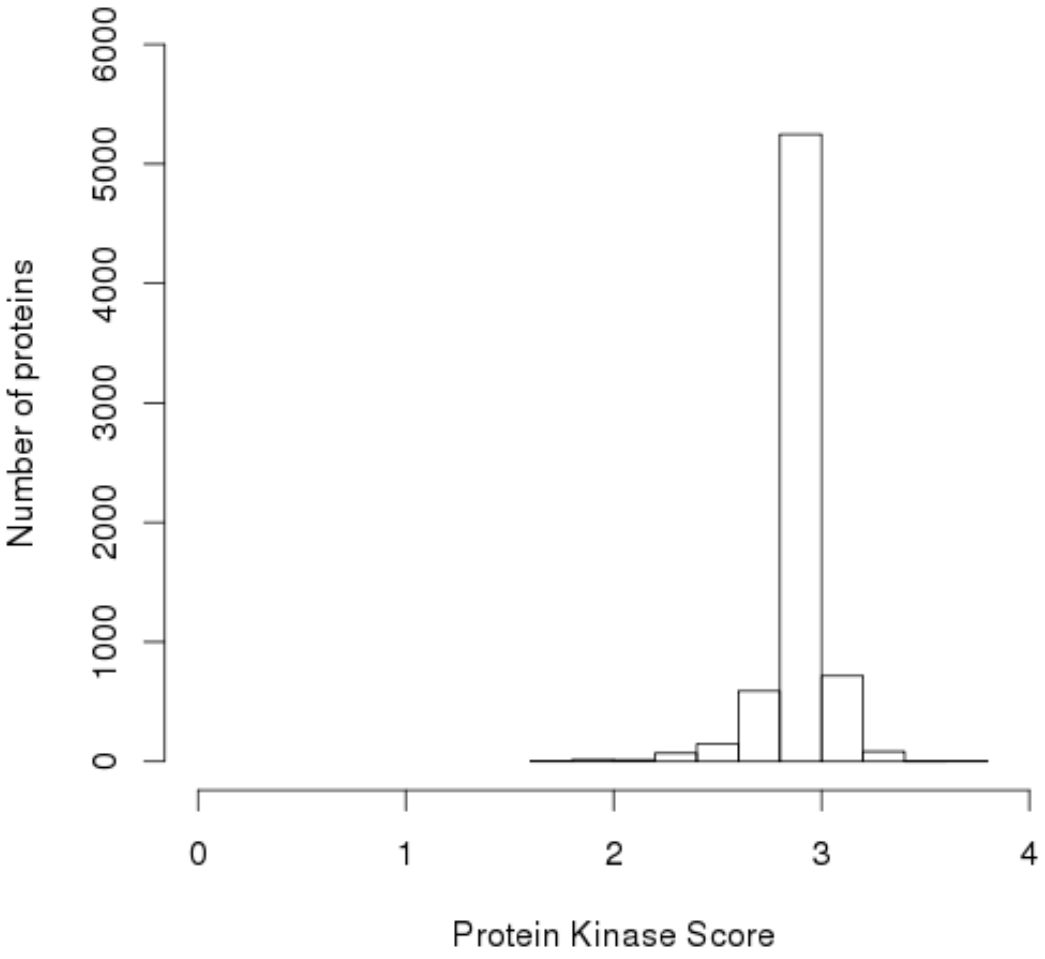

Fig. 2.18.: Histogram showing the distribution of final scores in the set of protein kinases from *O. sativa*

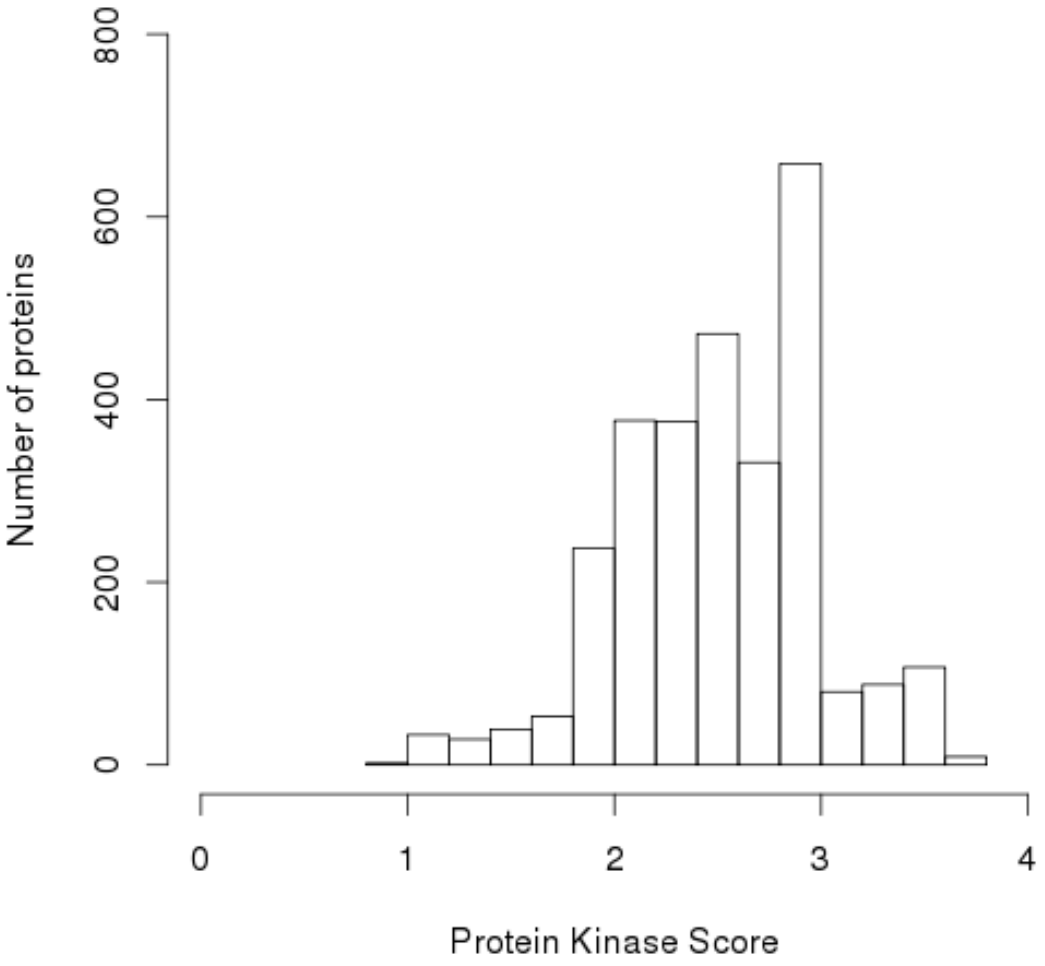

Fig. 2.19.: Histogram showing the distribution of final scores in the set of protein kinases from *P. patens*

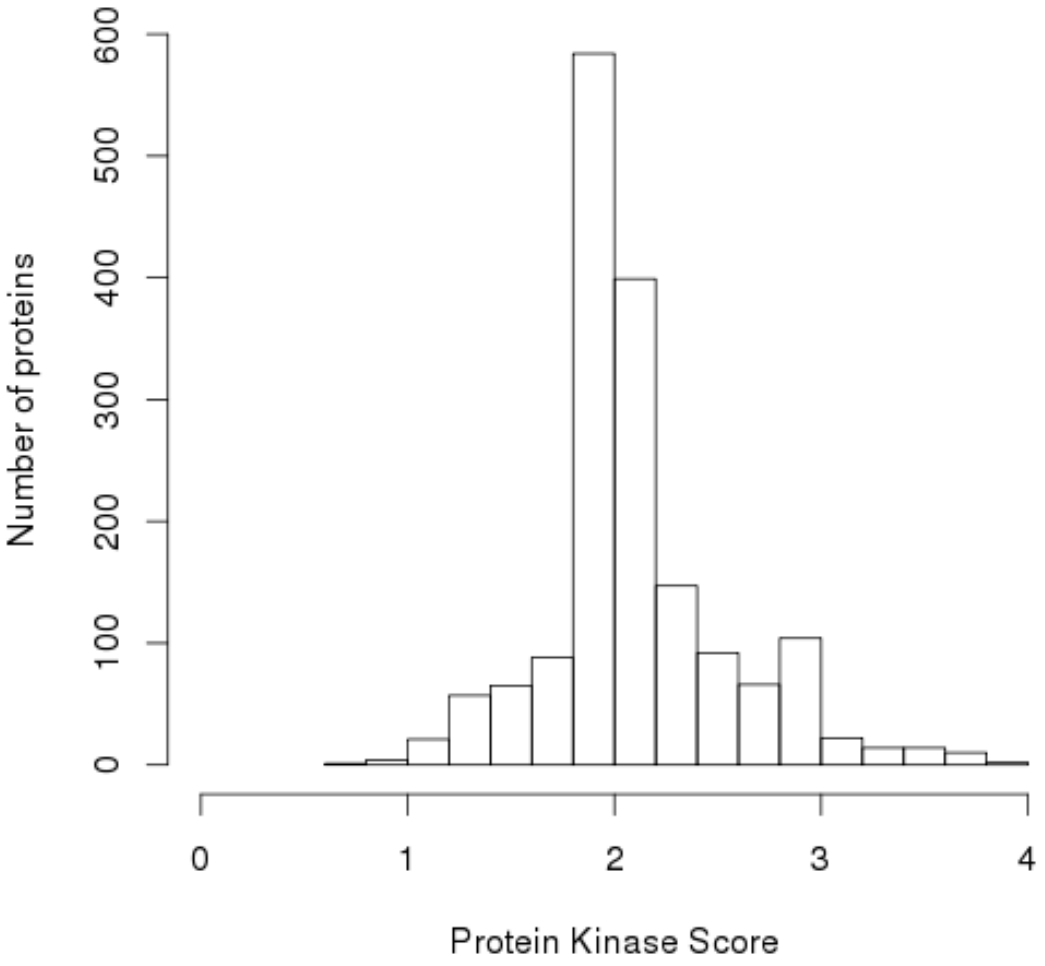

Fig. 2.20.: Histogram showing the distribution of final scores in the set of protein kinases from *C. reinhardtii*

3. FUNCTIONAL CLASSIFICATION AND ANALYSIS OF PROTEIN KINASES FROM *PHYSCOMITRELLA PATENS* AND *CHLAMYDOMONAS REINHARDTII*

3.1 Introduction

Protein kinases played an important role in the evolution of early land plants from an aquatic environment to a terrestrial environment due to their involvement with major stress response and signaling pathways (Zhu, 2000; McDonald and Linde, 2002; Cristina et al., 2010). Therefore, we can study the functional evolution of protein kinases in early plants to comprehend the changes that occured during early plant evolution. Unfortunately, the quality of *ab initio* gene models in non-model plants are questionable due to the reliance on computational gene predictors that do not use the unique characteristics of protein kinases to make gene model predictions (Li et al., 2005).

As discussed in the first chapter, there are no currently available methods to evaluate gene models. We discussed the development of a novel method to score protein kinase gene models in early plants such as *P. patens* and *C. reinhardtii* in the previous chapter. Using the scoring function, we had shortlisted 1422 and 2221 gene models in *C. reinhardtii* and *P. patens* respectively as protein kinases. In order to fully understand the impact the protein kinase family had on the development of early plants, we need to analyze the functions of these newly curated set of protein kinases and compare them with the protein kinases in the reference plants. Apart from performing functional analyses, we also need to estimate the expansion of the protein kinase family by categorizing the newly curated protein kinases from the

early plants, and the set of protein kinases from the reference plants. This can help us understand the manner of elaboration that occurred in the protein family.

3.2 Materials and Methods

3.2.1 Functional analysis using Blast2GO

To identify the functions of the curated set of protein kinases, we used the Blast2GO program (Conesa et al., 2005). Blast2GO is a software suite that enables the functional annotation of proteins using a combination of tools such as BLAST, InterPro, Gene Ontology and KEGG pathway analysis (Altschul et al., 1990; Apweiler et al., 2001; Consortium, 2004; Kanehisa et al., 2006). Thus, it provides a snapshot of the different functions and the functional pathways of a given set of proteins. In order to compare the functional arsenal of protein kinases from *P. patens* and *C. reinhardtii* with those from *A. thaliana* and *O. sativa*, we used Blast2GO to analyze the protein kinase sequences from each set of plants.

3.2.2 Tracking the expansion of protein kinase families

In order to find out the relative expansion and contraction of specific protein kinase families, we used the "hmmsearch" program from the HMMER suite of tools (Finn et al., 2011). Hmmsearch utilizes a set of sequences as the database, and a profile HMM as the query in order to find similarity between them. Profile HMMs for specific protein families in plants were previously published by Lehti Shiu et al (2012) (Lehti-Shiu and Shiu, 2012). The previous study was done to track protein family changes, but used the already available poor quality gene models. Therefore, it was necessary to perform evaluation of the existing gene models and then track the protein kinase family changes. Therefore, these profile HMMs were used as the query for the search, while the set of protein kinases from *P. patens* and *C. reinhardtii*, and *A. thaliana* and *O. sativa* were used as the sequence database respectively.

3.3 Results

3.3.1 Blast2GO results

Protein kinases sequences from early plants, and the reference plants were loaded separately as inputs to Blast2GO. The first step in the functional analysis was to perform a BLASTP search against the TAIR database for each set of protein kinases. Once BLAST results were obtained, an InterPro domain annotation was performed, and GO terms were mapped to each sequence based on the best BLAST hit, and the set of InterPro domains it was annotated with. The final annotation step verifies the GO terms assigned to each sequence by taking the intersection of the set of all annotations, and an enzyme-code mapping is done based on the GO terms.

In both sets of protein kinases, almost all sequences was annotated with at least one GO term (Figure 3.1 and 3.2). A majority of sequences had between 4 and 20 GO annotations. This means that there was sufficient evidence for the functional annotation to be correct, since there are multiple sources of evidence for the same annotation. Looking at specific GO terms, the third level GO terms were extracted from each set of protein kinases and the top 20 biological process, molecular function and cellular component terms were compared between them. Figure 3.3 shows the distribution of the top GO terms for the sequences from *P. patens* and *C. reinhardtii*. Focusing on the biological processes, we find that most GO terms denote different metabolic and cellular processes. Functions involving response to stress, and response to chemical stress were also present, as were responses to different kinds of stimuli. Looking at the same comparison for the protein kinases from *A. thaliana* and *O. sativa*, we found that the distribution of biological processes and functions remains similar, even though the number of sequences annotated with each term increased (Figure 3.4.

Next, we looked at the distribution of all biological process GO terms in each set of protein kinases. For the sequences from the early plants, a total of 28 different biological processes were affected, with processes related to serine/threonine protein kinase activity and phosphorylation having the most number of sequences annotated (Figure 3.5). Similarly, when looking at protein kinases from the reference plants, a total of 37 different biological processes were shown to be affected (Figure 3.6. This means that the number of biological process functions had increased in late land plants when compared to early plants. Interestingly, the protein kinases from the reference plants had more annotations related to stress and defense response processes than the early plants. For instance, terms such as "response to salt stress" and "defense response to bacterium" are completely missing from the functional annotation of early plants.

3.3.2 Hmmsearch results

In order to investigate the changes in specific protein kinase families between the early plants and the reference plants, a hmmsearch was done between each set of protein kinases and the profile HMM for different protein kinase families. The results are tabulated in Table 3.1. In the early plant group, RLK-Pelle kinases were the largest group, with the CMGC family having the second highest number of proteins. RLK-Pelle kinases had the highest number of proteins in the reference group as well, but the CK1 protein kinases had the next highest number of proteins. The CK1 group of protein kinases seemed to have undergone the most expansion, going from 182 to 1039 proteins. CK1 kinases function in DNA repair, transcription factor regulation, and signaling (Eide and Virshup, 2001). Alternately, the CMGC family of protein kinases went down from 553 to 340 in the reference group. Overall, 5 protein kinase families showed expansion in numbers, while 3 families had their number of proteins reduced when going from early to late land plants.

3.4 Discussion

This study was done to investigate the changes in the protein kinase functions and families of well-evolved plants when compared to early plants. Looking at the functional annotations of early plants, we found that the protein kinases from *A. thaliana* and *O. sativa* may have expanded their functional ability when compared to the protein kinases from *P. patens* and *C. reinhardtii*. In other words, protein kinases from the reference plants gained several functions as they completely moved from an aquatic to a terrestrial environment. We also found that while early plants were annotated with certain stress related functions, some of the stress and defense response functions may have evolved at a later evolutionary stage.

We also studied the changes in specific protein kinase families during evolution. While the protein kinase families of AGC, CAMK, CK1, RLK-Pelle and STE kinases expanded, the number of proteins in Aurora, CMGC, and TKL had reduced over time. CAMK and RLK-Pelle kinases have been known to regulate different types of stress responses (Afzal et al., 2008; Sheen, 1996). Therefore, it is possible that these protein kinases had duplicated to combat significant biotic and abiotic stress over the course of evolution.

