INTEGRATIVE BIOINFORMATICS FOR KINASE-CONCENTRIC PHOSPHORYLATION NETWORKS IN *ARABIDOPSIS THALIANA*

by

Mengxi Lv

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics and Computational Biology

Fall 2014

© 2014 Mengxi Lv
All Rights Reserved
INTEGRATIVE BIOINFORMATICS FOR KINASE-CONCENTRIC
PHOSPHORYLATION NETWORKS IN ARABIDOPSIS THALIANA

by

Mengxi Lv

Approved:

________________________________
Cathy H. Wu, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved:

________________________________
Errol Lloyd, Ph.D.
Chair of the Department of Computer and Information Sciences

Approved:

________________________________
Bahatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved:

________________________________
James G. Richards, Ph.D.
Vice Provost for Graduate and Professional Education
ACKNOWLEDGMENTS

I wish to thank my advisors Dr. Wu and Dr. Lee, and my committee member Dr. Cecilia Arighi for their continuous advice, guidance, and academic support during my master study. Especially, I would like to thank the postdoctoral researcher Dr. Karen Ross for many helpful discussions on my master research work. Also, I must thank my professional friends Jia Ren, Irem Celen, Zhiwen Li, and Qili Fei who have helped me throughout my master education.

I gratefully acknowledge the funding sources Grant No.1062520 from National Science Foundation and Grant No.R01GM080646 from National Institutes of Health, which support my thesis work.

This manuscript is dedicated to my parents Shuxian Lv and Yue’e Yao, and my younger brother Zhaoxi Lv for their unconditional love. (这篇论文献给我的父母吕书贤和姚月娥，和我的弟弟吕兆熙，感谢他们对我一直以来无尽的爱。)
TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................ vi
LIST OF FIGURES .................................................................................................... viii
ABSTRACT .................................................................................................................. x

Chapter

1 LITERATURE REVIEW ............................................................................................. 1
  1.1 Post-translational Modifications (PTMs) .............................................................. 1
  1.2 Plant Protein Kinase Family ................................................................................. 2
  1.3 Major Resources for PTMs, and Their Limitations ........................................... 3
  1.4 RLIMS-P and Protein Ontology ......................................................................... 9

2 INTRODUCTION ....................................................................................................... 12
  2.1 PID- and PDK1-regulated Signaling Pathways .................................................... 13
  2.2 BR Signaling Protein Phosphorylation Network ................................................. 13
  2.3 Project Summary .................................................................................................. 16

3 METHODS .................................................................................................................. 18
  3.1 Overview of the Workflow ................................................................................ 18
  3.2 Assessment of Plant Protein Phosphorylation Information .................................. 19
  3.3 Text mining of Phosphorylation Events with RLIMS-P .................................... 20
  3.4 Creation and Annotation of PRO Terms ............................................................. 21
  3.5 Data Mining of Phosphorylation Information ................................................... 22
  3.6 Visualization and Analysis of Integrated Phosphorylation Data ....................... 22
  3.7 Analysis of the BR-signaling Phosphorylation Network with External Tools .... 22

4 RESULTS .................................................................................................................... 24
  4.1 Assessment of the Availability of Protein Phosphorylation Information for Plants ........................................................................................................................... 24
  4.2 Identification of the Top 15 Most Literature-documented Kinase Proteins in Arabidopsis Using RLIMS-P .............................................................. 27
  4.3 PDK1 Phosphorylates PID to Regulate Auxin Polar Transport ....................... 31
4.4 Construction of BR-signaling Protein Phosphorylation Network
Integrating Text mining and Data mining ............................................ 32

4.4.1 Integration of text- and data-mining phosphorylation
information on BR signaling .......................................................... 32
4.4.2 BR-signaling protein phosphorylation knowledge map .......... 34
4.4.3 Functional analysis of BR-signaling protein phosphorylation
network ......................................................................................... 36
4.4.4 Phosphorylation of BSK5 by BIN2 in response to salt stress .... 38
4.4.5 SERK proteins integrate BR signaling pathway with other
biological processes ..................................................................... 40
4.4.6 Conservation of BR signaling pathway across plant species ..... 41

5 DISCUSSION AND FUTURE WORK ............................................. 49

5.1 Discussion ............................................................................... 49
5.2 Future Work ............................................................................ 54

REFERENCES ................................................................................. 56

Appendix

A MULTIPLE SEQUENCE ALIGNMENTS OF THE ARABIDOPSIS
ORTHOLOGS OF 8 BR-SIGNALING KEY COMPONENTS (BRI1,
BAK1, BSK1, CDG1, BUS1, BIN2, BZR1, AND BZR2) ....................... 63
LIST OF TABLES

Table 4.1 Numbers of substrates, phosphosites, kinases, and scientific articles (PMIDs) captured by three databases (UniProtKB, PhosPhAt, P3DB) and RLIMS-P. ................................................................. 25

Table 4.2 Numbers of phosphoproteins, phosphosites and PMIDs available in recombined databases and literature for 7 plant organisms .............. 26

Table 4.3 Numbers of kinase protein names identified by RLIMS-P for the plant model organisms................................................................. 27

Table 4.4 Total scientific publications (PMIDs) in PubMed, RLIMS-P-positive PMIDs, kinase proteins, substrate proteins and phosphosites identified by RLIMS-P. ................................................................. 28

Table 4.5 Summary table of PMIDs identified by RLIMS-P and PMIDs annotated in PRO for the top 15 kinases in Arabidopsis. .................................. 28

Table 4.6 Classification of the top 15 most occurring kinases in Arabidopsis ....... 29

Table 4.7 Summary numbers of BR-signaling protein phosphorylation knowledge map data .................................................................................. 33

Table 4.8 Phosphorylation-dependent PPIs in BR-signaling phosphorylation network ......................................................................................... 36

Table 4.9 Potential phosphorylation sites of BSK5 by BIN2............................. 39

Table 4.10 Conservation of phosphorylation-specific protein forms across 6 plant species. ......................................................................................... 42

Table 4.11 BLAST result for the orthologs of Arabidopsis BRI1 protein across 6 species ......................................................................................... 44

Table 4.12 BLAST result for the orthologs of Arabidopsis BAK1 protein across 6 species. ......................................................................................... 46

Table 5.1 The top 15 most curated kinase proteins identified by data mining........ 51
Table A.1 BLAST result for the orthologs of *Arabidopsis* BSU1 protein across 6 species ................................................................. 63

Table A.2 BLAST result for the orthologs of *Arabidopsis* BSU1 protein across 6 species ....................................................................... 64

Table A.3 BLAST result for the orthologs of *Arabidopsis* BIN2 protein across 6 species ....................................................................... 65

Table A.4 BLAST result for the orthologs of *Arabidopsis* BZR1 protein across 6 species ....................................................................... 65

Table A.5 BLAST result for the orthologs of *Arabidopsis* BZR2 protein across 6 species ....................................................................... 66
LIST OF FIGURES

Figure 1.1 PhosPhAt showing the sample view of substrate-phosphosite information of BIN2 (At4g18710.1) (http://phosphat.uni-hohenheim.de) ................................................................. 5

Figure 1.2 P3DB showing the sample view of kinase-substrate relationships of CPK21 (AT4G04720.1; http://www.p3db.org/substrateNetwork.php?type=kinase) ................................................................. 6

Figure 1.3 PlantsP showing the sample view of functional annotations of BRI1 (plantsp ID: 21289) (http://plantsp.genomics.purdue.edu/cgi-bin/detail.cgi?id=21289) ................................................................. 7

Figure 1.4 UniprotKB showing the sample view of protein phosphosite information of BZR2 (Q9LN63) (http://www.uniprot.org/uniprot/Q9LN63) ....... 8

Figure 1.5 PRO entry showing the sample view of unmodified protein form and GO annotation of At-BZR2/Phos:1 (PR:000035391; http://pir.georgetown.edu/cgi-bin/pro/entry_pro?id=PR:000035391) .... 10

Figure 1.6 PRO entry showing the sample view of phosphorylated protein form and GO annotation information of At-BZR2/Phos:1 (PR:000035391; http://pir.georgetown.edu/cgi-bin/pro/entry_pro?id=PR:000035391) .... 11

Figure 1.7 PRO entry showing the sample view of phosphorylated protein form and GO annotation information of At-BZR2/Phos:2 (PR:000035891; http://pir.georgetown.edu/cgi-bin/pro/entry_pro?id=PR:000035891) ... 11

Figure 2.1 A model of BR signaling pathway adopted from a review article (Kim et al., 2009) .................................................................................................................................................. 15

Figure 3.1 Work flow for construction of the protein phosphorylation knowledge map .................................................................................................................................................. 19

Figure 4.1 PDK1 phosphorylated PID participating in PID-regulated auxin polar transport process ............................................................................................................................................ 32

Figure 4.2 Network visualization of the BR-signaling phosphorylation knowledge map integrating text mining and data mining results. ......................... 35
Figure 4.3 Functional sub-network of BR-signaling phosphorylation knowledge map. ................................................................. 38

Figure 4.4 The BR signaling pathway cross talks with the salinity stress response through phosphorylation BSK5 by BIN2. ................................................ 39

Figure 4.5 Sub-network of SERKs family proteins involved in multiple biological processes. ................................................................. 41

Figure 4.6 Multiple sequence alignment of the orthologs of Arabidopsis BRI1 protein across 6 plants species ................................................................. 45

Figure 4.7 Multiple sequence alignment of BAK1 across 6 plants species. ............... 47

Figure A.1 Multiple sequence alignment of the orthologs of Arabidopsis BSK1 across 6 plants species ........................................................................ 63

Figure A.2 Multiple sequence alignment of the orthologs of Arabidopsis BSU1 across 6 plants species ........................................................................ 64

Figure A.3 Multiple sequence alignment of the orthologs of Arabidopsis BIN2 across 6 plants species ........................................................................ 65

Figure A.4 Multiple sequence alignment of the orthologs of Arabidopsis BZR1 across 6 plants species ........................................................................ 66

Figure A.5 Multiple sequence alignment of the orthologs of Arabidopsis BZR2 across 6 plants species ........................................................................ 67
ABSTRACT

Phosphorylation, mediated by various types of protein kinases, is one of the most well studied mechanisms regulating signal transduction. The extensive studies produce a large amount of phosphorylation data, which are scattered and distributed in scientific articles and databases. However, lack of integrative PTM resources, especially in plants, makes it challenging for biological scientists to connect those fragmented data for systematic understanding and global analysis of PTMs. Here we developed an integrative bioinformatics approach combining text mining, data mining, protein ontology, visualization, and analysis methods, thereby facilitating PTM information integration and knowledge discovery in plants.

