

**TEXT-MINING AND VISUALIZATION APPROACH HELP
INTERPRET EXPERIMENTAL DATA AND MAKE HYPOTHESIS**

by

Pan Teng

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

Spring 2015

© 2015 Pan Teng
All Rights Reserved

ProQuest Number: 1596901

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 1596901

Published by ProQuest LLC (2015). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

**TEXT-MINING AND VISUALIZATION APPROACH HELP
INTERPRET EXPERIMENTAL DATA AND MAKE HYPOTHESIS**

by

Pan Teng

Approved: _____
Carl J. Schmidt, 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

ACKNOWLEDGEMENTS

I would express my gratitude to my advisor Dr. Schmidt, for his continuous guidance, caring, patience and providing me the excellent atmosphere and chance for doing research. I would also like to thank Dr. Wu for taking me to the bioinformatics program, providing excellent training to help me transform from a chemist to a bioinformatician.

To my committee member Dr. Shanker, Dr. Arighi, Dr. Chen for their guidance, support and advice. I would also thank Dr. Tudor for being my mentor during my master study.

To my parents Xuefeng Teng and Naiqin Li, and all my family members, I sincerely thank you for your endless love and unconditional support.

To my friends and lab mates, thank all of you guys for your cheering me up, companion through good times as well as the hard ones.

I could never have done this without all of you.

TABLE OF CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	vii
ABSTRACT	x
 Chapter	
1 INTRODUCTION	1
2 RELATED WORK	4
2.1 Protein Phosphorylation Information Databases	4
2.1.1 PhosphoSite Plus	4
2.1.2 Phospho.ELM	5
2.1.3 P3DB	5
2.1.4 Protein Ontology (PRO)	5
2.2 Text-mining Approach for Protein Phosphorylation Information . . .	11
2.2.1 eGIFT	12
2.2.2 RLIMS-P	12
2.3 iPTMnet	14
2.4 WebGIVI	16
3 METHOD	17
3.1 Overview of the study	17
3.2 Pipeline	19
3.3 Database	19
3.4 Web Interface	20
3.5 Visualization	22
3.5.1 Visualizing the kinase-substrate pairs	23

3.5.2	Visualizing genes with concept terms	23
3.6	Evaluation iGep, Case Study	24
4	RESULT	25
4.1	iGep	25
4.2	Analysis of LMH Cell Heat Shock Response — A Case Study Using iGep	31
5	DISCUSSION AND FUTURE WORK	37
5.1	Discussion	37
5.2	Future Work	38
6	CONCLUSION	40
	REFERENCES	41

LIST OF TABLES

4.1	Kinase VRK1 and its substrates identified from input data	26
-----	---	--------------------

LIST OF FIGURES

2.1	PhosphoSitePlus Protein Page overview. Figure obtained from Hornbeck et. al. 2012[7]. This figure shows protein page overview for YAP1. Hippo Pathway from CST, shown lower left can be opened and downloaded. Lower right shows 3KYS PDF file which can be opened in a new Viewer window.	6
2.2	Substrate sequence logo generated by application from PhosphoSitePlus (in this case substrate of kinase AKT1). http://www.phosphosite.org/siteSearchMotifViewAction.do?x=30&y=13&background=automatic	7
2.3	Protein phosphorylation modification sites and domains for human AKT1 protein from PhosphoSitePlus multiple sequence alignment view. http://www.phosphosite.org/proteinAction.do?id=570&showAllSites=true	8
2.4	Figure took from Dinkel et. al. 2011[4]. Output example of Phospho.ELM for Cyclin dependent kinase inhibitor 1B (UniProt P46527).	9
2.5	P3DB result page of Arabidopsis thaliana ACG kinases including homologs to PKA, PKG and PKC http://www.p3db.org/protein.php?id=11616&ref=	10
2.6	Protein Ontology report for entry phosphorylated RAC-alpha serine/threonine-protein kinase isoform 1 phosphorylated 1 and GO annotation of hAKT1/iso:1/Phos:1 http://pir.georgetown.edu/cgi-bin/pro/entry_pro?id=PR:000028994	11
2.7	eGIFT summary page of gene AKT1	13
2.8	Higher ranked iTerms under category of functions and processes for gene AKT1	13

2.9	RLIMS-P result page for search of gene AKT1 http://research.bioinformatics.udel.edu/rlimsp/view.php?s=1764&abs=0 . .	14
2.10	RLIMS-P result page for article with PMID 16780593, about protein phosphorylation information where the kinase is AKT1 protein . .	15
3.1	Overview of the study design	18
3.2	Pipeline overview	18
3.3	Web interface of iGep	20
3.4	Sample code at http://iamceege.github.io/tooltipster/ demonstrating how to encode HTML markup directly by setting the title attribute.	22
4.1	Legend of results	25
4.2	Sample result table	27
4.3	An interactive tooltip pops up when user hover over link of VRK1 .	27
4.4	Using excel to view the CSV file downloaded	28
4.5	Left: Users can select layout of table from kinase centric, substrate centric, or genes couldn't identify any phosphorylation information for their corresponding substrate or kinase; Right: Download button to download the result to CSV format file	28
4.6	RLIMS-P result for article PMID: 15105425	29
4.7	Concept Map view from WebGIVI visualizing kinase and substrate relationships in sample data	30
4.8	Cytoscape view from WebGIVI visualizing kinase and substrate relationships in sample data	31
4.9	Cytoscape view from WebGIVI visualizing kinase and substrate relationships in LMH cell	32
4.10	Concept map view from WebGIVI visualizing kinase and substrate relationships in LMH cell	32

4.11	Fragment from the cytoscape view: CHAR connects several big groups in the phosphorylation network	33
4.12	Fragment from the cytoscape view: MUC1 connects several big groups in the phosphorylation network	34
4.13	Fragment from the cytoscape view: JDP2 can be phosphorylated by MAPK1, MAPK8, MAPK9, MAPK14 expressed	34
4.14	Fragment from the cytoscape view: JUN could be phosphorylated by PRKD1 and VRK1	35
4.15	Fragment from the cytoscape view: PRKD1 with lots of its substrate proteins	35

ABSTRACT

Protein phosphorylation plays a central roll in cellular signaling. Kinases are enzyme that participates in protein phosphorylation events, by catalyzing the transfer of phosphate group to a specific substrate. This phosphorylation typically affects substrate function, typically by either activating or inhibiting the substrates activity. Consequently, identifying kinase and substrate pairs in large-scale gene expression data will help the researcher in understanding the underlying biology of their experiments. With the continuous growth of scientific literature, it becomes more and more difficult for biologists to search for all of the information regarding kinases and substrates manually. To assist in this effort, we developed a web-based tool iGep (Integrating Gene Expression and Phosphorylation) to identify potential kinase and substrate pairs in gene expression data. Other functions including highlighting up and down regulated genes, linking users to PubMed literature describing particular phosphorylation events. In addition, users can visualize corresponding RLIMS-P, (a rule-based text-mining program for extracting protein phosphorylation information from literature) text evidence, sentences in the literature containing co-occurring kinase and substrates pairs and download all results.

Chapter 1

INTRODUCTION

As biology increases its use of high-throughput experimental techniques, it is essential to identify functional activities within large lists of genes and proteins. Often this involves mapping individual genes in these lists to ontology terms (REF DAVID and others), pathways (REFerence KEGG, Reactome, pathRings) and informative terms (REF eGIFT). These mappings provide a way to organize data and to generate new, testable hypotheses. One means to improve hypothesis generation would be to provide groupings at a level that implies direct consequences of changes in gene or protein levels. Identifying kinases and substrates in large-scale experimental data could directly assist the researcher in developing hypotheses to explain the biological processes under investigation. Protein phosphorylation is protein posttranslational modification event in which kinases transfer one or more phosphate group onto a protein substrate. This modification plays important roles in regulating many biological processes including metabolic and signaling pathways along with transcription. Identifying kinases and substrates in large gene lists could yield immediately testable hypotheses regarding the biology under investigation.

As literature continuously grow at a ultrafast speed, as well as other kinds of resources, biologists are overwhelmed by information. For example, 16,000 to 20,000 genes in average will be found expressed one chicken library, which our group usually works on. Although there are a lot of published databases, or even integrated source, once got such large list of genes, it's impossible for biologists to search for information or identify kinases-substrate pairs manually one by one. At the same time, it's hard to further intepret the gene list from the expression profile, to tell which one plays

important role in their experiment systems, potentially regulating the gene expression profile, responsible for the experiment conditions.

There is abundant knowledge of phosphorylation events and their consequences in the literature. This information is being captured in databases such as PhosphoSitePlus [7] which is the most comprehensive one. Many of these databases are manually curated and provide information for multiple species. RLIMS-P is a rule-based text-mining system for protein phosphorylation information that automatically extracts phosphorylation instances from literature, displays results with text evidence including abstract, color-coded entity mentions, and links to UniProtKB for normalized entities[19]. iPTMnet is an integrated resource for protein post-translational modifications. iPTMnet integrates information from multiple resources including: the text-mining tools eGIFT[21], RLIMS-P[19][25], and eFIP[20]; Protein Ontology (PRO)[15] along with other relevant PTM Resources including (PhosphoSitePlus[7], Phospho.ELM[4], PhosphoGRID[17], UniProt KnowledgeBase (UniProtKB)[22], etc.). By exploiting these resources life scientists can rapidly determine the relationships between kinases, substrates and the impact of phosphorylation events on protein activities.

To facilitate identification of kinases and their substrates in large data lists, we developed a web-based tool iGep (Integrating Gene Expression and Phosphorylation) to identify kinase and substrate pairs from gene expression data and retrieve relevant information about these entities from existing resources. Given iPTMnets comprehensiveness, we decided to use it as the knowledge base for our phosphorylation information retrieval. The system accepts tab-delimited expression data, using Entrez Gene IDs in the first column and $\log(2)$ of the ratios of gene expression levels determined for two distinct states (e.g. tumor vs. control) in the second column. The knowledge retrieval approach includes: (1) take the input Entrez ID, map it to all known orthologs; (2) retrieve all the phosphorylation events for these gene products from iPTMnet database; (3) identify extract kinase and substrate pairs in the input list; (4) retrieve literature evidence of the phosphorylation event from iPTMnet and provide links to the abstracts

in PubMed; (5) if there is RLIMS-P results stored in add links to RLIMS-P result page; (6) retrieve sentences from eGIFT[21].

The kinase and substrate pairs identified by the system iGep will be non-species specific. This tool is designed to provide the maximum amount of information about protein kinases and their substrates. By expanding the data retrieval to report information obtained from orthologous proteins orthologous proteins it allows the user to infer that such kinase-substrate pairs exist in the species being investigated. It is up to the user to determine if such relationships actually exist. For the rest of this paper, we will first describe other kinase and substrate network and phosphorylation databases. Then we will demonstrate our approach for identifying kinase and substrate pairs, providing further literature evidence. A case study using gene expression data of chicken LMH cells under heat stress will be loaded to the system, showing sample results and describing the functions of the web interface.

Chapter 2

RELATED WORK

2.1 Protein Phosphorylation Information Databases

With development of phosphoproteomic techniques protein phosphorylation data from multiple species is being generated by high-throughput mass spectrometry. Such data is being captured and organized into and each of these databases have their own aspects.

