

國 立 交 通 大 學

生 物 資 訊 研 究 所

博 士 論 文

蛋 白 激 酶 與 受 質 磷 酸 化 網 路 之 建 構

Discovery of Protein Kinase-Substrate

Phosphorylation Networks



研 究 生：李 宗 夷

指 導 教 授：黃 憲 達 博 士

中 華 民 國 九 十 七 年 七 月

蛋白激酶與受質磷酸化網路之建構

Discovery of Protein Kinase-Substrate Phosphorylation
Networks

研究生：李宗夷

Student : Tzong-Yi Lee

指導教授：黃憲達 博士

Advisor : Hsien-Da Huang

國立交通大學

生物資訊研究所

博士論文

A Thesis

Submitted to Institute of Bioinformatics

College of the Biological Science & Technology

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Ph.D.

in

Bioinformatics

July 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年七月

蛋白激酶與受質磷酸化網路之建構

學生:李宗夷

指導教授:黃憲達 博士

國立交通大學 生物資訊研究所

摘要

透過蛋白激酶 (protein kinase) 所催化的蛋白質磷酸化 (protein phosphorylation) 機制是一種常見可逆的蛋白質轉譯後修飾作用，扮演著訊息傳遞路徑上的重要角色。Manning 等作者在 2002 年發現了 518 個人類蛋白激酶基因，也提供了一個蛋白質磷酸化網路研究的切入點。隨著高通量的質譜儀蛋白質體學技術，實驗驗證的蛋白質磷酸化資料也快速的增加，但是，只有 20% 的磷酸化位置有註解是被哪個蛋白激酶催化的。為了完整地探討蛋白激酶如何調控細胞內的機制，需要詳盡且精確的方法來辨識受質 (substrate) 上面的磷酸化位置是被哪個特定蛋白激酶所催化。因此我們發展了一個叫作 RegPhos 的方法，整合了電腦模型與蛋白質相關性(包含蛋白質交互作用、功能相關性以及細胞內位置)來辨識某個磷酸化位置被哪個蛋白激酶催化。為了評估 RegPhos 方法的效能，四個已知的蛋白激酶 (CDK、PKC、PIKK 和 INSR) 的磷酸化資料被用來測試是否能正確的預測作用的蛋白激酶，RegPhos 跟單純用電腦模型的方法比起來，可以改善 5 到 10% 的準確度。這些完整且準確被分析預測蛋白激酶與受質的交互作用可以被用來建構細胞內從細胞膜上的受體蛋白激酶 (receptor kinase) 到細胞核內的轉錄因子 (transcription factor) 的磷酸化網路，並且用實驗表現證據 (如:基因微陣列資料) 來檢視蛋白激酶跟受質是否有統計上顯著的相似表現行為。

Discovery of Protein Kinase-Substrate Phosphorylation Networks

Student: Tzong-Yi Lee Advisor : Dr. Hsien-Da Huang

Institute of Bioinformatics, National Chiao Tung University

Abstract

Protein phosphorylation, catalyzed by protein kinases, is a ubiquitous reversible post-translational modification (PTM) and plays a crucial role in signaling pathway. Manning *et al.* have identified 518 human kinase genes, the so-called “kinome”, that provides a starting point for comprehensive analysis of protein phosphorylation networks. With the high-throughput mass spectrometry (MS) proteomics, the number of *in vivo* phosphorylation sites is increasing rapidly. However, only 20% of the experimentally verified phosphorylation sites have the annotation of catalytic kinases. To understand how protein kinases regulate their substrates in intracellular processes, it is necessary to link these sites to specific kinases. Therefore, we propose an approach that incorporates machine learning method with protein associations (protein-protein interactions, functional associations, and subcellular localization) for identifying the catalytic kinase for each experimental phosphorylated site. Four well-annotated kinase families, such as CDK, PKC, PIKK, and INSR, are used to test the ability to correctly predict which kinases are responsible for catalyzing them. The presented approach can improve 5 - 10% predictive accuracy more than purely using machine learning method. The identified kinase-substrate interactions are used to construct the intracellular phosphorylation network starting from receptor kinases to transcription factors. Moreover, the experimental expression evidence such as time-series microarray gene expression profiles is adopted to validate the syn-expression of kinase and substrate with statistical significance.

Table of Contents

Abstract	iv
Table of Contents	v
List of Figures	viii
List of Tables	xi
Chapter 1 Introduction	1
1.1 Biological Background	1
1.1.1 Protein Post-Translational Modifications (PTMs).....	1
1.1.2 Protein Phosphorylation.....	3
1.1.3 Signal Transduction Pathway.....	6
1.1.4 Mass Spectrometry-based Identification of Protein Phosphorylation ..	8
1.1.5 Phosphoproteomics	10
1.2 Motivation.....	12
1.3 Research Goals.....	13
1.3.1 Management of Heterogeneous Phosphorylation Databases and Related Information	13
1.3.2 Identification of Kinase-Specific Phosphorylation Sites	14
1.3.3 Discovery of Protein Kinase-Substrate Phosphorylation Networks ...	14
1.4 Organization of This Dissertation	15
Chapter 2 Information Repository of Protein Post-Translational Modifications 16	
2.1 Introduction.....	16
2.1.1 Protein Solvent Accessibility	17
2.1.2 Protein Intrinsic Disorder.....	17
2.1.3 Subcellular Localization	22
2.2 Related Works	23
2.3 Motivation and the Specific Aim	32
2.4 Materials and Methods.....	33
2.4.1 Integration of External PTM Databases.....	34
2.4.2 Computational Annotation of PTM Sites.....	35
2.4.3 Structural and Functional Annotations	41
2.4.4 Benchmark of PTM Prediction	43
2.5 Results.....	46
2.5.1 Performance of PTM Computational Models.....	46
2.5.2 Data Statistics.....	49
2.5.3 Data Access.....	51
2.5.4 Characteristics.....	54

2.6 Summary	56
Chapter 3 Identification of Kinase-Specific Phosphorylation Sites	58
3.1 Introduction.....	58
3.2 Related Works	63
3.2.1 Machine Learning Methods	63
3.2.2 Phosphorylation Site Prediction.....	70
3.3 Motivation and The Specific Aim.....	76
3.4 Materials and Methods.....	77
3.4.1 Data Preprocessing.....	77
3.4.2 Feature Extraction and Coding	82
3.4.3 Model Learning and Evaluation.....	83
3.4.4 Independent Test	84
3.5 Results.....	85
3.5.1 Structural Investigation of Phosphorylation Sites.....	85
3.5.2 Predictive Performance	91
3.5.3 Predictive Performance of Independent Test	94
3.5.4 Web-based Prediction Tool	95
3.6 Discussions	96
3.6.1 Kinase-specific Groups with Similar Consensus Motif.....	96
3.6.2 Comparison with Other Phosphorylation Prediction Tools	98
3.7 Summary	100
Chapter 4 Discovery of Protein Kinase-Substrate Phosphorylation Networks .	101
4.1 Introduction.....	101
4.2 Related Works	104
4.2.1 Discovery of Human Phosphorylation Networks	105
4.2.2 Human Kinase Interactome Resource.....	106
4.2.3 Modeling of Signal Transduction Networks.....	107
4.3 Motivation and Specific Aim.....	109
4.4 Materials	110
4.4.1 Protein Kinase and Phosphorylation Site Resource.....	110
4.4.2 Protein-Protein Interaction Databases.....	114
4.4.3 Functional Association Databases.....	115
4.4.4 Protein Subcellular Localization Databases.....	116
4.4.5 Gene Expression Database.....	119
4.5 Method	120
4.5.1 Identification of Kinase-Substrate Interactions	121
4.5.2 Construction of Phosphorylation Network	129
4.5.3 Expression Profile of Kinase and Substrate Genes.....	131

4.6 Results.....	135
4.6.1 Protein Kinases, Phosphoproteins, and Interacting Proteins	135
4.6.2 Subcellular Localization of Protein Kinases and Substrates.....	138
4.6.3 Expression Analysis of Kinase and Substrate.....	142
4.6.4 Predictive Performance	146
4.6.5 Statistics of Discovered Kinase-specific Substrate Interactions.....	147
4.7 Case Study	148
4.8 Web-based System of RegPhos.....	152
4.9 Summary	155
Chapter 5 Discussions.....	156
5.1 Characteristics.....	156
5.2 Limitations	158
5.2.1 How Reliable are Protein-Protein Interaction?	158
5.2.2 Time Complexity and Path Length of Signaling Pathway Construction	158
5.2.3 Visualization of Complex Phosphorylation Network	159
5.3 Perspectives.....	161
5.3.1 Phosphorylation Sites on Various Protein Isoforms.....	161
5.3.2 Downstream Genes of Transcription Factors.....	162
5.3.1 Dephosphorylation and Phosphatase	162
Chapter 6 Conclusion	164
References	166
Appendix I – Human Kinase Families	173

List of Figures

Figure 1.1 Schematic representation of several common post-translational modifications.	2
Figure 1.2 Schematic representation of protein phosphorylation.	4
Figure 1.3 Chemical formula of serine, threonine, and tyrosine (Lehninger <i>et al.</i> , 2005).	4
Figure 1.4 Activation of the insulin-receptor Tyr kinase by autophosphorylation (Lehninger <i>et al.</i> , 2005).	5
Figure 1.5 Overview of signal transduction pathways.	7
Figure 1.6 Insulin-induced signal transduction (Lehninger <i>et al.</i> , 2005).	8
Figure 1.7 Generic mass spectrometry (MS)-based proteomics experiment (Aebersold <i>et al.</i> , 2003).	9
Figure 1.8 Example of phosphopeptide MS/MS spectra.	10
Figure 1.9 Combined large-scale approaches to unravel phosphorylation driven signaling networks (Bentem <i>et al.</i> , 2007).	11
Figure 1.10 Schematic representation of dissertation organization.	15
Figure 2.1 Measurement of protein solvent accessibility.	17
Figure 2.2 Disorder in Calcineurin.	18
Figure 2.3 Example of a binding region and its positions relative to the regions of PONDR predicted disorder score (Garner, <i>et al.</i> , 1999).	19
Figure 2.4 Eukaryotic cellular compartments.	22
Figure 2.5 Different forms of ubiquitin and ubiquitin-modified proteins (Chernorudskiy, <i>et al.</i> , 2007).	29
Figure 2.6 A list of instances for the PDB file 1A52 (Zanzoni, A., <i>et al.</i> , 2007).	30
Figure 2.7 System flow for constructing dbPTM.	33
Figure 2.8 An example of 9-mer (window length n is set to 4) phosphorylated peptides and sequence logo.	35
Figure 2.9 System flow of KinasePhos-like method.	36
Figure 2.10 A 5x5 contingency table between two positions in PTM site.	38
Figure 2.11 The optimization of the threshold of the HMM bit score in the model of phosphorylated serine which is catalyzed by PKA.	40
Figure 2.12 Flowchart of constructing PTM benchmark dataset.	43
Figure 2.13 The profile hidden Markov model of N-linked (glcNAc) asparagine.	47
Figure 2.14 Search interface of dbPTM.	52
Figure 2.15 Browse interface of dbPTM.	53
Figure 2.16 Example of PTM site located in orthologous conserved region.	54

Figure 2.17 Example post-translational modification reactions and structures of.....	57
Figure 3.1 Consensus sequences for protein kinases (Lehninger <i>et al.</i> , 2005).	58
Figure 3.2 Structural environment of reversible modifications (Pang <i>et al.</i> , 2007)....	59
Figure 3.3 Phosphorylated insulin-receptor Tyr kinase (PDB: 1IR3) (Lehninger <i>et al.</i> , 2006).	60
Figure 3.4 An example of decision tree.	64
Figure 3.5 A schematic diagram of artificial neural network. Each circle in the hidden and output layer is a computation element known as a neuron (Haykin <i>et al.</i> , 1999).	66
Figure 3.6 An example of small profile HMM representing a short multiple alignment of five sequences with three consensus columns (Eddy <i>et al.</i> , 1998).....	67
Figure 3.7 Basic concept of support vector machine.	68
Figure 3.8 Principle of hyperplane in support vector machine.	69
Figure 3.9 The system flow of KinasePhos 1.0.	74
Figure 3.10 The system flow of KinasePhos 2.0.	75
Figure 3.11 The system flow of kinase-specific phosphorylation site prediction.....	77
Figure 3.12 Comparison of flanking amino acids between phosphorylated and non-phosphorylated sites.	86
Figure 3.13 Predictive accuracy of PKA, PKC, CK2, CDK, Src and EGFR models trained with different training features, based on various window sizes.	91
Figure 3.14 Web interface of KinasePhos.....	95
Figure 4.1 Kinase distribution by major groups in human and model systems (Manning <i>et al.</i> , 2002).	102
Figure 4.2 Phylogenetic tree of human kinome (Manning <i>et al.</i> , 2002).....	104
Figure 4.3 Effects of including network context (Linding <i>et al.</i> , 2007).	106
Figure 4.4 Annotation and visualization of PhosphoPOINT (Yang <i>et al.</i> , 2008).....	107
Figure 4.5 MAPK signal transduction pathways in yeast (Roberts <i>et al.</i> , 2000).....	108
Figure 4.6 System architecture of RegPhos.	121
Figure 4.7 System flow of identification of kinase-substrate interactions.....	122
Figure 4.8 Pseudocode of breadth-first search (BFS) algorithm.	125
Figure 4.9 Schematic representation of Cosine similarity between two vectors.	126
Figure 4.10 Schematic representation of phosphorylation network.....	129
Figure 4.11 Comparison of clustering results between Euclidean distance and Pearson correlation distance strategies.	132
Figure 4.12 The schematic representation of kinase, interacting proteins, and phosphoproteins.	135
Figure 4.13 Subcellular localization preference of kinase family and their substrates.	140

Figure 4.14 Comparison of Pearson correlation coefficient distribution between background gene pairs and kinase-substrate pairs.	142
Figure 4.15 Distribution of Pearson correlation coefficients of PKA-substrate pairs, CDC2-substrate pairs, and EGFR-substrate pairs based on 98 microarray series.	143
Figure 4.16 Distribution of Pearson correlation coefficients of PKA-substrate pairs, CDC2-substrate pairs, and EGFR-substrate pairs based on time-coursed microarray data.	144
Figure 4.17 Effects of including protein associations.	147
Figure 4.18 Example of the discovered phosphorylation networks.	148
Figure 4.19 Example of RegPhos-identified kinase-specific phosphorylation sites.	149
Figure 4.20 Validation of the RegPhos-identified kinase-specific phosphorylation sites using HPRD annotation.	150
Figure 4.21 Graphical visualization of substrate protein with catalytic kinases.	152
Figure 4.22 The expression profile of kinase and substrate genes.	153
Figure 4.23 Example of insulin signaling network in the construction of phosphorylation network.	154
Figure 5.1 Comparison of network visualization between pure interactions and complex interactions.	160
Figure 5.2 Schematic representation of phosphorylation site located in alternatively spliced exon.	161

List of Tables

Table 1.1 Some common and important post-translational modifications (Mann, M. and O.N. Jensen, 2003).....	3
Table 2.1 Summary of the web servers offering prediction of intrinsically disordered proteins.....	21
Table 2.2 Summary of PTM resource.	24
Table 2.3 Data statistics of the integrated PTM resource.....	34
Table 2.4 Comparisons between KinasePhos, NetPhos, DISPHOS and rBPNN.....	35
Table 2.5 The amino acids group used in MDD.....	37
Table 2.6 List of the integrated external databases and programs for structural and functional annotations.	42
Table 2.7 Several representative PTM prediction servers.	44
Table 2.8 Parameters and predictive performance of the PTM computational models.	48
Table 2.9 Data statistics of dbPTM.	49
Table 2.10 The statistics of the putative phosphorylation sites, sulfation sites, and glycosylation sites with different thresholds of the Accessible Surface Area (ASA) of residues.	50
Table 2.11 The statistics of literatures extracted from release 55.0 of Swiss-Prot knowledgebase in several common PTMs.....	51
Table 2.12 Advances and improvements in current dbPTM.	55
Table 3.1 The statistics of structural information in phosphorylated serine, threonine and tyrosine.....	62
Table 3.2 List of the previously developed phosphorylation site prediction tools.....	72
Table 3.3 The statistics of phosphorylation sites obtained from Phospho.ELM and Swiss-Prot.	78
Table 3.4 Statistics of non-redundant kinase-specific phosphorylation sites in Swiss-Prot and Phospho.ELM.	79
Table 3.5 Structural features of kinase-specific groups.	88
Table 3.6 Average cross-validation performance of several common kinase-specific groups with training features which reach highest accuracy.	93
Table 3.7 Performance of independent test in several common kinase-specific groups.	94
Table 3.8 The cross predictive specificity of the kinase-specific models with similar substrate motif.....	97
Table 3.9 Comparison of KinasePhos 3.0 with PredPhospho, GPS, PPSP,	

MetaPredictor, KinasePhos 1.0, and KinasePhos 2.0.	99
Table 4.1 Statistics of integrated experimental protein phosphorylation site databases.	111
Table 4.2 List of representative kinase families containing more than 10 substrates.	112
Table 4.3 Statistics of integrated protein-protein interaction databases.	114
Table 4.4 Statistics of integrated functional association databases.	116
Table 4.5 List of public databases of protein subcellular localization.	117
Table 4.6 List of human gene microarray platform of GEO used in this work.	119
Table 4.7 Statistics of integrated experimental protein phosphorylation sites.	135
Table 4.8 Statistics of kinases and their interacting proteins.	136
Table 4.9 Statistics of kinases and their interacting proteins and functionally associated proteins.	136
Table 4.10 The protein interacting neighbor of several representative human kinase families.	137
Table 4.11 Subcellular localization of human proteins, kinases and substrates.	139
Table 4.12 Subcellular localization of human kinase-specific substrates.	141
Table 4.13 Predictive performance of purely SVM-based prediction (KinasePhos).	145
Table 4.14 Cross classifying specificity among PKC, CDK, PIKK, and INSR families based on KinasePhos method.	146
Table 5.1 Comparison between RegPhos and NetworKIN.	157

Chapter 1 Introduction

Protein phosphorylation catalyzed by protein kinase is a ubiquitous reversible post-translational modification (PTM) found in eukaryotes as well as prokaryotes. With the increasing number of *in vivo* phosphorylation sites have been identified, the desire of map the network of protein kinase and substrate has motivated. To understand how protein kinases regulate their substrates in intracellular processes, it is necessary to link these sites to specific kinases. In this dissertation, we focus on the integration of heterogeneous phosphorylation site databases (**Chapter 2**), identification of kinase-specific phosphorylation sites (**Chapter 3**), and systematic discovery of kinase-substrate interactions in protein phosphorylation networks (**Chapter 4**). The comprehensive kinase-substrate interactions are used to construct the intracellular phosphorylation network starting from receptor kinases to transcription factors. Moreover, the experimental expression evidence such as time-series microarray gene expression profiles is adopted to validate the syn-expression of kinase and substrate with statistical significance.

1.1 Biological Background

Protein post-translational modifications (PTMs), involving several chemical groups such as acetyl, methyl, phosphoryl, hydroxyl, glycans, and lipids covalently attach to individual amino acid, alter protein's biochemical natures significantly and play key roles in a wide variety of cellular processes. Studies suggest that one-third to one-half of all proteins are modified by phosphorylation [1]. In signal transduction pathways, reversible phosphorylation is essential for the maintenance of signaling amplitude, duration and specificity. Until recently, high-throughput mass spectrometry-based method is widely used to identify the phosphopeptides with specific phosphorylated site. Therefore, the increasing number of experimentally verified phosphorylation sites can be adopted to investigate the systems biology of kinase and substrate in detail.

1.1.1 Protein Post-Translational Modifications (PTMs)

Protein Post-Translational Modification (PTM) is an extremely important cellular control

mechanism because it may alter proteins' physical and chemical properties, folding, conformation distribution, stability, activity, and consequently, their functions [2]. Several chemical groups such as acetyl, methyl, phosphoryl, hydroxyl, glycans, and lipids covalently attach to individual amino acids (**Figure 1.1**), alter protein's biochemical natures significantly and play key roles in a wide variety of cellular processes. Examples of the biological effects of protein modifications include phosphorylation for signal transduction, attachment of fatty acids for membrane anchoring and association, and glycosylation for changing protein half-life, targeting substrates, and promoting cell-cell and cell-matrix interactions. Although the modification of amino acids does occur before, during and after the said amino acids are incorporated into proteins by ribosomes, they are usually referred to misleadingly as post-translational modifications.

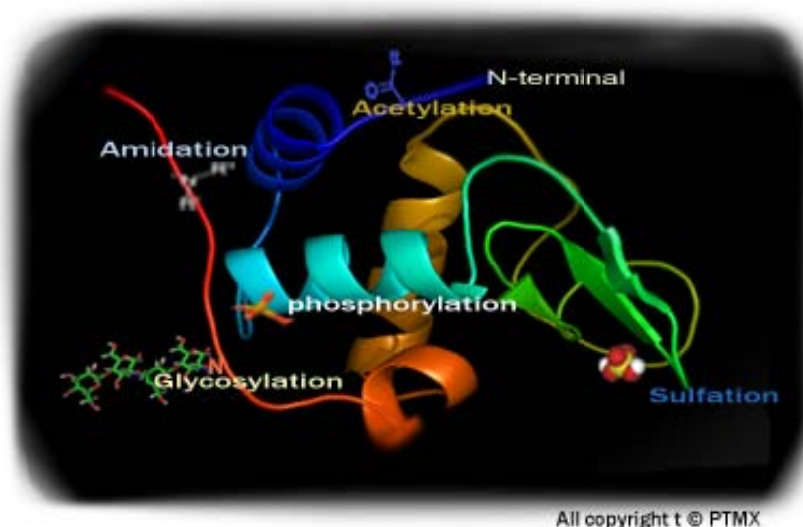


Figure 1.1 Schematic representation of several common post-translational modifications.

PTMs arise from the cleaving or forming of covalent bonds and can be classified into three categories based on the following processes: cleavage (including preand propeptide processing, initiator methionine removal, C-terminal processing), linkage (including attachment of chemical groups from the simple such as acetyl, methyl, phosphoryl, or hydroxyl, to more complex entities such as glycans or lipids) and cross-linking (including disulphide, thioether, and thioester bonds) [3]. By the statistics of RESID [4] modification database, there are more than 400 types of PTMs been discovered. The list of several common and important post-translational modifications is shown in **Table 1.1** [5] which contains mass difference, modified residues, occurring position, and description. In this dissertation, we

focus on the investigation of protein phosphorylation.

Table 1.1 Some common and important post-translational modifications (Mann, M. and O.N. Jensen, 2003).

PTM type	Δ Mass (Da)	Modified residue	Position	Description
Glycosylation				
O-linked (O-GlcNAc)	>800	S,T	anywhere	Reversible, cell-cell interaction and regulation of proteins
N-linked	203.2			
N-linked	>800	N		
Phosphorylation	79.98	S,T,Y,H,D	anywhere	Reversible, regulation of protein activity, signaling
Acetylation	42.04	S	N-term	Reversible, protein stability, regulation of protein function
		K	anywhere	
Methylation	14.03	K	anywhere	Regulation of gene expression, protein stability
Acylation				Reversible, cellular localization to membrane
farnesylation	204.36	C	anywhere	
myristoylation	210.36	G	N-term	
		K	Anywhere	
palmitoylation	238.41	C (S,T,K)	anywhere	
Hydroxyproline	16.00	P	anywhere	Protein stability and protein-ligand interactions
Deamidation	0.98	N,Q	anywhere	N to D, Q to E, possible regulator of protein-ligand and protein-protein interactions, also a common chemical artifact
Nitration	45.0	Y		Oxidative damage during inflammation
S-Nitrosylation	29.0	C	anywhere	
Ubiquitination	>1,000	K	anywhere	Reversible/irreversible, destruction signal,
Sumoylation		K	[ILFV]K.D	
Sulfation	79.96	Y	anywhere	Modulator of protein-protein and receptor-ligand interactions
Glycosylphosphatidylinositol (GPI) anchor,	>1,000	S,N,C	C-term	Membrane tethering of enzymes and receptors, mainly to outer leaflet of plasma membrane

1.1.2 Protein Phosphorylation

Post-translational phosphorylation is one of the most common protein modifications; one-third to one-half of all proteins in a eukaryotic cell are phosphorylated. Phosphoserine, threonine and tyrosine residues play critical roles in the regulation of many cellular processes. As shown in **Figure 1.2**, the catalytic site of a protein kinase hydrolyzes adenosine triphosphate (ATP) and transfers a phosphate moiety to the acceptor residue (S, T, Y in eukaryotes) in the substrate protein (**Figure 1.3**).

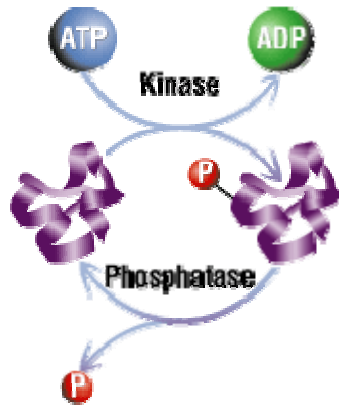


Figure 1.2 Schematic representation of protein phosphorylation.

In eukaryotes, protein phosphorylation is probably the most important regulatory event. Many enzymes and receptors are switched "on" or "off" by phosphorylation and dephosphorylation. The attachment of phosphoryl groups to specific amino acid residues of a protein is catalyzed by **protein kinases**; removal of phosphoryl groups is catalyzed by **protein phosphatases**. Phosphorylation is catalyzed by various specific protein kinases, whereas phosphatases *dephosphorylate*. Adding a phosphoryl (PO_3) to a polar R group of an amino acid might not seem like it would do much to a protein, but it can actually turn a nonpolar hydrophobic protein into a polar and extremely hydrophilic molecule. Phosphoserine, threonine, tyrosine residues in observed in eukaryotes, and histidine residue in observed in prokaryotes.

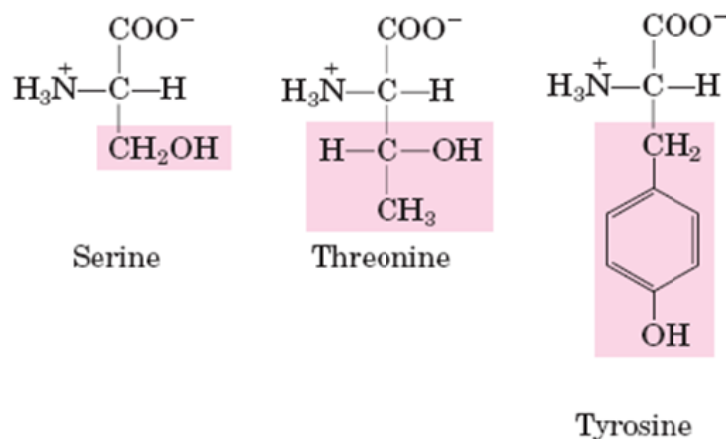


Figure 1.3 Chemical formula of serine, threonine, and tyrosine (Lehninger *et al.*, 2005).

An example of the important role that phosphorylation plays is the p53 tumor suppressor gene, which—when active—stimulates transcription of genes that suppress the cell cycle,

even to the extent that it undergoes apoptosis. However, this activity should be limited to situations where the cell is damaged or physiology is disturbed. To this end, the p53 protein is extensively regulated. In fact, p53 contains more than 18 different phosphorylation sites.

In **Figure 1.4a**, the inactive form of the Tyr kinase domain (PDB ID 1IRK), the activation loop (blue) sits in the active site, and none of the critical Tyr residues (black and red ball-and-stick structures) are phosphorylated [6]. This conformation is stabilized by hydrogen bonding between Tyr-1162 and Asp-1132. When insulin binds to the α chains of insulin receptors (**Figure 1.4b**), the Tyr kinase of each β subunit of the dimer phosphorylates three Tyr residues (Tyr-1158, Tyr-1162, and Tyr-1163) on the other β subunit (shown here; PDB ID 1IR3). (Phosphoryl groups are depicted here as an orange space-filling phosphorus atom and red ball-and-stick oxygen atoms.) The effect of introducing three highly charged P –Tyr residues is to force a 30 Å change in the position of the activation loop, away from the substrate-binding site, which becomes available to bind to and phosphorylate a target protein, shown here as a red arrow.

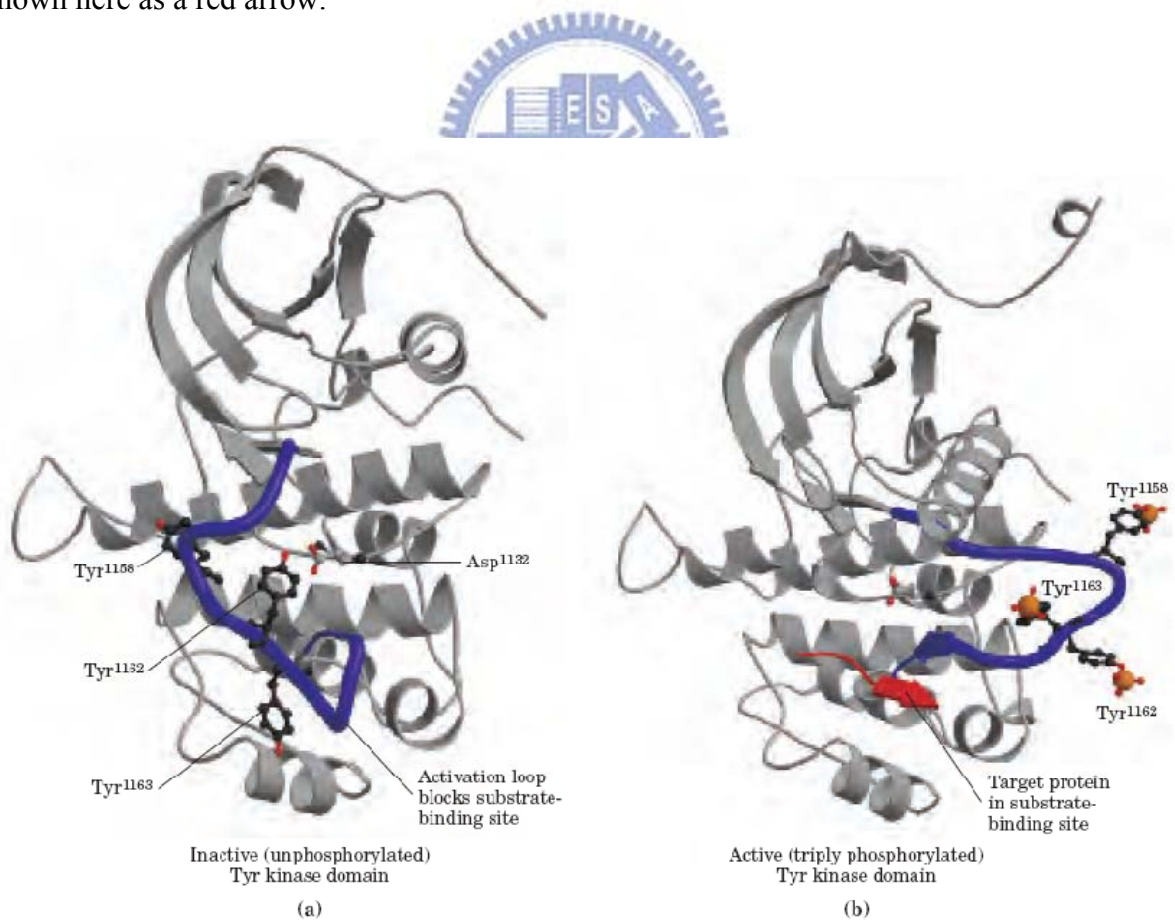


Figure 1.4 Activation of the insulin-receptor Tyr kinase by autophosphorylation (Lehninger *et al.*, 2005).

1.1.3 Signal Transduction Pathway

In biology, signal transduction refers to any process by which a cell converts one kind of signal or stimulus into another. Most processes of signal transduction involve ordered sequences of biochemical reactions inside the cell, which are carried out by enzymes, activated by second messengers, resulting in a signal transduction pathway. Intracellular signal transduction is the process by which chemical signals from outside the cell are passed through the cytoplasm to cellular systems, such as the nucleus or cytoskeleton, where appropriate responses to those signals are generated [7]. Such processes are usually rapid, lasting on the order of milliseconds in the case of ion flux, minutes for the activation of protein- and lipid-mediated kinase cascades, or hours and even days for gene expression. The number of proteins and other molecules participating in the events involving signal transduction increases as the process emanates from the initial stimulus, resulting in a "signal cascade," beginning with a relatively small stimulus that elicits a large response.

As shown in **Figure 1.5**, most signal transduction involves the binding of extracellular signaling molecules (or ligands) to cell-surface receptors that face outward from the plasma membrane and trigger events inside the cell. Also, intracellular signaling cascades can be triggered through cell-substratum interactions, as in the case of integrins, which bind ligands found within the extracellular matrix. The signaling molecules have been functionally classified as: hormones (e.g., melatonin), growth factors (e.g. epidermal growth factor), extra-cellular matrix components (e.g., fibronectin), cytokines (e.g., interferon-gamma), chemokines (e.g., RANTES), neurotransmitters (e.g., acetylcholine), and neurotrophins (e.g., nerve growth factor). A fundamentally different mechanism of signal transduction is carried out by the receptor enzymes. These proteins have a ligand-binding domain on the extracellular surface of the plasma membrane and an enzyme active site on the cytosolic side, with the two domains connected by a single transmembrane segment. Commonly, the receptor enzyme is a protein kinase that phosphorylates Tyr residues in specific target proteins; the insulin receptor is the prototype for this group. In plants, the protein kinase of receptors is specific for Ser or Thr residues.

phosphorylation plays crucial regulatory role in signal transduction pathway [8].

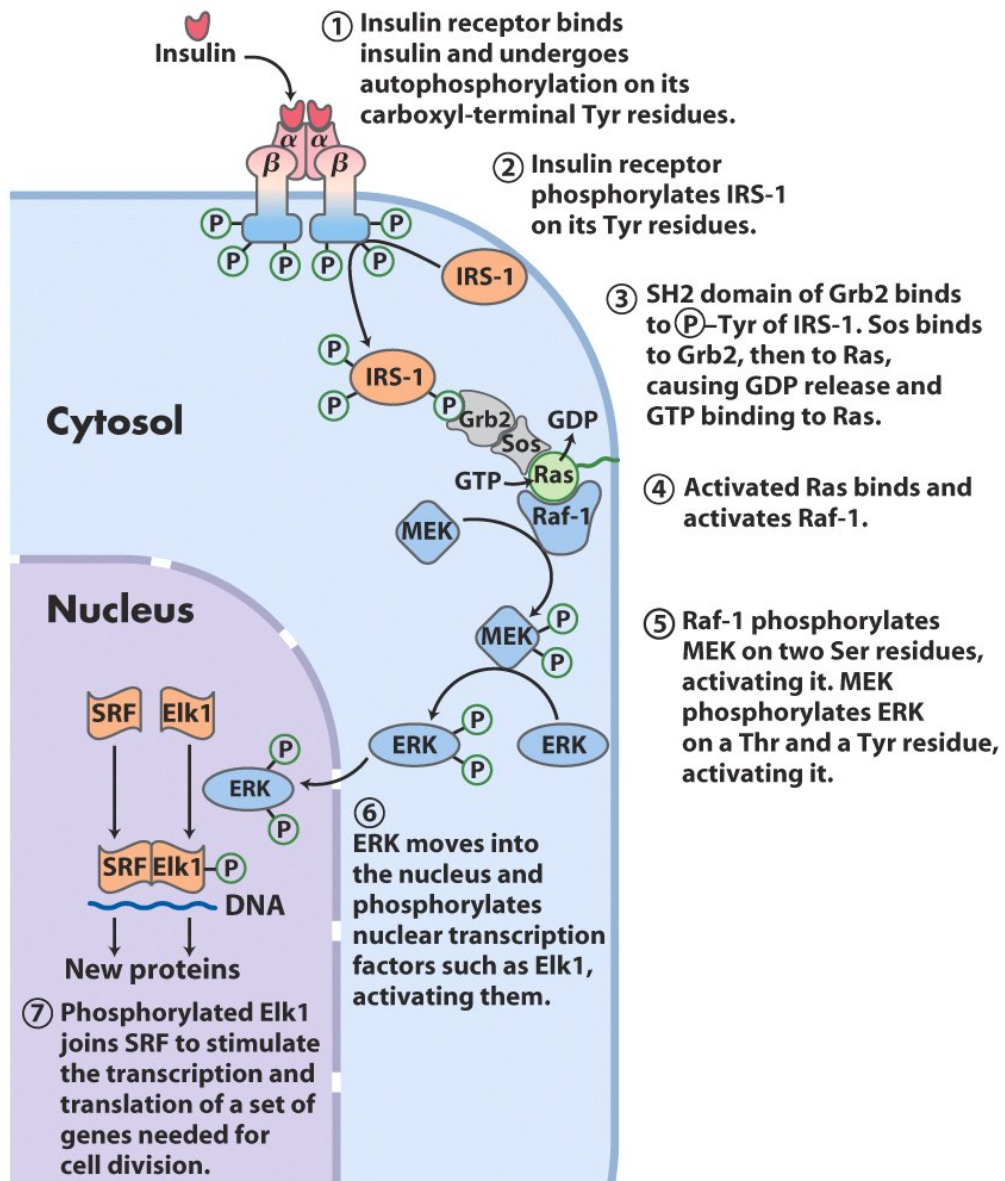


Figure 1.6 Insulin-induced signal transduction (Lehninger *et al.*, 2005).

1.1.4 Mass Spectrometry-based Identification of Protein Phosphorylation

Recent successes illustrate the role of mass spectrometry-based proteomics as an indispensable tool for molecular and cellular biology and for the emerging field of systems biology. So far, protein analysis (primary sequence, post-translational modifications (PTMs) or protein–protein interactions) by MS has been most successful when applied to small sets of proteins isolated in specific functional contexts [9]. The systematic analysis of the much

larger number of proteins expressed in a cell, an explicit goal of proteomics, is now also rapidly advancing, due mainly to the development of new experimental approaches. By definition, a mass spectrometer consists of an ion source, a mass analyzer that measures the mass-to-charge ratio (m/z) of the ionized analytes, and a detector that registers the number of ions at each m/z value. Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are the two techniques most commonly used to volatilize and ionize the proteins or peptides for mass spectrometric analysis [10]. ESI ionizes the analytes out of a solution and is therefore readily coupled to liquid-based (for example, chromatographic and electrophoretic) separation tools.

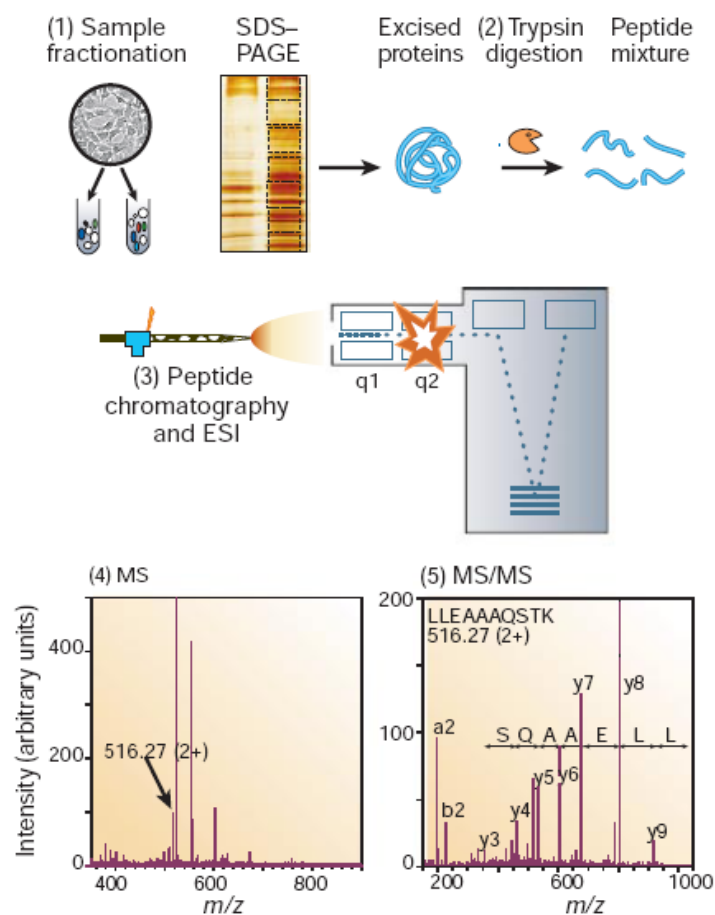


Figure 1.7 Generic mass spectrometry (MS)-based proteomics experiment (Aebersold *et al.*, 2003).

As shown in **Figure 1.7**, the typical proteomics experiment consists of five stages [9]. In stage 1, the proteins to be analyzed are isolated from cell lysate or tissues by biochemical fractionation or affinity selection. This often includes a final step of one-dimensional gel electrophoresis, and defines the ‘sub-proteome’ to be analyzed. MS of whole proteins is less

sensitive than peptide MS and the mass of the intact protein by itself is insufficient for identification. Therefore, proteins are degraded enzymatically to peptides in stage 2, usually by trypsin, leading to peptides with C-terminally protonated amino acids, providing an advantage in subsequent peptide sequencing. In stage 3, the peptides are separated by one or more steps of high-pressure liquid chromatography in very fine capillaries and eluted into an electrospray ion source where they are nebulized in small, highly charged droplets. After evaporation, multiply protonated peptides enter the mass spectrometer and, in stage 4, a mass spectrum of the peptides eluting at this time point is taken (MS1 spectrum, or ‘normal mass spectrum’). The computer generates a prioritized list of these peptides for fragmentation and a series of tandem mass spectrometric or ‘MS/MS’ experiments ensues (stage 5). These consist of isolation of a given peptide ion, fragmentation by energetic collision with gas, and recording of the tandem or MS/MS spectrum. The MS and MS/MS spectra are typically acquired for about one second each and stored for matching against protein sequence databases. **Figure 1.8** shows an example of MS/MS spectra which contains a phosphorylated serine.

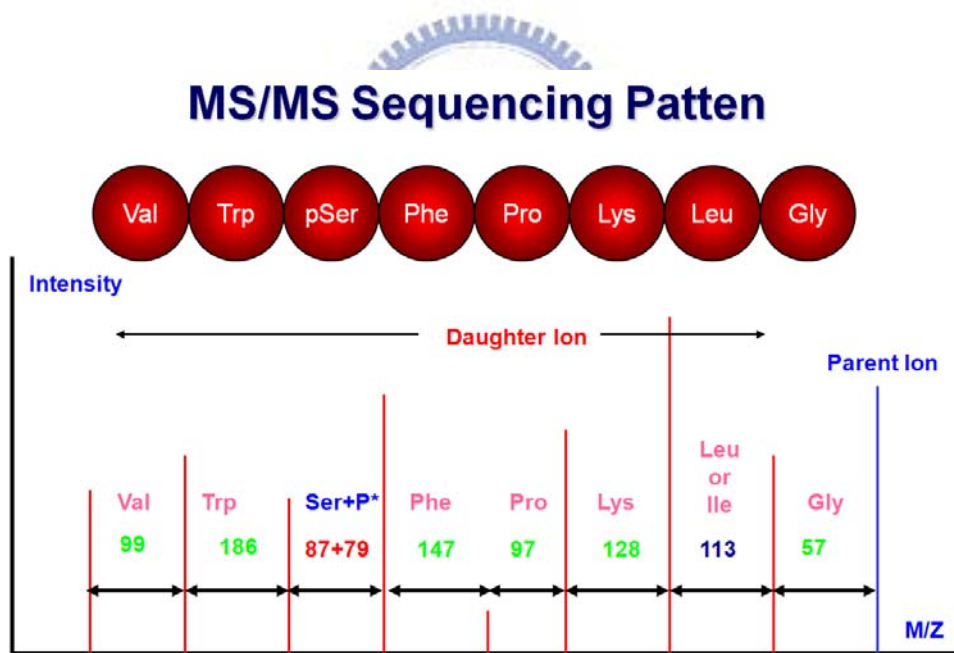


Figure 1.8 Example of phosphopeptide MS/MS spectra.

1.1.5 Phosphoproteomics

Phosphorylation is a key reversible modification occurring mainly on serine, threonine and tyrosine residues that can regulate enzymatic activity, subcellular localization, complex formation and degradation of proteins. Analysis of the entire cellular phosphoproteins panel,

the so-called phosphoproteome, has been an attractive study subject since the discovery of phosphorylation as a key regulatory mechanism of cell life [11]. The understanding of the regulatory role played by phosphorylation begins with the discovery and identification of phosphoproteins and then by determining how, where and when these phosphorylation events take place. Because phosphorylation is a dynamic process difficult to quantify, we must at first acquire an inventory of phosphoproteins and characterize their phosphorylation sites. Several experimental strategies can be used to explore the phosphorylation status of proteins from individual moieties to phosphoproteomes.

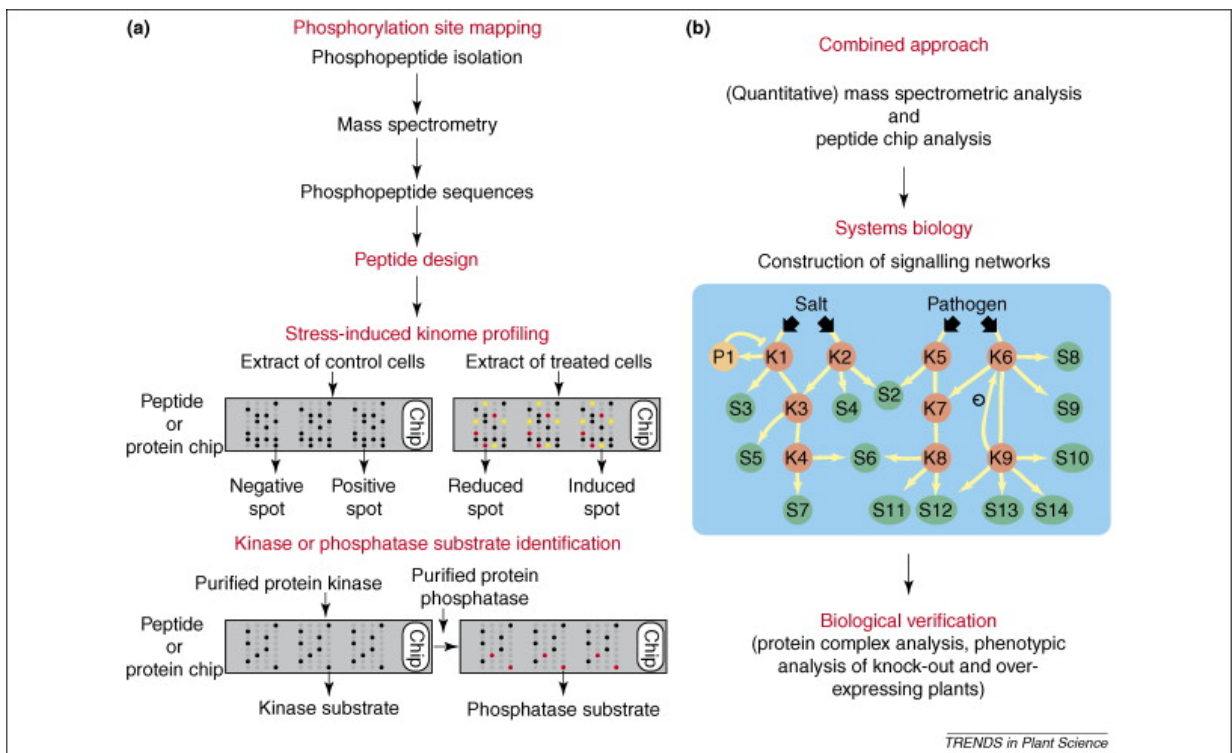
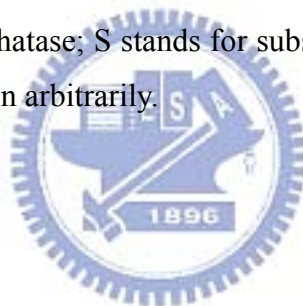


Figure 1.9 Combined large-scale approaches to unravel phosphorylation driven signaling networks (Bentem *et al.*, 2007).

As reviewed previously, mass spectrometry-based techniques have enabled the large-scale mapping of *in vivo* phosphorylation sites. Alternatively, methods based on peptide and protein microarrays have revealed protein kinase activities in cell extracts, in addition to kinase substrates (**Figure 1.9a**) [12]. On chips, protein kinase activities are measured by the incorporation of radioactive phosphate into the substrate peptide or protein that are spotted in small amounts, in duplicate or triplicate (as shown here). Yellow and red spots indicate peptides or proteins that are more intensely and less intensely phosphorylated, respectively. Using cell extracts, a more intensely phosphorylated (‘induced’) spot means that a kinase activity in the treated extract towards the peptide or protein is activated. On the contrary, a

less phosphorylated ('reduced') spot means that the responsible kinase is inactivated. In case of protein phosphatases, a chip pre-phosphorylated by a purified kinase (or, alternatively, a cell extract) could be used for target discovery. This is possible by analyzing which phosphorylated peptides or proteins are dephosphorylated on the chip, as indicated here by red spots.

A combined phosphoproteomic approach of mass spectrometry and microarray technology could enhance the construction of dynamic signaling networks (**Figure 1.9b**) [12]. The experimental data, ultimately, need to be combined by systems biology analysis, which translates the separate, large-scale datasets into signaling networks [13]. The predicted connections within and between signaling cascades need to be experimentally verified by, for instance, analysis of protein complexes and analysis of kinase or substrate knockout and over-expression. In the phosphorylation cascade, arrows indicate phosphorylation reactions and the circled minus sign indicates negative feedback phosphorylation. Only phosphoproteins in the signaling network are indicated. Abbreviations: K stands for protein kinase; P stands for protein phosphatase; S stands for substrate. Numbers behind each kinase, phosphatase and substrate are given arbitrarily.



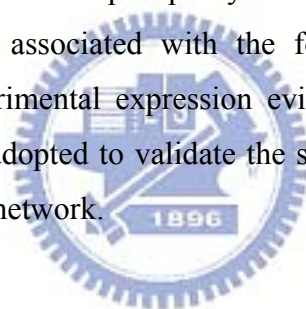
1.2 Motivation

Protein phosphorylation, which is catalyzed by kinase, plays a crucial role in intracellular signal transduction that is achieved by networks of proteins and small molecules that transmit information from the cell surface to the nucleus, where they ultimately effect transcriptional changes. Manning *et al.* have identified 518 human kinase genes, the so-called "kinome", that provides a starting point for comprehensive analysis of protein phosphorylation networks. How differential responses are generated by these networks is not obvious nor is the reason cells evolved a complicated mechanism for transducing signals. Thus, a full understanding of the mechanism of intracellular signal transduction remains a major challenge in cellular biology.

Mass spectrometry-based proteomics have enabled the large-scale mapping of *in vivo* phosphorylation sites. There are several phosphorylation site databases have been constructed previously. However, only 20% of the experimentally verified phosphorylation sites have the annotation of catalytic kinases. Experimental identification of kinase-specific

phosphorylation sites is an inconvenient work and usually limited by the availability of detailed data on the kinase-specific substrates. To fully investigate how protein kinases regulate the intracellular processes, it is necessary to comprehensively and accurately identify the kinase-specific substrates. *In silico* prediction could be a promising strategy to conduct preliminary analyses and could greatly reduce the number of potential targets that need further *in vivo* or *in vitro* confirmation.

With the increasing number of *in vivo* phosphorylation sites have been identified, the desire of map the network of protein kinase and substrate has motivated. The experimental kinase-specific substrates, ultimately, need to be combined by systems biology analysis, which translates the separate, large-scale datasets into signaling networks. Several works have been proposed to incorporate protein-protein interaction data with microarray data for constructing signaling pathway. However, no researchers incorporated the experimentally verified kinase-substrate interactions and the computationally identified kinase-substrate interactions to construct the intracellular phosphorylation network starting from receptor kinases to transcription factors, associated with the formation of protein subcellular localization. Moreover, the experimental expression evidence, such as gene microarray data and mass spectra, could be adopted to validate the syn-expression of the constructed kinase-substrate phosphorylation network.



1.3 Research Goals

In this dissertation, we focus on the integration of heterogeneous phosphorylation site databases, identification of kinase-specific phosphorylation sites, and systematic discovery of kinase-substrate network in human protein phosphorylation.

1.3.1 Management of Heterogeneous Phosphorylation

Databases and Related Information

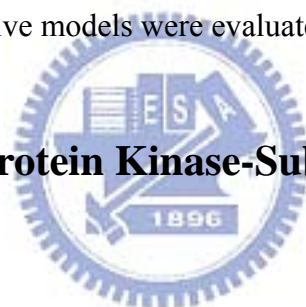
In this study, a variety of biological databases with heterogeneous data format need to be integrated, including phosphorylation site databases, protein sequence and knowledge databases, gene annotation databases, protein structure databases, protein domain databases, protein-protein interaction databases, biochemical pathway databases, and so on. The

inconsistent data format of these integrated biological databases increases the difficulty in the development of proposed system. Therefore, a data warehousing system should be incorporated to efficiently manage, maintain and update all the collected external databases.

1.3.2 Identification of Kinase-Specific Phosphorylation Sites

Experimental identification of phosphorylation sites is an inconvenient work and usually limited by the availability of detailed data on the kinase-specific substrates. *In silico* prediction could be a promising strategy to conduct preliminary analyses and could greatly reduce the number of potential targets that need further *in vivo* or *in vitro* confirmation. Therefore, we propose a method, namely KinasePhos, which incorporates machine learning methods to identify the phosphorylation sites with their catalytic kinase. Not only protein amino acids, but also the structural information such as secondary structure, solvent accessibility and protein disorder region were used to investigate the substrate specificity. Moreover, the constructed predictive models were evaluated by the independent test sets.

1.3.3 Discovery of Protein Kinase-Substrate Phosphorylation Networks



To fully investigate how protein kinases regulate the intracellular processes, it is necessary to comprehensively and accurately identify the kinase-specific substrates. Therefore, we propose a method, named RegPhos, incorporates computational model with protein associations (protein-protein interactions, functional associations, and subcellular localization) for identifying the catalytic kinase for each phosphoprotein with experimental phosphorylated sites. With the highly predictive performance of phosphorylation sites, a better understanding of relationships between protein kinases and substrates will be facilitated and engineered to analyze the therapeutic usefulness. The identified kinase-substrate interactions are used to comprehensively construct the intracellular phosphorylation network starting from receptor kinases to transcription factors, with the information of protein-protein interactions and subcellular localization. Moreover, the experimental expression evidence such as time-coursed gene microarray data is adopted to validate the syn-expression of kinase and substrate with statistical significance.

1.4 Organization of This Dissertation

There are three major parts in this dissertation, including the integration of heterogeneous phosphorylation site databases (**Chapter 2**), identification of kinase-specific phosphorylation sites (**Chapter 3**), and systematic discovery of protein kinase-substrate phosphorylation networks (**Chapter 4**). A variety of biological databases with heterogeneous data format need to be integrated, including phosphorylation site databases, protein sequence and knowledge databases, gene annotation databases, protein structure databases, protein domain databases, protein-protein interaction databases, biochemical pathway databases, and so on. We propose a method, named RegPhos, incorporates computational model with protein associations (protein-protein interactions, functional associations, and subcellular localization) for identifying the catalytic kinase for each phosphoprotein with experimental phosphorylated sites. The protein phosphorylation network of kinase and substrate in human was constructed using the experimentally verified and computationally identified kinase-substrate interactions. The gene microarray expression data is adopted to analyze the syn-expression of kinase and substrate genes in specific conditions. Moreover, the microarray data with time series can be used to recognize the dynamic behavior of kinase and their substrate.