For *Arabidopsis thaliana*, we identified a list of the top 15 most literature-documented kinase names through RLIMS-P, a text-mining tool to extract phosphorylation information. The resulting kinase list included PDK1 and PID, which are two kinases involved in a wide range of cellular activities and auxin polar transport respectively. Text-mining results demonstrated that phosphorylation of PID by PDK1 is a primary step to activate PID kinase’s regulatory role in auxin polar transport.

Moreover, given that 3 key Brassinosteroid (BR)-signaling components, BRI1, BAK1 and BIN2, were in the kinase list, we assume that BR-signaling pathway is enriched with phosphorylation events. BR is a plant steroid hormone essential not only for plant growth and development but also for the immune response. Previous studies have shown that BR directly binds to receptor-like kinase BRI1 and its co-receptor BAK1; and consequently activates the downstream signal transduction components
including BSK1, CDG1, BSU1, BIN2, BZR1 and BZR2. Those 8 key components are connected through a series of phosphorylation events, leading to the expression of genome-wide BR-responsive genes. Thus, we combined text mining and data mining data to construct a BR-signaling protein phosphorylation network. The combined data showed that the BR-signaling pathway was interconnected with multiple pathways such as the defense-, stress- and auxin-regulatory networks. Specifically, we hypothesized that BIN2 phosphorylated BSK5 on S329 was in response to salt stress; and the interplay of BRI1 with SERK family proteins was involved in the regulation of anther development, cell death and innate immunity.

This thesis work demonstrates that our integrative approach facilitates the integration of discrete and fragmented data and construction of networks enriched with phosphorylation information, thereby prompting hypothesis generation and knowledge discovery.
Chapter 1

LITERATURE REVIEW

1.1 Post-translational Modifications (PTMs)

PTMs play a fundamental role in a wide range of biological processes. Post-translational modification (PTM) refers to chemical alteration of translated proteins, typically through a modification at a specific amino acid residue(s) (Garavelli, 2004). Among a variety of PTMs known, phosphorylation, glycosylation, acetylation, myristoylation and ubiquitination are the most-well studied in plant systems (Podell and Gribskov, 2004; Wolschin et al., 2005; Chen and Tian, 2007; Fitchette et al., 2007; Chen and Sun, 2009; Choudhary et al., 2009). Those PTMs often work in a concerted manner to modulate the function, activity and cellular localization of a wide range of proteins, leading to complex biological responses (Copeland et al., 2008; Ray, 2009). For example, the involvement of phosphorylation in GlcNAcylation plays a critical role in insulin resistance, glucose toxicity and other important biological process (Copeland et al., 2008).

Phosphorylation, one of the most widespread protein modifications, is a reversible process controlled by the interplay between kinases and phosphatases. The process of phosphorylation involves the covalent attachment of phosphate to the hydroxyl group of specific serine, threonine or tyrosine residue(s), which modulates the properties of the protein (Champion et al., 2004). Phosphorylation is a fundamental cellular regulatory mechanism, by which a number of signal transduction pathways can be initiated and controlled (Wang et al., 2005; Yoshida and Parniske, 2005). For
example, the BR-signaling pathway, a case in this thesis work, plays an essential role not only for plant growth and development but also for immune responses. This process is regulated by a series of phosphorylation and dephosphorylation events, leading to the expression of BR-responsive genes (Wang et al., 2005).

1.2 Plant Protein Kinase Family

The genome of human, a representative example of mammalian system, encodes 518 kinases, which is around 2.5% of the total number of protein-encoding genes (20,687) in human genome (Manning et al., 2002; Consortium et al., 2012). Similarly, yeast possesses 129 (2.1%) kinase proteins out of a total 6,217 protein-encoding genes (Hunter, 2000; Breitkreutz et al., 2010). Comparatively, Arabidopsis thaliana genome encodes 989 serine/threonine kinases, constituting a remarkable higher percentage (3.9%) of its genome that contains 25,498 protein-encoding genes (Arabidopsis Genome, 2000; Champion et al., 2004). This reveals the prevalence of kinases and potentially significant roles of phosphorylation in Arabidopsis.

Plant protein kinome consists of around 600 receptor-like protein kinases (RLKs) and 400 non-receptor-like protein kinases (non-RLKs). For RLKs group, its phylogenetic relationships with plants non-RLKs and mammalian RLKs have been established (Shiu and Bleecker, 2001). In addition, a larger number of studies are conducted to characterize the diverse biological functions of non-RLKs through genetic approach knocking out specific kinase genes (Sugano et al., 1998; Benjamins et al., 2003). Based on sequence alignment of protein kinase catalytic domains, plant kinases have been classified into 5 major groups: (1) Calcium-dependent protein kinases and SNF1-related Protein Kinase (SnRK); (2) Mitogen-activated kinases (MAPK); (4) AGC kinases; (5) Cyclin-dependent protein kinases; and the (6) Shaggy-
like/GSK3 kinases (Hanks et al., 1995; (Zulawski et al., 2013). In Arabidopsis, the kinome are further classified into 5 classes, 27 groups, and 80 families (Tchieu et al., 2003).

Many plant kinases such as MAPK, GSK3, CDK, and Casein Kinase II (CK2) families are universal components in eukaryotes. Take MAPK family as an illustrative example. In yeast, animals and plants, MAPK cascade is a universal module regulating the intracellular transduction and amplification of extracellular signal, leading to the expression of the target genes. In this process, a series of phosphorylation events are involved: a MAPK kinase kinase (MAPKKK) phosphorylates the MAPK kinase (MAPKK), which subsequently activates the MAPK to control the transcription factors of the responsive genes (Yang et al., 2003). The responsive genes regulate a wide range of biological processes, including physiology, development and growth, and stress (Yang et al., 2003). However, there also exist plant-specific kinases distinct from those identified in animals or bacterium, like RLKs, and CDPKs as well as kinases that are absent from plants, such as the cyclic nucleotide-dependent protein kinases and conventional PTKs.

1.3 Major Resources for PTMs, and Their Limitations

Advances in phosphoproteomic experimental techniques provide us the ability to assess the prevalence of phosphorylation throughout the whole proteome, which is based on the large scale and high-throughput mass spectrometry phosphorylation experiments. The expanding volume of such data is increasingly being captured by bioinformatics databases, most of which focus on mammalian and bacterium systems. PhosphoSitePlus (Hornbeck et al., 2004) is the most comprehensive PTM resource, primarily covering human and mouse. Its manually curated PTM data consists of
phosphorylation, acetylation, methylation, ubiquitination and O-glycosylation information from literature and other data sources. Other public databases such as PHOSIDA (Gnad et al., 2007), Phospho.ELM (Diella et al., 2008), and PhosphoPep (Bodenmiller et al., 2008) mainly provide phosphorylation site information for human, yeast, *E.coli*, fly and other model organisms but no plant species.

Several resources exist for curating plant phosphorylation information. PhosPhAt (Durek et al., 2010), an *Arabidopsis thaliana* database, contains information covering phosphorylation sites and kinase-substrate relationships (lacking mapped phosphosite information) (Figure 1.1). P3DB (Gao et al., 2009) has the largest collection of plant information from several model plant organisms and covers phosphorylation sites and kinase-substrate relationships (lacking mapped phosphosite information) (Figure 1.2). PlantsP (Tchieu et al., 2003) contains phosphorylation data mainly on *Arabidopsis*, but its focus is on the functional annotation of protein kinases and phosphatases (Figure 1.3). UniProt Knowledgebase (UniProtKB) (Wu et al., 2006; UniProt, 2014) provides comprehensive information on proteins from a much wider range of organisms, but it does not specialize in the curation of phosphorylation information (Figure 1.4). In addition, it’s important to note that PhosPhAt and P3DB databases curate phosphosite information from the results of medium- to large-scale proteomic phosphorylation mass spectrometry studies; while the phosphosite information curated by UniprotKB database is curated from small-, medium- and large-scale phosphorylation experiments.
Figure 1.1 PhosPhAt showing the sample view of substrate-phosphosite information of BIN2 (At4g18710.1) (http://phosphat.uni-hohenheim.de)
Figure 1.2 P3DB showing the sample view of kinase-substrate relationships of CPK21 (AT4G04720.1; http://www.p3db.org/substrateNetwork.php?type=kinase)
Figure 1.3 PlantsP showing the sample view of functional annotations of BRI1 (plantsp ID: 21289) (http://plantsp.genomics.purdue.edu/cgi-bin/detail.cgi?id=21289)
Figure 1.4 UniprotKB showing the sample view of protein phosphosite information of BZR2 (Q9LN63) (http://www.uniprot.org/uniprot/Q9LN63).

In the above-mentioned plant resources, critical gaps remain for systematic study of PTMs. Those gaps include: (1) much of the scientific knowledge about phosphorylation is still buried in large and rapid expanding volumes of literature, which leads to a disconnection between information in existing databases and enriched information available in scientific literature; (2) the resources for plant PTM do not integrate kinase, substrate and phosphosite information in the specific biological context; (3) the resources for plant PTM do not connect PTMs across taxons for comparative analysis and knowledge discovery; and (4) the large amount of information available in the literature and public databases is fragmented and not integrated into a form that can be easily deciphered for knowledge discovery. Due to
those gaps, our understanding of phosphorylation events is still fragmented, which limits our global understanding of PTM-regulatory mechanisms for novel knowledge discovery in biology systems.

1.4 RLIMS-P and Protein Ontology

To address above-mentioned gaps, we used the text-mining tool RLIMS-P (Rule-based Literature Mining System for Protein Phosphorylation; http://research.bioinformatics.udel.edu/rlimsp/) to extract phosphorylation information from articles. The retrieved information was annotated in Protein Ontology (PRO, http://pir.georgetown.edu/pro/pro.shtml) describing the PTM-specific protein form/complex and its associated functional annotations.

RLIMS-P is a text-mining tool specifically designed to extract protein phosphorylation information on protein kinase, substrate and phosphorylation site from biomedical literature (Hu et al., 2005). Keyword and PMID could be used to search with RLIMS-P, and the result table displays the extracted kinases, substrates, and sites from each abstract. When user entering the page for one abstract, it shows a list of suggested Uniprot KB IDs for the extracted kinases and substrates as well as the highlights of the mention of kinases, substrates and sites as evidence.

Protein Ontology a newly developed bioinformatics resource, describes diverse protein forms, e.g. forms post-translationally modified on specific amino acid residues, as well as their associated functional annotations attributes (Natale et al., 2011; Natale et al., 2014). The ontology term defines protein form/complex by UniprotKB identifiers, PTM types and sites, modifying enzyme, and evidence source. The annotation denotes the PTM-state specific attributes using Gene Ontology terms, functional domains using Pfam IDs, related diseases using OMIM IDs, as well as
sequence types and features using SO terms. Figure 1.5 shows the PRO term of unmodified protein form and GO annotations of *Arabidopsis BZR2*. Figure 1.6 shows the phosphorylated form of BZR2 by ASK7 on specific phosphosites. Figure 1.7 shows the phosphorylated form BZR2 of Thus, PRO precisely represents protein forms and complexes and denotes them with functional annotations facilitating accessibility to research results on protein PTMs.