2.1.1 PhosphoSite Plus

PhosphoSite Plus[7] predicts phosphorylation sites in human and mouse proteins and is manually curated database covering different kinds of protein modifications including phosphorylation, acetylation, methylation, ubiquitination and O-glycosylation. Besides protein phosphorylation information, it also provides information about phospho-specific antibodies from Cell Signaling Technology that could be used for biological experiment verifications. Figure 2.1 shows protein page overview for YAP1. Sequence logo from PhosphoSitePlus is generated according to morphology pattern of modification site, which is a spatial combination of specific amino acids. A modification site is defined as the modified residue at 0 position, along with seven flanking amino acids N-terminal (from position -7 to -1) and C-terminal (from position +1 to +7). Figure 2.2 shows substrate sequence logo generated by application from PhosphoSitePlus (in this case substrate of kinase AKT1). The motif of [ST] at position 0 with R (Arg) at position -3 together shows the substrate preference of AKT1 protein. Figure 2.3 shows phosphorylation modification sites and domains for human AKT1 protein from PhosphoSitePlus multiple sequence alignment view. Ortholog residues

are from mouse, fruit fly, cow and rat. Reference evidence for the phosphorylation events are provided for further site information.

2.1.2 Phospho.ELM

The Phospho.ELM[4] resource (<http://phospho.elm.eu.org>) provides manually-curated, experimentally verified protein phosphorylation sites. There are currently 42,574 serine, threonine and tyrosine non-redundant phosphorylation sites from animal in the database. New features like structural disorder/order, accessibility information and conservation score has been implemented recently. Visualizing the conservation of particular phosphosites among species by a multiple sequence alignment useful for comparing phosphosites and generating hypothesis between species species. Figure 2.4 shows the result table from Phospho.ELM for Cyclin dependent kinase inhibitor 1B (UniProt P46527), containing phosphorylation residue, position, surrounding sequence, kinase responsible for the phosphorylation, literature evidence (PMID), type of source (HTP/LTP), conservation score, link to ELM database, binding domain for the phosphorylation residue, SMART/Pfam protein domains, IUPRED disorder score, link to PDB, and P3D accessibility score. Multiple sequence alignments shows the conservation of phosphorylation site, which is viewed by JALVIEW plugin. Phospho.ELM BLAST search allows users to submit a protein query (either UniProt Identifier/accession number or the actual sequence of the protein) to blast against the curated dataset for phosphorylated peptides with maximum of 11 amino acids.

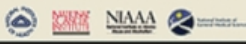
2.1.3 P3DB


P3DB[24] (Figure 2.5) has the largest collection of protein phosphorylation data from plants. The most updated version, P3DB 3.0, included altogether 47,923 phosphosites in 16,477 phosphoproteins curated across nine plant organisms from 32 studies.

2.1.4 Protein Ontology (PRO)

The Protein Ontology (PRO) represents proteins, protein isoforms, protein variants, protein modified forms and protein complexes[15]. It uses PRO terms to distinct

PhosphoSitePlus[®] from Cell Signaling Technology[®]

Home with grant support from  ABOUT PHOSPHOSITE USING PHOSPHOSITE CURATION PROCESS CONTACT

Advanced Search / Browse Functions: 

Protein Page:
YAP1 (human)

Overview

YAP1 an adaptor protein that binds to HER4 and the SH3 domain of the Yes tyrosine kinase. Associates with multiple transcription factors in the nucleus, and appears to be a co-transcriptional activator for the carboxyl-terminal fragment of ErbB-4 that translocates to the nucleus. Contains a WW domain that is found in various structural, regulatory and signaling molecules.

Protein type: Transcription, coactivator/corepressor

Cellular Component: cytoplasm; nucleus

Molecular Function: protein binding; transcription coactivator activity; transcription corepressor activity

Biological Process: cell proliferation; regulation of transcription

Reference #: **P46937** (UniProtKB)

Alt. Names/Synonyms: 65 kDa Yes-associated protein; YAP; YAP1; YAP2; YAP65; Yes-associated protein 1; Yes-associated protein 1, 65kDa; yes-associated protein 2; YKI; Yorkie homolog

Gene Symbols: YAP1

Molecular weight: 54,462 Da

Basal Isoelectric point: 5 **Predict pI for various phosphorylation states**

CST Pathways: **Hippo Signaling**

Protein-Specific Antibodies or siRNAs from Cell Signaling Technology[®]

Select Structure to View Below

YAP1

3KYS - D/B=50-171 (human)

Open Viewer

STRING | Scansite | Phospho.ELM | NetworkIN | Pfam | Phospho3D | DISEASE | Source | GeneCards | UniProt | Entrez-Gene | Ensembl Gene

Hippo Signaling

Low Cell density

High Cell density

YAP1 (human)
3KYS - D/B=50-171.

Get ChimeraX Script
Get PyMOL Script

T63-OH
S61-OH
S94-OH
K97-εNH2

☒ Atoms ☐ Solvent
☐ Spheres ☒ Surface
☒ Serine sites
☒ Threonine sites
☒ Lysine sites

Color: White
Opacity: 100%
Texture: Off
☐ Schematic
Background: Grey
☐ Antialias

LeftMouse: Rotate
Shift+Left: Scale
Ctrl+Left: Translate
RightMouse: Options...
Click: Select
c: Centre
+-: Clipping

Figure 2.1: PhosphoSitePlus Protein Page overview. Figure obtained from Hornbeck et. al. 2012[7]. This figure shows protein page overview for YAP1. Hippo Pathway from CST, shown lower left can be opened and downloaded. Lower right shows 3KYS PDF file which can be opened in a new Viewer window.

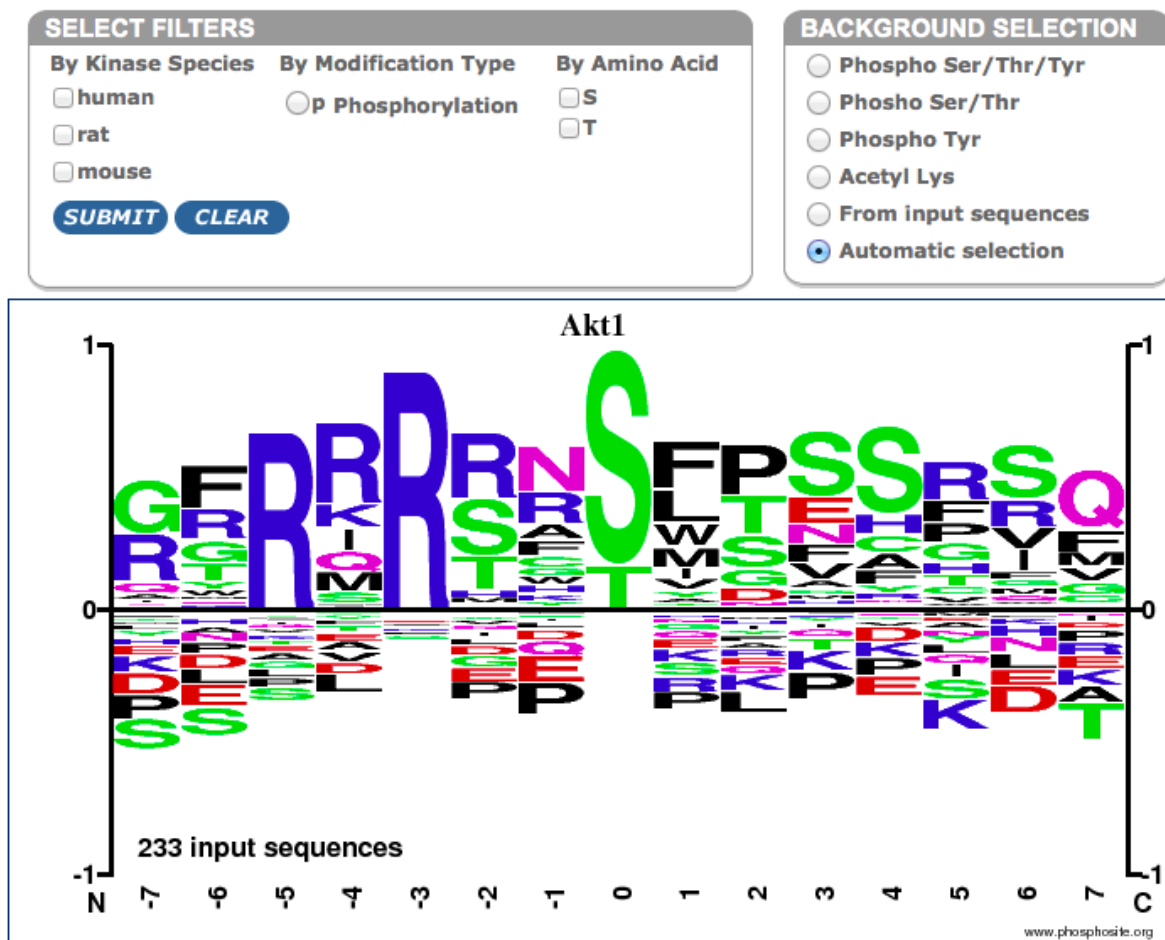


Figure 2.2: Substrate sequence logo generated by application from PhosphoSitePlus (in this case substrate of kinase AKT1). <http://www.phosphosite.org/siteSearchMotifViewAction.do?x=30&y=13&background=automatic>

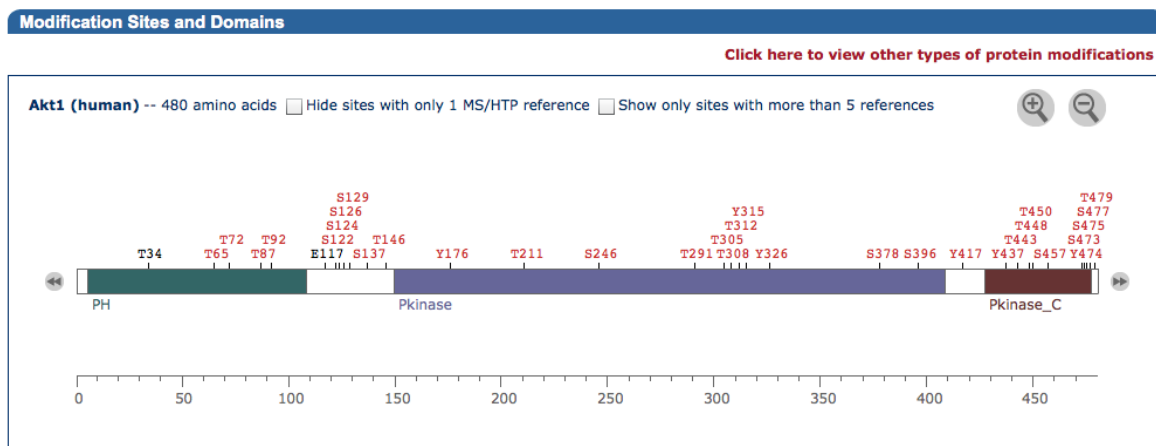


Figure 2.3: Protein phosphorylation modification sites and domains for human AKT1 protein from PhosphoSitePlus multiple sequence alignment view. <http://www.phosphosite.org/proteinAction.do?id=570&showAllSites=true>

proteins from family level, gene level, sequence level and modification level to present information to biologists. At family level, PRO terms describe proteins of a distinct gene family from the common ancestor. At gene level, PRO terms describe protein products of distinct genes. At sequence level, PRO terms describe protein products from distinct sequences according to their initial translation, e.g. sequences differed from alleles of a gene, RNA splice variants, alternative splicing, cleavage or ribosomal frameshifting. At modification level, PRO terms describe protein products from single mRNAs due to any kind of change occurred after initiation of translation and/or post-translational modifications, e.g. long isoform of smad2 can either be unmodified or be phosphorylated and contain phosphorylated residues. Thanks to these PRO terms at different levels, which was originally extended from the classification of proteins, it is possible to distinct different protein complexes or sub-complexes from different species within PRO through GO terms. Its also possible to differentiate active and inactive forms of a protein by assigning different terms to them, no matter theyre derived from post-translational modification or addition/substraction of one or more components.