Table 3.1.: Comparison of the number of protein kinases in each protein kinase family between the early plant group and the reference group

Protein kinase family	**Early plant group**	**Reference group**
AGC	49	120
CAMK	117	294
Aurora	44	9
CK1	182	1039
CMGC	553	340
RLK-Pelle	867	2145
STE	55	142
TKL	205	166

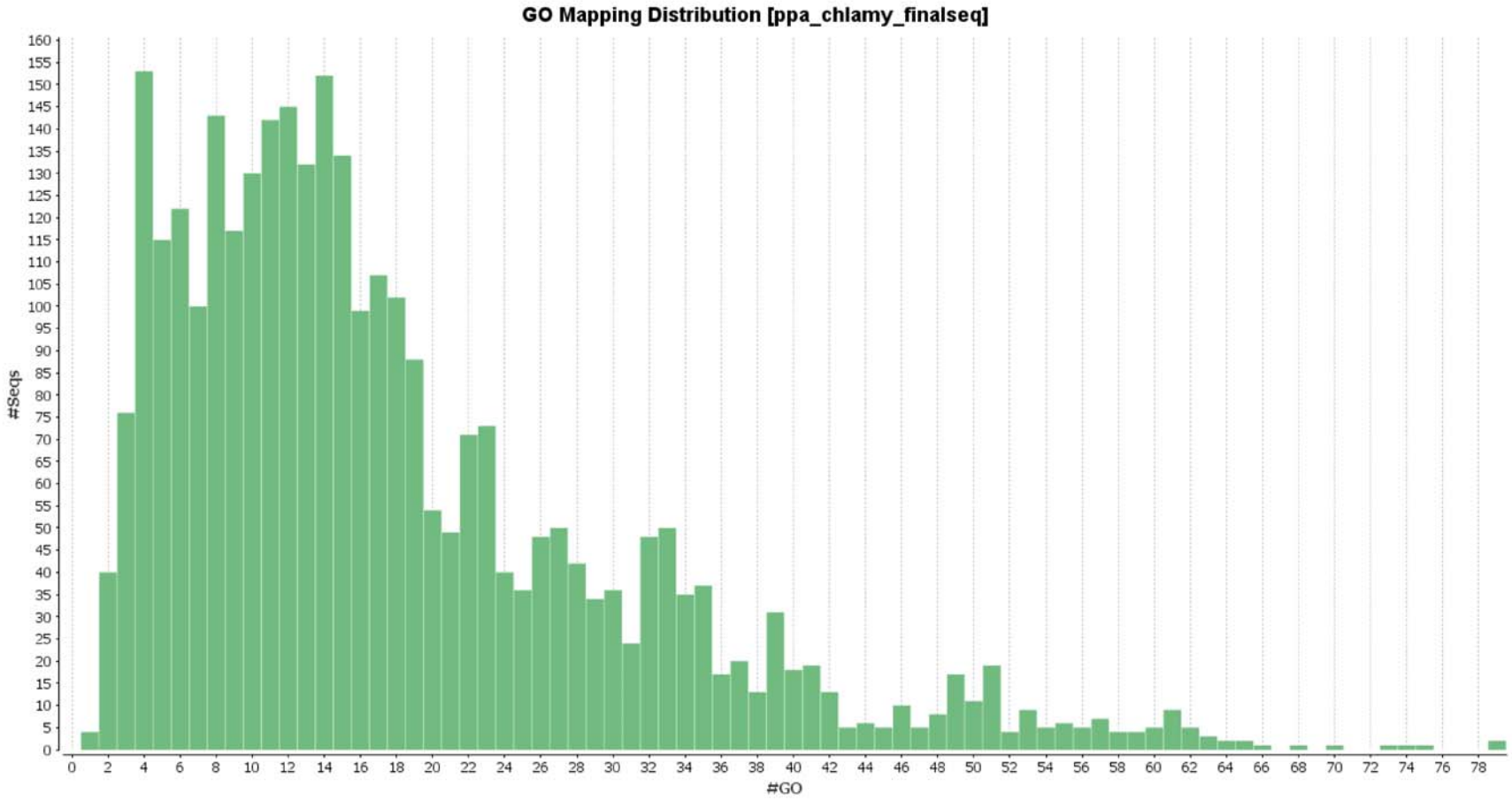

Fig. 3.1.: Distribution of the number of GO annotations for the protein kinase sequences of *P. patens* and *C. reinhardtii*.

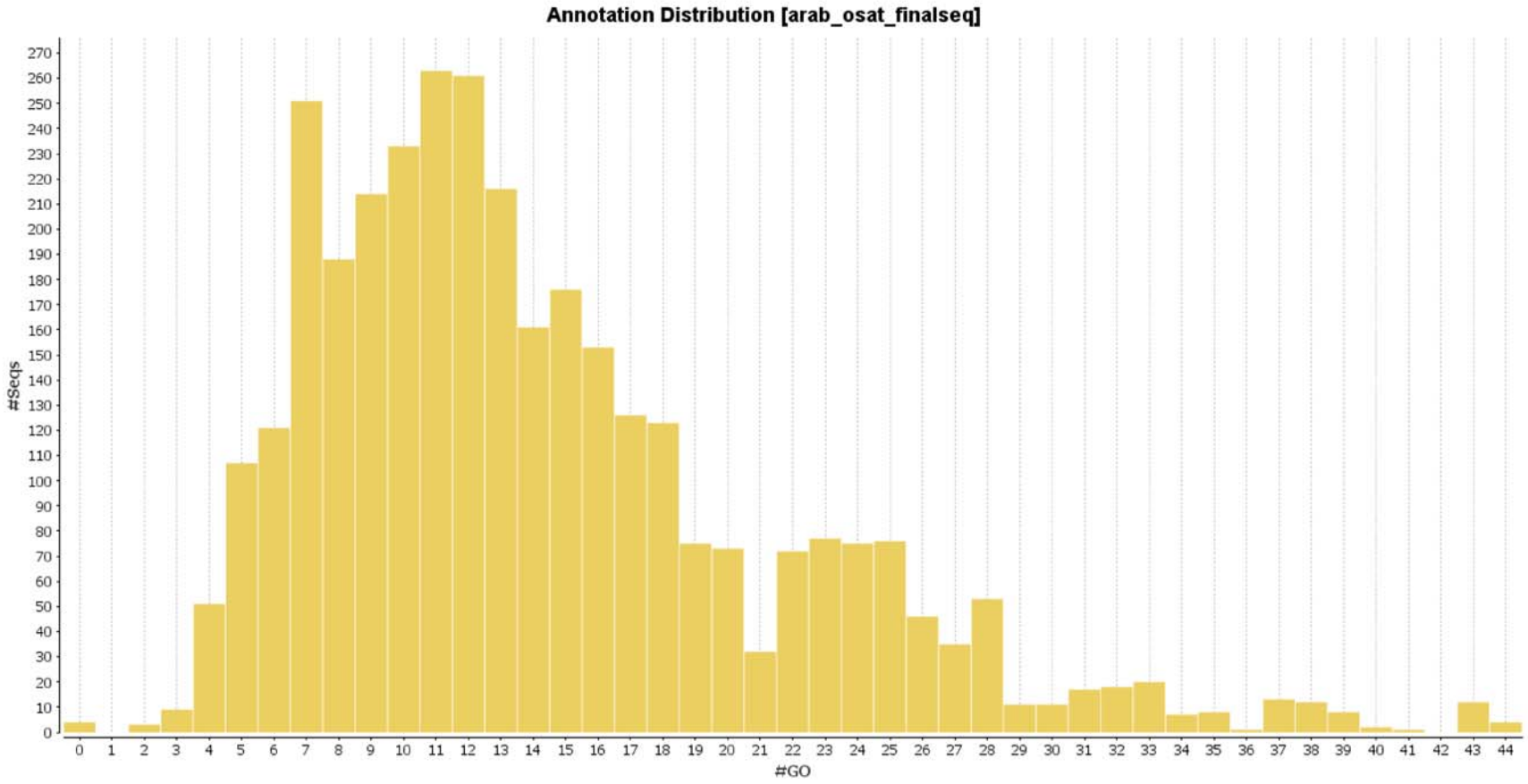

Fig. 3.2.: Distribution of the number of GO annotations for the protein kinase sequences of *A. thaliana* and *O. sativa*.

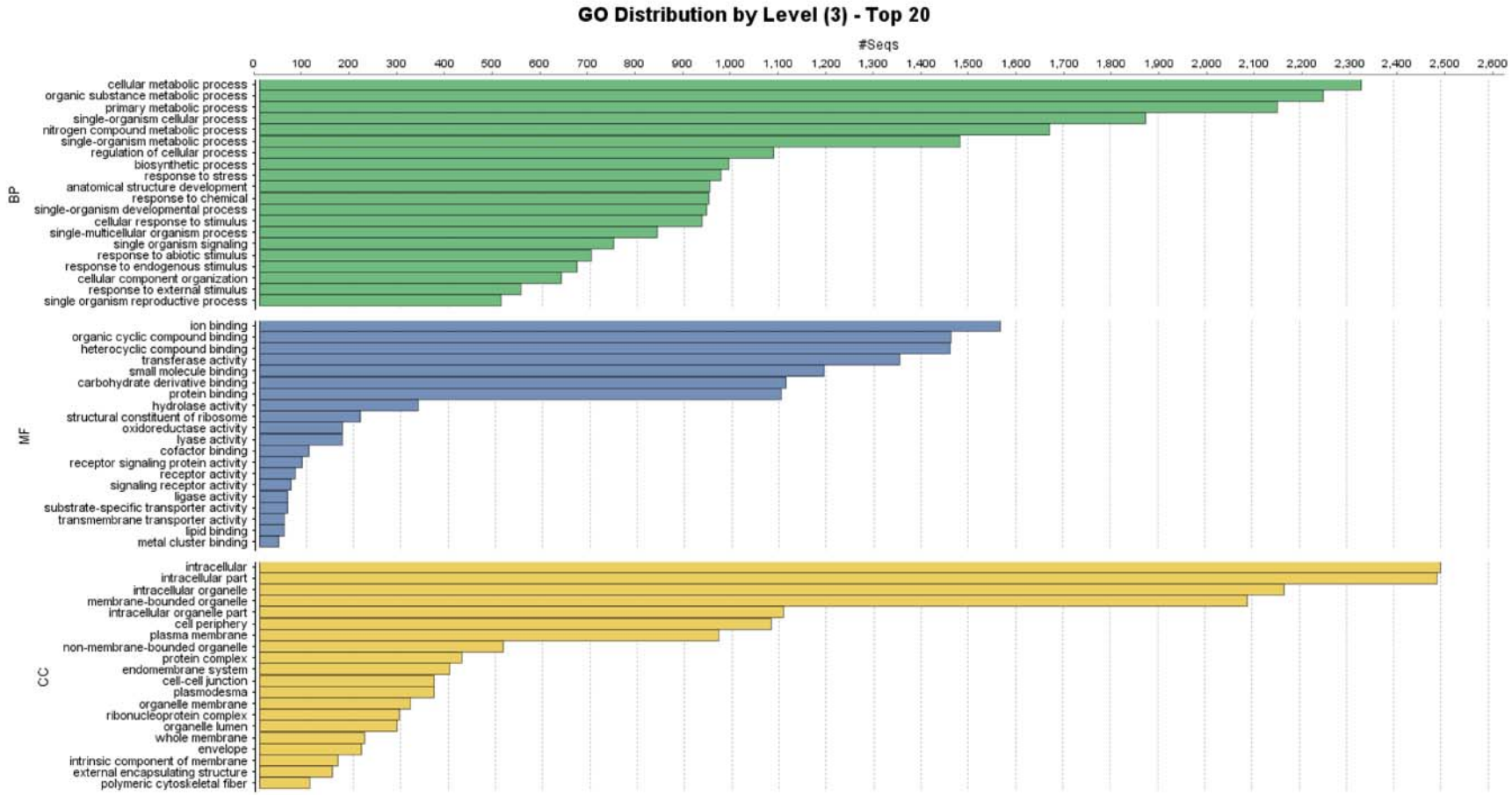

Fig. 3.3.: Top 20 biological process (BP), molecular function (FM) and cellular component (CC) GO term annotations for the protein kinases from *P. patens* and *C. reinhardtii*.

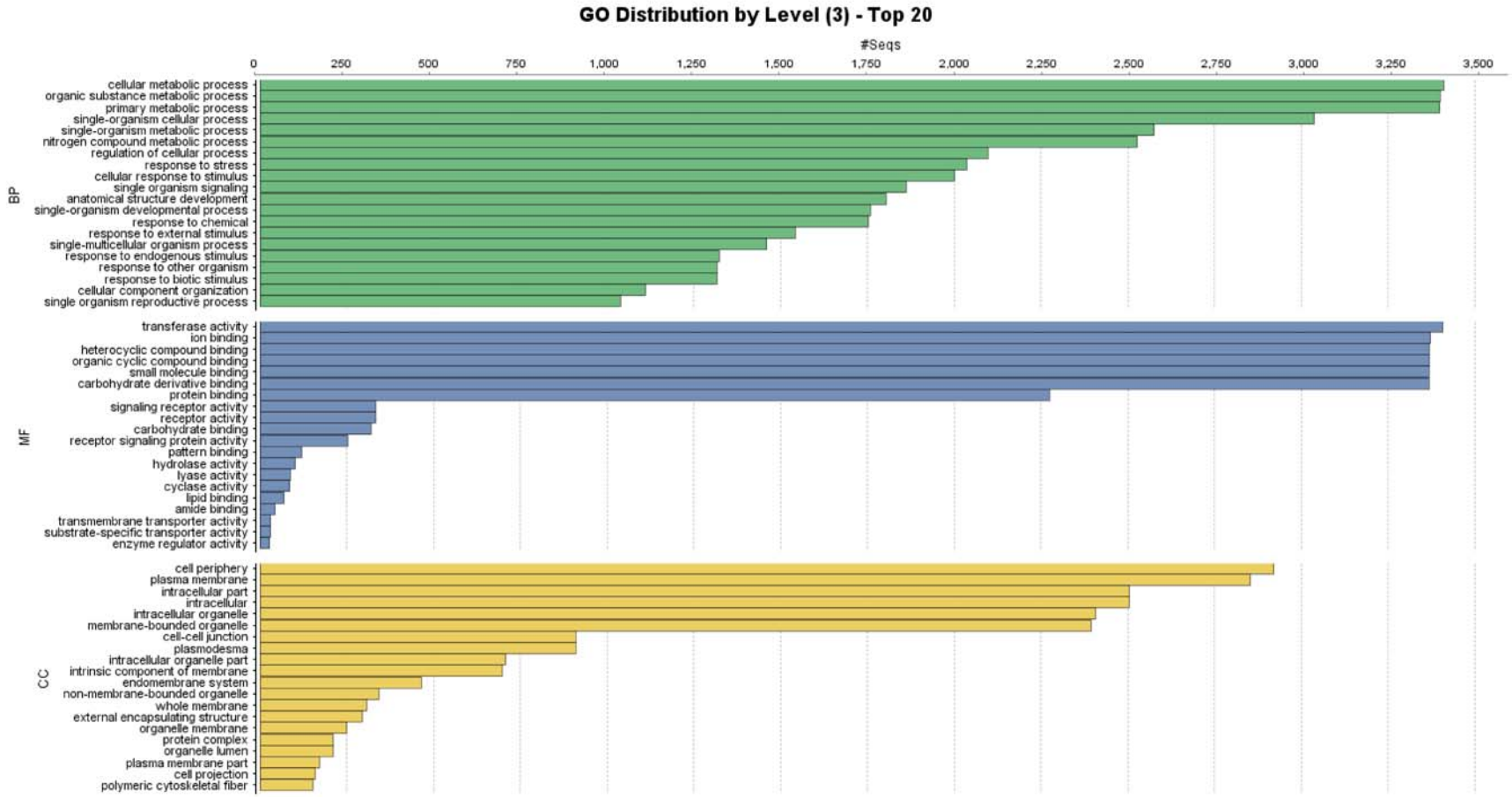

Fig. 3.4.: Top 20 biological process (BP), molecular function (FM) and cellular component (CC) GO term annotations for the protein kinases from *A. thaliana* and *O. sativa*.