Figure 1.5 PRO entry showing the sample view of unmodified protein form and GO annotation of At-BZR2/Phos:1 (PR:000035391; http://pir.georgetown.edu/cgi-bin/pro/entry_pro?id=PR:000035391)
Figure 1.6 PRO entry showing the sample view of phosphorylated protein form and GO annotation information of At-BZR2/Phos:1 (PR:000035391; http://pir.georgetown.edu/cgi-bin/pro/entry_pro?id=PR:000035391)

Figure 1.7 PRO entry showing the sample view of phosphorylated protein form and GO annotation information of At-BZR2/Phos:2 (PR:000035891; http://pir.georgetown.edu/cgi-bin/pro/entry_pro?id=PR:000035891)
Phosphorylation, mediated by various types of protein kinases, is one of the most well studied mechanisms regulating signal transduction. Phosphorylation data produced by both traditional genetic studies on specific proteins and newly emerged high-throughput experiments on the whole proteome have been increasing at a rapid pace. Given the wealth of such data scattered and distributed in scientific articles and databases, one particular challenge for biological scientists is to connect those fragmented data for systematic understanding and global analysis of the protein phosphorylation network. In this article, we developed an integrative bioinformatics approach combining text mining, data mining, Protein Ontology, visualization, and analysis methods, thereby facilitating PTM information integration and knowledge discovery in plants.

The integrative approach involves: (1) text mining via RLIMS-P, a tool that retrieves phosphorylation information from literature; (2) representing phosphorylated protein forms and complexes with their functional attributes in the Protein Ontology (PRO); (3) data mining of phosphorylation information from experiment- and prediction-based databases; (4) visualization of data integrated from text mining and data mining as a knowledge map; (5) comparative analysis of the knowledge map for hypothesis generation and novel knowledge discovery. Using the integrative approach described above, we developed two case studies focused on PID-PDK1 and the BR signaling pathway to demonstrate the application of this approach for hypothesis generation and knowledge discovery.
2.1 PID- and PDK1-regulated Signaling Pathways

The plant hormone auxin is an essential regulator in plant growth and development. Auxin is transported in a polar fashion concentrating in apex and root, which determines multiple plant developmental processes, such as vascular differentiation (Friml and Palme, 2002). Previous studies demonstrated the regulatory kinase role of PINOID (PID) in phosphorylation and the asymmetric distribution of the PINFORMED (PIN) family of proteins, which are carrier proteins of auxin (Friml et al., 2004). Additionally, loss-of-function mutants pid mutants, defective in initiating lateral organ development, can be partially rescued by local application of auxin, which suggests the positive regulatory role of PID in auxin polar transport process (Reinhardt et al., 2000).

The 3-phosphoinositide-dependent protein kinase-1 (PDK1) is an essential protein kinase regulator in mammals, which phosphorylates a number of targets involved in various physiological processes, such as cell proliferation, cell differentiation and cell death. In Arabidopsis, AtPDK1 was characterized as an ortholog of PDK1; and the autophosphorylation of AtPDK1 has been observed, which activates AtPDK1kinase activity to modify its downstream targets (Otterhag et al., 2006). However, in plants, PDK1-regulated downstream targets and signaling pathways are still largely unknown.

2.2 BR Signaling Protein Phosphorylation Network

A wide range of biological processes are involved in BR-signaling pathway, in which phosphorylation is the major mechanism transducing hormone signal from membrane receptor kinase to nuclear transcription factors. As Figure 2.1 displays, external hormone BR binds to the cell surface kinase receptor BRASSINOSTEROID
INSENSITIVE 1 (BRI1) (Hothorn et al. 2011; Kinoshita et al., 2005; Li et al., 1997; She et al., 2011; Wang et al., 2001), and activates BRI kinase activity (Wang et al., 2001). BRI1 then associates with and transphosphorylates the coreceptor BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase 1 [BAK1, also named SOMATIC EMBRYOGENESIS RECEPTOR KINASE 3 (SERK3)] (Li et al., 2002; Nam et al., 2002; Wang et al., 2008), which leads the BR signaling transduction to switch on. BRI1 phosphorylates BR-SIGNALING KINASE 1 (BSK1) and CONSTITUTIVE DIFFERENTIAL GROWTH 1 (CDG1) (Kim et al., 2011; Tang et al., 2008), which subsequently activate BRI1-SUPPRESSOR 1 (BSU1) (Kim et al., 2011; Kim et al., 2009; Mora-Garcia et al., 2004). BSU1 dephosphorylates and inactivates BRASSINOSTEROID INSENSITIVE 2 (BIN2) (Kim et al., 2011; Li et al., 2002). When BR levels are low, BIN2 is active to phosphorylate and repress two homologous transcription factors BRASSINAZOLERESISTANT 1 (BZR1) and BRASSINAZOLERESISTANT 2 [BZR2, also named BRI1-EMS-SUPPRESSOR 1 (BZR2)] BZR1 and BZR2 (He et al., 2002; Wang et al., 2002; Yin et al., 2002). When BR levels are high, BIN2 is inactivated, leading to the rapid dephosphorylation of BZR1 and BZR2 by protein phosphatase 2A (PP2A) (Tang et al., 2011). Thus, transcription factors BZR1 and BZR2 move into the nucleus, and bind to the genomic DNA to regulate BR-responsive gene expression (He et al., 2005; Sun et al., 2010; Yin et al., 2005).
Figure 2.1 A model of BR signaling pathway adopted from a review article (Kim et al., 2009). The solid arrows represent phosphorylation and activation; solid lines represent dephosphorylation and suppression; and dashed lines represent suppression relieved by pathway activation.
In the experimental organism *Arabidopsis*, the BR signaling pathway is well studied; however, few studies have been conducted on crop plants. The BR biosynthetic and insensitive mutants have been identified in tomato (Bishop et al., 1999; Koka et al., 2000), rice (Yamamuro et al., 2000), barley (Chono et al., 2003), and pea (Nomura et al., 1997). In addition, productive forward and reverse genetics studies in rice (*Oryza sativa*) have identified a number of BR-signaling components that are orthologous to *Arabidopsis* proteins as well as some that are rice-specific (Yamamuro et al., 2000; Bai et al., 2007; Koh et al., 2007; Li et al., 2009; Tanaka et al., 2009).

Recently, in rice autophosphorylation of osBRI1 has been observed *in vitro* (Zhao et al., 2013); and the BR-induced dephosphorylated osBZR1 was observed to have an decreased interaction with the shuttling protein 14-3-3 protein, leading to the movement of BZR1 into nucleus for regulating BR-responsive genes expression (Bai et al., 2007). This study result was as the same as the dephosphorylation impact in *Arabidopsis* (Zhang et al., 2012). Moreover, in tomato, autophosphorylation and transphosphorylation sites of SIBRI1 and SIBAK1 have been identified by large-scale mass spectroscopy experiment (Bajwa et al., 2013); and site-directed mutagenesis studies revealed the functional roles of the either conserved or tomato-specific phosphosites. Those facts suggest that BR-signaling pathway may primarily conserved across plants species, and more experimental studies, especially for phosphorylation events, are in need.

### 2.3 Project Summary

In this article, we described the identification of the top 15 most literature-documented kinase proteins using the text-mining tool RLISM-P. The textual
information of kinase-substrate relationships and their functional impacts was curated and annotated in the PRO framework. The text-mining results demonstrated that PDK1 could phosphorylate PID and activate PID kinase’s regulatory role in auxin polar transport. Next, we used BR signaling as a case study to demonstrate the application of integrating text mining and data mining to the construction of a protein phosphorylation knowledge map. The integrative approach captured kinase-substrate and protein-protein interaction (PPI) relationships for the 8 key BR-signaling components. In the resulting BR-signaling protein phosphorylation network, we predicted that phosphorylation events involving BSK5 and SERKs contributed to the interconnections between BR signaling and other signal pathways, such as the salinity stress- and male sporogenesis- regulatory networks. Our understanding of these signal transduction pathways would greatly benefit from the integration of the phosphorylation and its functional impacts information available in scientific literature and public databases.
Chapter 3

METHODS

3.1 Overview of the Workflow

A protein phosphorylation knowledge map was constructed by integrating bioinformatics techniques and resources including text mining, Protein Ontology, data mining, and visualization and analysis tools (Figure 3.1). First, we assessed the current state of phosphorylation events available in databases and literature in plants. Second, phosphorylation related articles restricted to Arabidopsis were collected through an RLIMS-P search. Those articles containing the top 15 most mentioned kinase proteins were curated retrieving protein forms, complexes and functional attributes, which were entered into Protein Ontology (PRO) via the RACE-PRO interface. Third, we developed 2 case studies on PID-PDK1 and BR-signaling pathway. In the case of BR-signaling pathway, we further combined experimental kinase-substrate data from PhosPhAt database as well as PPI relationships from PhosPhAt, STRING, IntAct, and BioGriD databases for the 8 key BR-signaling components. The combined data was visualized as a network for hypothesis generation of BR-signaling phosphorylation events. Last, comparative analysis of BR-signaling pathway within Arabidopsis and across plant species was performed by PRO and DAVID database as well as BLAST program. Details of each step of this workflow are presented in the following sections.
3.2 Assessment of Plant Protein Phosphorylation Information

Plant-specific data sets from PhosPhAt, P3DB, and UniProtKB databases were integrated by mapping proteins from PhosPhAt and P3DB databases to unique UniprotKB entries. The numbers including substrates, kinases, phosphorylation sites and PMIDs were calculated. The number of substrates with kinase information was calculated based on PhosPhAt-curated (http://phosphat.uni-hohenheim.de) kinase-
target information, of which the interactions tagged with phosphorylation or autophosphorylation experimental evidence were taken into consideration.

In addition, we used a Python script (available upon request) to extract the above data for plant species including: *Arabidopsis thaliana, Oryza sativa* (rice), *Medicago truncatula* (Barrel Medic), *Zea mays* (maize), *Nicotiana tabacum* (tobacco), *Solanum lycopersicum* (tomato), and *Solanum tuberosum* (potato) as well as other 9 species including *Capsicum annuum, Chlamydomonas reinhardtii, Chlorella fusca, Hordeum vulgare, Lotus japonicus, Medicago sativa, Sinapis alba, Sorghum bicolor, Spinacia oleracea* (combined as one row “Other”).