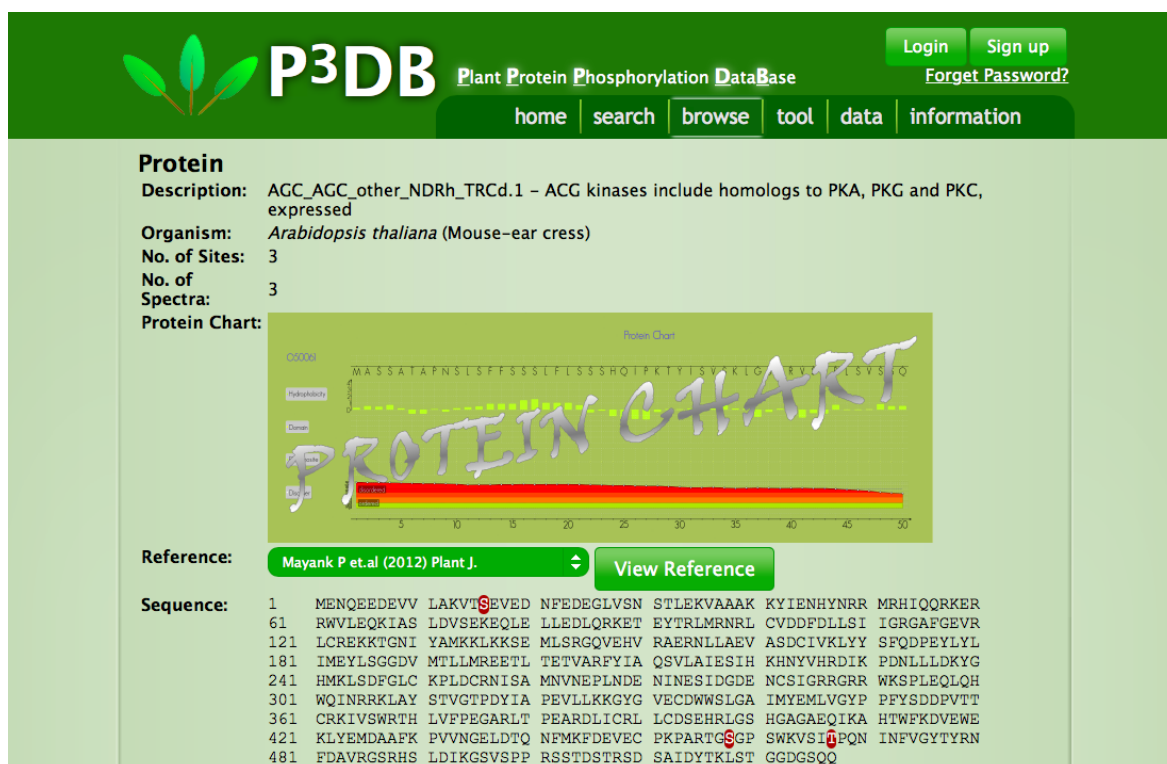


Figure 2.5: P3DB result page of *Arabidopsis thaliana* AGC kinases including homologs to PKA, PKG and PKC <http://www.p3db.org/protein.php?id=11616&ref=>

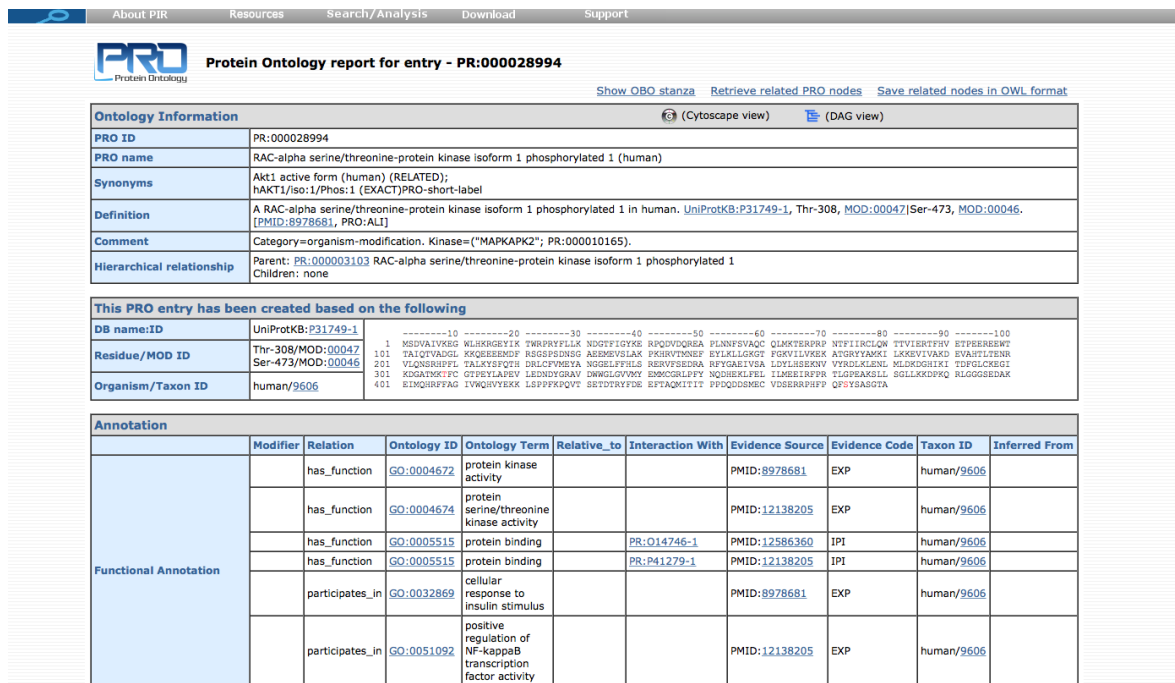


Figure 2.6: Protein Ontology report for entry phosphorylated RAC-alpha serine/threonine-protein kinase isoform 1 phosphorylated 1 and GO annotation of hAKT1/iso:1/Phos:1 http://pir.georgetown.edu/cgi-bin/pro/entry_pro?id=PR:000028994

In this way, PRO can represent protein and protein complexes from their evolutionary relationships and precisely define the objects in pathways or disease modeling. Figure 2.6 shows entity report from Protein Ontology for entry phosphorylated RAC-alpha serine/threonine-protein kinase isoform 1 phosphorylated 1 and GO annotation of hAKT1/iso:1/Phos:1.

2.2 Text-mining Approach for Protein Phosphorylation Information

With the development of phosphoproteomic technologies, protein phosphorylation data is growing rapidly and being reported by biologists and captured by bioinformatics databases. There have been various published databases or integrated sources as we mentioned above, however, literature continuously grow at a ultrafast speed, as

well as other kinds of resources. There is still abundant protein phosphorylation information scattered and remain buried in the literatures. Text-mining tools fulfill the gap, connect these information in existing databases with knowledge available in scientific literature.

2.2.1 eGIFT

eGIFT is a web-based text-mining tool, associating concept terms (iTerms) from sentences containing them with genes[21] <http://biotm.cis.udel.edu/eGIFT/index.php>. eGIFT helps out for searching PubMed for information about genes of interest by providing gene information in a form called iTerm. The iTerms are computed based on the frequency of co-occurrence of words in the literature about the a gene. Through iTerms, users can rapidly get an overall understanding of the gene without reading large amount of literature. eGIFT also can provide sentences containing the iTerms related to the specific genes of interest, which can potentially biologists gain better knowledge about unfamiliar genes. The result of eGIFT analysis is stored in a database allowing fast retrieving through database query. Figure 2.7 shows the summary report of eGIFT search of gene AKT1. Figure 2.8 shows higher ranked iTerms under category of functions and processes for gene AKT1.

2.2.2 RLIMS-P

RLIMS-P is a rule-based system designed for extracting protein phosphorylation information from literature[19]. <http://research.bioinformatics.udel.edu/rlimsp/>. It relies on three components: a database storing phosphorylation information extracted from full-length articles along with gene normalized to UniProtKB identifiers; web-interface. Additionally, RLIMS-P accepts Boolean operators (AND, OR, NOT) via keyword searches. The results are displayed in a sortable table, containing information about kinase, substrate, phosphorylation sites and PMIDs as evidence for the particular phosphorylation event. Figure 2.8 shows the RLIMS-P search result for gene AKT1 in a sortable table. Figure 2.9 shows the RLIMS-P result page for article

eGIFT

Home Gene Search Gene Analysis /Term Search Add Gene Page Guide Feedback

AKT1 - V-akt murine thymoma viral oncogene homolog 1

/Terms:

- See /Terms for this gene

Documents:

- See 9034 documents mentioning this gene (*Full Set ?*)
- See 5340 documents that eGIFT considers to be central to this gene (*About Set ?*)

Names found in the literature (click on a name to see documents containing it):

- akt • protein kinase b • rac • pkb • akt1 • akt-1 • v-akt • akt1 kinase • v-akt murine thymoma viral oncogene homolog 1 • akti • akt 1 • rac-alpha serine/threonine-protein kinase • xact • thymoma viral proto-oncogene 1 • akt-i • v-akt murine thymoma viral oncogene homolog-i • v-akt1 • akt i

Other names (not found in the literature):

- proto-oncogene c-akt • rac-pk-alpha • thymoma viral proto-oncogene i • thymoma viral proto-oncogene-1 • thymoma viral proto-oncogene-i • v akt1 • v-akt murine thymoma viral oncogene homolog i • v-akt murine thymoma viral oncogene homolog-i • v-akt1 • vakt1

Last updated: January 28, 2015

Copyright © 2008-2015 by University of Delaware | Computer and Information Sciences | Contact: **Catalina O Tudor**

Figure 2.7: eGIFT summary page of gene AKT1

eGIFT

Home Gene Search Gene Analysis /Term Search Add Gene Page Guide Feedback

/Terms for gene AKT1 - V-akt murine thymoma viral oncogene homolog 1 see documents

The following symbols -- **GOBP** **KWMP** **HY** -- represent GO terms, UniProt keywords, and species. You can click on an /Term to see sentences containing it. Clicking on the arrow next to the /Term will open a window with additional information. You can select to see /Terms co-occurring with a specific species, or you can select to see /Term of a specific category. Additionally, you can click on the /Term(s) for which you want to see documents containing them. For more information about this page, please visit the User Guide.