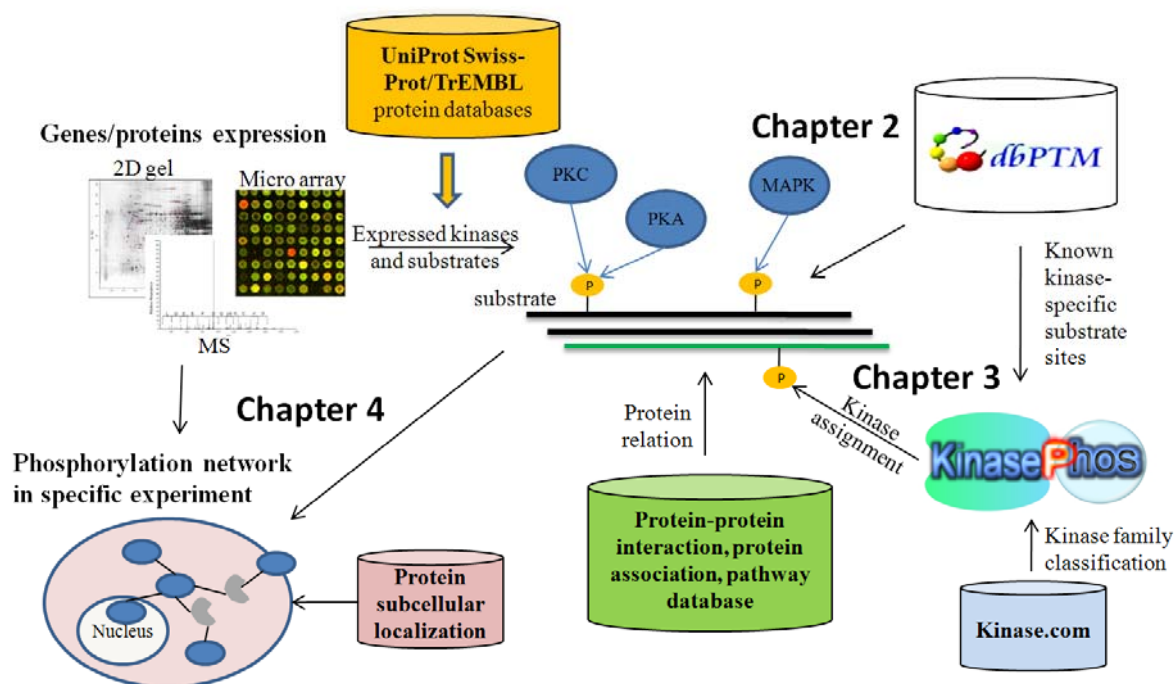


Figure 1.10 Schematic representation of dissertation organization

Chapter 2 Information Repository of Protein Post-Translational Modifications

2.1 Introduction

Protein Post-Translational Modification (PTM) is an extremely important cellular control mechanism because it may alter proteins' physical and chemical properties, folding, conformation distribution, stability, activity, and consequently, their functions [14]. Examples of the biological effects of protein modifications include phosphorylation for signal transduction, attachment of fatty acids for membrane anchoring and association, and glycosylation for changing protein half-life, targeting substrates, and promoting cell-cell and cell-matrix interactions. High-throughput proteomic studies produce a wealth of new information regarding post-translational modifications. With the accelerating progress in proteomics, biological knowledge bases containing a wealth of information, in particular protein modifications, are playing crucial roles in cell regulation research [3]. In this work, we not only provide the sequence-based information such as PTM site, functional domain and protein variant site, but also annotate the structure-based information including protein tertiary structure, protein secondary structure, surface accessibility and protein intrinsic disorder region.

A side chain of amino acid that undergoes enzymatic modification needs to be accessible on the surface of protein [15]. Several works have been proposed the links between the post-translational modifications and their solvent accessible surface area. Pang *et al.* investigated the structural environment of 8378 incidences in 44 types of post-translational modifications [15]. The information of surface accessibility, disorder region, and linker/domain are computationally annotated by several published programs, including ASA [16], GOR [17] and RVP-net [18] for surface accessibility, RONN[19] and DISEMBL [20] for disorder, PSIPRED [21] for secondary structure, and George et al. [22] for linker/domain. The introduction of structural information of protein is described as following.

2.1.1 Protein Solvent Accessibility

Residue solvent accessibility is usually measured by rolling a spherical water molecule over a protein surface and summing the area that can be accessed by this molecule on each residue (typical values range from 0-300 Å²). To allow comparisons between the accessibility of long extended and spherical amino acids, typically relative values are compiled (actual area as percentage of maximally accessible area). A simplified description distinguishes two states: exposed (here residues numbered 1-3 and 10-12) and buried (here residues 4-9) residues. Since the packing density of native proteins resembles that of crystals, values for solvent accessibility provide upper and lower limits to the number of possible inter-residue contacts.

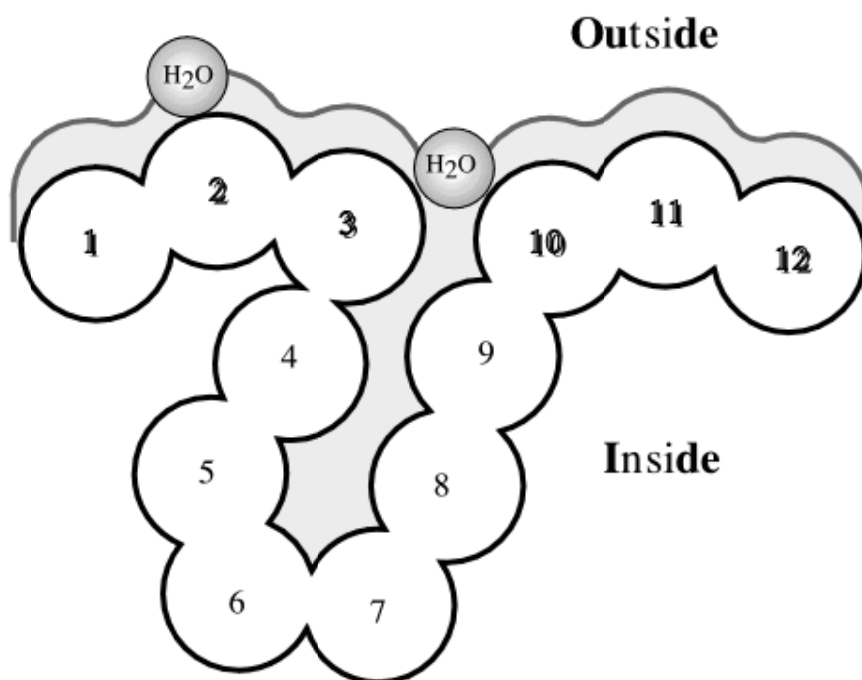


Figure 2.1 Measurement of protein solvent accessibility.²

2.1.2 Protein Intrinsic Disorder

Until the early 1990s, a widely, almost exclusively accepted concept of protein function was the well-known protein sequence → structure → function paradigm. According to this concept, a protein can achieve its biological function only upon folding into a unique,

² The figure was obtained from http://www.rostlab.org/papers/2003_rev_1d/paper.html

structured state, which represents a kinetically accessible and an energetically favorable conformation (usually the global energy minimum for the whole protein) determined by its amino acid sequence. This specific conformation has been referred to as the native state of the protein. However, recent discoveries of intrinsically disordered proteins (IDPs) [23] (known also as natively disordered, natively unfolded, and intrinsically unstructured proteins) have significantly broadened the view of the scientific community and increased the number of groups systematically studying these intriguing members of the protein world.

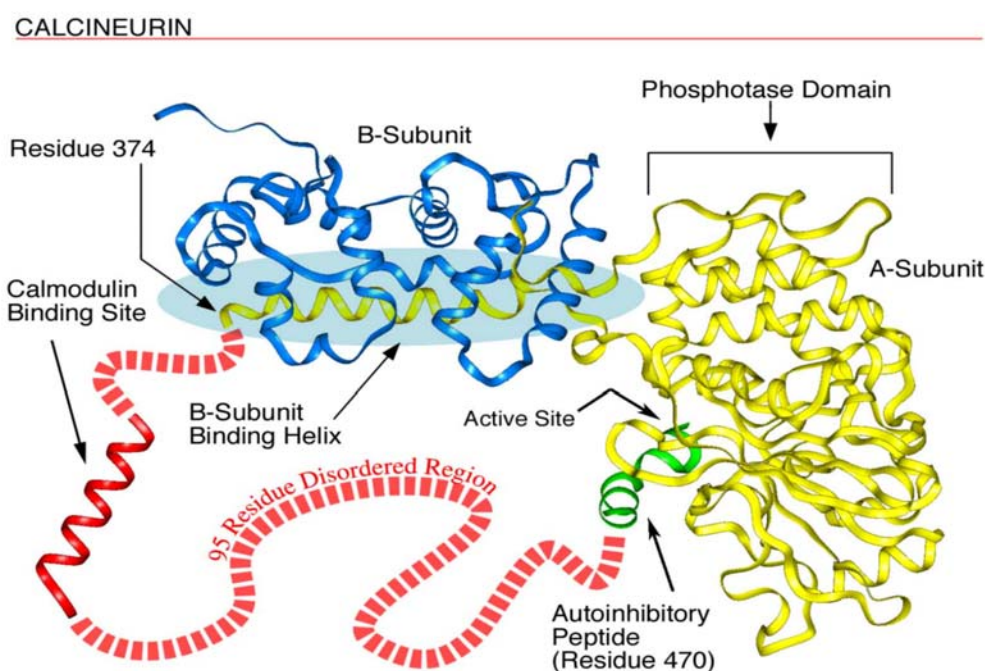


Figure 2.2 Disorder in Calcineurin.³

Intrinsically unstructured proteins are frequently involved in key biological processes such as cell cycle control, transcriptional and translational regulation, membrane fusion and transport, and signal transduction [24]. A high percentage of cell-signaling and cancer-associated proteins are predicted to have long disordered regions [25]. An investigation of the functions performed by intrinsically disordered regions reveals that they are often involved in molecular recognition and protein modifications including phosphorylation [26]. To provide a concrete example, the calmodulin binding site in

³ The figure was obtained from <http://genome.gsc.riken.go.jp/hgmis/publicat/hgn/v12n1/13trinity.html>

calcineurin (**Figure 2.2**)⁴ was shown to be extremely sensitive to protease digestion and thus to be a disordered ensemble; this disorderliness was confirmed in Kissinger's X-ray diffraction structure as indicated by missing coordinates in the same region. As shown in **Figure 2.2**, Calcineurin's α -subunit contains a globular phosphatase domain, a helical extension that binds the β -subunit, a disordered region not observed in the crystal structure, and an autoinhibitory peptide that binds in the phosphatase domain's active site. The α -subunit's intrinsically disordered region, containing 95 amino acids, connects the ends of the helical extension (residue 374) and the autoinhibitory peptide (residue 470) and includes a calmodulin binding site. This region probably is disordered at least in part to allow calmodulin to bind.

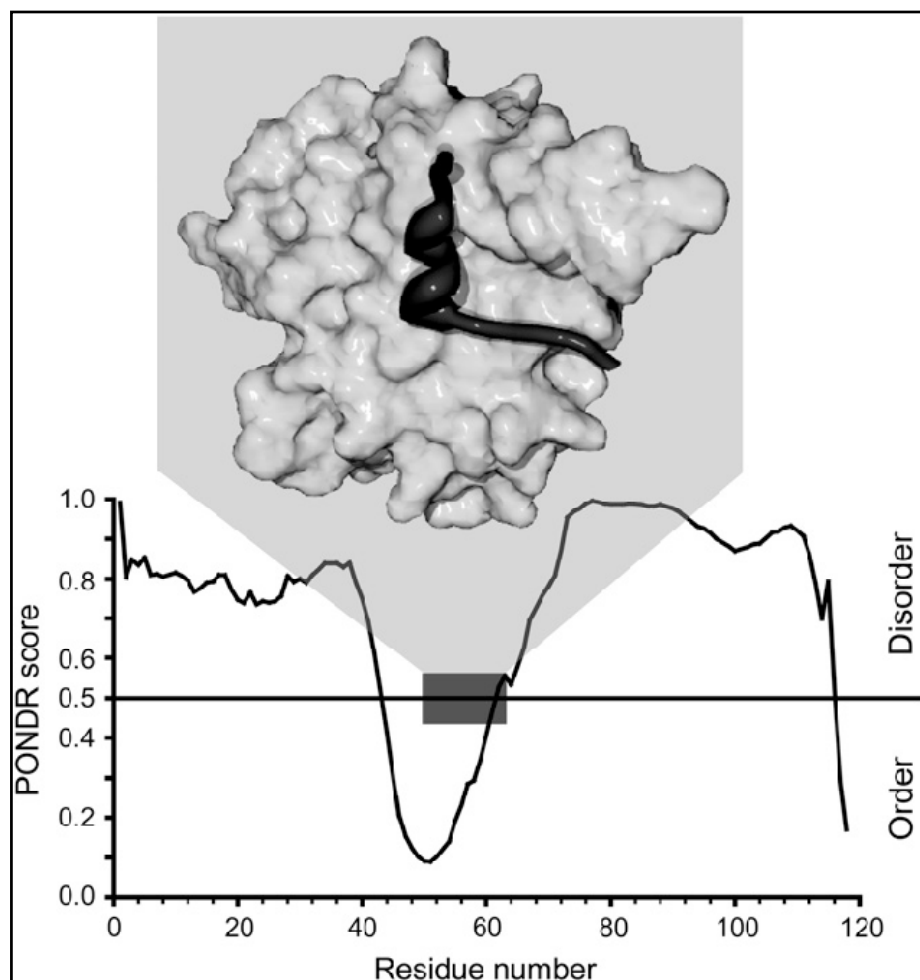


Figure 2.3 Example of a binding region and its positions relative to the regions of PONDR predicted disorder score (Garner, *et al.*, 1999).

Computational methods exploit the sequence signatures of disorder to predict whether a

⁴ The figure was obtained from <http://genome.gsc.riken.go.jp/hgmis/publicat/hgn/v12n1/13trinity.html>

protein is disordered given its amino acid sequence. The table below (**Table 2.1**), adapted from Ferron *et. al.* [27], shows the main features of tools for disorder prediction. Note that different tools use different definitions of disorder. Various predictors of intrinsic disorder have been used to facilitate prediction of functional properties of proteins. The first use of a disorder predictor to find protein-binding sites was performed by Garner *et al.* [28] who noticed that sharp dips in disorder prediction could indicate short loosely structured binding regions that undergo disorder-to-order transitions upon binding to a partner. Interestingly, these dips in disorder prediction were originally noticed for the 4E binding protein (4EBP1, see **Figure 2.3**) [28], which had been shown to be completely disordered by NMR [29]. However, a short stretch of 4EBP1 undergoes a disorder-to-order transition upon binding to eukaryotic translation initiation factor 4E [30].



Table 2.1 Summary of the web servers offering prediction of intrinsically disordered proteins.

Tool name	What is predicted	Method	URL
PONDR [31]	All regions that are not rigid including random coils, partially unstructured regions, and molten globules	Feed-forward neural network with separate N-/C-terminus predictor. Based on amino-acid compositions and physicochemical properties.	http://www.pondr.com
FoldIndex [32]	Regions that have a low hydrophobicity and high net charge (either loops or unstructured regions)	Charge/hydrophobicity score based on a sliding window.	http://bip.weizmann.ac.il/fldbin/findex
NORSp [33, 34]	Regions with No Ordered Regular Secondary Structure (NORS). Most, but not all, are highly flexible.	Rule-based using a set of several neural-networks. Amino acid compositions and sequence profiles used as features.	http://roslab.org/services/NORSp/
DISOPRED [35]	Regions devoid of ordered regular secondary structure	Feed-forward neural network (DISOPRED) and linear support vector machine (DISOPRED2) based on sequence profiles.	http://bioinf.cs.ucl.ac.uk/disopred/
Globplot [36]	Regions with high propensity for globularity on the Russell/Linding scale (propensities for secondary structures and random coils)	Autoregressive model based on amino-acid propensities for disorder/globularity.	http://globplot.embl.de/
DisEMBL [20]	LOOPS (regions devoid of regular secondary structure); HOT LOOPS (highly mobile loops); REMARK465 (regions lacking electron density in crystal structure)	Ensemble of feed-forward neural networks.	http://dis.embl.de/
IUPred [37]	Regions that lack a well-defined 3D-structure under native conditions	Linear model based on the estimated energy of pairwise interactions in a window around a residue.	http://iupred.enzim.hu/index.html
PreLink [38]	Regions that are expected to be unstructured in all conditions, regardless of the presence of a binding partner	Rule-based. Ratio of multinomial probabilities (for linker and structured regions) combined with the distance to the nearest hydrophobic cluster.	http://genomics.eu.org/spip/PreLink
RONN [39]	Regions that lack a well-defined 3D structure under native conditions	Feed-forward neural network in the space of distances to a set of prototype sequences of known fold state.	http://www.strubi.ox.ac.uk/RONN
DISpro	Protein intrinsically disordered regions	Recursive neural network based on sequence profiles, predicted secondary structure and relative solvent accessibility.	http://www.igb.uci.edu/servers/psss.html
SPRITZ [40]	Intrinsically disordered regions in proteins from sequence	Nonlinear support vector machine based on multiply aligned sequences. Separate predictors for short and long disorder regions.	http://protein.cribi.unipd.it/spritz/

2.1.3 Subcellular Localization

The eukaryotic cell is a composite system internally subdivided into membrane-enveloped compartments that perform particular functions [41]. Every subcellular compartment contains specific proteins, including enzymes, synthesized in the cytoplasm and translocated into the locations, where they carry out functional patterns. As shown in **Figure 2.4**, some major constituents of eukaryotic cells are: extracellular space, cytoplasm, nucleus, mitochondria, Golgi apparatus, endoplasmic reticulum (ER), peroxisome, vacuoles, cytoskeleton, nucleoplasm, nucleolus, nuclear matrix and ribosomes. The proteins which are involved in similar biological functions are closely located in the same subcellular localization. Therefore, knowing the localization of every protein is important for elucidating its interactions with other molecules and for understanding its biological function.

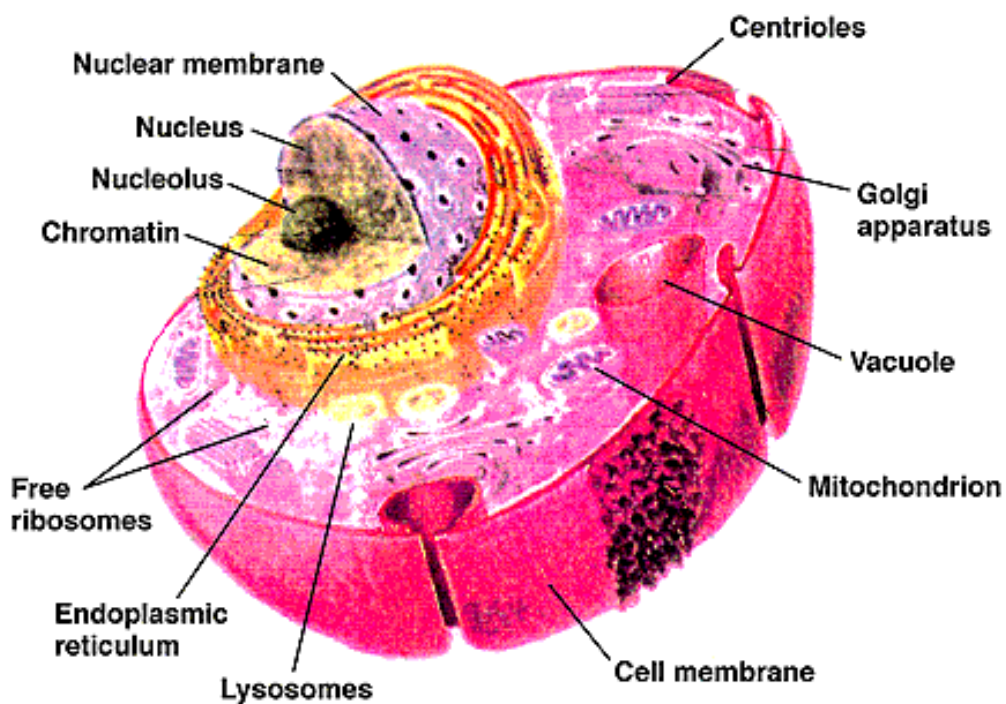


Figure 2.4 Eukaryotic cellular compartments.⁵

⁵ The figure was obtained from http://mendel.imp.ac.at/CELL_LOC/

2.2 Related Works

Taking the advantage of the high-throughput mass spectrometry in proteomics, several databases involved in protein modifications were established. UniProtKB/Swiss-Prot [42] includes as much modification information as available with consistency and structure, allowing easy retrieval by biologists. Phospho.ELM [2], PhosphoSite [43] and Phosphorylation Site Database [44] were developed for collecting experimentally verified phosphorylation sites. PHOSIDA [45] integrates thousands of high-confidence *in vivo* phosphorylation sites identified by mass spectrometry-based proteomics in various species. O-GLYCBASE [46] is a database of glycoproteins, most of which include experimentally verified O-linked glycosylation sites. Moreover, UbiProt stores experimental ubiquitylated proteins and ubiquitylation sites, which are implicated in protein degradation via an intracellular ATP-dependent proteolytic system [47]. The RESID protein modification database is a comprehensive collection of annotations and structures for protein modifications and cross-links including pre-, co-, and post-translational modifications [4]. Each RESID entry presents a protein with a chemically unique modification and indicates how the modification is currently annotated in the Swiss-Prot. The summary of published PTM databases is presented in **Table 2.2**. The detailed introduction about these PTM resources is illustrated as following.

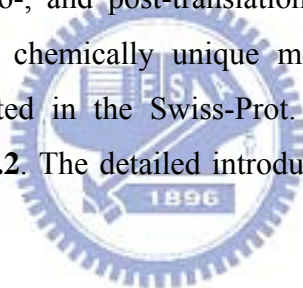


Table 2.2 Summary of PTM resource.

Resource	Reference	Description	URL
UniProt KB / Swiss-Prot	Farriol-Mathis, Garavelli <i>et al.</i> 2004	Experimental PTMs and putative PTMs (annotated as “by similarity”, “potential” or “probable” in the ‘MOD_RES’, “CARBOHYD”, “LIPID” and “CROSSLNK” fields)	www.expasy.org/sprot/
PhosphoELM	Diella, Cameron <i>et al.</i> 2004	Experimental phosphorylation sites	phospho.elm.eu.org
PhosphoSite	Hornbeck, Chabra <i>et al.</i> 2004	Experimental phosphorylation sites	www.phosphosite.org
Phosphorylation site database	Wurgler-Murphy, King <i>et al.</i> 2004	Experimental phosphorylation sites in prokaryotic organisms	vigen.biochem.vt.edu/xpd/xpd.htm
PHOSIDA	Gnad, Ren <i>et al.</i> 2007	In vivo phosphorylation sites which was identified by mass spectrometry-based Proteomics	www.phosida.com
HPRD	Peri, S. <i>et al.</i> 2003	Human PTMs with curated literatures	www.hprd.org
PhosPhAt	Heazlewood, Durek <i>et al.</i> 2008	mass spectrometry-based identified phosphorylation sites in Arabidopsis	phosphat.mpimp-golm.mpg.de
O-GLYCBASE	Gupta, Birch <i>et al.</i> 1999	Experimental glycosylation sites	www.cbs.dtu.dk/databases/OGLYCBASE/
UbiProt	Chernorudskiy, Garcia <i>et al.</i> 2007	Ubiquitylated protein and ubiquitylation sites	ubiprot.org.ru
RESID	Garavelli 2004	Protein modification annotations	www.ebi.ac.uk/RESID
Phospho3D	Zanzoni <i>et al.</i> 2007	3D structures of protein phosphorylation sites	cbm.bio.uniroma2.it/phospho3d/index.py

UniProtKB/Swiss-Prot Modifications

With the accelerating progress in proteomics, UniProt KB/Swiss-Prot knowledge base [48] is faced with the challenge of including this information in a consistent and structured way, in order to facilitate easy retrieval and promote understanding by biologist expert users as well as computer programs. The authors are therefore standardizing the annotation of PTM features represented in UniProt KB/Swiss-Prot [3]. Indeed, a controlled vocabulary has been associated with every described PTM. There are two types of PTM annotation, the experimentally validated PTM sites and the putative PTM sites. The putative PTMs are annotated as “by similarity”, “potential” or “probable” in the ‘MOD_RES’, “CARBOHYD”, “LIPID” and “CROSSLNK” fields.

Phospho.ELM

The fast growing number of research reports on protein phosphorylation points to a general need for an accurate database dedicated to phosphorylation to provide easily retrievable information on phosphoproteins. Phospho.ELM (<http://phospho.elm.eu.org>) [2], which was developed as part of the ELM (Eukaryotic Linear Motif) resource, is a resource containing experimentally verified phosphorylation sites that were manually curated from the literature. Phospho.ELM constitutes the largest searchable collection of phosphorylation sites available to the research community. The Phospho.ELM entries store information about substrate proteins with the exact positions of residues known to be phosphorylated by cellular kinases. Additional annotation includes literature references, subcellular compartment, tissue distribution, and information about the signaling pathways involved as well as links to the molecular interaction database MINT.

The current release of Phospho.ELM (version 7.0, July 2007) contains 4078 phospho-protein sequences covering 12 025 phospho-serine, 2362 phospho-threonine and 2083 phospho-tyrosine sites [49]. The entries provide information about the phosphorylated proteins and the exact position of known phosphorylated instances, the kinases responsible for the modification (where known) and links to bibliographic references. The database entries have hyperlinks to easily access further information from UniProt [50], PubMed, SMART, ELM, MSD as well as links to the protein interaction databases MINT and STRING. A new BLAST search tool, complementary to retrieval by keyword and UniProt accession number, allows users to submit a protein query (by sequence or UniProt accession) to search against

the curated data set of phosphorylated peptides.

PhosphoSite

PhosphoSite is a curated, web-based bioinformatics resource dedicated to physiologic sites of protein phosphorylation in human and mouse. PhosphoSite is populated with information derived from published literature as well as high-throughput discovery programs. PhosphoSite provides information about the phosphorylated residue and its surrounding sequence, orthologous sites in other species, location of the site within known domains and motifs, and relevant literature references. Links are also provided to a number of external resources for protein sequences, structure, post-translational modifications and signaling pathways, as well as sources of phospho-specific antibodies and probes. As the amount of information in the underlying knowledgebase expands, users will be able to systematically search for the kinases, phosphatases, ligands, treatments, and receptors that have been shown to regulate the phosphorylation status of the sites, and pathways in which the phosphorylation sites function. As it develops into a comprehensive resource of known *in vivo* phosphorylation sites, PhosphoSite will be a valuable tool for researchers seeking to understand the role of intracellular signaling pathways in a wide variety of biological processes.



Phosphorylation Site Database

Phosphorylation Site Database (<http://vigen.biochem.vt.edu/xpd/xpd.htm>) [44] provides ready access to information from the primary scientific literature concerning those proteins from prokaryotic organisms, i.e., the members of the domains Archaea and Bacteria, that have been reported to undergo covalent phosphorylation on the hydroxyl side chains of serine, threonine, and/or tyrosine residues. Where known, the sequence of the site(s) of phosphorylation and the functional consequences of phosphorylation also are included. Active links enable users to quickly access further information concerning the phosphoprotein of interest from PubMed, GenBank, SWISS-PROT, and PIR.

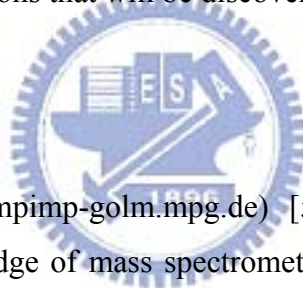
PHOSIDA

PHOSIDA (<http://www.phosida.com>), a phosphorylation site database, integrates thousands of high-confidence *in vivo* phosphorylation sites identified by mass spectrometry-based

proteomics in various species. For each phosphorylation site, PHOSIDA lists matching kinase motifs, predicted secondary structures, conservation patterns, and its dynamic regulation upon stimulus. Using support vector machines, PHOSIDA also predicts phosphorylation sites.

HPRD

Human Protein Reference Database (HPRD, <http://www.hprd.org>) [51] is an object database that integrates a wealth of information relevant to the function of human proteins in health and disease. Data pertaining to thousands of protein-protein interactions, posttranslational modifications, enzyme/substrate relationships, disease associations, tissue expression, and subcellular localization were extracted from the literature for a nonredundant set of 2750 human proteins. Almost all the information was obtained manually by biologists who read and interpreted >300,000 published articles during the annotation process. This unified bioinformatics platform will be useful in cataloging and mining the large number of proteomic interactions and alterations that will be discovered in the postgenomic era.



PhosPhAt

The PhosPhAt (<http://phosphat.mpimp-golm.mpg.de>) [52] database provides a resource consolidating our current knowledge of mass spectrometry-based identified phosphorylation sites in Arabidopsis and combines it with phosphorylation site prediction specifically trained on experimentally identified Arabidopsis phosphorylation motifs. The database currently contains 1187 unique tryptic peptide sequences encompassing 1053 Arabidopsis proteins. Among the characterized phosphorylation sites, there are over 1000 with unambiguous site assignments, and nearly 500 for which the precise phosphorylation site could not be determined. The database is searchable by protein accession number, physical peptide characteristics, as well as by experimental conditions (tissue sampled, phosphopeptide enrichment method). For each protein, a phosphorylation site overview is presented in tabular form with detailed information on each identified phosphopeptide. An analysis of the current annotated Arabidopsis proteome yielded in 27,782 predicted phosphoserine sites distributed across 17,035 proteins. These prediction results are summarized graphically in the database together with the experimental phosphorylation sites in a whole sequence context.

O-GLYCBASE

O-GLYCBASE (<http://www.cbs.dtu.dk/databases/OGLYCBASE/>) is a database of glycoproteins with O-linked glycosylation sites. Entries with at least one experimentally verified O-glycosylation site have been compiled from protein sequence databases and literature. Each entry contains information about the glycan involved, the species, sequence, a literature reference and http-linked cross-references to other databases. Version 4.0 contains 179 protein entries, an approximate 15% increase over the last version. Sequence logos representing the acceptor specificity patterns for GalNAc, GlcNAc, mannosyl and xylosyl transferases are shown.

UbiProt

UbiProt (<http://ubiprot.org.ru>) [47] Database is a public resource offering comprehensive information on ubiquitylated proteins. Post-translational protein modification with ubiquitin, or ubiquitylation, is one of the hottest topics in a modern biology due to a dramatic impact on diverse metabolic pathways and involvement in pathogenesis of severe human diseases. As shown in **Figure 2.5** [47], Ubiquitylation may result in addition of a single ubiquitin moiety or a branched multi-ubiquitin chain to the target protein lysine(s). Note that a ubiquitin molecule possesses 7 inner lysine residues that can serve as attachment sites of the next ubiquitin moiety, resulting in the formation of the chains that have a different structure and topology. Functionally significant amino acids are marked as follows: Kn and Kn' – lysine residue(s) that can serve as attachment sites of the ubiquitin moiety; G76 – ubiquitin C-terminal glycine residue participating in the isopeptide bond formation.

A great number of eukaryotic proteins were found to be ubiquitylated. However, data about particular ubiquitylated proteins are rather disembodied. To fill a general need for collecting and systematizing experimental data concerning ubiquitylation, a knowledge base of ubiquitylated proteins, UbiProt Database, have been developed. The database contains retrievable information about overall characteristics of a particular protein, ubiquitylation features, related ubiquitylation and de-ubiquitylation machinery and literature references reflecting experimental evidence of ubiquitylation. The resource can serve as a general reference source both for researchers in ubiquitin field and those who deal with particular ubiquitylated proteins which are of their interest. Further development of the UbiProt Database is expected to be of common interest for research groups involved in studies of the

ubiquitin system.

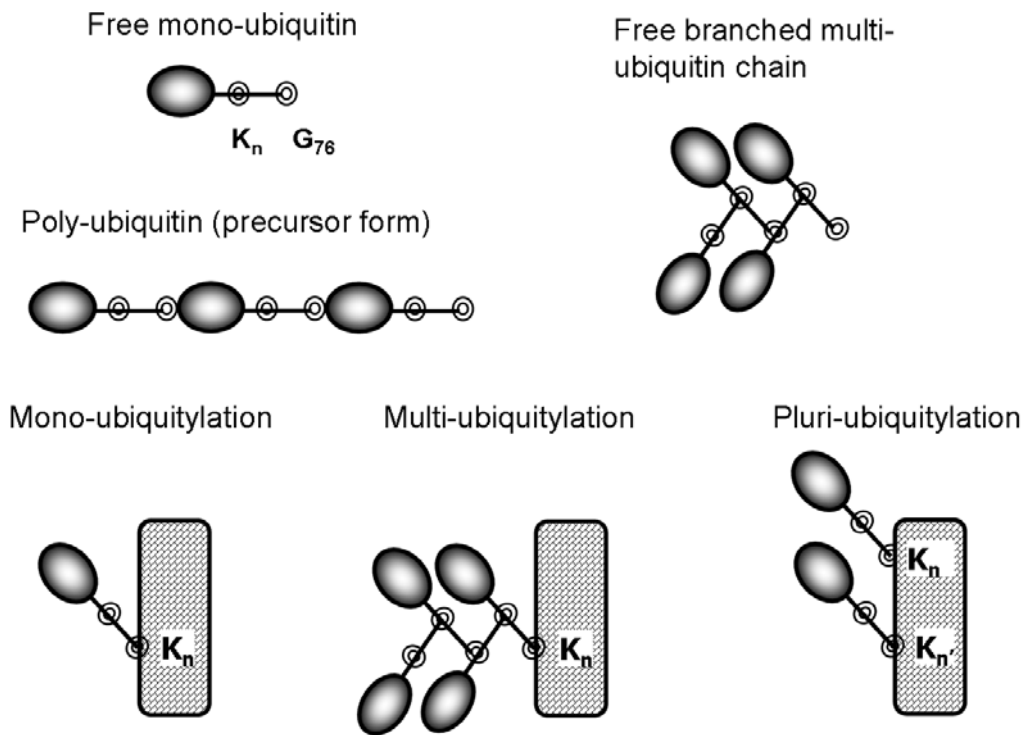


Figure 2.5 Different forms of ubiquitin and ubiquitin-modified proteins (Chernorudskiy, et al., 2007).

RESID

The RESID [4] Database of Protein Modifications is a comprehensive collection of annotations and structures for protein modifications and cross-links including pre-, co-, and post-translational modifications. The database provides: systematic and alternate names, atomic formulas and masses, enzymatic activities that generate the modifications, keywords, literature citations, Gene Ontology (GO) cross-references, protein sequence database feature table annotations, structure diagrams, and molecular models. This database is freely accessible on the Internet through resources provided by the European Bioinformatics Institute (<http://www.ebi.ac.uk/RESID>), and by the National Cancer Institute--Frederick Advanced Biomedical Computing Center (<http://www.ncifcrf.gov/RESID>). Each RESID Database entry presents a chemically unique modification and shows how that modification is currently annotated in the protein sequence databases, Swiss-Prot and the Protein Information Resource (PIR). The RESID Database provides a table of corresponding equivalent feature annotations that is used in the UniProt project, an international effort to combine the resources

of the Swiss-Prot, TrEMBL and PIR. As an annotation tool, the RESID Database is used in standardizing and enhancing modification descriptions in the feature tables of Swiss-Prot entries. As an Internet resource, the RESID Database assists researchers in high-throughput proteomics to search monoisotopic masses and mass differences and identifies known and predicted protein modifications.

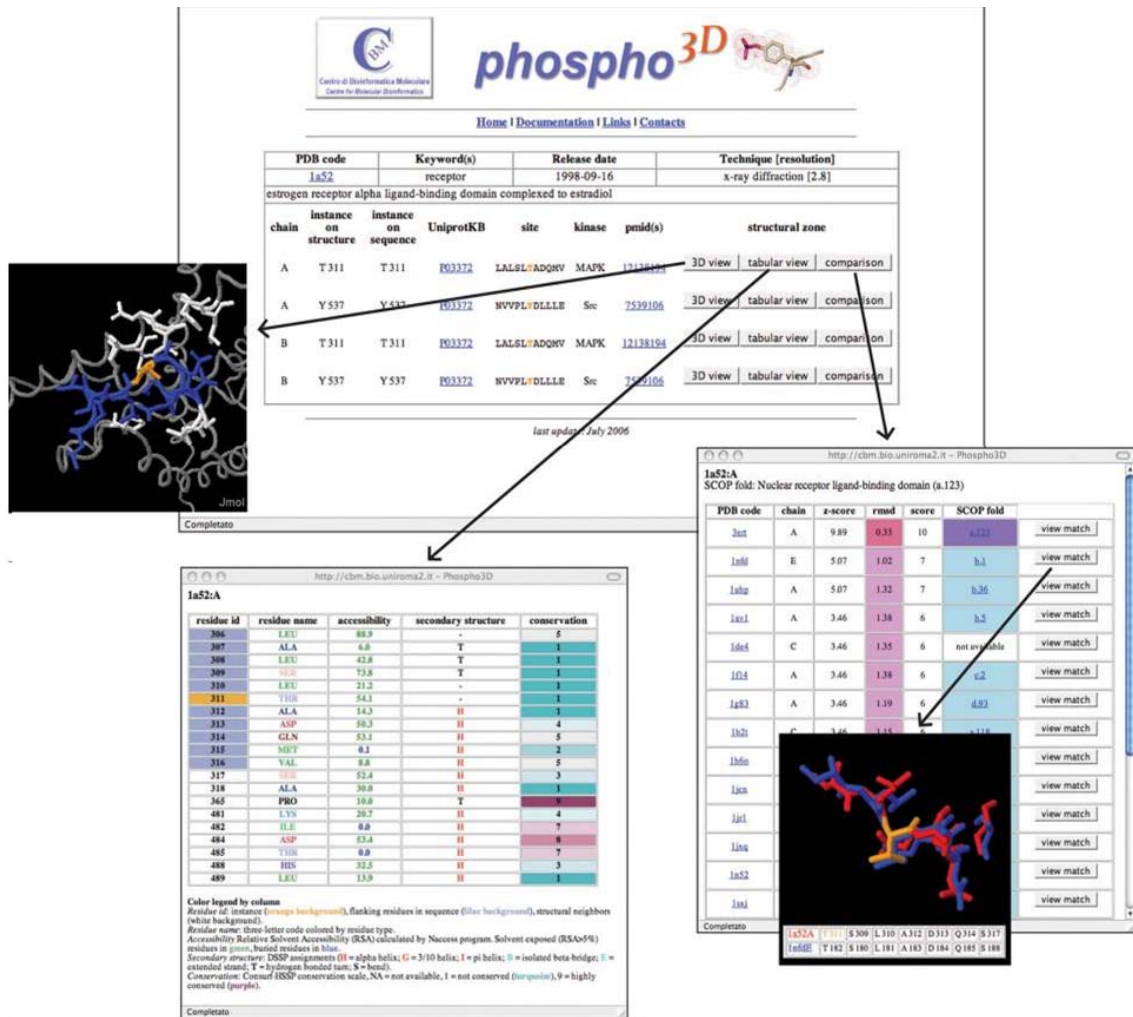


Figure 2.6 A list of instances for the PDB file 1A52 (Zanzoni, A., *et al.*, 2007).

Phospho3D

Since the amount of data produced by screening assays is growing continuously, the development of computational tools for collecting and analyzing experimental data has become a pivotal task for unraveling the complex network of interactions regulating eukaryotic cell life. The authors presented Phospho3D [53] (<http://cbm.bio.uniroma2.it/phospho3d>), a database of 3D structures of phosphorylation sites, which stores information retrieved from the phospho.ELM database and is enriched with

structural information and annotations at the residue level. The database also collects the results of a large-scale structural comparison procedure providing clues for the identification of new putative phosphorylation sites. As shown in **Figure 2.6**, in the central panel a list of instances for the PDB file 1A52 is shown. For each of them, users can visualize the corresponding *zone* via the Jmol viewer, the annotation at the residue level and the results of the large-scale local structural comparison. For each structural match the score, the Z-score, and the root-mean-square deviation (RMSD) are reported along with the SCOP fold [54] of the matching PDB files.



2.3 Motivation and the Specific Aim

With the high-throughput mass spectrometry in proteomics, biological knowledge bases containing a wealth of protein modifications are established. The annotating format of protein modifications from various resources is different. Therefore, we are inspired to integrate all the data of protein modifications and store them in consistent and structured way, in order to facilitate easy retrieval and promote understanding by biologist expert users as well as computer programs.

In this study, we develop a knowledge base, namely dbPTM, which collects the known protein post-translational modification information from external biological data sources. Since only a small fraction of UniProtKB/Swiss-Prot proteins are annotated with experimentally verified post-translational modifications, we also developed computational tools [55, 56] to comprehensively identify phosphorylation sites, glycosylation sites and sulfation sites against the UniProtKB/Swiss-Prot proteins. Protein structural properties and functional information, such as the solvent accessibility of residues, protein disorder regions, protein variations, non-synonymous single nucleotide polymorphism (SNP), protein tertiary structures, and protein functional domains, are provided for researchers who investigating the protein post-translational modification mechanisms. Besides, the PTM related literature, protein conservations and substrate specificity are also provided in the resource. Web query interface and graphical visualization were designed and implemented to facilitate access to the database content.

Currently, computational identification of protein modifications becomes a promising strategy to conduct preliminary analyses for protein functions and its roles in biological systems. A variety of computational tools have been developed for more than ten PTM types including phosphorylation, glycosylation, acetylation, methylation, sulfation, sumoylation and so on. In order to evaluate these computational tools, we compiled a PTM benchmark containing all available sites for each type of PTM. The PTM benchmark can provide a standard for evaluating performance of the computational prediction tools developed for identification of protein post-translation modification sites.

2.4 Materials and Methods

The data generation flow of the dbPTM is briefly depicted in **Figure 2.7**. The data generation flow comprises the three major components: integration of external Post-Translational Modification (PTM) databases, computational annotation of PTM sites, and structural or functional annotations. The experimentally validated PTM data sources were extracted from UniProt KB/Swiss-Prot [3], Phospho.ELM [2], PHOSIDA [45], O-GLYCBASE [46], and UbiProt [47]. The experimentally verified PTM sites were used to generate computer models to further identify putative PTM sites against the Swiss-Prot proteins. Additional structural properties and functional information, such as protein tertiary structures, protein secondary structures, solvent accessibility of residues, protein disorder region, protein functional domains, protein variations and non-synonymous SNP are also annotated to the Swiss-Prot proteins. The detailed data generation flow is described below.

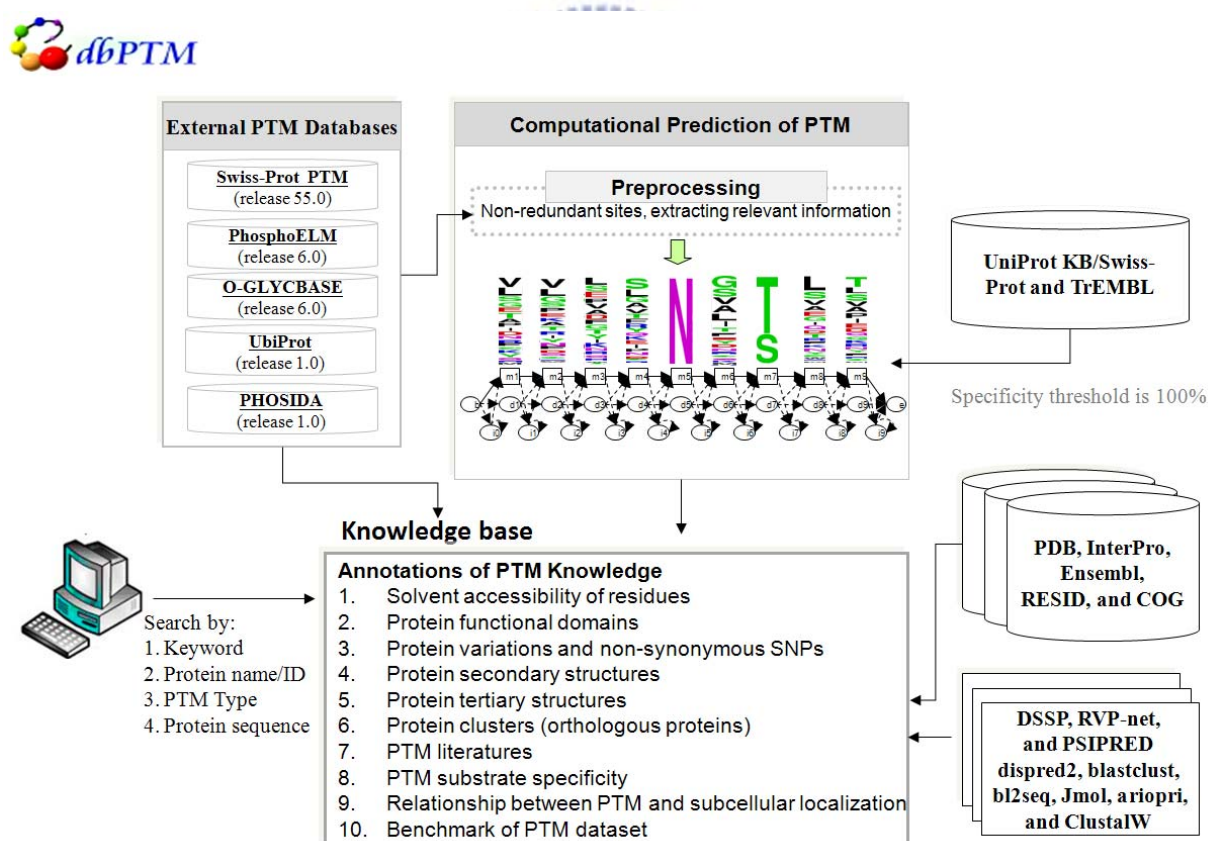


Figure 2.7 System flow for constructing dbPTM.

2.4.1 Integration of External PTM Databases

Five external biological databases related to protein post-translational modification information, UniProt KB/Swiss-Prot [3], Phospho.ELM [2], PHOSIDA [45], O-GLYCBASE [46], and UbiProt [47], are integrated into the proposed resource. Both the experimentally validated PTM sites and the putative PTM sites, which are annotated as “by similarity”, “potential” or “probable” in the ‘MOD_RES’, “CARBOHYD”, “LIPID” and “CROSSLNK” fields, have been extracted from the UniProt KB/Swiss-Prot database. As shown in **Table 2.3**, release 55.0 of UniProt KB/Swiss-Prot contributes 17957 experimental validated PTM sites within 8086 proteins, and 124933 putative PTM sites within 29356 proteins. The Phospho.ELM entries store information about substrate proteins with the exact positions of residues are known to be phosphorylated by cellular kinases. 16428 experimentally verified phosphorylation sites within 4026 proteins were obtained from Phospho.ELM version 2 [2]. PHOSIDA stores more than 6600 in vivo phosphorylation sites which were identified by mass spectrometry-based proteomics on 2244 proteins in response to EGF stimulation. O-GLYCBASE [46] Version 6.00 provides 242 glycoproteins containing 2,765 experimentally verified O-linked, N-linked, and C-linked glycosylation sites. Moreover, 185 glycoproteins in O-GLYCBASE are corresponded to Swiss-Prot proteins, which have 2,353 experimentally verified glycosylation sites. UbiProt, which contains 417 ubiquitylated proteins and 165 ubiquitylation sites, was also integrated into dbPTM.

Table 2.3 Data statistics of the integrated PTM resource.

Resource	Version	Description	Statistics
Swiss-Prot	55.0	Experimental Post-Translational Modifications (PTMs)	17,957 PTM sites within 8,086 proteins
		Putative PTMs (annotated as “by similarity”, “potential” or “probable” in the ‘MOD_RES’, “CARBOHYD”, “LIPID” and “CROSSLNK” fields)	124,933 PTM sites within 29,356 proteins
PhosphoELM	7.0	Experimental phosphorylation sites	16,428 phosphorylation sites within 4,026 proteins
PHOSIDA	1.0	In vivo phosphorylation sites which was identified by mass spectrometry-based Proteomics	More than 6600 phosphorylation sites on 2244 proteins in response to EGF stimulation
O-GLYCBASE	6.0	Experimental glycosylation sites	2,353 PTM sites within 185 glycoproteins
UbiProt	1.0	Ubiquitylated protein and ubiquitylation sites	417 Ubiquitylated proteins and 165 ubiquitylated sites

2.4.2 Computational Annotation of PTM Sites

To provide the post-translational modification information of the PTM un-annotated proteins available from Swiss-Prot, we adopted computational tools for identifying the post-translational modifications of the Swiss-Prot proteins. Our previous work, namely KinasePhos [56], incorporated the profile Hidden Markov Model (HMM) to identify kinase-specific phosphorylation sites with about 87% prediction accuracy [55], which was compared with several phosphorylation prediction tools such as NetPhos [57], DISPHOS [58], and rBPNN [59] (Table 2.4).

Table 2.4 Comparisons between KinasePhos, NetPhos, DISPHOS and rBPNN.

Residue types	NetPhos	DISPHOS	rBPNN	KinasePhos
Serine	0.69	0.75	No data	0.86
Threonine	0.72	0.80	No data	0.91
Tyrosine	0.61	0.82	No data	0.84
Total or average	No data	No data	0.87	0.87

First of all, the PTMs should be categorized by their modification types and be investigated each type of PTM with enough samples in advance. Based on the KinasePhos-like method, removing the redundancy of PTM sites from various PTM databases is important. Before the model training, the positive and negative set should be constructed. Here we define the PTM residues as positive set, while those non-PTM residues in the same protein from which positive sites were taken are regarded as negative set, instead of using proteins randomly picked from the Swiss-Prot/tremble databases. In general, we make the equal sizes of positive set and negative set. After that, we employed a traditional sliding window strategy to represent the PTM or non-PTM peptides. Given the window length n , a fragment of $2n+1$ residues centering on PTM site was adopted to represent a PTM peptide.

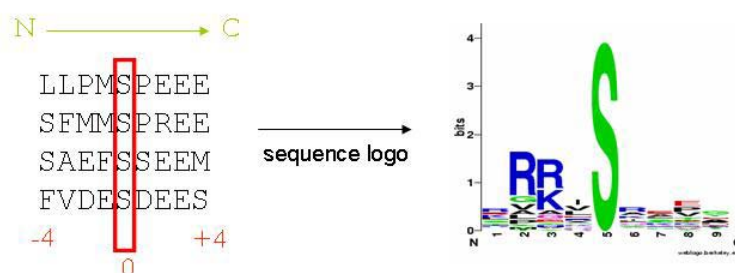


Figure 2.8 An example of 9-mer (window length n is set to 4) phosphorylated peptides and sequence logo.

As shown in **Figure 2.8**, for example, the phosphorylated residue was define as the position 0 and the positions (-4 ~ -1) and (+1 ~ +4) designated the residues surrounding the phosphorylation residue, such as serine. However, the serines, threonines and tyrosines, which are not annotated as phosphorylation residues, within the experimentally validated phosphorylated proteins are selected as negative sets, i.e., the non-phosphorylated sites. Different values of n varying from 4 to 10 were used to determine the optimized window length. For the sake of the observation of the amino acid distribution surrounding the PTM residues, we make up the $(2n+1)$ -mer sequence logos[60, 61] of the phosphorylation sites which is shown in **Figure 2.8**. The sequence logos are a graphical representation of an amino acid or nucleotide multiple sequence alignment. Each logo consists of stacks of symbols, one stack presents each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of the symbols within the stack indicates the relative frequency of each amino or nucleotide at that position. **Figure 2.9** shows the system flow of KinasePhos-like method.

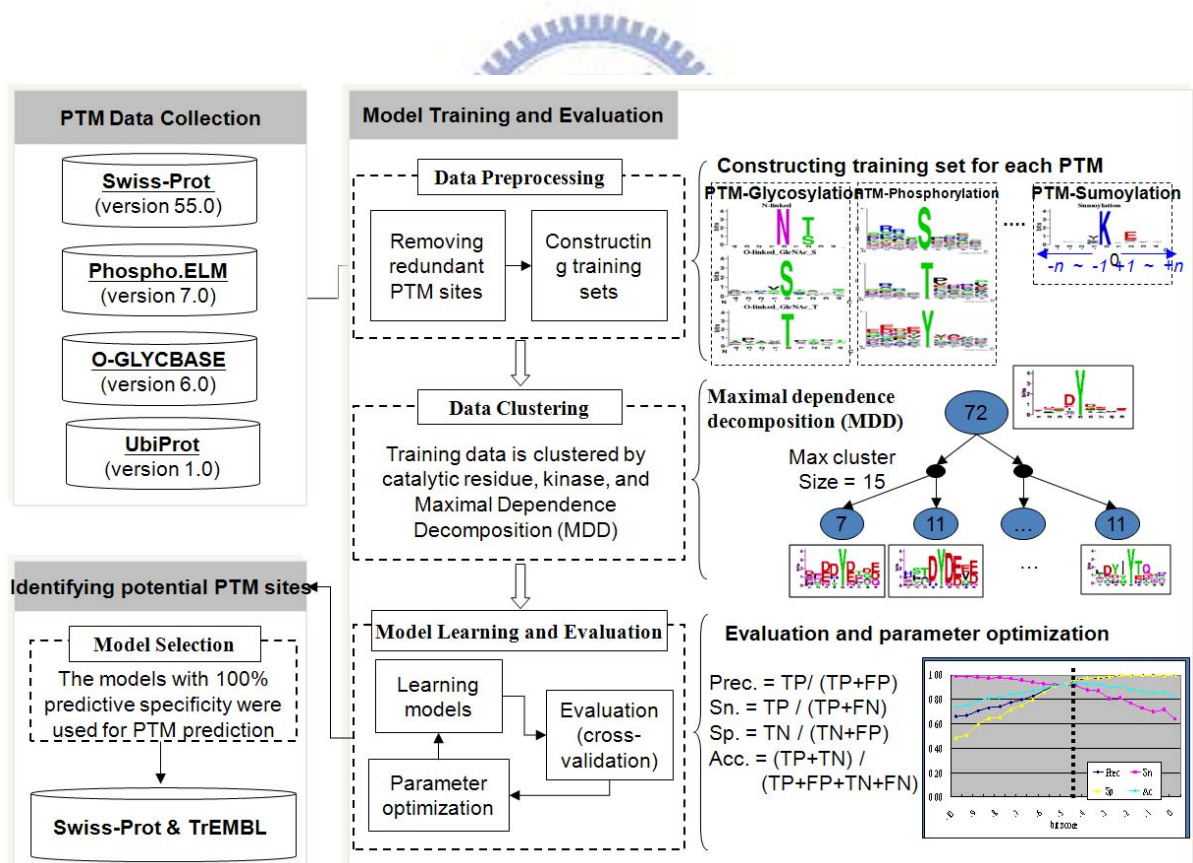


Figure 2.9 System flow of KinasePhos-like method.

The positive set for training might contain several homologous sites from homologous proteins. If the training data are highly similar with too many homologous sites, the prediction

accuracy will be overestimated. To avoid the overestimation, we filtered the identical training sequences from homologous proteins in positive set. Thus, we obtained a high quality training set with non-redundant positive set for model training.

The PTM site sequences in the positive sets with a larger size could be alternatively clustered by MDD method in order to increase the predictive sensitivity and specificity of the models. The Maximal Dependence Decomposition (MDD) [62] employs statistical χ^2 -test to group an set of aligned signal sequences to moderate a large group into subgroups that capture the most significant dependencies between positions. In previous work, MDD was proposed to group the splice sites during the identification process of splice site prediction [62]. However, in our study, we group protein sequences instead of nucleotides. In order to reduce the data complexity of the phosphorylated sites when applying MDD, we categorize the twenty types of amino acids into five groups such as neutral, acidic, basic, aromatic and imino groups, as the mapping given in **Table 2.5**. Then, we implement the MDD algorithm in JAVA programming language for amino acids and apply it to cluster PTM site sequences with large data sets.