To assess the plant-specific phosphorylation information available in literature, we collected a set of PubMed abstracts that are tagged with the Mesh term “Plants” and used RLIMS-P to extract phosphorylation information, including kinases, substrates, and phosphosites. Furthermore, We assessed the protein phosphorylation information available in literature for the above-mentioned plant species. For each species, we used the scientific names, authority names, synonym names, and common names as keywords searching with RLIMS-P. The returned RLIMS-P report table was downloaded for counting the distinct substrates, kinases, sites, sites with kinase information and PMID.

3.3 Text mining of Phosphorylation Events with RLIMS-P

We collected the phosphorylation information of the top 15 most frequently documented *Arabidopsis* kinase proteins using RLIMS-P (http://research.bioinformatics.udel.edu/rlimsp/). The literature containing phosphorylation information restricted to *Arabidopsis thaliana* was obtained by
searching RLIMS-P with keyword “Arabidopsis” and by setting filters on the search page to include only those documents relevant to the organism “Arabidopsis thaliana”. Given the returned result table, we used a Python script (available upon request) to identify the 15 most frequently occurring kinase names and extract their phosphorylation information including kinase, substrate, phosphosites and text evidence. We mapped the top 15 kinase protein names to UniprotKB identifiers based on literature context. Of the top 15 kinase protein names, CK2 and Casein Kinase II refer to the same complex so we combined these results under the name CK2; in addition mapk represents a protein kinase family but all of its RLIMS-P positive articles focused on MPK3 and MPK6. As MPK6 was already in the list of top 15 kinase protein names, we replaced mapk with MPK3.

3.4 Creation and Annotation of PRO Terms
As in a previous study (Ross et al., 2013), we curated the full-length articles for the above 15 kinase proteins and retrieved phosphorylation events (kinase, substrate and phosphosites) and their functional impact (e.g. interacting partners, protein activity, subcellular localization, pathways). All retrieved information was entered into RACE-PRO (Rapid Annotation interfaCE for PRotein Ontology; http://pir.georgetown.edu/cgi-bin/pro/race_pro), a web interface for users to define and annotate a Protein Ontology term. In the definition subsection, users can enter the protein form with its sequence, PTM types and sites, modifying enzyme, and evidence source. In the annotation subsection, users can add functional annotations using GO terms, functional domains using Pfam IDs, related diseases using OMIM IDs, as well as sequence types and features using SO terms. In our study, as we focus on the study of the functional impact of phosphorylation, we only add GO terms in the functional
annotation subsection. The protein entries in RACE-PRO framework were used to generate PRO terms. All terms (OBO file) and annotations (PAF file) generated in this study can be found in PRO release 40.1.

3.5 Data Mining of Phosphorylation Information

The experimental kinase-substrate relationships of 8 key BR-signaling components were retrieved from PhosPhAt; and the experimental PPIs were retrieved from the PhosPhAt, STRING (Franceschini et al., 2013), IntAct (Orchard et al., 2014), and BioGriD (Chatr-Aryamontri et al., 2013) databases. We focused exclusively on physical interactions documented in experiments and filtered out PPIs predicted using techniques such as spoke expansion, coexpression, and genetic association.

Phosphorylation site data for the 8 BR-signaling key proteins were retrieved from UniprotKB, PhosPhAt, P3DB and the PRO PAF file. Predicted phosphorylation sites were excluded.

3.6 Visualization and Analysis of Integrated Phosphorylation Data

We combined the phosphorylation information from text- and data-mining efforts in an Excel file and excluded the redundant data-mining results that were captured by text mining. As in a previous study (Ross et al., 2013), the integrated phosphorylation events were visualized as a network by Cytoscape version 3.0.0 (Saito et al., 2012).

3.7 Analysis of the BR-signaling Phosphorylation Network with External Tools

We clustered the nodes (proteins/genes) based on DAVID (Database for Annotation, Visualization and Integrated Discovery) (Huang da et al., 2009a, b) functional clustering results and PRO functional annotations. To avoid duplicate
classification and thus simplify the analysis, we manually rebinned the GO terms of all the nodes such that each node was assigned solely to its best functional sub-network.

Reciprocal best-hit BLAST approach was used for identification of orthologs of *Arabidopsis* 8 BR-signaling key proteins in other crop plants, including rice, tomato, potato, maize and *Glycine max* (soybean). For each of the *Arabidopsis* 8 key proteins, we used BLAST (http://www.uniprot.org/blast/) searching its sequences against all sequences in the UniprotKB database; then we compared the sequence from the target organism with the maximal pairwise blast score against the UniprotKB database. If the best hit in this search was the original *Arabidopsis* sequence, then we defined the candidate sequence as an ortholog. Note that the orthologs to AtBSL1, the closest homolog to AtBSU1 (Mora-Garcia et al., 2004), were also considered as the orthologs to AtBSU1. SIBRI1 and SIBAK1 in tomato as well as OsBRI1, OsBAK1, OsGSK1 and OsBZR1 in rice were identified from published articles (Bai et al., 2007; Koh et al., 2007; Li et al., 2009; Bajwa et al., 2013; Zhao et al., 2013) instead of BLAST approach.
Chapter 4

RESULTS

4.1 Assessment of the Availability of Protein Phosphorylation Information for Plants

To evaluate the availability of plants phosphorylation information in databases and scientific literature, we summarized the phosphorylation data by integrating the data sets from databases of UniprotKB, P3DB and PhosPhAt as well as by mining the plant-related articles with RLIMS-P. Summarized as Table 4.1, RLIMS-P extracted 3,880 substrates and 601 phosphosites from 4,749 plant-related articles. In contrast, the combined databases curated a small set of articles (109 PMIDs) extracting a larger number of substrates (10,311) and phosphosites (27,711). PhosPhAt and P3DB databases, specialized in curating the results from whole phosphoproteomic experiments, exclusively contributed 92.51% and 94.78% to the total data of substrates and the sites, which reflects the contribution of high-throughput proteomic phosphorylation study. However, limited kinase information is integrated in the combined databases. Only 5.6% (574/10,311) of the total substrates are pair with kinase information, which is limited to Arabidopsis species; moreover, no phosphosite in databases is paired with kinase information. In contrast, RLIMS-P paired 20.2% (784/3,880) and 30.1% (181/601) of the corresponding total substrates and phosphosites with kinase information. Those facts revealed the potential of text mining tool RLIMS-P in addressing the lag in elucidating kinase information for the huge number of phosphoproteins and phosphosites available in current databases.
Table 4.1 Numbers of substrates, phosphosites, kinases, and scientific articles (PMIDs) captured by three databases (UniProtKB, PhosPhAt, P3DB) and RLIMS-P.

<table>
<thead>
<tr>
<th></th>
<th>Databases</th>
<th>RLIMS-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMIDs</td>
<td>109</td>
<td>4,749</td>
</tr>
<tr>
<td>Substrates</td>
<td>10,311(^1)</td>
<td>3,880</td>
</tr>
<tr>
<td>Kinase</td>
<td>147</td>
<td>925</td>
</tr>
<tr>
<td>Sites</td>
<td>27,711(^2)</td>
<td>601</td>
</tr>
<tr>
<td>Substrate w/ kinase information</td>
<td>574</td>
<td>784</td>
</tr>
<tr>
<td>Sites w/ kinase information</td>
<td>0</td>
<td>181</td>
</tr>
</tbody>
</table>

\(^1\)9,539 (of 10,311) substrates are exclusively from P3DB/PhosPhAt
\(^2\)26,263 (of 27,711) sites are exclusively from P3DB/PhosPhAt.

Then, we continued with assessing the availability of phosphorylation information by plant species that have been integrated by the combined databases. In addition to experimental plant Arabidopsis, crop plants such as *Oryza sativa* (rice), *Medicago truncatula* (Barrel Medic), *Zea mays* (maize), *Nicotiana tabacum* (tobacco), *Solanum tuberosum* (potato), and *Solanum lycopersicum* (tomato) as well as some other plant species (combined in the last row in Table 4.1, see Methods) are also incorporated by the current data resources. As the Databases columns of Table 4.1 show, *Arabidopsis* (contribute 66.7% and 68.5% to the total data sets of substrates and sites respectively) and rice (31.2%, 29.5%) are the 2 organisms having the most substrates and phosphosites data curated by combined databases. For *Arabidopsis* and rice, the remarkable larger numbers of substrate and sites from databases than those from RLIMS-P search again reveal the facts that high-throughput experiments results are often not described in the abstracts, which are processed by RLIMS-P. For example, in rice P3DB only curated one large-scale phosphoproteomic study results from supplementary materials (Nakagami et al., 2010) retrieving 3219
phosphoproteins. However, the combined databases contain little phosphorylation information from other species, such as the agronomically important plant species tobacco, maize, tomato, and potato. For those species, in contrast, RLIMS-P identified 1330 phosphoproteins (as compared to 92 in the databases) and 208 phosphosites (as compared to 128 in the databases) (Table 4.2). Those results revealed the huge number of phosphorylation data buried in scientific papers and also suggest that text-mining tool RLIMS-P could capture plant phosphorylation information complementary to the information currently represented in databases.

Table 4.2 Numbers of phosphoproteins, phosphosites and PMIDs available in recombined databases and literature for 7 plant organisms

<table>
<thead>
<tr>
<th></th>
<th>Databases</th>
<th>RLIMS-P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Phosphoproteins</td>
<td>Phosphosites</td>
</tr>
<tr>
<td><strong>Arabidopsis thaliana</strong></td>
<td>6874</td>
<td>18,976</td>
</tr>
<tr>
<td><strong>Oryza sativa</strong></td>
<td>3219</td>
<td>8179</td>
</tr>
<tr>
<td><strong>Medicago truncatula</strong></td>
<td>113</td>
<td>411</td>
</tr>
<tr>
<td><strong>Zea mays</strong></td>
<td>88</td>
<td>117</td>
</tr>
<tr>
<td><strong>Nicotiana tabacum</strong></td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td><strong>Solanum tuberosum</strong></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td><strong>Solanum lycopersicum</strong></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td>13</td>
<td>17</td>
</tr>
</tbody>
</table>

Due to the limited kinase information in current plant resources, we used RLISM-P to assess the current state of kinase study by plant species that are discussed in Table 4.2. For each species, we employed RLIMS-P to mine a subset of relevant scientific articles and manually curated the results for summarizing the numbers of
distinct kinase names (Table 4.3). As expected, most kinase information is from
*Arabidopsis*, followed by rice, tobacco, and maize. The large amount of textual kinase
information reveals the extensive phosphorylation studies and also the necessity to
integrate scattered information for systematic analysis. Thus, RLIMS-P, a text-mining
tool that can retrieve phosphorylation events with a specific focus on kinase-substrate
relationships with phosphosite information, could be applied for addressing the gap
between information in public databases and knowledge buried in scientific literature.