Select species Submit Select category Submit See documents for selected /Terms

Functions and Processes show all (142) | show top 10 | close go to top

Expand	Rank	/Term
▶ <input type="checkbox"/>	1	phosphorylation GOBP
▶ <input type="checkbox"/>	4	activator KWMP
▶ <input type="checkbox"/>	18	constitutive activity
▶ <input type="checkbox"/>	21	migration → (cell migration)
▶ <input type="checkbox"/>	22	apoptosis GOBP
▶ <input type="checkbox"/>	31	cell survival
▶ <input type="checkbox"/>	33	proliferation → (cell proliferation)
▶ <input type="checkbox"/>	36	nucleotide exchanger
▶ <input type="checkbox"/>	40	cell migration GOBP
▶ <input type="checkbox"/>	42	receptor (ror) KWMP

Figure 2.8: Higher ranked iTerms under category of functions and processes for gene AKT1


Show Selected	PubMed ID	PTM enzyme	Phosphorylated Protein (Substrate)	No. of Sentences	Text Evidence
<input type="checkbox"/>	16780593 PMC1524731	akt1, cyclin e-cdk2, akt, full, his-akt, s10 akt1, p27, akt1 (prep3), carboxy-terminal his-tagged akt1 (akt1-prep2), akt1 (prep2), full length wild type akt1, full length akt1, growth factor	hp27, s10 antibody, p27, p27 s10a (lane 4), hp27s10a, p27s10, hp27t157a, hp27s10t, bsa, p27 [30,37-42], p27t187, specific s10 antibody, caspase-9, gsk	21	
<input type="checkbox"/>	23526884 PMC3601961	pi3k, p110gamma, p110alpha, an anti-p85 antibody, pi 3-kinase, p110alpha down-regulates endogenous, p110, p110beta, recombinant p110alpha, cytokine, p85	interleukin 3 (il-3), betac subunit, tropomyosin, pi3k protein kinase, gm-csf, il-3 receptors, betac, gm-csf/il-3 betac receptor, granulocyte macrophage colony stimulating factor (gm-csf) receptors, ser585(a) pi3k, aml blasts, p85, p85 subunit, betaic, akt, gsk, pi3k, ser585, aml, pips (figure s2c), cytokine receptors, gsk-3	22	
<input type="checkbox"/>	24670416 PMC3966770	gef-h1, catenin delta-1, mapk2 (erk1), integrin beta-4, rho gtpase -activating protein 31, afap, glycogen synthase kinase-3 alpha, mick, catenin alpha-1, mrck alpha	file, mapk1, gef-h1, 14, myosin light chain proteins, mrck alpha, tak1, t202 and/or y204 and/or mapk1, fyn, raf1, rho guanine nucleotide exchange factor 12, jam-a, catenin alpha-1, map kinase p38 alpha, akt1, map kinase kinases, integrin, mapk2, afap, araf, braf, kegg, go terms dna, drug bank, mrckalpha, catenin delta-1	25	
<input type="checkbox"/>	21869924 PMC3160084	akt, mammalian target of rapamycin (mtor) complex 2 (mtorc2), foxo1, foxo3a, gsk3alpha, akt isoforms, myrakt1 and 3, mtorc1, myrakt2, myrakt3, s6 kinase 1 (s6k1), akt1, akt2, myrakt1	mdm2, akt1, foxo1/3a, foxo1, gsk3alpha, pras40, initiation factor 4e-binding protein 1 (4e-bp1), phosphoproteins 4, ribosomal protein s6 (rps6), gsk3, 4e-bp1, rps6, akt, foxo3a, gsk-3 beta, rps7, isoform-specific	12	
<input type="checkbox"/>	20361045 PMC2845649	pdk1-ifpc, pdk1, ifpn-akt1, akt, akt1, ifpn-akt1 complex	ifpn-akt1, akt, pdk1-ifpc, gsk3, akt1, s6 (s235/236), ifpn-akt1 (r25a), mtorc1	22	
<input type="checkbox"/>	21592956 PMC3137030	mtorc1, akt, pkb, mtorc2, protein kinase b, akt isoforms	pras40, akt substrate glycogen synthase kinase 3beta, akt, gsk-3 beta, akt2, akt1, p70s6k, gsk3beta, (protein kinase b), pleckstrin, akt isoforms, mtorc1 substrates p70s6k, 4e-bp1, insulin, igf-1, mtorc1 substrate p70s6k, gsk3, tsc2	17	
<input type="checkbox"/>	24949720 PMC4064967	pdk1, nahs, pp242, akt kinases, mtorc2, mtorc1, pdk1 and/or mtorc2, pip3, h2s, akt	akt, akt1, pdk1, bim, to-total mtor, mtorc2, bcl-2, p70s6k, mtor, s6k	25	
<input type="checkbox"/>	24516643 PMC3916429	pdk1, mtorc1, t-loop, erk1/2, erk1	akt, s6k, akt1, aktt308, aktthr308, post-translational	15	

Figure 2.9: RLIMS-P result page for search of gene AKT1 <http://research.bioinformatics.udel.edu/rlimsp/view.php?s=1764&abs=0>

with PMID 16780593, about protein phosphorylation information where the kinase is AKT1 protein. The participating kinases are marked in green, substrates are marked in blue and the phosphorylation sites are marked in red.

2.3 iPTMnet

iPTMnet is an integrated database resource, combining information from three different types of resources: text-mining tools (eGIFT[21], RLIMS-P[19][25], and eFIP[20]); Protein Ontology (PRO)[15] and other relevant PTM Resources (PhosphoSitePlus[7], Phospho.ELM[4], PhosphoGRID[17], UniProt KnowledgeBase (UniProtKB)[22], etc.). The PTM relations in iPTMnet is PRO curated, giving the most accurate annotation for protein forms. We're using iPTMnet database as the knowledge base for extracting phosphorylation information in this study.



CONSORTIUM MEMBER

Protein Information Resource

[About PIR](#)
[Databases](#)
[Search/Analysis](#)
[Download](#)
[Support](#)

[Previous Page](#)
[RLIMS-P Home](#)
[Login](#)

Text Evidence

Choose a specific section: All

PubMed Information

16780593

2006

Lucas P Nacusi, Robert J Sheaff

Cell Division

Full Text

No.	PTM enzyme	Substrate	Site	Sentence
1	cyclin e-cdk2	hp27	Ser-10	6 (Results 4)
2	akt1	p27	Ser, Ser-10	2 (Figure 5)
3	s10 akt1	hp27s10a	Ser	54 (Discussion 1)
4	akt1	hp27s10a	Ser	25 (Discussion 1)
5	his-akt	p27 s10a (lane 4)	Ser-10	13 (Results 7)
6	p27	s10 antibody	Ser-10	50 (Discussion 1)
7	akt1	hp27	Ser-10	12 (Results 4)
8	akt1	p27	Thr-157	13 (Results 1) 7 (Results 7) 15 (Discussion 1) 9 (Figure 2)
9	akt1, full	p27	Ser-10	2 (Figure 8)
10	p27	hp27s10a	Thr-187	42 (Discussion 1)
11	akt1 (prep3)	p27	Ser-10	15 (Figure 7)
12	akt1	s10 antibody	Ser-10	30 (Discussion 1)
13	akt, full	p27 (mouse p27)	Ser-10	11 (Figure 8)
14	his-akt	p27	Ser-10	10 (Results 7)
15	full length akt1	hp27s10a		15 (Figure 8)
16	akt1	hp27s10t		11 (Figure 5)
17	akt1 (prep2)	bsa		12 (Figure 7)
18	akt1 (prep2)	p27		12 (Figure 7)
19	full length wild type akt1	p27		6 (Figure 8)

Text Evidence

Akt1 phosphorylates both human and mouse p27 (RESULTS 1)

- Akt1 phosphorylates both human and mouse p27
- Numerous reports indicate Akt1 directly phosphorylates human p27 [30,37-42]
- Several potential sites have been suggested , based mainly on the failure of Akt1 to phosphorylate purified p27 containing site specific mutations [30,39-42] .
- Consistent with earlier reports recombinant Akt1 phosphorylated both GSK and human p27 (Figure 2B ; lanes 2 and 3) .
- However , Akt1 also phosphorylated mouse p27 and p27 T157A to a similar extent (Figure 2B ; lanes 4 and 5) , suggesting Akt1 targets a sites other than T157 .

P27S10 phosphorylation is required to target a second site (RESULTS 4)

- P27S10 phosphorylation is required to target a second site
- Phospho-peptide mapping suggests Akt1 targets multiple sites (Figure 3A) .
- Thus , we expected that mutating S10 to alanine would not completely ablate p27 phosphorylation .
- Both hp27 and hp27 S10A were phosphorylated equally well by cyclin E-CDK2 (Figure 5A , middle panel) suggesting hp27 S10A structure is not

Figure 2.10: RLIMS-P result page for article with PMID 16780593, about protein phosphorylation information where the kinase is AKT1 protein

2.4 WebGIVI

WebGIVI is a web-based application for visualizing data which can be accessed at (<http://raven.anr.udel.edu/~sunliang/webgivi/index.php>). It provides two of views for visualizing data: a force-directed layout from CytosCape.js and a concept map view from D3.js. It can be used to explore gene-iTerm pairs or a customized two column tab delimited data. It was built with a database contains iTerms for each gene from eFIPs, allowing users to explore shared iTerms among genes from an input gene list.

Chapter 3

METHOD

3.1 Overview of the study

This text-mining and visualization approach for understanding experimental data is conducted with bioinformatics techniques including text-mining, data-mining and visualization. Figure 3.1 shows the overview of the idea for this approach. With the iPTMnet integrating text-mining tools results and databases, we were able to use it as protein phosphorylation knowledge base. Figure 3.2 shows the pipeline of the study. The online tool is named as iGep (Integrating Gene Expression and Phosphorylation). It contains three parts: (i) the analyzing pipeline constructed with PHP scripts extracts phosphorylation information, kinase substrate relations, up/down-regulation computing, evidence extraction, and generating the result table; (ii) the database, which stores all protein phosphorylation information including kinases and substrates genes Entrez ID, UniProtAC, gene name, phosphorylation site, text evidence, source of the evidence, gene long name, synonyms, orthologs Entrez IDs, orthologs UniProtAC numbers preprocessed from iPTMnet database and UniProt ID translation table; (iii) a web interface, which allows users to submit their own two-column tab delimited list, where the first column is the Entrez ID of the gene, the second column is the $\log(2)$ of the ratios of gene expression determined for two distinct states (e.g. case vs. control).

Since protein kinases and substrates detected in a experiment system may have a many to many relationship the result table has three types of view formats: view by substrate; view by kinase and view by Entrez IDs. The table is also downloadable in CSV format for further analysis.

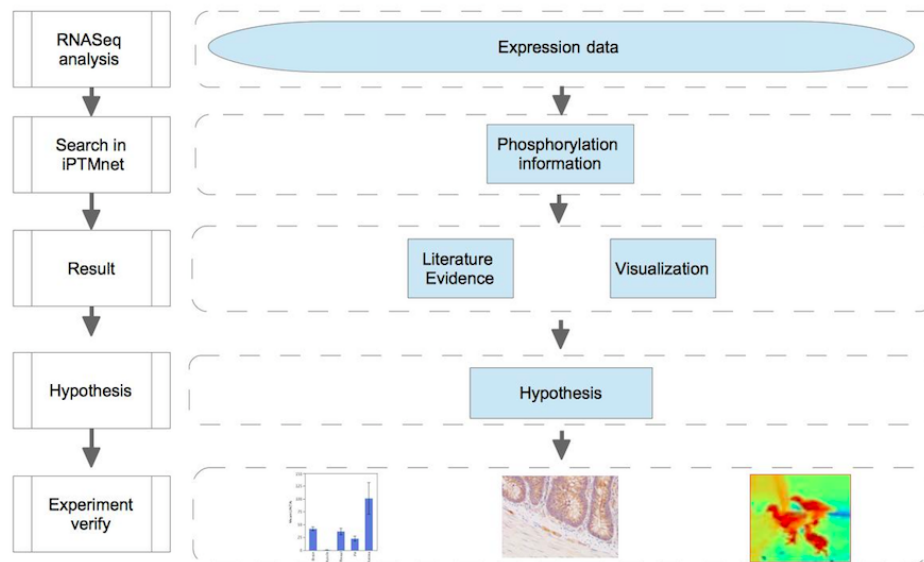


Figure 3.1: Overview of the study design

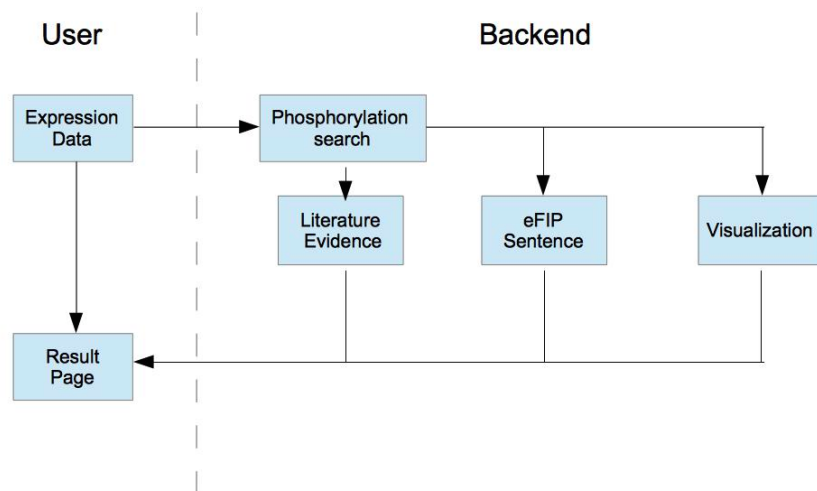


Figure 3.2: Pipeline overview

3.2 Pipeline

As phosphorylation events can be conserve between species, we retrieve protein phosphorylation information from gene orthologs of the input gene list. The iGep system consists of several customized modules to implement the goal of kinase-substrate pair relationship extraction, including (i) retrieve orthologs for the given gene from a preprocessed table in the database including information mapping from one Entrez ID to orthologs UniProt AC numbers (because iPTMnet database store protein phosphorylation information using UniProt AC number as protein identifiers); (ii) phosphorylation information, retrieve protein phosphorylation information for all the genes in the input list, and determine if their corresponding kinase or substrate was also in the input list, if so, add the kinase and substrate pair to cache and store the information; (iii) literature evidence, including the source of the evidence (RLIMS-P, PhosphoSite Plus, UniProt, etc.) from iPTMnet database, can be retrieved by kinase and substrate UniProt Accession Numbers together; (iv) sentence co-occurrence extraction, from pre-computed literature sentence results of eFIP system, using kinase and substrate gene names iTerms from eFIP, retrieve sentences containing iTerms for both kinase and substrate gene names.