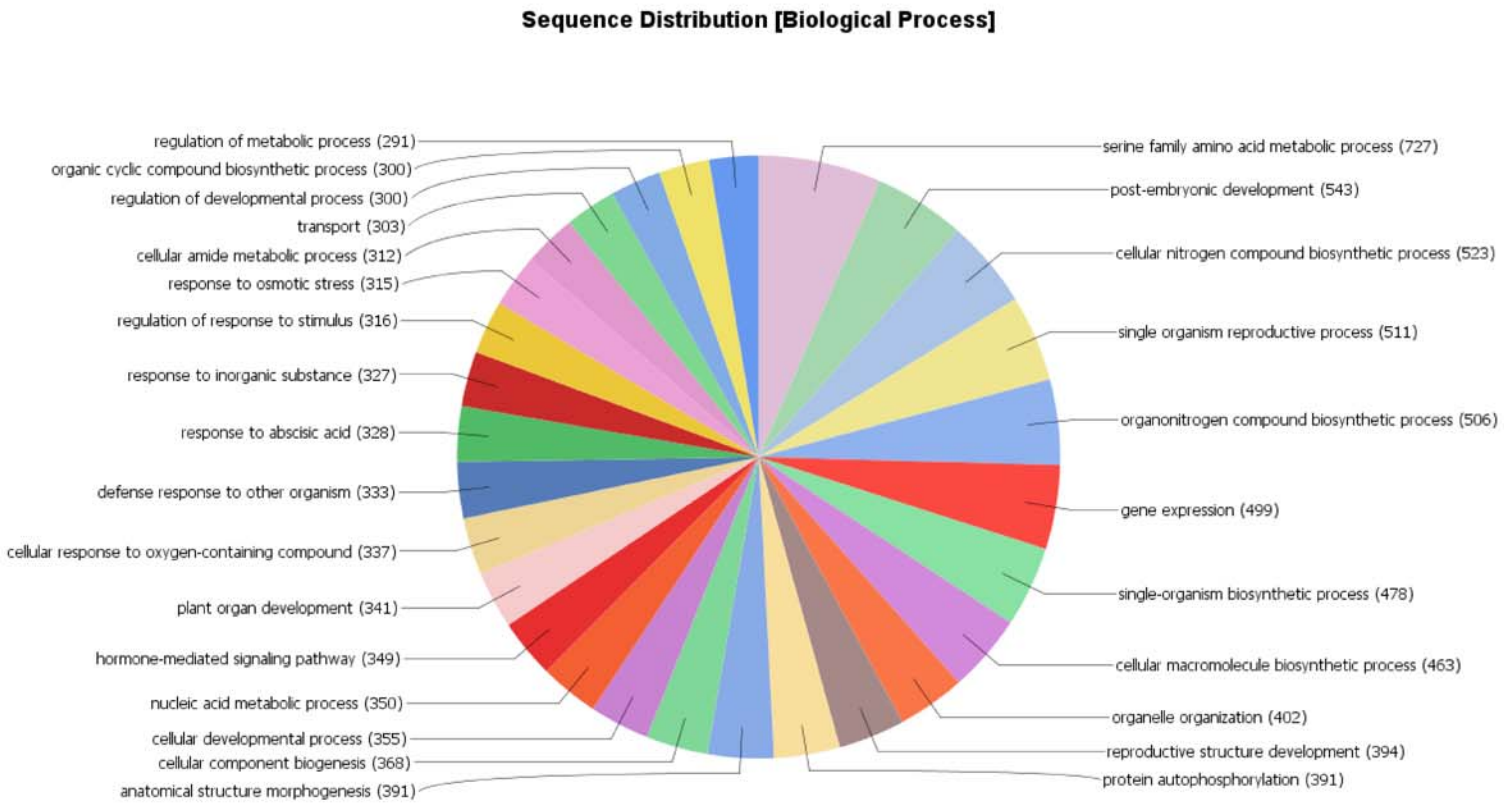

Fig. 3.5.: Total sequence distribution of biological process GO term annotations for the protein kinases from *P. patens* and *C. reinhardtii*.

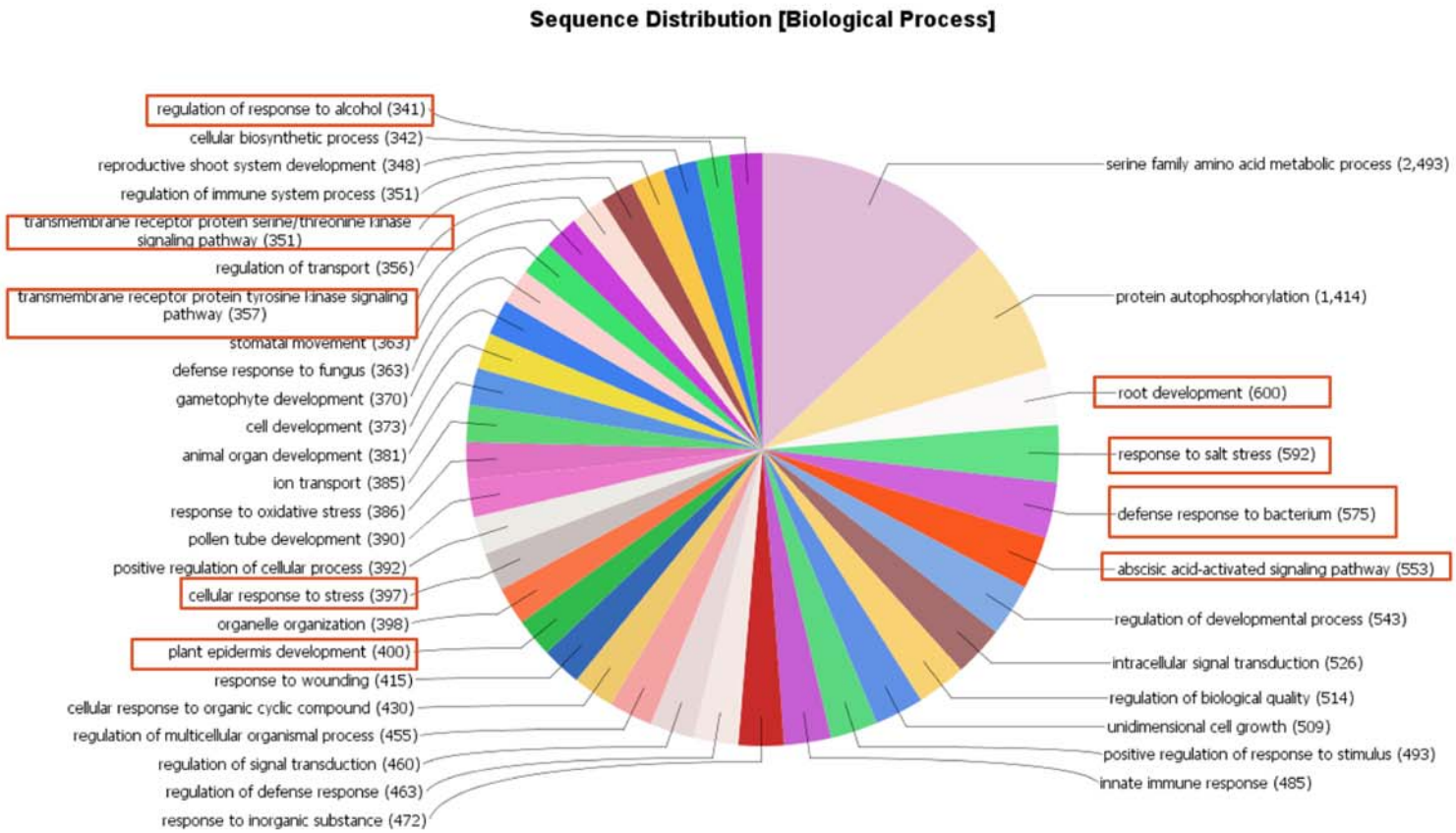

Fig. 3.6.: Total sequence distribution of biological process GO term annotations for the protein kinases from *A. thaliana* and *O. sativa*. Functions that are marked indicate late functional elaboration in the plant evolutionary timeline.

4. *DE NOVO* ASSEMBLY AND ANNOTATION OF THE GIANT RAGWEED (*AMBROSIA TRIFIDA*) TRANSCRIPTOME

4.1 Introduction

Giant ragweed (*Ambrosia trifida*) is one of the most problematic annual weeds in corn and soybean production across the eastern corn belt in the United States, and in some parts of Canada (Abul-Fatih and Bazzaz, 1979; Bassett and Crompton, 1982). It is a member of the Asteraceae family. Other common names of the weed include great ragweed, tall ambrosia and crown-weed wild hemp. It is usually found growing in ditches adjacent to roads, meadows and riverbanks (Abul-Fatih and Bazzaz, 1979). It is very adaptive to a variety of environments, and is resistant to a variety of weed control measures (Baysinger and Sims, 1991). Prior to the introduction of genetically-modified glyphosate-resistant crops, giant ragweed was the most troublesome weed for Midwestern crop varieties (Harrison et al., 2001). Due to the rapid growth cycle of giant ragweed seedlings, it is very competitive with crops and, if left unchecked, can dominate any cropping system.

In order to study the glyphosate resistance mechanism in giant ragweed, it is essential that we study the gene expression differences between the resistant and sensitive plants, and identify the genes responsible for the resistance. The first version of the glyphosate-sensitive (GS) biotype of giant ragweed transcriptome was published in 2012 (Lai et al., 2012). But no gene annotations were provided, making it difficult to identify the key genes involved in glyphosate resistance. This existing transcriptome was determined using older 454 sequencing technology and a substantially lower depth of coverage than is typical of more modern approaches. In this study, the transcrip-

tome of giant ragweed was sequenced using Illumina HiSeq sequencing technology, and annotated using the Trinotate *de novo* sequence assembly pipeline (Haas et al., 2005).

4.2 Materials and Methods

4.2.1 Plant material

Glyphosate-resistant (GR) and GS biotype seeds of giant ragweed were collected from Noble County, Indiana and Darke County, Ohio respectively. Greenhouse dose-response studies originally proposed by Stachler (2008) were used to characterize their resistance and susceptibility (Stachler, 2008). After allowing the seeds to grow in the greenhouse, plants at the five-node growth stage were selected for hebicide treatment.

4.2.2 Herbicide treatment

All glyphosate solutions for plant treatment were prepared using Touchdown HiTech (N-(phosphonomethyl) glycine, in form of the monopotassium salt) (Syngenta Crop Protection, Inc., Greensboro, NC 27419). The herbicide was sprayed at the recommended field rate of 0.7 kg ae ha^{-1}. Due to the absence of surfactant from the formulation, a non-ionic spreader-sticker adjuvant surfactant (NIS), (AttachTM) at 0.25% v/v and 1.0 % w/v Ammonium Sulfate (AMS) was added. Glyphosate was sprayed on the plants using a compressed-air bench top track sprayer equipped with a flat fan 80015E Tee Jet tip (Spraying Systems Co., Wheaton, IL 60189) with a nozzle pressure of 249 kPa delivering a volume of 187 L of spray solution ha^{-1}.

4.2.3 mRNA extraction

Leaf material was harvested from *A. trifida* obtained from Indiana and Ohio, and used for RNA extraction using a protocol modified from Eggermont et al (Eggermont et al., 1996). 2 cm diameter leaf disks from the first fully developed leaf

were punched out, frozen in liquid nitrogen, and total RNA was extracted in a 2 ml test tube. Each time point contained leaf disks from four separate plants. SDS and phenol-chloroform mixture was used for primary extraction. RNA was purified with subsequent chloroform extractions and lithium chloride precipitations. DNA contamination was removed by DNaseI treatment. RNA concentrations were determined with a Nanodrop photometer and the quality assessed with the RNA 6000 nanochip of an Agilent Bioanalyzer. Samples with RIN values (RNA Integrity Number) above 8 were used for library construction. Sequencing libraries were constructed using the Illumina TruSeq RNA library kit with paired-end barcoding. Steps in this procedure include isolation of poly-A containing mRNA and fragmentation to small pieces which were transcribed into first and second strand cDNA and ligated to adapter oligonucleotides and subsequently amplified by PCR.

4.2.4 Sequencing

Sequencing libraries were constructed using the Illumina TruSeq RNA library kit with paired-end barcoding. Steps in this procedure include isolation of poly-A containing mRNA and fragmentation to small pieces, which were transcribed into first and second strand cDNA, ligated to adapter oligonucleotides, and subsequently amplified by PCR. Between 31 x 10^6 and 88 x 10^6 raw reads (101 bases length) were generated via Illumina sequencing from each RNA sample. Sequence data totaling 50 Gbases has been deposited in the NCBI Sequence Read Archive (SRA) database under accession SRX759962.

4.2.5 RNA-Seq assembly and annotation

RNA was assembled from paired-end reads using the Trinity package (version r2012-10-05) (Grabherr et al., 2011). The resulting assembly contained 246,544 predicted transcript sequences derived from 145,713 assemblies (Trinity components). Trinotate (version r2013-02-25) was used to annotate the transcript assembly with

predicted protein functions (Haas et al., 2005). Trinotate is an annotation program that is specifically designed to work with *de novo assembled transcriptomes*. It uses a combination of methods for functional annotation, such as NCBI-BLAST, HMMER, Pfam, eggNog, TMHMM, signalP and the Gene Ontology database (Altschul et al., 1990; Finn et al., 2011; Bateman et al., 2000; Jensen et al., 2008; Sonnhammer et al., 1998; Petersen et al., 2011; Consortium, 2004). The completeness of the transcriptome was evaluated using CEGMA (Core Eukaryotic Genes Mapping Approach) and BUSCO (Benchmarking Universal Single-Copy Orthologs) (Parra et al., 2007; Simão et al., 2015).

4.2.6 Transcriptome Quality Improvement

Since the RNA-seq data obtained was a part of a single-replicate study, we wanted to find out if using data from other sources could improve the overall quality of the transcriptome. First, we used MIRA (Mimicking Intelligent Read Assembly) to combine 454 RNA-Seq data from a previously published giant ragweed transcriptome (Chevreux et al., 1999). The second study was to find if the genome of a related plant can be used to extend the transcriptome sequence length. For this, we used a program called PASA (Program to Assemble Spliced Alignments) to make the sunflower genome that was recently published as the reference genome for sequence assembly (Haas et al., 2003). For each case, we then analyzed the resulting hybrid transcriptome assembly and used Trinotate for annotation.

4.3 Results

Since giant ragweed does not have a published genome, a *de novo* transcriptome assembly was performed. RNA-seq assembly was done using Trinity, there were a total of 246,544 predicted Trinity isoforms. As mentioned in the previous section, Trinotate was used to annotate the *de novo* assembly. Since Trinotate uses BLAST as one of the methods for annotation, we can estimate the number of predicted transcripts based on

the number of Trinity isoforms annotated with at least one functional hit. Based on the number of transcripts annotated by Trinotate using BLAST comparisons, giant ragweed has slightly more than 54,500 predicted transcripts, which is more than *Arabidopsis thaliana* and less than *Oryza sativa* (Table 4.1) (Altschul et al., 1990). An E-value cutoff of $1x10^{-20}$ was used as a similarity cut-off in the BLASTP searches.

4.3.1 Transcriptome completeness

To evaluate the completeness of the transcriptome, Core Eukaryotic Genes Mapping Approach (CEGMA) analysis was first performed (Parra et al., 2007). The CEGMA gene set consists of approximately 450 proteins that are highly conserved and found universally in most eukaryotes, and can therefore, be used to gauge how complete the transcriptome is. The annotated transcriptome of *A. trifida* was compared against a set of core eukaryotic genes, and it was found that 97% (241 out of 248) of the core genes were present and complete, and 100% (248 out of 248) were present and partially represented. We also quantified the completeness of the transcriptome using a similar analysis pipeline called BUSCO, which assesses the quality of the assembly based on gene content from single-copy orthologs from OrthoDB, a database of eukaryotic orthologs (Simão et al., 2015; Kriventseva et al., 2008) . When compared to a plant lineage dataset of core genes, the giant ragweed transcriptome was estimated to be 94% complete. These results suggests that the transcriptome is relatively complete.

4.3.2 eggNog annotation

eggNog is a database of functionally annotated orthologous genes, similar to Clusters of Orthologous Groups (COG) and EuKaryotic Orthologous Groups (KOG) (Jensen et al., 2008; Koonin et al., 2004; Tatusov et al., 2003). eggNog annotations provide a snapshot of the representation of the protein functional categories in the transcriptome. Based on the eggNog annotations of the giant ragweed transcrip-

tome, we find that the most common annotated function is that of serine/threonine protein kinase, followed by leucine-rich repeat protein, and WD-40 repeat protein (Table 4.2). The result is along expected lines since the serine/threonine protein kinases are one of the largest groups of proteins in plants. The cytochrome P450 family of proteins, which is also a large protein family, is also among the top five most annotated functions.

4.3.3 Gene Ontology annotation

Trinotate incorporates Gene Ontology (GO) annotations into the results, allowing comparisons with protein function results obtained by eggNog, and analysis of predicted cellular localizations and biological processes of the proteins in the predicted proteome (Grabherr et al., 2011; Consortium, 2004). 56,345 predicted transcript isoforms were annotated with at least one GO term. GO terms are hierarchical in nature; the parent terms are generalized, while the child terms are more specialized in nature. GO terms at the third hierarchical level were thus extracted from the hierarchy of annotations predicted for the transcriptome. In total, 90,612 cellular component, 121,057 biological process, and 108,272 molecular function annotations were assigned. Figures 4.1, 4.2 and 4.3 show the top 25 GO terms each for molecular function, biological process and cellular component respectively.

4.3.4 Pfam annotation

Approximately 3500 predicted Trinity transcripts had Pfam domain annotations in the Trinotate results. The domain with the highest number of hits was the Protein kinase domain, followed by the Chlorophyll A-B binding protein and the Tyrosine kinase domain respectively (Table 4.3). Considering that plants have only a few known tyrosine kinases, it was suprising that so many transcripts were annotated with the tyrosine kinase domain. However, there is a possibility that the domain

annotations indicate the presence of a large number of Tyrosine kinase-like proteins, which are known to be present in plants, and are part of a diverse family of proteins.