Table 2.5 The amino acids group used in MDD.

Group name	Amino acids
Neutral	threonine (T), valine (V), leucine (L), isoleucine (I), methionine (M), glycine (G), alanine (A), serine (S), cysteine (C)
Acid	aspartic acid (D), asparagine (N), glutamic acid (E), glutamine (Q)
Basic	lysine (K), arginine (R), histidine (H)
Aromatic	phenylalanine (F), tyrosine (Y), tryptophan (W)
Imino	proline (P)

To perform the null hypothesis test of independence on a pair of i -th and j -th positions of a PTM site, we formed a 5 x 5 contingency table, as shown in **Figure 2.10**, by counting the observed number X_{mn} of PTM site sequence where the i -th amino acid A_i was m and j -th amino acid A_j was n (for simplicity, we have encoded neutral, acid, basic, aromatic, imino as 1, 2, 3, 4, 5, respectively) from a sample of X PTM site sequences. The numbers X_{mR} and X_{Cn} in Figure are row sums and column sums, respectively. It is clear that $\sum_{m=1}^5 X_{mR} = \sum_{n=1}^5 X_{Cn} = X$. The test statistic used is as follows:

$$\chi^2(A_i, A_j) = \sum_{m=1}^5 \sum_{n=1}^5 \frac{(X_{mn} - E_{mn})^2}{E_{mn}},$$

where

$$E_{mn} = \frac{X_{mR} X_{Cn}}{X}$$

is the expected number of amino acids in which the i -th position A_i is m and the j -th position A_j is n from a sample of X PTM site sequences when the null hypothesis of independence was true. To determine the rejection region for the null hypothesis, we have specified a numerical value α for the Type I error of the test, according to a χ^2 -distribution with degrees of freedom $(5-1) \times (5-1) = 16$, and then the critical point, K , was computed as follows:

$$P(\text{null hypothesis is rejected when it is true}) = P(\chi^2(A_i, A_j) \geq K \mid \text{null hypothesis}) = \alpha.$$

PTM

↓

L L P M **S** P E E E
 S F M M **S** P R E E
 S A E F **S** S E E M
 F V D E **S** D E E S

A₄ A₃ A₂ A₁ A₊₁ A₊₂ A₊₃ A₊₄

↓

A contingency table between two positions in PTM site

$A_i \setminus A_j$	Neutral	Acid	Basic	Aromatic	Imino	Total
Neutral	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{1R}
Acid	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{2R}
Basic	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	X_{3R}
Aromatic	X_{41}	X_{42}	X_{43}	X_{44}	X_{45}	X_{4R}
Imino	X_{51}	X_{52}	X_{53}	X_{54}	X_{55}	X_{5R}
Total	X_{C1}	X_{C2}	X_{C3}	X_{C4}	X_{C5}	X

Figure 2.10 A 5x5 contingency table between two positions in PTM site.

The MDD is a recursive process to divide the positive sets into tree-like subgroups. When applying MDD to cluster the sequences of a positive set, a parameter, i.e., the minimum-cluster-size, should be set. If the size of a subgroup is less than the minimum-cluster-size, the subgroup will not be divided any more. The MDD process terminates when all the subgroup sizes are less than the minimum-cluster-size. When considering a MDD-clustered data set, for instance, MDD-clustered PKA catalytic serine (S_PKA), the model are trained separately from the subgroups of the phosphorylated sites

resulted by MDD. Each model is used to search in the given protein sequences for the phosphorylated sites. A positive prediction of the model group is defined by at least one of the model makes a positive prediction, whereas a negative prediction is defined as all the models make negative predictions.

Profile Hidden Markov Models (HMMs) are trained from the PTM site sequences aligned without gaps of the positive sets. An HMM describes a probability distribution over a potentially infinite numbers of sequences [63]. It can be used to detect distant relationships between amino acids sequences. Here, we use the software package HMMER[63] (version 2.3.2) to build the models, to calibrate the models and to search the putative PTM sites against the protein sequence. The emission and transition probabilities are generated from each of the training set to capture the characteristics of the training sequences. All residue types of the PTM sites with enough data set were taken to train the HMM; moreover, as well as the sets of the kinase-specific or MDD-clustered sets of PTM sites.

After the models are trained, it is necessary to evaluate whether the models are fitted or not. The following measures of the predictive performance of the models are then calculated: Precision (Pr) = $TP / (TP+FP)$, Sensitivity (Sn) = $TP / (TP+FN)$, Specificity (Sp) = $TN / (TN+FP)$ and Accuracy (Ac) = $(TP + TN) / (TP+FP+TN+FN)$, where TP, TN, FP and FN represent true positive, true negative, false positive and false negative predictions, respectively. In general, we make the equal sizes of the positive samples and the negative samples during the cross-validation processes.

To evaluate the trained models, two cross-validation methods, k-fold cross-validation and leave-one-out cross-validation, are applied in this study. For a large positive set, i.e., the number of a positive set of PTM sites is equal or greater than thirty sites, the k-fold cross-validation is used to evaluate the model trained from the data set. The size of the negative set, which is constructed by randomly selected from the corresponding non-PTM sites, is equal to the size of positive set. The experiments are repeated for 20 times and the average precision, sensitivity, specificity and accuracy are calculated. Furthermore, in order to avoid a skewed sampling during the cross-validation process, for a small positive set (less than 30), the leave-one-out cross-validation is alternatively applied. Similarly, the negative set in this cross-validation is constructed by the same strategy as the k-fold cross-validation.

For each training set of PTM sites, the best performed model is selected and used to identify the PTM sites within the input protein sequences by HMMsearch [63]. To search the

hits of a model, HMMER returns both a HMMER bit score and an expectation value (E-value). The score is the base two logarithm of the ratio between the probability that the query sequence is a significant match and the probability that it is generated by a random model. The E-value represents the expected number of sequences with a score greater than or equal to the returned HMMER bit scores. While decreasing the E-value threshold favors finding true positives, increasing the E-value threshold favors finding true negatives. We select the HMMER score as the criteria to define a HMM match. A search of a model with the HMMER score greater than the threshold t of bit score is defined as a positive prediction, i.e., a HMM recognizes a PTM site. The threshold t of each model is decided by maximizing the accuracy measure during a variety of cross-validations with the HMM bit score value range from 0 to -10. For instance, **Figure 2.11** depicts the optimization of the threshold of the HMM bit scores in the model of phosphorylated serine which is catalyzed by PKA (S_PKA). The threshold of the S_PKA model is set to -4.5 to maximize the accuracy measure of the model.

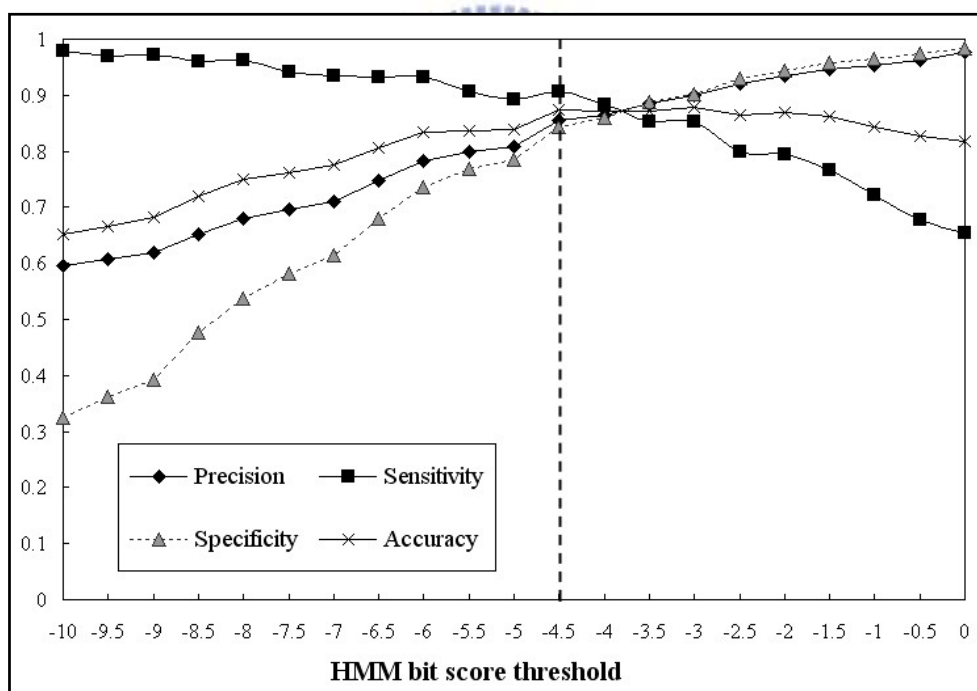


Figure 2.11 The optimization of the threshold of the HMM bit score in the model of phosphorylated serine which is catalyzed by PKA.

KinasePhos-like method was applied to 20 types of PTM with over 30 experimentally verified PTM sites, which were learned the computational models and then adopted to identify potential PTM sites against all Swiss-Prot proteins. The learned models were evaluated using k -fold cross validation. To reduce the number of false positive predictions

when the potential PTM sites were fully detected against the Swiss-Prot protein sequences, the predictive parameters were set to ensure a predictive specificity of 100%.

2.4.3 Structural and Functional Annotations

In order to provide more effective information about protein structural and functional annotations relevant to protein post-translational modification, a variety of biological databases, such as Swiss-Prot [64], Ensembl [65], InterPro [66], Protein Data Bank [67], and RESID [4], are integrated.

Protein variation is the change of amino acids in polypeptides. As shown in **Table 2.6**, Swiss-Prot contributes 32,101 protein variants corresponding to 6,115 proteins, where 47 variant residues are located at the PTM sites and 267 variant residues are located surrounding 236 PTM sites (-4 ~ +4 AA). Furthermore, Single Amino acid Polymorphism (SAP) is the amino acid variation corresponding to the genetic variation as the definition of non-synonymous Single Nucleotide Polymorphism (SNP) in genomic sequence. The amino acid variants may have an impact on protein folding, active sites, or the overall solubility and stability of a protein. SAP is the type of variation most frequently related to human diseases [64]. Therefore, when the amino acid variations occur in the post-translational modification sites or the surrounding residues, they may affect the recognition of PTM sites by catalytic kinases. 23,378 human non-synonymous single nucleotide polymorphisms (SNP) located at 7,230 Swiss-Prot human proteins were obtained from the variation part of Ensembl database [65].

InterPro provides 1113928 entries corresponding to 247238 Swiss-Prot proteins. We found that about 65% of Swiss-Prot annotated PTM sites are located at InterPro annotated protein functional domains. The RESID [4] protein modifications database is integrated into dbPTM to provide PTM related information such as mass difference, chemical formula, enzymatic activities, literature citations, Gene Ontology (GO) cross-references, structure diagrams, and molecular models.

The latest version of Protein Data Bank (PDB) contains 31,721 tertiary structures corresponding to 6,806 Swiss-Prot protein entries (**Table 2.6**). For the proteins with known tertiary structures, the DSSP [68] program was used to extract the true secondary structure and solvent accessibility for those 6,808 Swiss-Prot proteins. Solvent accessibility of amino acids residues is important for both the structure and function of proteins, especially the

post-translational modifications studied in this investigation. Protein secondary structure is the regular arrangement of amino acid residues in a segment of a polypeptide chain, where each amino acid is assigned a structure state, helix (H), strand (E) or coil (C). There are 1,124 experimentally verified PTMs have the true secondary structure and solvent accessibility.

Table 2.6 List of the integrated external databases and programs for structural and functional annotations.

Database		
Name	Description	Statistics
Swiss-Prot [42, 64]	Protein variants	32,101 variants corresponding to 6,115 proteins
RESID [4]	Annotations of Post-Translational Modification (PTM)	431 PTM annotations
InterPro [66]	Protein domain	1,113,928 entries can be corresponded to 247,238 Swiss-Prot entries
Protein Data Bank [67]	Protein structures	30,937 entries can be corresponded to 10,274 Swiss-Prot proteins
COG [69]	Clusters of orthologous groups of proteins	138,458 proteins form 4873 COGs in 66 genomes of unicellular organisms. The eukaryotic orthologous groups (KOGs) include proteins from 7 eukaryotic genomes consisting of 4852 clusters of orthologs, which include 59,838 proteins.
Program		
Name	Description	Version
KinasePhos [56]	Identifying Kinase-specific phosphorylation sites	Release 1.0
DSSP [68]	Calculating the secondary structure and solvent accessibility of residues	April 1,2000
RVP-net [70]	Predicting the solvent accessibility of residues	Release 1.0
PSIPRED [21]	Predicting the protein secondary structures	Release 2.45
DISOPRED2 [35]	Predicting the protein disorder region	Version 2.1
Jmol ⁶	An open-source Java viewer for chemical structures in 3D	Release 11.2.4
Weblogo [60]	Generating sequence logo for PTM substrates	Release 2.8.2
Blast [71]	The programs BLASTCLUST and BL2SEQ were used to remove the redundant PTM sites	Release 2.2.12
ClustalW [72]	Multiple sequences alignment in orthologous protein clusters	Release 1.83

However, only ~ 4% of Swiss-Prot proteins have the known tertiary structures. For proteins without known tertiary structures, two previously published tools, RVP-net [70],

⁶ Jmol: <http://www.jmol.org/>

PSIPRED [21] and DISOPRED [35], were applied to predict the solvent accessibility, secondary structure and protein disorder region, respectively. RVP-net [70] presents a feed-forward type neural network which can predict a real value ranging from 0% to 100% of Accessible Surface Areas (ASA) for amino acid residues, based on their neighborhood information. We applied the RVP-net program [70] to fully predict the real-valued ASA for the amino acid residues of all Swiss-Prot proteins. By selecting a suggested threshold [70] (i.e. 25%), the residues with larger ASA values are viewed as surface residues. Moreover, dynamically disordered regions appear to be relatively abundant in eukaryotic proteomes. The DISOPRED server allows users to submit a protein sequence, and returns a probability estimate of each residue in the sequence being disordered.

2.4.4 Benchmark of PTM Prediction

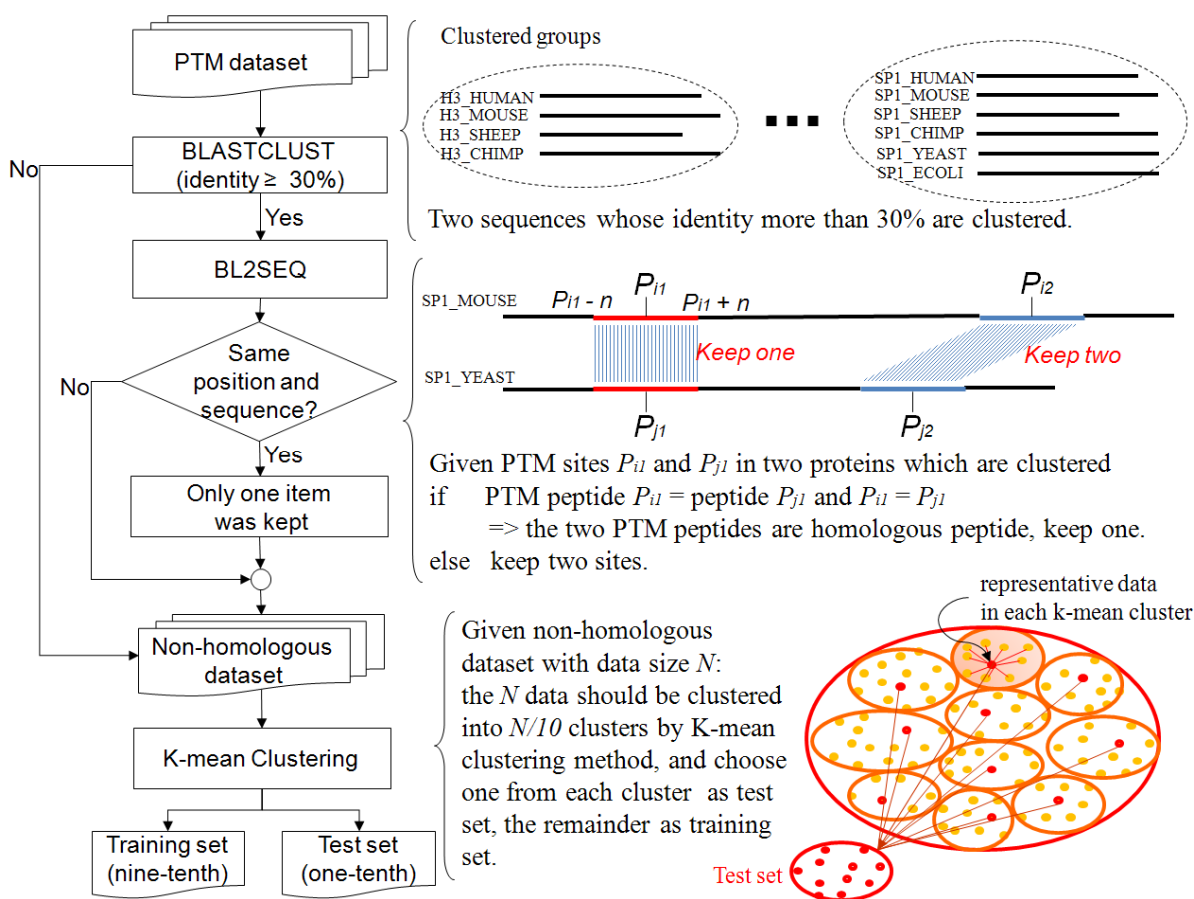


Figure 2.12 Flowchart of constructing PTM benchmark dataset.

With the recent exponential increase in some PTM sites identified by mass spectrometry, the opportunity has arisen to analyze the motifs surrounding each PTM site and use these motifs

to identify potential PTM sites in proteins. Up to now, about 40 PTM site prediction servers have been developed and made publicly available through the internet. Several representative prediction servers of PTMs are show in **Table 2.7** which displays the method and predictive performance of each PTM prediction server.

Table 2.7 Several representative PTM prediction servers.

PTM type	Prediction server	Method	Predictive performance
Glycosylation	NetOGlyc 3.1 [73]	NN	Sn = 76%, Sp = 93%
	NetNGlyc 1.0	NN	N/A
	DictyOGlyc [74]	NN	
Phosphorylation	KinasePhos [56]	MDD+HMM	Ac = 87%
	GPS [75]	Group-based scoring	Sn = 91.8%, Sp = 85%
	PPSP [76]	Bayesian theory	Ac ~ = 88%
	NetPhosK [77]	NN	
N-terminal acetylation	PredPhospho [78]	SVM	Ac = 83%~95%
	NetAcet [79]	NN	Sn = 75%, Sp = 92%
Methylation	MeMo [80]	SVM	Ac of lysine = 67.1% Ac of argine = 86.7%
	PAIL [81]	Bayesian Discriminant Method	Ac = 89%
Sulfation	Sulfinator [82]	HMM	Ac = 98%
Palmitoylation	NBA-Palm [83]	Naïve Bayes Algorithm	Ac = 86%
	CSS-Palm [84]	Clustering and Scoring Strategy	Sn=82%, Sp=83%
N-terminal myristoylation	NMT [85]		Sn = 95%
	Myrist [86]	HMM	Ac ~ = 97%
	Myristoylator [87]	NN	Sn= 93.8%, Sp= 97.9%
Sumoylation	SUMOsp [88]	GPS+MotifX	Ac = 92.71%
Glycosylphosphatidylinositol (GPI) anchoring	GPI-SOM [89]	NN	
	big-II [90]	N/A	Ac = 83%

Abbreviations: Sn, sensitivity; Sp, specificity; Ac, accuracy; NN, neural network; MDD, maximal dependence decomposition; HMM, hidden markov model; SVM, support vector machine.

A PTM benchmark comprising the experimental sites for each PTM type was built to provide a standard for evaluating the predictive performance of various prediction tools. **Figure 2.12** shows the process for compiling the PTM benchmark, which is based on the previous work of Chen *et al.* [80]. To eliminate the redundancy, the protein sequences containing the same type of PTM sites were grouped by a threshold of 30% identity using BLASTCLUST [71]. If the identify of two protein sequences is greater than 30%, then the

fragment sequences of the substrates were re-aligned with BL2SEQ. If the fragment sequences of two substrates with the same location are identical, then only one of the substrate sequences was included in the benchmark. After the reduction of homologous dataset, the non-homologous dataset could be further categorized into training set and test set. Usually, one-tenth of the non-homologous dataset is extracted as the independent test set. The remainder (nine-tenth) is defined as the training set. To avoid the biased sampling of test set, we adopted the k-mean clustering method to cluster the non-homologous dataset into $N/10$ clusters, given non-homologous dataset with data size N . k-mean clustering method can let the PTM peptide sequences that are similar to each other are clustered together. Then, select one point from each cluster and join them into the test set. These selected points that from each cluster are uniformly distributed. Therefore, the constructed test set could be not biased.

K-means Clustering Algorithm

The k-means algorithm (J.A. Hartigan and M.A. Wong, 1979) is an algorithm to cluster n objects based on attributes into k partitions, $k < n$. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data. It assumes that the object attributes form a vector space. The objective it tries to achieve is to minimize total intra-cluster variance, or, the squared error function

$$V = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

where there are k clusters S_i , $i = 1, 2, \dots, k$ and μ_i is the centroid or mean point of all the points .

The most common form of the algorithm uses an iterative refinement heuristic known as Lloyd's algorithm. Lloyd's algorithm starts by partitioning the input points into k initial sets, either at random or using some heuristic data. It then calculates the mean point, or centroid, of each set. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters, and algorithm repeated by alternate application of these two steps until convergence, which is obtained when the points no longer switch clusters (or alternatively centroids are no longer changed).

2.5 Results

dbPTM integrates several databases to accumulate known protein modifications, as well as the putative protein modifications predicted by a series of accurately computational tools. KinasePhos [55, 56], which incorporates the profile hidden Markov model (HMM) to identify kinase-specific phosphorylation sites, is integrated into dbPTM. Moreover, dbPTM is a knowledge base for protein post-translational modification, which comprises the modified sites, solvent accessibility of substrate, protein secondary and tertiary structures, protein domains and protein variations. Literature related to PTM, protein conservations and substrate specificity are also provided in the resource.

2.5.1 Performance of PTM Computational Models

KinasePhos-like method was applied to 20 types of PTM with over 30 experimentally verified PTM sites, which were learned the profile hidden Markov models and then adopted to identify potential PTM sites against all Swiss-Prot proteins. The profile hidden Markov models of each PTM are constructed by HMMER package [63] (version 2.3.2); for example, a HMM of N-linked (GlcNAc) asparagine is shown in **Figure 2.13b**, which is difficult for biologist to understand what the profile HMM parameters say. Therefore, **Figure 2.13a** shows the graphical representation of profile hidden Markov model which let users understand the framework of HMM concretely. The HMM is mainly composed of three kinds of state, such as match state, insertion state, and deletion state. The HMM in **Figure 2.13a** contains nine match states (squares labeled m1, m2, ..., m9), each of which has 20 residue emission probabilities, shown with sequence logos. The first column of **Figure 2.13b** is the node number (1 ...9) corresponding to the match state (m1 ...m9) of **Figure 2.13a**. Insertion states (circles labeled i0 – i9) also have 20 emission probabilities each. Deletion states (circles labeled d1-d9) are ‘mute’ states that have no emission probabilities. A begin and end state are included (b,e). State transition (m->m, m->i, m->d, i->m, i->i, d->m, d->d, b->m, m->e) probabilities are shown as arrows.

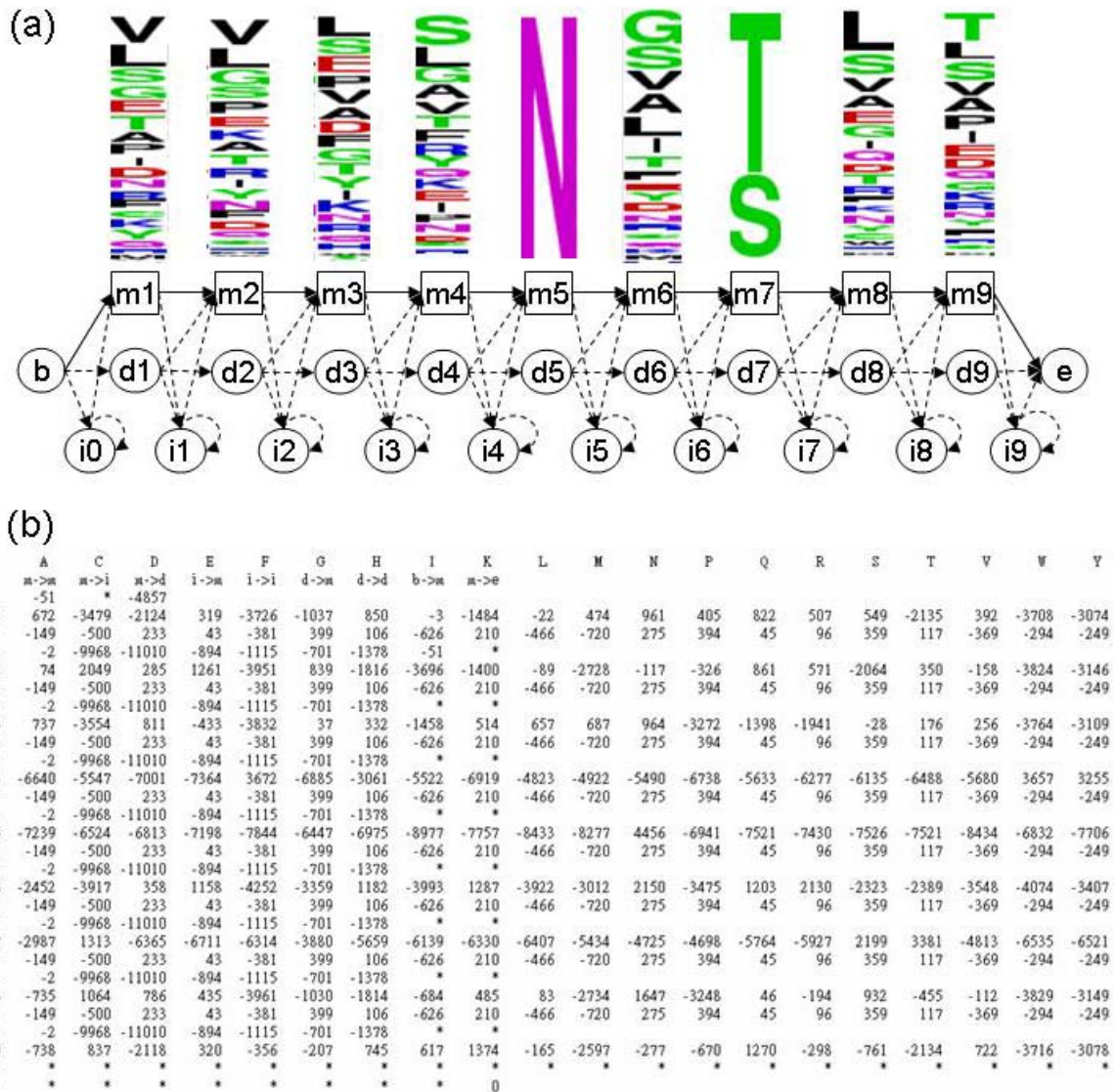


Figure 2.13 The profile hidden Markov model of N-linked (glcNAc) asparagine.

After the evaluation of the learned models using k -fold cross validation, the parameters of the predictive models that achieved the best predictive accuracy are listed in

Table 2.8, which contains the number of known PTM sites, method, cross-validation method, precision (Pr), sensitivity (Sn), specificity (Sp), accuracy (Ac). Twenty PTMs were trained with the computational models. The parameters, including window length and HMMER bit score, were optimized iteratively during cross-validation. The models with good predictive accuracy will be selected to implement the PTM prediction. In contrast, those PTM models with poor predictive performance will be improved in two ways which include changing the machine learning method and considering other features such as secondary structure and solvent accessibility.



Table 2.8 Parameters and predictive performance of the PTM computational models.

PTM Types	Substrates	No. of PTM sites	Window length	HMM bit score	Pr.	Sn.	Sp.	Ac.
N-linked glycosylation	Asparagines (GlcNAc)	3019	-6 ~ +6	-4.5	0.85	0.98	0.83	0.91
	Serine (GalNAc)	212	-6 ~ +6	-5	0.80	0.85	0.79	0.82
	Serine (GlcNAc)	35	-6 ~ +6	-6	0.81	0.71	0.83	0.77
O-linked glycosylation	Serine (Man)	79	-6 ~ +6	-5	0.88	0.74	0.90	0.82
	Threonine (GalNAc)	386	-6 ~ +6	-4.5	0.81	0.75	0.82	0.79
	Threonine (GlcNAc)	42	-6 ~ +6	-4	0.77	0.82	0.76	0.79
	Threonine (Man)	83	-6 ~ +6	-7	0.83	0.88	0.81	0.85
	Lysine (Gal)	46	-6 ~ +6	-5	1.00	1.00	1.00	1.00
C-linked glycosylation	Tryptophan (Man)	49	-6 ~ +6	-0.5	1.00	0.98	1.00	0.99
Phosphorylation	Serine (kinase-specific)	4382	-6 ~ +6	-5.5	0.88	0.84	0.88	0.86
	Threonine (kinase-specific)	1030	-6 ~ +6	-4	0.91	0.92	0.91	0.91
	Tyrosine (kinase-specific)	901	-6 ~ +6	-5	0.86	0.81	0.87	0.84
	Histidine	41	-6 ~ +6	-3	0.90	0.80	0.91	0.86
Acetylation	Alanine	403	0 ~ +6	-6	0.64	0.72	0.60	0.66
	Lysine	292	0 ~ +6	-6	0.77	0.73	0.79	0.76
	Methionine	199	0 ~ +6	-4	0.83	0.75	0.85	0.80
	Serine	402	0 ~ +6	-4	0.59	0.84	0.42	0.63
	Threonine	58	0 ~ +6	-6	0.85	0.53	0.90	0.71
Methylation	Arginine	180	-6 ~ +6	-1	0.97	0.78	0.98	0.88
	Lysine	407	-6 ~ +6	0	0.83	0.60	0.88	0.74
N-myristoyl glycine	Glycine	100	-6 ~ +6	-10	0.99	0.91	0.99	0.95
N-palmitoyl csteine	Cysteine	58	-6 ~ +6	-5	0.88	0.93	0.88	0.91
S-palmitoyl csteine	Cysteine	169	-6 ~ +6	-4	0.94	0.70	0.95	0.83
S-farnesyl cysteine	Cysteine	63	-6 ~ +6	-4	0.78	0.89	0.75	0.82
S-geranylgeranyl cysteine	Cysteine	52	-6 ~ 0	-6	0.69	0.88	0.61	0.74
Hydroxylation	Proline	635	-6 ~ +6	-4	0.82	0.88	0.81	0.84
	Lysine	83	-6 ~ +6	-3	0.97	0.84	0.98	0.91
Amidation	Asparagine	77	-6 ~ +6	-5	1.00	1.00	1.00	1.00
	Glycine	143	-6 ~ +6	-5	1.00	0.96	1.00	0.98
	Isoleucine	72	-6 ~ +6	-4	0.92	0.85	0.92	0.88
	Leucine	263	-6 ~ +6	-4	0.92	0.92	0.92	0.92
	Methionine	88	-6 ~ +6	-8	1.00	1.00	1.00	1.00
	Phenylalanine	433	-6 ~ +6	-1	0.97	0.99	0.97	0.98
	Proline	95	-6 ~ +6	-7	0.96	1.00	0.96	0.98
	Tyrosine	88	-6 ~ +6	-7	1.00	1.00	1.00	1.00
Sulfation	Tyrosine	162	-6 ~ +6	-4.5	0.96	0.91	0.96	0.94
Sumoylation	Lysine	77	-6 ~ +6	-5	0.86	0.75	0.88	0.81
Ubiquitination	Lysine	284	-6 ~ +6	-5	0.82	0.67	0.85	0.76
Pyrrolidone Carboxylic Acid	Glutamate acid	598	0 ~ +6	-4	0.76	0.69	0.79	0.74
4-carboxyglutamate	Glutamate	371	-6 ~ +6	-4	0.92	0.90	0.93	0.91
Nitration	Tyrosine	47	-6 ~ +6	-3	0.85	0.65	0.81	0.73
S-diacylglycerol cysteine	Cysteine	36	-6 ~ +6	-5	1.00	0.94	1.00	0.97
Average					0.87	0.82	0.86	0.84

Abbreviations: Pr., precision; Sn., sensitivity; Sp., specificity; Ac., accuracy.

2.5.2 Data Statistics

Table 2.9 Data statistics of dbPTM.

PTM types	Substrates	Number of experimental sites	Number of putative sites from Swiss-Prot	Number of putative sites in dbPTM
N-linked Glycosylation	Asparagine and lysine	3,036	72,125	479,955
O-linked Glycosylation	Lysine, proline, serine, threonine, and tyrosine	1,896	2,558	386,545
C-linked Glycosylation	Tryptophan	49	31	4,015
Phosphorylation	Serine, threonine, tyrosine, aspartate, histidine or cysteine	22,363	27,200	1,815,472
Acetylation	N-terminal of some residues and side chain of lysine or cysteine	2,071	5,143	1,206
Amidation	Generally at the C-terminal of a mature active peptide after oxidative cleavage of last glycine	2,150	1,117	24,352
Hydroxylation	Generally of asparagine, aspartate, proline or lysine	1,033	1,074	9,743
Methylation	Generally of N-terminal phenylalanine, side chain of lysine, arginine, histidine, asparagine or glutamate, and C-terminal cysteine	746	2,846	18,716
Pyrrolidone Carboxylic Acid	N-terminal glutamine which has formed an internal cyclic lactam.	598	584	12,322
Gamma-Carboxyglutamic Acid	Glutamate	371	361	1,924
Farnesylation	Cysteine	61	216	5,349
Myristoylation	Glycine	108	765	10,998
Palmitoylation	Cysteine	210	3,582	27,841
Geranylgeranylation	Cysteine	47	819	14,317
S-diacylglycerol cysteine	Cysteine	36	1,529	8,977
GPI anchoring	C-terminal asparagine, aspartate, and serine	27	681	-
Deamidation	Asparagine and glutamine (needs to be followed by a G)	38	26	2,022
Sulfation	Serine, threonine, and tyrosine	165	570	15,654
Sumoylation	Glycyl lysine isopeptide (Lys-Gly)(interchain with G-Cter in SUMO)	77	259	10,342
Ubiquitylation	Glycyl lysine isopeptide (Lys-Gly)(interchain with G-Cter in ubiquitin)	286	516	8,865
Nitration	Tyrosine	47	5	1,432
ADP-ribosylation	Arginine	3	203	-
Formylation	Of the N-terminal methionine	28	35	-
Citrullination	Arginine	27	91	-
Bromination	Tryptophan	18	3	-
FAD	Tyrosine, histidine, and cysteine	12	116	-
S-nitrosylation	Cysteine	9	93	-
Others		889	2,358	-
Total		36,466	124,933	2,860,047

Table 2.9 summarizes the statistics of the experimental PTM sites and the predicted PTM sites in the updated dbPTM. This updated dbPTM had a total of 36466 experimental PTM sites. The experimental PTM sites obtained from Swiss-Prot, Phospho.ELM, O-GLYCBASE and UbiProt were categorized by PTM type, and the number of non-redundant PTM sites was calculated. For instance, the database contains 22363 experimental phosphorylation sites and 2071 experimental acetylation sites. Besides the experimental PTM sites, a machine learning method was adopted to build predictive models for twenty types of PTM. The computational predictions are described in detail in our previous works [19, 55, 56, 91]. These models were used to search the potential PTM sites against Swiss-Prot protein sequences. As listed in **Table 2.9**, 2860047 sites for all PTM types were detected.

Table 2.10 The statistics of the putative phosphorylation sites, sulfation sites, and glycosylation sites with different thresholds of the Accessible Surface Area (ASA) of residues.

Accessible Surface Area (ASA)	No. of phosphorylated serine, threonine and tyrosine	No. of sulfated tyrosine	No. of N-linked glycosylated asparagine	No. of C-linked glycosylated tryptophane
≥ 0%	1,346,067	189,457	38,416	5,478
≥ 25%	652,756	13,315	33,836	51
≥ 50%	32,816	2	4,973	0
≥ 75%	7	0	7	0

Statistics of PTM Sites and Solvent Accessibility

The numbers of putative phosphorylation and sulfation sites, where the ASA of the substrates are greater than 25% (defined as the residue locating at the protein surface), are 652,756 and 13,315, respectively. There are a total of 33, 887 predicted N-linked glycosylations of asparagine and C-linked glycosylations of tryptophan.

Statistics of PTM Sites and Referable Literatures

PTM profile, which annotates the PTM sites and related literatures, can help biologist to understand the relationship between the protein function and PTMs. With the comprehensive annotation of PTMs from dbPTM, the experimentally verified and computational detected PTMs of a protein can be provided to users. However, the relationship between protein function and PTMs is not understood while only provide the PTM sites. Thus, the related literatures about the protein and PTMs are extracted from literature databases and integrated in PTM profiles. In release 51.2 of Swiss-Prot knowledgebase, there are totally 490,996 literatures against 243,975 proteins. Based on searching keyword in literature titles, the

number of literatures about several common PTMs is listed in **Table 2.11**. Users not only get the PTM information, but also look into the relationship between PTM sites and protein functions. There are totally 490,996 literatures against 243,975 proteins.

Table 2.11 The statistics of literatures extracted from release 55.0 of Swiss-Prot knowledgebase in several common PTMs.

PTM type	Keyword	No. of literatures	No. of proteins
Glycosylation	Glyco, glycosylation, glycosylated, O-linked, N-linked, C-linked, carbohyd, carbohydrate	1,793	1,422
Phosphorylation	Phospho, phosphorylated, phosphorylation	5,944	4,177
Acetylation	Acetyl, acetylated, acetylation	992	842
Methylation	Methyl, methylated, methylation	310	222
Palmitoylation	Palmitoyl, palmitoylated, palmitoylation	171	146
Myristoylation	Myristoyl, myristoylated, myristoylation	109	103
Hydroxylation	Hydroxyl, hydroxylated, hydroxylation	121	105
Amidation	Amid, amided, amidation	380	358
Deamidation	Deamidation	32	24
Nitration	Nitrated, S-Nitrosylation, nitration	40	40
Ubiquitination	Ubiquitin, ubiquitinated, ubiquitination	450	316
Sumoylation	SUMO, sumoylated, Sumoylation	132	102
Sulfation	Sulfo, sulfated, sulfation	72	62
Glycosylphosphatidylinositol (GPI) anchor	GPI, GPI-anchor, GPI-anchoring	219	141

2.5.3 Data Access

To facilitate the use of the dbPTM resource, we developed a website for users to browse and search for content. As depicted in **Figure 2.14**, the database can be queried using the protein name, gene name, Swiss-Prot ID or accession, the input of protein sequences for homology search against Swiss-Prot protein sequence database. Both tabular and graphical visualizations of the experimental and predicted PTM sites are displayed, revealing an overview of the post-translational modification sites, solvent accessibility, protein variations, protein secondary structures and protein functional domains in a protein sequence. A summary table shows the details of all PTM types, and the number of PTM sites categorized by substrate amino acid.

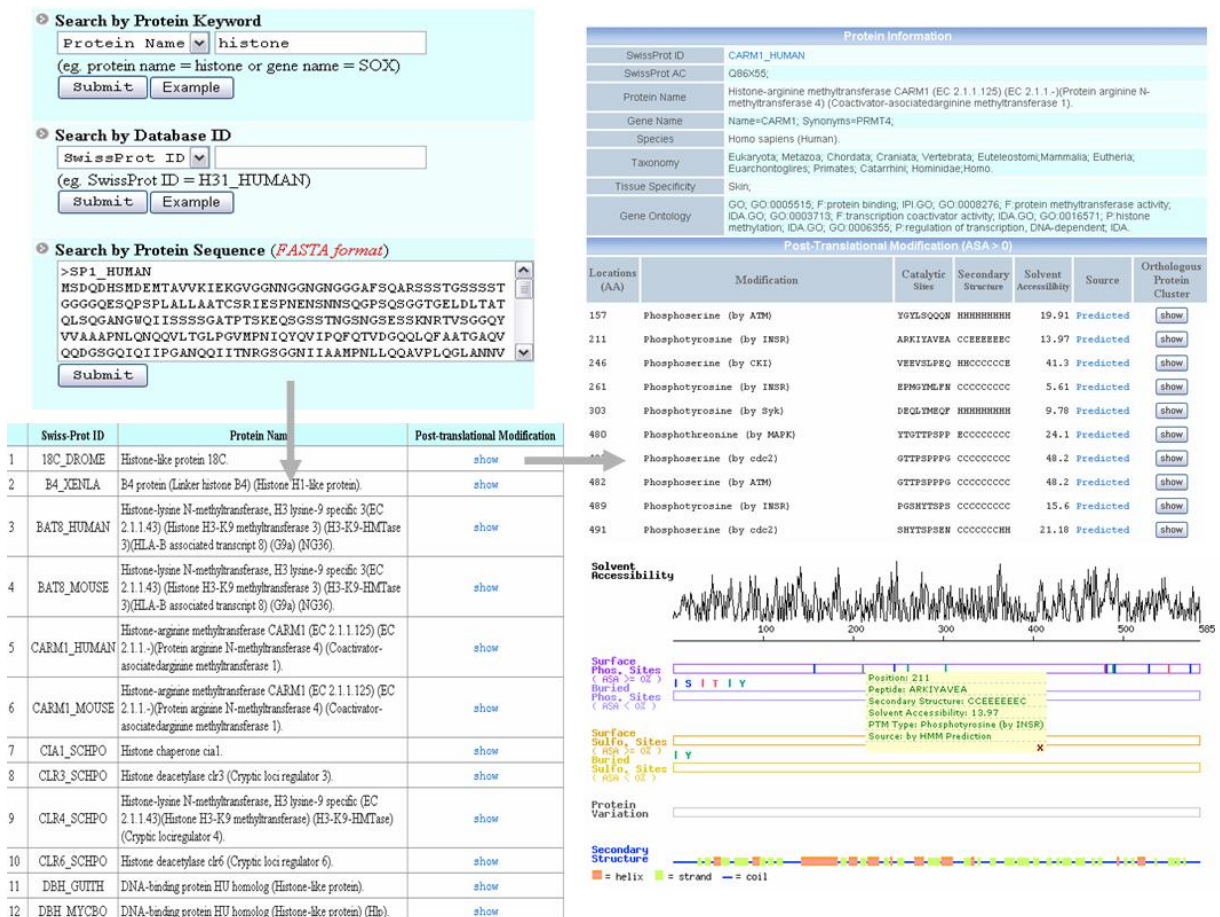


Figure 2.14 Search interface of dbPTM.

Substrate Specificity Investigation

Substrate specificity is the preference of amino acids surrounding the modification sites, which is usually investigated for the identification of particular modification type. To provide the substrate specificity in each type of PTM in detail, the experimentally validated sites in each type of PTM were initially categorized by amino acid types of the substrates. For instance, protein phosphorylation sites can be categorized into subgroups of serine, threonine, tyrosine and histidine. Given a window length n , the fragment of $2n+1$ residues centering on PTM site (position 0) is extracted, and the positional frequencies of amino acids are calculated and presented as sequence logos [60], allowing the sequence entropy to be computed by summing over the height of the letter stacks along the sequence positions. The structural information, such as solvent accessibility and secondary structure surrounding the modified sites, are adopted to calculate the positional solvent accessibility and the matrix of positional secondary structure.

As indicated in Figure 2.15, users can choose the acetylation of lysine (K), for instance,

to obtain more detailed information, including the position of the modified amino acid, the location of the modification in protein sequence, the modified chemical formula and the mass difference. The Jmol program generates the visualization of the formula structure. In particular, the subcellular localization distribution of proteins with acetyllysine was provided to investigate the relationship between them. Furthermore, the sequence logo presents the substrate site specificity, including the composition of amino acid surrounding the modification site [60]. All the experimental PTM sites and putative PTM sites are available and downloadable in the web interface. The PTM benchmark for computational studies is also downloadable.

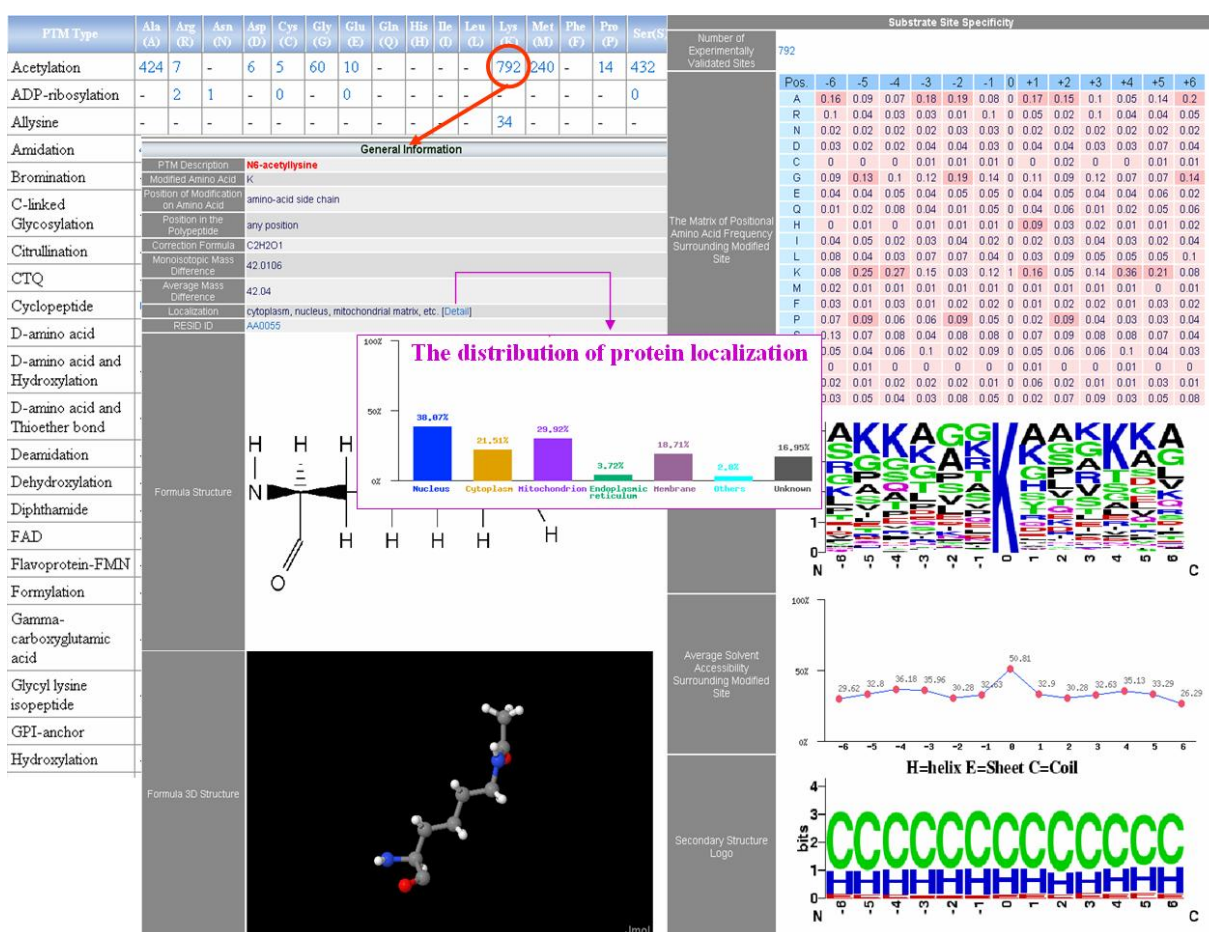


Figure 2.15 Browse interface of dbPTM.

Orthologous Conserved PTM Sites

Moreover, the Clusters of Orthologous Groups of proteins (COGs) [69] was integrated to observe whether a PTM sites located in the conserved regions of protein orthologous sequences. The alignment of the protein sequences in each cluster is provided in the resource.

In **Figure 2.16**, an experimentally verified acetyllysine located in a protein-conserved region indicates an evolutionary influence in which orthologous sites in other species could be involved in the same type of PTM.

Post-Translational Modification (ASA > 0)						
Locations (AA)	Modification	Catalytic Sites	Secondary Structure	Solvent Accessibility	Source	Orthologous Protein Cluster
2	Phosphoarginine (Potential)	---ARTKQT	---CCCCCC	55.07	SwissProt	show
6	Phosphothreonine (by PKC)	RTKQTARKS	CCCCCCCCCC	29.95	Predicted	show
9	N6-methyllysine	QTARKSTGG	CCCCCCCCCC	50.32	SwissProt	show
10	Phosphoserine (by PKG)	TARKSTGGK	CCCCCCCCCC	27.08	Predicted	show
11	Phosphothreonine (by PKC)	ARKSTGGKA	CCCCCCCCCC	53.65	Predicted	show
14	N6-acetyllysine	STGGKAPRK	CCCCCCCCCC	57.47	SwissProt	show
23	N6-acetyllysine	QLATKAARK	CCCCCCCCCC	34.15	SwissProt	show
27	N6-methyllysine	KAARKSAPA	CCCCCCCCCC	46.36	SwissProt	show
36	N6-methyllysine	TGGVKKPHR	CCCCCCCCCC	64.95	SwissProt	show

H31_CANAL	MARTKQTARKSTGGKAPRKQLASKAARKSAPST--GGVKKPHRYKPGTVALREIRRFQKSTELLIRKLPFQRLVREIAQDFK-
H31_DEBHA	MARTKQTARKSTGGKAPRKQLASKAARKSAPST--GGVKKPHRYKPGTVALREIRRFQKSTELLIRKLPFQRLVREIAQDFK-
H31_SCHPO	MARTKQTARKSTGGKAPRKQLASKAARKAAPT--GGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFK-
H31_USTHA	MARTKQTARKSTGGKAPRKQLATKAARKSAPAA--GGVKKPHRYKPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFK-
H31_HUSPA	MARTKQTARKSTGGKAPRKQLATKAARKSAPAT--GGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFK-
H31_RAT	MARTKQTARKSTGGKAPRKQLATKAARKSAPAT--GGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFK-
H31_HUMAN	MARTKQTARKSTGGKAPRKQLATKAAARKSAPAT--GGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFK-
H31_MOUSE	MARTKQTARKSTGGKAPRKQLATKAARKSAPAT--GGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFK-
H31_BOVIN	MARTKQTARKSTGGKAPRKQLATKAARKSAPAT--GGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFK-
H31_CHLRE	MARTKQTARKSTGGKAPRKQLATKAARK-TPAT--GGVKKPHRYRPGTVALREIRKYQKSTELVIRKLPFQRLVREIAQDFK-
H31_TETPY	MARTKQTARKSTGAKAPRKQLASKAARKSAPAT--GGIKKPHRFPGTVALREIRKYQKSTOLLIRKLPFQRLVRDIAHEFK-
H31_TETTH	MARTKQTARKSTGAKAPRKQLASKAARKSAPAT--GGIKKPHRFPGTVALREIRKYQKSTOLLIRKLPFQRLVRDIAHEFK-
H31_STYLE	-----NTGAKAPREQLANKAARKTAQVAQSGGVKKPHRFPGTVALREIRKFQKSTELLIRKLPFQRLVREIAQEQYK-
H31_ENCCU	MARTKQSARKTTGGKAPRKQLSAKSARKGVSPASSAGAKK--SRYRPGSVLKEIRRYQKSTDFLIRLPPQACRSVVKESCN

Figure 2.16 Example of PTM site located in orthologous conserved region.

2.5.4 Characteristics

The proposed server enables both wet-lab biologists and bioinformatics researchers to easily explore the information about protein post-translational modifications. dbPTM not only accumulates the experimentally verified PTM sites with relevant literature references, but also computationally annotates twenty types of PTM sites on Swiss-Prot proteins without any previously annotated PTM sites. As indicated in **Table 2.12**, the proposed knowledge base provides effective information relating to each type of PTM, including orthologous conserved regions, relationship between PTMs and subcellular localization, and the substrate specificity such as the frequency of amino acids, the average solvent accessibility and the frequency of secondary structure surrounding the modified site. Moreover, the proposed PTM benchmark

can be adopted to compare the predictive performance of various tools involved in the same type of PTM prediction, based on the same testing set.

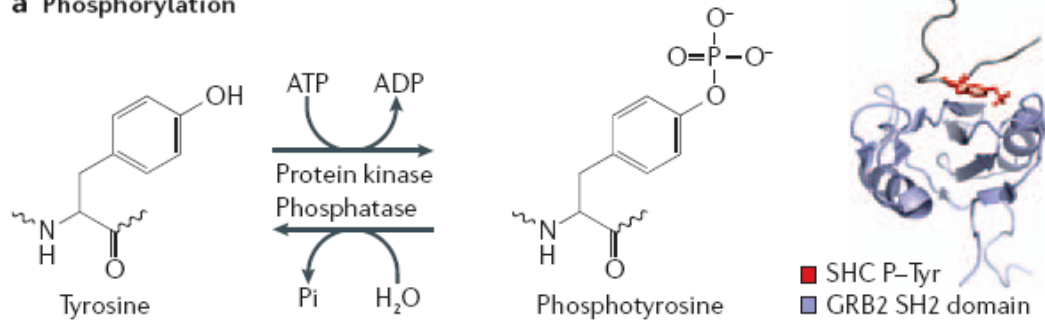
Table 2.12 Advances and improvements in current dbPTM.

Features	dbPTM [19]	dbPTM update
Protein entry	Swiss-Prot (release 46)	Swiss-Prot (release 55)
Experimental PTM resource	Swiss-Prot, Phospho.ELM, and O-GLYCBASE	Swiss-Prot, Phospho.ELM, O-GLYCBASE, and UbiProt
Computationally predicted PTMs	Phosphorylation, glycosylation, and sulfation	About 25 types of PTM (phosphorylation, glycosylation, sulfation, acetylation, methylation, sumoylation, hydroxylation, etc.)
Experimental structural properties	Protein Data Bank (PDB)	Protein Data Bank (PDB)
Computational structural properties	RVP-net and PSIPRED	RVP-net and PSIPRED
PTM annotation	RESID (373 PTM annotations)	RESID (431 PTM annotations)
Protein domain	InterPro	InterPro
Protein variation	Swiss-Prot and Ensembl	Swiss-Prot and Ensembl
PTM literature	none	Swiss-Prot, Phospho.ELM, O-GLYCBASE, and UbiProt
Substrate specificity	none	Amino acid frequency, solvent accessibility, and secondary structure surrounding modified sites
Protein clusters	none	COG and ClustalW
PTM Benchmark	none	Providing the benchmark of PTM test set to comparing the predictive performance base on the same dataset
Relationship between PTM and subcellular localization	none	Analyzing the relationship between PTM and subcellular localization
Graphical visualization	PTM, solvent accessibility, secondary structure, protein variation, protein domain, and tertiary structure	PTM, solvent accessibility, secondary structure, protein variation, protein domain, tertiary structure, orthologous conserved regions and substrate site specificity

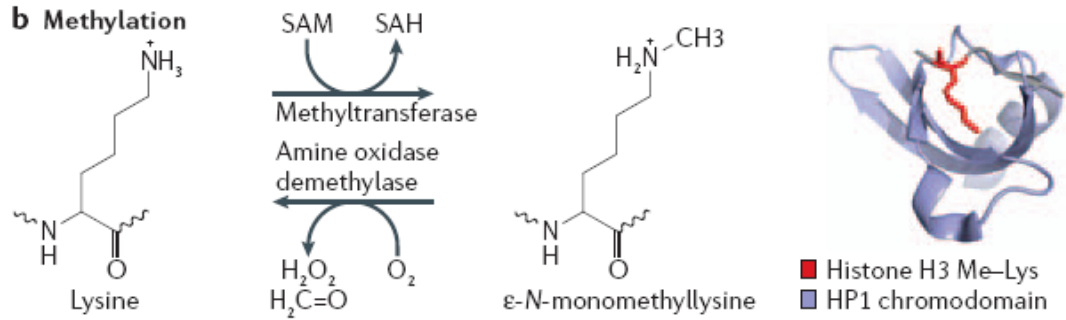
2.6 Summary

The proposed resource, dbPTM, not only integrates the experimentally validated post-translational modification information, but it also computationally annotates the Swiss-Prot proteins for putative phosphorylation, glycosylation and sulfation sites. Furthermore, the PTM related protein structural properties and functional information, such as solvent accessibility of amino acid residues, protein variations, protein secondary structures, protein tertiary structures and protein domains, are provided to facilitate the research of protein post-translational modifications. dbPTM also provides comprehensive and effective PTM information about substrate specificity and their roles in biological systems. The PTM benchmark has much potential to become a performance evaluation standard for computational studies of protein post-translational modification. Previous investigations have indicated that many protein modifications create binding sites for specific protein-protein interaction domains to regulate cellular behavior [92]. As shown in **Figure 2.17** [92], interaction domains often recognize short peptide motifs that are embedded in target proteins, but do not bind stably until the peptide has acquired an appropriate PTM. Such domains usually have a conserved binding pocket for the modified residue and a more variable surface that selectively engages the flanking amino acids, and thereby distinguishes between different peptide motifs with the same PTM^{6–9}. Both the domains and the peptide motifs that they recognize are modular in design and can therefore, in principle, be incorporated into many different proteins. Future work of dbPTM will combine information about protein-protein interaction domains, such as InterDom [93].