3.3 Database

To support fast online computing and retrieval, we store protein phosphorylation information from iPTMnet (kinase gene name, kinase Uniprot AC number, substrate gene name, substrate UniProt AC number, source of the phosphorylation information, phosphorylation site), ortholog information (gene Entrez ID, Orthologs Entrez ID, Orthologs UniProt AC number, gene short name) computed regarding to the source translation table downloaded from UniProt website, and other related information in a SQLite database. Its also been reported by developers that SQLite often works faster than client/server RDBMS SQL databases, without having issues on concurrency. Since the database to store pre-computed information and performing searches requested from webpage, instead of intensively write on the database through webpages, we

ENTER YOUR EXPRESSION DATA HERE

Enter Your tab delimited expression data here..

[click here for sample input]

Default: Expression ratio ≥ 1.5 (log2 based) is up-regulated, expression ratio ≤ -1.5 (log2 based) is down-regulated

Customized value (optional):

Define your own up-regulated ratio (\geq) ..

Define your own down-regulated ratio (\leq) ..

Submit

Figure 3.3: Web interface of iGep

decided to use SQLite as the database engine, which is easier to install and use and further updates. Literature evidence for the phosphorylation information are stored as the PubMed ID (PMID) numbers. Data can be imported to SQLite databases in several formats, including tab delimited text files. Building indexes on the columns used as key for the table will increase performance of database dramatically.

3.4 Web Interface

The web interface of the tool is held on an Apache HTTP server. Figure 3.3 shows how the web interface we developed for the application look, a webpage will take users input data. The input format for data will be a two-column tab delimited list, where the first column is the Entrez ID of the gene, the second column is the $\log(2)$ of the ratios of gene expression determined for two distinct states (e.g. case vs. control) . After users submit their data, the iGep program, which was mostly written in PHP and JavaScript, from the back end of the application will search the iPTMnet

database for protein phosphorylation information and produce a result table in a new HTML page. Users can define the up-regulated ratio and down-regulated ratio of the expression data themselves, where as the default value for up-regulation is greater or equal to 1.5, and down-regulation is smaller or equal to -1.5. In the database, the source of protein phosphorylation is also stored with the particular phosphorylation event. A single phosphorylation event from specific kinase to specific target substrate on its specific phosphorylation site may have been reported from various sources. If the phosphorylation event is reported by RLIMS-P, we mark that in our results to provide links to RLIMS-P webpage for the analysis result page of the particular full-length. Since the kinase and substrate relationship is a many-to-many relationship, the result page provides options of view the table by kinase or substrate, as well as view the Entrez IDs that dont have corresponding kinase or substrate from the input list.

PHP library SQLite3 was used to execute SQL commands and retrieve information from database. JavaScript were used to draw the ratio bar representing how many genes from the input gene list had phosphorylation information after iGep analysis. To enable selection of different view format of the table result, either kinase centric or substrate centric, tables for the particular format were all produced, but only the kinase centric one was shown by default. Substrate centric table, no-result table were hidden by adding "display: none" in the CSS script. After selecting the needed view format of result table, the script written in JavaScript will modify the result page without refreshing it. Function for downloading the table was implemented by linking the Download button to a separate PHP script, writing the CSV file and enable downloading using fputcsv function in PHP language.

When hovering over gene names in the result table, a tooltip will popup and provide links to NCBI and UniProt pages for the particular gene. This function was implemented by using a jQuery plugin called tooltipster. Its a flexible jQuery plugin enhanced with CSS, that enables users to interact with the tooltip as well as other powerful functions. There are two ways for bounding data to the tooltip, first is define the content of tooltip in JavaScript using HTML tags, which means developer can


```

<head>

...

<script>
    $(document).ready(function() {
        $('.tooltip').tooltipster({
            contentAsHTML: true
        });
    });
</script>
</head>
<body>

<div class="tooltip" title="&lt;img src='my-image.png' /&gt; &lt;strong&gt; This text is in bold case
!&lt;/strong&gt;">
    This div has a tooltip with HTML when you hover over it!
</div>

</body>

```

Figure 3.4: Sample code at <http://iamceege.github.io/tooltipster/> demonstrating how to encode HTML markup directly by setting the title attribute.

insert things even like images and tags with text format. The other way is directly encode the HTML markup in the title attribute and set the "contentAsHTML" option to "true".

Every cell from kinase-substrate evidence column in the result table is set to show 4 PMIDs in maximum; if there is more PMID evidence for the particular phosphorylation event, they are not shown by default. Users may click on the more + link to show the hidden PMIDs and less - to toggle them and hide them again. Up-regulated gene will be marked by adding an HTML element a red triangle pointing upwards by its side and down-regulated gene will be marked by adding a green triangle pointing downwards. Genes neither up or down regulated will not have any colored triangle mark by its side. In this way, users will distinguish up/down/neutral regulated genes at a glance.

3.5 Visualization

After successfully extracting kinase-substrate pairs from the experiment data, results must be visualized for the user. For kinase and substrate pairs, it is possible to

draw networks from their relationship. To present the kinase-substrate pairs as well as other information in a more comprehensive way, we used WebGIVI, (<http://raven.anr.udel.edu/~sunliang/webgivi/index.php>) as the visualization tool to visualize the result. WebGIVI provides two kinds of view, Cytoscape view from Cytoscape.js and concept map view from D3.js. Cytoscape.js and D3.js are JavaScript libraries for data visualization and analysis, Cytoscape is very good at visualizing smaller dataset such as molecular interaction networks and biological pathways, whereas concept map view from D3.js is better at visualizing larger dataset. More advanced functions from WebGIVI includes: (i) pre- filter function, after extracting the iTerms for the given gene, users are able to filter out irrelative iTerms prior to visualizing the gene-iTerm relationship. Hovering over the genes (ii) sort function, which will sort all iTerms based on the frequency of shared genes or their alphabetical order. (iii) cut-off function, by defining the cut-off of specific frequency, iTerms meets the requirement will get highlighted. (iv) clear function, will clear up all previous selections. (v) download function, to download gene-iTerms pairs from result of WebGIVI.

3.5.1 Visualizing the kinase-substrate pairs

To visualize the kinase-substrate pair data, first copy the columns of the kinase gene name and substrate gene name from the downloaded CSV file, then paste it to the WebGIVI tool, use the custom data function. Under force-directed layout, networks of interactions from the kinase and substrate pairs identified by iGep can be easily visualized and connected together, helping users gain insights to the protein interactions beyond the one-to-many relationship shown in the table result.

3.5.2 Visualizing genes with concept terms

Since WebGIVI links genes with iTerms, submitting a list of kinase/substrate will help users make better understand of their result. To visualize genes with its concept terms iTerms computed from eGIFT, WebGIVI accepts single columned Entrez Gene ID list for analysis.

3.6 Evaluation iGep, Case Study

To evaluate the system, we conducted a case study using a gene list from LMH cell, looking for the response of heat shock from protein phosphorylation information. LMH cells were purchased from ATCC (Manassas, Virginia). Six T-75 (falcon) flasks of cells were cultured in Waymouths MB medium, along with 10% heat inactivated fetal bovine serum, coated with 0.1% gelatin. Cells were cultured at 37 Celsius degree, in 5% CO₂ and passaged every 2-3 days. Before applying heat stress to cells, they were grown to 80% confluence. These six flasks of cells were then separated to two groups, one group for experiment control and the other for heat shock. Control group were maintained at 37 Celsius degree, whereas the heat shock group were heated up to 43 Celsius degree for 2.5 hours.

Chapter 4

RESULT

4.1 iGep

iGep now is available at: http://annotation.dbi.udel.edu/text_mining/doc/pan/path/index.html.

We started with sample data could be found at http://annotation.dbi.udel.edu/text_mining/doc/pan/path/sampleInput.txt. Its a two-column tab-delimited txt file, first column is Entrez gene ID and second column is the $\log(2)$ of the ratios of gene expression determined for two distinct states (e.g. case vs. control). Users can play with the tool with this sample data, or upload their own data. On the top part of the webpage, there is legend of the result table, Figure 4.1 shows an example of the result legend. The blue ratio bar show how many of genes from the input experiment gene list has corresponding substrate or kinase expressed together and identified by iGep. Each kinase might have multiple substrates expressed in the experiment data, and vise

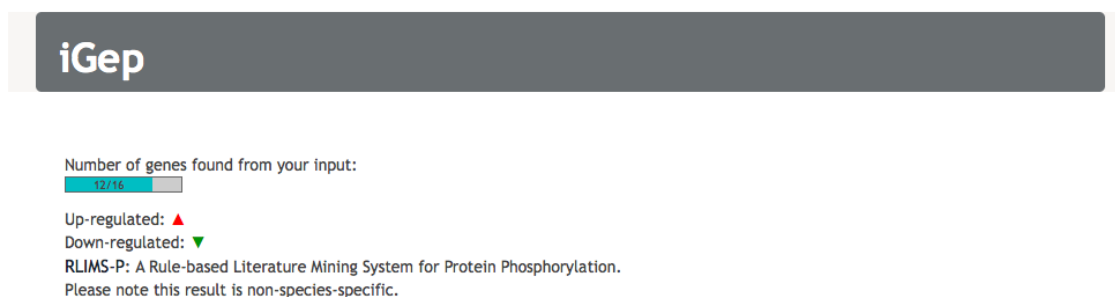


Figure 4.1: Legend of results

Table 4.1: Kinase VRK1 and its substrates identified from input data

Kinase	Substrate	Kinase-Substrate Evidence
VRK1	VRK1	11883897 RLIMS-P
	JUN	15378002 RLIMS-P
	ATF2	15105425 RLIMS-P

versa. So the actual kinase-substrate pairs identified by iGep may be greater than the number of genes found from the input having corresponding kinase or substrate.

The lower part of the webpage shows protein phosphorylation information retrieved by iGEP pipeline in a table format. Figure 4.2 shows a sample result table from iGEP. The default table layout will be kinase centric, with kinase genes listed in the first column, their corresponding substrate genes listed in the second column, literature evidence for the particular kinase-substrate pair listed in the third column, and link redirecting users to sentences extracted from eGIFT based on mentioning of the kinase- substrate pairs gene names in the fourth column.

When hovering over VRK1, an interactive tooltip will popup (Shown in Figure 4.3), containing links to the NCBI homepage of database search for VRK1 <http://www.ncbi.nlm.nih.gov/gquery/?term=VRK1> and UniProt search result for VRK1 and UniProt search result for VRK1 as gene name or protein name. This will allow users to obtain more knowledge about VRK1 by request.