4.3.5 TMHMM predictions

Transmembrane helices are a part of the structure of membrane proteins. Approximately one-third of all currently mapped gene sequences in the Protein Data Bank (PDB) are known to encode membrane proteins (Hildebrand et al., 2004). They typically function as transporters for various specific molecules across the biological membrane. Trinotate results include the prediction of TMHMM which indicates the presence of transmembrane helices in the translation products of the predicted Trinity transcripts. We found that 8211 Trinity transcripts had a TMHMM prediction, with a protein length of 51.66 amino acids and 2.308 helices per protein on average. The number of helices ranged from 1 to 16, and the predicted protein length varied between 11.14 to 352.56 amino acids respectively.

4.3.6 Improving the transcriptome using long read sequence data

We investigated whether the previously published transcriptome of giant ragweed, which was based on the 454 sequencing platform, can be used to improve the quality of the short read sequence data (Lai et al., 2012). Transcriptome data sequenced using 454 sequence technology was obtained from The Compositae Genome Project, and the MIRA (Mimicking Intelligent Read Assembly) program was used to combine the data with the Illumina giant ragweed data (Chevreux et al., 1999). While the original Trinity assembly contained 249,598 predicted transcript isoforms, the combined 454-Illumina transcriptome contained 142,395 transcripts. This dataset was then annotated using Trinotate, and a total of 54,596 annotations were found. 48,270 transcript isoforms were annotated with at least one GO term, and a total of 102,223 cellular component, 110,806 biological process and 87,667 molecular function annotations were determined. In eggNog annotations, the top annotated function remained

Serine/threonine protein kinase with 2303 transcripts. The reduction in the number of predicted genes together with the increased level of annotation suggests that combining these datasets slightly improves the transcriptome assembly. The relatively small change again suggests that the Trinity assembly is fairly complete.

4.3.7 Comparison with the sunflower transcriptome

Since the complete genome of giant ragweed is not available, we calculated the coverage of the giant ragweed transcriptome versus the transcriptome of *Helianthus annuus* (sunflower), a close relative (Gill et al., 2014). We used a Perl script to estimate the percentage of coverage relative to the sunflower transcriptome, and found that close to 26% of the 246,544 predicted Giant Ragweed transcripts had matches in the sunflower transcriptome, based on BLAST comparisons using a conservative E-value cutoff of $1x10^{-20}$. This shows that the giant ragweed transcriptome sequences have coverage greater than $1x10^{-20}$. On the other hand, around 77% of the sunflower transcriptome sequences had coverage greater than an E-value of $1x10^{-20}$ in the giant ragweed transcriptome. Considering the divergence between giant ragweed and sunflower, this suggests that the ragweed transcriptome reported here is nearly complete.

4.3.8 Improving the transcriptome using sunflower as the reference genome

We wanted to find out whether the recently published sunflower genome could be used to improve the quality of the giant ragweed transcriptome (Gill et al., 2014). To test this, a software pipeline called PASA (Program to Assemble Spliced Alignments) was used (Haas et al., 2003). The sunflower genome was used as the reference for the giant ragweed transcriptome data. However, the transcriptome quality was not appreciably improved, and the number of transcripts remained relatively high. This was probably due to the fact that the sunflower genome is only about 80% similar to giant ragweed in sequence similarity.

4.4 Discussion

The genomics of giant ragweed has been of great interest recently due to the increase in prevalence of herbicide resistance. The present work seeks to publicize the availability of the annotated transcriptome of giant ragweed, The transcriptome of giant ragweed was assembled using the Trinity pipeline, and subsequently annotated using Trinotate. This would help a great deal in identifying the source of glyphosate resistance. The development of tools such as Trinotate could lead to a deluge in annotations of genome and transcriptome sequences of non-model organisms. The transcriptome was annotated using BLAST, Gene Ontology, and eggNog identifiers. We tried to improve the quality of the transcriptome using the recently published sunflower genome and transcriptome sequences. Even though the use of the sunflower genome as reference did not lead to the reduction in the number of Trinity transcripts, the use of transcriptome sequences lead to a notable reduction. Finally, we attempted to use the previously published long-read transcriptome sequences of giant ragweed to improve the transcriptome quality. This lead to only a slight reduction in the number of transcripts from which we can infer that the transcriptome sequence we have is fairly complete.

Table 4.1.: The number of annotated genes in different species. The data for *A. thaliana*, *O. sativa* and *Z. mays* were obtained from PlantGDB (Duvick et al., 2008).

Species	**No. of gene annotations**
Ambrosia trifida	54,596
Arabidopsis thaliana	37,761
Oryza sativa	68,464
Zea mays	136,522

Table 4.2.: Top 25 functional annotations of the giant ragweed transcriptome using eggNog

COG/NOG	**No. of transcripts**	**Functional annotation**
COG0515	4470	Serine/threonine protein kinase
COG4886	804	Leucine-rich repeat (LRR) protein
COG2319	680	FOG: WD40 repeat
COG0666	340	FOG: Ankyrin repeat
NOG12793	278	Calcium ion binding protein
COG2124	239	Cytochrome P450
COG0724	235	RNA-binding proteins (RRM domain)
COG0513	233	Superfamily II DNA and RNA helicases
COG0699	213	Predicted GTPases (dynamin-related)
COG0631	213	Serine/threonine protein phosphatase
COG0457	201	FOG: TPR repeat
COG0477	198	Permeases of the major facilitator superfamily
COG1028	189	Dehydrogenases with different specificities
NOG318082	188	Transposable element
COG0474	182	Cation transport ATPase
COG1100	176	GTPase SAR1 and related small G proteins
NOG237917	172	Protein involved in lipid transport
COG2939	167	Carboxypeptidase C (cathepsin A)
NOG251664	149	Delta-Like 3 (Drosophila) protein
COG0484	138	DnaJ-class molecular chaperone with C-terminal Zn finger
COG0596	136	Predicted hydrolases or acyltransferases (alpha/beta hydrolase)
NOG280712	125	Disease resistance protein
COG0154	125	Asp-tRNAAsn/Glu-tRNA Fln amidotransferase A subunit
COG2940	119	Proteins containing SET domain

Table 4.3.: Top 25 Pfam domain annotations of the predicted proteins in the giant ragweed transcriptome

PFAM Domain ID	**Function**	**Frequency**
PF00069.20	Protein kinase domain	865
PF00504.16	Chlorophyll A-B binding protein	417
PF07714.12	Tyrosine kinase	406
PF00400.27	WD40 repeat	374
PF00101.15	Ribulose bisphosphate carboxylase, small chain	295
PF00067.17	Cytochrome P450	290
PF00076.17	RNA recognition motif	286
PF12338.3	Ribulose-1,5-bisphosphate carboxylase small subunit	256
PF13504.1	Leucine rich repeat	222
PF01946.12	Thi4 family	199
PF01535.15	Pentatricopeptide repeat	178
PF00481.16	Protein phosphatase 2C	163
PF00646.28	F-box domain	133
PF00106.20	short chain dehydrogenase	129
PF00226.26	DnaJ domain	128
PF00249.26	Myb-like DNA-binding domain	128
PF00270.24	DEAD/DEAH box helicase	125
PF00153.22	Mitochondrial carrier protein	119
PF00847.15	AP2 domain	111
PF00005.22	ABC transporter	109
PF00025.16	ADP-ribosylation factor family	109
PF00501.23	AMP-binding enzyme	106
PF00004.24	ATPase family associated with various cellular activities (AAA)	105
PF00149.23	Calcineurin-like phosphoesterase	105
PF08263.7	Leucine rich repeat N-terminal domain	101

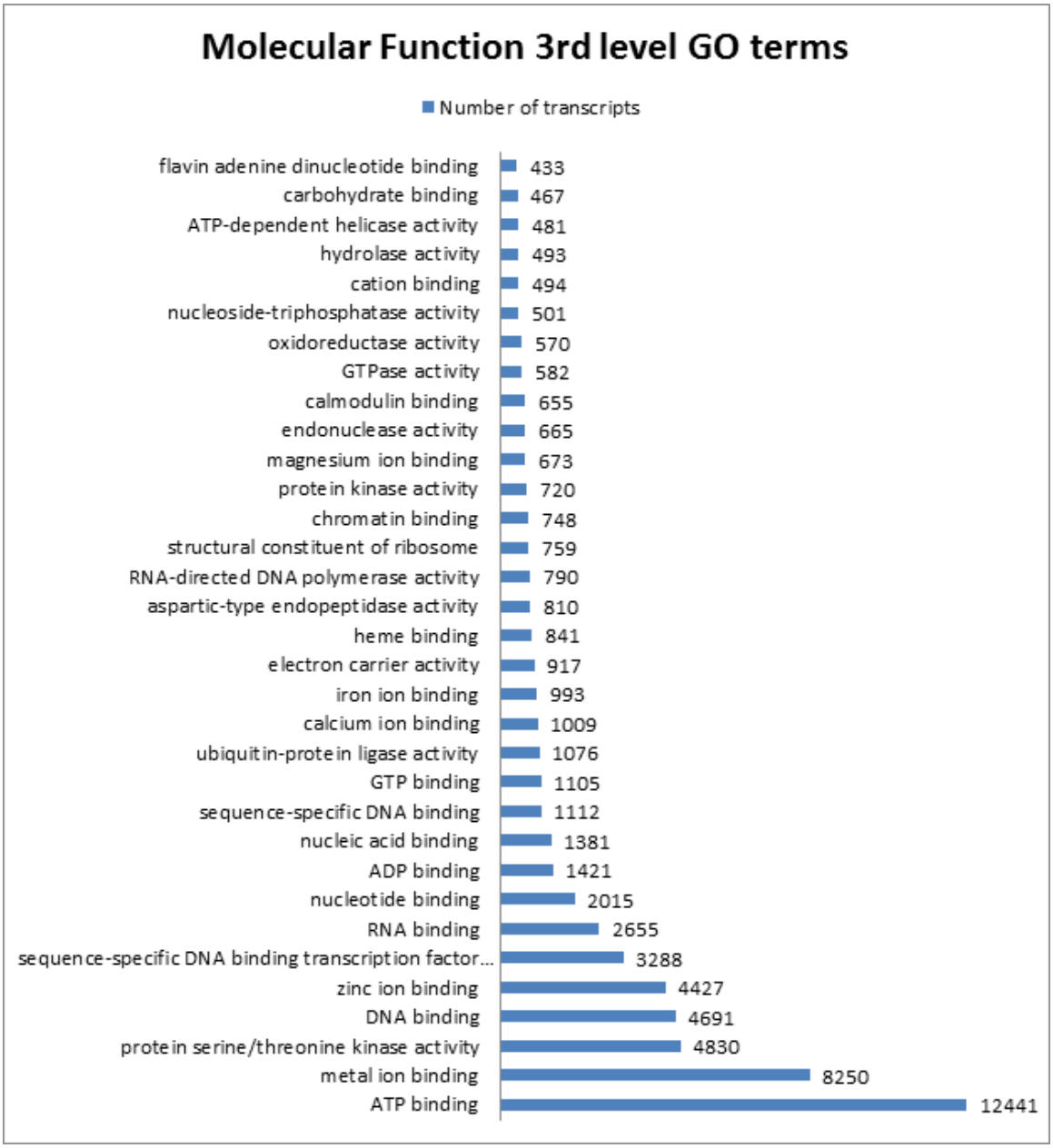

Fig. 4.1.: Top 25 Molecular Function third-level annotations found using Gene Ontology

Fig. 4.2.: Top 25 Biological Process third-level annotations found using Gene Ontology

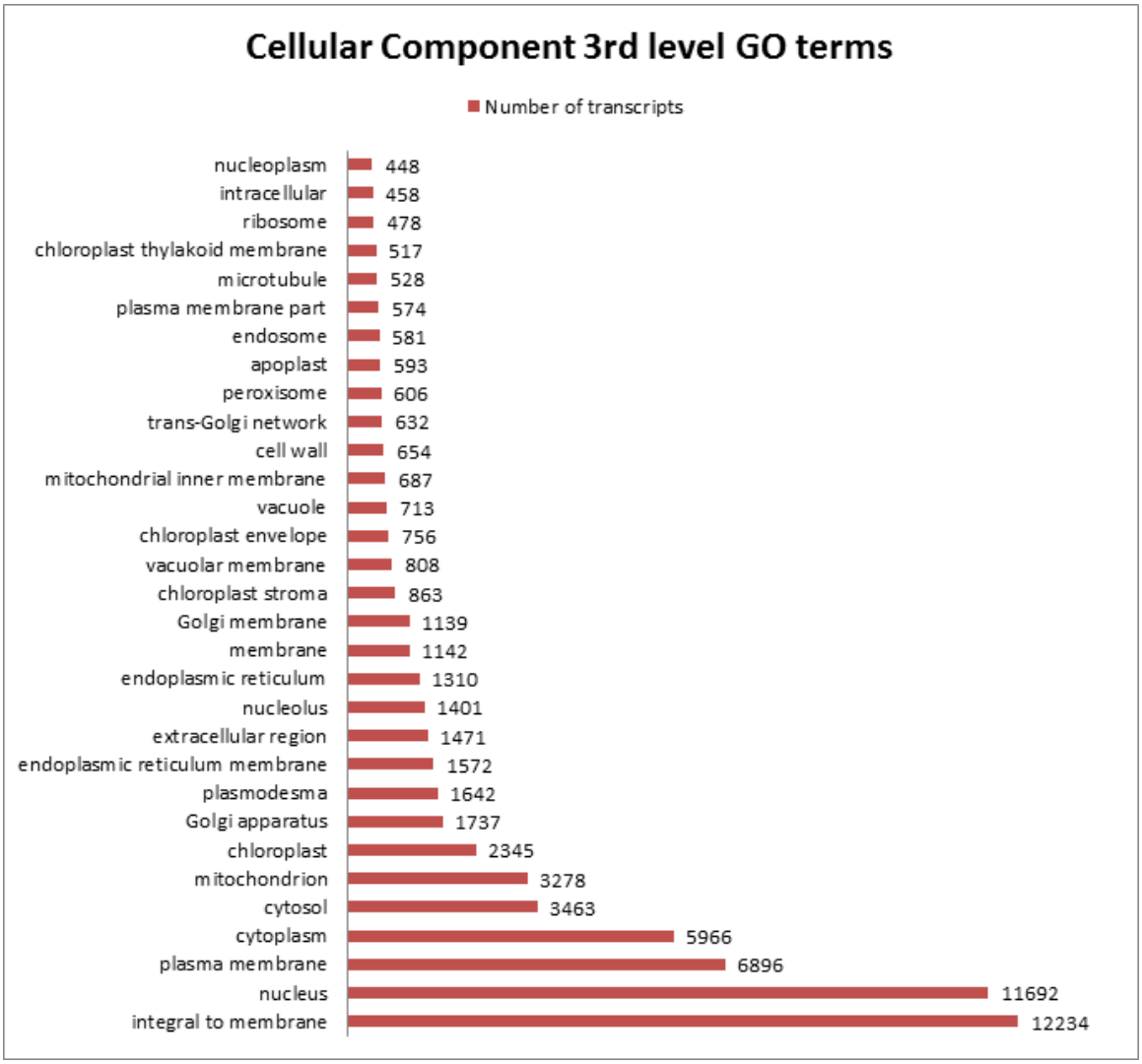

Fig. 4.3.: Top 25 Cellular Component third-level annotations found using Gene Ontology

5. INVESTIGATION OF THE MECHANISM OF RESISTANCE TO GLYPHOSATE IN GIANT RAGWEED (*AMBROSIA TRIFIDA*)

5.1 Introduction

Resistance to herbicides, especially glyphosate, in weeds has been a major issue across the world recently. In the past decade, there has been a rise in reports of glyphosate-resistant weeds across 17 countries, including Brazil, Canada, Australia and the United States (Heap, 1997). Due to the use of glyphosate-resistant cropping systems for over two decades, and overuse of the herbicide, there has been a strong selective pressure for giant ragweed to develop resistance to glyphosate (Duke and Powles, 2008; Nandula, 2010). Glyphosate-resistant giant ragweed is a huge problem for farmers since it results in the failure of glyphosate-ready cropping systems, thus leading to huge yield losses (Foresman and Glasgow, 2008). One way to gain insight into resistance mechanisms and the adaptation of giant ragweed to the presence of glyphosate, is to identify genes whose expression differs between glyphosate sensitive and resistant biotypes.

There are no significant phenotypic differences between the glyphosate-sensitive (GS) and glyphosate-resistant (GR) biotypes of giant ragweed prior to herbicide treatment. But when sprayed with glyphosate, certain varieties of giant ragweed plants resistant to glyphosate exhibit a hypersensitive response, with rapid necrosis of the mature leaves of the plant within the first 12 hours of treatment (Figure 5.1) (Segobye, 2013). GR plants thus had a unique response when treated with glyphosate, and resumed normal growth from axillary meristems and started to reproduce. The progression of the response and symptoms resemble a typical hypersensitive response similar to

that observed on some plants after pathogen attack. GS plants do not exhibit rapid leaf necrosis but their leaves become chlorotic, then necrotic, and eventually, the plants die over a two week period.

As mentioned in the first chapter, glyphosate is an inhibitor of 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) which is an important enzyme in the shikimic acid pathway involved in biosynthesis of aromatic amino acids. When sprayed on leaves, the herbicide is absorbed and transported throughout the plant by both passive and active transport (Hetherington et al., 1999). The competitive inhibition of the EPSPS enzyme leads to shikimic acid accumulation and disruption of the production of the aromatic amino acids tryptophan, phenylalanine and tyrosine (de María et al., 2006).