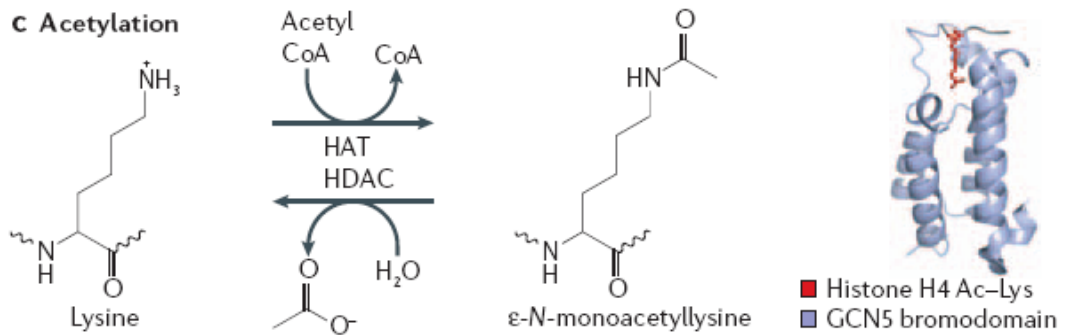
a Phosphorylation



b Methylation



c Acetylation



d Ubiquitylation

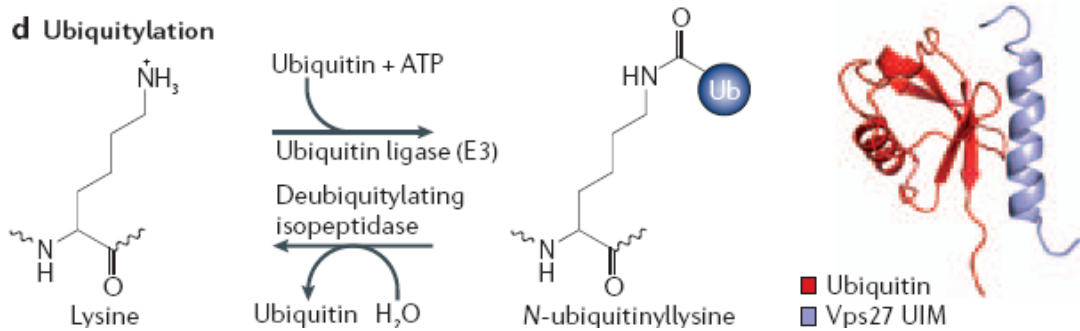


Figure 2.17 Example post-translational modification reactions and structures of protein-interaction-domain–ligand complexes (Seet, B.T., et al., 2006).

Chapter 3 Identification of Kinase-Specific Phosphorylation Sites

3.1 Introduction

Protein phosphorylation, which is an important reversible mechanism in post-translational modifications, is involved in many essential cellular processes including cellular regulation, cellular signal pathways, metabolism, growth, differentiation, and membrane transport [59]. Phosphorylation of substrate sites at serine, threonine, and tyrosine residues of eukaryotic proteins is performed by members of the protein kinase family. Additionally, phosphorylation on histidine plays an important role in signal transduction in prokaryotes known as two-component histidine kinase [94]. It is estimated that one-third of proteins are phosphorylated and around half of kinome are disease- or cancer-related by chromosomal mapping [95].




Protein kinase	Consensus sequence and phosphorylated residue*
Protein kinase A	-X-R-(R/K)-X-(S/T)-B-
Protein kinase G	-X-R-(R/K)-X-(S/T)-X-
Protein kinase C	-(R/K)-(R/K)-X-(S/T)-B-(R/K)-(R/K)-
Protein kinase B	-X-R-X-(S/T)-X-K-
Ca ²⁺ /calmodulin kinase I	-B-X-R-X-X-(S/T)-X-X-X-B-
Ca ²⁺ /calmodulin kinase II	-B-X-(R/K)-X-X-(S/T)-X-X-
Myosin light chain kinase (smooth muscle)	-K-K-R-X-X-S-X-B-B-
Phosphorylase b kinase	-K-R-K-Q-I-S-V-R-
Extracellular signal-regulated kinase (ERK)	-P-X-(S/T)-P-P-
Cyclin-dependent protein kinase (cdc2)	-X-(S/T)-P-X-(K/R)-
Casein kinase I	-(Sp/Tp)-X-X-(X)-(S/T)-B
Casein kinase II	-X-(S/T)-X-X-(E/D/Sp/Yp)-X-
β -Adrenergic receptor kinase	-(D/E) _n -(S/T)-X-X-X-
Rhodopsin kinase	-X-X-(S/T)-(E) _n -
Insulin receptor kinase	-X-E-E-E-Y-M-M-M-M-K-K-S-R-G-D-Y-M-T-M-Q-I-G-K-K-K- L-P-A-T-G-D-Y-M-N-M-S-P-V-G-D-
Epidermal growth factor (EGF) receptor kinase	-E-E-E-E-Y-F-E-L-V-

Figure 3.1 Consensus sequences for protein kinases (Lehninger *et al.*, 2005).

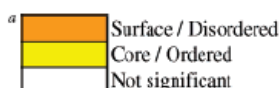
The Ser, Thr, or Tyr residues that are phosphorylated in regulated proteins occur within common structural motifs, called consensus sequences, that are recognized by specific protein kinases (**Figure 3.1**) [6]. Some kinases are basophilic, preferring to phosphorylate a residue having basic neighbors; others have different substrate preferences, such as for a residue near

a Pro residue. Primary sequence is not the only important factor in determining whether a given residue will be phosphorylated, however. Protein folding brings together residues that are distant in the primary sequence; the resulting three-dimensional structure can determine whether a protein kinase has access to a given residue and can recognize it as a substrate. Another factor influencing the substrate specificity of certain protein kinases is the proximity of other phosphorylated residues.

With the recent exponential increase in protein phosphorylation sites identified by mass spectrometry (MS), many researches are undertaken to identify the kinase-specific phosphorylation sites using consensus sequences. Our previous work, KinasePhos 1.0, incorporated profile hidden Markov model (HMM) for identifying kinase-specific phosphorylation sites prediction, whose overall predictive accuracy is about 87% [55, 56]. Recently, version 2.0 of KinasePhos incorporated the protein coupling pattern as a feature for training computer models for identifying phosphorylation sites [91]. In this work, we propose a new method that incorporates support vector machine (SVM) with protein structural information such as surface accessibility, secondary structure and protein disorder region for identifying phosphorylation sites.



PTM Types		Phosphoserine	Phosphothreonine	Phosphotyrosine	N6-acetyllysine	Phosphohistidine	O-GlcNAc (Ser)	O-GlcNAc (Thr)	4-aspartylphosphate
AA		S	T	Y	K	H	S	T	D
No. PTM Data		3255	725	645	77	29	28	23	21
Average No. of PTM per Protein		2.3	1.4	1.5	2.0	1.2	1.6	1.9	1.0
Preferred Secondary Structure ^b		C (84%)	C (82%)	C (58%)	C (68%)	H (78%)	C (71%)	C (82%)	E (96%)
Surface Accessibility	ASA_3	Orange	Orange	Orange	Orange	Yellow	White	White	White
	ASA_5	Orange	Orange	Orange	Orange	Yellow	White	White	White
	ASA_7	Orange	Orange	Orange	Orange	Yellow	White	White	White
	ASA_9	Orange	Orange	Orange	Orange	Yellow	White	White	White
	KD_9	Orange	Orange	Orange	Orange	Yellow	White	White	White
	GOR_17	Orange	Orange	Orange	Orange	Yellow	White	White	White
	RVP-net	Orange	Orange	Orange	Orange	Yellow	White	White	White
Disorder v.s. Order	RONN	Orange	Orange	Orange	Orange	White	White	Orange	Yellow
	DCOILS	Orange	Orange	Orange	Orange	White	White	Orange	Yellow
	REM465	Orange	Orange	Orange	Orange	White	White	Orange	Yellow
	HOTLOOPS	Orange	Orange	Orange	Orange	White	White	Orange	Yellow
Linker or Domain ^c		NHL	NHL	NHL	D	NS	NS	NS	L

^a  Surface / Disordered
Core / Ordered
Not significant

^b C, coils or loops; E, extended (β -sheets); H, helices. ^c L, general linker; HL, helical linker; NHL, non-helical linker; D, domain; NS, not significant.

Figure 3.2 Structural environment of reversible modifications (Pang *et al.*, 2007).

Structural Properties of Phosphorylated Sites

A side chain of amino acid that undergoes enzymatic modification needs to be accessible on the surface of protein [15]. Several works have been proposed the links between the post-translational modifications and their solvent accessible surface area. Pang *et al.* investigated the structural environment of 8378 incidences in 44 types of post-translational modifications [15]. As shown in **Figure 3.2** [15], the structural environment of reversible modifications indicated that protein phosphorylation prefers to occur in regions that are intrinsically disorder and easily accessible. The information of surface accessibility, disorder region, and linker/domain are computationally annotated by several published programs, including ASA [16], GOR [17] and RVP-net [18] for surface accessibility, RONN[19] and DISEMBL [20] for disorder, PSIPRED [21] for secondary structure, and George et al. [22] for linker/domain.

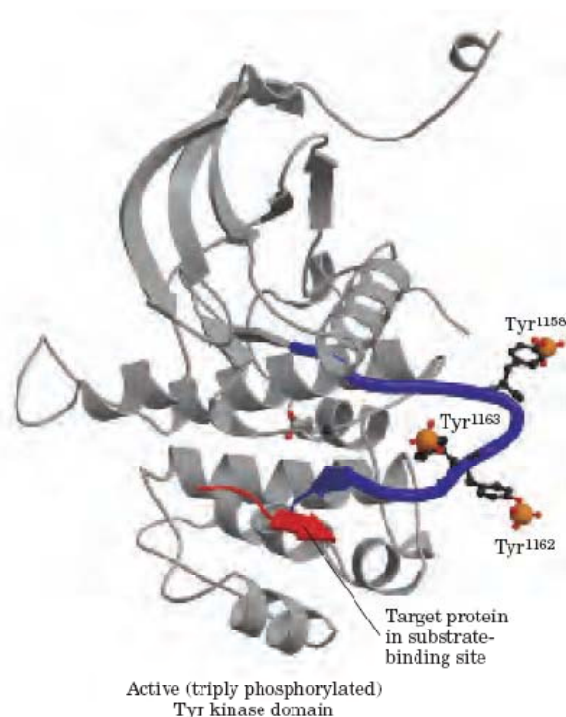


Figure 3.3 Phosphorylated insulin-receptor Tyr kinase (PDB: 1IR3) (Lehninger *et al.*, 2006).

It has been observed that protein phosphorylation prefers to occur in regions that are intrinsically disorder and easily accessible, as an example of phosphorylated insulin-receptor Tyr kinase (PDB: 1IR3) in **Figure 3.3**. In other study, the solvent accessibility has been used to aid the detection of phosphorylation, glycosylation, and tyrosine sulfation sites, whose residues with solvent accessibility above a threshold are identified as surfaced modification sites [19]. Arthur *et al.* incorporated homology modeling of protein tertiary structure and

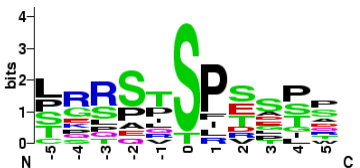
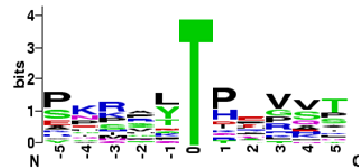
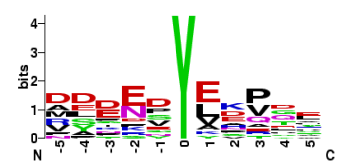
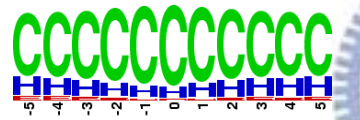
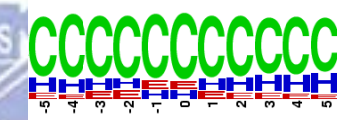

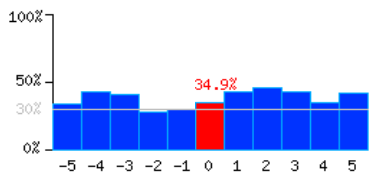
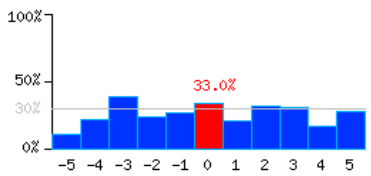
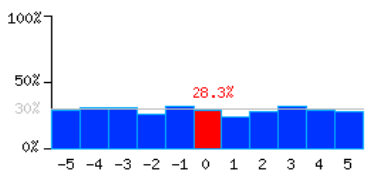
solvent accessibility calculation of predicted structure for identifying phosphorylation sites [96]. As a result, the preference of surface accessibility could provide the useful indication for the prediction of protein methylation.

In order to investigate the preference of solvent accessible surface area (ASA) surrounding phosphorylation sites in tertiary structures, the collected experimental phosphorylation sites should be mapped to the correct position of protein entries in Protein Data Bank (PDB) [67]. The preference of secondary structure surrounding phosphorylation sites is also taken into account. DSSP [68] is a database of secondary structure assignments (and much more) for all protein entries in the Protein Data Bank (PDB). DSSP also provides the program for calculating the solvent accessibility and standardizing secondary structure of PDB entries. **Table 3.1** lists the mapping hits of phosphorylated residues between Swiss-Prot and PDB in detail, which composed of 1078 serines, 404 threonines, and 432 tyrosines. In the case of phosphoserine, there are 77.3% of phosphorylated sites located in coil region, 17.5% sites were observed in helical region, and 5.2% located in sheet. As given in **Table 3.1**, the average percentage of solvent accessible surface area for phosphoserine is 34.9%. Moreover, the flanking positions -4, -3, +1, +2, and +3 have higher surface accessibility.

In the case of phosphothreonine, out of 404 phosphorylated sites covered by the PDB hits, of the 68.6% sites are observed in coil regions, 12.8% sites are observed in helical regions, and 18.6% are in sheet regions. The average percentage of accessible surface area for phosphorylated threonine is 33.0%, which is higher exposed to surface than flanking regions. In phosphotyrosine, out of 432 phosphorylated sites covered by the PDB hits, of the 42.5% sites are observed in coil regions, 27.3% sites are observed in helical regions, and 30.2% are in sheet regions. The average percentage of accessible surface area for phosphorylated threonine is 28.3%, which is slightly exposed to surface.

Although the number of experimental phosphorylated sites which locate in the protein regions with tertiary structure is not enough to be investigated the preferences of solvent accessibility and secondary structure in each kinase group, it seems that protein phosphorylation site prefers to occur on the exposed and coil regions. Even though phosphorylated sites may not always be in surface-accessible regions, surface-accessible amino acids would have a higher likelihood been modified.

Table 3.1 The statistics of structural information in phosphorylated serine, threonine and tyrosine.

Phosphorylated residue	Serine	Threonine	Tyrosine
Number of experimental phosphorylated sites with PDB structure	1,078	404	432
Sequence logo of amino acids surrounding phosphorylated sites			
Distribution of secondary structure on phosphorylated site	17.5% helix, 5.2% sheet, 77.3% coil	12.8% helix, 18.6% sheet, 68.6% coil	27.3% helix, 30.2% sheet, 42.5% coil
Sequence logo of secondary structure surrounding phosphorylated sites			
Average percentage of accessible surface area on phosphorylated site	34.9%	33.0%	28.3%
Average accessible surface area surrounding phosphorylated sites			

3.2 Related Works

In this section, several common machine learning methods which have been frequently used to phosphorylation site prediction are described over here. Following, some representative prediction servers of protein post-translational modifications are listed and briefly introduced.

3.2.1 Machine Learning Methods

Machine learning is programming computers to optimize a performance criterion using example data or past experience. In bioinformatics, machine learning is usually referred to classification which learns predictive model from training data sets for distinguishing between different exemplars based on their differentiating patterns. Several common machine learning algorithms such as *k*-Nearest Neighbor (KNN), decision tree, Bayesian decision theory (BDT), neural network (NN), hidden Markov model (HMM), and support vector machine (SVM) are described as follows.

k-Nearest Neighbor (KNN)

Arguably the simplest method is the *k*-Nearest Neighbor classifier (Cover and Hart, 1967). Here the *k* points of the training data closest to the test point are found, and a label is given to the test point by a majority vote between the *k* points. This method is highly intuitive and attains – given its simplicity – remarkably low classification errors, but it is computationally expensive and requires a large memory to store the training data.

Decision Tree

Another intuitive class of classification algorithms are *decision trees*. As shown in **Figure 3.4**, these algorithms solve the classification problem by repeatedly partitioning the input space, so as to build a tree whose nodes are as pure as possible (that is, they contain points of a single class). Classification of a new test point is achieved by moving from top to bottom along the branches of the tree, starting from the root node, until a terminal node is reached. Decision trees are simple yet effective classification schemes for small datasets. The computational complexity scales unfavorably with the number of dimensions of the data. Large datasets tend to result in complicated trees, which in turn require a large memory for storage. The C4.5

implementation by Quinlan (1992) is frequently used and can be downloaded at <http://www.rulequest.com/Personal>.

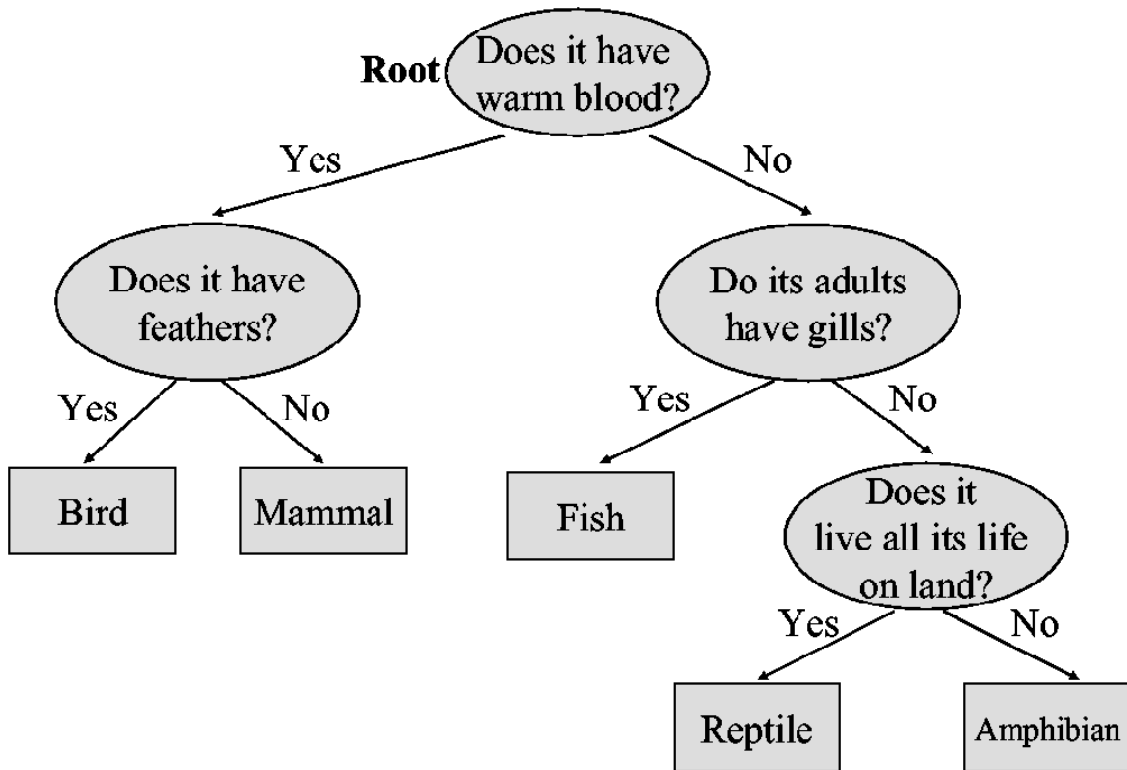


Figure 3.4 An example of decision tree.

Bayesian Decision Theory (BDT)

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. This approach is based on quantifying the tradeoffs between various classification decisions using probability and the costs that accompany such decisions. It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known. Suppose that we have an unclassified data x that belongs to one of two certain categories: C_1 (defined as phosphorylation sites) and C_2 (defined as non-phosphorylation sites). Suppose that we know both the prior probabilities $P(C_j)$ and the conditional densities $p(x|C_j)$. In addition, the posterior probability of x for these two categories can be denoted as: $p(C_1|x)$ and $p(C_2|x)$, which are called *Bayes' formula*:

$$P(C_j | x) = \frac{p(x | C_j)P(C_j)}{p(x)},$$

where in this case of two categories

$$p(x) = \sum_{j=1}^2 p(x | C_j)P(C_j).$$

Then the probability of wrong prediction is:

$$P(\text{error} | x) = \begin{cases} p(C_1 | x), & x \in C_2 \\ p(C_2 | x), & x \in C_1 \end{cases}$$

To minimize the expectation of error probability that is defined as [97]:

$$P(\text{error}) = \int P(\text{error} | x)p(x)dx$$

It is obvious that one should choose the more probable category as the prediction result, which can be formulated by the Bayesian Decision Rule:

$$\text{predict } x \text{ as } \begin{cases} C_1, & \text{if } P(C_1 | x) > P(C_2 | x) \\ C_2, & \text{otherwise} \end{cases}$$

Furthermore, by definition the loss function $\lambda(\alpha_i | C_j)$, where $\alpha_i, i = 1,2$ is the finite set of possible solution. Thus, the expected loss (risk) of taking action α_i is:

$$R(\alpha_i | x) = \sum_{l=1}^2 \lambda(\alpha_i | C_l)P(C_l | x)$$

In this condition, the goal of optimization becomes to minimize the overall risk for every x . Similar to the rationale of Bayesian Decision Rule, we can obtain the best performance by computing $R(\alpha_i | x)$ for each solution α_i and choose that for which has the minimal overall risk [97].

Neural Network (NN)

Neural network (NN) is one of the most commonly used approaches to classification. Artificial neural network (ANN) is a computational model inspired by the connectivity of neurons in animate nervous systems [98]. A simple scheme for ANN is shown in **Figure 3.5** [98]. Each circle denotes a computational element referred to as a *neuron*, which computes a weighted sum of its inputs, and possibly performs a nonlinear function on this sum. If certain classes of nonlinear functions are used, the function computed by the network can approximate any function (specifically a mapping from the training patterns to the training targets), provided enough neurons exist in the network and enough training examples are

provided.

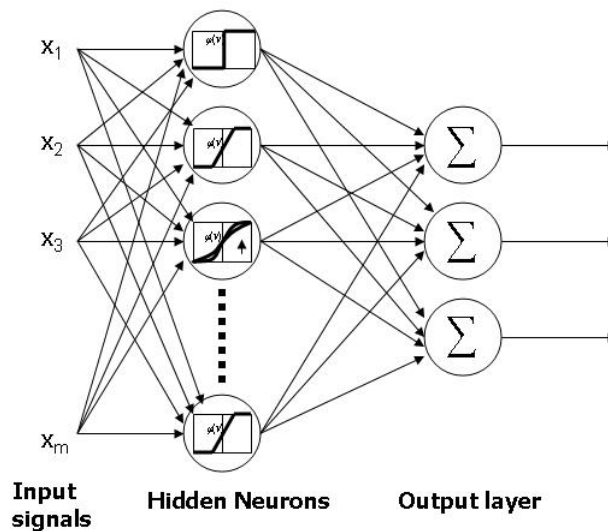


Figure 3.5 A schematic diagram of artificial neural network. Each circle in the hidden and output layer is a computation element known as a neuron (Haykin *et al.*, 1999).

ANN is capable of classifying highly complex and nonlinear biological sequence patterns, where correlations between positions are important. Not only does the network recognize the patterns seen during training, but it also retains the ability to generalize and recognize similar, though not identical patterns. Artificial neural network algorithms have been extensively used in biological sequence analysis. An artificial neural network library ANNLIB [99], which were implemented in C program language, is available.

Hidden Markov Model (HMM)

A hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications. A HMM can be considered as the simplest dynamic Bayesian network. The key idea is that an HMM is a finite model that describes a probability distribution over an infinite number of possible sequences. The HMM is composed of some number of *states*, which might correspond to positions in a three-dimensional structure or columns of a multiple alignment. Each state “emits” symbols (residues) according to *symbol emission probabilities*, and the states are interconnected by *state transition probabilities*. Starting from *initial state* and a sequence of states is generated by moving from state to state according to the state transition

probabilities until an *end state* is reached. Each state then emits symbols according to that state's emission probability distribution, creating an observable sequence of symbols.

The *state path* is a Markov chain, meaning that what state we go to next depends only on what state we're in. Since we're only given the observed sequence, this underlying state path is hidden - these are residue labels that we'd like to infer. The state path is a *hidden Markov chain*, whose probability $P(S, \pi | HMM, \theta)$ that an HMM with parameters θ generates a state path π and an observed sequence S is the product of all the emission probabilities and transition probabilities that were used. Why are they called hidden Markov models? The sequence of states is a Markov chain, because the choice of the next state to occupy is dependent on the identity of the current state. However, this state sequence is not observed; it is hidden. Only the symbol sequence that these hidden states generate is observed. The most likely state sequence must be inferred from an alignment of the HMM to the observed sequence.

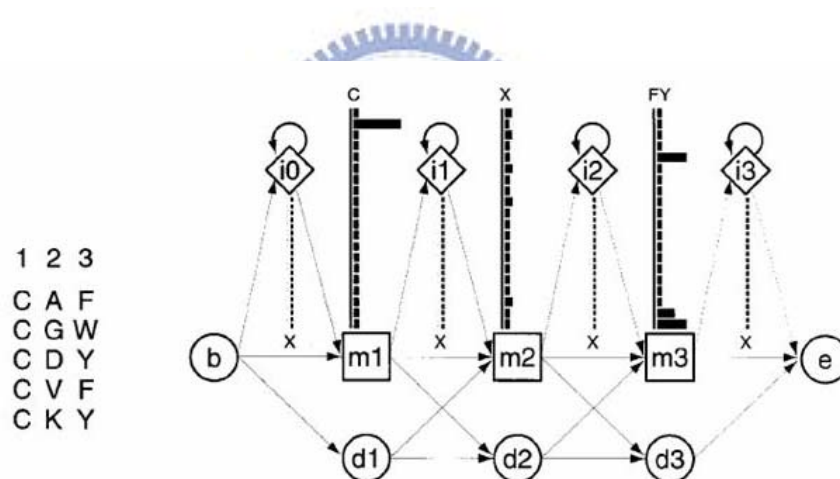


Figure 3.6 An example of small profile HMM representing a short multiple alignment of five sequences with three consensus columns (Eddy *et al.*, 1998).

Hidden Markov models now provide a coherent theory for profile methods, namely Profile hidden Markov models (profile HMMs) [63], which are statistical models (maximum likelihood) of multiple sequence alignments. They capture position-specific information about how conserved each column of the alignment is, and which residues are likely. An example of small profile HMM is shown in **Figure 3.6** [63]. The three columns are modeled by three match states (squares labeled m1, m2, and m3), each of which has 20 residue emission probabilities, shown with black bars. Insert states (diamonds labeled i0 - i3) also have 20 emission probabilities each. Delete states (circles labeled d1-d3) are 'mute' states that have no

emission probabilities. A begin and end state are included (b,e). State transition probabilities are shown as arrows.

Support Vector Machine (SVM)

Support vector machine (SVM) [100] is a useful technique for data classification. A classification task usually involves with training and testing data which consist of some data instances. Each instance in the training set contains one “target value” (class label) and several “attributes” (features). The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes. The basic concept of SVM is to transform the samples into a high dimensional space and find a separating hyperplane with the maximal margin between two classes in the space (**Figure 3.7**).

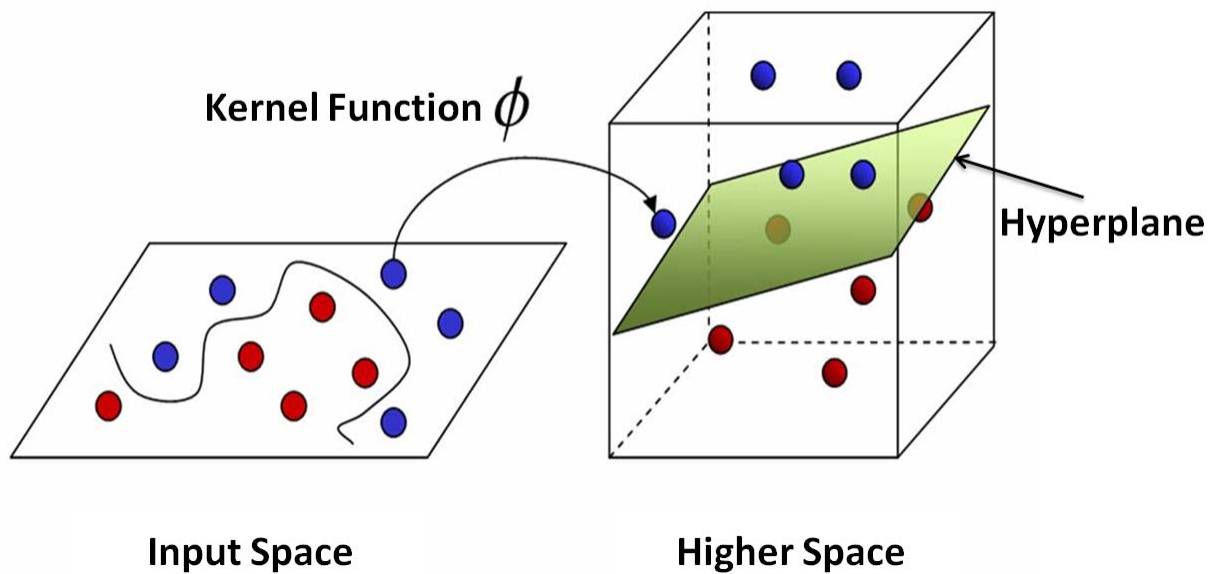


Figure 3.7 Basic concept of support vector machine.⁷

Basically, SVM is a binary classifier. Given training vectors $x_i, i = 1, \dots, l$ and a vector y defined as: $y_i = 1$ if x_i is in class I, and $y_i = -1$ if x_i is in the class II. The support vector technique tries to find the separating hyperplane $w^T x_i + b = 0$ with the largest distance between two classes, measured along a line perpendicular to this hyperplane, which require the solution of following optimization problems (**Figure 3.8**):

⁷ The figure was obtained from <http://www.imtech.res.in/raghava/rbpred/svm.jpg>

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^l \xi_i \quad \text{subject to} \quad y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

Here training vectors x_i are mapped into a higher dimensional space by the function ϕ . Constraints $y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i$ allow that training data may not be on the correct side of the separating hyperplane $w^T x_i + b = 0$. Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. C is the penalty parameter of the error term to be optimized. Furthermore, $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function.

Four basic kernel functions are listed as follows:

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
- Radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$
- Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$

Here, γ , r , and d are kernel parameters. Most commonly used kernel functions are RBF kernel.

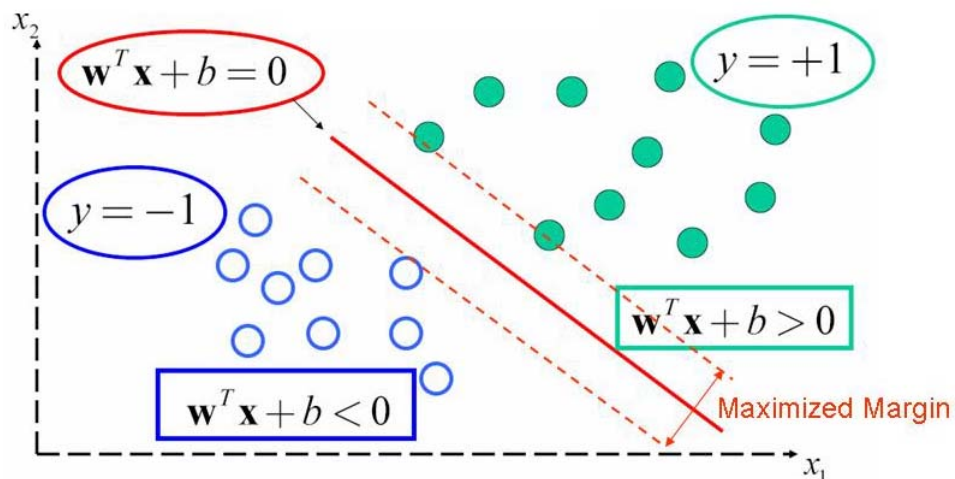


Figure 3.8 Principle of hyperplane in support vector machine. ⁸

Recently, SVM has been successfully applied in solving many biological problems, such as predicting protein subcellular localization [101], protein secondary structures [102], tumor

⁸ The figure was obtained from <http://www.imtech.res.in/raghava/rbpred/>

classification [103] and phosphorylation sites [78], which shown to be an effective machine learning method. A public SVM library, namely LIBSVM [104], was available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.

Boosting

The basic idea of boosting and ensemble learning algorithms in general is to iteratively combine relatively simple *base hypotheses* – sometimes called *rules of thumb* – for the final prediction. One uses a so-called *base learner* that generates the base hypotheses. In boosting the base hypotheses are linearly combined. In the case of two-class classification, the final prediction is the weighted majority of the votes. The combination of these simple rules can boost the performance drastically. It has been shown that Boosting has strong ties to support vector machines and large margin classification (Rˆatsch, 2001, Meir and Rˆatsch, 2003). Boosting techniques have been used on very high dimensional data sets and can quite easily deal with than hundred thousands of examples. Research papers and implementations can be downloaded from <http://www.boosting.org>.

3.2.2 Phosphorylation Site Prediction

With the recent exponential increase in protein phosphorylation sites identified by mass spectrometry (MS), many researches are undertaken to identify the kinase-specific phosphorylation sites. The summary of tool name, reference, material, method, number of kinase group, and predictive performance of the previously developed phosphorylation site prediction tools is shown in **Table 3.2**. Our previous work, KinasePhos 1.0, incorporated profile hidden Markov model (HMM) for identifying kinase-specific phosphorylation sites prediction, whose overall predictive accuracy is about 87% [55, 56]. Version 2.0 of KinasePhos incorporated support vector machine (SVM with)the protein coupling pattern for identifying phosphorylation sites [91]. NetPhos [57] developed neural networks to predict phosphorylation sites on serine, threonine and tyrosine residues; however, it cannot provide information on the kinases involved and NetPhosK [77] applied an artificial neural network algorithm to predict 17 PK groups-specific phosphorylation sites. DISPHOS [58] took advantage of the position-specific amino acid frequencies and disorder information to improve the discrimination between phosphorylation sites and non-phosphorylation sites. Scansite 2.0 [105] identified short protein sequence motifs that are recognized by modular

signaling domains, phosphorylated by protein serine/threonine, tyrosine kinases or mediate specific interactions with protein or phospholipid ligands. PredPhospho [78] predicts phosphorylation sites limited to four protein major kinase families, such as CDK, CK2, PKA and PKC, and four protein kinase groups (AGC, CAMK, CMGC and TK) with predictive accuracy 83-95% and 76-91%, respective. GPS [75, 106], is a group-based phosphorylation site predicting and scoring platform which clustered the 216 unique protein kinases in 71 groups. PPSP [76] developed an approach based on Bayesian decision theory for predicting the potential phosphorylation sites accurately for around 70 protein kinase groups. PHOSIDA [45], incorporated support vector machine with surface accessibility and evolutionary conservation, made 91.75%, 81.06%, and 76.19% accuracies in serine, threonine, and tyrosine, respectively. Recently, a proficient meta-predictor [107] adopted weighted voting strategy to organize and process the predictions produced by several other predictors, including GPS, KinasePhos, NetPhosK, PPSP, PredPhospho and Scansite.



Table 3.2 List of the previously developed phosphorylation site prediction tools.

Tool	Reference	Material	Feature	Method	Kinase group	Proposed predictive performance			
						Overall	PKA	PKC	CK2
NetPhos	Blom et al., 1999	PhosphoBase	sequence	ANN	-	Sn=69%~96%	-	-	-
Scansite	Obenauer et al., 2003	Swiss-Prot+TrEMBL+Genpept+Ensembl	sequence	PSSM (motif-based service)	-				
DISPHOS	Lakoucheva et al., 2004	Swiss-Prot+PhosphoBase	predicted protein disordered region and secondary structure	Logistic regression models	-	Serine Ac=76% Threonine Ac=81% Tyrosine Ac=83%	-	-	-
rBPNN	Berry et al., 2004	PhosphoBase	sequence	BPNN, decision tree, rBBFNN	-	BPNN: Ac=89.65±1.64, rBBFNN:Ac=87.77±1.05 C4.5: Ac=90.43±2.03	-	-	-
AutoMotif	Plewczynski et al., 2005	Swiss-Prot (12 types of PTM)	sequence	SVM	-	Precision > 70% (12 types of PTM)	Sn=41% Pre=75%	Sn=17% Pre=83%	Sn=11% Pre=53%
PredPhospho	Jong Hun Kim et al., 2004	Swiss-Prot+PhosphoBase	sequence	SVM	4	Ac = 76 - 91%	Ac=89.98% Sn=88.32% Sp=91.11%	Ac=82.9% Sn=78.71% Sp=85.79%	Ac=91.47% Sn=83.9% Sp=96.43%
NetPhosK	Blom et al., 2004	Swiss-Prot, PhosphoBase, PhosphoSite	sequence	ANN	17	Sn = 84% Sp = 76%	-	-	-
GPS	Feng-Feng Zhou et al., 2004	PhosphoBase, Phospho.ELM	sequence	Clustering or Segmentation	71	Sn = 94.44% Sp = 97.14%	-	-	-
KinasePhos	Huang et al., 2005	PhosphoBase, Swiss-Prot	sequence	MDD + HMM	18	Serine Ac = 86% Threonine Ac = 91% Tyrosine Ac = 84%	Sn = 0.91 Sp = 0.86	Sn = 0.80 Sp = 0.87	Sn = 0.87 Sp = 0.85
Li et al.	Li et al., 2005	PhosphoBase	sequence	kNN measured by Manhattan distance	-	-	Sn=~87.36% Sp=~99.07%	-	Sn=~67.88% Sp=~99.16%

PPSP	Yu Xue et al., 2006 March	Phospho.ELM	sequence	BDT	68	Na	Sn=88.88% Sp=90.57%	Na	Sn=82.99% Sp=87.59%
pkaPS	Neuberger et al., 2007 January	UniProt+Phospho.ELM	sequence	simplified kinase-substrate binding model		Na	Sn=~96% Sp=~94%	Na	Na
KinasePhos 2.0	Wong, Lee et al., 2007	Swiss-Prot+Phospho.ELM	Sequence + coupling pattern	SVM	58	Serine Ac = 90% Threonine Ac = 93% Tyrosine Ac = 88%	Sn = 0.92 Sp = 0.89	Sn = 0.84 Sp = 0.86	Sn = 0.87 Sp = 0.86
GANNPhos	Tang et al., 2007	Phospho.ELM	sequence	GA+NN		S: Ac=81.3~81.8%, Sn=80.5~80.9%, Sp=82.7~83.5% T: Ac=77.5~81.2% , Sn=74.3~77.6%, Sp=83.1~86.4% Y: Ac=74~80.2% , Sn=72.5~76.6%, Sp=77.3~85.6%	Na	Na	Na
AutoMotif 2.0	Plewczynski et al., 2007	UniProt(06.2007)+ Swiss-Prot	sequence	SVM	-	Precision > 90%	Sn=14% Precision=86%	Sn=5% Precision=100%	Sn=6% Precision=80%
PHOSIDA	Gnad, Ren et al. 2007	PHOSIDA	Sequence+ASA+ evolutionary conservation	SVM	-	Serine Ac = 91.75% Threonine Ac = 81% Tyrosine Ac = 76.2%	-	-	-
MetaPredPS	Ji wan et al., 2008	Swiss-Prot+PhosphoSite+ Phospho.ELM	-	voting from GPS, KinasePho, NetPhosK, PPSP, PredPhospho, Scansite	-	-	Sn=88.3% Sp=82.8% Ac=85%	Sn=77.3% Sp=79.1% Ac=78.4%	Sn=87.8% Sp=90.4% Ac=89.3%

Abbreviation: ANN, artificial neural networks; BPNN, back propagation neural network; PSSM, position-specific scoring matrix; SVM, support vector machine; MDD, maximal dependency decomposition; HMM, hidden Markov model; KNN, k-Nearest Neighbor; BDT, Bayesian decision theory; GA, genetic algorithm; ASA, accessible surface area; Ac, accuracy; Sn, sensitivity; Sp, specificity; Pre, precision.

KinasePhos 1.0

The known phosphorylation sites from public domain data sources are categorized by their annotated protein kinases. Based on the concepts of profile Hidden Markov Model (HMM), computational models are learned from the kinase-specific groups of the phosphorylation sites. The Maximal Dependence Decomposition (MDD) [62], employs statistical χ^2 -test to group an set of aligned signal sequences to moderate a large group into subgroups that capture the most significant dependencies between positions, was adopted to group the phosphorylation site sequences of each kinase group with data size more than 50. Based on k -fold cross-validation and Jackknife cross-validation, the average predictive accuracy of phosphorylated serine, threonine, and tyrosine are 86%, 91%, and 84%, respectively. After evaluating the learned models, we select the model with highest accuracy in each kinase-specific group and provide a web-based prediction tool for identifying protein phosphorylation sites. The main contribution here is that we develop a kinase-specific phosphorylation site prediction tool with both high sensitivity and specificity. The proposed web server is freely available at <http://KinasePhos.mbc.nctu.edu.tw/>.

KinasePhos

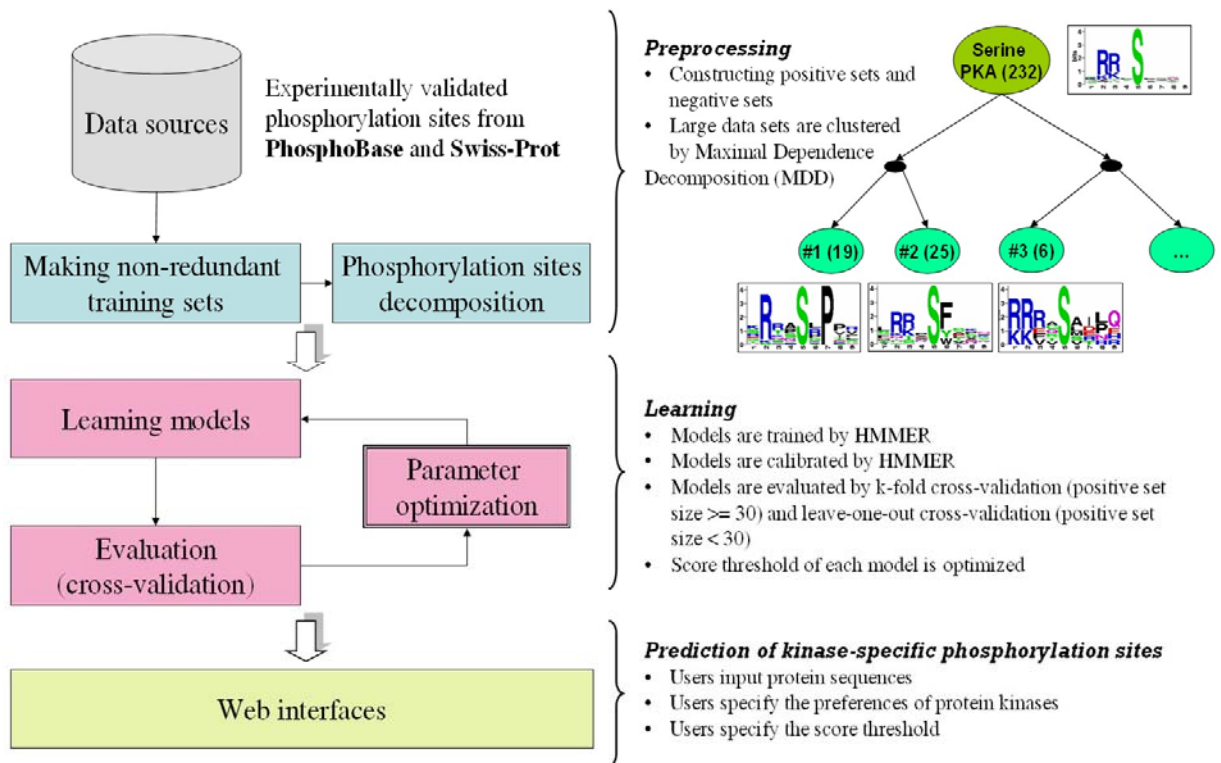


Figure 3.9 The system flow of KinasePhos 1.0.

KinasePhos 2.0

This work proposed a kinase-specific phosphorylation site prediction server which incorporates support vector machines (SVM) with two features, i.e., protein sequence profiles surrounding the modified sites and coupling patterns surrounding the modified sites [91]. The coupling pattern of proteins, which is firstly used for analyzing the protein thermostability [108]. Protein coupling pattern is a novel feature used for identifying phosphorylation sites. The coupling pattern $[XdZ]$ denotes the amino acid coupling-pattern of amino acid types X and Z that are separated by d amino acids. The differences or quotients of coupling strength C_{XdZ} between the positive set of phosphorylation sites and the background set of whole protein sequences from Swiss-Prot are computed to determine the number of coupling patterns for training SVM models. After the evaluation based on k -fold cross-validation and Jackknife cross-validation, the average predictive accuracy of phosphorylated serine, threonine, tyrosine and histidine are 90%, 93%, 88% and 93%, respectively. KinasePhos 2.0 performs better than other tools previously developed. The proposed web server is freely available at <http://KinasePhos2.mbc.nctu.edu.tw/>.

KinasePhos 2.0

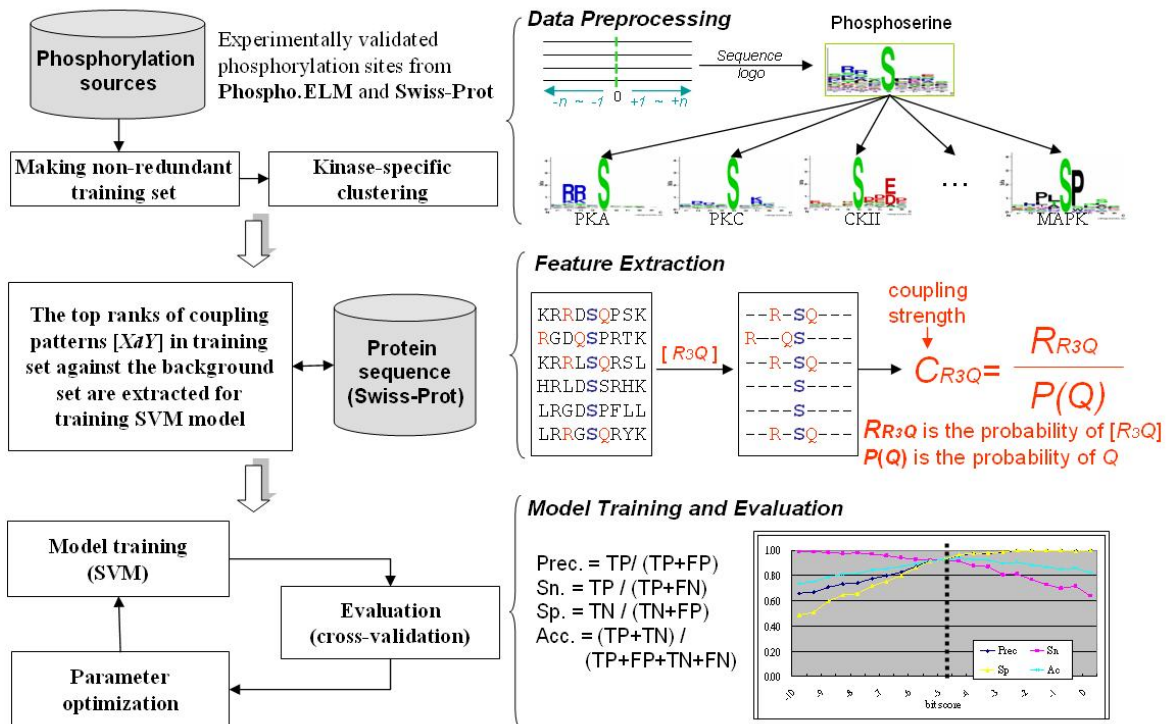


Figure 3.10 The system flow of KinasePhos 2.0.

3.3 Motivation and The Specific Aim

Protein phosphorylation is a ubiquitous and important post-translational modification, responsible for modulating protein function, stability, localization, and cellular signaling network. Experimental identifications of kinase-specific phosphorylation sites on substrates *in vivo* and *in vitro* are the foundation of understanding the mechanisms of phosphorylation dynamics and important for the biomedical drug design. However, these experiments are often time-consuming, labor-intensive, and expensive. Thus, *in silico* prediction of phosphorylation sites with high predictive performance could be a promising strategy to conduct preliminary analyses and could heavily reduce the number of potential targets that need further *in vivo* or *in vitro* confirmation.

We propose a method, namely KinasePhos, which incorporates support vector machine (SVM) to construct the computational models for identifying the kinase-specific phosphorylation sites. It has been observed that protein phosphorylation prefers to occur in regions that are intrinsically disorder and easily accessible. Not only protein amino acids, but also the structural information such as secondary structure, solvent accessibility and protein disorder region were used for analysis. The constructed models were evaluated based on k-fold cross-validation. Moreover, the independent test set, which was constructed based on the proposed benchmark, was used to evaluate whether the constructed model over-fitted the training set. With the highly predictive performance of kinase-specific phosphorylation sites, a better understanding of relationships between protein kinases and substrates will be facilitated and engineered to analyze the therapeutic usefulness.

3.4 Materials and Methods

Figure 3.11 depicts the system flow of the proposed method, which consists of four major analyzing processes such as data preprocessing, feature extraction and coding, model training and evaluation, and independent test. The detailed descriptions are illustrated in following subsection.

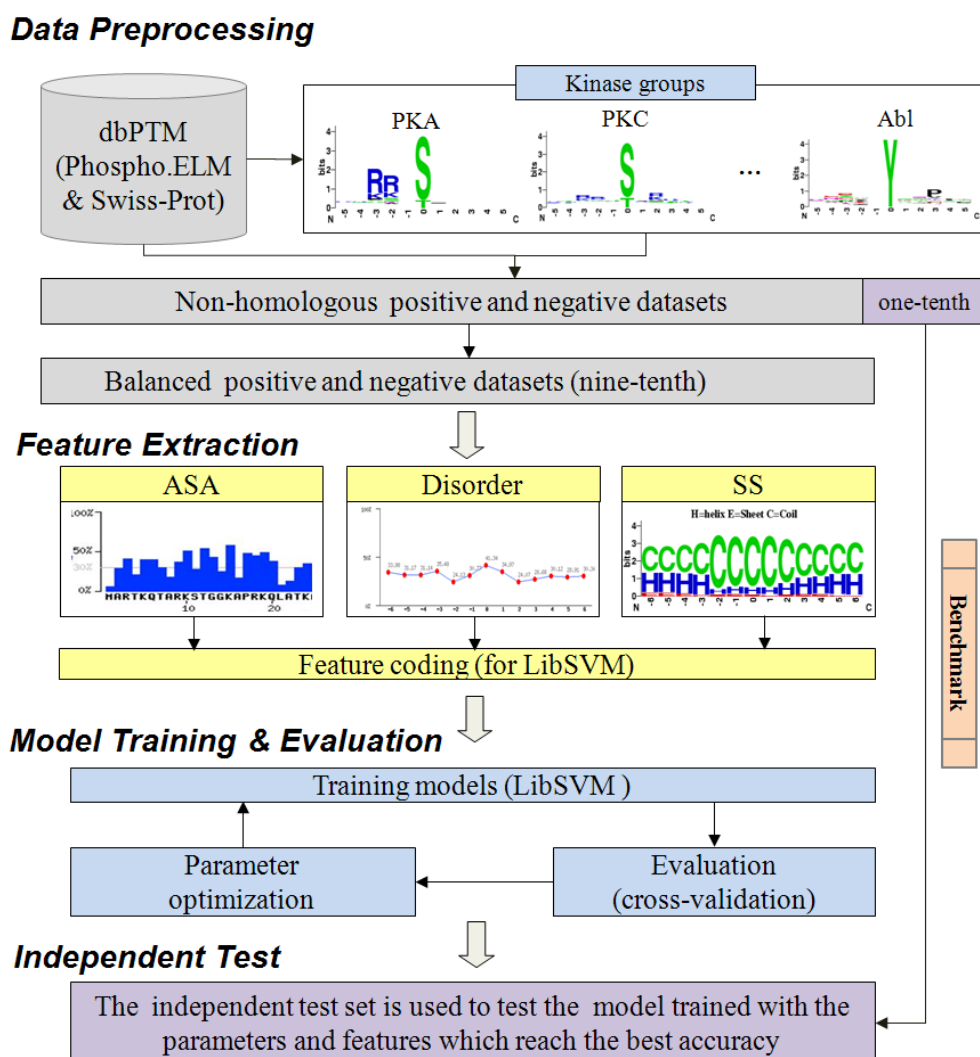


Figure 3.11 The system flow of kinase-specific phosphorylation site prediction.

3.4.1 Data Preprocessing

The experimentally validated phosphorylation sites are extracted from version 7.0 of Phospho.ELM [49] and release 55.0 of UniProtKB/Swiss-Prot [48], containing totally 16525 experimental phosphorylation sites within 5484 proteins and 24328 experimental

phosphorylation sites within 8606 proteins, respectively. After removing the redundant data between Phospho.ELM and Swiss-Prot, the number of serine (S), threonine (T), and tyrosine (Y) phosphorylated sites are 22640, 4982, and 3175, respectively, as given in **Table 3.3**. It notices that the sum of serine, threonine, and tyrosine is not equal to total number of phosphorylation sites because there are several phosphorylation sites located on other kinds of residue.

Table 3.3 The statistics of phosphorylation sites obtained from Phospho.ELM and Swiss-Prot.