Selection and download function (Figure 4.4), where users can select the layout format of the result table, either kinase centric, substrate centric, or view the genes with no information for their corresponding substrate or kinases by Entrez gene IDs. Kinases and substrates from the experiment data usually are in many-to-many relationship, but in the CSV file, we split them from one to many or many to one into one-to-one for better display of results.

From the sample result table shown in Figure 4.2, the first row looks like Table 4.1. In this particular row, means VRK1 was expressed in the experimental data.

View by ▼ Download			
Kinase	Substrate	Kinase-Substrate Evidence	Co-occurring Sentences
VRK1 ▲	VRK1 ▲	11883897 RLIMS-P	Sentence
	JUN ▲	15378002 RLIMS-P	Sentence
	ATF2 ▼	15105425 RLIMS-P	Sentence
PLK1 ▲	BUB1B ▲	17785528 RLIMS-P	Sentence
	BCL2L1	24621501 RLIMS-P	Sentence
PLK3 ▲	ATF2 ▼	21098032 RLIMS-P	Sentence
	JUN ▲	21296815 RLIMS-P	Sentence
	BCL2L1	21840391 RLIMS-P	Sentence
	VRK1 ▲	19103756 RLIMS-P 20068231	Sentence
PRKCE ▼	PRKCE ▼	18237277 18604201 18669648 18691976 11964154 16810323 11062054 less -	Sentence
PTK2 ▼	PTK2 ▼	11468287 RLIMS-P 12738990 RLIMS-P 15817454 RLIMS-P 16195476 RLIMS-P more +	Sentence
DAPK1	DAPK1	20220139 RLIMS-P 22988864 RLIMS-P 11579085 15729359 RLIMS-P	Sentence
	DAPK3	15367680 15611134 18239682	Sentence
DAPK3	DAPK3	16325270 RLIMS-P 15611134 17158456 18239682 more +	Sentence
CHEK1	CHEK1	20053762 RLIMS-P 21289283 RLIMS-P 22357623 RLIMS-P	Sentence

Figure 4.2: Sample result table



Figure 4.3: An interactive tooltip pops up when user hover over link of VRK1

	A	B	C	D	E	F	G	H	I	J	K	L
	Kinase Ent	Kinase Name	Kinase Up/Down	Substrate	Substrate	Substrate Up/Down	PMID					
1	423443	VRK1	up-regulated	423443	VRK1	up-regulated	11883897					
2	423443	VRK1	up-regulated	3725	JUN	up-regulated	15378002					
3	423443	VRK1	up-regulated	1386	ATF2	down-regulated	15105425					
4	431670	DAPK1	neutral	431670	DAPK1	neutral	20220139; 22988864; 11579085; 15729359					
5	431670	DAPK1	neutral	428342	DAPK3	neutral	15367680; 15611134; 18239682					
6	421409	PRKCE	down-regulated	421409	PRKCE	down-regulated	18237277; 18604201; 18669648; 18691976; 11964154; 16810323; 11062054					
7	428342	DAPK3	neutral	428342	DAPK3	neutral	16325270; 15611134; 17158456; 18239682; 15367680; 20854903					
8	396416	PTK2	down-regulated	396416	PTK2	down-regulated	11468287; 12738990; 15817454; 16195476; 14500712; 9790958; 19294408;					
9	5347	PLK1	up-regulated	378922	BUB1B	up-regulated	17785528					
10	5347	PLK1	up-regulated	373954	BCL2L1	neutral	24621501					
11	1263	PLK3	up-regulated	1386	ATF2	down-regulated	21098032					
12	1263	PLK3	up-regulated	3725	JUN	up-regulated	21296815					
13	1263	PLK3	up-regulated	373954	BCL2L1	neutral	21840391					
14	1263	PLK3	up-regulated	423443	VRK1	up-regulated	19103756; 20068231					

Figure 4.4: Using excel to view the CSV file downloaded

View by ▼	Download
Kinase	g Sentences
Substrate	
No Result	
Sentence	

Figure 4.5: Left: Users can select layout of table from kinase centric, substrate centric, or genes couldn't identify any phosphorylation information for their corresponding subastrate or kinase; Right: Download button to download the result to CSV format file

Text Evidence Choose a specific section:

PubMed Information

15105425	2004	Ana Sevilla, Claudio R Santos, Francisc...	The Journal of biological chemistry	Full Text
----------	------	--	-------------------------------------	-----------

Text Evidence

Abstract (ABSTRACT 1)

1 Human vaccinia-related kinase 1 (VRK1) activates the ATF2 transcriptional activity by **novel phosphorylation** on **Thr-73** and **Ser-62** and cooperates with JNK .

4 We have studied **the phosphorylation** of **the transcription factor ATF2** , which regulates gene expression by forming dimers with proteins with basic region-leucine zipper domains and recognizing cAMP-response element or AP1 sequences implicated in cellular responses to stress .

5 **VRK1 phosphorylates ATF2** mainly on **Thr-73** , stabilizing the ATF2 protein and increasing its intracellular level .

9 **VRK1 and JNK** , which **phosphorylates ATF2** in **Thr-69** and **Thr-71** , have an additive effect on ATF2 -dependent transcription at suboptimal doses .

Select/deselect: ☒kinase ☒substrate ☒site ☒phospho keywords

Gene Normalization + ?

Protein	Name	UniProtKB AC	Annotation No.
PTM enzyme	jnk	P45983/MK08_HUMAN	1
	vrk1	A1L4K2/A1L4K2_HUMAN	
Substrate	atf2	P15336/ATF2_HUMAN	1, 2, 3
		A4D7V5/A4D7V5_HUMAN	

Figure 4.6: RLIMS-P result for article PMID: 15105425

VRK1 can act as kinase, phosphorylating itself (autophosphorylate), JUN and ATF2. 11883897 is the PMID number, which is the literature evidence for VRK1s autophosphorylation. 15105425 is the PMID number, as well as the identifier for the article talking about phosphorylation event between VRK1 and ATF2. RLIMS-P link next to 15105425 links to which is the RLIMS-P result page for analysis of this article, shown in Figure 4.5. In this figure, the left panel represents information of the PubMed, PTM enzyme and substrate in this article, and gene normalization result with links redirecting users to UniProtKB knowledgebase; the right panel highlights kinase gene names in green, substrate gene name in blue, and phosphorylation site in red.

To visualize the connection between these kinase and substrate pairs in a network, we copied the kinase name column and substrate name column from the downloaded CSV file, and upload it to WebGIVI. Figure 4.7 and Figure 4.8 shows how kinases and substrates are connected. In figure 4.8, Yellow nodes are protein kinases and red nodes are protein substrates.

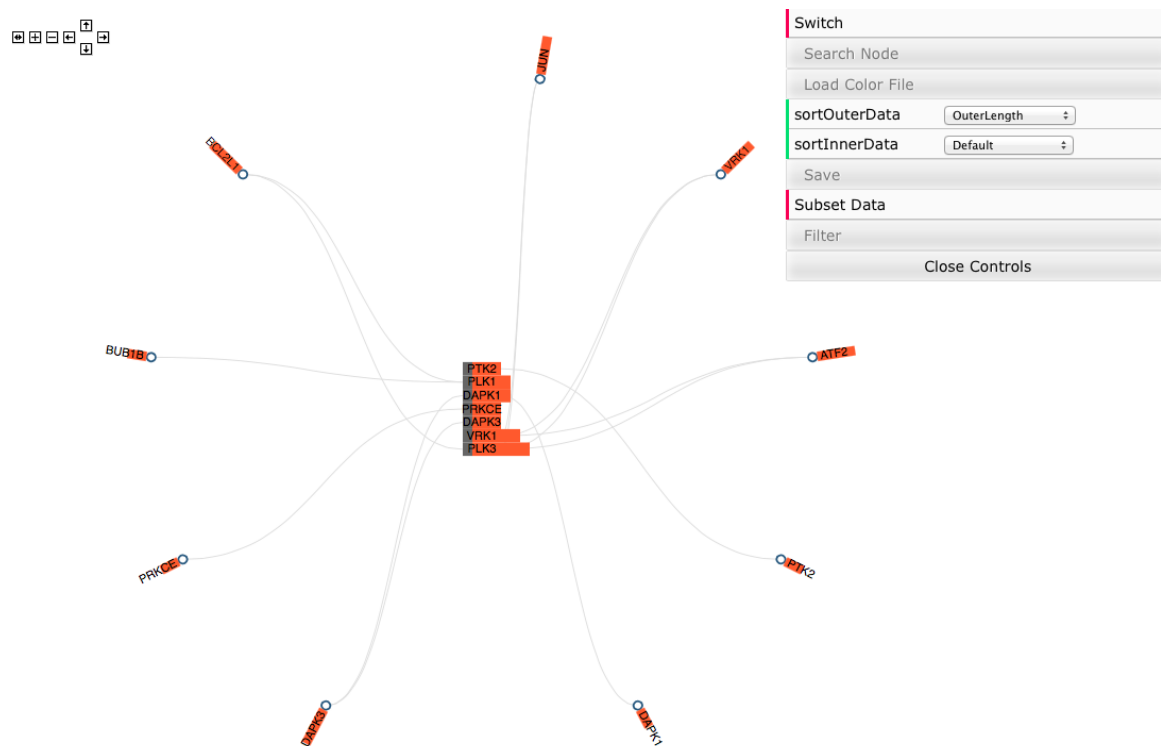


Figure 4.7: Concept Map view from WebGIVI visualizing kinase and substrate relationships in sample data

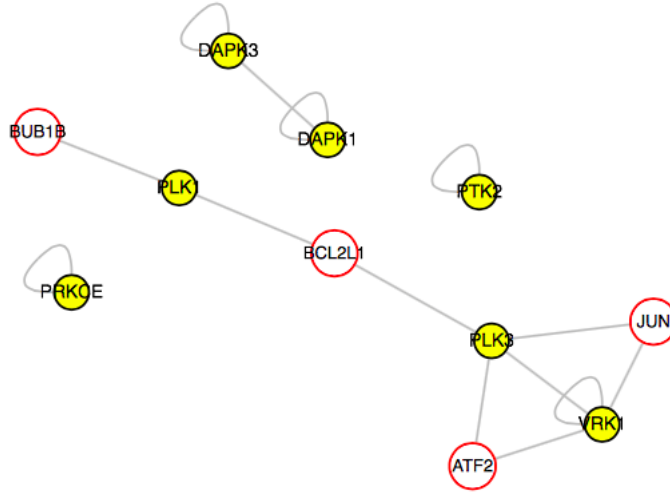


Figure 4.8: Cytoscape view from WebGIVI visualizing kinase and substrate relationships in sample data

4.2 Analysis of LMH Cell Heat Shock Response — A Case Study Using iGep

To evaluate the system, we conducted a comparative case study using iGEP. In this study, we submitted the entire expressed gene list from LMH cell, extracting the impact of heat stress on kinases and their substrates.

There are 13643 genes expressed from the LMH cell. After running iGEP, we identified 876 genes as potential kinases and kinase substrates. This yielded a total of 1679 kinase and substrate pairs. From iGEP result, there are 122 kinase-substrate pairs with either kinase or substrate, or both of them differentially expressed.

In Figure 4.9, we can see several groups of genes that might identify some key genes regulating cellular response to heat stress. In the Cytoscape network (Figure 4.9), we can see CAMK2A, PRKCB, PRKD1 and their substrates form three big groups. Calcium/calmodulin-dependent protein kinase II alpha (CAMK2A) belongs to the serine/threonine protein kinases family, is a calcium calmodulin-dependent protein kinase.