The mechanism of resistance to glyphosate in other common weeds such as Malaysian goosegrass, Italian ryegrass, and Rigid ryegrass have been identified (Gomes et al., 2014; Jasieniuk et al., 2008; Preston et al., 2009). However, the glyphosate resistance mechanism in giant ragweed is yet to be determined. In this study, we compared the gene expression differences between the resistant and sensitive plants using a time course experiment and identified sets of genes that could be involved in glyphosate resistance. We also investigate the presence of single nucleotide polymorphisms (SNPs) in the EPSPS gene in order to verify target-site mutation as a possible mechanism of resistance.

5.2 Materials and Methods

5.2.1 RNA-Seq and assembly

mRNA extraction, sequencing and assembly was done using the procedure described in Chapter 4. Trinity was used for *de novo* transcriptome assembly, and the resulting transcripts were annotated using Trinotate.

5.2.2 Transcriptome analysis

After the RNA-seq reads were mapped to the assembly sequences, we estimated the counts per million transcripts (CPM) value using RSEM (RNA-Seq by Estimation Maximization) (version 1.2.9) (Li and Dewey, 2011). Since we observed clear systemic changes in gene expression even at the first time point, we used a set of genes previously published in rice analyses as controls to normalize the expression values (Jain, 2009). A list of rice genes with stable expression levels over many conditions were identified by Jain (2009). The set of 25 genes given in the paper were used as queries in a TBLASTN search against the assembled ragweed sequences. 21 of the initial 25 had matches in the ragweed transcriptome. The expression levels (CPM) for these genes in both the resistant and sensitive varieties were compared across all four time points. Only 12 genes showed a relatively stable CPM value, while the rest of the genes varied excessively across the time points (more than 1.5 fold up or down). These 12 genes were considered for the normalization (Table 5.1).

To normalize the expression levels, a scale factor was determined for each standard gene with respect to the time zero point (scale = CPM_t/CPM_0). An average scale factor for each time point was then calculated as the simple average of the scale factors for each of the standard genes at each time point. By definition, the scale factor for the zero time point is one, corrected CPM for all genes were then calculated by multiplying the raw expression level by the scale factor for the respective time point. Gene level counts that were less than 1 CPM in all time points were excluded from further analysis. Expression ratios were then calculated for each assembly, comparing the expression levels in the glyphosate resistant and sensitive strains at each time point. Numbers larger than 1 therefore reflect genes (assemblies) with higher expression in the resistant variety. All values were adjusted by the addition of a pseudo count of 0.5 CPM before calculating expression ratios. Assemblies with expression ratios greater than 4, or less than -4 were further examined in the study.

5.2.3 SNP analysis

Single nucleotide polymorphisms (SNPs) are variations in genetic sequences between different individuals, where each variation specifically occurs at a particular position in the genome. Presence of SNPs between two populations could explain differences in disease resistance. In order to determine if mutation in the EPSPS gene could be a possible mechanism of resistance, we used the annotations of the transcriptome done in the previous chapter to identify copies of the EPSPS gene, and calculated the number of SNPs present in the genes in both the resistant and sensitive biotypes. We used samtools mpileup for variant calling and to estimate the SNPs in each copy of the EPSPS gene.

5.3 Results

5.3.1 Transcriptome analysis

A preliminary time course study of the transcriptome level response of resistant and sensitive biotypes of giant ragweed to glyphosate treatment was performed. As mentioned previously, mRNA was extracted from each biotype and sequenced, using the Illumina TruSeq technology, for four time points pre-treatment (0 hour), and 3 hours, 8 hours, and 12 hours, after treatment with glyphosate. RSEM was used to estimate the number of read counts per million transcripts (CPM), and control genes identified from rice were used for the normalization. After normalization, genes that were differentially expressed between resistant and sensitive plants were identified and compared across the four time points.

Genes that were differentially expressed between resistant and sensitive plants were identified and compared across the four time points. Looking at the results, two striking observations can be made.

1. There is a difference in gene expression patterns between the resistant and sensitive plants even before the plants were sprayed with herbicide (Table 5.2).

2. The response to glyphosate is very rapid, and a large number of genes were significantly up or down-regulated within the first three hours (Table 5.3).

The top differentially expressed transcripts in resistant and sensitive plants before treatment are shown in Tables 5.4 and 5.5 respectively. The genes with at least a two-fold change in expression level were identified in resistant and sensitive plants, and subjected to pathway analysis using agriGO to identify pathways with significantly over-represented genes (cutoff $P < 1e^{-7}$) (Consortium, 2004; Du et al., 2010). Pathways with terms such as response to other organisms and lipid biosynthetic process both of which are known to be related to pathogen response were the most significantly over-represented (Figure 5.2). Contrastingly, pathways that are over-represented in the sensitive biotype are annotated with terms like response to stress, response to oxidative stimulus and lignin biosynthesis, which are known stress response indicators (Figure 5.3) [24]. The most significant GO terms for GR and GS giant ragweed at the 0 hour time point are tabulated in Tables 5.6 and 5.7 respectively. This leads us to speculate that, not only do resistant giant ragweed plants react to glyphosate treatment in a manner resembling pathogen defense reactions, but they are already primed by alterations in stress response processes to hyper-react. This is consistent with the rapid necrosis reaction observed in resistant giant ragweed biotypes used in this study.

5.3.2 EPSPS gene expression comparison

To test the hypothesis if over-expression of the EPSPS gene could be a possible mechanism of resistance, we compared the gene expression of the EPSPS gene between GR and GS giant ragweed. Based on the Trinotate annotations obtained, we identified two copies of the EPSPS gene in the giant ragweed transcriptome - comp144227_c0_-

seq1 and comp163996_c0_seq1 (Table 5.8). Using the normalized gene expression results, we compared the expression of each gene copy across the four time points for both the resistant and sensitive biotype (Table 5.9). The basal level of expression in the first copy is higher than the second in both GR and GS giant ragweed. While the expression of the first copy shows a marginal increase in the GR biotype during the first three hours post-treatment, it increases rapidly at the later time points. Contrastingly, the expression of the same gene copy in the GS giant ragweed shows a dramatic decrease in the first three hours, and maintains the low level of expression at the 8 and 12 hour time points. The second EPSPS gene copy on the other hand shows minimal change in gene expression across time points in both GR and GS giant ragweed.

5.3.3 SNP analysis

In order to test if target site mutation could be a possible mechanism of resistance to glyphosate, we performed SNP analysis on the EPSPS gene. We then used samtools mpileup to compare the SNPs in the EPSPS gene copies between the resistant and sensitive plants. We found 29 SNPs in the first copy of the EPSPS gene (Table 5.10), and 17 SNPs in the second copy (Table 5.11). In the first copy, all SNPs were in the GS biotype of giant ragweed, and only 1 of the 29 was found to alter the amino acid sequence. Upon further inspection., it was discovered that the amino acid change occurs before the first Met residue. Therefore, there is little possibility that the amino acid change affects the predicted protein in any way. In the second EPSPS gene copy, 10 SNPs were found in the GR biotype, 3 were found in the GS biotype, and 4 were found in both. 6 out of the 17 SNPs were found to cause amino acid changes in the predicted protein. Interestingly, all 6 were discovered in the GR biotype.

5.4 Discussion

The use of RNA-Seq and transcriptome analysis of GR and GS giant ragweed seems like a powerful approach to understand the mechanism of resistance. Even though this was a single-replicate study, preliminary results from the differential gene expression comparison and SNP analysis indicate that there could be multiple mechanisms that lead to glyphosate resistance in giant ragweed. We performed a time course study to quantitatively measure the impact of glyphosate on GR and GS giant ragweed across four time points - 0 (pre-treatment), 3, 8 and 12 hours after treatment with glyphosate. Looking at the differential gene expression before treatment with glyphosate, the genes expressed higher in GR plants seem to play important roles in pathogen resistance, while highly differentially expressed genes in GS plants play major roles in stress response. We can infer that GR plants possibly utilize a pathogen-response pathway to prevent the uptake of glyphosate, resulting in a hypersensitive-like response after glyphosate treatment.

Using the transcriptome annotation done in the previous chapter, we isolated two copies of the crucial EPSPS gene. Analyzing the gene expression of the EPSPS genes across the time points in both GR and GS plants, we found that while the expression level of the second gene copy remained relatively unchanged, the first gene copy showed dramatic increase in expression in later time points in GR plants. In GS plants, the expression of the first gene copy went down from the first time point to the second time point, and showed little change at later time points. This could indicate that the overexpression of the EPSPS gene could also be a possible resistance mechanism.

Finally, we did SNP analysis on the two EPSPS gene copies for both GR and GS biotypes of giant ragweed. We found that the first gene copy contained no SNPs in GR plants, and no SNPs that could have an effect on the protein sequence in GS

plants. For the second gene copy, there were a total of 17 SNPs, of which 6 possibly affect the predicted amino acid sequence. In contrast to the SNPs found in the first gene copy, the second gene copy had SNPs in both GR and GS biotypes, even though the 6 that affect the protein sequence were all from the GR biotype. This means that target-site mutation could potentially be an additional mechanism of resistance to glyphosate.

The complete transcriptome assembly of giant ragweed has been deposited in the NCBI BioProject database (http://www.ncbi.nlm.nih.gov/bioproject/) and is publicly available under accession PRJNA267208. The preliminary time-course experiment presented here identified groups of genes that may explain glyphosate resistance in giant ragweed. A more extensive transcriptome analysis study, with multiple replicates of sensitive and resistant giant ragweed biotypes, from a broader range of geographic sources, and with shorter time intervals will be useful to overcome the limitations of this preliminary study.

Table 5.1.: Normalization of expression values using control genes from rice. 12 genes from the list of 25 genes identified by Jain (2009) showed relatively stable expression across all time points, and thus were used for determining the scaling factor for the normalization (Jain, 2009)

	Initial average expression (CPM)											Scaled expression values							
Seq Name	**R03**	**R33**	**R83**	**R123**	**S03**	**S33**	**S83**	**S123**	**Gene name**	**Mean R**	**Mean S**	**R03**	**R33**	**R83**	**R123**	**S03**	**S33**	**S83**	**S123**
LOC_Os07g34589	92.97	130.37	153.91	138.51	90.76	78.69	121.71	142.69	Protein translation factor SUI1 homolog	128.94	108.4625	0.721033	1.01109	1.193656	1.074221	0.836787	0.725504	1.122139	1.31557
NM_001065286	182.57	244.85	103.47	136.94	175.15	133.2	202.35	240.41	Conserved hypothetical protein	166.9575	187.7775	1.093512	1.466541	0.619739	0.820209	0.932753	0.70935	1.077605	1.280292
LOC_Os04g35910	10.9	13.04	19.94	20	9.49	14.38	19.88	21.08	Coiled-coil domain containing 55	15.97	16.2075	0.68253	0.816531	1.248591	1.252348	0.585531	0.887244	1.226593	1.300632
LOC_Os01g05490	87.3	100.14	33.86	25.81	98.41	68.5	73.38	62.6	Triosephosphate isomerase	61.7775	75.7225	1.413136	1.620979	0.548096	0.41779	1.299614	0.904619	0.969065	0.826703
LOC_Os08g03290	519.16	487.42	235.41	209.67	749.7	664.05	768.51	622.39	Glyceraldehyde-3-phosphate dehydrogenase	362.915	701.1625	1.430528	1.343069	0.648664	0.577739	1.069224	0.94707	1.096051	0.887654
LOC_Os01g70780	33.59	32.51	41.98	31.06	51.52	38.63	52.9	64.43	SVP1-like protein 2	34.785	51.87	0.965646	0.934598	1.206842	0.892914	0.993252	0.744746	1.019857	1.242144
LOC_Os07g11290	5.84	7.65	10.93	9.83	4	5.96	15.4	11.38	Expressed protein	8.5625	9.185	0.682044	0.893431	1.276496	1.148029	0.435493	0.648884	1.676647	1.238977
LOC_Os04g53620	1147.96	1395.11	1766.81	1772.33	1549.44	1577.01	1345.8	929.58	Polyubiquitin	1520.553	1350.458	0.754962	0.917502	1.161953	1.165583	1.147345	1.16776	0.996551	0.688345
LOC_Os08g03390	53.36	57.83	58.61	56.47	50.29	56.69	95.99	86.41	Pre-mRNA-splicing factor SLU7	56.5675	72.345	0.943298	1.022318	1.036107	0.998276	0.695141	0.783606	1.326837	1.194416
NM_001057599	9.88	8.04	2.23	2.7	6.07	5.46	6.15	6.71	Atypical receptor-like kinase MARK	5.7125	6.0975	1.72954	1.40744	0.390372	0.472648	0.99549	0.895449	1.00861	1.100451
LOC_Os08g12750	7.29	7	5.08	7.51	9.3	10.57	14.39	15.06	Serine/threonine protein kinase	6.72	12.33	1.084821	1.041667	0.755952	1.11756	0.754258	0.857259	1.167072	1.221411
LOC_Os04g51370	13.88	17.12	20.6	21.42	11.35	15.18	10.89	16.06	Protein kinase domain containing protein	18.255	13.37	0.76034	0.937825	1.128458	1.173377	0.848915	1.135378	0.81451	1.201197
Final Scale Factor												1.000	1.134	1.088	1.069	1.000	1.040	1.461	1.432

Table 5.2.: Gene expression differences between resistant and sensitive biotypes of giant ragweed before treatment with glyphosate. The number of genes that are expressed more than 4-fold higher in glyphosate-resistant giant ragweed (Resistant +) or more than 4-fold higher in glyphosate-sensitive giant ragweed (Sensitive +) are shown.

Pre-treatment	
Resistant +	318
=	35079
Sensitive +	70

Table 5.3.: Gene expression differences between resistant and sensitive biotypes of giant ragweed after treatment with glyphosate. After treatment with glyphosate, the number of differentially expressed genes increases rapidly within the first three hours, and continues to increase at later time points. (+) denotes at least 4-fold higher expression level, (=) denotes similar expression level, (-) denotes at least 4-fold lower expression level, and PT stands for post-treatment.

		3 hours PT			**8 hours PT**			**12 hours PT**		
		Sensitive								
		(+)	(=)	(-)	(+)	(=)	(-)	(+)	(=)	(-)
Resistant	(+)	62	550	31	101	3342	323	412	5471	329
	(=)	552	33020	1014	2643	26654	597	1273	25339	696
	(-)	18	181	39	58	1632	117	22	1710	215

Table 5.4.: Genes with greater than four-fold higher expression in resistant plants compared to sensitive plants. Genes expressed higher in resistant plants tend to play important roles in pathogen response regulation.

Gene identifier	**Gene Annotation**	**GO Annotation**	**Expression ratio R/S**
comp148939_c0	Glycosyl hydrolase superfamily protein	GO:0009725 response to hormone stimulus	26.846
comp166081_c1	Alpha/beta-hydrolases superfamily protein	GO:0005515 protein binding	26.344
comp149865_c0	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein	GO:0012505 endomembrane system	7.665
comp142951_c0	Lipid transfer protein 12	GO:0008289 lipid binding	6.532
comp159731_c0	Glutathione S-transferase family protein	GO:0006457 protein folding	4.906
comp167561_c0	Protein kinase superfamily protein	GO:0006468 protein amino acid phosphorylation	4.777
comp158185_c0	Ethylene-forming enzyme	GO:0009815 1-aminocyclopropane-1-carboxylate oxidase activity	4.581
comp171245_c0	Pleiotropic drug resistance 7	GO:0005886 plasma membrane	4.076

Table 5.5.: Genes with greater than four-fold higher expression levels in sensitive plants compared to resistant plants. Genes expressed at a higher level in sensitive plants seem to impact control of stress response.

Gene identifier	**Gene Annotation**	**GO Annotation**	**Expression ratio S/R**
comp161591_c0	Metallathionein 2B	GO:0006508 proteolysis	12.926
comp165624_c0	Thioredoxin superfamily protein	GO:0009535 chloroplast thylakoid membrane	9.706
comp144176_c0	NADH-ubiquinone oxidoreductase (complex I), chain 5 protein	GO:0009507 chloroplast	7.694
comp150391_c0	Cysteine-rich domain superfamily protein	GO:0009507 chloroplast	4.972
comp165059_c0	Fe superoxide dismutase 2	GO:0019430 removal of superoxide radicals	4.702
comp163658_c0	Unknown protein involved in response to salt stress	GO:0003677 DNA binding	4.258

Table 5.6.: Highly significant GO terms determined by BlastX search against *Arabidopsis thaliana* using agriGO for glyphosate resistant giant ragweed. All matching *Arabidopsis* genes with E-value less than $1e^{-20}$ and percentage identity greater than 40% were retained.

GO term	**Description**	**P**	**FDR**
GO:0006457	protein folding	5.20E-07	4.20E-05
GO:0019748	secondary metabolic process	2.30E-06	9.20E-05
GO:0051707	response to other organism	7.70E-05	0.0021
GO:0009607	response to biotic stimulus	0.00011	0.0023

Table 5.7.: Highly significant GO terms determined by BlastX search against *Arabidopsis thaliana* using agriGO for glyphosate sensitive giant ragweed. All matching *Arabidopsis* genes with E-value less than $1e^{-20}$ and percentage identity greater than 40% were retained.