Data source	Version	Number of phosphorylated proteins	Number of phosphorylation sites			
			Serine (S)	Threonine (T)	Tyrosine (Y)	Total
Phospho.ELM	7.0	5,484	12,082	2,361	2,081	16,525
Swiss-Prot *	55.0	8,606	18,320	3,982	2,003	24,328
Combined (non-redundant)		9,966	22,640	4,982	3,175	30,818

*The entries which contain residues annotated as “phosphoserine,” “phosphothreonine,” and “phosphotyrosine” in the “MOD_RES” field are extracted and the entries annotated as “by similarity,” “potential,” and “probable” are excluded.

The collected experimental phosphorylation sites are further categorized according to the annotations of catalytic kinases.

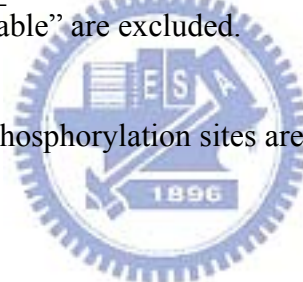


Table 3.4 shows the number of phosphorylated serine, threonine, and tyrosine in each kinase-specific group. The kinase-specific groups whose data size is more than ten are selected to construct the computational models. Otherwise, the kinase groups whose data size is smaller than ten are used to construct the positional weighted matrix for scan the phosphorylation sites.



Table 3.4 Statistics of non-redundant kinase-specific phosphorylation sites in Swiss-Prot and Phospho.ELM.

Kinase family	Number of phosphoprotein	Number of substrate site			
		Total	Serine	Threonine	Tyrosine
Protein kinase A(PKA)	286	458	401	56	1
Protein kinase C(PKC)	274	485	403	80	2
Casein kinase 2(CK2)	192	368	305	61	2
Mitogen-activated protein kinases(MAPK)	192	333	230	103	0
Cell Division Control 2 (CDC2)	138	328	167	161	0
Cyclin dependent kinase (CDK)	82	153	91	62	0
CaM kinase 2 (CAMK2)	77	119	91	28	0
protein Kinase B(PKB)	79	114	91	23	0
Ataxia telangiectasia mutated(ATM)	63	102	94	8	0
Casein Kinase 1(CK1)	52	101	73	27	1
Glycogen synthase 3 kinase (GSK)	50	87	62	35	0
G protein-coupled receptor kinase(GRK)	27	85	59	26	0
p21-activated kinase (PAK)	37	59	51	8	0
Aurora kinase (Aurora)	28	57	46	11	0
Phosphoinositide-dependent protein kinase(PDK)	28	54	23	31	0
Ribosomal S6 Kinase(RSK)	38	51	49	2	0
Polo-like kinase (PLK)	21	50	36	14	0
I kappa B kinase(IKK)	16	49	48	1	0
cGMP-dependent protein kinas (PKG)	26	47	37	10	0
Rho-associated protein kinase (ROCK)	25	41	14	26	1
AMP-activated protein kinase(AMPK)	28	38	35	3	0
MAP kinase-activated protein kinase (MAPKAPK)	20	36	32	4	0
Beta Adrenergic Receptor Kinase (BARK)	17	35	19	16	0
DNA dependent protein kinase(DNA-PK)	19	35	19	16	0
Mammalian STE20-like kinase (MST)	32	34	15	19	0
Checkpoint kinase 2 (CHK2)	11	33	24	9	0
Checkpoint kinase 1 (CHK1)	7	16	13	3	0
MAP kinase kinase (MAP2K)	14	30	7	14	9
CaM kinase 1 (CAMK1)	22	26	22	4	0
Death-associated protein kinase (DAPK)	15	24	11	13	0
Serum and Glucocorticoid Responsive Kinase (SGK)	13	24	18	6	0
MAP kinase kinase kinase (MAP3K)	14	22	6	16	0
Phosphorylase kinase(PHK)	11	20	18	2	0
LKB1 kinase (LKB)	18	20	1	19	0
Serine/threonine-protein kinase IPL1 (IPL1)	14	19	16	3	0
Interferon-induced, double-stranded RNA-activated protein kinase (PKR)	6	17	7	10	0
FKBP12-rapamycin-associated protein (FRAP)	5	15	6	9	0
p21-activated kinase 2 (PAK2)	9	15	13	2	0
Mitogen- and stress-activated protein kinase (MSK)	5	12	12	0	0
Protein Kinase D	9	13	10	3	0
NimA-Related Kinase (NEK)	7	13	8	5	0
Microtubule Affinity Regulating Kinase	4	10	10	0	0

(MARK)					
Myosin Light Chain Kinase (MLCK)	7	10	5	5	0
Dual-specificity Tyrosine Regulated Kinase (DYRK)	11	13	7	0	6
Proto-oncogene tyrosine-protein kinase Src (Src)	101	171	0	0	171
Epidermal growth factor receptor (EGFR)	29	67	0	0	67
Lymphocyte specific protein tyrosine kinase(LCK)	36	64	6	5	53
Abl Protein Tyrosine Kinase (Abl)	39	56	0	0	56
Proto-oncogene tyrosine-protein kinase FYN (Fyn)	31	56	0	0	56
Spleen tyrosine kinase (SYK)	22	51	0	0	51
Tyrosine-protein kinase LYN (Lyn)	28	50	0	0	50
Janus kinase (Jak)	22	46	0	0	46
Insulin Receptor kinase (InsR)	14	46	0	0	46
platelet derived growth factor receptor (PDGFR)	12	32	0	0	32
Insulin-like growth factor I receptor (IGF1R)	7	31	0	0	31
Met proto-oncogene tyrosine kinase (MET)	6	31	0	0	31
Tec protein tyrosine kinase family (Tec)	14	30	0	0	30
Fibroblast growth factor receptor (FGFR)	8	30	0	0	30
Anaplastic lymphoma kinase (ALK)		22	0	0	22
Ephrin receptor (EPH)	13	22	0	0	22
C-SRC kinase (CSK)	11	22	0	0	22
Vascular Endothelial Growth Factor Receptors (VEGFR)	6	23	0	0	23
Tyrosine-protein kinase ZAP-70 (ZAP70)	9	20	0	0	20
Insulin receptor (IR)	12	18	2	2	14
Bruton's tyrosine kinase (BTK)	8	18	0	0	18
Hemopoietic cell kinase (Hck)	12	17	2	3	12
Focal adhesion kinase (FAK)	12	15	0	0	15
Proto-oncogene tyrosine-protein kinase receptor ret (Ret)	3	14	0	0	14
TRK transforming tyrosine kinase protein (TRK)	4	13	0	0	13
Discoidin Domain Receptor kinase (DDR)	11	13	0	0	13
Integrin Linked Kinase (ILK)	9	11	0	0	11
Proto-oncogene tyrosine-protein kinase Fes/Fps (Fes)	3	9	0	0	9
Proto-oncogene tyrosine-protein kinase FGR (Fgr)	4	8	0	0	8
Platelet-derived growth factor, FMS (Fms)	2	8	0	0	8
Non-receptor tyrosine-protein kinase TYK2 (TYK2)	5	8	0	0	8
Proto-oncogene tyrosine-protein kinase YES (YES)	4	6	0	0	6

The combined experimental verified phosphorylation sites (non-redundant) are defined as the positive data set. On the other hand, the serine, threonine and tyrosine, which are not annotated as phosphorylated sites within the experimental validated phosphorylated proteins, are defined as the negative data set. However, the positive data set may contain several homologous sites among orthologous proteins. To avoid the overestimation of predictive performance, the positive data set was further removed the homologous sequences with a given window size $2n+1$ (from upstream n to downstream n residues centering the phosphorylated site) among orthologous proteins, where n varies from 4 to 10. Referred to the homology reduction of MeMo [83], two phosphorylated protein sequences with more than 30% identity were specified to re-align the fragment sequences with residues of window length $2n+1$ centered on modified sites using BL2SEQ. If two fragment sequences were similar with 100% identity, and the phosphorylated sites from the two proteins were at the same position corresponding whole protein, then only one site was kept, while the other one was discarded. The homology reducing process was also carried out on negative data set.

After the homology reduction, randomly sampled nine-tenth of the non-homologous positive datasets is defined as the positive training set. To avoid the skew classifying ability for positive or negative set, the balanced negative training set is extracted from the non-homologous negative datasets. However, the negative training set, if is randomly selected in one time, may be not random sampling enough. Therefore, thirty negative training sets are constructed by randomly extracting from the non-homologous negative datasets. The average predictive performance of the thirty sets of training data is calculated after cross-validation. On the other hand, randomly sampled one-tenth of the non-homologous positive datasets is defined as the positive independent test set. The negative independent test set is also randomly sampled from the non-homologous negative datasets, which is balanced to positive independent test set. Sometimes, the trained model can classify the training data very well, but not effective for the independent test set. It might indicate that the trained model is over fitting for the training data. Thus, the constructed independent test set not only can be used to evaluate the predictive performance of the trained model, also can be used to measure whether the trained model is over fitting for the training data. To avoid the skew sampling of independent test set in one time, the independent test is executed in ten rounds.

3.4.2 Feature Extraction and Coding

Since the flanking sequences (position -5 ~ +5) of the phosphorylation sites (position 0) are graphically visualized as sequence logos [61], the conservation of amino acids in the phosphorylation sites can be observed. 11-mer sequences (-5 ~ +5) of kinase-specific phosphorylation sites are extracted and constructed as training sets.

This study not only takes the flanking amino acids (AA) as the training feature, but also takes the solvent accessible surface area (ASA) and secondary structure (SS) surrounding the phosphorylated sites into account. The fragment of amino acids with window length $2n+1$ centered on phosphorylated site are extracted from positive and negative training sets. An orthogonal binary coding scheme is used to transform amino acids into numeric vectors, which is the so called 20-dimensional vector coding. For example, glycine is encoded as “10000000000000000000,” alanine is encoded as “01000000000000000000,” and so on. The number of feature vector representing the flanking amino acids surrounding phosphorylated site is $(2n+1) \times 20$. Different values of n varying from 4 to 10 are used to determine the optimized window length. Furthermore, the positional weighted matrix (PWM) of amino acids surrounding the phosphorylated sites is calculated for four phosphorylated residues by using non-homologous training data. The positional weighted matrix (PWM) is the relative frequency of amino acids in a position surrounding the phosphorylated sites, which is also used to encode the fragment sequences.

Because most of the experimental phosphorylated proteins don't have the corresponded protein tertiary structures in PDB, an effective tool, named RVP-Net[18, 70], is used to compute the ASA value based on protein sequence. The computed ASA value is the percentage of accessible surface area for each amino acid on protein sequence. RVP-net incorporated the neural network to predict real value of ASAs for residues based on neighborhood information, which could reach 18.0 – 19.5% mean absolute error, defined as per residue absolute difference between the predicted and experimental values of relative ASA.[18] The full-length protein sequences with experimental phosphorylated sites are inputted to RVP-Net to compute the ASA value for all residues. The ASA values of amino acids surrounding the phosphorylated site are extracted and scaled in 0 to 1.

In the investigation of secondary structure surrounding the phosphorylated sites, PSIPRED[21] is used to compute secondary structure based on protein sequence. PSIPRED is a simple and reliable secondary structure prediction method, incorporating two feed-forward

neural networks which perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated - BLAST) [71]. PSIPRED 2.0 achieved an average Q₃ score of 80.6% across all 40 submitted target domains with no obvious sequence similarity to structures present in PDB, which ranked PSIPRED top out of 20 evaluated methods[109]. The output of PSIPRED includes three symbols “H,” “E” and “C” which stand for helix, sheet and coil, respectively. The full-length protein sequences with phosphorylated sites are inputted to PSIPRED to determine the secondary structure for all residues, respectively. The orthogonal binary coding scheme is used to transform three symbols of secondary structure into numeric vectors. For example, helix is encoded as “100,” sheet is encoded as “010,” and coil is encoded as “001.”

3.4.3 Model Learning and Evaluation

There are mainly four types of features such as amino acid (AA), secondary structure (SS), accessible surface area (ASA) and disorder region (DIS) been evaluated the discriminatory power between phosphorylated and non-phosphorylated sites. The support vector machine (SVM) is applied to create the computational models with the encoded amino acids and structural features, secondary structure and accessible surface area. With the binary classification, the concept of SVM is mapping the input samples onto a higher dimensional space through a kernel function, and then seeking a hyper-plane that discriminates the two classes with maximal margin and minimal error. A public SVM library, namely LibSVM [110], is adopted to train the predictive model with the positive and negative training sets which are encoded according to different types of training features. Radial basis function (RBF) $K(S_i, S_j) = \exp(-\gamma \|S_i - S_j\|^2)$ is selected as the kernel function of SVM.

Referred to previous work, KinasePhos [55], incorporated profile hidden Markov models (HMMs) for identifying kinase-specific phosphorylation sites. It shows that the HMM can perform accurate prediction for phosphorylation sites. Therefore, HMMER [63] is used to train the HMMs from the fragments of amino acids surrounding the phosphorylated sites. An HMM describes a probability distribution over a potentially infinite numbers of sequences, which can be used to detect distant relationships between amino acid sequences. The emission and transition probabilities of HMM are generated from the positive training set to capture the characteristics of the phosphorylated sites.

To evaluate the predictive performance of the trained models, k-fold cross-validation is

performed on phosphorylated lysine and arginine, and Jackknife cross-validation is adapted to phosphorylated glutamate and asparagine whose data size is smaller than 30. The following measures of predictive performance of the trained models are defined:

$$\text{Precision (Pre)} = \frac{TP}{TP + FP},$$

$$\text{Sensitivity (Sn)} = \frac{TP}{TP + FN},$$

$$\text{Specificity (Sp)} = \frac{TN}{TN + FP},$$

$$\text{Accuracy (Ac)} = \frac{TP + TN}{TP + FP + TN + FN},$$

where TP , TN , FP and FN are true positive, true negative, false positive and false negative, respectively. Matthews Correlation Coefficient (MCC) is defined as

$$\text{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}.$$

Because there are thirty negative training sets, the average precision, sensitivity, specificity, accuracy, and MCC are computed for each model trained with different window lengths and features. Moreover, the parameters of the predictive models, including window length, the cost value and gamma value of the SVM models, and bit score of HMM models, are optimized for achieving the best predictive accuracy. Finally, the window size and features that achieve the highest accuracy are adopted to construct the prediction models for independent test.

3.4.4 Independent Test

The prediction performance of the trained models might be overestimated because of the over-fitting for training set. To estimate the real prediction performance, about one-tenth part of the non-homologous data set are randomly selected as the independent test set, which will be used to evaluate the predictive performance of the trained models which reach the best accuracy based on the cross-validation. Because the number of training set in several kinase-specific groups is not efficient, the independent test set is constructed only for the groups that contain more than 10 phosphorylated sites. However, the performance of

independent test may be good by chance. To avoid the skew sampling of independent test set, the process of independent test is executed in ten rounds. Therefore, the construction of positive and negative training set, feature extraction, model training and evaluation, and independent test are implemented in ten rounds. The average performance of independent test will be computed. The independent test sets of lysine and arginine are not only adopted to test our method but also used to test other previously proposed protein phosphorylation prediction tools. Moreover, the experimentally verified phosphorylation sites with catalytic kinase from Human Protein Reference Database (HPRD) [51] are used to evaluate the predictive performance of the constructed models.

3.5 Results

In this section, the structural preferences of each kinase-specific group are investigated in detail. The predictive performances of cross-validation and independent test are discussed for each kinase-specific group. Finally, the selected models with highest predictive accuracy are used to implement a predictive system for protein kinase-specific phosphorylation sites.

3.5.1 Structural Investigation of Phosphorylation Sites

Previous studies have already shown that phosphorylation sites are mainly located in parts of proteins without regular structure [26, 58]. To verify this observation on the basis of large-scale studies and to enable users to investigate the structural properties of each kinase-specific group of interest, we employed several tools to large-scale analysis. The structural propensity of phosphorylated site was compared to non-phosphorylated site. The flanking amino acids of the non-redundant combined phosphorylated sites are graphically visualized as sequence logo, which can be easily investigated the conservation of amino acids surrounding the phosphorylated sites. WebLogo [60, 61] is used for creating the graphical sequence logo for the relative frequency of the corresponding amino acid at each position surrounding the phosphorylated sites, with a given window $-n \sim +n$ (position 0 is the phosphorylated site). **Figure 3.12** shows that the amino acids surrounding phosphorylated sites have higher conservation than non-phosphorylated sites. Moreover, the amino acids surrounding the phosphorylation sites in each kinase group are listed to investigate the kinase-specific substrate specificity.

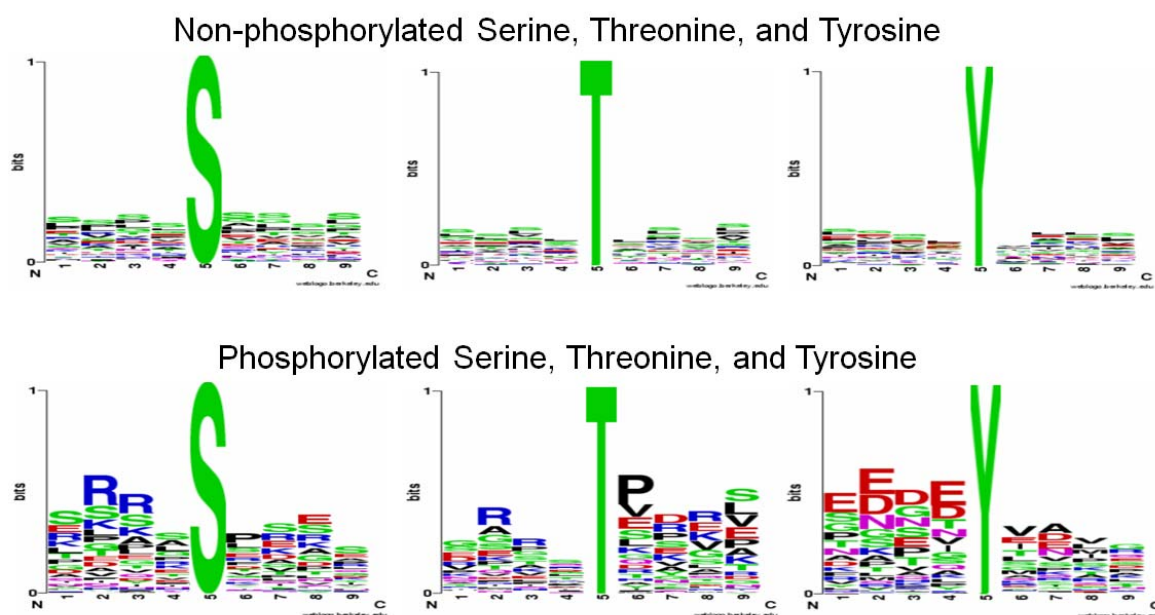


Figure 3.12 Comparison of flanking amino acids between phosphorylated and non-phosphorylated sites.

Structural Investigation of Kinase-specific Phosphorylation Site

Due to various types of the annotated catalytic kinase, the investigation of structure features, such as surface accessibility, secondary structure and intrinsic disorder regions, are presented for each kinase-specific group. **Table 3.5** lists several common kinase-specific groups, including PKA, PKC, PKB, ATM, CaMK2, CDK, CDC2, CK1, CK2, MAPK, Aurora, Abl, EGFR, InsR, and Src, due to the abundance of enough experimental verified data. The flanking amino acids of the non-redundant combined phosphorylated sites categorized by their catalytic kinase are graphically visualized as sequence logo, which can be easily investigated the conservation of amino acids surrounding the phosphorylated sites. WebLogo [60, 61] is used for creating the graphical sequence logo for the relative frequency of the corresponding amino acid at each position surrounding the phosphorylated sites, with a given window $-5 \sim +5$ (position 0 is the phosphorylated site).


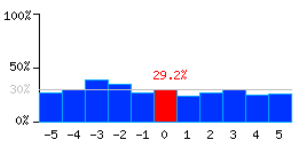
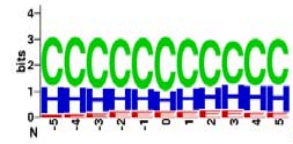
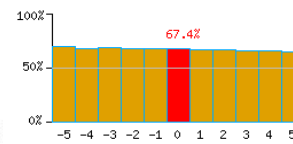
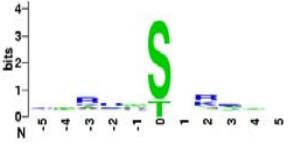
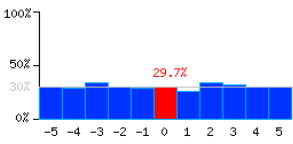

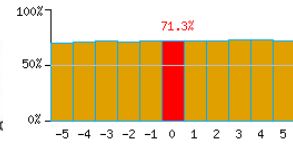
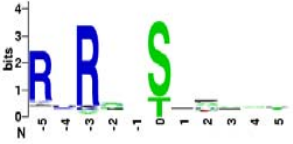
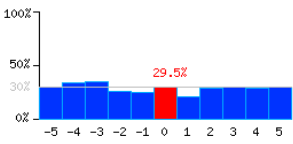
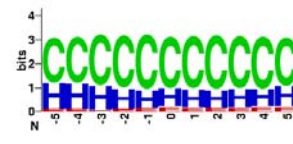
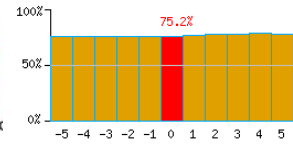
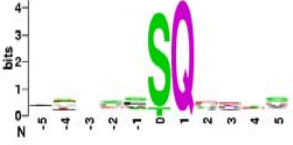
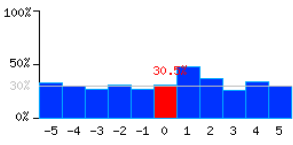
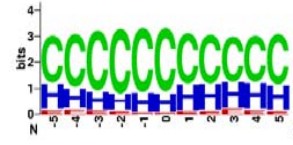
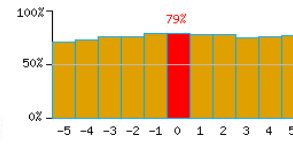

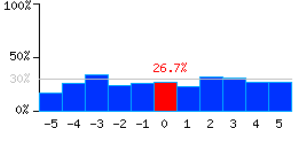

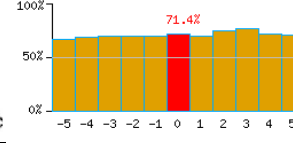

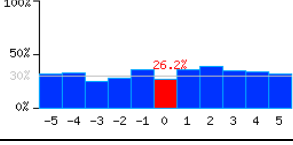
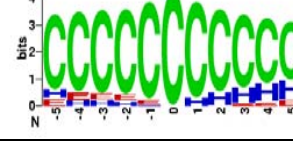
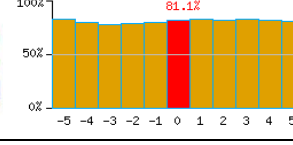
As the representation of sequence logo, there are obvious conserved amino acids surrounding the modified sites in most kinase-specific groups. In the case of serine/threonine kinase groups, for example, PKA group have enriched arginine (R) in position -2 and -3, which get the same consensus motif in **Figure 3.1** [6]. Group PKB, CaMK2, and Aurora also have the enriched arginine surrounding the phosphorylated sites. PKC group have slightly enriched arginine surrounding the phosphorylated sites. Several kinase groups are

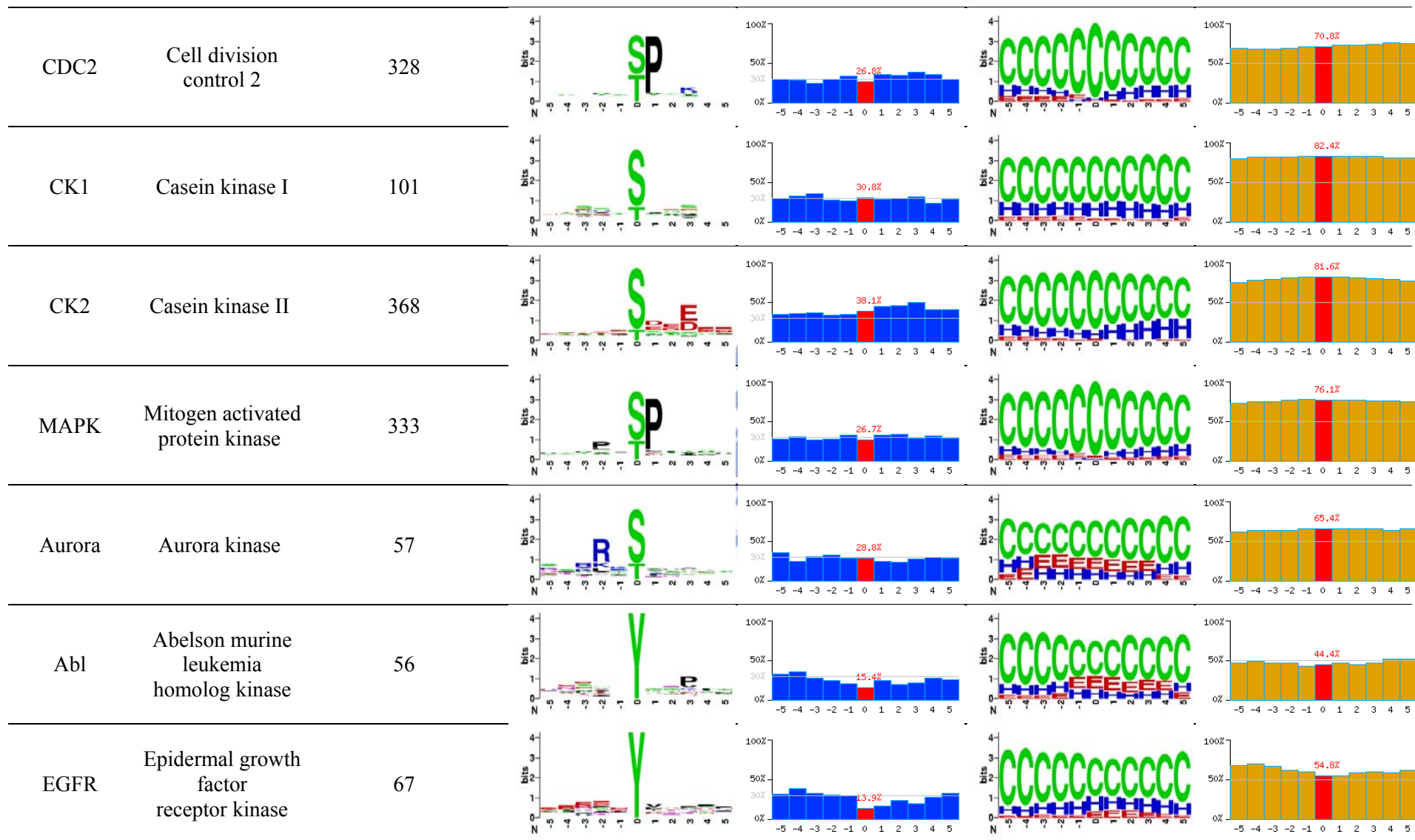
proline-directed phosphorylated sites, such as CDK, CDC2, and MAPK. Moreover, ATM kinase is involved in glutamine (Q) in position +1. Although **Figure 3.1** shows that the InsR and EGFR have consensus motifs, in the case of tyrosine kinase groups. However, most of them have no obvious conserved amino acids surrounding the phosphorylated sites.

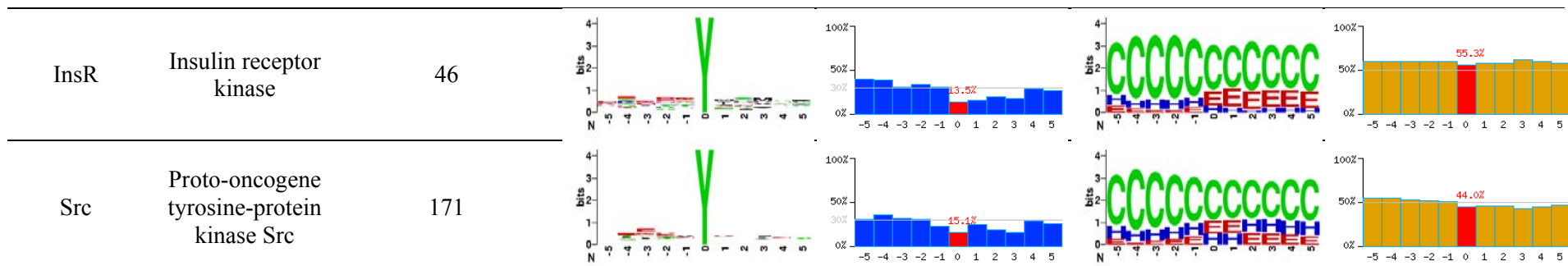
Due to the number of experimental phosphorylated sites which are located in the protein regions with tertiary structure of PDB [67] is not sufficient, RVP-Net [18, 70], PSIPRED [21], and DISOPRED2 [35] are used to compute the ASA value, secondary structure, and intrinsic disorder region based on protein sequence, respectively. The average percentage of ASA, sequence logo of secondary structure, and average percentage of disorder region within 11-mer window (-5 ~ +5) are also shown in **Table 3.5**. In the investigation of solvent accessibility, most of the methylated sites are located in the highly accessible surface area, besides the methylated asparagines. The average solvent accessible surface area surrounding the methylated lysine is highly similar to the observation in protein tertiary structure. In the observation of secondary structure surrounding the phosphorylated sites, most of the phosphorylated sites are likely occurred on coil (loop).



Table 3.5 Structural features of kinase-specific groups.

Kinase group	Description	Number of phosphorylated sites*	Amino acids of surrounding residue	Average surface accessibility of flanking residues	Secondary structure of flanking residues	Average disorder rate of flanking residues
PKA	Protein kinase A	458				
PKC	Protein kinase C	485				
PKB	Protein kinase B	114				
ATM	Ataxia telangiectasia mutated kinase	102				
CaMK2	Calcium/calmodulin-dependent protein kinase II	119				
CDK	Cyclin dependent kinase	123				





*The statistics of phosphorylated sites is the non-redundant experimentally verified phosphorylation sites extracted from Phospho.ELM and UniProtKB/Swiss-Prot.



3.5.2 Predictive Performance

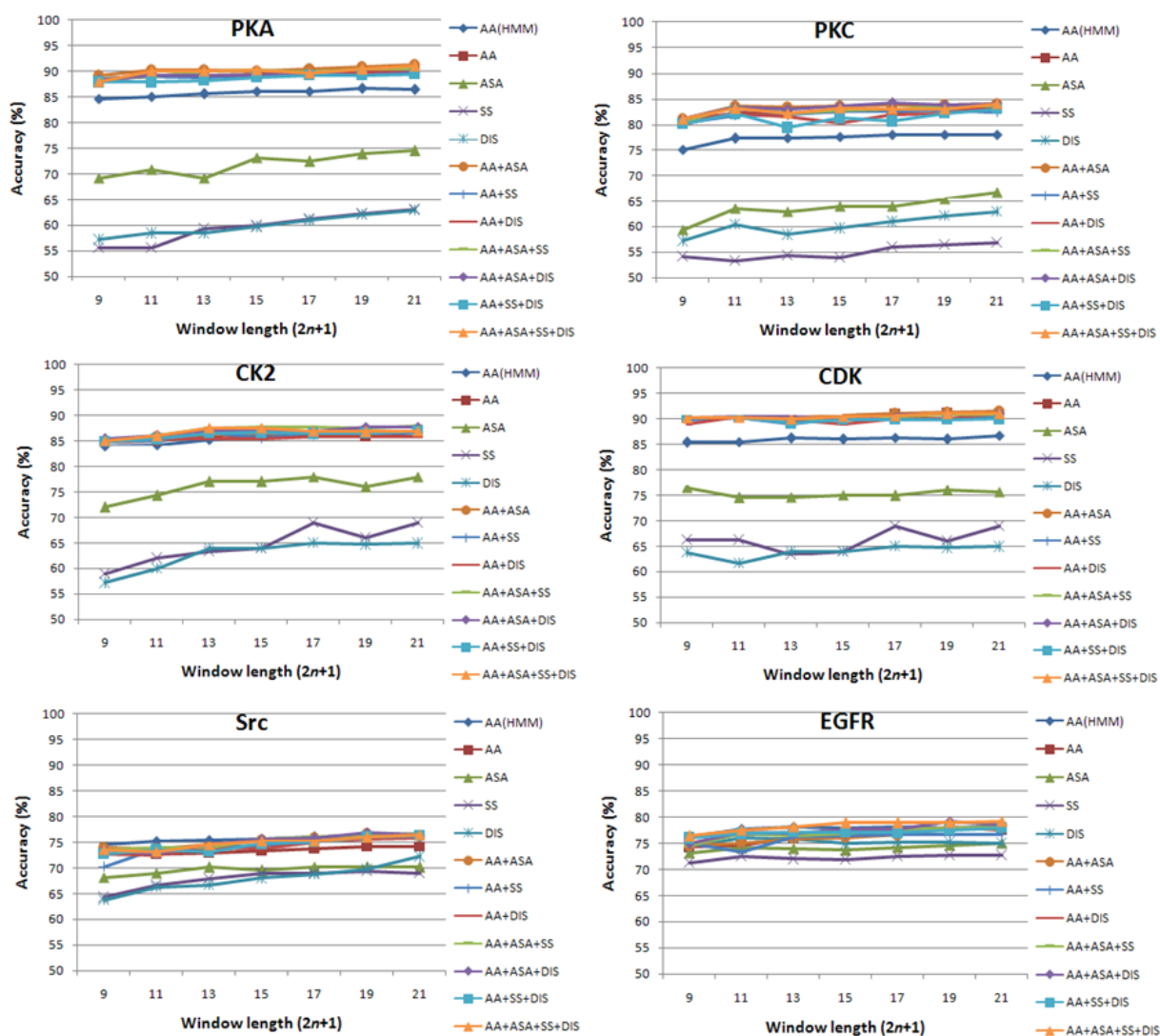


Figure 3.13 Predictive accuracy of PKA, PKC, CK2, CDK, Src and EGFR models trained with different training features, based on various window sizes.

To investigate what kinds of window length and feature can be adopted to construct the model which achieves the best prediction performance in each kinase-specific group, the models trained with different window lengths and various features are evaluated based on k-fold cross-validation. There are four major types of features which are including amino acid (AA), accessible surface area (ASA), secondary structure (SS), and intrinsic disorder (DIS). The feature of amino acids surrounding the phosphorylated sites is encoded with 20-dimensional vector and positional weighted matrix, which are named “AA(20D)” and “AA(PWM),” respectively. The features of accessible surface area and secondary structure are encoded with the ASA values and 3-dimensional vector, respectively. **Figure 3.13** illustrates the predictive accuracy of the models trained with different training features, based on various window sizes

$2n+1$, where n varies from 4 to 10. Especially, the feature of amino acids surrounding the phosphorylated sites is also trained with profile hidden Markov model. Investigating into the models trained with individual features, the model trained with ASA values performs slightly better than the models trained with SS or DIS in overall kinase groups, whose amino acids surrounding the phosphorylated site are not conserved. In general, the kinase-specific model trained with AA performs strongly better than the model trained with ASA, and the model trained with SS performs worst. In PKA, PKC, CK2 and CDK groups, the performance of the model trained with AA is usually better than the model trained with ASA, SS and DIS, because their flanking amino acids are conserved. However, the performance of the model trained with AA in Src and EGFR groups is not absolutely better than ASA, SS, or DIS, because the amino acids surrounding the phosphorylated site are not highly conserved. As you can see, the performance of the model trained with SS is generally worst. As far as various window sizes are concerned in each kinase group, the predictive accuracy is increased with window size increasing from 4 to 10.

The predictive performance of the model trained with the combination of AA, ASA, SS and DIS features is also evaluated. As previous illustration, the feature of ASA can perform higher than 70% accuracy on most kinase groups. Therefore, the models trained with the combination of AA and ASA can perform better than the model trained individually with AA or ASA. However, the predictive accuracy of the model trained with the combination of AA(20D) and ASA is not better than the model trained individually with ASA. Because AA(20D) encoding method is 20 times the number of dimensions in ASA, the weight of AA feature is higher than ASA features in phosphorylation prediction. Thus, the predictive performance is mainly dominated by the AA feature. On the other hand, the number of dimensions in AA(PWM) is equal to ASA, which makes the weight of ASA and AA balanced in the classification between phosphorylated and un-phosphorylated sites. The average cross-validation performance of the models trained with different window sizes and features which achieve the highest accuracy are listed in **Table 3.6**. The training features which achieve the highest accuracy is the combination of AA(PWM) and ASA. To consider the overall performance of the models trained with different window sizes, $-6 \sim +6$ is selected as the feasible window size for the four phosphorylated residues. The average precision, sensitivity, specificity, accuracy and Matthews Correlation Coefficient of the models trained with the selected features and window sizes are given in **Table 3.6**. The overall predictive accuracy of the kinase-specific groups is 89.6%.

Table 3.6 Average cross-validation performance of several common kinase-specific groups with training features which reach highest accuracy.

Kinase type	Kinase group	Number of non-homologous training set	Training features	Window length	Pre (%)	Sn (%)	Sp (%)	Acc (%)	MCC
Serine/ threonine kinase	PKA	373	AA+ASA+SS+DIS	-5 ~ +5	91.2	89.4	91.0	90.2	0.82
	PKC	386	AA+ASA	-6 ~ +6	87.1	80.8	86.2	83.5	0.67
	CK2	299	AA+SS	-6 ~ +6	87.9	87.1	88.1	87.6	0.75
	MAPK	286	AA+ASA	-4 ~ +4	93.9	90.2	93.2	91.7	0.83
	CDC2	172	AA+ASA+DIS	-5 ~ +5	93.1	88.2	92.6	90.4	0.71
	CDK	105	AA+ASA	-5 ~ +5	97.1	92.7	96.3	94.5	0.89
	PKB	98	AA+ASA	-5 ~ +5	91.4	91.9	91.5	91.7	0.83
	CaMK2	94	AA+DIS	-6 ~ +6	82.7	77.9	84.3	81.1	0.62
	ATM	90	AA	-4 ~ +4	98.9	96.1	98.1	97.1	0.94
	CK1	87	AA+DIS	-8 ~ +8	74.2	77.2	75.6	76.4	0.53
Tyrosine kinase	Aurora	49	AA+ASA	-5 ~ +5	80.2	78.1	81.1	79.6	0.59
	Src	142	AA+ASA	-9 ~ +9	77.0	77.1	76.7	76.9	0.54
	Abl	48	AA+ASA	-8 ~ +8	73.8	72.8	74.0	73.4	0.47
	EGFR	55	AA+ASA+DIS	-10 ~ +10	74.8	78.7	75.5	77.1	0.54
	InsR	41	AA+SS+DIS	-9 ~ +9	75.0	77.4	75.2	76.3	0.53

Abbreviation: AA, amino acid; ASA, accessible surface area; SS, secondary structure; DIS, disorder; Pre, precision; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews Correlation Coefficient.

3.5.3 Predictive Performance of Independent Test

Based on the proposed benchmark of constructing the training data and test data, nine-tenth part and one-tenth part of experimental data of several kinase groups which contain more than 50 experimental phosphorylation sites are defined as the training set and independent test set, respectively. After the evaluation of cross-validation, the independent test sets are used to evaluate whether the constructed models are over-fitting for their training data. As given in **Table 3.7**, the number of test set, precision, sensitivity, specificity, and accuracy in each kinase group are listed. In general, kinase groups with enough training data, such as PKA, PKC, CK2, and MAPK, will perform robustly for classifying phosphorylation sites from non-phosphorylation sites. However, CaMK2, CK1, and EGFR perform worse than the cross-validation accuracy in 10%.

Table 3.7 Performance of independent test in several common kinase-specific groups.

Kinase	Number of positive test set	Number of negative test set	Pr	Sn	Sp	Acc
PKA	41	41	86.4	92.7	85.4	89.0
PKC	44	44	89.7	79.5	90.9	85.2
PKB	11	11	100	90.9	100	95.4
ATM	9	9	100	100	100	100
CaMK2	10	10	75.0	60.0	80.0	70.0
CDK	11	11	100	90.9	100	95.4
CDC2	19	19	88.9	84.2	89.5	86.8
CK1	9	9	63.6	77.8	55.6	66.7
CK2	34	34	82.9	85.3	82.4	83.8
MAPK	32	32	90.3	87.5	90.6	89.1
Aurora	5	5	66.7	80.0	60.0	70.0
Abl	5	5	80.0	80.0	80.0	80.0
EGFR	6	6	66.7	66.7	66.7	66.7
Src	16	16	73.3	68.7	75.0	71.9

Abbreviation: Pr, precision; Sn, sensitivity; Sp, specificity; Acc, accuracy.



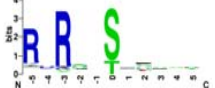






The average performance of the independent test is slightly worse than the performance of cross-validation. If the performance of independent test is strongly worse than cross-validation, it indicates that the constructed model may be over fitting for the training data. This independent test shows that the constructed models may be not over fitting to the training data. The independent test sets of several representative kinase groups are also used to test other phosphorylation site predictors.

3.6 Discussions

3.6.1 Kinase-specific Groups with Similar Consensus Motif

In order to assess the cross predictive specificity of the kinase-specific models containing the similar substrate site motifs, we take a particular group as the positive set and the other groups as the negative sets one by one. The higher specificity the cross-validation, the less incorrect prediction of the phosphorylation sites in other groups. As given in **Table 3.8**, the number in the parenthesis besides the kinase name indicates the size of the positive set. For example, the first row gives that there are 414 phosphorylated sites in kinase PKA group. The sensitivity (S_n) of the PKA model is 89.4%. The specificity are given in the table, for instance, in the first column the specificity (S_p) of PKC, PKB, CaMK2 and Aurora sets corresponding to the PKA model are 51.4%, 27.5%, 39.2% and 38.6%, respectively. The S_p values marked in red color indicate they are relatively lower between the kinases PKA, PKC, PKB, CaMK2 and Aurora in basophilic group. Similarly, the S_p values in green color indicate they are relatively lower between the kinases CKI and CKII in acidophilic group. The S_p values in blue color indicate they are relatively lower between the kinases CDC2 and MAPK in proline-directed group. We observe that the specificity corresponding to the kinase-specific data sets in the same kinase group, such as basophilic, acidophilic and proline-directed groups, are relatively lower than the specificity corresponding to the kinase-specific data sets in other groups.

Table 3.8 The cross predictive specificity of the kinase-specific models with similar substrate motif.

Positive set	Negative set	PKA (414)	PKC (430)	PKB (109)	CaMK2 (104)	Aurora (54)	CK1 (96)	CK2 (333)	CDC2 (191)	MAPK (318)
PKA (414)		Sn = 89.4	51.4	27.5	39.2	38.6	84.7	97.3	97.6	98.1
PKC (430)		34.3	Sn = 80.8	36.2	43.7	62.1	85.6	96.8	87.4	93.4
PKB (109)		47.8	83.2	Sn = 91.9	52.2	81.2	97.9	98.1	100	100
CaMK2 (104)		56.2	69.6	49.2	Sn = 77.9	58.3	93.2	95.4	98.0	93.5
Aurora (54)		41.8	71.3	70.2	68.9	Sn = 78.1	86.7	91.2	98.4	94.2
CK1 (96)		84.3	85.5	92.4	88.8	82.4	Sn = 77.2	75.8	94.0	96.2
CK2 (333)		95.2	95.4	98.1	89.5	90.7	72.4	Sn = 87.1	100	97.6
CDC2 (191)		98.1	96.2	100	95.7	100	100	99.3	Sn = 88.2	67.6
MAPK (318)		97.6	98.6	98.9	99.1	96.4	96.2	98.9	46.2	Sn = 90.2

Abbreviation: Pr, precision; Sn, sensitivity; Sp, specificity; Acc, accuracy.

3.6.2 Comparison with Other Phosphorylation Prediction

Tools

The proposed method is compared with several previously developed phosphorylation prediction tools, such as PredPhospho [78], GPS [75, 106], PPSP [76], MetaPredictor [107], KinasePhos 1.0 [55, 56], and KinasePhos 2.0 [91]. As given in **Table 3.9**, the number of kinases, sensitivity and specificity of prediction and the overall predictive performance of these tools are compared. GPS, PPSP, PredPhospho, KinasePhos 1.0 and 2.0, and the proposed methods all support the identification of kinase-specific phosphorylation sites. Although only the kinase groups containing at least 10 experimental phosphorylation sites were selected to evaluate the average predictive performance, this proposed version of KinasePhos provided about 100 kinase-specific models. The predictive performance of three representative kinases such as PKA, PKC and CK2 are compared. As given in **Table 3.9**, the cross-validation performances of three representative kinases in KinasePhos 3.0 are similar to PredPhospho, GPS, PPSP, and KinasePhos 1.0 and 2.0. In particular, KinasePhos 2.0 provides the predictive model for phosphohistidine, whose predictive accuracy is 93%. In this version, the kinase-specific group was not further categorized into serine, threonine, and tyrosine, while the predictive performance was not decreased. The overall predictive accuracy of the kinase-specific groups with at least 10 phosphorylation sites of the proposed method is 89%.

Table 3.9 Comparison of KinasePhos 3.0 with PredPhospho, GPS, PPSP, MetaPredictor, KinasePhos 1.0, and KinasePhos 2.0.

Tools	PredPhospho	GPS	PPSP	MetaPredictor	KinasePhos 1.0	KinasePhos 2.0	KinasePhos 3.0
Reference	Kim et al., 2004	Zhou et al., 2004	Xue et al., 2006	Ji Wan et al., 2008	Huang and Lee et al., 2005	Wong and Lee et al. 2007	-
Method	SVM	MCL+GPS	BDT	Voting from GPS, KinasePho, NetPhosK, PPSP, PredPhospho, Scansite	MDD+HMM	SVM	SVM
Material	PhosphoBase + UniProtKB/SwissProt	Phospho.ELM	Phospho.ELM	Phospho.ELM (5.0) + PhosphoSite (July 2006) + UniProtKB/SwissProt (51.1)	PhosphoBase + UniProtKB/SwissProt	Phospho.ELM + UniProtKB/SwissProt	Phospho.ELM (7.0) + UniProtKB/SwissProt (55.0)
Training features	AA	AA	AA	-	AA	AA+CP	AA+ASA+SS+DIS
No. of kinases	4 groups	71 groups	68 groups	-	18	58	101 groups
Kinase PKA	Sn = 0.88 Sp = 0.91	Sn = 0.89 Sp = 0.91	Sn = 0.90 Sp = 0.92	Sn = 0.88 Sp = 0.83	Sn = 0.91 Sp = 0.86	Sn = 0.92 Sp = 0.89	Sn = 0.89 Sp = 0.91
Kinase PKC	Sn = 0.79 Sp = 0.86	Sn = 0.82 Sp = 0.83	Sn = 0.82 Sp = 0.86	Sn = 0.77 Sp = 0.79	Sn = 0.80 Sp = 0.87	Sn = 0.84 Sp = 0.86	Sn = 0.81 Sp = 0.86
Kinase CK2	Sn = 0.84 Sp = 0.96	Sn = 0.83 Sp = 0.88	Sn = 0.83 Sp = 0.90	Sn = 0.88 Sp = 0.90	Sn = 0.87 Sp = 0.85	Sn = 0.87 Sp = 0.86	Sn = 0.87 Sp = 0.88
Serine	Acc = 0.81	-	-	-	Acc = 0.86	Acc = 0.90	-
Threonine	Acc = 0.77	-	-	-	Acc = 0.91	Acc = 0.93	-
Tyrosine	-	-	-	-	Acc = 0.84	Acc = 0.88	-
Histidine	-	-	-	-	-	Acc = 0.93	-
Overall performance	Acc = 0.76~0.91	-	-	-	Acc = 0.87	Acc = 0.91	Acc = 0.89

Abbreviation: SVM, support vector machine; MCL, Markov cluster algorithm; GPS, group-based phosphorylation scoring method; BDT, Bayesian decision theory; MDD, maximal dependence decomposition; HMM, hidden Markov model; CP, coupling pattern; AA, amino acid; ASA, accessible surface area; SS, secondary structure; DIS, disorder region; Sn, sensitivity; Sp, specificity; Acc, accuracy.

3.7 Summary

In general, the previous works of phosphorylation site prediction focused on flanking residues of phosphorylation sites; like our previous work (KinasePhos 1.0 and 2.0). Herein, KinasePhos 3.0 comprehensively investigates structural properties in each kinase-specific phosphorylated site. The protein structural properties, such as accessible surface area (ASA), secondary structure, and intrinsic disorder, were considered in the model training of each kinase groups. Because most of the experimentally verified kinase-specific phosphorylation sites do not located in the protein regions with known structure from PDB [67], the effective prediction tools RVP-net [70], PSIPRED [21], and DISOPRED2 [35] are adopted to compute the accessible surface area of residues, secondary structures, and protein disorder regions, respectively. The cross-validation demonstrates that the structural properties can improve the predictive accuracy ranging from 1% to 10%. The models trained with various features, including sequence profiles and structural features, were evaluated by 5-fold and Jackknife cross-validation, the predictive performance of the models trained with the combination of sequence and structural features are better than the models trained only with sequence. The overall accuracy of 101 kinase groups is 89.4%. Moreover, the independent test shows that the constructed model of kinase-specific groups were not over-fitting to training data. Finally, the constructed SVM models with best predictive accuracy were used to implement the web-based prediction tool.

Chapter 4 Discovery of Protein Kinase-Substrate

Phosphorylation Networks

4.1 Introduction

Protein phosphorylation catalyzed by protein kinases is the most widespread and well-studied signaling mechanism in eukaryotic cells. It was estimated that one-third to one-half of all proteins in a eukaryotic cell are phosphorylated [1]. Phosphorylation can regulate almost every property of a protein and is involved in all fundamental cellular processes. Cataloging and understanding protein phosphorylation is not easy task: many kinases may be expressed in a cell, and one-third of all intracellular proteins may be phosphorylated, representing as many as 20,000 distinct phosphoprotein states [111]. Manning et al. [95] have identified 518 human kinases, and every active protein kinase phosphorylates a distinct set of substrates in a regulated manner. Defining the kinase complement of the human genome, the kinome, has provided an excellent starting point for understanding the scale of the problem.

With the high-throughput mass spectrometry (MS) proteomics, the number of *in vivo* phosphorylation sites is increasing rapidly. However, about 20% of the experimentally verified phosphorylation sites have the annotation of catalytic kinases. To fully investigate how protein kinases regulate the intracellular processes, it is necessary to comprehensively and accurately identify the kinase-specific substrates. Therefore, we were inspired to integrate experimentally verified phosphorylation data and computational techniques for identifying physiological substrates of the protein kinases and studying phosphorylation network in cell. Due to the fact that signaling proteins are modular in the sense that they contain domains (catalytic or interaction) and linear motifs (phosphorylation or binding sites), which mediate interactions between proteins [92], the protein-protein interaction and protein association are incorporated. It also exploits both the inherent propensity of kinase catalytic domains to phosphorylate particular sequence motifs and contextual information regarding the physical interaction, functional association, cellular co-localization and coexpression of kinases and substrates.

Intracellular signal transduction is the process by which chemical signals from outside the cell are passed through cytoplasm to nucleus or cytoskeleton, where appropriate responses

to those signals are generated [7]. Deciphering the complex network of protein kinase and substrate is necessary for a thorough and therapeutically applicable understanding of the functioning of a cell in physiological and pathological states. Therefore, the comprehensive kinase-substrate interactions are used to construct the intracellular phosphorylation network starting from receptor kinases to transcription factors. Moreover, the gene expression data is adopted to validate the syn-expression of kinase and substrate with statistical significance.

Human Kinome

Manning *et al.* [95] have catalogued the protein kinase complement of the human genome, the so-called “kinome”, using public and proprietary genomic, complementary DNA, and expressed sequence tag (EST) sequences. This provides a starting point for comprehensive analysis of protein phosphorylation in normal and disease states, as well as a detailed view of the current state of human genome analysis through a focus on one large gene family. There are 518 putative protein kinase genes been identified, of which 71 have not previously been reported or described as kinases, and we extend or correct the protein sequences of 56 more kinases. New genes include members of well-studied families as well as previously unidentified families, some of which are conserved in model organisms. Classification and comparison with model organism kinomes identified orthologous groups and highlighted expansions specific to human and other lineages. The authors also identified 106 protein kinase pseudogenes.

Group	Families	Subfamilies	Yeast kinases	Worm kinases	Fly kinases	Human kinases	Human pseudogenes	Novel human kinases
AGC	14	21	17	30	30	63	6	7
CAMK	17	33	21	46	32	74	39	10
CK1	3	5	4	85	10	12	5	2
CMGC	8	24	21	49	33	61	12	3
Other	37	39	38	67	45	83	21	23
STE	3	13	14	25	18	47	6	4
Tyrosine kinase	30	30	0	90	32	90	5	5
Tyrosine kinase-like	7	13	0	15	17	43	6	5
RGC	1	1	0	27	6	5	3	0
Atypical-PDHK	1	1	2	1	1	5	0	0
Atypical-Alpha	1	2	0	4	1	6	0	0
Atypical-RIO	1	3	2	3	3	3	1	2
Atypical-A6	1	1	1	2	1	2	2	0
Atypical-Other	7	7	2	1	2	9	0	4
Atypical-ABC1	1	1	3	3	3	5	0	5
Atypical-BRD	1	1	0	1	1	4	0	1
Atypical-PIKK	1	6	5	5	5	6	0	0
Total	134	201	130	454	240	518	106	71

Figure 4.1 Kinase distribution by major groups in human and model systems (Manning *et al.*, 2002).

Most protein kinases contain a conserved catalytic domain belonging to the eukaryotic

protein kinase (ePK) superfamily (all other protein kinases are classified as atypical protein kinases, or aPKs). As shown in **Figure 4.1**, ePKs are classified into 9 major groups, and are subdivided into families, and sometimes subfamilies, based on the sequence of their ePK domains, including AGC, CAMK, CK1, CMGC, Other, STE, TK, TKL, and RGC. Manning *et al.* also identified 13 atypical protein kinase (aPK) families, which contain proteins reported to have biochemical kinase activity, but which lack sequence similarity to the ePK domain, and their close homologs. To compare related kinases in human and model organisms and to gain insights into kinase function and evolution, we classified all kinases into a hierarchy of groups, families, and subfamilies. Kinases were classified primarily by sequence comparison of their catalytic domains, aided by knowledge of sequence similarity and domain structure outside of the catalytic domains, known biological functions, and a similar classification of the yeast, worm, and fly kinomes. Phylogenetic comparison of the human kinome with those of yeast, worm, and fly confirms that most kinase families are shared among metazoans and defines classes that are expanded in each lineage. Of 189 subfamilies present in human, 51 are found in all four eukaryotic kinomes, and these presumably serve functions essential for the existence of a eukaryotic cell. Comparison with the draft mouse genome indicates that more than 95% of human kinases have direct orthologs in mouse; additional orthologs may emerge as that genome sequence is completed.

Figure 4.2 [95] shows a phylogenetic tree that depicts the relationships between members of the complete superfamily of human protein kinases. The 518 human protein kinases control protein activity by catalyzing the addition of a negatively charged phosphate group to other proteins. Most protein kinases belong to a single superfamily of enzymes whose catalytic domains are related in sequence and structure. The main diagram illustrates the similarity between the protein sequences of these catalytic domains. Each kinase is at the tip of a branch, and the similarity between various kinases is inversely related to the distance between their positions on the tree diagram. Most kinases fall into small families of highly related sequences, and most families are part of larger groups. The seven major groups are labeled and colored distinctly. Other kinases are shown in the center of the tree, colored gray. The relationships shown on the tree can be used to predict protein substrates and biological function for many of the over 100 uncharacterized kinases presented here. The inset diagram shows trees for seven atypical protein kinase families. These proteins have verified or strongly predicted kinase activity, but have little or no sequence similarity to members of the protein kinase superfamily. A further eight atypical protein kinases in small families of one or

two genes are not shown.