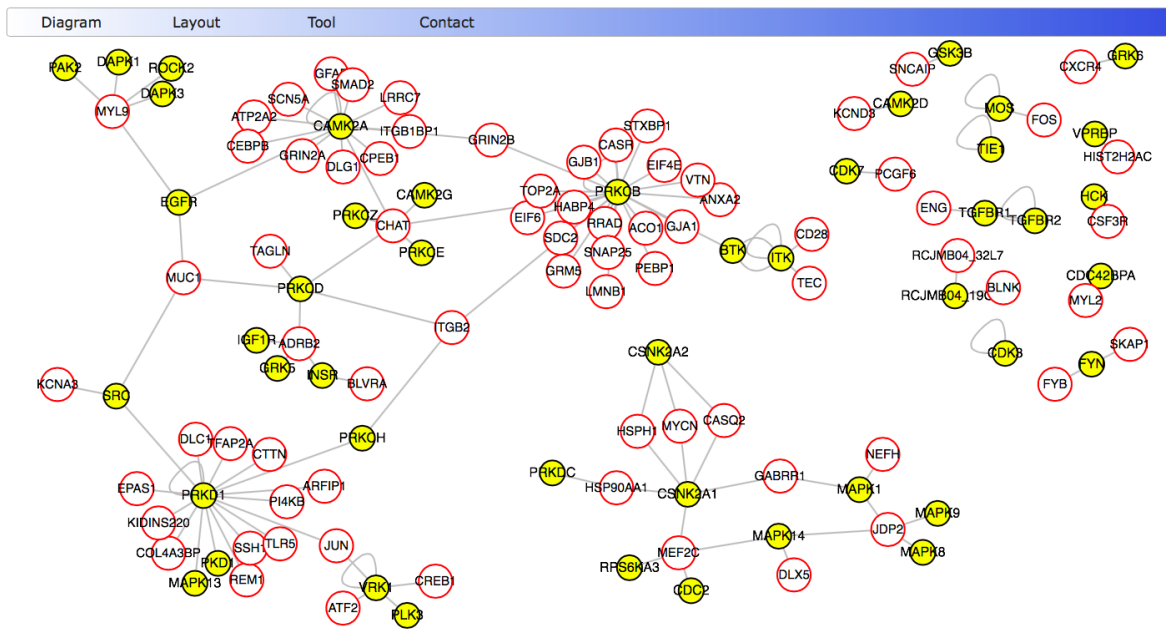


Figure 4.9: Cytoscape view from WebGIVI visualizing kinase and substrate relationships in LMH cell

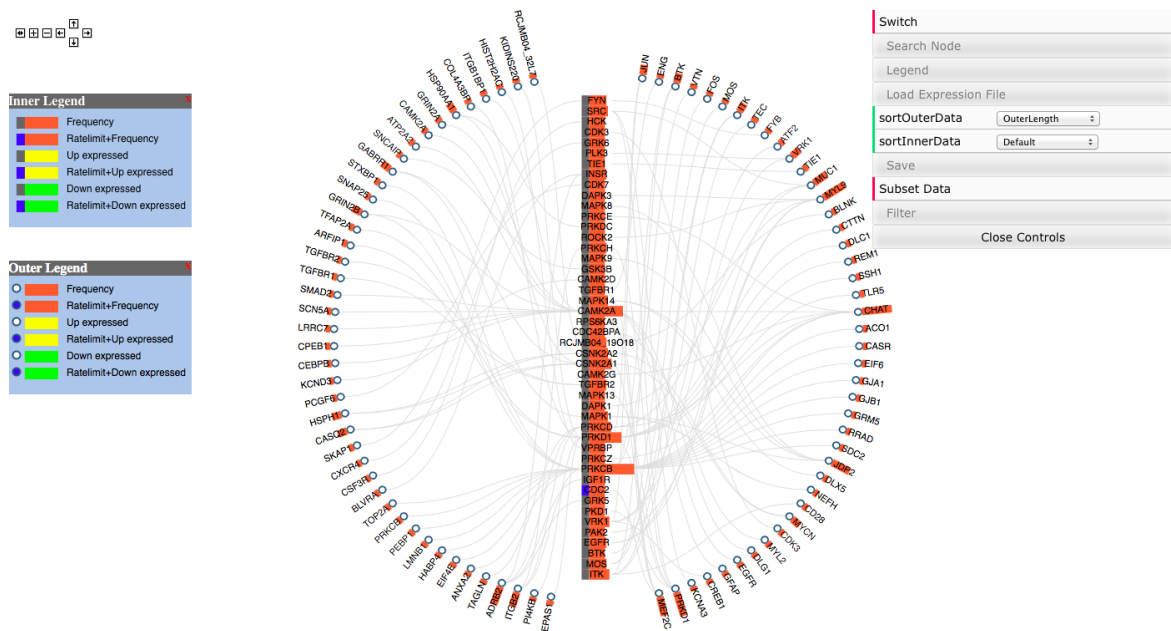


Figure 4.10: Concept map view from WebGIVI visualizing kinase and substrate relationships in LMH cell

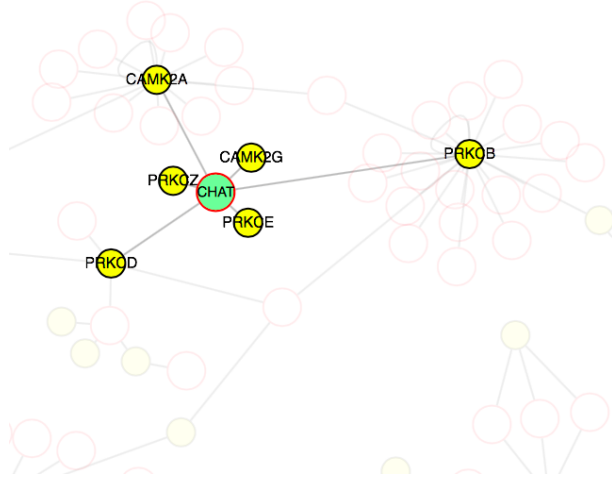


Figure 4.11: Fragment from the cytoscape view: CHAT connects several big groups in the phosphorylation network

Protein Kinase C, beta (PRKCB), Protein kinase D1 (PRKD1) all participate in calcium signaling. For example, PRKD1 phosphorylates REM1 resulting in an increase of calcium channel activity[2][10]. Possibly, under heat stress, calcium is necessary for intracellular signalling.

CAMK2A , CAMK2G and PRKCD. CHAT (choline acetyl-transferase) synthesizes the neurotransmitter acetylcholine in cholinergic neurons[6][5]. However, the physiological role of acetylcholine in non-neuron cells is unknown.

MUC1 is transmembrane glycoprotein that can be aberrantly overexpressed in carcinoma cells. Its been reported by Li Y et. al 2001[13] that, EGF-R mediates phosphorylation of MUC1, and this will induces MUC1 bind to c-SRC in cells. This relation was captured by iGEP in the visualization approach shown in Figure 4.12. PKCdelta interacts with MUC1, and phosphorylating MUC1 increases the binding of beta-catenin to E-cadherin. Hence, PKCdelta regulates interactions between MUC1 and the beta-catenin signaling pathway[16]. Since LMH cell are hepatocellular carcinoma cells, its either possible that MUC1 participates in regulating cell division and other biological processes in response to heat stress.

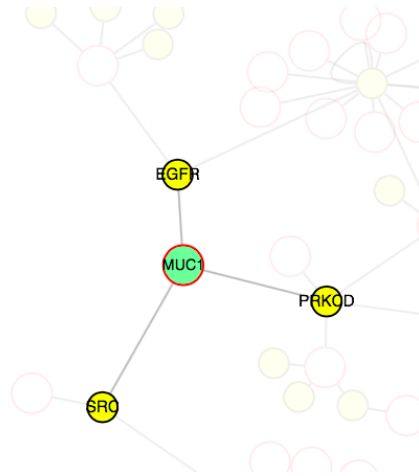


Figure 4.12: Fragment from the cytoscape view: MUC1 connects several big groups in the phosphorylation network

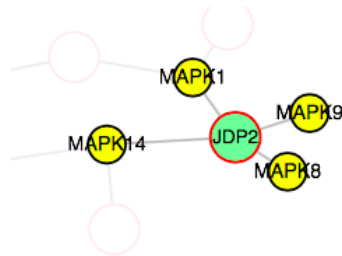


Figure 4.13: Fragment from the cytoscape view: JDP2 can be phosphorylated by MAPK1, MAPK8, MAPK9, MAPK14 expressed

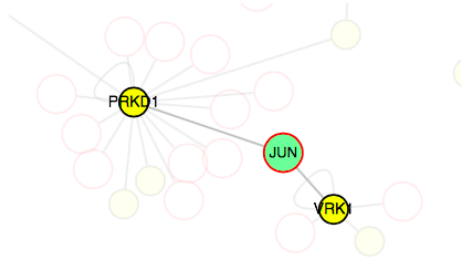


Figure 4.14: Fragment from the cytoscape view: JUN could be phosphorylated by PRKD1 and VRK1

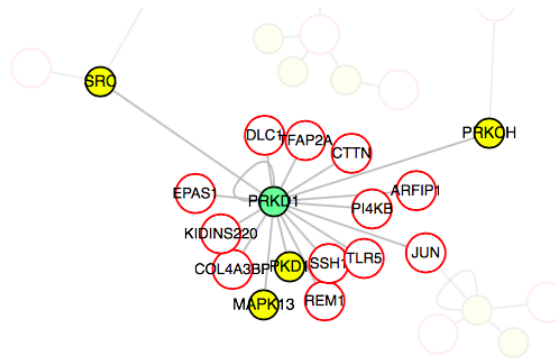


Figure 4.15: Fragment from the cytoscape view: PRKD1 with lots of its substrate proteins

Four kinases that could phosphorylate the transcription factor, Jun dimerization protein 2 (JDP2) were expressed in response to heat stress: MAPK1, MAPK8, MAPK9, and MAPK14 (Figure 4.13). Transcription factor JDP2 functions in controlling senescence and differentiation via direct interactions with histones and DNA. Conceivably, phosphorylation by these kinases could play a role in modulating JDP2 activity.

Protein kinase D could phosphorylate c-Jun protein alternative sites at N-Terminal, where c-jun protein can regulate cell cycle and apoptosis[9]. Figure 4.14 shows JUN, PRKD1, VRK1 were detected by iGep.

Figure 4.15 shows Protein kinase D1, along with numerous substrates were differentially expressed in the response to heat stress. Previous study shows that Protein kinase D1 is a stress-activated kinase, regulating various biological processes[18]. PKD1 is activated during oxidative stress, protecting cells at an early stage exposure to oxygen radicals[1]. Here in this experiment, cells are likely exposed oxidative stress as a result by heat stress.

Chapter 5

DISCUSSION AND FUTURE WORK

5.1 Discussion

Phosphorylation plays a central role in regulating many biological processes. There is no available tool for identifying kinase-substrate pairs in expression data. Although there are abundant protein phosphorylation captured and reported by databases, and updated regularly, its still difficult for biologists to scan the whole expressed gene list may have more than ten thousands of genes to retrieve kinase and substrate pairs. In this article, we described our approach, combining text mining and visualization approaches to help interpret experimental data. Biologists gain insights of biological process through getting know more about the kinase and substrates functions, as well as reading the literature evidence retrieved during analysis.

As its been reported that protein phosphorylation sites conserves not only within gene families, but also across species. Specific phosphorylation events on a particular protein, phosphosite is not always well studied because of the technical and laborious demand. The specific model biologist is using for their experiment may not been studied very well when compared to other popular species like human, mouse, E.coli, fruit fly, etc. We successfully built a web-based tool for extracting kinase- substrate pairs from the knowledgebase we built among species.