Table 4.3 Numbers of kinase protein names identified by RLIMS-P for the plant
model organisms

<table>
<thead>
<tr>
<th>Plant Name</th>
<th>Number of kinase</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>312</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>126</td>
</tr>
<tr>
<td><em>Nicotiana tabacum</em></td>
<td>91</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>64</td>
</tr>
<tr>
<td><em>Solanum lycopersicum</em></td>
<td>45</td>
</tr>
<tr>
<td><em>Solanum tuberosum</em></td>
<td>38</td>
</tr>
<tr>
<td><em>Medicago truncatula</em></td>
<td>5</td>
</tr>
</tbody>
</table>

4.2 Identification of the Top 15 Most Literature-documented Kinase Proteins in
*Arabidopsis* Using RLIMS-P

Rich kinase information has been identified by RLIMS-P for *Arabidopsis*, in
which we further constructed the kinase-centric phosphorylation network. As shown in
Table 4.4, there are 33,638 articles in the PubMed database relevant to *Arabidopsis*.
Of these articles, RLIMS-P identified 1023 (3.04%) articles as containing
phosphorylation events and 478 articles as containing phosphorylation events with
kinase information. In the 478 articles, RLIMS-P identified 465 distinct protein kinase
names; 600 substrate names; 211 substrates with phosphosite information; 485 kinase-substrate relationships; and 96 substrates with kinase and phosphosite information.

Table 4.4 Total scientific publications (PMIDs) in PubMed, RLIMS-P-positive PMIDs, kinase proteins, substrate proteins and phosphosites identified by RLIMS-P.

<table>
<thead>
<tr>
<th>PMIDs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>RLIMS-P-positive PMIDs</td>
<td>1023</td>
</tr>
<tr>
<td>RLIMS-P-positive PMIDs w/ kinase information</td>
<td>478</td>
</tr>
<tr>
<td>Kinase</td>
<td>465</td>
</tr>
<tr>
<td>Substrate</td>
<td>600</td>
</tr>
<tr>
<td>Site w/ substrate</td>
<td>211</td>
</tr>
<tr>
<td>Substrate w/ kinase</td>
<td>485</td>
</tr>
<tr>
<td>Substrate w/ kinase and site</td>
<td>96</td>
</tr>
</tbody>
</table>

Among the 465 *Arabidopsis* kinase names retrieved by RLIMS-P, we identified the top 15 most literature-mentioned protein names including 2 kinase complexes and 13 monomeric kinase proteins (Table 4.5). Casein kinase 2 (CK2) and SNF1-related protein kinase 1 (Snrk1) are complexes so we denoted them by the UniprotKB identifiers of their catalytic subunits. The other 13 protein names were mapped to unique UniprotKB identifiers.

Table 4.5 Summary table of PMIDs identified by RLIMS-P and PMIDs annotated in PRO for the top 15 kinases in *Arabidopsis*.

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Uniprot ID</th>
<th>RLIMS-P-positive PMIDs</th>
<th>PMIDs curated in PRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>CK2</td>
<td>Q08467</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>CKA1</td>
<td>Q08467</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>CKA2</td>
<td>Q08466</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>MPK6</td>
<td>Q39026</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>PDK1</td>
<td>Q9XF67</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>
According to *Arabidopsis* kinome classification in the PlantsP database (http://plantsp.genomics.purdue.edu/index.html), the 15 kinase proteins/complexes in were classified into 3 classes, 5 groups and 10 families (Table 4.6). STN8, BRI1 and BAK1 are in the class Transmembrane Receptor Kinase and Related non-Transmembrane Kinases and group Leucine Rich Repeat Receptor Kinase, which reflects their initiating roles in signal transduction pathways. PHYA, a member of Mitogen Activated Protein Kinase Kinase Kinase (MAP3K) family, is supposed to be an upstream kinase of MPK3 and MPK6, which are in the Mitogen Activated Protein Kinase (MAPK) family.

### Table 4.6 Classification of the top 15 most occurring kinases in *Arabidopsis*

<table>
<thead>
<tr>
<th>Class</th>
<th>Group</th>
<th>Family</th>
<th>Protein Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Transmembrane Protein Kinases</td>
<td>Mitogen Activated Protein Kinase Kinase Kinase (MAP3K)</td>
<td>MAP3K</td>
<td>PHYA</td>
</tr>
<tr>
<td>Calcium Response Kinase</td>
<td>SNF1 Related Protein Kinase (SnRK)</td>
<td>KIN10</td>
<td>KIN11</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------------------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>IRE/NPH/PI dependent/S6 Kinase</td>
<td>PID</td>
<td>PHOT1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PDK1</td>
<td></td>
</tr>
<tr>
<td>MAPK/CDC/CK2/GSK Kinases</td>
<td>MAPK Family</td>
<td>MPK3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Casein Kinase II Family</td>
<td>MPK6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CKA1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GSK3/Shaggy Like Protein Kinase</td>
<td>CKA2</td>
<td></td>
</tr>
<tr>
<td>ATN1/CTR1/EDR1/GmPK6 like Kinase</td>
<td>ATN1/CTR1/EDR1/GmPK6 like Kinase</td>
<td>CTR1</td>
<td></td>
</tr>
<tr>
<td>Transmembrane Receptor Kinase and Related non-Transmembrane Kinases</td>
<td>Leucine Rich Repeat Kinase II &amp; X</td>
<td>BAK1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leucine Rich Repeat Kinase VII</td>
<td>STN8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leucine Rich Repeat Kinase XI &amp; XII</td>
<td>BR11</td>
<td></td>
</tr>
</tbody>
</table>

The top 15 protein kinases were mentioned in 106 RLIMS-P-positive abstracts and for 84 of these articles we curated the full-length text (Table 4.5). Note that the 22 articles that we chose not to pursue further included 12 articles that contained a brief mention of a phosphorylated form that had been discussed in depth in an article that we had already processed, and 10 reviews articles not containing new experimental data. We extracted the phosphorylation information for each kinase including substrates, phosphosites and consequences of phosphorylation on target protein function, activity and/or subcellular localization. The extracted information was then entered into RACE-PRO and 348 PRO entries were generated for.
We applied the phosphorylation information we extracted and curated to two case studies. In the first case, protein kinases PDK1 and PID were used to demonstrate the role of text mining in data integration and knowledge discovery. Then, we continue with the second case of the BR signaling pathway, in which text mining, data mining, protein ontology, and visualization and analysis were integrated for construction and comparative analysis of a protein phosphorylation network.

4.3 PDK1 Phosphorylates PID to Regulate Auxin Polar Transport

Both PDK1 and PID are in the list of the top 15 most documented kinase proteins in the scientific literature. PDK1 is a central kinase phosphorylating a number of different targets involved in various physiological process including phospholipid signaling, cell growth and cell death (Otterhag et al., 2006). Activity of the protein kinase PID has been implicated in the phosphorylation of PINFORMED (PIN) family and B subfamily of ABC transporters (ABCBs), which interact and coordinate in auxin polar transport regulation (Benjamins et al., 2003; Huang et al., 2010; Henrichs et al., 2012). However, the mechanism of activating PID kinase activity is still unknown. Figure 4.1 shows that PDK1 directly phosphorylates PID; furthermore, the PRO term of At-PID/Phos:1 describes the phosphorylation sites are within PID activation loop and its autophosphorylation is induced (Zegzouti et al., 2006). Thus we hypothesized that PDK1 participates in the PID-regulated auxin polar transport process via phosphorylation of PID.
Figure 4.1 PDK1 phosphorylated PID participating in PID-regulated auxin polar transport process. Kinase-substrate relationships captured by text mining are depicted with solid blue arrows, and is_a relationships with dashed grey lines. The kinase proteins are depicted in blue-pink nodes, and phosphorylated forms from PRO with grey nodes with black borders.

4.4 Construction of BR-signaling Protein Phosphorylation Network Integrating Text mining and Data mining

Among the top 15 Arabidopsis kinases, BRI1, BAK1 and BIN2 are key components involved in BR-signaling pathway, which indicates the enrichment of phosphorylation studies and the significance of phosphorylation in the BR hormone signal response. Below we used BR as a use case to demonstrate the integrative approach combining text mining and data mining to construct phosphorylation networks for knowledge discovery.

4.4.1 Integration of text- and data-mining phosphorylation information on BR signaling

We collected phosphorylation information for the other 5 key BR-signaling components via RLIMS-P. RLIMS-P searches with keywords “CDG1”, “BSK1”, “BSU1”, “BZR1” and “BZR2” returned 1, 4, 8, 18 and 19 articles, respectively. Because these proteins are highly related as major components in BR signaling, they are often discussed together in the same BR-relevant articles. When this overlap is taken into account, 88 articles related to the 8 key BR-signaling components are identified by RLIMS-P. We curated the full-text of 39 of those articles and created 103
PRO terms including 15 *Arabidopsis*-gene terms, 2 *Arabidopsis*-complex terms and 46 *Arabidopsis* phosphorylated protein terms.

To enrich the BR-signaling phosphorylation network, we collected 29 experimentally validated kinase-substrate relationships from the PhosPhAt database (Table 4.7). Of the 55 kinase-substrate relationships found from text and data mining, 26 were exclusively identified by text mining, 10 were exclusively identified by data mining, and 19 were identified by both data mining and text mining. The 10 data-mining exclusive kinase-substrate relationships consist of data from large-scale phosphorylation experiments discussed in literature supplementary materials so they are not flagged by RLIMS-P. Thus, data mining could provide complementary phosphorylation information to text mining results, while a lot of phosphorylation information is in the literature is missed by databases but accessible through text mining.

In addition, we collected experimental PPI relationships for the 8 key BR-signaling proteins from databases including PhosPhAt, STRING (Franceschini et al., 2013), IntAct (Orchard et al., 2014), and BioGriD (Chatr-Aryamontri et al., 2013). Of the 98 data-mining experimental PPIs, 38 PPIs were already annotated in PRO as kinase-substrate relationships or with Gene Ontology (GO) (Ashburner et al., 2000) term GO:0005515 (protein binding); and 60 PPIs were included for below analysis (Table 4.7). The text- and data-mining information we collected revealed the extensive interplay of various types of biological interactions.