GO term	**Description**	**P**	**FDR**
GO:0009699	phenylpropanoid biosynthetic process	1.10E-26	1.20E-23
GO:0019748	secondary metabolic process	7.40E-25	4.00E-22
GO:0009698	phenylpropanoid metabolic process	5.70E-24	2.00E-21
GO:0019438	aromatic compound biosynthetic process	1.30E-23	3.50E-21
GO:0006725	cellular aromatic compound metabolic process	1.90E-23	4.10E-21
GO:0042398	cellular amino acid derivative biosynthetic process	1.30E-21	2.40E-19
GO:0006952	defense response	3.00E-18	4.60E-16
GO:0006950	response to stress	4.60E-18	6.20E-16
GO:0006575	cellular amino acid derivative metabolic process	5.50E-18	6.60E-16
GO:0050896	response to stimulus	3.00E-17	3.20E-15
GO:0006519	cellular amino acid and derivative metabolic process	1.20E-15	1.20E-13
GO:0051707	response to other organism	1.00E-13	9.30E-12
GO:0009607	response to biotic stimulus	1.10E-13	9.50E-12
GO:0051704	multi-organism process	5.50E-12	4.30E-10
GO:0009808	lignin metabolic process	7.30E-11	5.30E-09
GO:0009807	lignan biosynthetic process	1.70E-10	1.10E-08
GO:0009806	lignan metabolic process	1.70E-10	1.10E-08
GO:0006468	protein amino acid phosphorylation	2.90E-10	1.80E-08
GO:0042221	response to chemical stimulus	5.50E-10	3.20E-08
GO:0009809	lignin biosynthetic process	6.00E-10	3.30E-08
GO:0006979	response to oxidative stress	6.80E-09	3.50E-07
GO:0016310	phosphorylation	8.40E-09	4.20E-07
GO:0008152	metabolic process	1.40E-08	6.80E-07
GO:0006796	phosphate metabolic process	7.10E-08	3.20E-06
GO:0006793	phosphorus metabolic process	7.30E-08	3.20E-06

Table 5.8.: EPSPS gene copies in the giant ragweed transcriptome and their annotations

Sequence ID: comp144227_c0_seq1
PFAM annotation: PF00275.15: EPSP synthase
GO annotation:
GO:0003866: 3-phosphoshikimate 1-carboxyvinyltransferase activity,
GO:0009073: aromatic amino acid family biosynthetic process
Predicted Gene Model:
MNLASLSCNQTINKRSLAVAASVATTEKSSVEEIVLKPIKEISGTVNLPGSKS
LSNRILLLAALAEGTTVVDNLLNSDDVHYMLGALRALGLNVEENGEIKRAT
VEGCGGVFPVGKEAKDEIQLFLGNAGTAMRPLTAAVTAAGGNSSYILDG
VPRMRERPIGDLVTGLKQLGADVDCFLGTNCPPVRVAANGGLPGGKVKL
SGSISSQYLTALLMAAPLALGDVEIEIIDKLISVPYVEMTLKLMERFGVSVEHS
DSWDKFYVRGGQKYKSPGNAYVEGDASSASYFLAGAAITGGTVTVEGC
GTSSLQGDVKFAEVLGQMGAEVTWTENSVTVKGPPRNASGRGHLRPVDV
NMNKMPDVAMTLAVVALYADGPTAIRDVASWRVKETERMIAICTELRKLG
ATVEEGPDYCVITPPEKLNVTAIDTYDDHRMAMAFSLAACADVPVTIKDPG
CTRKTFPDYFEVLERFTKH*
Sequence ID: comp163966_c0_seq1
PFAM annotation: PF00275.15 EPSP synthase
GO annotation:
GO:0003866: 3-phosphoshikimate 1-carboxyvinyltransferase activity,
GO:0009073: aromatic amino acid family biosynthetic process
Predicted Gene Model:
MAAHVSNVAQNIQTSNSIFNNLSKSQTPSSKSSPFLSFGSKYKTPFTHFSFS
SNNRKLFTVSASVAATSAIPEIVLQPIKEISGTVNLPGSKSLSNRILLLAALSQ
GTTVVDNLLNSDDVHYMLGALRTLGLRVDEDGAIKRAVVEGCGGVFPV
GREAKDEIQLFLGNAGTAMRPLTAAVTAAGGNSSYILDGVPRMRERPIGD
LVTGLKQLGADVDCFLGTNCPPVRVVGGGGLPGGKVKLSGSISSQYLT
ALLMASPLALGDVEIEIIDKLISIPYVEMTIKLMERFGVSVEHSDSWDRFFIKG
GQKYKSPGNAYVEGDASSASYFLAGAAITGGTITVEGCGTSSLQGDVK
FAEVLGQMGAEVTWTENSVTVKGPARDASGRKHLRAVDVNMNKMPDV
AMTLAVVALYADGPTAIRDVASWRVKETERMIAICTELRKLGATVEEGPD
YCVITPPERLNVAAIDTYDDHRMAMAFSLAACADVPVTIKDPACTRKTFP
DYFEVLQRFTKH*

Table 5.9.: Comparison of the expression of the EPSPS gene copies across the four time points in GR and GS giant ragweed

Sequence ID	**FPKM value**							
	R0	**R3**	**R8**	**R12**	**S0**	**S3**	**S8**	**S12**
comp144227_c0	52.21	66.21	298.66	798.74	64.98	18.97	16.51	22.23
comp163966_c0	4.72	3.7	2.11	7.77	19.17	8.75	6.75	5.59

Table 5.10.: SNPs found in the first copy of the EPSPS gene (comp144227) in glyphosate-sensitive giant ragweed. Amino acid changes in italics indicate amino acid changes in the protein sequence. The GR biotype had no SNPs.

Position	**Consensus nucleotide**	**Modified nucleotide**	**Consensus amino acid**	**Modified amino acid**
111	A	C	asn	*lys*
273	CT	C	asn	asn
276	AG	A	lys	lys
360	CT	T	ile	ile
378	AG	G	leu	leu
468	TC	C	asp	asp
501	CT	C	ala	ala
508	TC	T	leu	leu
558	CT	C	cys	cys
618	TC	C	asn	asn
621	A	G	ala	ala
720	CT	C	ile	ile
723	CT	T	gly	gly
756	A	C	ala	ala
768	T	C	cys	cys
789	AG	A	pro	pro
990	TC	T	ser	ser
1002	TC	C	ser	ser
1017	GA	G	lys	lys
1029	CA	A	arg	arg
1179	TC	C	phe	phe
1194	AT	A	gly	gly
1209	AG	G	glu	glu
1314	TC	T	asp	asp
1371	GA	G	arg	arg
1416	TC	T	ala	ala
1602	CG	G	thr	thr
1611	TC	C	thr	thr
1746	T	C	leu	leu

Table 5.11.: SNPs found in the second copy of the EPSPS gene (comp163996) in giant ragweed. Amino acid changes in italics indicate amino acid changes in the protein sequence.

Position	**Consensus nucleotide**	**Modified nucleotide**	**Consensus amino acid**	**Modified amino acid**	**Giant ragweed biotype**
204	TC	T	tyr	*his*	R
247	CG	G	arg	*thr*	R
274	CT	T	val	*ala*	R
305	G	T	val	val	S
623	T	G	thr	thr	S
866	TC	C	leu	leu	R
872	TC	C	asp	asp	R
897	TC	T	leu	leu	RS
956	AG	A	ser	ser	RS
960	GA	G	glu	*lys*	R
1052	CT	T	ser	ser	RS
1097	TC	C	ile	ile	S
1340	TC	T	asp	asp	R
1415	AT	A	thr	thr	R
1428	CT	C	pro	*ser*	R
1447	GC	C	thr	*ser*	R
1487	CT	T	tyr	tyr	RS

Fig. 5.1.: Comparison of resistant (left) and sensitive (right) giant ragweed biotypes 12 hours after glyphosate treatment. Note the rapid necrosis in the resistant biotype.

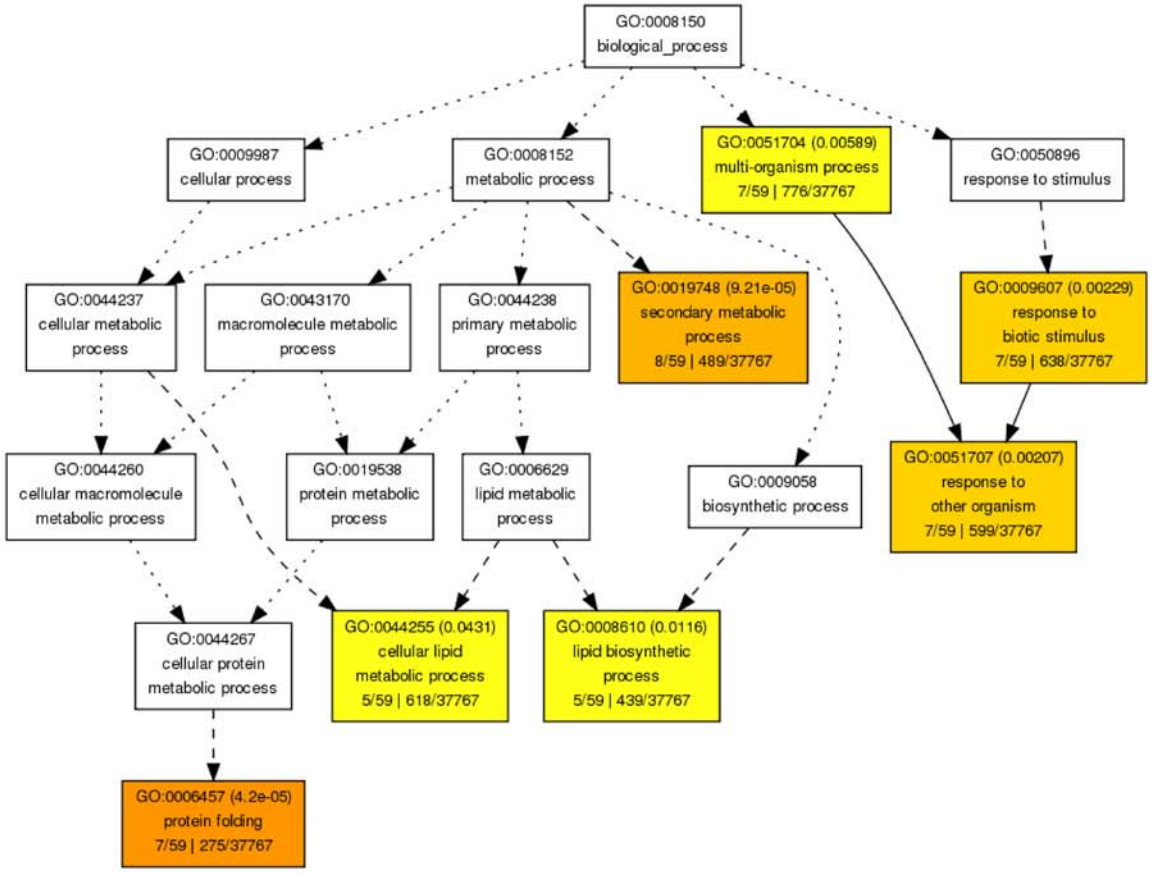

Fig. 5.2.: agriGO analysis of genes expressed higher in GR compared to GS. Gene Ontology Biological Process terms with P value less than $1e^{-7}$ are shown.

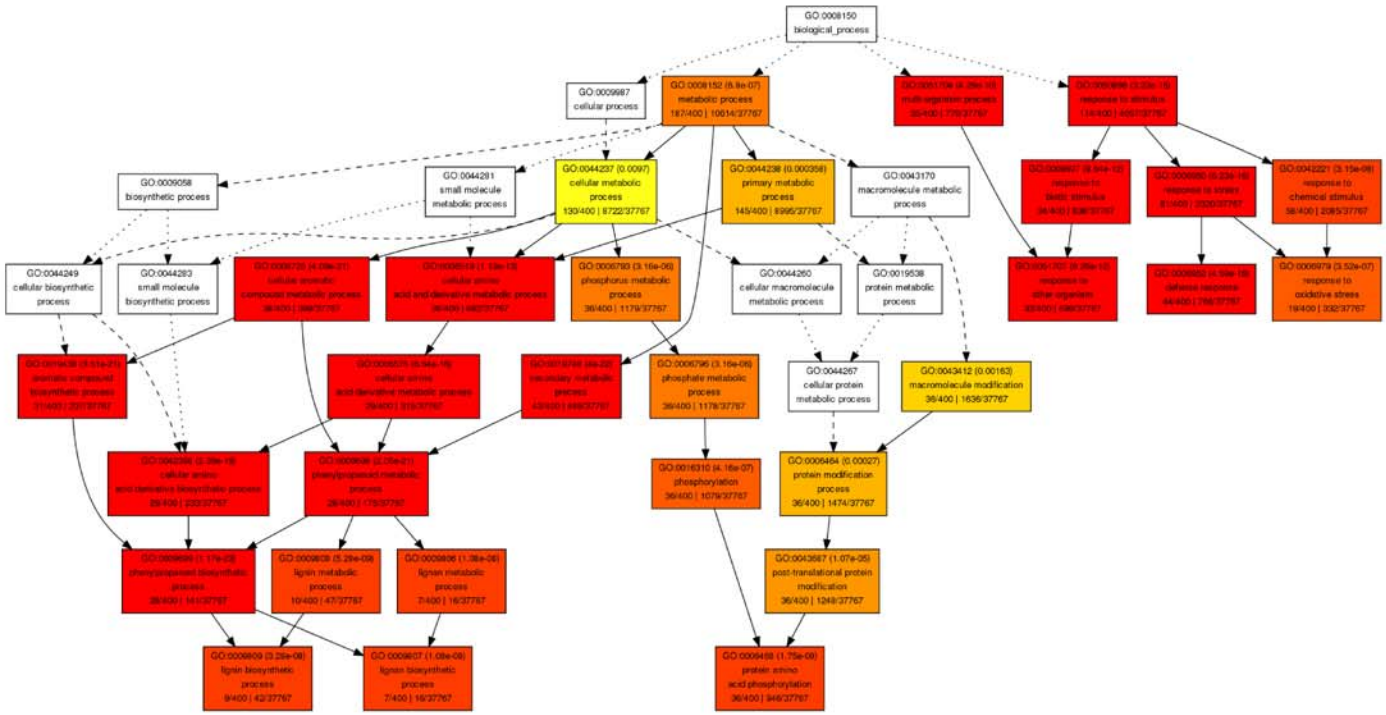

Fig. 5.3.: agriGO analysis of genes expressed higher in GS compared to GR. Gene Ontology Biological Process terms with P value less than $1e^{-7}$ are shown.

REFERENCES

REFERENCES

Abul-Fatih, H. and Bazzaz, F. A. (1979). The biology of ambrosia trifida l. *New phytologist*, 83(3):817–827.

Afzal, A. J., Wood, A. J., and Lightfoot, D. A. (2008). Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Molecular plant-microbe interactions*, 21(5):507–517.

Alkan, C., Sajjadian, S., and Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature methods*, 8(1):61–65.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D. R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J. A., and Zdobnov, E. M. (2001). The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic acids research*, 29(1):37–40.

Asai, S. and Yoshioka, H. (2008). The role of radical burst via mapk signaling in plant immunity. *Plant signaling & behavior*, 3(11):920–922.

Bairoch, A. (1991). Prosite: a dictionary of sites and patterns in proteins. *Nucleic acids research*, 19(Suppl):2241.

Bassett, I. and Crompton, C. (1982). The biology of canadian weeds. 55. ambrosia trifida l. *Canadian journal of plant science*, 62:1003–1010.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000). The pfam protein families database. *Nucleic acids research*, 28(1):263–266.

Baysinger, J. A. and Sims, B. D. (1991). Giant ragweed (ambrosia trifida) interference in soybeans (glycine max). *Weed science*, pages 358–362.

Becraft, P. W. (1998). Receptor kinases in plant development. *Trends in plant science*, 3(10):384–388.

Bialik, S. and Kimchi, A. (2006). The death-associated protein kinases: structure, function, and beyond. *Annual review of biochemistry*, 75:189–210.

Blanc, G. and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The plant cell*, 16(7):1667–1678.

Boerboom, C. M., Wyse, D. L., and Somers, D. A. (1990). Mechanism of glyphosate tolerance in birdsfoot trefoil (lotus corniculatus). *Weed science*, pages 463–467.

Bradshaw, L. D., Padgette, S. R., Kimball, S. L., and Wells, B. H. (1997). Perspectives on glyphosate resistance. *Weed technology*, pages 189–198.

Buchanan, B. B., Gruissem, W., and Jones, R. L. (2015). *Biochemistry and molecular biology of plants*. John Wiley & Sons, New York, NY, USA.

Chaves, M. M., Maroco, J. P., and Pereira, J. S. (2003). Understanding plant responses to droughtfrom genes to the whole plant. *Functional plant biology*, 30(3):239–264.

Chevreux, B., Wetter, T., and Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. In *German conference on bioinformatics*, volume 99, pages 45–56. Heidelberg.

Clarke, A., Desikan, R., Hurst, R. D., Hancock, J. T., and Neill, S. J. (2000). No way back: nitric oxide and programmed cell death in arabidopsis thaliana suspension cultures. *The plant journal*, 24(5):667–677.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676.