In the case study of LMH cells response to heat stress, we analyzed the whole expression gene list using iGEP. We submitted the experimental data containing a gene list expressed in the experiment (13643 genes). By going over the literature evidence for kinase-substrate pairs identified by iGEP from the input gene list, and sentences extracted based on the kinase-substrate gene name co-occurrence, we gained insights

into how heat stress may regulate phosphorylation reactions by controlling the levels of kinases and their substrates. In the result table, we were able to identify 876 genes from the input list are either kinases or kinase substrates. These can form 1679 kinase and substrate pairs due to their many-to-many relationship. Using the visualization approach, we were able to construct a protein phosphorylation network.

We also hypothesized several genes played important role in response to heat stress by regulating calcium channel activity. Pretreatment of maize seeds with calcium chloride (CaCl₂) solution, raising calcium content of maize seedlings enhances ABA-induced thermotolerance[14]; ; pretreatment *Agrostis stolonifera* (a cool-season grass, creeping bent grass) with calcium chloride (CaCl₂) will induce tolerance to subsequent heat stress[12]. This suggests that calcium levels may play a role in heat stress response.

The results also showed the close connection between oxidative stress and heat stress. Several genes were detected that might participate in regulating gene expression in response to heat stress including MUC1, JDP2, JUN, PRKD1 as well as MAPK signaling pathway. Previously, these genes were reported to be responsive for heat stress. iGEP successfully helped us propose a hypothesis that these genes might also be heat stress responsive.

However, there are still limitations to this work. The literature articles sometimes are not providing more functional information of the phosphorylation event. However, when analyzing data, after successfully identified kinase and substrate pairs from the experiment data, biologists usually tend to get more knowledge about the functional impact of the phosphorylation event. Some duplicate articles appear many-times because of a lot of mentions of protein phosphorylation information, without further explanation of any functional impact. Also as the network gets bigger it will take greater time for the data to be processed.

5.2 Future Work

PathRings is a visualization approach for pathway analysis at <http://raven.anr.udel.edu/%7Esunliang/PathRings/>. Its a web application to assist biologists

explore and analyze experimental data interactively. It implemented visualizing pathways hierarchically, and allows search pathways from Reactome. Pathway information and sub-cellular localization information of the protein will benefit biologists for both understanding data and experimental design. Since phosphorylation often regulates gene expression profile, so integrating protein phosphorylation information to PathRings will assist biologists gain better knowledge of the experiment system from pathway level. Protein phosphorylation can be treated as a kind of protein interaction, which could be easily visualized by adding arrows and define different types of protein by shape, color size, etc.

Building word cloud based on iTerm of the differentially expressed genes will provide significant functional or other information for a group of genes. Refining iTerms to particular categories, filtering out iTerms that are not informative or related can increase accuracy of word cloud.

Finally, this work mainly focuses on protein phosphorylation information and kinase-substrate pair relation extraction, however, there are other kinds of PTMs involved in biological processes as well. Expanding the tool to include other kinds of PTM will potentially help biologists gain further insights and have a deeper understanding.

Chapter 6

CONCLUSION

In conclusion, we developed a web-based tool, iGEP, with interactive web interface, combining text-mining, data-mining and visualization approach helping biologists interpret experiment data. Result of the pipeline can be presented in a kinase-centric or substrate-centric table, with marks on whether up/down regulated, and visualized in WebGIVI. It supports a download function, where data can be downloaded in a CSV format file, for further analysis. We conducted the case study for LMH cells in response to heat shock, gained interesting insights from cross-species phosphorylation information.

REFERENCES

- [1] Arunkumar Asaithambi, Arthi Kanthasamy, Hariharan Saminathan, Vellareddy Anantharam, and Anumantha G Kanthasamy. Protein kinase d1 (pkd1) activation mediates a compensatory protective response during early stages of oxidative stress-induced neuronal degeneration. *Molecular Neurodegeneration*, 6(43), 2011.
- [2] Mei Bai, Sunita Trivedi, Charles R. Lane, Yinhai Yang, Steven J. Quinn, and Edward M. Brown. Protein kinase c phosphorylation of threonine at position 888 in ca^{2+} -sensing receptor (car) inhibits coupling to ca^{2+} store release. *Journal of Biological Chemistry*, 1998.
- [3] Shyh-Shin Chiou, Sophie Sheng-Wen Wang, Deng-Chyang Wu, Ying-Chu Lin, Li-Pin Kao, Kung-Kai Kuo, Chun-Chieh Wu, Chee-Yin Chai, Cheng-Lung Steve Lin, Cheng-Yi Lee, Yu-Mei Liao, Kenly Wuputra, Ya-Han Yang, Shin-Wei Wang, Chia-Chen Ku, Yukio Nakamura, Shigeo Saito, Hitomi Hasegawa, Naoto Yamaguchi, Hiroyuki Miyoshi, Chang-Sheng Lin, Richard Eckner, and Kazunari K. Yokoyama. Control of oxidative stress and generation of induced pluripotent stem cell-like cells by jun dimerization protein 2. *Cancers*, 5:959–984, 2013.
- [4] Holger Dinkel, Claudia Chica, Allegra Via, Cathryn M Gould, Lars J Jensen, Toby J Gibson, and Francesca Diella. Phospho.elm: a database of phosphorylation sites—update 2011. *Nucleic Acids Res*, 39(Database issue):D261–7, Jan 2011.
- [5] Tomas Dobransky, Dyanne Brewer, Gilles Lajoie, and R Jane Rylett. Phosphorylation of 69-kda choline acetyltransferase at threonine 456 in response to amyloid-beta peptide 1-42. *J Biol Chem*, 278(8):5883–93, Feb 2003.
- [6] Tomas Dobransky, Wanda L. Davis, Gong-Hua Xiao, and R. Jane Rylett. Expression, purification and characterization of recombinant human choline acetyltransferase: phosphorylation of the enzyme regulates catalytic activity. *Biochem. J.*, 349:141–151, 2000.
- [7] Peter V Hornbeck, Jon M Kornhauser, Sasha Tkachev, Bin Zhang, Elzbieta Skrzypek, Beth Murray, Vaughan Latham, and Michael Sullivan. Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res*, 40(Database issue):D261–70, Jan 2012.

- [8] Yu-Chang Huang, Hitomi Hasegawa, Shin-Wei Wang, Chia-Chen Ku, Ying-Chu Lin, Shyh-Shin Chiou, Ming-Feng Hou, Deng-Chyang Wu, Eing-Mei Tsai, Shigeo Saito, Naoto Yamaguchi, and Kazunari K Yokoyama. Jun dimerization protein 2 controls senescence and differentiation via regulating histone modification. *J Biomed Biotechnol*, 2011:569034, 2011.
- [9] Cliff Hurd, Richard T Waldron, and Enrique Rozengurt. Protein kinase d complexes with c-jun n-terminal kinase via activation loop phosphorylation and phosphorylates the c-jun n-terminus. *Oncogene*, 21:2154–2160, 2002.
- [10] Bong Sook Jhun, Jin O-Uchi, Weiye Wang, Chang Hoon Ha, Jinjing Zhao, Ji Young Kim, Chelsea Wong, Robert T Dirksen, Coeli M B Lopes, and Zheng Gen Jin. Adrenergic signaling controls rgk-dependent trafficking of cardiac voltage-gated l-type ca²⁺ channels through pkd1. *Circ Res*, 110(1):59–70, Jan 2012.
- [11] Louise N Johnson. The regulation of protein phosphorylation. *Biochem Soc Trans*, 37(Pt 4):627–41, Aug 2009.
- [12] Jane Larkindale and Marc R Knight. Protection against heat stress-induced oxidative damage in arabidopsis involves calcium, abscisic acid, ethylene, and salicylic acid. *Plant Physiol*, 128(2):682–95, Feb 2002.
- [13] Y Li, J Ren, W Yu, Q Li, H Kuwahara, L Yin, K L Carraway, 3rd, and D Kufe. The epidermal growth factor receptor regulates interaction of the human df3/muc1 carcinoma antigen with c-src and beta-catenin. *J Biol Chem*, 276(38):35239–42, Sep 2001.
- [14] Gong Ming, Li Yong-Jun, and Chen Shun-Zhong. Absciscic acid-induced thermotolerance in maize seedlings is mediated by calcium and associated with antioxidant systems. *Journal of Plant Physiology*, 153:488–496, 1997.
- [15] Darren A Natale, Cecilia N Arighi, Judith A Blake, Carol J Bult, Karen R Christie, Julie Cowart, Peter D’Eustachio, Alexander D Diehl, Harold J Drabkin, Olivia Helfer, Hongzhan Huang, Anna Maria Masci, Jia Ren, Natalia V Roberts, Karen Ross, Alan Ruttenberg, Veronica Shamovsky, Barry Smith, Meher Shruti Yerramalla, Jian Zhang, Aisha AlJanahi, Irem Çelen, Cynthia Gan, Mengxi Lv, Emily Schuster-Lezell, and Cathy H Wu. Protein ontology: a controlled structured network of protein entities. *Nucleic Acids Res*, 42(Database issue):D415–21, Jan 2014.
- [16] Jian Ren, Yongqing Li, and Donald Kufe. Protein kinase c delta regulates function of the df3/muc1 carcinoma antigen in beta-catenin signaling. *J Biol Chem*, 277(20):17616–22, May 2002.
- [17] Chris Stark, Ting-Cheng Su, Ashton Breitkreutz, Pedro Lourenco, Matthew Dahabieh, Bobby-Joe Breitkreutz, Mike Tyers, and Ivan Sadowski. Phosphogrid: a

- database of experimentally verified in vivo protein phosphorylation sites from the budding yeast *saccharomyces cerevisiae*. *Database (Oxford)*, 2010:bap026, 2010.
- [18] Susan F. Steinberg. Regulation of protein kinase d1 activity. *Molecular Pharmacology*, 81:284–291, 2012.
 - [19] Manabu Torii, Gang Li, Zhiwen Li, Rose Oughtred, Francesca Diella, Irem Celen, Cecilia N Arighi, Hongzhan Huang, K Vijay-Shanker, and Cathy H Wu. Rlims-p: an online text-mining tool for literature-based extraction of protein phosphorylation information. *Database (Oxford)*, 2014, 2014.
 - [20] Catalina O Tudor, Cecilia N Arighi, Qinghua Wang, Cathy H Wu, and K Vijay-Shanker. The efip system for text mining of protein interaction networks of phosphorylated proteins. *Database (Oxford)*, 2012:bas044, 2012.
 - [21] Catalina O Tudor, Carl J Schmidt, and K Vijay-Shanker. egift: Mining gene information from the literature. *BMC Bioinformatics*, 11(418):1471–2105, 2010.
 - [22] UniProt Consortium. Activities at the universal protein resource (uniprot). *Nucleic Acids Res*, 42(Database issue):D191–8, Jan 2014.
 - [23] Sally-Anne Whiteman, Liliya Serazetdinova, Alexandra M E Jones, Dale Sanders, John Rathjen, Scott C Peck, and Frans J M Maathuis. Identification of novel proteins and phosphorylation sites in a tonoplast enriched membrane fraction of *arabidopsis thaliana*. *Proteomics*, 8(17):3536–47, Sep 2008.
 - [24] Qiuming Yao, Huangyi Ge, Shangquan Wu, Ning Zhang, Wei Chen, Chunhui Xu, Jianjiong Gao, Jay J Thelen, and Dong Xu. Pdb 3.0: From plant phosphorylation sites to protein networks. *Nucleic Acids Res*, 42(Database issue):D1206–13, Jan 2014.
 - [25] X Yuan, Z Z Hu, H T Wu, M Torii, M Narayanaswamy, K E Ravikumar, K Vijay-Shanker, and C H Wu. An online literature mining tool for protein phosphorylation. *Bioinformatics*, 22(13):1668–9, Jul 2006.