<table>
<thead>
<tr>
<th>Relation type</th>
<th>No. of</th>
<th>No. of</th>
<th>Data</th>
<th>No. of</th>
<th>No. of</th>
</tr>
</thead>
</table>

Table 4.7 Summary numbers of BR-signaling protein phosphorylation knowledge map data
<table>
<thead>
<tr>
<th>Table</th>
<th>nonredundant relations</th>
<th>nonredundant proteins</th>
<th>source</th>
<th>relations</th>
<th>proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinase-substrate</td>
<td>55</td>
<td>66</td>
<td>RLIMS-P, PhosPhAt</td>
<td>45</td>
<td>68</td>
</tr>
<tr>
<td>Phosphorylation-dependent PPI</td>
<td>12</td>
<td>19</td>
<td>PRO</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>PPI</td>
<td>60</td>
<td>57</td>
<td>BioGriD, IntAct, PhosPhAt, STRING</td>
<td>98</td>
<td>78</td>
</tr>
<tr>
<td>Is_a</td>
<td>35</td>
<td>51</td>
<td>PRO</td>
<td>35</td>
<td>51</td>
</tr>
<tr>
<td>has_part</td>
<td>4</td>
<td>5</td>
<td>PRO</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

### 4.4.2 BR-signaling protein phosphorylation knowledge map

In Figure 4.2, integrated text- and data-mining results for the BR-signaling pathway were visualized as a protein phosphorylation knowledge map, which consists of 121 proteins/genes (nodes) and 166 edges (relationships). The 8 key BR-signaling components—BRI1, BAK1, BSK1, CDG1, BSU1, BIN2, BZR1, BZR2—are linked directly in a signal transduction pathway from cell membrane to nucleus via phosphorylation events identified by text mining. In addition, PP2A, a central regulatory phosphatase in BR signaling transduction, was integrated into the knowledge map through data-mining PPIs with transcription factors BZR1 and BZR2. Thus, both text mining and data mining contributed critical pieces to the BR-signaling protein phosphorylation knowledge map.
Figure 4.2 Network visualization of the BR-signaling phosphorylation knowledge map integrating text mining and data mining results. Text-mining kinase-substrate relationships are depicted with solid blue arrows, is _ a relationships with dashed grey lines, has_part relationships with solid grey lines, data-mining kinase-substrate relationships with dashed blue arrows, text-mining PPIs with solid green lines, and data-mining PPIs with dashed green lines. In addition, the 8 BR-signaling key proteins are depicted in pink nodes, the 6 BR-signaling key kinases in pink-blue nodes, and phosphorylated forms form PRO in grey nodes with black borders. The node size corresponds to the number of directly connected edges.
The BR-signaling phosphorylation network includes 12 phosphorylation-dependent PPI relationships (Table 4.8), which describes the PPIs in the phosphorylation-state specific biological context. In contrast, data mining PPIs from current public resources only curated PPI on gene-level without specific modified form of binding proteins. In Table 4.8, BRI1 is phosphorylated on S858, S891 and T872 and subsequently binds to BAK1. Also, BSU1 phosphorylated on S395, S444 and S764 specifically interacts with and inactivates BIN2/Phos:1, which is an autophosphorylated form on Y200 and plays a negative role in BR signaling pathway.

Table 4.8 Phosphorylation-dependent PPIs in BR-signaling phosphorylation network

<table>
<thead>
<tr>
<th>Protein Form #1</th>
<th>Protein Form #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRI1/iso:1/phos:5</td>
<td>BAK1/iso:1</td>
</tr>
<tr>
<td>BSU1/Phos:1</td>
<td>BIN2/Phos:1</td>
</tr>
<tr>
<td>BIN2/Phos:1</td>
<td>BSL1</td>
</tr>
<tr>
<td>MYBL/Phos:1</td>
<td>BZR2</td>
</tr>
<tr>
<td>ASK3/Phos:1</td>
<td>BZR2</td>
</tr>
<tr>
<td>BZR1/phos:1</td>
<td>14-3-3 proteins</td>
</tr>
<tr>
<td>BZR1/Phos:3</td>
<td>14-3-3 proteins</td>
</tr>
<tr>
<td>BZR2/phos:1</td>
<td>14-3-3 proteins</td>
</tr>
<tr>
<td>BON1/Phos:1</td>
<td>BIR</td>
</tr>
<tr>
<td>PUB12/Phos:1</td>
<td>FLS2</td>
</tr>
<tr>
<td>PUB13/Phos:1</td>
<td>FLS2</td>
</tr>
<tr>
<td>BKI1/Phos:1</td>
<td>14-3-3 proteins</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
</tr>
</tbody>
</table>

4.4.3 Functional analysis of BR-signaling protein phosphorylation network

A number of studies demonstrated the interconnection of BR-signaling pathway with multiple biological processes; however, the regulatory mechanisms and the interconnecting components are largely unknown. We applied functional
annotation clustering analysis to investigate the cross talk between BR signaling and other various signaling pathways as well as biological processes. Within the knowledge map, we subdivided the 121 nodes in the BR signaling knowledge map into 9 functional sub-networks: 1 BR-signaling network, 7 functional sub-networks and 1 uncharacterized functional sub-network.

Figure 4.3 presents a clear picture of how sets of proteins/genes participate in a wide range of plant developmental and physiological processes. As expected, BR signal transduction was involved in regulation of plant development, protein modification, innate immunity, stress-responsive network and stomata development. Additionally it is highly integrated with the gibberellin-, Jasmonic acid-, abscisic acid- and auxin- signaling pathways through both direct interactions between signaling proteins and transcriptional regulation of key components of these pathways. Furthermore, we noticed a sub-network of uncharacterized proteins/gens without GO annotation, and thus further functional studies of those proteins/genes are needed for complete understanding of the global ramifications of BR signaling regulatory network.
4.4.4 Phosphorylation of BSK5 by BIN2 in response to salt stress

BSK5 was known to play a redundant role with BSK1 mediating BR signaling, while a recent functional study of BSK5 revealed that it was required for response to abiotic stresses including salt and drought (Li et al., 2012). Figure 4.4 displays the interconnection between the BR signaling and stress signal regulatory networks, which depends on BR-SIGNALING KINASE 5 (BSK5). RLIMS-P captured the
phosphorylation of BSK5 by BIN2 and BRI1 (solid blue arrows) but without phosphosite information. There is no clear BRI phosphorylation recognition motif mentioned in literature. However, previous studies demonstrated that most of the known BIN2 substrates contain repeats of a short conserved phosphorylation motif [S/T] XXX [S/T] (S/T corresponds to serine/threonine, and X represents any amino acid) (Peng et al., 2010), in which the C-terminal Ser/Thr residue is phosphorylated by an unknown kinase to generate a the candidate N-terminal phosphorylation site for BIN2 (Table 4.9). To obtain potential BSK5 phosphosite information, we searched PhosPhAt for curated phosphosite information and S329, shown to be phosphorylated in a high-throughput experiment (Reiland et al., 2009) matched the phosphorylation recognition motif of BIN2. This supports the hypothesis that phosphorylation of BSK5 by BIN2 on S329 might initiate or regulate salt- and drought- stress signaling.

![Diagram](image)

**Figure 4.4** The BR signaling pathway cross talks with the salinity stress response through phosphorylation BSK5 by BIN2. Node and edge colors and symbols are the same as in Figure 4.2.

**Table 4.9** Potential phosphorylation sites of BSK5 by BIN2

<table>
<thead>
<tr>
<th>Substrate</th>
<th>PRO ID</th>
<th>Kinase</th>
<th>Site</th>
<th>Motif</th>
<th>Reference</th>
</tr>
</thead>
</table>

39
4.4.5 **SERK proteins integrate BR signaling pathway with other biological processes**

There are 5 members in the SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASE (SERK) family including SERK1, SERK2, SERK3/BAK1, SERK4 and SERK5. It was known that, addition to BRI1-coreceptor BAK1/SERK3, other members of the SERK family redundantly promote BR signaling, while SERK2 plays a minor role in this process (Gou et al., 2012). Genetic experiments reported that SERK1 and SERK2 were involved in the regulation of anther development (Albrecht et al., 2005; Colcombet et al., 2005); and BAK1/SERK3 and SERK4 were involved in the suppression of cell death. Also, BAK1/SERK3 served as receptor kinase regulating innate immunity and pathogen defense in a BR-independent manner. Figure 4.5 displays the subnetwork of SERK proteins directly interacting with BRI1 by multiple types of edges (relationships). PPIs of SERK2 and SERK4 with BRI captured through data mining, auto-phosphorylation of SERK1 and SERK3, phosphorylation of SERK1 and BAK1/SERK3 by BRI1, and phosphorylation of BRI1 by SERK1 and SERK3 retrieved by text mining. Thus integration of text mining and data mining results connected fragmented phosphorylation information buried in articles and databases to create a network that can be easily used for novel knowledge discovery.
4.4.6 Conservation of BR signaling pathway across plant species

Extensive studies on BR-signaling pathway in *Arabidopsis* provided excellent bases for the understanding of the mechanisms that regulate signal cascade. Combined text- and data-mining approach captured a wealth of phosphorylation information representing a multiply cross-talking regulatory network. However, transferring the knowledge from experimental plant *Arabidopsis* to agronomically important plants is still challenging due to the lack of indications about whether the primary BR-signaling protein phosphorylation network is conserved. For generating novel knowledge on BR-signaling pathway in other species, we examined the conservation of orthologous protein sequences and the phosphorylation-state specific forms of the 8 key proteins:

- Cell death
- Male sporogenesis
- Tapetum Development
- Microspore maturation

Figure 4.5 Sub-network of SERKs family proteins involved in multiple biological processes. Node and edge colors and symbols are the same as in Figure 4.2.
components for 5 agronomically important and whole-proteome established plant species, including *Glycine max* (soybean), potato, tomato, rice, and maize.

All five organisms had orthologs of BRI1, BAK1, BSK1, BIN2, BSU1, BZR1, and BZR2 (Appendix Table A.1-A.5). Interestingly, none of them had an ortholog of the *Arabidopsis* CDG1. Furthermore, we explore the conservation of the phosphorylation-state specific PRO entry that has a particular combined phosphosites (Table 4.10). Often those PRO entries are associated with functional GO annotation. Although there was almost no experimental phosphorylation information for any of the five species, the sites were considered as putative phosphosites if the aligned orthologous protein sequence matched at the positions of the known *Arabidopsis* phosphorylation sites (serine for threonine substitutions and vice versa were permitted) (Table 4.9; Supplementary materials A.1-A.5). As Table 4.9 summarized, many BR-signaling phosphorylation events are conserved across the 6 plants species; there are some species-specific features as well. For BSK1, BIN2 and BZR1 the *Arabidopsis* phosphorylated forms are all conserved across the 6 species; and the phosphorylated BZR2 form is conserved in all species except maize (11/12 phosphosites are conserved in maize). This suggested the potential of their conserved biological functions in BR-signaling pathway across plant species.

Table 4.10 Conservation of phosphorylation-specific protein forms across 6 plant species. + represents GO term: positive regulation of brassinosteroid mediated signaling pathway (GO: 1900459); — represents GO term: negative regulation of brassinosteroid mediated signaling pathway (GO:1900458); * represents conserved form, X represents non-conserved form, and NA represents nonexistent ortholog.