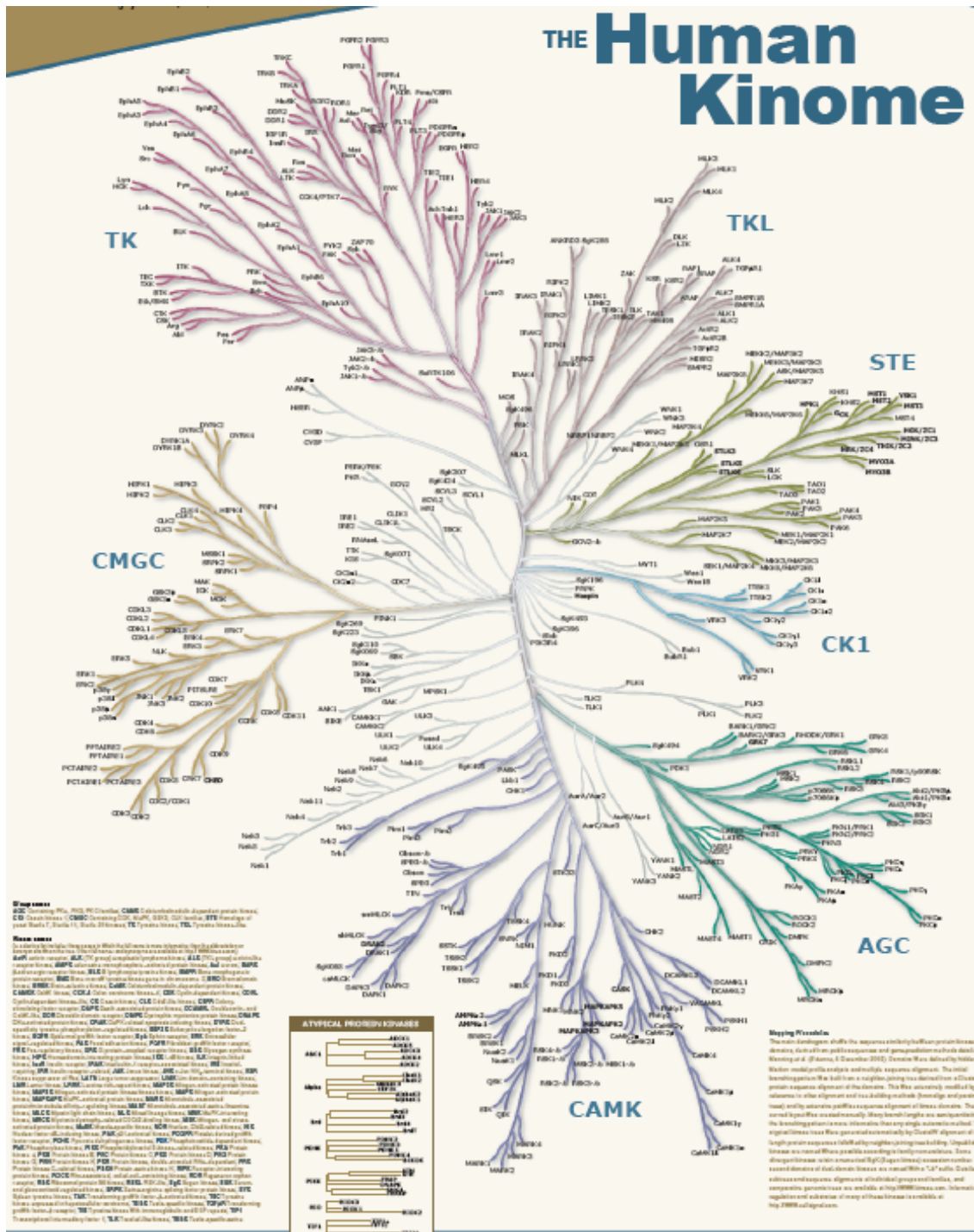


Figure 4.2 Phylogenetic tree of human kinome (Manning *et al*, 2002).

4.2 Related Works

Due to the high-throughput of mass spectrometry-based proteomics, there are several databases storing experimentally verified phosphorylation sites with catalytic kinase, such as

Phospho.ELM [2], PhosphoSite [43], UniProtKB/Swiss-Prot [42], Phosphorylation Site Database [44], and PHOSIDA [45]. The resource can be utilized for constructing the phosphorylation network between kinase and substrate proteins. The experimental data, ultimately, need to be combined by systems biology analysis, which translates the separate, large-scale datasets into signaling networks [13].

4.2.1 Discovery of Human Phosphorylation Networks

Protein kinases control cellular decision processes by phosphorylating specific substrates. Thousands of *in vivo* phosphorylation sites have been identified, mostly by proteome-wide mapping. However, systematically matching these sites to specific kinases is presently infeasible, due to limited specificity of consensus motifs, and the influence of contextual factors, such as protein scaffolds, localization, and expression, on cellular substrate specificity. Linding *et al.* [112] proposed a method, namely NetworKIN⁹, that augments motif-based predictions with the network context of kinases and phosphoproteins. In the first step, the authors use neural networks (NetPhosK [113]) and position-specific scoring matrices (ScanSite [105]) to assign each phosphorylation site to one or more kinase families, based on the intrinsic preference of kinases for consensus substrate motifs. In the second stage, the context for each substrate is represented by a probabilistic protein network extracted from the STRING database [114], which integrates information from curated pathway databases, cooccurrence in abstracts, physical protein interaction assays, mRNA expression studies, and genomic context. This approach captures both direct and indirect interactions; for example, phosphorylation events mediated by scaffolds are predicted, as the scaffolding protein provides a path in the probabilistic network between the substrate and kinase.

NetworKIN pinpoints kinase responsible for specific phosphorylation and yields a 2.5-fold improvement in the accuracy with which phosphorylation networks can be constructed. As shown in **Figure 4.3**, manually curated data sets of CDK, PKC, PIKK, and INSR *in vivo* phosphorylation sites were used to assess the prediction accuracy (the fraction of predictions that are known to be correct) and sensitivity (the fraction of known sites that are correctly predicted) of NetworKIN and solely motif-based methods (NetPhosK and Scansite). This shows that including the cellular context (in the form of a protein association network) leads to a significant improvement in accuracy. Notably, the accuracy of

⁹ NetworKIN URL: <http://networkin.info/index.php>

NetworkKIN predictions is likely to be an underestimate since not all the kinases that target each phosphorylation site in the set of test proteins may currently be known from experiments.

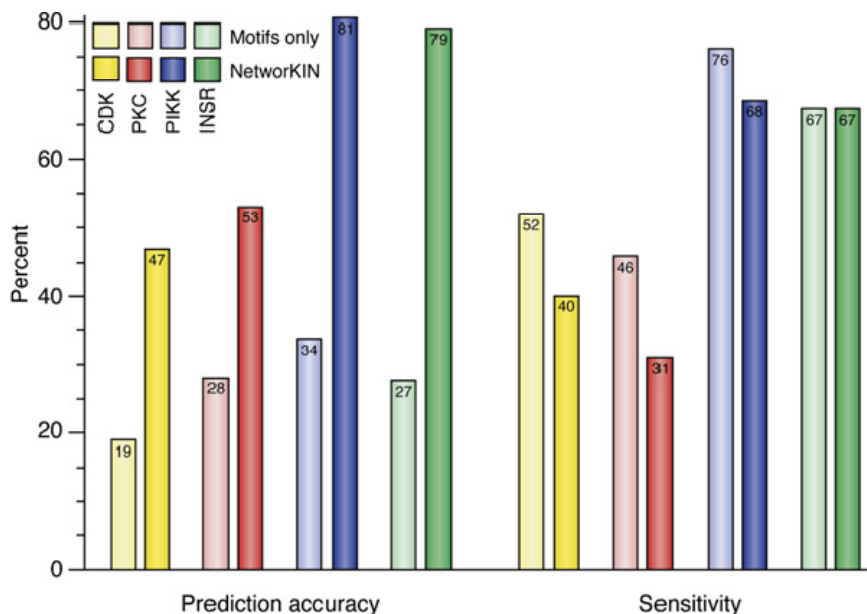


Figure 4.3 Effects of including network context (Linding *et al.*, 2007).

4.2.2 Human Kinase Interactome Resource

PhosphoPOINT [115] is comprehensive human kinase interactome and phospho-protein database, which collects 4,195 phospho-proteins with a total of 15,738 phosphorylation sites. PhosphoPOINT annotates the interactions among kinases, with their downstream substrates and with interacting (phospho)-proteins to modulate the kinase-substrate pairs. PhosphoPOINT integrates various gene expression profiles and Gene Ontology (GO) cellular component information to evaluate each kinase and their interacting (phospho)-proteins/substrates. Integration of cSNPs that cause amino acids change with the proteins with the phospho-protein dataset reveals that 64 phosphorylation sites result in a disease phenotypes when changed; the linked phenotypes include schizophrenia and hypertension. PhosphoPOINT also provides a search function for all phospho-peptides using about 300 known kinase/phosphatase substrate/binding motifs. Altogether, PhosphoPOINT provides robust annotation for kinases, their down-stream substrates and their interaction (phospho)-proteins and this should accelerate the functional characterization of kinome-mediated signaling. **Figure 4.4** [115] shows Auroa kinase as an example to illustrate the search result.

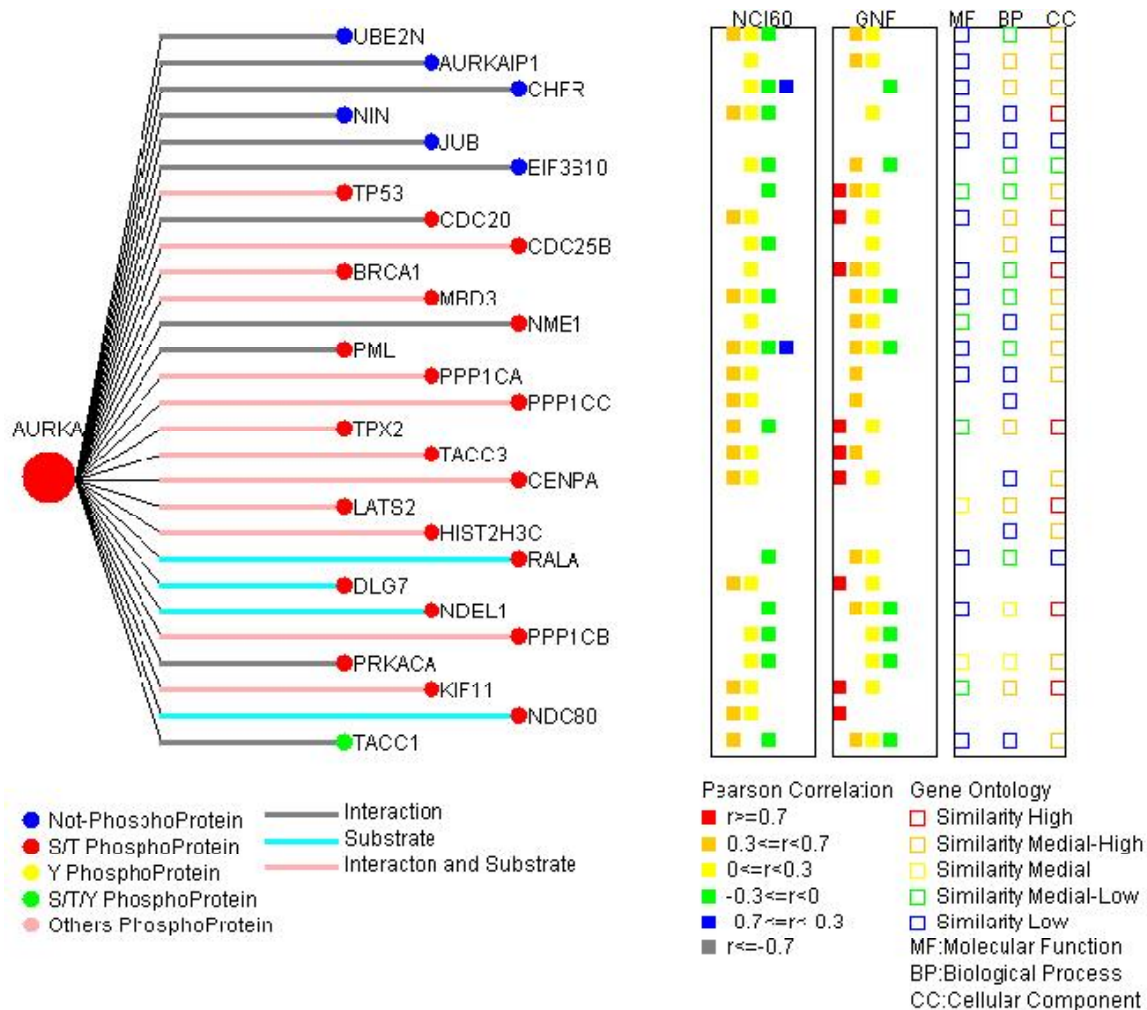


Figure 4.4 Annotation and visualization of PhosphoPOINT (Yang *et al.*, 2008).

4.2.3 Modeling of Signal Transduction Networks

Signaling pathways have been an active research area in recent history. There are many studies in which signaling pathways were modeled using various approaches. Previously, signaling pathways were modeled through modular kinetic simulations of biochemical networks [116] and detailed integration of biochemical properties of the pathways [117]. In another study, Bayesian Networks were applied to multi-variable cell data to infer signaling pathways [118]. Correlating cancer based mRNA expression levels, autocrine receptor signaling loops were also discovered [14]. Another approach to model cellular pathways was developed based on perturbations of critical pathway components [15]. These were analyzed using DNA microarrays, quantitative proteomics, and databases of known physical interactions.

Steffen *et al.* [119] have developed a computational approach for generating static

models of signal transduction networks which utilizes protein-interaction maps generated from large-scale two-hybrid screens and expression profiles from DNA microarrays. Networks are determined entirely by integrating protein-protein interaction data with microarray expression data, without prior knowledge of any pathway intermediates. In effect, this is equivalent to extracting subnetworks of the protein interaction dataset whose members have the most correlated expression profiles. The authors show that their technique accurately reconstructs MAP Kinase signaling networks in *Saccharomyces cerevisiae*. This approach should enhance the ability to model signaling networks and to discover new components of known networks. More generally, it provides a method for synthesizing molecular data, either individual transcript abundance measurements or pairwise protein interactions, into higher level structures, such as pathways and networks.

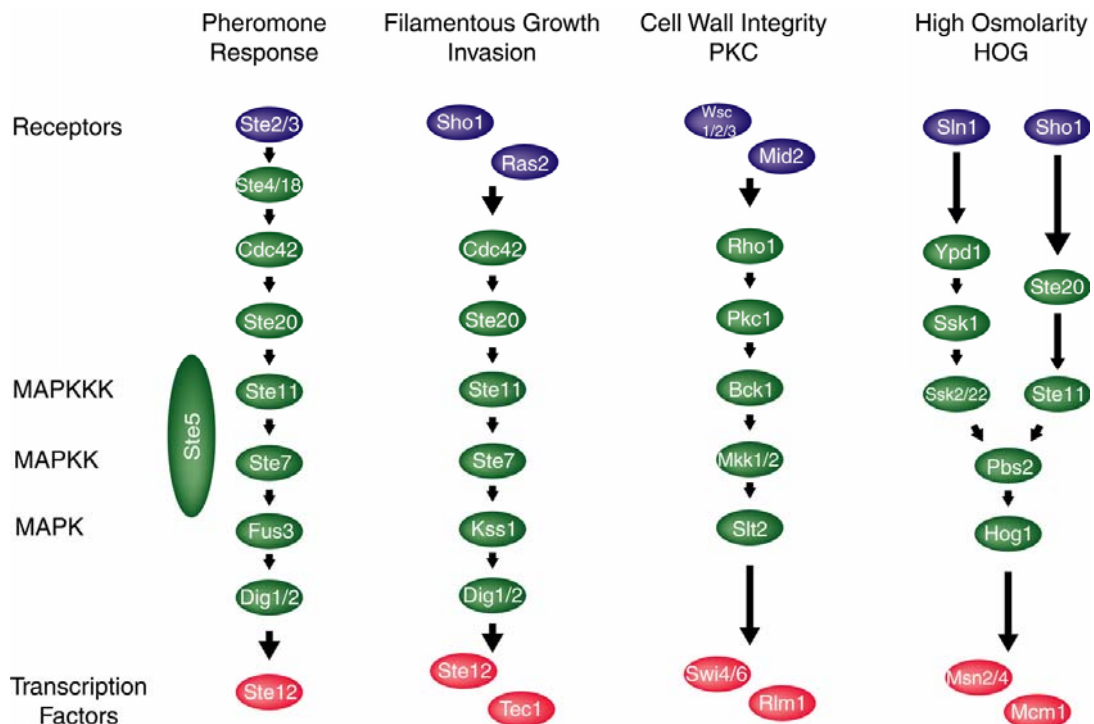


Figure 4.5 MAPK signal transduction pathways in yeast (Roberts *et al.*, 2000).

The proposed approach is calibrated using the yeast MAPK (mitogen-activated protein kinases) pathways involved in pheromone response, filamentous growth, and maintenance of cell wall integrity (**Figure 4.5**). These pathways are activated by G protein-coupled receptors and characterized by a core cascade of MAP kinases that activate each other through sequential binding and phosphorylation reactions; they are among the most thoroughly studied net works in yeast and are therefore excellent benchmarks against which to test our

approach. As shown in **Figure 4.5**, membrane proteins are depicted in blue, transcription factors in red, and intermediate proteins in green. Figure is adapted from [120].

Signaling pathways have been an active research area in recent history. There are many studies in which signaling pathways were modeled using various approaches. Previously, signaling pathways were modeled through modular kinetic simulations of biochemical networks and detailed integration of biochemical properties of the pathways [12]. In another study, Bayesian Networks were applied to multi-variable cell data to infer signaling pathways [13]. Correlating cancer based mRNA expression levels, autocrine receptor signaling loops were also discovered [14]. Another approach to model cellular pathways was developed based on perturbations of critical pathway components [15]. These were analyzed using DNA microarrays, quantitative proteomics, and databases of known physical interactions.

4.3 Motivation and Specific Aim

Protein phosphorylation catalyzed by kinase plays crucial regulatory role in intracellular signal transduction that is achieved by networks of proteins and small molecules that transmit information from the cell surface to the nucleus, where they ultimately effect transcriptional changes. How differential responses are generated by these networks is not obvious nor is the reason cells evolved a complicated mechanism for transducing signals. Thus, a full understanding of the mechanism of intracellular signal transduction remains a major challenge in cellular biology. Manning *et al.* have identified 518 human kinase genes, the so-called “kinome”, that provides a starting point for comprehensive analysis of protein phosphorylation networks.

Mass spectrometry-based proteomics have enabled the large-scale mapping of *in vivo* phosphorylation sites. However, only 20% of the experimentally verified phosphorylation sites have the annotation of catalytic kinases. To fully investigate how protein kinases regulate the intracellular processes, it is necessary to comprehensively and accurately identify the kinase-specific substrates. Therefore, we propose a method, RegPhos, incorporates computational model with protein associations (protein-protein interactions, functional associations, and subcellular localization) for identifying the catalytic kinase for each phosphoprotein with experimental phosphorylated sites. To observe the expressed relationship between kinase and substrate, the gene expression microarray data is adopted to observe the expression of kinase and substrate genes in specific conditions, for instance,

the normal tissue and cancerous tissue.

With the increasing number of in vivo phosphorylation sites have been identified, the desire to map the phosphorylation network of protein kinase and substrate has motivated. The experimental kinase-specific substrates, ultimately, need to be combined by systems biology analysis, which translates the separate, large-scale datasets into signaling networks. Therefore, we incorporated the experimentally verified kinase-substrate interactions with computationally identified kinase-substrate interactions to construct the intracellular phosphorylation network starting from receptor kinases to transcription factors, associated with the formation of protein subcellular localization. Moreover, the experimental expression evidence, such as gene microarray data and mass spectra, are adopted to validate the syn-expression of the constructed kinase-substrate phosphorylation network with statistical significance.

4.4 Materials

To construct the intracellular phosphorylation network between protein kinase and substrate, we propose a method, namely RegPhos, which incorporates computational models with protein associations (protein-protein interaction, functional associations, and protein subcellular localizations) for assigning the potential kinase for the experimental phosphorylation sites without annotated catalytic kinase. Moreover, the gene expression microarray data is adopted to validate the syn-expression of kinase and substrate.

4.4.1 Protein Kinase and Phosphorylation Site Resource

The experimental verified phosphorylation sites are extracted from dbPTM which has integrated version 7.0 of Phospho.ELM [2], release 55.0 of UniProtKB/Swiss-Prot [3], and version 1.0 of PHOSIDA [45]. As shown in **Table 4.1**, Phospho.ELM contains 16428 experimental phosphorylation sites within 4026 phosphoproteins, Swiss-Prot contains 24328 experimental phosphorylation sites within 8606 phosphoproteins, and PHOSIDA consists of 6600 in vivo phosphorylation sites within 2244 phosphoproteins. Especially, Human Protein Reference Database (HPRD), which integrates a wealth of information relevant to the function of human proteins in health and disease, is integrated in this work. Data pertaining to thousands of protein-protein interactions, posttranslational modifications, enzyme/substrate

relationships, disease associations, tissue expression, and subcellular localization were extracted from the literature for a non-redundant set of 25661 human proteins. In release 7.0 of HPRD, there are totally 16972 PTMs within 2830 protein entries, of 7438 PTMs are phosphorylation sites within 1774 proteins.

Because this work focuses on constructing human phosphorylation network, the phosphorylation sites in human proteins are represented in **Table 4.1**. After removing the redundant data among these databases, the number of human phosphorylation sites and phosphoprotein are 19817 and 5083, respectively.

Table 4.1 Statistics of integrated experimental protein phosphorylation site databases.

Database	Version	All species		Human	
		Number of phosphoprotein	Number of phosphosite	Number of phosphoprotein	Number of phosphosite
Phospho.ELM	7.0	4,026	16,428	3,354	11,278
UniProtKB/Swiss-Prot	55.0	8,606	24,328	3,746	11,862
PHOSIDA	1.0	N/A	N/A	2,212	8,969
HPRD	7.0	-	-	1,774	7,438
Combined (NR)	-	-	-	5,083	19,817

Abbreviation: NR, non-redundant.

Manning *et al.* [95] have identified 518 known protein kinase genes been identified, of which 71 have not previously been reported or described as kinases, and we extend or correct the protein sequences of 56 more kinases. These human kinase annotations extracted from KinBase [95] are used to unify the kinase names among the external phosphorylation site databases which contain various names for a kinase. The 518 kinases are major nodes in the construction of human phosphorylation networks. Due to the classification of kinase identified by Manning *et al.*, 518 kinases are categorized by their annotated family or subfamily, including totally 221 kinase families¹⁰. Several representative kinase families are listed in **Table 4.2**. Because the collection of experimentally verified phosphorylated sites from PhosphoELM, UniProtKB/Swiss-Prot, and PHOSIDA involved in various species, the number of phosphorylated sites in each kinase family is calculated in human and other species, as well as the number of phosphorylated proteins.

¹⁰ 221 kinase families: <http://140.113.239.26/RegPhos/statistics.php>

Table 4.2 List of representative kinase families containing more than 10 substrates.

Group	Kinase family	Kinase subfamily	Description	Kinase member	Human		All species	
					Phosphosite	Phosphoprotein	Phosphosite	Phosphoprotein
AGC	PKB		Protein kinase B	AKT1,AKT2,AKT3	89	63	114	79
AGC	GRK	GRK	G-protein coupled Receptor Kinase	GPRK7,RHOK,GPRK6,GPRK5,GPRK4	64	19	85	27
AGC	GRK	BARK	Beta Adrenergic Receptor Kinase	BARK1,BARK2	32	14	35	17
AGC	PKA		Protein kinase A	PKACa, PKACb, PKACg	232	151	458	286
AGC	PKC		Protein kinase C	PKCh, PKCa, PKCb, PKCd, PKCe, PKCg, PKCi, PKCt, PKCz	280	168	485	274
Atypical	PIKK	ATM	Ataxia telangiectasia mutated	ATM,ATR	67	34	102	63
CAMK	CAMK2		CAMK family 2	CaMK2a,CaMK2b,CaMK2g,CaMK2d	56	36	119	77
CK1	CK1		Cell Kinase 1	CK1a,CK1d,CK1e,CK1g2,CK1g3,CK1a2,CK1g1	63	33	101	52
CMGC	CDK	CDK	Cyclin Dependent Kinase	CDK4,CDK5,CDK6,CDK8,CDK9,CDK10,CDK11	64	34	123	66
CMGC	CDK	CDK7	Cyclin Dependent Kinase subfamily 7	CDK7	17	11	30	22
CMGC	CDK	CDC2	Cell Division Control 2	CDC2,CDK2,CDK3	226	95	328	138
CMGC	CK2		Casein kinase II	CK2a1,CK2a2,CK2a1-rs	241	123	368	192
CMGC	MAPK		Mitogen Activated Protein Kinase	Erk1(MAPK3),Erk2(MAPK1),Erk3(MAPK6),Erk4(MAPK4),Erk5(MAPK7),Erk7(MAPK15),JNK1(MAPK8),JNK2(MAPK9),JNK3(MAPK10),NLK,p38a(MAPK14),p38b(MAPK11),p38g(MAPK12),p38d(MAPK13)	248	140	333	192
CMGC	MAPK	JNK	JNK subfamily of MAPK	JNK1(MAPK8),JNK2(MAPK9),JNK3(MAPK10)	47	27	66	40
CMGC	MAPK	p38	p38 subfamily of MAPK	p38a(MAPK14),p38b(MAPK11),p38g(MAPK12),p38d(MAPK13)	62	35	66	38
CMGC	MAPK	ERK	Extracellular signal-Regulated protein Kinase	Erk1(MAPK3),Erk2(MAPK1),Erk3(MAPK6),Erk4(MAPK4),Erk5(MAPK7),Erk7(MAPK15)	138	88	178	112

Other	AUR	Aur	Aurora Kinase	AurA, AurB, AurC	42	19	57	28
Other	IKK		I kappa Kinase	IKKa,IKKb,IKKe,TBK1	43	12	49	16
TK	Abl		Abelson murine leukemia homolog	ABL1(Abl),ABL2(ARG)	36	26	56	39
TK	EGFR		Epidermal Growth Factor Receptor	EGFR,ErbB2,ErbB3,ErbB4	48	22	67	29
TK	InsR		Insulin Receptor and associated Kinases	INSR,IRR	30	9	46	14
TK	Src	Lck	Proto-oncogene tyrosine-protein kinase Lck	LCK	48	25	64	36
TK	Src	LYN	Tyrosine-protein kinase LYN	LYN	33	20	50	28
TK	Src	Src	Proto-oncogene tyrosine-protein kinase Src	SRC	108	68	171	101
TK	Syk	SYK	Spleen tyrosine kinase	SYK	38	17	51	22
TK	Syk	ZAP70	70 kDa zeta-associated protein, Syk-related tyrosine kinase	ZAP70	16	8	20	9
TK	Tec		Tec protein tyrosine kinase family	TXK,TEC,ITK,BMX,BTK	26	13	30	14



4.4.2 Protein-Protein Interaction Databases

To enhance the identification of kinase substrates, the physical protein-protein interaction data is used to explore the predictive accuracy in the proposed method. This work extract human protein-protein interactions from DIP [121, 122], MINT [123], IntAct [124], and HPRD [51], as shown in **Table 4.3**. The Database of Interacting Proteins¹¹ (DIP) is a database that documents experimentally determined protein-protein interactions. It provides the scientific community with an integrated set of tools for browsing and extracting information about protein interaction networks. As of April 2008, the DIP catalogs approximately 56000 unique interactions among 19000 proteins from > 180 organisms; the vast majority from yeast, *Helicobacter pylori* and human. Tools have been developed that allow users to analyze, visualize and integrate their own experimental data with the information about protein-protein interactions available in the DIP database. Because the reliability of experimental evidence varies widely, methods of quality assessment have been developed and utilized to identify the most reliable subset of the interactions. This CORE set of DIP can be used as a reference when evaluating the reliability of high-throughput protein-protein interaction data sets, for development of prediction methods, as well as in the studies of the properties of protein interaction networks.

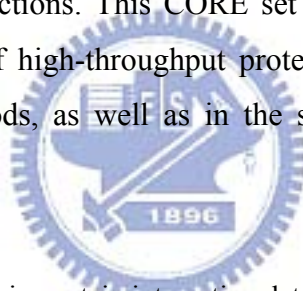


Table 4.3 Statistics of integrated protein-protein interaction databases.

Database	Data source	Version	All species		Human	
			Number of protein	Number of interaction	Number of protein	Number of interaction
DIP	Physical interaction	2008-04-30	19,765	56,493	1,224	1,794
MINT	Physical interaction	2008-04-30	28,817	105,899	6,106	20,832
IntAct	Physical interaction	2008-04-11	63,121	163,909	~15,000	~28,500
HPRD	Physical interaction	Release 7.0	-	-	38,167	25,661

The Molecular INTERaction database¹² (MINT) aims at storing, in a structured format, information about molecular interactions (MIs) by extracting experimental details from work published in peer-reviewed journals. At present the MINT team focuses the curation work on physical interactions between proteins. Genetic or computationally inferred interactions are

¹¹ DIP URL: <http://dip.doe-mbi.ucla.edu>

¹² MINT URL: <http://mint.bio.uniroma2.it/mint/>

not included in the database. Up to April 30th 2008, there are totally 105899 interactions between 28,817 proteins. The new version of MINT is based on a completely remodeled database structure, which offers more efficient data exploration and analysis, and is characterized by entries with a richer annotation. The whole dataset can be freely accessed online in both interactive and batch modes through web-based interfaces and an FTP server. MINT now includes, as an integrated addition, HomoMINT, a database of interactions between human proteins inferred from experiments with ortholog proteins in model organisms.

IntAct¹³ is an open source database and software suite for modeling, storing and analyzing molecular interaction data. The data available in the database originates entirely from published literature and is manually annotated by expert biologists to a high level of detail, including experimental methods, conditions and interacting domains. At present, the database features over 163000 binary interactions extracted from over 2200 scientific publications and makes extensive use of controlled vocabularies. The web site provides tools allowing users to search, visualize, and download data from the repository.

4.4.3 Functional Association Databases

To capture the biological context of a substrate, we use a network of functional associations extracted from the STRING¹⁴ database [114]. This network is based on four fundamentally different types of evidence: genomic context (gene fusion, gene neighborhood, and phylogentic profiles), primary experimental evidence (physical protein interactions and gene coexpression), manually curated pathway databases, and automatic literature mining. Information on protein-protein interactions is still mostly limited to a small number of model organisms, and originates from a wide variety of experimental and computational techniques. The underlying infrastructure includes a consistent body of completely sequenced genomes and exhaustive orthology classifications, based on which interaction evidence is transferred between organisms. Although primarily developed for protein interaction analysis, the resource has also been successfully applied to comparative genomics, phylogenetics and network studies, which are all facilitated by programmatic access to the database backend and the availability of compact download files. As of release 7.1, STRING has almost doubled to

¹³ IntAct URL: <http://www.ebi.ac.uk/intact>

¹⁴ STRING URL: <http://string.embl.de>

373 distinct organisms, and contains more than 1.5 million proteins for which associations have been pre-computed.

Table 4.4 Statistics of integrated functional association databases.

Database	Data source	Version	All species		Human	
			Number of protein	Number of interaction	Number of protein	Number of interaction
STRING	Physical interaction and functional association	7.1	~1,500,000	77,147,159	16,050	1,397,066
GOA	Cellular component, molecular function, and biological process	2008-04-30	3,977,963	29,269,200	35,423	183,316

The Gene Ontology Annotation (GOA) database¹⁵ [125] aims to provide high-quality electronic and manual annotations to the UniProt Knowledgebase (Swiss-Prot, TrEMBL and PIR-PSD) using the standardized vocabulary of the Gene Ontology (GO) [126]. As a supplementary archive of GO annotation, GOA promotes a high level of integration of the knowledge represented in UniProt with other databases. GOA provides annotated entries for nearly 60,000 species and is the largest and most comprehensive open-source contributor of annotations to the GO Consortium annotation effort. By integrating GO annotations from other model organism groups, GOA consolidates specialized knowledge and expertise to ensure the data remain a key reference for up-to-date biological information. Furthermore, the GOA database fully endorses the Human Proteomics Initiative by prioritizing the annotation of proteins likely to benefit human health and disease. The GOA data set can be used to enhance the annotation of particular model organism or gene expression data sets, although increasingly it has been used to evaluate GO predictions generated from text mining or protein interaction experiments. Up to April 30th 2008, GOA totally stores 29,269,200 functional associations between 3,977,963.

4.4.4 Protein Subcellular Localization Databases

The eukaryotic cell is a composite system internally subdivided into membrane-enveloped compartments that perform particular functions [41]. Every subcellular compartment contains specific proteins, including enzymes, synthesized in the cytoplasm and translocated into the

¹⁵ GOA URL: <http://www.ebi.ac.uk/GOA>

locations, where they carry out functional patterns. Therefore, knowing the localization of every protein is important for elucidating its interactions with other molecules and for understanding its biological function. Some major constituents of eukaryotic cells are: extracellular space, cytoplasm, nucleus, mitochondria, Golgi apparatus, endoplasmic reticulum (ER), peroxisome, vacuoles, cytoskeleton, nucleoplasm, nucleolus, nuclear matrix and ribosomes. The proteins which are involved in similar biological functions are closely located in the same subcellular localization. Protein phosphorylation plays crucial regulatory role in intracellular signal transduction networks from the receptors of cell surface to the transcription factors of nucleus, where they ultimately effect transcriptional changes. In order to identify phosphorylation cascade, the information of protein subcellular localization is used in the construction of phosphorylation network. **Table 4.5** shows the list of public databases of protein subcellular localization, including LOCATE [127], DBSubLoc [128], Organelle DB [129], and PSORTdb [130].

Table 4.5 List of public databases of protein subcellular localization.

Database	Species	Statistics	Statistics of human
LOCATE	Human and mouse	122,765 protein isoforms	64,637 protein isoforms
DBSubLoc	All	64,051 proteins	30,633 proteins
Organelle DB	138 organisms	30,188 genes	4,233 genes
PSORTdb	Bacterial	~2000 proteins	-
UniProtKB	All	487,934 proteins	16,052 proteins

LOCATE¹⁶ [127] is a curated, web-accessible database that houses data describing the membrane organization and subcellular localization of mouse and human proteins. The membrane organization is predicted by the high-throughput, computational pipeline MemO [131]. The subcellular locations were determined by a high-throughput, immunofluorescence-based assay and by manually reviewing peer-reviewed publications. The database now contains high-quality localization data for 20% of the mouse proteome and general localization annotation for nearly 36% of the mouse proteome. The proteome annotated in LOCATE is from the RIKEN FANTOM Consortium Isoform Protein Sequence [132] sets which contains 58128 mouse within 29682 transcript units and 64637 human protein isoforms within 26583 transcript units.

DBSubLoc¹⁷ [128] is a database of protein subcellular localization which contains

¹⁶ LOCATE URL: <http://locate.imb.uq.edu.au/>

¹⁷ DBSubLoc URL: <http://www.bioinfo.tsinghua.edu.cn/dbsubloc.html>

proteins from primary protein database SWISS-PROT and PIR. By collecting the subcellular localization annotation, the information are classified and categorized by cross references to taxonomies and Gene Ontology database. Based on sequence alignment, non-redundant subsets of the database have been built, which may provide useful information for subcellular localization prediction. The database now contains >60,000 protein sequences including approximately 30,000 protein sequences in the non-redundant data sets.

Organelle DB¹⁸ [129] is a web-accessible relational database presenting a supplemented catalog of organelle-localized proteins and major protein complexes. Since its release in 2004, Organelle DB has grown by 20% to encompass over 30,000 proteins from 138 eukaryotic organisms. Each protein in Organelle DB is presented with its subcellular localization, primary sequence and a detailed description of its function, as available. All records in Organelle DB have been annotated using controlled vocabulary from the Gene Ontology consortium. Protein localization data are inherently visual, and Organelle DB is a significant repository of biological images, housing 1500 micrographs of yeast cells carrying stained proteins. Organelle View offers a dimensional representation of a yeast cell; users can search Organelle View for proteins of interest, and the organelles housing these proteins will be highlighted in the cell image.

PSORTdb¹⁹ [130] is a web-accessible database of SubCellular Localization (SCL) for bacteria that contains both information determined through laboratory experimentation and computational predictions. The dataset of experimentally verified information (approximately 2000 proteins) was manually curated by us and represents the largest dataset of its kind. Earlier versions have been used for training SCL predictors, and its incorporation now into this new PSORTdb resource, with its associated additional annotation information and dataset version control, should aid researchers in future development of improved SCL predictors. The second component of this database contains computational analyses of proteins deduced from the most recent NCBI dataset of completely sequenced genomes. Analyses are currently calculated using PSORTb, the most precise automated SCL predictor for bacterial proteins. Both datasets can be accessed through the web using a very flexible text search engine, a data browser, or using BLAST, and the entire database or search results may be downloaded in various formats.

Moreover, UniProtKB [48] also have the annotation of subcellular localization for

¹⁸ Organelle DB URL: <http://organelledb.lsi.umich.edu>

¹⁹ PSORTdb URL: <http://db.psорт.org/>

protein entries in Swiss-Prot and TrEMBL. Based on manually curated literatures, there are 487934 proteins contain the annotation of subcellular localization, 16052 of them are human proteins.

4.4.5 Gene Expression Database

The Gene Expression Omnibus²⁰ (GEO) [133] at the National Center for Biotechnology Information (NCBI) is the largest fully public repository for high-throughput molecular abundance data, primarily gene expression data. The database has a flexible and open design that allows the submission, storage and retrieval of many data types. These data include microarray-based experiments measuring the abundance of mRNA, genomic DNA and protein molecules, as well as non-array-based technologies such as serial analysis of gene expression (SAGE) and mass spectrometry proteomic technology. GEO currently holds over 235,000 submissions for over 100 organisms. In this work, the human gene expression samples of Affymetrix GeneChip Human Genome U133 Array Set HG-U133A platform (GPL96) and Affymetrix GeneChip Human Genome U133 Plus 2.0 Array (GPL570), consisting of 22283 probe set for 12678 genes and 54681 probe sets for 18433 genes, respectively, are used to explore the coexpression of kinase and substrate genes.

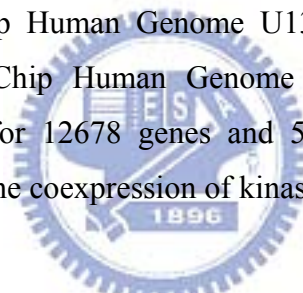


Table 4.6 List of human gene microarray platform of GEO used in this work.

Platform	Title	Type	Probe sets	Genes	Date	Samples
GPL96	Affymetrix GeneChip Human Genome U133 Array Set HG-U133A	in situ oligonucleotide	22283	12678	Feb. 19, 2002	16033
GPL570	Affymetrix GeneChip Human Genome U133 Plus 2.0 Array	in situ oligonucleotide	54681	18433	Nov 07, 2003	14046

The U133 set includes 2 arrays with a total of 44928 entries and was indexed 29-Jan-2002. The set includes over 1,000,000 unique oligonucleotide features covering more than 39,000 transcript variants, which in turn represent greater than 33,000 of the best

²⁰ GEO URL: <http://www.ncbi.nlm.nih.gov/geo>

characterized human genes. Sequences were selected from GenBank, dbEST, and RefSeq. Sequence clusters were created from Build 133 of UniGene (April 20, 2001) and refined by analysis and comparison with a number of other publicly available databases including the Washington University EST trace repository and the University of California, Santa Cruz golden-path human genome database (April 2001 release). In addition, ESTs were analyzed for untrimmed low-quality sequence information, correct orientation, false priming, false clustering, alternative splicing and alternative polyadenylation.

Complete coverage of the Human Genome U133 Set plus 6,500 additional genes for analysis of over 47,000 transcripts. All probe sets represented on the GeneChip Human Genome U133 Set are identically replicated on the GeneChip Human Genome U133 Plus 2.0 Array. The sequences from which these probe sets were derived were selected from GenBank®, dbEST, and RefSeq. The sequence clusters were created from the UniGene database (Build 133, April 20, 2001) and then refined by analysis and comparison with a number of other publicly available databases, including the Washington University EST trace repository and the University of California, Santa Cruz Golden-Path human genome database (April 2001 release). In addition, there are 9,921 new probe sets representing approximately 6,500 new genes. These gene sequences were selected from GenBank, dbEST, and RefSeq. Sequence clusters were created from the UniGene database (Build 159, January 25, 2003) and refined by analysis and comparison with a number of other publicly available databases, including the Washington University EST trace repository and the NCBI human genome assembly (Build 31).

4.5 Method

How can one bridge the gap from transcript abundances and protein-protein interaction data to pathway models? To construct the complete phosphorylation network, the comprehensive and reliable information of protein kinase-substrate interactions is needed. This work proposes a method, RegPhos, not only integrates the experimentally verified phosphorylation sites which have the annotation of catalytic kinase, but also incorporates the computational models with protein associations to identify the catalytic kinase for the experimental phosphorylation sites which have not the annotation of kinase. The system architecture of RegPhos is shown in **Figure 4.6**, including the collection of experimental kinase-substrate resource, identification of kinase-substrate interactions, integration of gene expression data, and construction of

intracellular phosphorylation networks. Microarray expression data is then used to rank all paths according to the degree of similarity in the expression profiles of pathway members.

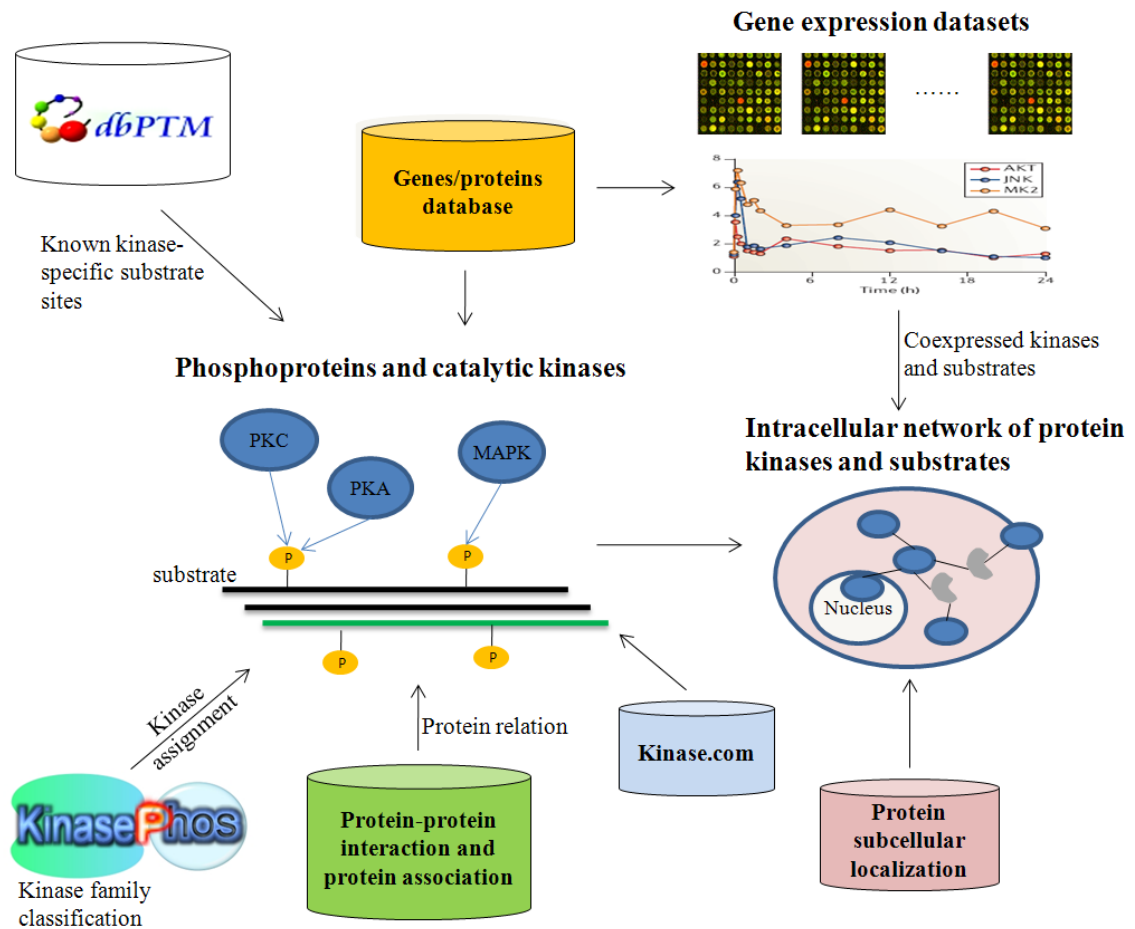


Figure 4.6 System architecture of RegPhos.

4.5.1 Identification of Kinase-Substrate Interactions

With the integration of experimental phosphorylation sites, there are totally 18,823 experimental verified phosphorylation sites within 4983 human proteins, of 3535 phosphorylation sites (~20%) have the annotation of catalytic kinases. Most of the experimental phosphorylation sites (~80%) do not have the annotation of catalytic kinases. Although most of human phosphorylation sites in PHOSIDA have the annotation of kinases based on the consensus motif of kinases, the annotations are still needed to be verified by more information, such as protein-protein interactions, subcellular localization, and functional associations. Therefore, the enriched kinase-substrate interactions could be used to construct the complete intracellular phosphorylation networks.

To identify the catalytic kinase for each experimentally verified phosphorylation site without annotated kinase, we propose a method which incorporates computational models with protein-protein interaction, protein subcellular localization, and gene expression data for assigning the potential kinase. The system flow is shown in **Figure 4.7**, including two types of measurement. First is the model-based measurement for kinase-specific phosphorylation site prediction (as described previously in Chapter 3). Second is using the functional association such as protein-protein interaction, functional association, and subcellular co-localization to identify the catalytic kinase for a substrate protein. Finally, the experimentally validated phosphorylation sites with annotated catalytic kinase are used to evaluate the performance and decide the cutoff.

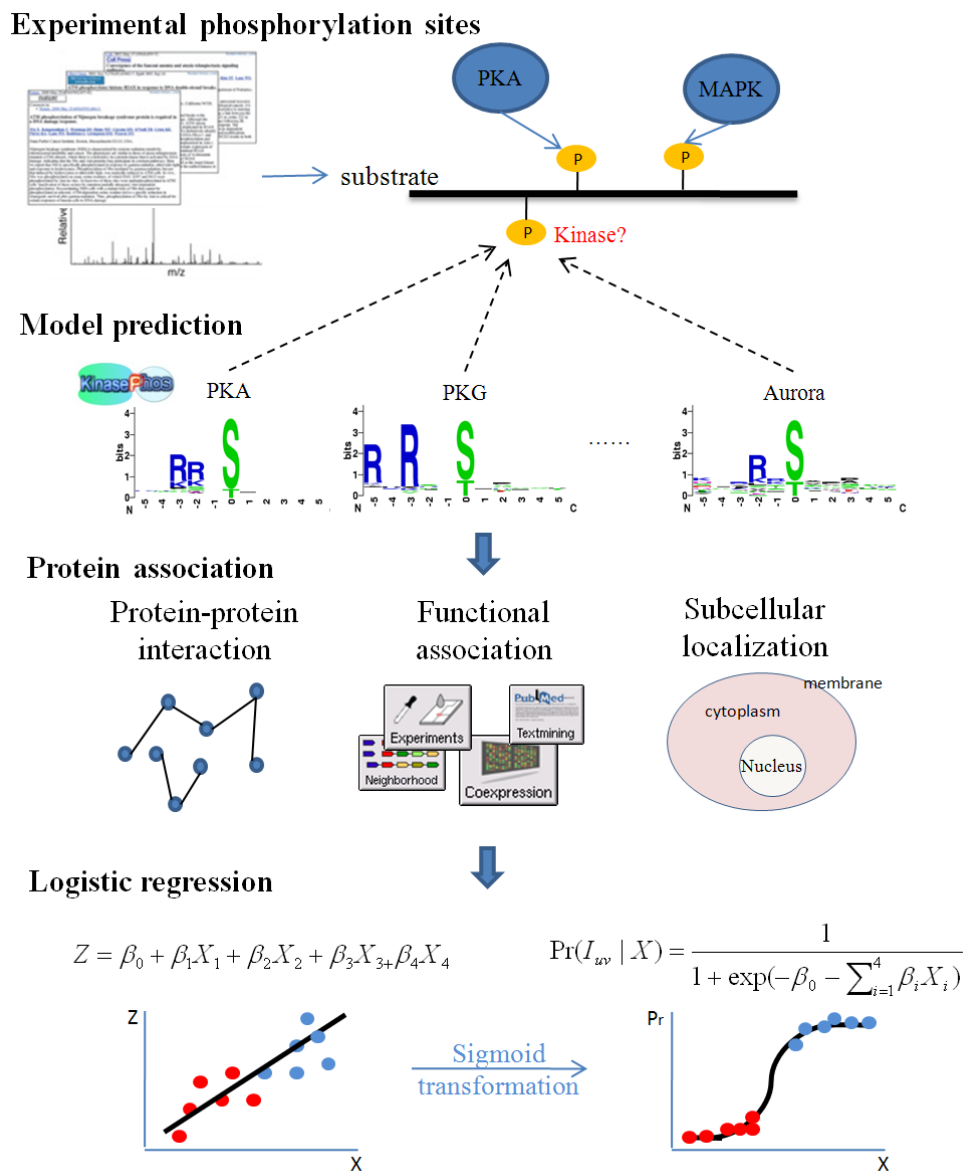


Figure 4.7 System flow of identification of kinase-substrate interactions.

4.5.1.1 Computational Annotation of Kinase-Specific Phosphorylation

Sites

The proposed kinase-specific phosphorylation site prediction method, namely KinasePhos, is used to identify the candidate kinase families for the phosphorylation sites without annotated catalytic kinases. As illustrated in Chapter 3, the support vector machine (SVM) is applied to create the computational models with the encoded amino acids and structural features, secondary structure and accessible surface area. With the binary classification, the concept of SVM is mapping the input samples onto a higher dimensional space through a kernel function, and then seeking a hyper-plane that discriminates the two classes with maximal margin and minimal error. A public SVM library, namely LibSVM [110], is adopted to train the predictive model with the positive and negative training sets which are encoded according to different types of training features. Radial basis function (RBF) $K(S_i, S_j) = \exp(-\gamma \|S_i - S_j\|^2)$ is selected as the kernel function of SVM.

There are more than 100 kinase families been constructed the predictive models, whose average predictive accuracy is approaching 90%. In general, each kinase-specific phosphorylation site prediction model has a cut-off value of score and use the value to decide whether a phosphorylation site is catalyzed by the kinase family. However, a phosphorylation may be predicted as the substrate site that was catalyzed by more than one kinase family because several kinase families have the similar substrate specificity. For instance, as shown in **Figure 4.7**, the amino acid motifs of PKA, PKG and Aurora, which have conserved arginine (R) in upstream position -2 or -3 of phosphorylated site, are similar. There may be a lot of false positives in the kinase assignment of phosphorylation site. Therefore, it needs the experimental evidence of functional association, such as protein-protein interaction or signaling pathway, to be used to reduce the false positive predictions.

4.5.1.2 Exploration of Protein Associations

To explore the possibility of using functional association to enhance the identification of kinase-specific substrates, we developed an integrative computational approach, RegPhos, which combines computational kinase-specific phosphorylation site prediction models and protein association networks to predict which protein kinases target experimentally identified phosphorylation sites in vivo (**Figure 4.7**). The association context for each substrate is

investigated by the information of manually curated protein-protein interaction databases (physical protein interaction assays, curated pathway, cooccurrence in literature abstracts), cellular colocalization, and mRNA coexpression signature. This approach captures both direct and indirect interactions; for example, phosphorylation events mediated by scaffolds are predicted, as the scaffolding protein provides a path in the indirect connection between the substrate and kinase. The use of indirect links between kinases and their substrates enables unobvious predictions that would be very difficult to spot by manually inspecting the available evidence.

Exploring the Protein-Protein Interactions

To identify the direct and indirect connection between kinase and substrate, a graph searching algorithm, Breadth-first search (BFS), is adopted. BFS is one of the simplest algorithms for searching a graph and the archetype for many important graph algorithms. Given a graph $G = (V, E)$ where V represents the set of proteins and E is the set of physical interactions between proteins, and a distinguished source vertex s , BFS systematically explores the edges of G to discover every vertex that is reachable from s . The brief procedure of BFS, contain four major stpes, is listed as bellow:

1. Put the source node on the queue.
2. Pull a node from the beginning of the queue and examine it.
 - If the searched element is found in this node, quit the search and return a result.
 - Otherwise push all the (so-far-unexamined) successors (the direct child nodes) of this node into the end of the queue, if there are any.
3. If the queue is empty, every node on the graph has been examined -- quit the search and return "not found".
4. Repeat from Step 2.

The breadth-first search (BFS) procedure assumes that the input graph $G = (V, E)$ is represented using adjacency lists. It maintains several additional data structures with each vertex in the graph. The pseudocode of BFS is shown in **Figure 4.8**, which is implemented in C programming language. The depth of interacting neighbor is decided by the investigation of experimentally verified kinase-substrate interactions.

```

void BFS(VLink G[], int v) {
    int w;
    VISIT(v);           /*visit vertex v*/
    visited[v] = 1;    /*mark v as visited : 1 */
    ADDQ(Q, v);
    while(!EMPTYQ(Q)) {
        v = DELQ(Q);   /*Dequeue v*/
        w = FIRSTADJ(G, v); /*Find first neighbor, return -1 if no neighbor*/
        while(w != -1) {
            if(visited[w] == 0) {
                VISIT(w); /*visit vertex w*/
                ADDQ(Q, w); /*Enqueue current visited vertex w*/
                visited[w] = 1; /*mark w as visited*/
            }
            w = NEXTADJ(G, v); /*Find next neighbor, return -1 if no neighbor*/
        }
    }
}

```

Figure 4.8 Pseudocode of breadth-first search (BFS) algorithm.

Evaluating the Functional Association between Kinase and Substrate

To capture the biological context of a substrate, we use a network of functional associations extracted from the STRING²¹ database [114]. This network is based on four fundamentally different types of evidence: genomic context (gene fusion, gene neighborhood, and phylogenetic profiles), primary experimental evidence (physical protein interactions and gene coexpression), manually curated pathway databases, and automatic literature mining. Referred to NetworKIN [112], it was found that physical protein interactions play the dominant role among the primary experimental data, whereas gene coexpression contributes only very little. As the curated pathway databases generally contain few errors, a confidence score of 0.9 is assigned to this type of evidence, Physical protein interactions were imported and merged from numerous repositories, and the reliability of each individual interaction was assessed based on the promiscuity of the interaction partners using a scoring schemes described elsewhere (Von Mering et al., 2005).