Consortium, G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl 1):D258–D261.

Cristina, M. S., Petersen, M., and Mundy, J. (2010). Mitogen-activated protein kinase signaling in plants. *Annual review of plant biology*, 61:621–649.

Das, M., Reichman, J. R., Haberer, G., Welzl, G., Aceituno, F. F., Mader, M. T., Watrud, L. S., Pfleeger, T. G., Gutiérrez, R. A., Schäffner, A. R., and Olszyk, D. M. (2010). A composite transcriptional signature differentiates responses towards closely related herbicides in arabidopsis thaliana and brassica napus. *Plant molecular biology*, 72(4):545–556.

De Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., and Hulo, N. (2006). Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic acids research*, 34(suppl 2):W362–W365.

de María, N., Becerril, J. M., García-Plazaola, J. I., Hernández, A., de Felipe, M. R., and Fernández-Pascual, M. (2006). New insights on glyphosate mode of action in nodular metabolism: Role of shikimate accumulation. *Journal of agricultural and food chemistry*, 54(7):2621–2628.

De Smet, I., Voß, U., Jürgens, G., and Beeckman, T. (2009). Receptor-like kinases shape the plant. *Nature cell biology*, 11(10):1166–1173.

Delye, C. (2013). Unravelling the genetic bases of non-target-site-based resistance (ntsr) to herbicides: a major challenge for weed science in the forthcoming decade. *Pest management science*, 69(2):176–187.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agrigo: a go analysis toolkit for the agricultural community. *Nucleic acids research*, page gkq310.

Duke, S. O. and Powles, S. B. (2008). Glyphosate: a once-in-a-century herbicide. *Pest management science*, 64(4):319–325.

Duret, L., Mouchiroud, D., and Gautier, C. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in gc-rich isochores. *Journal of molecular evolution*, 40(3):308–317.

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M. D., Lawrence, C. J., Lushbough, C., and Brendel, V. (2008). Plantgdb: a resource for comparative plant genomics. *Nucleic acids research*, 36(suppl 1):D959–D965.

Eggermont, K., Goderis, I. J., and Broekaert, W. F. (1996). High-throughput rna extraction from plant samples based on homogenisation by reciprocal shaking in the presence of a mixture of sand and glass beads. *Plant molecular biology reporter*, 14(3):273–279.

Eide, E. J. and Virshup, D. M. (2001). Casein kinase i: another cog in the circadian clockworks. *Chronobiology international*, 18(3):389–398.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). Hmmer web server: interactive sequence similarity searching. *Nucleic Acids Research*.

Foresman, C. and Glasgow, L. (2008). Us grower perceptions and experiences with glyphosate-resistant weeds. *Pest management science*, 64(4):388–391.

Ge, X., d'Avignon, D., Ackerman, J., Ostrander, E., and Sammons, R. (2013). Applications of 31p nmr spectroscopy to glyphosate studies in plants: insights into cellular uptake and vacuolar sequestration correlated to herbicide resistance. *Herbicides: biological activity, classification and health and environmental implications*.

Ge, X., d'Avignon, D. A., Ackerman, J. J., and Sammons, R. D. (2010). Rapid vacuolar sequestration: the horseweed glyphosate resistance mechanism. *Pest management science*, 66(4):345–348.

Gill, N., Buti, M., Kane, N., Bellec, A., Helmstetter, N., Berges, H., and Rieseberg, L. H. (2014). Sequence-based analysis of structural organization and composition of the cultivated sunflower (helianthus annuus l.) genome. *Biology*, 3(2):295–319.

Gomes, M. P., Smedbol, E., Chalifour, A., Hénault-Ethier, L., Labrecque, M., Lepage, L., Lucotte, M., and Juneau, P. (2014). Alteration of plant physiology by glyphosate and its by-product aminomethylphosphonic acid: an overview. *Journal of experimental botany*, 65(17):4691–4703.

Goodswen, S. J., Kennedy, P. J., and Ellis, J. T. (2012). Evaluating high-throughput ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. *PloS one*, 7(11):e50609.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechology*, 29(7):644–652.

Graham, L. K. and Wilcox, L. W. (2000). The origin of alternation of generations in land plants: a focus on matrotrophy and hexose transport. *Philosophical transactions of the royal society of london B: Biological sciences*, 355(1398):757–767.

Guigó, R., Agarwal, P., Abril, J. F., Burset, M., and Fickett, J. W. (2000). An assessment of gene prediction accuracy in large dna sequences. *Genome research*, 10(10):1631–1642.

Guo, H. and Ecker, J. R. (2004). The ethylene signaling pathway: new insights. *Current opinion in plant biology*, 7(1):40–49.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., and White, O. (2003). Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research*, 31(19):5654–5666.

Haas, B. J., Wortman, J. R., Ronning, C. M., Hannick, L. I., Smith, R. K., Maiti, R., Chan, A. P., Yu, C., Farzad, M., Wu, D., White, O., and Town, C. D. (2005). Complete reannotation of the arabidopsis genome: methods, tools, protocols and the final release. *BMC biology*, 3(1):1–19.

Haig, D. (2010). What do we know about charophyte (streptophyta) life cycles? *Journal of phycology*, 46(5):860–867.

Haig, D. and Wilczek, A. (2006). Sexual conflict and the alternation of haploid and diploid generations. *Philosophical transactions of the royal society of london B: Biological sciences*, 361(1466):335–343.

Halford, N. G. and Hardie, D. G. (1998). Snf1-related protein kinases: global regulators of carbon metabolism in plants? *Plant molecular biology*, 37(5):735–748.

Hanks, S. K. and Hunter, T. (1995). Protein kinases 6. the eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *The FASEB journal*, 9(8):576–596.

Hanks, S. K. and Quinn, A. M. (1991). Protein kinase catalytic domain sequence database: Identification of conserved features of primary structure and classification of family members. *Methods in enzymology*, 200:38–62.

Harrison, S. K., Regnier, E. E., Schmoll, J. T., and Webb, J. E. (2001). Competition and fecundity of giant ragweed in corn. *Weed Science*, 49(2):224–229.

Haussler, D. (1998). Computational genefinding. *Trends in biotechnology*, 16:12–15.

Heap, I. (1997). International survey of herbicide-resistant weeds. In *Western society of weed science (USA)*.

Hetherington, P., Reynolds, T., Marshall, G., and Kirkwood, R. (1999). The absorption, translocation and distribution of the herbicide glyphosate in maize expressing the cp-4 transgene. *Journal of experimental botany*, 50(339):1567–1576.

Hildebrand, P. W., Preissner, R., and Frömmel, C. (2004). Structural features of transmembrane helices. *FEBS letters*, 559(1-3):145–151.

Hirayama, T. and Shinozaki, K. (2007). Perception and transduction of abscisic acid signals: keys to the function of the versatile plant hormone aba. *Trends in plant science*, 12(8):343–351.

Hoff, K. and Stanke, M. (2015). Current methods for automated annotation of protein-coding genes. *Current Opinion in insect science*, 7:8–14.

Jabbari, K. and Bernardi, G. (1998). Cpg doublets, cpg islands and alu repeats in long human dna sequences from different isochore families. *Gene*, 224(1):123–128.

Jain, M. (2009). Genome-wide identification of novel internal control genes for normalization of gene expression during various stages of development in rice. *Plant science*, 176(5):702–706.

Jasieniuk, M., Ahmad, R., Sherwood, A. M., Firestone, J. L., Perez-Jones, A., Lanini, W. T., Mallory-Smith, C., and Stednick, Z. (2008). Glyphosate-resistant italian ryegrass (lolium multiflorum) in california: distribution, response to glyphosate, and molecular evidence for an altered target enzyme. *Weed science*, 56(4):496–502.

Jasieniuk, M., Brûlé-Babel, A. L., and Morrison, I. N. (1996). The evolution and genetics of herbicide resistance in weeds. *Weed science*, pages 176–193.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., and Bork, P. (2008). eggnog: automated construction and annotation of orthologous groups of genes. *Nucleic acids research*, 36(suppl 1):D250–D254.

Johnson, D. A., Akamine, P., Radzio-Andzelm, E., Madhusudan, , and Taylor, S. S. (2001). Dynamics of camp-dependent protein kinase. *Chemical reviews*, 101(8):2243–2270.

Jones, J., Goldsbrough, P., and Weller, S. (1996). Stability and expression of amplified epsps genes in glyphosate resistant tobacco cells and plantlets. *Plant cell reports*, 15(6):431–436.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in kegg. *Nucleic acids research*, 34(suppl 1):D354–D357.

Karol, K. G., McCourt, R. M., Cimino, M. T., and Delwiche, C. F. (2001). The closest living relatives of land plants. *Science*, 294(5550):2351–2353.

Kaul, S., Koo, H., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L., Feldblyum, T., Nierman, W., Benito, M.-I., Lin, X., Town, C., Venter, J., Fraser, C., Tabata, S., Nakamura, Y., Kaneko, T., Sato, S., Asamizu, E., Kato, T., Kotani, H., Sasamoto, S., Ecker, J., Theologis, A., Federspiel, N., Palm, C., Osborne, B., Shinn, P., Dewar, K., Kim, C., Buehler, E., Dunn, P., Chao, Q., Chen, H., Theologis, A., Osborne, B., Vysotskaia, V., Lenz, C., Kim, C., Hansen, N., Liu, S., Buehler, E., Alta, H., Sakano, H., Dunn, P., Lam, B., Pham, P., Chao, Q., Nguyen, M., Yu, G., Chen, H., Southwick, A., Lee, J., Miranda, M., Toriumi, M., Davis, R., Federspiel, N., Palm, C., Conway, A., Conn, L., Hansen, N., Hootan, A., Lam, B., Wambutt, R., Murphy, G., Dsterhft, A., Stiekema, W., Pohl, T., Entian, K.-D., Terryn, N., Volckaert, G., Salanoubat, M., Choisne, N., Artiguenave, F., Weissenbach, J., Quetier, F., Rieger, M., Ansorge, W., Unseld, M., Fartmann, B., Valle, G., Wilson, R., Sekhon, M.,

Pepin, K., Murray, J., Johnson, D., Hillier, L., de la Bastide, M., Huang, E., Spiegel, L., Gnoj, L., Habermann, K., Dedhia, N., Parnell, L., Preston, R., Marra, M., McCombie, W., Chen, E., Martienssen, R., Mayer, K., Lemcke, K., Haas, B., Haase, D., Rudd, S., Schoof, H., Frishman, D., Morgenstern, B., Zaccaria, P., Mewes, H.-W., White, O., Creasy, T., Bielke, C., Maiti, R., Peterson, J., Ermolaeva, M., Pertea, M., Quackenbush, J., Volfovsky, N., Wu, D., Salzberg, S., Bevan, M., Lowe, T., Rounsley, S., Bush, D., Subramaniam, S., Levin, I., Norris, S., Schmidt, R., Acarkan, A., Bancroft, I., Brennicke, A., Eisen, J., Bureau, T., Legault, B.-A., Le, Q.-H., Agrawal, N., Yu, Z., Copenhaver, G., Luo, S., Preuss, D., Pikaard, C., Paulsen, I., Sussman, M., Britt, A., Selinger, D., Pandey, R., Chandler, V., Jorgensen, R., Mount, D., Pikaard, C., Juergens, G., Meyerowitz, E., Dangl, J., Jones, J., Chen, M., Chory, J., and Somerville, C. (2000). Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815.

Kemp, M. S., Moss, S. R., and Thomas, T. H. (1990). Herbicide resistance in alopecurus myosuroides. In *ACS symposium series-american chemical society (USA)*, Washington, DC, USA.

Kern, A. J. and Dyer, W. E. (1998). Compartmental analysis of herbicide efflux in susceptible and difenzoquat-resistantavena fatual. suspension cells. *Pesticide biochemistry and physiology*, 61(1):27–37.

Kobayashi, M., Ohura, I., Kawakita, K., Yokota, N., Fujiwara, M., Shimamoto, K., Doke, N., and Yoshioka, H. (2007). Calcium-dependent protein kinases regulate the production of reactive oxygen species by potato nadph oxidase. *The plant cell*, 19(3):1065–1080.

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the royal society of london B: Biological sciences*, 279(1749):5048–5057.

Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Rogozin, I. B., Smirnov, S., Sorokin, A. V., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome biology*, 5(2):R7–R7. gb-2004-5-2-r7[PII].

Kornev, A. P., Haste, N. M., Taylor, S. S., and Ten Eyck, L. F. (2006). Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proceedings of the national academy of sciences*, 103(47):17783–17788.

Kovtun, Y., Chiu, W.-L., Tena, G., and Sheen, J. (2000). Functional analysis of oxidative stress-activated mitogen-activated protein kinase cascade in plants. *Proceedings of the national academy of sciences*, 97(6):2940–2945.

Kriventseva, E. V., Rahman, N., Espinosa, O., and Zdobnov, E. M. (2008). Orthodb: the hierarchical catalog of eukaryotic orthologs. *Nucleic acids research*, 36(suppl 1):D271–D275.

Lai, Z., Kane, N. C., Kozik, A., Hodgins, K. A., Dlugosch, K. M., Barker, M. S., Matvienko, M., Yu, Q., Turner, K. G., Pearl, S. A., Bell, G. D. M., Zou, Y., Grassa, C., Guggisberg, A., Adams, K. L., Anderson, J. V., Horvath, D. P., Kesseli, R. V.,

Burke, J. M., Michelmore, R. W., and Rieseberg, L. H. (2012). Genomics of compositae weeds: Est libraries, microarrays, and evidence of introgression. *American journal of botany*, 99(2):209–218.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., and Huala, E. (2012). The arabidopsis information resource (tair): improved gene annotation and new tools. *Nucleic acids research*, 40(D1):D1202–D1210.

Lee, J. S., Wang, S., Sritubtim, S., Chen, J.-G., and Ellis, B. E. (2009). Arabidopsis mitogen-activated protein kinase mpk12 interacts with the mapk phosphatase ibr5 and regulates auxin signaling. *The plant journal*, 57(6):975–985.

Lehti-Shiu, M. D. and Shiu, S.-H. (2012). Diversity, classification and function of the plant protein kinase superfamily. *Phil. Trans. R. Soc. B*, 367(1602):2619–2639.

Lespinet, O., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome research*, 12(7):1048–1059.

Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):1.

Li, H., Liu, J.-S., Xu, Z., Jin, J., Fang, L., Gao, L., Li, Y.-D., Xing, Z.-X., Gao, S.-G., Liu, T., Li, H.-H., Li, Y., Fang, L.-J., Xie, H.-M., Zheng, W.-M., and Hao, B.-L. (2005). Test data sets and evaluation of gene prediction programs on the rice genome. *Journal of computer science and technology*, 20(4):446–453.

Liu, J., Ishitani, M., Halfter, U., Kim, C.-S., and Zhu, J.-K. (2000a). The arabidopsis thaliana sos2 gene encodes a protein kinase that is required for salt tolerance. *Proceedings of the national academy of sciences*, 97(7):3730–3734.

Liu, Q., Zhang, Y., and Chen, S. (2000b). Plant protein kinase genes induced by drought, high salt and cold stresses. *Chinese science bulletin*, 45(13):1153–1157.

Lohrmann, J. and Harter, K. (2002). Plant two-component signaling systems and the role of response regulators. *Plant physiology*, 128(2):363–369.

Malik, J., Barry, G., and Kishore, G. (1989). The herbicide glyphosate. *BioFactors*, 2(1):17–25.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934.

McClendon, C. L., Kornev, A. P., Gilson, M. K., and Taylor, S. S. (2014). Dynamic architecture of a protein kinase. *Proceedings of the national academy of sciences*, 111(43):E4623–E4631.

McDonald, B. A. and Linde, C. (2002). Pathogen population genetics, evolutionary potential, and durable resistance. *Annual review of phytopathology*, 40(1):349–379.

Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., Terry, A., Salamov, A., Fritz-Laylin, L. K., Maréchal-Drouard, L., Marshall, W. F., Qu, L.-H., Nelson, D. R., Sanderfoot, A. A., Spalding, M. H., Kapitonov, V. V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S. M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C.-L., Cognat, V., Croft, M. T., Dent, R., Dutcher, S., Fernández, E., Fukuzawa, H., González-Ballester, D., González-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P. A., Lemaire, S. D., Lobanov, A. V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J. V., Moseley, J., Napoli, C., Nedelcu, A. M., Niyogi, K., Novoselov, S. V., Paulsen, I. T., Pazour, G., Purton, S., Ral, J.-P., Riaño-Pachón, D. M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S. L., Allmer, J., Balk, J., Bisova, K., Chen, C.-J., Elias, M., Gendler, K., Hauser, C., Lamb, M. R., Ledford, H., Long, J. C., Minagawa, J., Page, M. D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A. M., Yang, P., Ball, S., Bowler, C., Dieckmann, C. L., Gladyshev, V. N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R. T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L., Huang, Y. W., Jhaveri, J., Luo, Y., Martínez, D., Ngau, W. C. A., Otillar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., Grigoriev, I. V., Rokhsar, D. S., and Grossman, A. R. (2007). The chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, 318(5848):245–250.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2016). Panther version 10: expanded protein families and functions, and analysis tools. *Nucleic acids research*, 44(D1):D336–D342.