<table>
<thead>
<tr>
<th><em>Arabidopsis</em></th>
<th>PRO</th>
<th>Soybean</th>
<th>Rice</th>
<th>Potato</th>
<th>Tomato</th>
<th>Maize</th>
</tr>
</thead>
</table>

42
Auto-phosphorylation of BRI1 and trans-phosphorylation BRI1 by its coreceptor BAK1 play a critical role in initiating the BR signal cascade. However, a detailed analysis of BRI1 phosphorylation has not been reported in any plant other than *Arabidopsis*. Of the 8 phosphorylated BRI1 forms, four are conserved, whereas four are specific only to *Arabidopsis* (Table 4.10; Table 4.11 and Figure 4.6). One of the completely conserved forms is At-BRI1/Phos:1 (PR: 000028357), an auto-phosphorylated protein form with five phospho-tyrosine residues (Y831, Y956,
Y1052, Y1057 and Y1072) that regulates BRI1 kinase activity and BR signaling. The residue that corresponds to Y831 in the tomato ortholog of BRI1 (Y839) has been experimentally shown to be phosphorylated (Bajwa et al., 2013).

Unlike the BRI1 auto-phosphotyrosine form, At-BRI1/Phos:3 (PR: 000028361), the form trans-phosphorylated by BAK1/SERK3, was not conserved in any other organisms. In Arabidopsis, this form plays a role in increasing BRI1 kinase activity and BR signaling output and thus, its lack of conservation was surprising. Upon closer inspection, we found that of the five serines and threonines phosphorylated in At-BRI1/Phos:3 (S838, T846, S858, S1166, and T1180), four were in fact conserved in all species. The exception was T1180, which was located in a relatively less well-conserved region of the protein near the C-terminus. In protein forms with multiple phosphorylated residues, some of the phosphorylation sites may be more important for function than other. It would be interesting to investigate the requirement for T1180 phosphorylation in BAK1/SERK3 trans-activation of BRI1.

Table 4.11 BLAST result for the orthologs of Arabidopsis BRI1 protein across 6 species

<table>
<thead>
<tr>
<th>BRI1</th>
<th>Uniprot AC</th>
<th>Identify</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>O22476</td>
<td>100.00%</td>
<td>0</td>
</tr>
<tr>
<td>Glycine max</td>
<td>C6FF79</td>
<td>69.00%</td>
<td>0</td>
</tr>
<tr>
<td>Oryza sativa subsp. japonica</td>
<td>Q942F3</td>
<td>53.00%</td>
<td>0</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>Q8GUQ5</td>
<td>67.00%</td>
<td>0</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>A4LAP6</td>
<td>66.00%</td>
<td>0</td>
</tr>
<tr>
<td>Zea mays</td>
<td>K7V4X2</td>
<td>54.00%</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 4.6 Multiple sequence alignment of the orthologs of Arabidopsis BR11 protein across 6 plants species. The kinase information recorded in PRO ERO is shown in parentheses. Experimentally validated phosphosites are represented with red, and predictive phosphosites are represented with grey.

Orthologs to Arabidopsis BAK1 across the 5 species were identified (Table 4.10; Table 4.12 and Figure 4.7). The sites in in vivo phosphorylated form of BAK1 (At-BAK1/Phos:2; PR:000035954) are conserved across other 5 species, which suggest those sites are potential phosphosites in other species. Moreover, some of the
those potential sites in tomato have been experimentally validated to be phosphosites (Bajwa et al., 2013). However, another autophosphorylated form of BAK1 (At-BAK1/Phos:1; PR: 000028347), a positive regulator in BR signaling pathway (GO:1900459), do not exist in other 5 species in view of the Y610 phosphosites in this Arabidopsis phosphorylated form are not conserved in other 5 species. This suggests that those crop plants may employ alternative mechanisms instead of phosphorylation of BAK1 in Y610 to positively regulate BR signaling pathway.

Table 4.12 BLAST result for the orthologs of Arabidopsis BAK1 protein across 6 species.

<table>
<thead>
<tr>
<th>BAK1</th>
<th>Uniprot AC</th>
<th>Identify</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>Q94F62</td>
<td>100.00%</td>
<td>0</td>
</tr>
<tr>
<td>Glycine max</td>
<td>I1MDV1</td>
<td>85.00%</td>
<td>0</td>
</tr>
<tr>
<td>Oryza sativa subsp. japonica</td>
<td>Q6Z4U4</td>
<td>79.00%</td>
<td>0</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>G0XZA6</td>
<td>85.00%</td>
<td>0</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>S5YNV6</td>
<td>85.00%</td>
<td>0</td>
</tr>
<tr>
<td>Zea mays</td>
<td>K7TQF3</td>
<td>78.00%</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 4.7 Multiple sequence alignment of BAK1 across 6 plants species. The kinase information recorded in PRO entry is shown in parentheses. Site colors are as same as Figure 4.6.

Phosphorylation BSU1 by CDG1 is an essential step to inactive downstream BIN2 to regulate transcription factors BZR1 and BZR2 for BR-responsive gene expression. According to the PRO entry of CDG1 (Table 4.6), phosphorylated CDG1 (At-CDG1/Phos:1; PR:000028337) will interact with (GO term: protein binding; GO:0005515) its substrate BSU1 (At-BSU1, PR:000028340), leading to the phosphorylation of BSU1 (At-BSU1/Phos:1; PR:000028342) (Kim et al., 2011). However, based on BLAST results we did not identify orthologs of CDG1 and 1 (of 3) phosphorylation sites in the phosphorylated form of BSU1 were not conserved, which suggested that the other 6 organisms might have distinct kinase to fulfill the role of
CDG1 for activating BSU1 as well as positively regulating BR-signaling downstream cascade.
Chapter 5

DISCUSSION AND FUTURE WORK

5.1 Discussion

Phosphorylation plays a central regulatory role in multiple biological processes; however, currently there is a lack of integrative bioinformatics resources for a systematic approach to global analysis of phosphorylation networks, especially in plants. In this article, we combined text mining, data mining, ontology, visualization and analysis to construct phosphorylation networks in Arabidopsis. The resulting phosphorylation network connects the fragmented phosphorylation information embedded in literature and databases for hypothesis generation. Meanwhile, there are some critical issues challenging our approach for systematic study and analysis of plant protein phosphorylation events, which are in need of further study.

In our study of the literature documentation of Arabidopsis kinases, we began with a list of kinase names retrieved from a set of Arabidopsis relevant articles flagged by RLIMS-P, which mines scientific articles for phosphorylation-related information including kinase names, substrate names, and phosphosites. In the work presented, we identified the 15 kinase names most frequently occurring on the list and developed two case studies based on these results.

However, there exist several biases we have to point out. First, a number of phosphorylation experiments are conducted in other plant model organisms such as rice, tomato, and maize, and those kinases will be ignored by our approach that merely focuses on Arabidopsis. As shown in table 4.2, there are respectively 177, 145 and 108
kinases in tobacco, rice and maize documented in the literature. In rice, we identified the list of the top 15 most literature-documented kinase proteins as we described in *Arabidopsis*. The rice list had several kinases in common with the *Arabidopsis* list, including CK2, MAPK family and SNRK1 (data not shown), whereas osM KK6, osCK1, and osCDPK5 are not among the *Arabidopsis* top 15 kinase list. The conservation and specificity of the kinome across plant species reveals a variety of phosphorylation events in plant systems in need of comparative analysis.

Second, the 465 kinase names identified by RLIMS-P contain duplicate and ambiguous terms, which might cause incorrect ranking of kinase names. To evaluate the kinase names identified by RLIMS-P, we manually reviewed the 465 kinase names and identified several limitations. (1) sometimes RLIMS-P extracted several distinct names that actually all refer to the same kinase. For example, RLIMS-P identified such terms as “atmpk6”, “map kinases 3 and 6”, “*Arabidopsis* thaliana mitogen-activated protein kinases mpk3 and mpk6”, and “*Arabidopsis* mapks mpk3, mpk6;” all of those names refer to the kinase MPK6. If we take this into consideration, the actual number of articles discussing MPK6 phosphorylation is 30, much larger than the 10 articles reported Table 4.5. (2) A number of articles describe experiments conducted on kinase families instead of specific kinases. For example, RLISM-P identified kinase names including “mpk”, “mapk”, ”atmpk proteins”, and “amp-activated protein kinase (ampk)”’, which all refer to the MAPK family. This subset of articles often discussed the parallel phosphorylation pathways regulated by MAPK family members such as MPK3, MPK4 and MPK6. (3) Another limitation is the RLIMS-P performance. We observed that RLIMS-P identified “mpk6.26” as a kinase name,
whereas the correct term is “mpk6” and “.26” belongs to the next sentence. This type of limitation will be addressed in future versions of RLIMS-P.

To evaluate the bias of the top 15 Arabidopsis kinase list identified by RLIMS-P, we compared it with the list of top 15 Arabidopsis most curated kinases by data mining. We downloaded the kinase-substrate interactions that are exclusively tagged with the evidence of phosphorylation or autophosphorylation from PhosPhAt database (http://phosphat.uni-hohenheim.de). In Table 5.1, the top 15 kinase list identified by RLIMSP covered 11 kinase proteins, including MPK6, MPK3, BRI1, STN8, OST1, BAK1, PDK1, BIN2 and PHOT1. The other 4 kinase proteins that are only in the data mining top 15 kinase list fell into 2 categories: (1) For CKPK6, CDKA-1 and CDKF-1, RLIMS-P retrieved CDPK family (CKPK6) and CDK family (CDKA-1 and CDKF-1) as kinase names, both of which were actually among the top 20 most documented kinase identified by RLIMSP; (2) STN7 was always discussed with other STN family members and retrieved e.g. “STN7, STN8”. Thus, the high overlapping (11/15) in kinase lists between RLIMS-P search and databases curation demonstrated the RLIMS-P precision, and also pointed out the necessity to collect phosphorylation data based on kinase family instead of single kinase.