Moreover, the Gene Ontology Annotation (GOA) database [125], which aims to provide high-quality electronic and manual annotations to the UniProt Knowledgebase using the standardized vocabulary of the Gene Ontology (GO) [126], is used to investigate the functional association between substrate and candidate kinase. By integrating GO annotations from other model organism groups, GOA consolidates specialized knowledge and expertise to ensure the data remain a key reference for up-to-date biological information. There are three

²¹ STRING URL: <http://string.embl.de>

major types of annotation in GO, including cellular component, molecular function, and biological process. Each GO term specifies a specific cellular component, molecular function, or biological process. To evaluate the similarity of functional association between substrate and candidate kinase proteins, the Cosine similarity, which is usually adopted in text mining, is used. With the task of text clustering, Cosine similarity is a simple measure endows documents with the same composition but different sizes to be treated identically which makes this the most popular measure for clustering text documents [134]. Due to this property, term vectors can be normalized to the unit sphere. Given a kinase k with GO term vector $\vec{V}_k = (G_1, G_2, \dots, G_m)$, where m is the number of GO term related to kinase k . If there are n candidate substrates S_1, S_2, \dots, S_n with GO term vectors $\vec{X}_i = (G_{i1}, G_{i2}, \dots, G_{im})$, $i = 1, \dots, n$, the Cosine similarity of GO terms between kinase k and substrate S_i is calculated as follows:

$$\text{sim}(\vec{V}_k, \vec{X}_i) = \frac{\vec{V}_k^T \vec{X}_i}{\|\vec{V}_k\| \cdot \|\vec{X}_i\|}.$$

A schematic representation of Cosine similarity is illustrated in **Figure 4.9**, the Cosine similarity between two GO term vectors is identical to calculate the cosine angle between two vectors. As the angle between the vectors shorten, the cosine angle approaches 1, meaning that the two vectors are getting closer, meaning that the similarity of whatever is represented by the vectors increases. Therefore, the cosine similarity between vectors A and B is calculated as follows:

$$\text{Sim}(A, B) = \cos \theta = \frac{A \bullet B}{\|A\| \cdot \|B\|} = \frac{x_1 \cdot x_2 + y_1 \cdot y_2}{\sqrt{x_1^2 + y_1^2} \cdot \sqrt{x_2^2 + y_2^2}}.$$

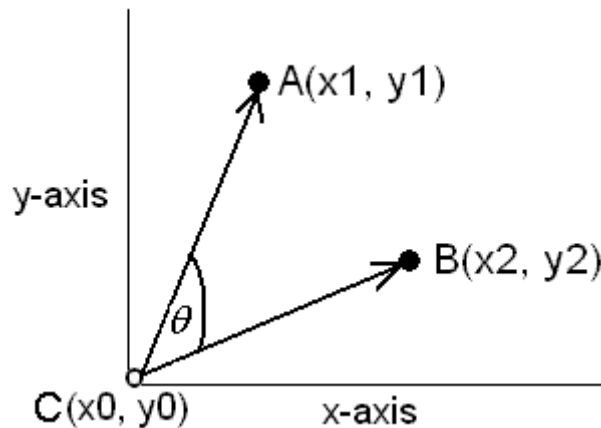


Figure 4.9 Schematic representation of Cosine similarity between two vectors.

Checking the Subcellular Co-localization of Kinase and Substrate

The eukaryotic cell is a composite system internally subdivided into membrane-enveloped compartments that perform particular functions [41]. Some major constituents of eukaryotic cells are: extracellular space, cytoplasm, nucleus, mitochondria, Golgi apparatus, endoplasmic reticulum (ER), peroxisome, vacuoles, cytoskeleton, nucleoplasm, nucleolus, nuclear matrix and ribosomes. The proteins which are involved in similar biological functions are closely located in the same subcellular localization. Knowing the localization of every protein is important for elucidating its interactions with other molecules and for understanding its biological function. Protein phosphorylation plays crucial regulatory role in intracellular signal transduction networks from the receptors of cell surface to the transcription factors of nucleus, where they ultimately effect transcriptional changes. In order to identify phosphorylation cascade, the information of protein subcellular localization is used in the construction of phosphorylation network.

4.5.1.3 Logistic Regression

Logistic regression was adopted to evaluate the confidence value of protein-protein interaction [135]. In this study we utilized a modified version of the Sharan *et al.* [136] method for evaluating the confidence values of the discovered kinase-substrate interactions. Since the framework is based on the functional enrichment of proteins, we have based the confidence evaluation on this methodology. In the logistic regression model, we incorporate four sets of variables for a given interaction set, including (1) the prediction score of the kinase-specific SVM model, (2) the depth of interaction between kinase and substrate was observed, (3) the confidence score of the STRING functional association, and (4) the binary (0/1) protein subcellular localization data of interacting pairs. Here in addition to the previously presented first three random variables [136], we also incorporate the protein subcellular localization data into the logistic model. This is very straightforward since in most of the signaling cascades the proteins would transmit the signal from the membrane, where the signal is initiated, towards to the nucleus, where the final product is transcribed. Although proteins travel in a cell and can coexist in multiple compartments, this classification may eliminate the false negatives.

Given the four variables, $X = (X_1, X_2, X_3, X_4)$, represented the four types of variables, and the positive and negative training data sets, a linear model $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

could be optimized the parameters β_0, \dots, β_4 to maximize the likelihood of training data. β_0 is called the "intercept" and $\beta_1, \beta_2, \beta_3$, and β_4 , are called the "regression coefficients" of X_1, X_2, X_3 , and X_4 , respectively. the probability of a kinase-substrate interaction $Pr(I_{uv})$ under the logistic distribution is given by

$$\Pr(I_{uv} | X) = \frac{1}{1 + \exp(-\beta_0 - \sum_{i=1}^4 \beta_i X_i)},$$

where β_0, \dots, β_4 , are parameters of the distribution. The positive and negative can be used to define the cutoff value of confidence score which can reach the best classifying accuracy.

4.5.1.4 Performance Evaluation

To evaluate the predictive performance of the proposed method, the experimentally verified kinase-specific phosphorylation sites are used to cutoff value and test the prediction accuracy. The following measures of predictive performance of the trained models are defined:

$$\text{Precision (Pre)} = \frac{TP}{TP + FP}, \text{ Sensitivity (Sn)} = \frac{TP}{TP + FN}, \text{ Specificity (Sp)} = \frac{TN}{TN + FP},$$

$$\text{Accuracy (Ac)} = \frac{TP + TN}{TP + FP + TN + FN}, \text{ where } TP, TN, FP \text{ and } FN \text{ are true positive, true negative, false positive and false negative, respectively.}$$

The proposed method is test by the experimentally verified phosphorylation sites of PKC, CDK, PIKK, and INSR kinase families from HPRD database. Moreover, the kinase groups with similar motif of substrate sites are used to test the predictive performance, including arginine-directed kinase families PKA, PKB, PKC, and Aurora from HPRD database.

4.5.2 Construction of Phosphorylation Network

After the identification of catalytic kinase of experimental phosphorylation sites, the enriched kinase-substrate interactions are used to construct the complete phosphorylation network. Graph-based method is adopted to formalize the construction of intracellular phosphorylation network to a shortest path problem in graph theory. Moreover, the cellular localization of proteins is used to constrain the search of phosphorylation network.

Graph-based Definition of Phosphorylation Network

Due to the graph-based method, the intracellular protein phosphorylation network are visualized as an directed graph $G = (V, E)$, where $x, y \in V$ and $(x, y) \in E$. Let x and y represent kinase and substrate proteins, respectively, and $(x, y) \in E$ represent a phosphorylation interaction when kinase x phosphorylates substrate y . In this work, V refers to all human proteins in UniProtKB [48], and E refers to all kinase-substrate interactions in knowledgebase including experimentally verified kinase-specific phosphorylations and RegPhos-identified kinase-substrate interactions. Each edge has the weighted score from 0 to 1, 1 for the experimentally verified kinase-substrate phosphorylation and logistic regression probability value for the RegPhos-identified kinase-substrate interaction.

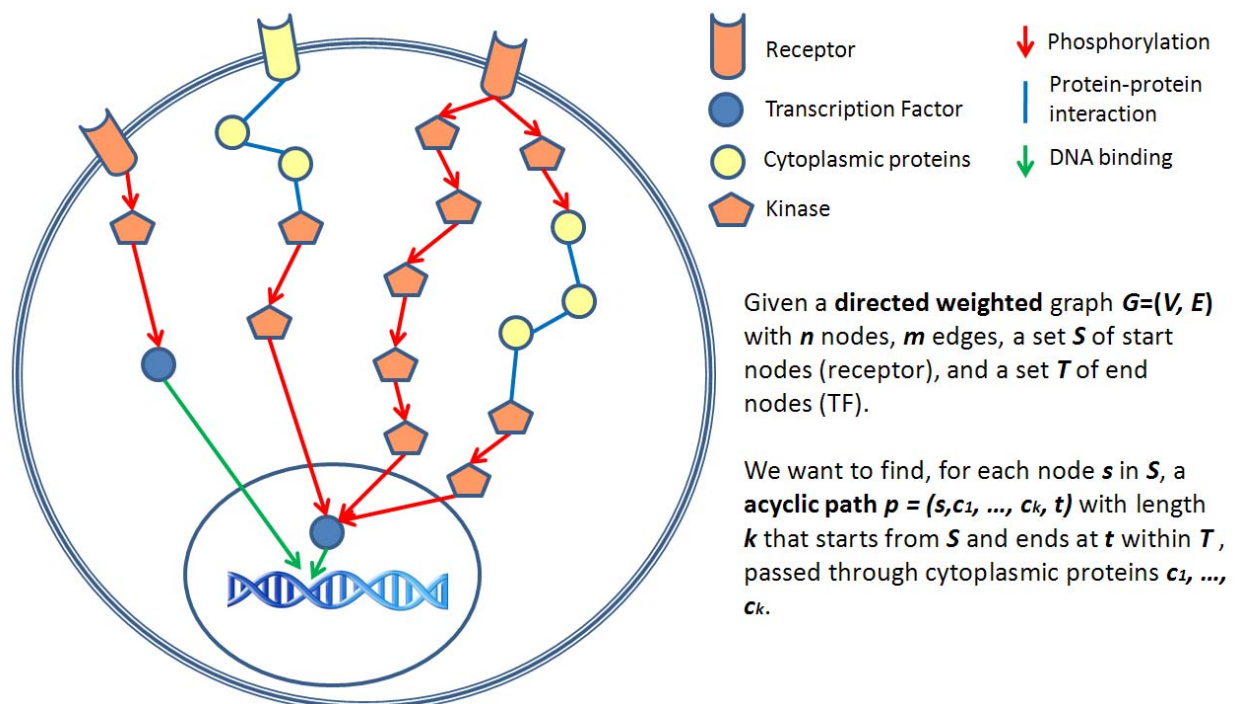


Figure 4.10 Schematic representation of phosphorylation network.

Identification of Signaling Pathway from Receptor Kinase to Transcription Factor

Due to the annotation of cellular localization databases, there are 84 cell membrane-associated kinases being the start points of the phosphorylation networks. With the annotation of TRANSFAC version 11.0 [137], there are 1364 transcription factors in human. To identify the phosphorylation networks starting from membrane receptor to transcription factor in nucleus, the graph-based definition can be refined as follows: given a **directed weighted** graph $G=(V, E)$ with n nodes, m edges, a set S of start nodes (receptor), and a set T of end nodes (TF). As shown in **Figure 4.10**, we want to find, for each node s in S , a **acyclic path** $p = (s, c_1, \dots, c_k, t)$ with length k that starts from S and ends at t within T , passed through cytoplasmic proteins c_1, \dots, c_k . We restrict attention to simple paths that was constrained the order of occurrence of proteins in a defined path length 8.



4.5.3 Expression Profile of Kinase and Substrate Genes

How can one bridge the gap from transcript abundances and protein-protein interaction data to pathway models? Clustering expression data into groups of genes that share profiles is a proven method for grouping functionally related genes, but does not order pathway components according to physical or regulatory relationships. Here we present an automated approach for modeling signal transduction networks in human by integrating protein-protein interaction, protein subcellular localization, and gene expression data. Our program draws all possible linear paths of a specified length through the interaction map starting at any membrane protein and ending on any transcription factor. Microarray expression data is then used to rank all paths according to the degree of similarity in the expression profiles of pathway members. Linear pathways that have common starting points and endpoints and the highest ranks are then combined into the final model of the branched networks.

4.5.3.1 Normalization of Gene Expression Samples

All statistical analyses were accomplished using R program language. Gene expression data were processed and normalized using Bioconductor Affy package²², based on the Robust Multichip Average (RMA) method [138] for single-channel Affymetrix chips. All 22,283 probe sets based on RMA summary measure were used in class comparison analyses.

4.5.3.2 Distance Function

Two major distance function were used to measure how closely related are kinase and substrate genes:

Euclidean Distance

This kind of distance strategy calculates the length of two separate points in n-directional space by their absolute differences. For example, Euclidean distance is measure by following definition:

Given two points $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$, Euclidean distance is

²² Bioconductor package: <http://www.bioconductor.org/>

$$\sqrt{\sum_{i=1}^n (a_i - b_i)^2}.$$

Pearson Correlation Coefficient

Contrasting to Euclidean distance, Pearson correlation coefficient accounts for the trends of two expression profile. For instance, Pearson correlation measures the similarity in shape between two profiles by the following formula:

Given two points $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$, Pearson correlation coefficient similarity is $\frac{1}{n} \sum_{i=1}^n \left(\frac{a_i - \bar{a}}{\sigma_a} \right) \left(\frac{b_i - \bar{b}}{\sigma_b} \right)$, where \bar{a} and \bar{b} are the mean of A and B, and σ_a and σ_b are the standard deviation of A and B. Pearson correlation distance is

$$1 - \frac{1}{n} \sum_{i=1}^n \left(\frac{a_i - \bar{a}}{\sigma_a} \right) \left(\frac{b_i - \bar{b}}{\sigma_b} \right).$$

These two kinds of distance strategies will lead to different clustering results. As shown in **Figure 4.11**, different distances will render different classifications because we are asking for grouping based on different features: trends in the case of correlation and absolute differences in the case of Euclidean distance.

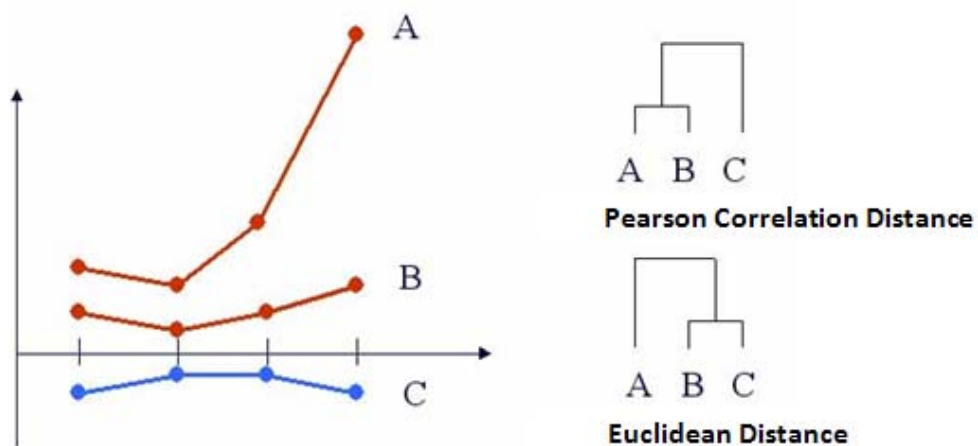


Figure 4.11 Comparison of clustering results between Euclidean distance and Pearson correlation distance strategies.²³

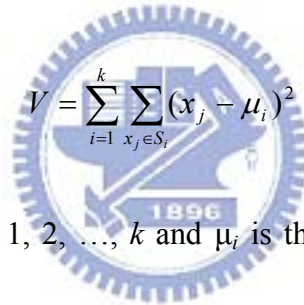
²³ The figure was obtained from <http://gepas.bioinfo.cipf.es/cgi-bin/tutoX?c=clustering/clustering.config>

4.5.3.3 Clustering of Syn-expressed Genes

Clustering aims to group data with similar characteristics together. Some clustering algorithms are usually used in gene expression analysis, including hierarchical clustering and k-means clustering. Gene coexpression was measured by calculating the Pearson correlation coefficient between two genes across all data sets in the Gene Expression Omnibus repository for the organism in question.

K-means Clustering Algorithm

The k-means algorithm (J.A. Hartigan and M.A. Wong, 1979) is an algorithm to cluster n objects based on attributes into k partitions, $k < n$. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data. It assumes that the object attributes form a vector space. The objective it tries to achieve is to minimize total intra-cluster variance, or, the squared error function


$$V = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

where there are k clusters S_i , $i = 1, 2, \dots, k$ and μ_i is the centroid or mean point of all the points .

The most common form of the algorithm uses an iterative refinement heuristic known as Lloyd's algorithm. Lloyd's algorithm starts by partitioning the input points into k initial sets, either at random or using some heuristic data. It then calculates the mean point, or centroid, of each set. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters, and algorithm repeated by alternate application of these two steps until convergence, which is obtained when the points no longer switch clusters (or alternatively centroids are no longer changed).

Hierarchical Clustering

In hierarchical clustering, a series of partitions takes place, which may run a single cluster containing all objects to n clusters, each contains a single object. Hierarchical clustering is subdivided into agglomerative methods, which proceed by series of fusions of the n objects into group, and divisive method, which separate n objects successively into finer groupings.

One of the simplest agglomerative hierarchical clustering methods is single linkage, also known as the nearest neighbor technique. The feature of the method is that distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered. The minimum value of these distances is said to be the distance between two clusters. At each stage of hierarchical clustering, the clusters whose distance is minimal are merged.



4.6 Results

The investigation of subcellular localization, protein interacting neighbor, and expression profiles of protein kinases and their substrate are illustrated as follows. The predictive performance of the proposed method is also discussed in this section. Finally, the statistics of the identified kinase-substrate interactions are listed.

Table 4.7 Statistics of integrated experimental protein phosphorylation sites.

Database	Version	All species		Human	
		Number of phosphoprotein	Number of phosphosite	Number of phosphoprotein	Number of phosphosite
Phospho.ELM	7.0	4,026	16,428	3,354	11,278
UniProtKB/Swiss-Prot	55.0	8,606	24,328	3,746	11,862
HPRD	7.0	-	-	1,774	7,438
Combined (NR)	-	-	-	4,825	18,031

Abbreviation: NR, non-redundant.

4.6.1 Protein Kinases, Phosphoproteins, and Interacting Proteins

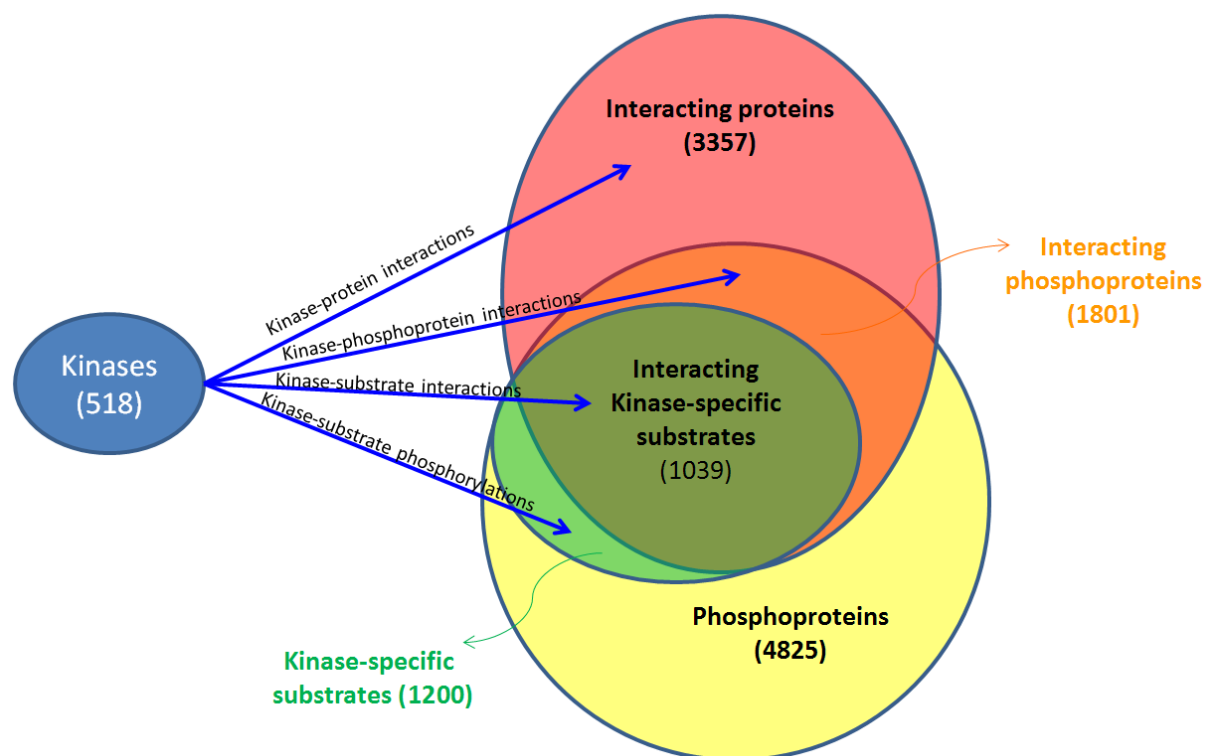


Figure 4.12 The schematic representation of kinase, interacting proteins, and phosphoproteins.

There are totally 518 known kinase genes identified by Manning *et al.* [95]. With the collected experimentally verified phosphorylation sites from version 7.0 of Phospho.ELM, release 55.0 of UniProtKB/Swiss-Prot, and release 7.0 of HPRD, there are totally 18031 experimental phosphorylation sites within 4825 human phosphoproteins (**Table 4.7**). With the annotation of catalytic kinases, there are 3550 kinase-specific phosphorylation sites within 1200 human phosphoproteins, catalyzed by 356 protein kinases. With the collected protein-protein interactions from DIP, MINT, IntAct, and HPRD, most of the 518 kinases (~80%) have interacting proteins. As shown in **Figure 4.12**, there are four types of interactions between kinases, interacting proteins and substrates, including kinase-protein interactions, kinase-phosphoprotein interactions, kinase-substrate phosphorylations, and kinase-substrate interactions.

Table 4.8 Statistics of kinases and their interacting proteins.

Interaction Type	Number of interactions	Number of kinase	Number of interacting proteins
Kinase-protein interactions	10,056	451	3,357
Kinase-phosphoprotein interactions	7,155	430	1,801
Kinase-substrate phosphorylations	6,015	356	1,200
Kinase-substrate interactions	5,443	342	1,039

The number of kinases and interacting proteins in the four types of interactions is listed in **Table 4.8**. There are totally 3357 proteins interacting with 451 human protein kinases, of 1801 interacting proteins contain experimental phosphorylation sites (interacting phosphoproteins). The 7155 kinase-phosphoprotein interactions could be used to indicate the potential kinase-specific substrates.

Table 4.9 Statistics of kinases and their interacting proteins and functionally associated proteins.

Interaction Type	Number of interactions	Number of kinase	Number of interacting proteins
Kinase-protein interactions	11,235	453	3,569
Kinase-phosphoprotein interactions	7,838	434	1,872
Kinase-substrate phosphorylations	6,015	356	1,200
Kinase-substrate interactions	5,922	352	1,056

To fully investigate the interacting proteins of human proteins kinases, the functional

association database, STRING, is integrated for enhance the protein interaction resource. This association is based on four fundamentally different types of evidence: genomic context (gene fusion, gene neighborhood, and phylogentic profiles), primary experimental evidence (physical protein interactions and gene coexpression), manually curated pathway databases, and automatic literature mining. The protein associations, whose confidence score are more than 0.9, are adopted. The number of kinases and interacting proteins in the four types of interactions is listed in **Table 4.9**.

Table 4.10 The protein interacting neighbor of several representative human kinase families.

Kinase family	Kinase members	Number of substrates	Number of interacting proteins			
			Depth=1	Depth=2	Depth=3	Depth> 4
PKA	PKACa, PKACb, PKACg	194	123	39	25	7
PKC	PKCh, PKCa, PKCb, PKCd, PKCe, PKCg, PKCi, PKCt, PKCz	231	175	41	6	9
CK2	CK2a1, CK2a2, CK2b, CK2a1-rs	158	120	28	9	1
CDK	CDC2, CDK2, CDK3, CDK4, CDK5, CDK6, CDK7, CDK8, CDK9, CDK10, CDK11,	157	135	15	2	5
Src	Src	92	68	19	3	2
EGFR	EGFR	27	25	0	1	1
InsR	InsR	14	12	0	1	1

This approach captures both direct and indirect interactions. The use of indirect links between kinases and their substrates enables unobvious predictions that would be very difficult to spot by manually inspecting the available evidence. To investigate the interacting depth of indirect connection between kinase and substrate, the number of interacting substrates in each kinase group is observed in different interacting depth. As shown in **Table 4.10**, the number of interacting substrates in PKA, PKC, CK2, CDK, Src, EGFR, and InsR families are listed with various interacting depth. For instance, PKA family, consisting of

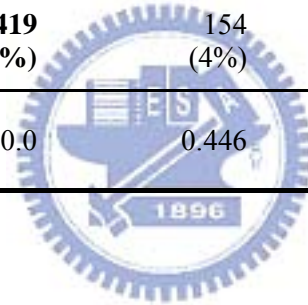
PKACa, PKACb and PKACg kinases, has 123 (63%) directly interacting substrates. About 37% of PKA-specific substrates are indirect connection to PKA kinases. Base on the statistics of interacting depth between kinase and substrate, most of the substrates (~95%) are connecting to kinases within interacting depth 3.

4.6.2 Subcellular Localization of Protein Kinases and Substrates

Protein phosphorylation can control intracellular translocation and trafficking of proteins. Due to the annotation of cellular component from the collected protein subcellular localization databases, the cellular distribution of human proteins including kinase and substrate proteins can be investigated in detail. There are 18609 human proteins in UniProt KB/Swiss-Prot, 13146 of which contain the localization information. **Table 4.11** shows the statistics of human protein cellular localization categorized mainly into nucleus, cytoplasm, Golgi apparatus, mitochondrion, endoplasmic reticulum (ER), and cell membrane. With the annotation from UniProtKB, there are 339 kinases have the information of subcellular localization, which are mainly located in nucleus (34.5%), cytoplasm (52.8%), and cell membrane (31%). Most of the kinases located in cell membrane are receptor tyrosine kinase (RTK). However, many kinases not only locate in a specific cellular localization, like PKA, PKC, MAPK kinase groups, which translocate between cytoplasm and nucleus. Moreover, the substrate proteins are mainly located in nucleus (44.6%) and cytoplasm (36.7%).

Table 4.11 Subcellular localization of human proteins, kinases and substrates.

Localization	all	Nucleus	Cytoplasm	Golgi	Mitochondrion	ER	Cell membrane	Other
Human proteins	13146	3953 (30%)	3082 (23.4%)	482 (3.7%)	733 (5.6%)	576 (4.4%)	1932 (14.7%)	4103 (31.2%)
Kinase	339	117 (34.5%)	179 (52.8%)	10 (2.9%)	9 (2.7%)	6 (1.8%)	105 (31%)	84 (24.8%)
P-value		0.0415	0.0	0.801	0.997	0.997	0.0	0.996
Substrate	3863	1724 (44.6%)	1419 (36.7%)	154 (4%)	131 (3.4%)	136 (3.5%)	439 (11.4%)	723 (18.7%)
P-value		0.0	0.0	0.446	0.984	0.803	0.976	0.999



To easily categorize the subcellular localization for kinase and substrate, the localization of substrates is classified into nuclear and cytoplasmic substrates. The subcellular localizations of each human kinase-specific substrate proteins extracted from Phospho.ELM and UniProtKB/Swiss-Prot are schematically represented in **Figure 4.13**. We mapped localizations from Swiss-Prot to the kinase-specific substrates, which resulted in 3863 phosphoproteins that are described as localizing to either the cytoplasm or the nucleus. Based on these statistics, we found 33 kinase groups that show a statistically significant preference for either cytoplasmic or nuclear substrates. For membrane-associated kinases (such as EGFR, INSR, and the Src family kinases), it almost exclusively was cytoplasmic substrates. Although receptor tyrosine kinases (RTKs) can occasionally translocate to the nucleus, there are very few nuclear substrates. However, we cannot exclude the possibility that the available phosphorylation data sets do not currently cover the cellular states where RTKs are active in the nucleus.

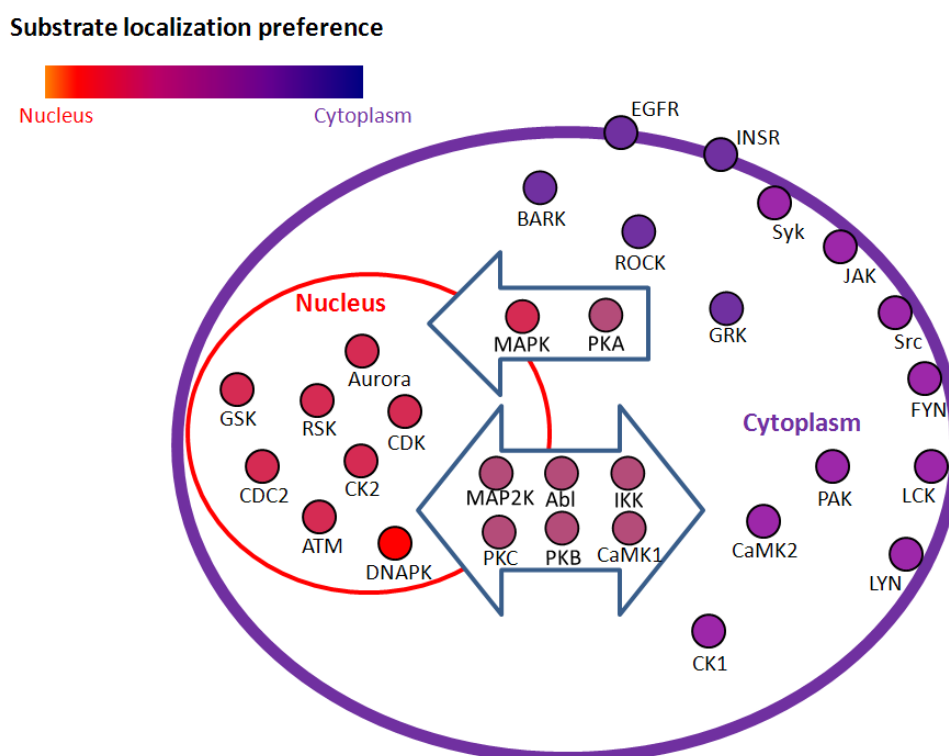


Figure 4.13 Subcellular localization preference of kinase family and their substrates.

In contrast, we find no kinases that are predicted to exclusively phosphorylate nuclear proteins. For the kinase groups that are primarily localized to the nucleus (DNAPK, ATM, CDK, CDC2, CK2, RSK, GSK and Aurora), were about 2-fold more nuclear than cytoplasmic targets. There are at least three possible explanations for this: (1) all nuclear kinases are synthesized in the cytosol and may phosphorylate cytosolic proteins prior to entering the

nucleus, (2) nuclear kinases may have access to cytosolic substrates during mitosis when the nuclear membrane is absent, and (3) many kinases may shuttle between the nucleus and the cytosol. This is exemplified by PKA and MAPK family, which, upon activation, translocate from the cytosol to the nucleus or the perinuclear region. However, PKB, PKC, Abl, IKK, and MAP2K families are both fairly pleiotropic kinases, which in the phosphorylation network show a weak preference for cytoplasmic substrates. The statistics of substrate localization preference of kinase families is listed in **Table 4.12**. The statistically significant (P -value < 0.05) localization preference of kinase family is marked in bold.

Table 4.12 Subcellular localization of human kinase-specific substrates.

Kinase group	Kinase localization	All substrates	Cytoplasmic substrates	Nuclear substrates	Cytoplasmic and Nuclear substrates
PKA	Cytoplasm, Nucleus	151	96	74	21
PKC	Cytoplasm, Nucleus	168	105	81	26
PKB	Cell membrane, Cytoplasm, Nucleus	63	49	32	19
GRK	Cytoplasm	19	18	2	2
ROCK	Cytoplasm	15	15	1	1
BARK	Cytoplasm	14	14	1	1
CaMK2	Cytoplasm	36	29	11	6
CaMK1	Cytoplasm, Nucleus	14	5	8	2
CK1	Cytoplasm	33	29	14	10
ATM	Nucleus	34	11	32	9
DNAPK	Nucleus	13	3	12	2
RSK	Nucleus	31	15	25	9
CK2	Nucleus	123	46	91	17
CDK	Nucleus	121	34	79	30
CDC2	Nucleus	95	37	66	17
GSK	Nucleus	34	15	23	9
MAPK	Cytoplasm, Nucleus	140	59	91	29
JNK	Cytoplasm, Nucleus	27	13	22	9
P38	Cytoplasm, Nucleus	35	15	22	4
ERK	Nucleus	88	41	63	18
Aurora	Nucleus	19	8	14	4
IKK	Cytoplasm, Nucleus	12	10	8	6
PAK	Cytoplasm	25	19	6	1
MAP2K	Cytoplasm, Nucleus	13	9	6	2
Abl	Cytoplasm, Nucleus	26	18	13	5
EGFR	Cell membrane Nucleus	22	18	0	4
InsR	Cell membrane	9	9	0	0
JAK	Membrane associated	17	17	6	6
Src	Membrane associated	68	61	22	16
FYN	Membrane associated	21	16	9	5
LCK	Membrane associated	25	22	1	1
LYN	Membrane associated	20	17	3	3
SYK	Membrane associated	17	15	1	1
Total		3863	1661	2195	612

Despite the caveats of possible biases in the various data sets, the putative kinase-substrate interactions are consistent with localization data for the substrates and kinases. The cell membrane-linked kinases show clear preference for cytoplasmic substrates, the predominantly nuclear kinases are biased toward nuclear substrates, and the kinases that shuttle between the cytosol and the nucleus exhibit a more even distribution of substrates.

4.6.3 Expression Analysis of Kinase and Substrate

In this work, the human gene expression samples of Affymetrix GeneChip Human Genome U133 Array Set HG-U133A platform (GPL96), consisting of 22283 probe set for 12678 genes, are used to explore the co-coexpression analysis of kinase and substrate genes. However, the first problem we faced is what kind of microarray experiment should be selected for investigating the co-expression of kinase and substrate genes. Without any specific interest and limitation, we decide to focus on the experimental series of microarray with the raw data. Totally 2714 samples within 98 experiment series (GSE), including *Large-scale analysis of the 79 human normal tissue transcriptome* (GSE1133), *Colon cancer progression* (GSE1323), *Lung tissue from smokers with severe emphysema* (GSE1650), *Lung cancer cell line response to motexafin gadolinium: time course* (GSE2189), *Epidermal growth factor effect on cervical carcinoma cell line: time course* (GSE6783), etc., were processed and normalized using Bioconductor Affy package, based on the Robust Multichip Average (RMA) method [138].

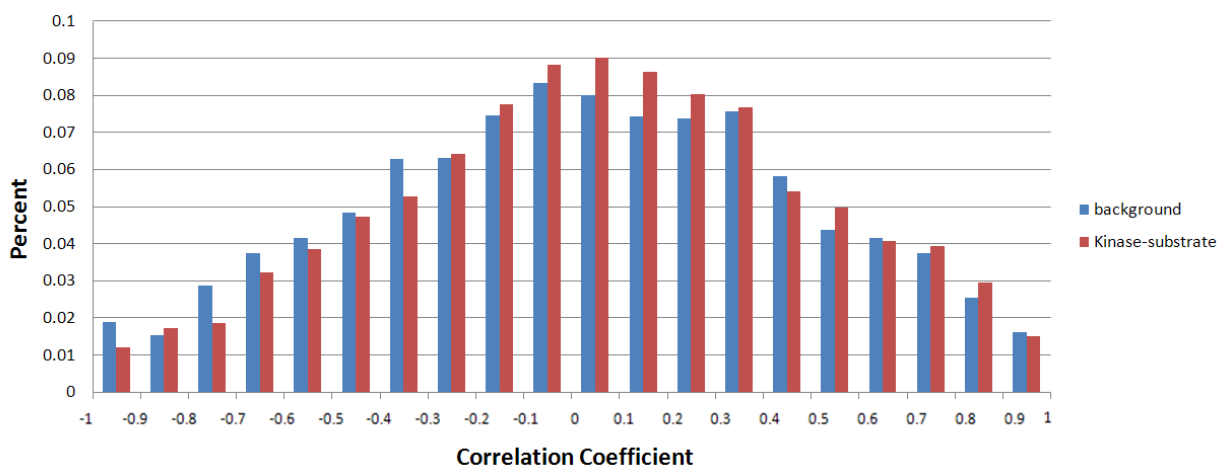


Figure 4.14 Comparison of Pearson correlation coefficient distribution between background gene pairs and kinase-substrate pairs.

Pearson correlation coefficient is used to analyze the expression pattern of two genes. To

investigate the statistically significant syn-expressed pair of kinase and substrate genes, all the pairs of genes are calculated for background correlation. However, it is time-expensive for calculating all pairs of genes. Therefore, the random sampling is adopted to extract 100,000 gene pairs as the background set for estimating the distribution of Pearson correlation coefficients of background gene pairs. All the 6015 experimentally verified kinase-substrate pairs are calculated the Pearson correlation coefficients. As shown in **Figure 4.14**, the distribution of correlation coefficients of background gene pairs is similar to normal distribution, based on central limit theorem. In the case of kinase-substrate pairs, the correlation distribution is slightly skew to right-side. It indicated that the kinases do not have high similarity of expression pattern to their substrates. The average correlation coefficients of background gene pairs and kinase-substrate pairs are 0.019 and 0.031.

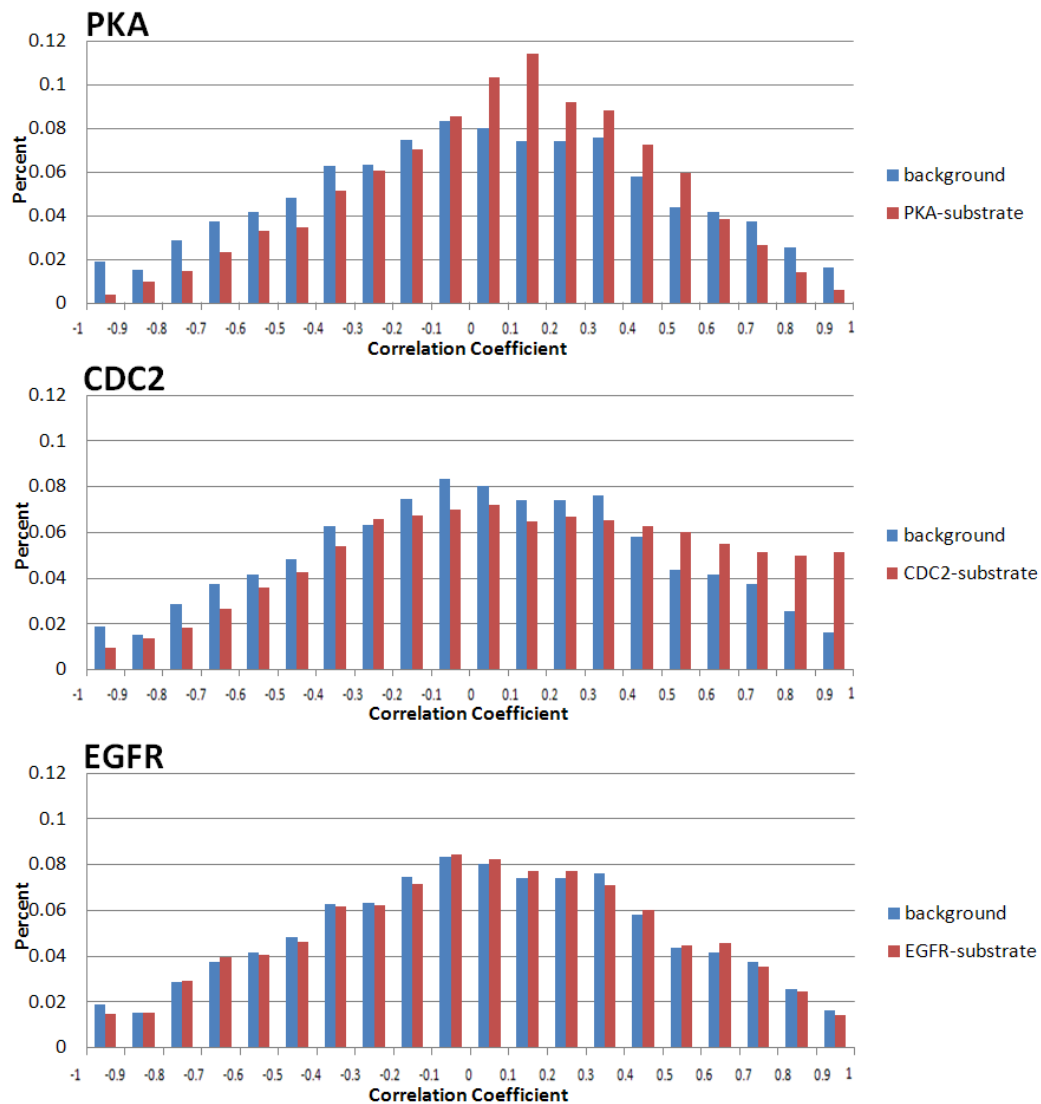


Figure 4.15 Distribution of Pearson correlation coefficients of PKA-substrate pairs, CDC2-substrate pairs, and EGFR-substrate pairs based on 98 microarray series.

The distribution of Pearson correlation coefficient of specific kinase-substrate pairs is also investigated. **Figure 4.15** shows the distribution of correlation coefficient of PKA-substrate pairs, CDC2-substrate pairs, and EGFR-substrate pairs, based on 98 microarray series. Most of the PKA-substrate pairs (40%) belong to the low positive correlation ($0 < r < 0.4$), with the average correlation coefficient 0.08. In particular, about 65% of CDC2-substrate pairs have the positive correlation, with $\sim 20\%$ high positive correlation ($r > 0.7$). The average correlation coefficient of CDC2-substrate pairs is 0.14. In the case of EGFR-substrate pairs, the distribution of correlation coefficient is similar to the distribution of all kinase-substrate pairs. The average correlation coefficient of EGFR-substrate pairs is 0.028.

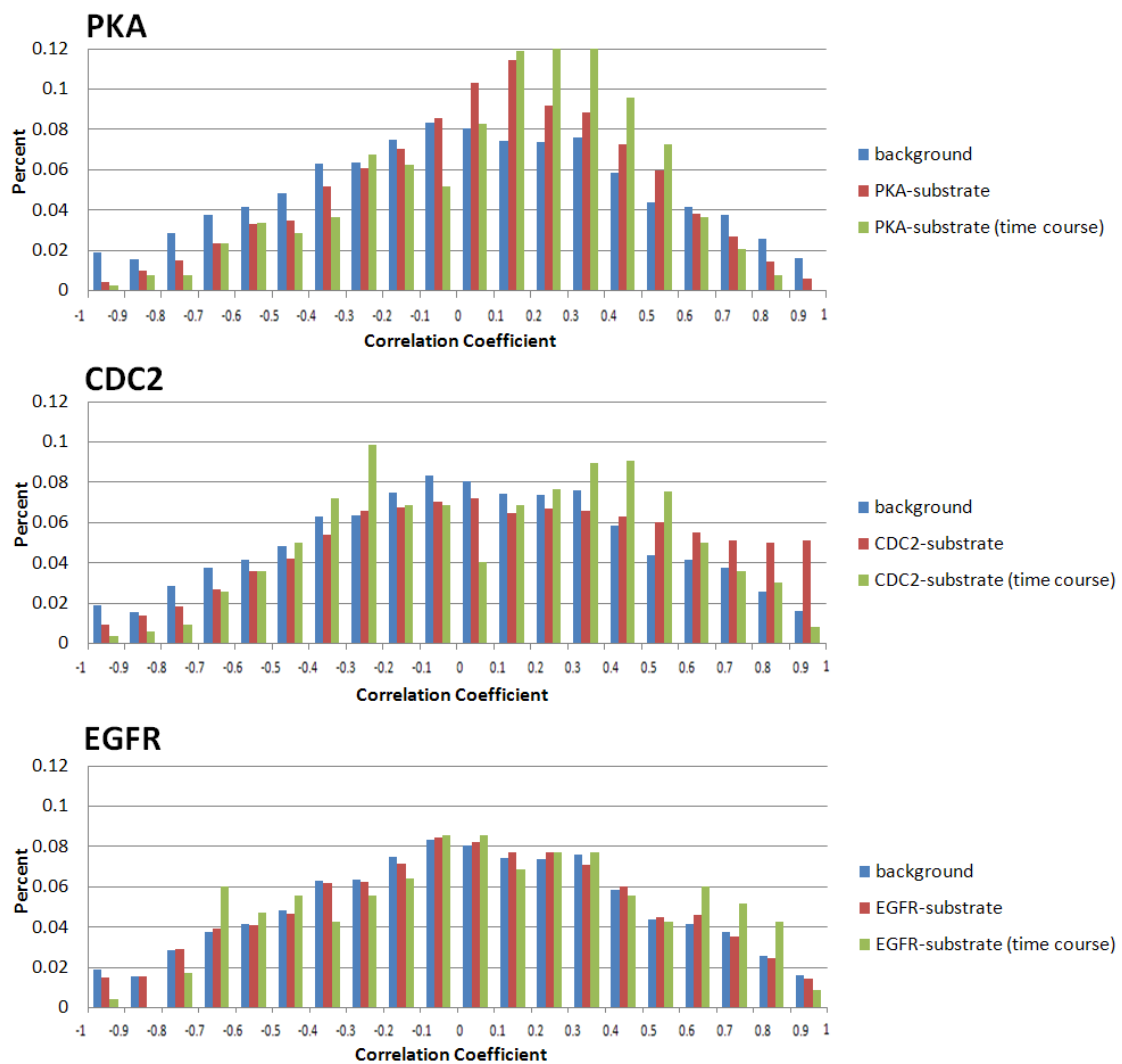
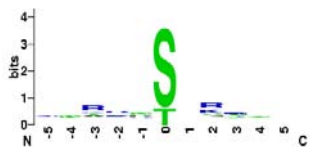

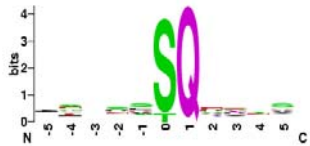
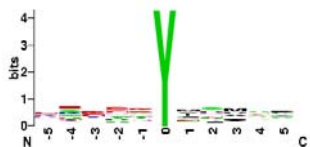


Figure 4.16 Distribution of Pearson correlation coefficients of PKA-substrate pairs, CDC2-substrate pairs, and EGFR-substrate pairs based on time-coursed microarray data.

Moreover, the distribution of Pearson correlation coefficient of specific kinase-substrate

pairs is investigated based on time-coursed microarray data. **Figure 4.16** shows the distribution of correlation coefficient of PKA-substrate pairs, CDC2-substrate pairs, and EGFR-substrate pairs based on 9 time-coursed microarray series, including *Esophageal cell response to low pH* (GSE2144), *Lung cancer cell line response to motexafin gadolinium* (GSE2189), *Cyanobacterial metabolite apratoxin A cytotoxic effect on colon adenocarcinoma cells* (GSE2742), *Interleukin 13 effect on bronchial cell line* (GSE3183), *Endotoxin effect on leukocytes* (GSE3284), *Blood response to various beverages* (GSE3846), *Androgen receptor modulator effect* (GSE4636), *Glucocorticoid receptor activation effect on breast cancer cells* (GSE4917), and *Epidermal growth factor effect on cervical carcinoma cell line* (GSE6783). The average correlation coefficient of PKA-substrate pairs is up to 0.12. The proportion of PKA-substrate pairs belonged to the low positive correlation ($0 < r < 0.4$) is increased from 40% to 45%. In the case of EGFR-substrate pairs, the average correlation coefficient of EGFR-substrate pairs is raised from 0.028 to 0.08. The proportion of EGFR-substrate pairs belonged to high positive correlation ($r > 0.6$) is approaching 16%. However, based on time-coursed microarray data, the average correlation coefficient of CDC2-substrate pairs is decreased to 0.10.

Table 4.13 Predictive performance of purely SVM-based prediction (KinasePhos).

Kinase family	Sequence logo	Number of positive data	Number of negative data	Pr	Sn	Sp	Acc
PKC		160	149	84.8	84.2	83.8	84.0
CDK		100	209	79.3	92.0	88.5	89.6
PIKK		37	272	60.0	89.1	91.9	91.5
INSR		12	297	14.5	75.0	82.2	81.9

Abbreviation: Pr, precision; Sn, sensitivity; Sp, specificity; Acc, accuracy.

4.6.4 Predictive Performance

To compare the predictive performance of RegPhos with NetworKIN [112], we also adopted the same data set to test the ability of RegPhos to correctly predict which kinases are responsible for catalyzing each of 667 known phosphorylation sites from four well-annotated kinase families, including CDK, PKC, PIKK, and INSR from HPRD database. The classifying specificity of each pair of PKC, CDK, PIKK, and INSR families are listed in **Table 4.14**. As given in table, the number in the parenthesis besides the kinase name indicates the size of the positive set. For example, the first row gives that there are 160 phosphorylated sites in kinase PKA set. The sensitivity (S_n) of the PKA model is 84.2%. The specificity are given in the table, for instance, in the first row the specificity (S_p) of CDK, PIKK and INSR sets corresponding to the PKA model are 81.9%, 89.1% and 83.3%, respectively. Similarly, the cross specificity values among PKC, CDK, PIKK, and INSR are generally higher than 80%. However, the specificity of INSR model is slightly weak when differentiating PKC substrates from INSR substrates. The higher specificity the cross-validation, the less incorrect prediction of the phosphorylation sites in other groups.

Table 4.14 Cross classifying specificity among PKC, CDK, PIKK, and INSR families based on KinasePhos method.

	PKC (160)	CDK (100)	PIKK (37)	INSR (12)
PKC model	$S_n=84.2\%$	81.9%	89.1%	83.3%
CDK model	86.9%	$S_n=92.8$	94.6%	91.7%
PIKK model	89.4%	96.0%	$S_n=89.1\%$	91.7%
INSR model	77.5%	88.0%	86.5%	$S_n=75.0\%$

Using only computational model (KinasePhos), we obtained the predictive accuracies 84%, 89.6%, 91.5% and 81.9% in PKC, CDK, PIKK, and INSR, respectively. Although the kinase families used for benchmarking have by necessity been studied more than most kinases, the predictive power of the consensus sequence motifs for CDK, PKC, PIKK, and INSR are representative for many other kinase families. By incorporating contextual information of protein association, the prediction accuracy improves to 84.1%, 91.6%, 91.9% and 91.9% in PKC, CDK, PIKK and INSR, respectively, because of the improvement of specificity (see **Figure 4.17**). However, there are slight drops in predictive sensitivity. These results highlight the importance of including contextual information in identifying kinase-substrate relationships for experimentally verified phosphorylation sites without annotated catalytic kinases.

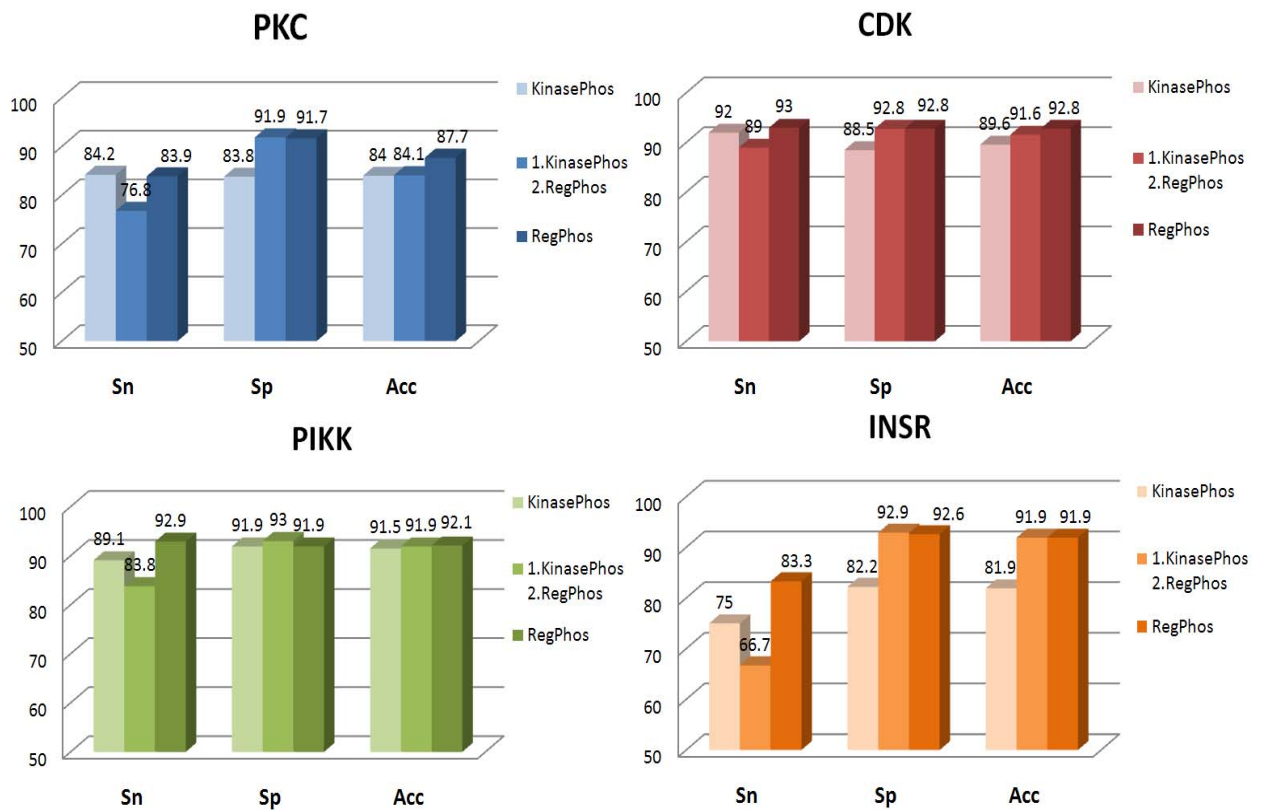


Figure 4.17 Effects of including protein associations.

4.6.5 Statistics of Discovered Kinase-specific Substrate Interactions

With the experimental verified kinase-specific phosphorylation sites extracted from version 7.0 of Phospho.ELM [2], release 55.0 of UniProtKB/Swiss-Prot [3], and release 7.0 of HPRD [51], there are 18031 experimentally verified human phosphorylation sites within 4825 phosphoproteins. Out of 3550 experimental sites have the annotation of catalytic kinases, which cover 356 kinases. In order to fully construct the intracellular phosphorylation networks, the 14481 experimental phosphorylation sites without annotated kinases are systematically discovered the catalytic kinases by the proposed method (RegPhos). In first step, 101 kinase group models (by support vector machine) with 89% overall predictive accuracy are used to scan the putative kinase-specific phosphorylation site. Secondly, the protein association including protein-protein interaction, functional association and cellular localization is adopted to help the discovery of catalytic kinase. The number of RegPhos-identified kinase-specific phosphorylation sites is 12,037.