Mikołajczyk, M., Awotunde, O. S., Muszyńska, G., Klessig, D. F., and Dobrowolska, G. (2000). Osmotic stress induces rapid activation of a salicylic acid–induced protein kinase and a homolog of protein kinase ask1 in tobacco cells. *The plant cell*, 12(1):165–178.

Mittler, R. (2006). Abiotic stress, the field environment and stress combination. *Trends in plant science*, 11(1):15–19.

Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C., and Bernardi, G. (1991). The distribution of genes in the human genome. *Gene*, 100:181–187.

Nakagami, H., Pitzschke, A., and Hirt, H. (2005). Emerging map kinase pathways in plant stress signalling. *Trends in plant science*, 10(7):339–346.

Nandula, V. K. (2010). *Glyphosate resistance in crops and weeds: history, development, and management*. John Wiley & Sons, Hoboken, NJ, USA.

Nasiri, J., Haghnazari, A., and Alavi, M. (2011). Evaluation of prediction accuracy of genefinders using mouse genomic dna. *Trends in bioinformatics*, 4:10–22.

Nishiyama, T., Fujita, T., Shin-I, T., Seki, M., Nishide, H., Uchiyama, I., Kamiya, A., Carninci, P., Hayashizaki, Y., Shinozaki, K., Kohara, Y., and Hasebe, M. (2003). Comparative genomics of physcomitrella patens gametophytic transcriptome and arabidopsis thaliana: implication for land plant evolution. *Proceedings of the national academy of sciences*, 100(13):8007–8012.

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I. V., and Dubchak, I. (2014). The genome portal of the department of energy joint genome institute: 2014 updates. *Nucleic acids research*, 42(D1):D26–D31.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R. L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., and Buell, C. R. (2007). The tigr rice genome annotation resource: improvements and new features. *Nucleic acids research*, 35(suppl 1):D883–D887.

Parra, G., Bradnam, K., and Korf, I. (2007). Cegma: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9):1061–1067.

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785–786.

Petit, J. R., Jouzel, J., Raynaud, D., Barkov, N. I., Barnola, J.-M., Basile, I., Bender, M., Chappellaz, J., Davis, M., Delaygue, G., Delmotte, M., Kotlyakov, V. M., Legrand, M., Lipenkov, V. Y., Lorius, C., PEpin, L., Ritz, C., Saltzman, E., and Stievenard, M. (1999). Climate and atmospheric history of the past 420,000 years from the vostok ice core, antarctica. *Nature*, 399(6735):429–436.

Picardi, E. and Pesole, G. (2010). Computational methods for ab initio and comparative gene finding. *Data mining techniques for the life sciences*, pages 269–284.

Pline-Srnic, W. (2006). Physiological mechanisms of glyphosate resistance 1. *Weed technology*, 20(2):290–300.

Powles, S. B. and Preston, C. (2006). Evolved glyphosate resistance in plants: biochemical and genetic basis of resistance 1. *Weed technology*, 20(2):282–289.

Powles, S. B. and Yu, Q. (2010). Evolution in action: plants resistant to herbicides. *Annual review of plant biology*, 61:317–347.

Pratley, J., Urwin, N., Stanton, R., Baines, P., Broster, J., Cullis, K., Schafer, D., Bohn, J., and Krueger, R. (1999). Resistance to glyphosate in lolium rigidum. i. bioevaluation. *Weed science*, pages 405–411.

Preston, C. and Wakelin, A. M. (2008). Resistance to glyphosate from altered herbicide translocation patterns. *Pest management science*, 64(4):372–376.

Preston, C., Wakelin, A. M., Dolman, F. C., Bostamam, Y., and Boutsalis, P. (2009). A decade of glyphosate-resistant lolium around the world: mechanisms, genes, fitness, and agronomic management. *Weed science*, 57(4):435–441.

Project, I. R. G. S. (2005). The map-based sequence of the rice genome. *Nature*, 436(7052):793–800.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). Interproscan: protein domains identifier. *Nucleic acids research*, 33(suppl 2):W116–W120.

Reddy, A. R., Chaitanya, K. V., and Vivekanandan, M. (2004). Drought-induced responses of photosynthesis and antioxidant metabolism in higher plants. *Journal of plant physiology*, 161(11):1189–1202.

Rensing, S. A., Ick, J., Fawcett, J. A., Lang, D., Zimmer, A., Van de Peer, Y., and Reski, R. (2007). An ancient genome duplication contributed to the abundance of metabolic genes in the moss physcomitrella patens. *BMC evolutionary biology*, 7(1):1.

Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E. A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin-I, T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S.-i., Yamaguchi, K., Sugano, S., Kohara, Y., Fujiyama, A., Anterola, A., Aoki, S., Ashton, N., Barbazuk, W. B., Barker, E., Bennetzen, J. L., Blankenship, R., Cho, S. H., Dutcher, S. K., Estelle, M., Fawcett, J. A., Gundlach, H., Hanada, K., Heyl, A., Hicks, K. A., Hughes, J., Lohr, M., Mayer, K., Melkozernov, A., Murata, T., Nelson, D. R., Pils, B., Prigge, M., Reiss, B., Renner, T., Rombauts, S., Rushton, P. J., Sanderfoot, A., Schween, G., Shiu, S.-H., Stueber, K., Theodoulou, F. L., Tu, H., Van de Peer, Y., Verrier, P. J., Waters, E., Wood, A., Yang, L., Cove, D., Cuming, A. C., Hasebe, M., Lucas, S., Mishler, B. D., Reski, R., Grigoriev, I. V., Quatrano, R. S., and Boore, J. L. (2008). The physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319(5859):64–69.

Rizzon, C., Ponger, L., and Gaut, B. S. (2006). Striking similarities in the genomic distribution of tandemly arrayed genes in arabidopsis and rice. *PLoS computational biology*, 2(9):e115.

Rochaix, J.-D. (1995). Chlamydomonas reinhardtii as the photosynthetic yeast. *Annual review of genetics*, 29(1):209–230.

Romeis, T. (2001). Protein kinases in the plant defence response. *Current opinion in plant biology*, 4(5):407–414.

Rupprecht, J. (2009). From systems biology to fuelchlamydomonas reinhardtii as a model for a systems biology approach to improve biohydrogen production. *Journal of biotechnology*, 142(1):10–20.

Saijo, Y., Hata, S., Kyozuka, J., Shimamoto, K., and Izui, K. (2000). Over-expression of a single ca2+-dependent protein kinase confers both cold and salt/drought tolerance on rice plants. *The plant journal*, 23(3):319–327.

Sammons, R. D. and Gaines, T. A. (2014). Glyphosate resistance: state of knowledge. *Pest management science*, 70(9):1367–1377.

Schaefer, D. G. and Zrÿd, J.-P. (1997). Efficient gene targeting in the moss physcomitrella patens. *The plant journal*, 11(6):1195–1206.

Schramp, M., Hedman, A., Li, W., Tan, X., and Anderson, R. (2012). *PIP Kinases from the Cell Membrane to the Nucleus*, pages 25–59. Springer Netherlands, Dordrecht.

Segobye, K. (2013). *Biology and ecology of glyphosate-resistant giant ragweed (Ambrosia trifida L.)*. PhD thesis, PURDUE UNIVERSITY, West Lafayette, IN, USA.

Shah, D. M., Horsch, R. B., Klee, H. J., Kishore, G. M., Winter, J. A., Tumer, N. E., Hironaka, C. M., Sanders, P. R., Gasser, C. S., Aykent, S., Siegel, N. R., Rogers, S. G., and Fraley, R. T. (1986). Engineering herbicide tolerance in transgenic plants. *Science*, 233(4762):478–481.

Shaner, D. L. (2000). The impact of glyphosate-tolerant crops on the use of other herbicides and on resistance management. *Pest management science*, 56(4):320–326.

Sheen, J. (1996). Ca2 plus-dependent protein kinases and stress signal transduction in plants. *Science*, 274(5294):1900.

Shiu, S.-H. and Bleecker, A. B. (2001). Plant receptor-like kinase gene family: diversity, function, and signaling. *Science signaling*, 113(re22):1–13.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, page btv351.

Sinha, A. K., Jaggi, M., Raghuram, B., and Tuteja, N. (2011). Mitogen-activated protein kinase signaling in plants under abiotic stress. *Plant signaling & behavior*, 6(2):196–203.

Sonnhammer, E. L. L., Heijne, G. v., and Krogh, A. (1998). A hidden markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, ISMB '98, pages 175–182, Palo Alto, California, USA. AAAI Press.

Stachler, J. M. (2008). *Characterization and Management of Glyphosate-Resistant Giant Ragweed (Ambrosia trifida (L.) and Horseweed [Conyza canadensis (L.) Cronq.]*. PhD thesis, The Ohio State University, Columbus, OH, USA.

Stock, A. M., Robinson, V. L., and Goudreau, P. N. (2000). Two-component signal transduction. *Annual review of biochemistry*, 69(1):183–215.

Stone, J. M. and Walker, J. C. (1995). Plant protein kinase families and signal transduction. *Plant physiology*, 108(2):451–457.

Suh, H., Hepburn, A. G., Kriz, A. L., and Widholm, J. M. (1993). Structure of the amplified 5-enolpyruvylshikimate-3-phosphate synthase gene in glyphosate-resistant carrot cells. *Plant molecular biology*, 22(2):195–205.

Takahashi, F., Yoshida, R., Ichimura, K., Mizoguchi, T., Seo, S., Yonezawa, M., Maruyama, K., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2007). The mitogen-activated protein kinase cascade mkk3–mpk6 is an important part of the jasmonate signal transduction pathway in arabidopsis. *The plant cell*, 19(3):805–818.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003). The cog database: an updated version includes eukaryotes. *BMC bioinformatics*, 4(1):1.

Tenney, A. E., Brown, R. H., Vaske, C., Lodge, J. K., Doering, T. L., and Brent, M. R. (2004). Gene prediction and verification in a compact genome with numerous small introns. *Genome research*, 14(11):2330–2335.

Testa, A. C., Hane, J. K., Ellwood, S. R., and Oliver, R. P. (2015). Codingquarry: highly accurate hidden markov model gene prediction in fungal genomes using rna-seq transcripts. *BMC genomics*, 16(1):1.

Tuteja, N. (2007). Abscisic acid and abiotic stress signaling. *Plant signaling & behavior*, 2(3):135–138.

Unver, T., Bakar, M., Shearman, R. C., and Budak, H. (2010). Genome-wide profiling and analysis of festuca arundinacea mirnas and transcriptomes in response to foliar glyphosate application. *Molecular genetics and genomics*, 283(4):397–413.

Vivancos, P. D., Driscoll, S. P., Bulman, C. A., Ying, L., Emami, K., Treumann, A., Mauve, C., Noctor, G., and Foyer, C. H. (2011). Perturbations of amino acid metabolism associated with glyphosate-dependent inhibition of shikimic acid metabolism affect cellular redox homeostasis and alter the abundance of proteins involved in photosynthesis and photorespiration. *Plant physiology*, 157(1):256–268.

Wang, Z., Chen, Y., and Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics proteomics & bioinformatics*, 2(4):216–221.

Whitaker, J. R., Burton, J. D., York, A. C., Jordan, D. L., and Chandi, A. (2013). Physiology of glyphosate-resistant and glyphosate-susceptible palmer amaranth (amaranthus palmeri) biotypes collected from north carolina. *International journal of agronomy*, 2013.

Widholm, J. M., Chinnala, A., Ryu, J.-H., Song, H.-S., Eggett, T., and Brotherton, J. E. (2001). Glyphosate selection of gene amplification in suspension cultures of 3 plant species. *Physiologia plantarum*, 112(4):540–545.

Yang, K.-Y., Liu, Y., and Zhang, S. (2001). Activation of a mitogen-activated protein kinase pathway is involved in disease resistance in tobacco. *Proceedings of the national academy of sciences*, 98(2):741–746.

Yao, H., Guo, L., Fu, Y., Borsuk, L. A., Wen, T.-J., Skibbe, D. S., Cui, X., Scheffler, B. E., Cao, J., Emrich, S. J., Ashlock, D. A., and Schnable, P. S. (2005). Evaluation of five ab initio gene prediction programs for the discovery of maize genes. *Plant molecular biology*, 57(3):445–460.

Yoshida, R., Hobo, T., Ichimura, K., Mizoguchi, T., Takahashi, F., Aronso, J., Ecker, J. R., and Shinozaki, K. (2002). Aba-activated snrk2 protein kinase is required for dehydration stress signaling in arabidopsis. *Plant and cell physiology*, 43(12):1473–1483.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., Li, R., Xu, Z., Li, S., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., Geng, J., Li, G., Shi, J., Liu, J., Lv, H., Li, J., Wang, J., Deng, Y., Ran, L., Shi, X., Wang, X., Wu, Q., Li, C., Ren, X., Wang, J., Wang, X., Li, D., Liu, D., Zhang, X., Ji, Z., Zhao, W., Sun, Y., Zhang, Z., Bao, J., Han, Y., Dong, L., Ji, J., Chen, P., Wu, S., Liu, J., Xiao, Y., Bu, D., Tan, J., Yang, L., Ye, C., Zhang, J., Xu, J., Zhou, Y., Yu, Y., Zhang, B., Zhuang, S., Wei, H., Liu, B., Lei, M., Yu, H., Li, Y., Xu, H., Wei, S., He, X., Fang, L., Zhang, Z., Zhang, Y., Huang, X., Su, Z., Tong, W., Li, J., Tong, Z., Li, S., Ye, J., Wang, L., Fang, L., Lei, T., Chen, C., Chen, H., Xu, Z., Li, H., Huang, H., Zhang, F., Xu, H., Li, N., Zhao, C., Li, S., Dong, L., Huang, Y., Li, L., Xi, Y., Qi, Q., Li, W., Zhang, B., Hu, W., Zhang, Y., Tian, X., Jiao, Y., Liang, X., Jin, J., Gao, L., Zheng, W., Hao, B., Liu, S., Wang, W., Yuan, L., Cao, M., McDermott, J., Samudrala, R., Wang, J., Wong, G. K.-S., and Yang, H. (2005). The genomes of oryza sativa: a history of duplications. *PLoS biology*, 3(2):e38.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829.

Zhu, J.-K. (2000). Genetic analysis of plant salt tolerance using arabidopsis. *Plant physiology*, 124(3):941–948.

VITA

VITA

Education

- Ph.D. Biological Sciences (Computational Life Sciences), Purdue University, West Lafayette, IN - August 2016
- M.S. Bioinformatics, Indiana University, Bloomington, IN - June 2011
- B.Tech. Industrial Biotechnology, SASTRA University, Tamil Nadu, India - May 2009

Papers

- **Padmanabhan KR**, Segobye K, Weller SC, Schulz B, Gribskov M. Preliminary investigation of glyphosate resistance mechanism in giant ragweed using transcriptome analysis. F1000Research, 2016, 5:1354 (doi: 10.12688/f1000research.8932.1)

Patents

- B. Schulz, S.C. Weller, M. Gribskov, **K. Padmanabhan**, K. Segobye, "Diagnostic Tools for Herbicide Resistance in Weeds", Application No. 61/910,770 (Technology), 2013

Presentations

- **K. Padmanabhan**, K. Segobye, S.C. Weller, B. Schulz, M. Gribskov, Transcript- -ome Analysis of Giant Ragweed, Office of Interdisciplinary Graduate Programs Spring Reception, Purdue University, 2015 (Poster)
- **K. Padmanabhan**, K. Segobye, M. Gribskov, B. Schulz, S.C. Weller, Molecular Analysis of Glyphosate Resistance in Giant Ragweed, NCWSS Annual Meeting, Minneapolis, MN, December 3, 2014 (Oral)
- **K. Padmanabhan**, N.B. Best, M. Gribskov, S.C. Weller, B. Schulz, Transcriptome Analysis of Glyphosate Resistance in Giant Ragweed, Joint Annual Meeting of WSSA and CWSS, Vancouver, BC, February 4, 2014 (Oral)

Awards

- Purdue Graduate Student Government Travel Award, Purdue University, 2015
- Summer Institutes in Statistical Genetics Travel Award, University of Washington, 2011
- Student Innovator Award, Purdue University, 2014
- Best Oral Presentation, Department of Biological Sciences, Purdue University, 2014
- Dr. P. T. Gilham Graduate Award, Purdue University, 2011
- Academic Excellence Award, SASTRA University, 2006

Work Experience

- Graduate Administrative Professional, School of Environmental and Ecological Engineering, Purdue University, January 2015 to May 2016
- Data Science Intern, Monsanto Inc., May 2015 to August 2015

Teaching Experience

- Teaching Assistant, BCHM 695, Introduction to R and Bioconductor, Summer 2014
- Teaching Assistant, BIOL 203/204, Human Anatomy and Physiology, Fall 2012 to Spring 2013

Leadership

- President, Biology Graduate Student Council, Jul 2014 to Jun 2015
- Webmaster, Biology Graduate Student Council, Jul 2013 to Jun 2015
- Treasurer and Secretary, Biology Graduate Student Council, Jul 2012 to Jun 2014
- Webmaster, Society of Industrial and Applied Mathematics Aug 2012 to May 2013