Table 5.1 The top 15 most curated kinase proteins identified by data mining

<table>
<thead>
<tr>
<th>UniprotKB ID</th>
<th>Gene name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q39026</td>
<td>MPK6</td>
</tr>
<tr>
<td>Q39023</td>
<td>MPK3</td>
</tr>
<tr>
<td>Q22476</td>
<td>BRI1</td>
</tr>
<tr>
<td>Q42479</td>
<td>CDPK6 *</td>
</tr>
<tr>
<td>Q9LZV4</td>
<td>STN8</td>
</tr>
<tr>
<td>Q940H6</td>
<td>OST1</td>
</tr>
<tr>
<td>Q94F62</td>
<td>BAK1</td>
</tr>
</tbody>
</table>
* represents the kinases that are exclusively among the top 15 most curated kinases identified by data mining.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q9XF67</td>
<td>PDK1</td>
</tr>
<tr>
<td>Q9S713</td>
<td>STN7*</td>
</tr>
<tr>
<td>Q39011</td>
<td>BIN2</td>
</tr>
<tr>
<td>O48963</td>
<td>PHOT1</td>
</tr>
<tr>
<td>P24100</td>
<td>CDKA-1*</td>
</tr>
<tr>
<td>O64682</td>
<td>PID</td>
</tr>
<tr>
<td>O80345</td>
<td>CDKF-1*</td>
</tr>
<tr>
<td>Q9M9E9</td>
<td>SRK2C</td>
</tr>
</tbody>
</table>

Given the article set flagged by RLIMS-P, we curated and created PRO terms annotating the textual phosphorylation knowledge in a standard and computer-readable format. In future studies, those PRO terms will provide an easy and convenient way to obtain the knowledge. In the case study of kinases PDK1 and PID, we connected the pieces of fragmented phosphorylation information on PID and PDK1 within one biological process, in which PDK1 can activate PID kinase activity through phosphorylation and might be involved in PID-regulated auxin polar transport. Thus, the systematic mining and curation of phosphorylation knowledge embedded in literature could integrate the scattered information for hypothesis generation and novel knowledge discovery.

In the case of BR-signaling pathway, we demonstrated the application of the integrative approach for construction of a protein phosphorylation network. We combined multiple approaches including text mining, data mining, ontology, visualization and analysis to construct BR-signaling protein phosphorylation network in *Arabidopsis*. The resulting network is enriched with multiple types of relationships including kinase-substrate, phosphorylation-state specific PPI, PPI, is_a and has_part relationships. Specifically, the interplay of multiple types of interactions of SERK
family proteins and BR-signaling key components demonstrates the cross-talk of BR-signaling pathway with the biological processes regulated by SERK family.

We also hypothesized the S329 of BSK5 is phosphorylated by kinase BIN2 in response to salt stress. Notably, text mining also captured the phosphorylation of BSK5 by BRI1 but lacking phosphosite information. As a transmembrane receptor-like kinase, BRI1 is speculated to initiate salt stress response through the perception of stress signaling and phosphorylation of BSK5. However, currently no clear BRI1 phosphorylation recognition motif has been discussed (Oh et al., 2000; Wu et al., 2012) and the alignment of 8 PRO terms of BRI1 phosphorylated forms did not contain any obviously conserved motif (data now shown). The sites and functional impact of BSK5 phosphorylation by BRI1 is still unknown.

Finally, the conservation of the BR-signaling pathway was examined across seven plant model organisms. BR-signaling pathway have proven primarily conserved based on previous studies; however, we observed that the BSU1 phosphorylation event by CDG1 was not conserved and its associated GO annotation in PRO could not be applied in the other 6 species. PRO employs controlled vocabularies and standard format to represent the textual phosphorylation events and their functional annotations. If the phosphorylation-state specific forms of orthologs are conserved, PRO provides a guide to the prediction of the functional annotation across species. Thus, our phosphorylation network could be applied across species for comparative analysis of phosphorylation-specific protein forms, PPIs, functional annotations in a biological pathway for hypothesis generation and knowledge discovery.

In conclusion, we developed a bioinformatics workflow, which involves text mining, data mining, protein ontology, and visualization and analysis to create
phosphorylation-focused knowledge maps for the plant science community. This integrative approach was applied to the extraction and curation of the phosphorylation data for the top 15 kinases described in the literature. In the case of PDK1-PID, connecting pieces of textual phosphorylation information, we showed that PDK1 directly phosphorylates PID and participates in the PID-regulated biological processes. In the case of the BR signaling pathway, combining text mining and data mining, we hypothesized that S329 in BSK5 was phosphorylated by BIN2 in response to salt stress. Also, we showed that SERK family proteins interacted with BRI1 via multiple types of interactions, which indicated the potential cross-talk of the BR signaling with multiple other SERKs-regulated pathways. Our integrative bioinformatics approach takes knowledge from discrete resources and provides the network visualization enriched with PTMs for novel knowledge discovery, which can be applied to future plant biology studies.

5.2 Future Work

Our study is conducted on specific kinase proteins in Arabidopsis; however, there are a number of phosphorylation experiments carried out within kinase protein families. For example, in the RLISM-P search results, the MAPK family was in the list of top 15 most mentioned kinase names, and the identified subset of articles discussed phosphorylation of its family members MPK3 and MPK6. In Arabidopsis, PlantP database has classified the 1264 kinases into 81 families and 110 phosphatases into 12 families. In order to systematically analyze phosphorylation events across kinase families, we are going to use RLIMS-P search focusing on specific kinase families to explore the conservation and uniqueness of phosphorylation-mediating mechanisms.
For example, we will search RLIMS-P using “MAPK family” as keywords to extract related phosphorylation data.

In addition, besides Arabidopsis thaliana, a number of phosphorylation studies have been conducted and curated by public databases for other plant model organisms. Future studies will likely collect and curate phosphorylation in other plant species for comparative analysis.

Finally, this thesis work is conducted with a focus on phosphorylation analysis; however, our approach is applicable to other types of PTMs involved in numerous biological processes. As an example, during the curation of the RLIMS-P-positive literature, we found an article containing protein acetylation information, which is also a critical mechanism regulating signaling pathways. With RLIMS-P expanding to capture other PTMs, it could be feasible to add information about these other PTMs to our knowledge map.
REFERENCES


STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research 41, D808-815.


Appendix A

MULTIPLE SEQUENCE ALIGNMENTS OF THE ARABIDOPSIS ORTHOLOGS OF 8 BR-SIGNALING KEY COMPONENTS (BRI1, BAK1, BSK1, CDG1, BUS1, BIN2, BZR1, AND BZR2)

Table A.1 BLAST result for the orthologs of Arabidopsis BSU1 protein across 6 species

<table>
<thead>
<tr>
<th>Ortholog</th>
<th>Uniprot AC</th>
<th>Identify</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>Q944A7</td>
<td>100.00%</td>
<td>0</td>
</tr>
<tr>
<td>Glycine max</td>
<td>I1JJV5</td>
<td>88.00%</td>
<td>0</td>
</tr>
<tr>
<td>Oryza sativa subsp. japonica</td>
<td>Q10S67</td>
<td>76.00%</td>
<td>0</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>K4BVV3</td>
<td>86.00%</td>
<td>0</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>M1AM44</td>
<td>86.00%</td>
<td>0</td>
</tr>
<tr>
<td>Zea mays</td>
<td>C4J4J0</td>
<td>76.00%</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure A.1 Multiple sequence alignment of the orthologs of Arabidopsis BSK1 across 6 plants species. The kinase information recorded in PRO entry is shown in parentheses. Site colors are as same as Figure 4.6.
Table A.2 BLAST result for the orthologs of Arabidopsis BSU1 protein across 6 species

<table>
<thead>
<tr>
<th>Species</th>
<th>BLAST Identity</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>100.00%</td>
<td>0</td>
</tr>
<tr>
<td>Glycine max</td>
<td>52.00%</td>
<td>0</td>
</tr>
<tr>
<td>Oryza sativa subsp. japonica</td>
<td>52.00%</td>
<td>0</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>54.00%</td>
<td>0</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>54.00%</td>
<td>0</td>
</tr>
<tr>
<td>Zea mays</td>
<td>76.00%</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure A.2 Multiple sequence alignment of the orthologs of Arabidopsis BSU1 across 6 plants species. The kinase information recorded in PRO entry is shown in parentheses. Site colors are as same as Figure 4.6.
Table A.3 BLAST result for the orthologs of *Arabidopsis* BIN2 protein across 6 species

<table>
<thead>
<tr>
<th>BIN2</th>
<th>Uniprot AC</th>
<th>Identify (%)</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>Q39011</td>
<td>100.00</td>
<td>0</td>
</tr>
<tr>
<td><em>Glycine max</em></td>
<td>I1JU95</td>
<td>89.00</td>
<td>0</td>
</tr>
<tr>
<td><em>Oryza sativa subsp. japonica</em></td>
<td>Q9LWN0</td>
<td>89.00</td>
<td>0</td>
</tr>
<tr>
<td><em>Solanum lycopersicum</em></td>
<td>K4B8K4</td>
<td>90.00</td>
<td>0</td>
</tr>
<tr>
<td><em>Solanum tuberosum</em></td>
<td>M1CR16</td>
<td>90.00</td>
<td>0</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>B4F9G4</td>
<td>87.00</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure A.3 Multiple sequence alignment of the orthologs of *Arabidopsis* BIN2 across 6 plants species. The kinase information recorded in PRO entry is shown in parentheses. Site colors are as same as Figure 4.6.

Table A.4 BLAST result for the orthologs of *Arabidopsis* BZR1 protein across 6 species

<table>
<thead>
<tr>
<th>BZR1</th>
<th>Uniprot AC</th>
<th>Identify (%)</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>Q8S307</td>
<td>100.00</td>
<td>0</td>
</tr>
<tr>
<td><em>Glycine max</em></td>
<td>I1K7U0</td>
<td>68.00</td>
<td>6.00E-94</td>
</tr>
<tr>
<td><em>Oryza sativa subsp. japonica</em></td>
<td>Q0D559</td>
<td>57.00</td>
<td>3.00E-70</td>
</tr>
</tbody>
</table>

65
Solanum lycopersicum  
K4DGX5  64.00%  9.00E-97
Solanum tuberosum  
M1B8X3  64.00%  2.00E-96
Zea mays  
B6TXW1  47.00%  2.00E-61

Figure A.4 Multiple sequence alignment of the orthologs of Arabidopsis BZR1 across 6 plants species. The kinase information recorded in PRO entry is shown in parentheses. Site colors are as same as Figure 4.6.

Table A.5 BLAST result for the orthologs of Arabidopsis BZR2 protein across 6 species

<table>
<thead>
<tr>
<th>BZR2</th>
<th>Uniprot AC</th>
<th>Identity</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>Q9LN63</td>
<td>100.00%</td>
<td>0</td>
</tr>
<tr>
<td>Glycine max</td>
<td>1IK7U0</td>
<td>68.00%</td>
<td>4.00E-95</td>
</tr>
<tr>
<td>Oryza sativa subsp. japonica</td>
<td>Q0D559</td>
<td>57.00%</td>
<td>1.00E-71</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>Q94ET3</td>
<td>63.00%</td>
<td>2.00E-94</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>M1B8X3</td>
<td>63.00%</td>
<td>2.00E-93</td>
</tr>
<tr>
<td>Zea mays</td>
<td>B6TXW1</td>
<td>47.00%</td>
<td>5.00E-64</td>
</tr>
</tbody>
</table>

66
Figure A.5 Multiple sequence alignment of the orthologs of *Arabidopsis* BZR2 across 6 plants species. The kinase information recorded in PRO entry is shown in parentheses. Site colors are as same as Figure 4.6.