4.7 Case Study

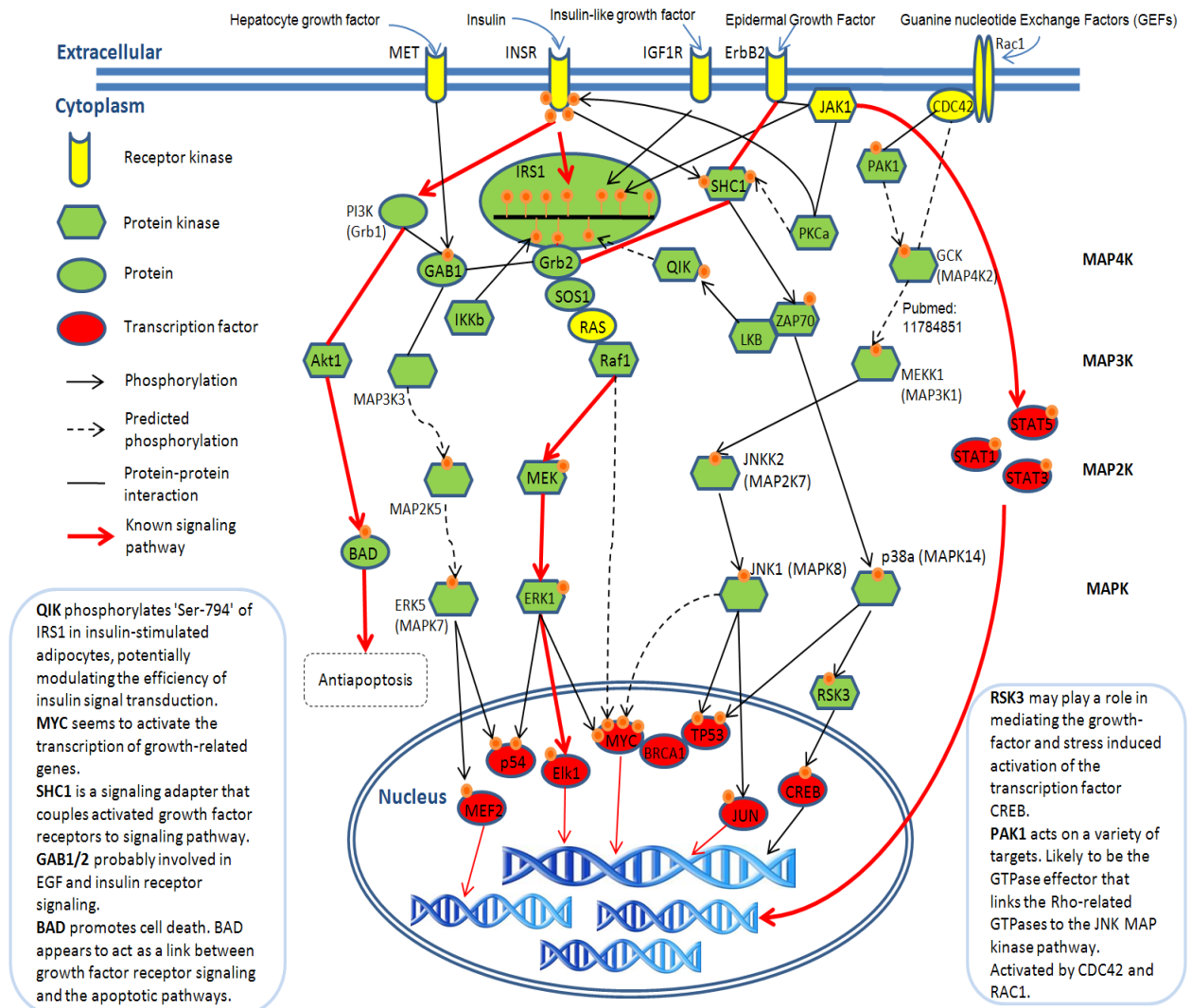


Figure 4.18 Example of the discovered phosphorylation networks.

To demonstrate the effectiveness of the proposed method, the discovered phosphorylation networks associated with the insulin signaling pathway are represented in **Figure 4.18**. Insulin regulates both metabolism and gene expression: the insulin signal passes from the plasma membrane receptor to insulin-sensitive metabolic enzymes and to the nucleus, where it stimulates the transcription of specific genes. The well-known insulin signaling pathway, $INSR \rightarrow IRS1 \text{ --- } Grb2 \text{ --- } SOS1 \text{ --- } RAS \text{ --- } Raf1 \rightarrow MEK \rightarrow ERK1 \rightarrow Elk1$, can be successfully identified by the presented graph-based phosphorylation network searching method (“ \rightarrow ” stands for phosphorylation and “ --- ” stands for protein-protein interaction). Due to the protein-protein interactions can be allowed in the network searching, so many insulin receptor (INSR) related signaling pathways have been discovered, which contain about

100000 pathways in depth 8. After the validation of time-coursed microarray data, the discovered INSR-related phosphorylation networks can be decreased to about 2000 networks.

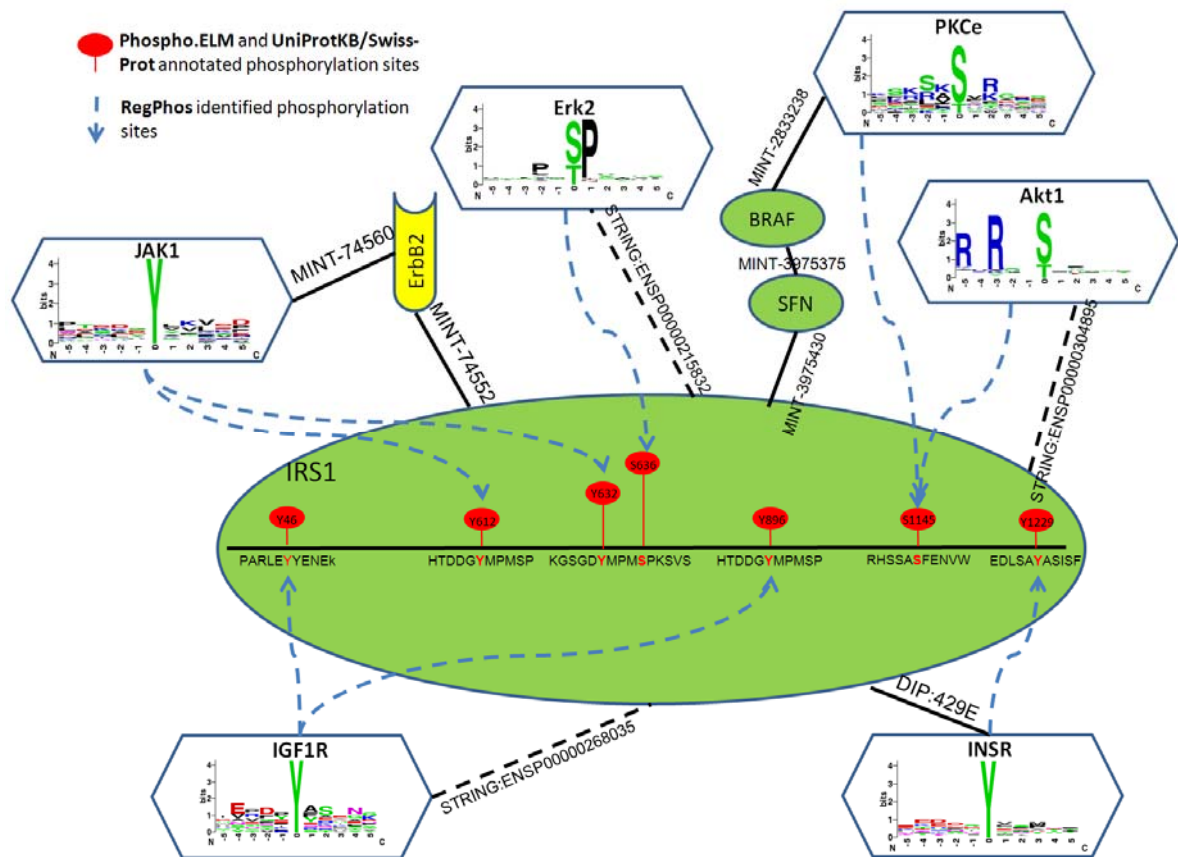


Figure 4.19 Example of RegPhos-identified kinase-specific phosphorylation sites.

Insulin receptor substrate 1 (IRS1), which may mediate the control of various cellular processes by insulin [139], were used to present the RegPhos-identified kinase-specific phosphorylation sites. With the annotation of Phospho.ELM [49] and UniProtKB/SwissProt [48], IRS1 has totally 32 experimentally verified phosphorylation sites. Some of the experimental phosphorylation sites don't have the annotation of catalytic kinases. Based on the trained threshold of logistic regression probability score in each kinase group, these phosphorylation sites were annotated the potential catalytic kinases. As illustrated in **Figure 4.19**, seven kinase-specific phosphorylation sites with their protein associations are identified. For instance, the tyrosine phosphorylation sites “Y612” and “Y632” were potentially catalyzed by *Janus kinase 1* (JAK1), with the indirect protein-protein interaction which was linked by *v-erb-b2 erythroblastic leukemia viral oncogene homolog 2* (ErbB2). The tyrosine phosphorylation sites “Y46” and “Y896” were catalyzed by *Insulin-like Growth Factor I Receptor* (IGF1R), with the directly functional association annotated by STRING [114]. Phosphoserine “S636” was catalyzed by MAPK group, and a functional association shows

that *Mitogen-Activated Protein Kinase 1* (MAPK1 or Erk2) was directly link to IRS1. Phosphotyrosine “Y1229” was catalyzed by insulin receptor (InsR) with the direct protein-protein interaction (DIP:429E) of DIP [122]. Some phosphorylation sites were identified by more than two kinases, for example, phosphoserine “S1145” was potentially catalyzed by *v-akt murine thymoma viral oncogene homolog 1* (Akt1) with directly functional association or was potentially catalyzed by *protein kinase C epsilon* (PKCε) with indirect protein-protein interaction in depth 3, passing through *Stratifin* (SFN) and *B-Raf proto-oncogene serine/threonine-protein kinase* (BRAF).

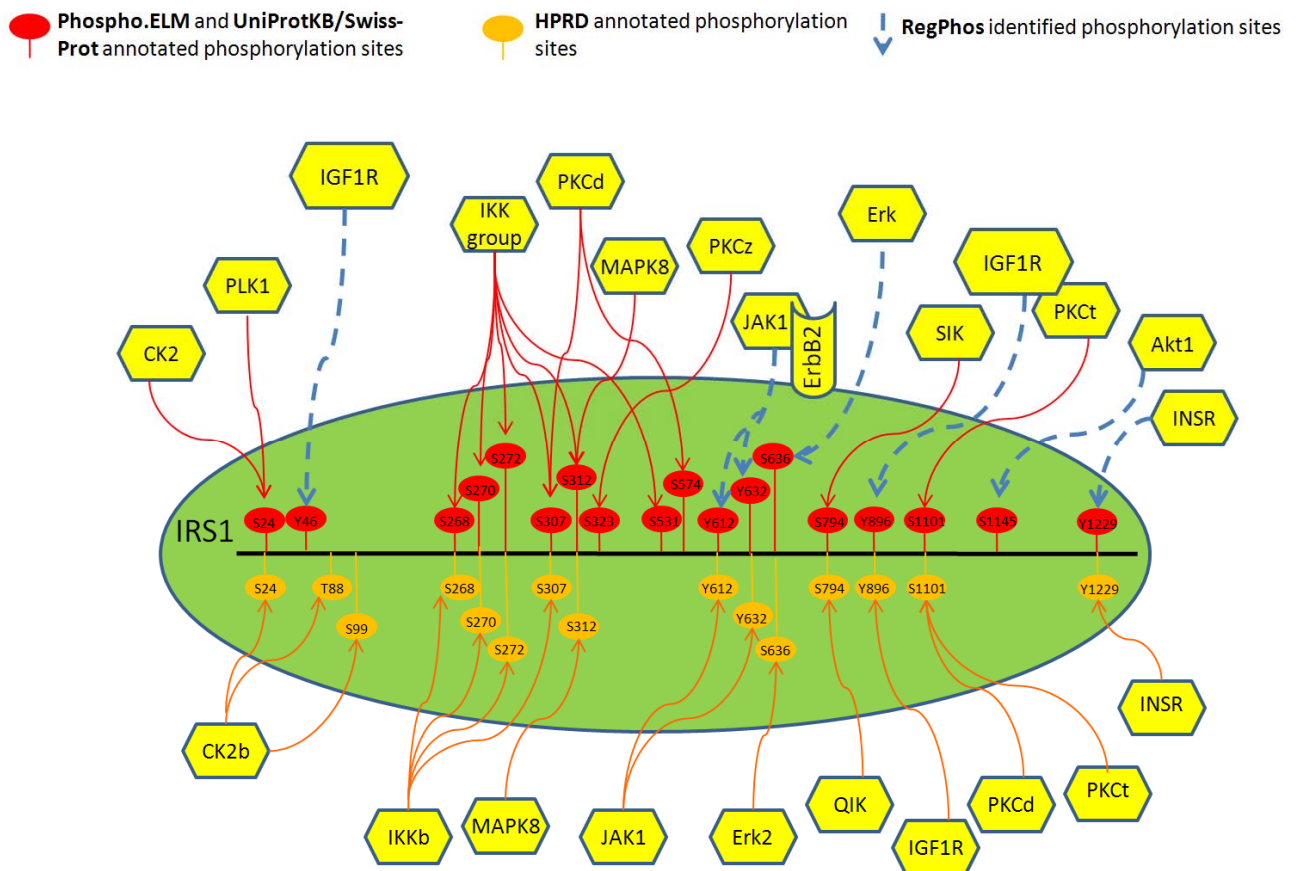


Figure 4.20 Validation of the RegPhos-identified kinase-specific phosphorylation sites using HPRD annotation.

The annotation of kinase-specific phosphorylation sites in HPRD [51] was used to validate the correction of the RegPhos-identified phosphorylation sites. As shown in **Figure 4.20**, the RegPhos-identified results can be verified by HPRD. IRS1 plays an important role in insulin signaling or insulin-like growth factor signaling [140], phosphotyrosine “Y896” were indeed catalyzed by *Insulin-like Growth Factor I Receptor* (IGF1R), by the annotation of HPRD.

Phosphorylation events often occur in a cascade, in which activity of one kinase is dependent on the upstream activity of another. One of the best-studied examples of this is the regulation of the mitogen activated protein kinase (MAPK)-signaling cascade, as the suffix pathway $MEK \rightarrow ERK1$ in insulin signaling pathway. MAPK signaling has no fewer than five levels of kinase regulation [141], MAP4K, MAP3K, MAP2K, MAPK, and MAPKAPK [142]. Furthermore, there is considerable cross talk between signaling cascades involving other phosphoregulators , as the $INSR \rightarrow IRS1 \text{ — Grb2}$ and $IGF1R \rightarrow IRS1 \text{ — Grb2}$, resulting in a network of phosphoregulators rather than a linear cascade.



4.8 Web-based System of RegPhos

To facilitate the investigation of protein kinase and their substrate, a web-based system was implemented for users to efficiently browse the protein kinase and their substrate proteins in a user-friendly manner. Three major functions, including browsing kinase or substrate, constructing phosphorylation network, and microarray expression analysis, are provided in the proposed system. The box of “quick search” can let users input their interested kinase name or substrate name, as shown in **Figure 4.21**, users can investigate into the protein description, subcellular localization, functional domain, tertiary structure, and phosphorylation sites with catalytic kinase of CEBPB. All the experimentally verified kinase-specific phosphorylation sites and RegPhos-identified kinase-specific phosphorylation sites are provided to users. The JMol viewer is adapted for the visualization of PDB structure.

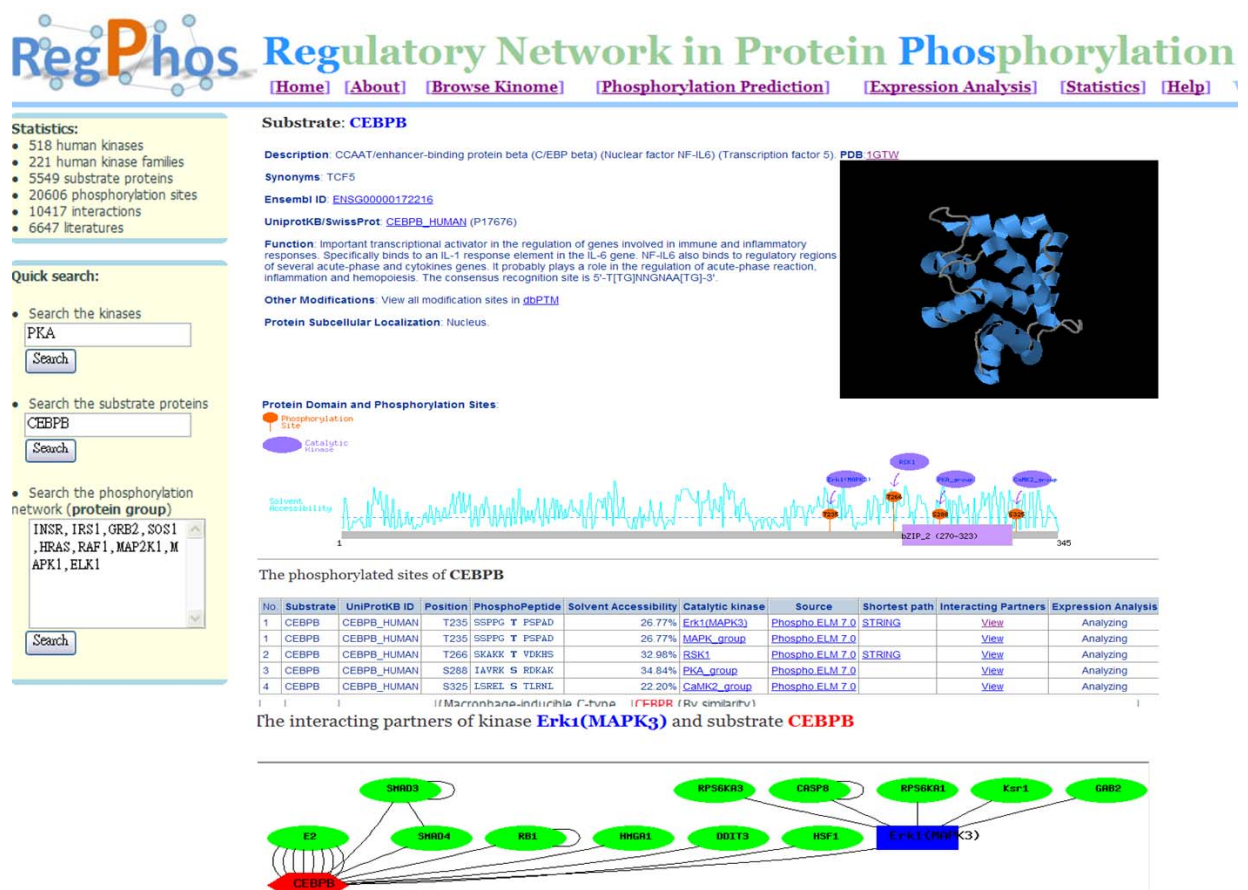


Figure 4.21 Graphical visualization of substrate protein with catalytic kinases.

To investigate the expression correlation of kinase and substrate, the human gene expression samples of Affymetrix GeneChip Human Genome U133 Array Set HG-U133A

platform (GPL96), consisting of 22283 probe set for 12678 genes, are used to explore the co-expression analysis of kinase and substrate genes. However, the first problem we faced is what kind of microarray experiment should be selected for investigating the co-expression of kinase and substrate genes. Without any specific interest and limitation, we decide to focus on the experimental series of microarray with the raw data. Totally 2714 samples within 98 experiment series (GSE) are provided in the web-based system. The Pearson correlation coefficient of gene expression pattern between kinase and substrate are calculated in all 98 experiment series. As shown in **Figure 4.22**, the expression correlation of kinase CDC2 and substrate p53 in 98 experiment series are provided, and users can investigate into the expression pattern of CDC2 and p53 genes in detail.

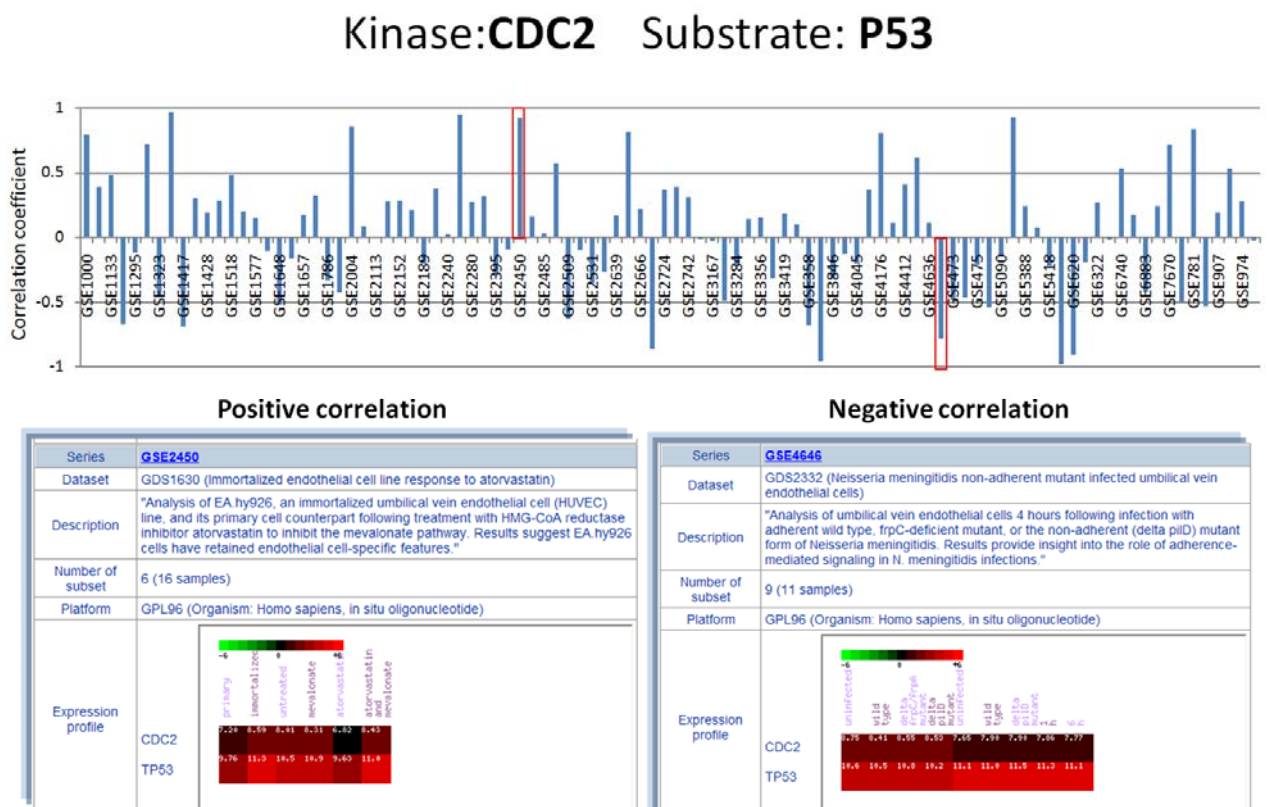


Figure 4.22 The expression profile of kinase and substrate genes.

The proposed system can let users input a group of protein names to be constructed the phosphorylation network associated with the information protein subcellular localization. To fully investigate how protein kinase control the intracellular processes, the experimentally verified kinase-specific phosphorylation sites and the discovered kinase-substrate interactions identified by RegPhos are incorporated to construct the phosphorylation networks starting

from receptor kinases associated with membrane to transcription factors located in nucleus. However, the phosphorylation-driven signal transduction pathway is not always the phosphorylation cascade. Some protein-protein interactions are involved in the signal transduction pathway, such as IRS1-GRB2 interaction, GRB2-SOS1 interaction, SOS1-HRAS interaction, and HRAS-RAF1 interaction in insulin signaling pathway. **Figure 4.23** shows an example of insulin signaling network in the construction of phosphorylation network. A group of proteins associated with insulin signaling pathway are inputted to construct the network from membrane-associated proteins to nuclear proteins.

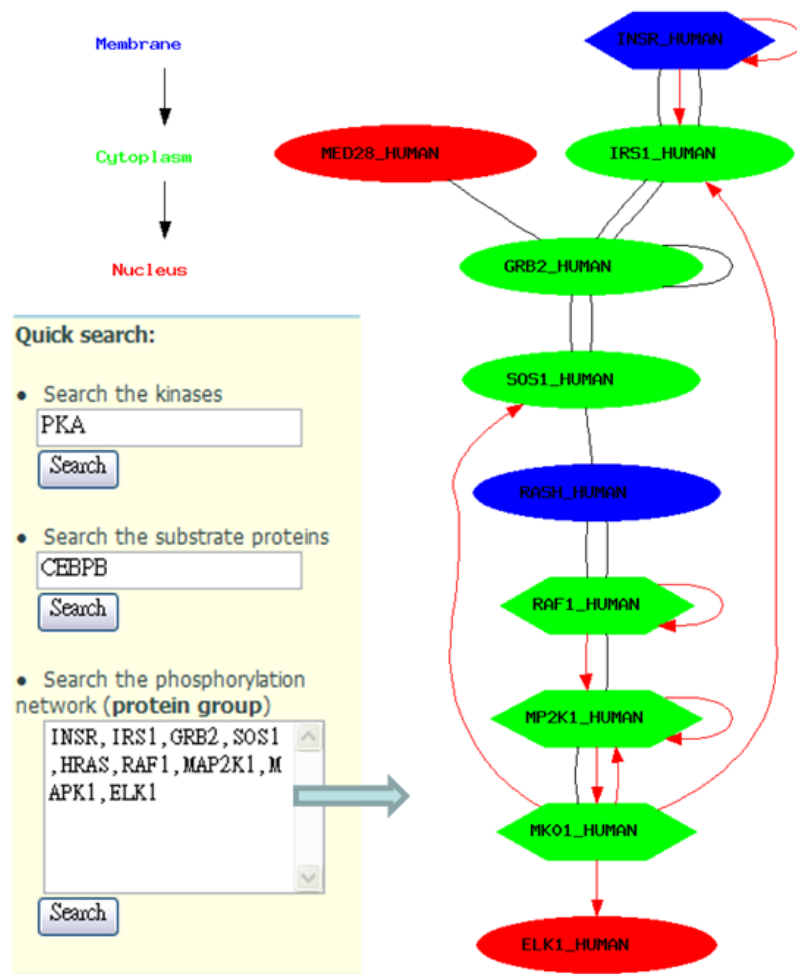


Figure 4.23 Example of insulin signaling network in the construction of phosphorylation network.

4.9 Summary

The desire of mapping phosphorylation networks has motivated the development of computational methods to investigate the substrate specificity of kinase-specific phosphorylation sites, based on experimental identification of the consensus sequence motifs recognized by the active site of kinase catalytic domains. However, only 20% experimental phosphorylation sites have the annotation of catalytic kinases, covering 350 kinases (67%). The presented method is designed to link experimentally validated phosphorylation sites to protein kinases. Due to the fact that signaling proteins are modular in the sense that they contain domains (catalytic or interaction) and linear motifs (phosphorylation or binding sites), which mediate interactions between proteins [92], the protein-protein interaction and protein association are incorporated. It also exploits both the inherent propensity of kinase catalytic domains to phosphorylate particular sequence motifs and contextual information regarding the physical interaction, functional association, cellular co-localization and coexpression of kinases and substrates.

Investigating into the predictive power of the context of protein associations, physical protein interactions play the dominant role among the primary experimental data, whereas gene coexpression contributes only very little. Physical protein interactions were imported and merged from numerous repositories, and the reliability of each individual interaction was assessed based on the promiscuity of the interaction partners. Gene coexpression was measured by calculating the Pearson correlation coefficient between two genes across 98 human gene expression experiment series of Affymetrix GeneChip Human Genome U133 Array Set HG-U133A platform (GPL96) collected from Gene Expression Omnibus repository. After the evaluation, the improved predictive power gained from using context of protein association underlines the importance of kinase-substrate interactions in the specificity of protein phosphorylation within cells. The predictive specificity of kinase groups with similar consensus motifs can be improved by the consideration of protein association. We would also suggest that this underlines the utility of protein association data in modeling cellular processes. The identified kinase-substrate interactions were adopted to fully construct the intracellular phosphorylation networks. Furthermore, GEO microarray expression data were used to validate whether the kinase and substrate genes in the constructed phosphorylation networks have syn-expression pattern.

Chapter 5 Discussions

5.1 Characteristics

To fully investigate how protein kinases regulate the intracellular processes, the comprehensive and accurate identification of the kinase-specific substrates is necessary. Therefore, we propose a method, RegPhos, incorporates computational model with protein associations (protein-protein interactions, functional associations, and subcellular localization) for identifying the catalytic kinase for each phosphoprotein with experimental phosphorylated sites. To observe the expressed relationship between kinase and substrate, the gene expression microarray data is adopted to observe the expression of kinase and substrate genes in specific conditions, for instance, the normal tissue and cancerous tissue.

With the increasing number of in vivo phosphorylation sites have been identified, the desire of map the network of protein kinase and substrate has motivated. The experimental kinase-specific substrates, ultimately, need to be combined by systems biology analysis, which translates the separate, large-scale datasets into signaling networks. Therefore, we incorporated the experimentally verified kinase-substrate interactions with computationally identified kinase-substrate interactions to construct the intracellular phosphorylation network starting from receptor kinases to transcription factors, associated with the formation of protein subcellular localization. Moreover, the experimental expression evidence, such as gene microarray data, was adopted to validate the syn-expression of the constructed kinase-substrate phosphorylation network with statistical significance.

Comparison between RegPhos and NetworKIN

Rune Linding and the authors have developed an approach, NetworKIN [112], that augments motif-based predictions with the network context of kinases and phosphoproteins. As given in **Table 5.1**, the comparison between RegPhos and NetworKIN are listed. NetworKIN collected the experimental phosphorylation data from Phospho.ELM and adopted NetPhosK and Scansite to the phosphorylation site prediction on 20 kinase families encompassing 112 individual kinases. The protein association database STRING, which integrates information from curated pathway databases, co-occurrence in abstracts, physical protein interaction assays, mRNA expression studies, and genomic context, is used to investigate the direct and

indirect interactions between kinase and substrate. NetworKIN pinpoints kinases responsible for specific phosphorylation and yields a 2.5-fold improvement in the accuracy with which phosphorylation networks can be constructed. T

Table 5.1 Comparison between RegPhos and NetworKIN.

Method	NetworKIN	RegPhos
Species	Human	Human
Phosphorylation resource	Phospho.ELM	Phospho.ELM (7.0), UniProtKB/SwissProt (55.0), HPRD (7.0) and PHOSIDA (1.0)
Number of kinase families	20 kinase families encompassing 112 individual kinases	101 kinase families covering 300 kinases
Kinase-specific phosphorylation site prediction	1.NetPhosK (neural network) 2.Scansite (position-specific matrix)	KinasePhos (SVM model trained with sequence and structural features) Blast (for individual kinase whose substrate site are less than 10)
Protein association context	Protein functional association database STRING	1.Protein-protein interaction (DIP, MINT, IntAct, and HPRD) 2.Functional association (STRING) 3.Cellular localization (LOCATE, PSORTdb, OrganelleDB, UniProtKB, and GOA)
Method	Two-staged prediction: 1. Kinase-specific phosphorylation site prediction 2.Protein association context	Logistic regression of 1.Kinase-specific phosphorylation site prediction score 2.Interacting depth of Protein-protein interaction 3.Confidence score of functional association 4.Cellular localization
Gene expression analysis	-	98 experiment series of Affymetrix HG-U133A platform (GPL96)
Predictive performance	52% sensitivity and 64% accuracy for classifying 282 phosphorylation sites of PKC, CDK, PIKK, and INSR	89% sensitivity and 91% accuracy for classifying 309 phosphorylation sites of PKC, CDK, PIKK and INSR from HPRD (independent test)
Phosphorylation network	Only kinase-substrate pairs	1.Using graph-based method to construct phosphorylation networks starting from membrane receptor to transcription factor 2.Using time-coursed microarray data to validate the discovered phosphorylation networks

To compare the predictive power between RegPhos and NetworKIN, the similar dataset of four well-known types of kinase group, such as PKC, CDK, PIKK and INSR, were used to evaluate the classifying power of RegPhos. There are totally 309 phosphorylation sites, which

were independent to training data, extracted from HPRD. By using logistic regression model to integrate the phosphorylation site prediction with protein associations (protein-protein interactions, functional associations, and subcellular localization), the predictive accuracy of RegPhos is higher than the NetworKIN, especially in INSR group. Finally, the constructed kinase-substrate phosphorylation network with statistically significant co-expression of time-coursed microarray data were provided to users.

5.2 Limitations

The proposed method, RegPhos, was used to link the protein kinase to experimentally validated phosphorylation sites. Although the predictive power of RegPhos is effective based on the independent test, there are several limitations about this study.

5.2.1 How Reliable are Protein-Protein Interaction?

Data of protein–protein interactions provide valuable insight into the molecular networks underlying a living cell. However, their accuracy is often questioned, calling for a rigorous assessment of their reliability. The high-throughput methods are believed to contain many false positives, i.e. interactions that are identified in the experiment but never take place in the cell [143]. It is therefore essential to obtain an estimate of the reliability of the interactions documented by the various methods [144]. Elinat Sprinzak *et al.* [145] have developed an intelligible mean to assess directly the rate of true positives in a data set of experimentally determined interacting protein pairs. They show that the reliability of high-throughput yeast two-hybrid assays is about 50%, and that the size of the yeast interactome is estimated to be 10,000–16,600 interactions. To assess the quality of the data we can use two measures in future analysis: the fraction of interacting proteins that were documented as localized in the same cellular compartment, and the fraction of interacting proteins that were annotated as having a common cellular-role.

5.2.2 Time Complexity and Path Length of Signaling Pathway

Construction

Given a graph $G=(V, E)$ with n nodes, m edges, a set S of start nodes (receptor), and a set T of end nodes (TF). When searching acyclic path $p = (s, c_1, \dots, c_k, t)$ with length k that starts from

S and ends at t within T in human protein-protein interaction network, the time complexity is approaching to $O(n^k)$. To accomplish the path searching in a reasonable time, in general, the length of path is defined no more than eight [119]. To address the NP-hard graph search problem, Alon et al. (1995) devised a novel randomized algorithm, called *color coding*, for finding simple paths and simple cycles of a specified length k , within a given graph. Scott *et al.* [146] have adopted and extended the efficient techniques, color coding algorithm, for finding paths in a graph to the problem of identifying pathways in protein interaction networks. The authors presented linear-time algorithms for finding path in a given network under several biologically motivated constraints, and demonstrated that the algorithm was very efficient, computing optimal paths of length 8 within minutes and paths of length 10 in about three hours.

5.2.3 Visualization of Complex Phosphorylation Network

In computer science, the graphical visualization of a graph without any overlap of nodes or edges is close to NP-hard problem. Therefore, in this work, we applied an excellent and popular package namely Graphviz²⁴ to graphically visualize the constructed networks. To present a signaling pathway starting from membrane protein to transcription factor in nucleus, the order of protein occurrence was constrained by the cellular localization of proteins. **Figure 5.1** shows the comparison of Graphviz visualization between pure network and complex, using insulin signaling pathway as an example. If the constructed signaling pathway contains pure interactions across the proteins, the graphical visualization could be illustrated in a reasonable layout. However, the network which contains complex interactions across proteins is visualized in an uneasily interpretable representation. Therefore, it is needed to improvement the visualization of complex network in an easily interpretable representation.

²⁴ Graphviz URL: <http://www.graphviz.org/>

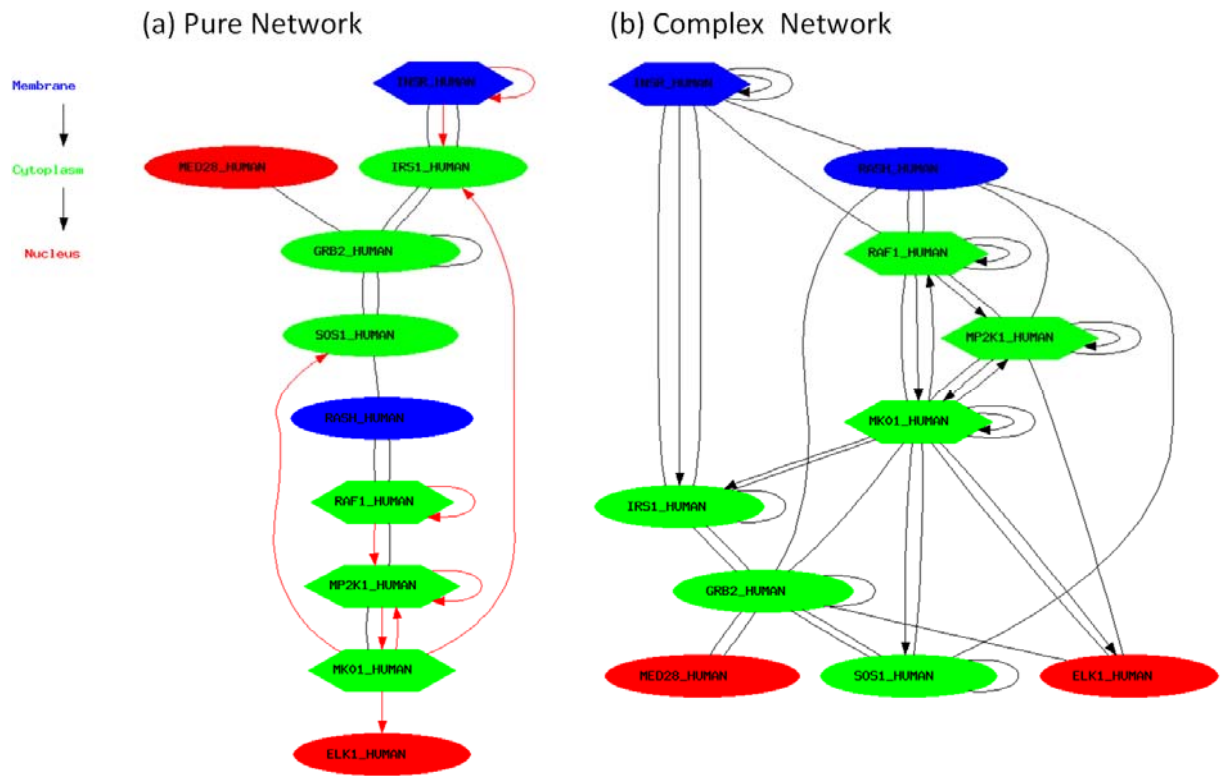
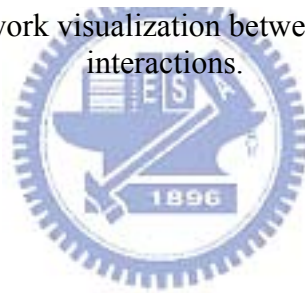


Figure 5.1 Comparison of network visualization between pure interactions and complex interactions.



5.3 Perspectives

Despite those limitations in the proposed method, combining multiple data types (i.e., experimentally validated kinase-specific phosphorylation sites, computationally identified kinase-substrate interactions, and protein association context) is essential for constructing phosphorylation networks and is, as we show in case study, also sufficiently accurate to allow meaningful, theoretical and experimental investigations.

5.3.1 Phosphorylation Sites on Various Protein Isoforms

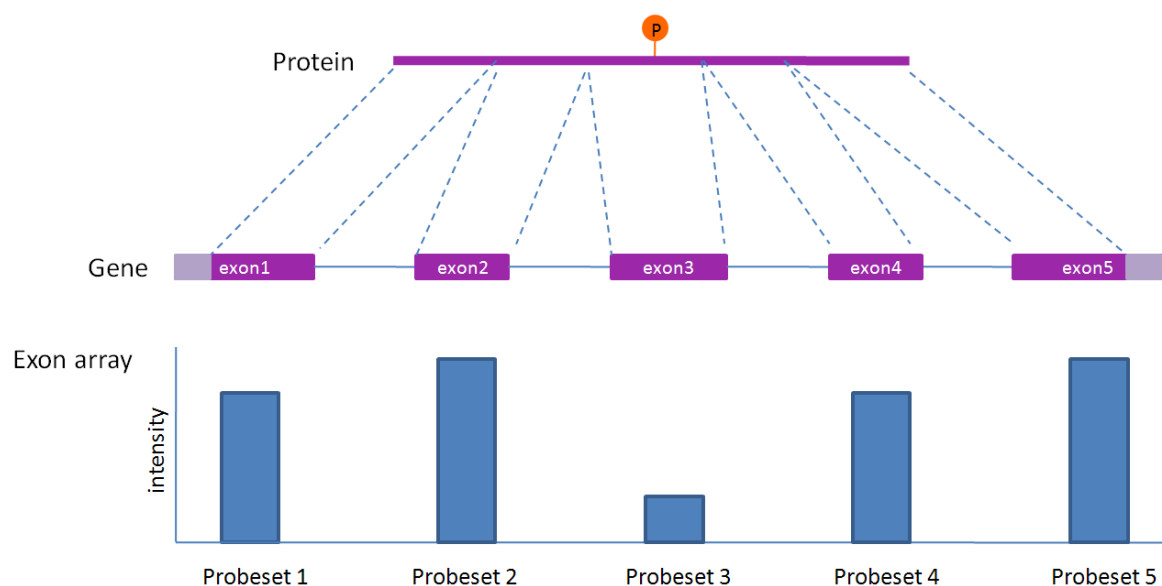


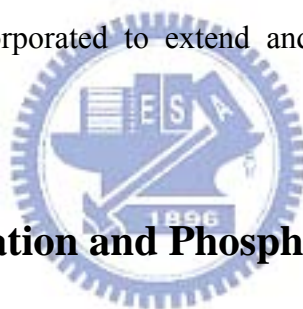
Figure 5.2 Schematic representation of phosphorylation site located in alternatively spliced exon.

With the alternative slicing in mRNA processing, one gene has more than one protein isoforms. Alternative splicing may make the essential phosphorylation sites un-occurred specifically in a protein isoform. Then, the protein function, may be involved in an intracellular signaling pathway, is affected by the missed phosphorylation sites. Therefore, a repository of protein isoforms with experimental phosphorylation sites should be constructed. Moreover, the newly developed exon array, Affymetrix Human Exon 1.0 ST Array (HG16), can be used to investigate the different isoforms in mRNA level in specific experimental condition. Up to May 15th 2008, there are 220 and 352 samples in two exon array platforms GPL5155 and GPL5160 in GEO, respectively. **Figure 5.2** shows the schematic representation

of phosphorylation site located in alternatively spliced exon with the experimental exon expression evidence.

5.3.2 Downstream Genes of Transcription Factors

Intracellular signal transduction is the process by which chemical signals from outside the cell are passed through the cytoplasm to nucleus, and affect the activity of transcription factors to regulate their target gene expression. This work focuses on the protein kinase-substrate phosphorylation network starting from membrane-associated proteins to transcription factors in nucleus. However, the constructed network just ends at transcription factor. The target genes of transcription factor may be more important to biologist. With the annotation of TRANSFAC [137], there are about 1300 transcription factors in human. Due to the statistics from the collected human experimental phosphorylation sites, ~ 40% of transcription factors contain phosphorylation sites. Therefore, the experimentally verified downstream genes of transcription factors can be incorporated to extend and complete the signal transduction network in cellular system.



5.3.1 Dephosphorylation and Phosphatase

Protein phosphorylation is a reversible post-translational modification implicated in many areas of biology. A phosphatase, act in opposition to protein kinases, is an enzyme that removes a phosphate group from its substrate by hydrolysing phosphoric acid monoesters into a phosphate ion and a molecule with a free hydroxyl group [142]. Protein kinases and phosphatases can regulate the phosphorylation status of the protein complement of a cell, and in turn, regulate the activity of their target phosphoproteins in cellular processes. The presence or absence of a phosphate group can change the conformation of the target protein, thereby modifying its activity. Defining the entire complement of these proteins gives us an opportunity to view the system as a whole. Forrest et al. [142] have identified 162 candidate protein phosphatases for the investigation of phosphoregulation. Phosphorylation events often occur in a cascade, in which activity of one kinase or phosphatase is dependent on the upstream activity of another. One of the best-studied examples of this is the regulation of the mitogen activated protein kinase (MAPK)-signaling cascade. MAPK signaling has no fewer than five levels of kinase regulation, MAP4K, MAP3K, MAP2K, MAPK, and MAPKAPK

[147] and one level of phosphatase regulation (MKP) [148]. Therefore, phosphatase is necessary in the signaling pathway and is needed to be considered in the investigation of protein phosphorylation networks.



Chapter 6 Conclusion

Protein phosphorylation catalyzed by kinase plays crucial regulatory role in intracellular signal transduction that transmits information from the cell surface to the nucleus, where they ultimately effect transcriptional changes. With the full annotation of human kinome identified by Manning *et al.*, there is a starting point for comprehensive analysis of intracellular protein phosphorylation networks. Mass spectrometry-based proteomics have enabled the large-scale mapping of *in vivo* phosphorylation sites. In order to fully and accurately investigate the phosphorylation networks, the experimentally validated phosphorylation site databases have been integrated. However, only 20% experimental phosphorylation sites have the annotation of catalytic kinases, covering 350 kinases (67%). Experimental identification of kinase-specific phosphorylation sites is an inconvenient work and usually limited by the availability of detailed data on the kinase-specific substrates. *In silico* prediction could be a promising strategy to conduct preliminary analyses and could greatly reduce the number of potential targets that need further *in vivo* or *in vitro* confirmation.

The presented method, namely RegPhos, was designed to link experimentally validated phosphorylation sites to protein kinases. Due to the fact that signaling proteins are modular in the sense that they contain domains (catalytic or interaction) and linear motifs (phosphorylation or binding sites), which mediate interactions between proteins, the protein-protein interaction, protein functional association, and cellular localization are incorporated. Investigating into the predictive power of the context of protein associations, physical protein interactions play the dominant role among the primary experimental data, whereas gene coexpression contributes un-robust correlation between kinase and substrate genes. Physical protein interactions were imported and merged from numerous repositories, and the reliability of each individual interaction was assessed based on the promiscuity of the interaction partners. After the evaluation, the improved predictive power gained from using context of protein association underlines the importance of kinase-substrate interactions in the specificity of protein phosphorylation within cells. The predictive specificity of kinase groups with similar consensus motifs can be improved by the consideration of protein association. We would also suggest that this underlines the utility of protein association data in modeling cellular processes.

To complete the intracellular processes about protein kinases and phosphorylation, the

identified kinase-substrate interactions were adopted to fully construct the intracellular phosphorylation networks starting from membrane receptor to transcription factors. The discovered phosphorylation networks were validated by calculating the Pearson correlation coefficient of gene expression patterns between kinase and substrate genes across 9 time-coursed experiment series of Affymetrix GeneChip Human Genome U133 Array Set HG-U133A platform (GPL96) collected from Gene Expression Omnibus repository. As illustrated in case study, the discovered phosphorylation networks with highly correlated expression pattern demonstrated that they may be involved in insulin signaling pathway or EGF signaling pathway.



References

1. Hubbard, M.J. and P. Cohen, *On target with a new mechanism for the regulation of protein phosphorylation*. Trends Biochem Sci, 1993. **18**(5): p. 172-7.
2. Diella, F., et al., *Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins*. BMC Bioinformatics, 2004. **5**(1): p. 79.
3. Farriol-Mathis, N., et al., *Annotation of post-translational modifications in the Swiss-Prot knowledge base*. Proteomics, 2004. **4**(6): p. 1537-50.
4. Garavelli, J.S., *The RESID Database of Protein Modifications as a resource and annotation tool*. Proteomics, 2004. **4**(6): p. 1527-33.
5. Mann, M. and O.N. Jensen, *Proteomic analysis of post-translational modifications*. Nat Biotechnol, 2003. **21**(3): p. 255-61.
6. Lehninger AL, N.D., Cox MM *Lehninger Principles of Biochemistry*. Fourth Edition ed. 2005: W. H. Freeman. 1100.
7. Helikar, T., et al., *Emergent decision-making in biological signal transduction networks*. Proc Natl Acad Sci U S A, 2008. **105**(6): p. 1913-8.
8. Pawson, T., *Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems*. Cell, 2004. **116**(2): p. 191-203.
9. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics*. Nature, 2003. **422**(6928): p. 198-207.
10. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules*. Science, 1989. **246**(4926): p. 64-71.
11. Delom, F. and E. Chevet, *Phosphoprotein analysis: from proteins to proteomes*. Proteome Sci, 2006. **4**: p. 15.
12. de la Fuente van Bentem, S. and H. Hirt, *Using phosphoproteomics to reveal signalling dynamics in plants*. Trends Plant Sci, 2007. **12**(9): p. 404-11.
13. Janes, K.A. and M.B. Yaffe, *Data-driven modelling of signal-transduction networks*. Nat Rev Mol Cell Biol, 2006. **7**(11): p. 820-8.
14. Diella, F., et al., *Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins*. BMC Bioinformatics, 2004. **5**: p. 79.
15. Pang, C.N., A. Hayen, and M.R. Wilkins, *Surface accessibility of protein post-translational modifications*. J Proteome Res, 2007. **6**(5): p. 1833-45.
16. Moelbert, S., E. Emberly, and C. Tang, *Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins*. Protein Sci, 2004. **13**(3): p. 752-62.
17. Naderi-Manesh, H., et al., *Prediction of protein surface accessibility with information theory*. Proteins, 2001. **42**(4): p. 452-9.
18. Ahmad, S., M.M. Gromiha, and A. Sarai, *Real value prediction of solvent accessibility from amino acid sequence*. Proteins, 2003. **50**(4): p. 629-35.
19. Lee, T.Y., et al., *dbPTM: an information repository of protein post-translational modification*. Nucleic Acids Res, 2006. **34**(Database issue): p. D622-7.
20. Linding, R., et al., *Protein disorder prediction: implications for structural proteomics*. Structure, 2003. **11**(11): p. 1453-9.
21. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server*. Bioinformatics, 2000. **16**(4): p. 404-5.
22. George, R.A. and J. Heringa, *An analysis of protein domain linkers: their classification and role in protein folding*. Protein Eng, 2002. **15**(11): p. 871-9.
23. Dunker, A.K., et al., *Intrinsically disordered protein*. J Mol Graph Model, 2001. **19**(1): p. 26-59.

24. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm*. J Mol Biol, 1999. **293**(2): p. 321-31.
25. Iakoucheva, L.M., et al., *Intrinsic disorder in cell-signaling and cancer-associated proteins*. J Mol Biol, 2002. **323**(3): p. 573-84.
26. Dunker, A.K., et al., *Intrinsic disorder and protein function*. Biochemistry, 2002. **41**(21): p. 6573-82.
27. Ferron, F., et al., *A practical overview of protein disorder prediction methods*. Proteins, 2006. **65**(1): p. 1-14.
28. Garner, E., et al., *Predicting Binding Regions within Disordered Proteins*. Genome Inform Ser Workshop Genome Inform, 1999. **10**: p. 41-50.
29. Fletcher, C.M. and G. Wagner, *The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein*. Protein Sci, 1998. **7**(7): p. 1639-42.
30. Mader, S., et al., *The translation initiation factor eIF-4E binds to a common motif shared by the translation factor eIF-4 gamma and the translational repressors 4E-binding proteins*. Mol Cell Biol, 1995. **15**(9): p. 4990-7.
31. Romero, P., et al., *Sequence complexity of disordered protein*. Proteins, 2001. **42**(1): p. 38-48.
32. Smyth, E., et al., *Solution structure of native proteins with irregular folds from Raman optical activity*. Biopolymers, 2001. **58**(2): p. 138-51.
33. Liu, J., H. Tan, and B. Rost, *Loopy proteins appear conserved in evolution*. J Mol Biol, 2002. **322**(1): p. 53-64.
34. Liu, J. and B. Rost, *NORSp: Predictions of long regions without regular secondary structure*. Nucleic Acids Res, 2003. **31**(13): p. 3833-5.
35. Ward, J.J., et al., *The DISOPRED server for the prediction of protein disorder*. Bioinformatics, 2004. **20**(13): p. 2138-9.
36. Linding, R., et al., *GlobPlot: Exploring protein sequences for globularity and disorder*. Nucleic Acids Res, 2003. **31**(13): p. 3701-8.
37. Dosztanyi, Z., et al., *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content*. Bioinformatics, 2005. **21**(16): p. 3433-4.
38. Coeytaux, K. and A. Poupon, *Prediction of unfolded segments in a protein sequence based on amino acid composition*. Bioinformatics, 2005. **21**(9): p. 1891-900.
39. Yang, Z.R., et al., *RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins*. Bioinformatics, 2005. **21**(16): p. 3369-76.
40. Vullo, A., et al., *Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W164-8.
41. Pierleoni, A., et al., *BaCellLo: a balanced subcellular localization predictor*. Bioinformatics, 2006. **22**(14): p. e408-16.
42. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003*. Nucleic Acids Res, 2003. **31**(1): p. 365-70.
43. Hornbeck, P.V., et al., *PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation*. Proteomics, 2004. **4**(6): p. 1551-61.
44. Wurgler-Murphy, S.M., D.M. King, and P.J. Kennelly, *The Phosphorylation Site Database: A guide to the serine-, threonine-, and/or tyrosine-phosphorylated proteins in prokaryotic organisms*. Proteomics, 2004. **4**(6): p. 1562-70.
45. Gnad, F., et al., *PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites*. Genome Biol, 2007. **8**(11): p. R250.

46. Gupta, R., et al., *O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins*. Nucleic Acids Res, 1999. **27**(1): p. 370-2.
47. Chornorudskiy, A.L., et al., *UbiProt: a database of ubiquitylated proteins*. BMC Bioinformatics, 2007. **8**: p. 126.
48. Boutet, E., et al., *UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase*. Methods Mol Biol, 2007. **406**: p. 89-112.
49. Diella, F., et al., *Phospho.ELM: a database of phosphorylation sites--update 2008*. Nucleic Acids Res, 2008. **36**(Database issue): p. D240-4.
50. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase*. Nucleic Acids Res, 2004. **32**(Database issue): p. D115-9.
51. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans*. Genome Res, 2003. **13**(10): p. 2363-71.
52. Heazlewood, J.L., et al., *PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor*. Nucleic Acids Res, 2008. **36**(Database issue): p. D1015-21.
53. Zanzoni, A., et al., *Phospho3D: a database of three-dimensional structures of protein phosphorylation sites*. Nucleic Acids Res, 2007. **35**(Database issue): p. D229-31.
54. Lo Conte, L., et al., *SCOP: a structural classification of proteins database*. Nucleic Acids Res, 2000. **28**(1): p. 257-9.
55. Huang, H.D., et al., *Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites*. J Comput Chem, 2005. **26**(10): p. 1032-41.
56. Huang, H.D., et al., *KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites*. Nucleic Acids Res, 2005. **33**(Web Server issue): p. W226-9.
57. Blom, N., S. Gammeltoft, and S. Brunak, *Sequence and structure-based prediction of eukaryotic protein phosphorylation sites*. J Mol Biol, 1999. **294**(5): p. 1351-62.
58. Iakoucheva, L.M., et al., *The importance of intrinsic disorder for protein phosphorylation*. Nucleic Acids Res, 2004. **32**(3): p. 1037-49.
59. Berry, E.A., A.R. Dalby, and Z.R. Yang, *Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms*. Comput Biol Chem, 2004. **28**(1): p. 75-85.
60. Crooks, G.E., et al., *WebLogo: a sequence logo generator*. Genome Res, 2004. **14**(6): p. 1188-90.
61. Schneider, T.D. and R.M. Stephens, *Sequence logos: a new way to display consensus sequences*. Nucleic Acids Res, 1990. **18**(20): p. 6097-100.
62. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA*. J Mol Biol, 1997. **268**(1): p. 78-94.
63. Eddy, S.R., *Profile hidden Markov models*. Bioinformatics, 1998. **14**(9): p. 755-63.
64. Yip, Y.L., et al., *The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants*. Hum Mutat, 2004. **23**(5): p. 464-70.
65. Hubbard, T., et al., *Ensembl 2005*. Nucleic Acids Res, 2005. **33**(Database issue): p. D447-53.
66. Mulder, N.J., et al., *InterPro: an integrated documentation resource for protein families, domains and functional sites*. Brief Bioinform, 2002. **3**(3): p. 225-35.
67. Deshpande, N., et al., *The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema*. Nucleic Acids Res, 2005. **33**(Database issue): p. D233-7.
68. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*. Biopolymers, 1983. **22**(12): p. 2577-637.

69. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes*. BMC Bioinformatics, 2003. **4**: p. 41.
70. Ahmad, S., M.M. Gromiha, and A. Sarai, *RVP-net: online prediction of real valued accessible surface area of proteins from single sequences*. Bioinformatics, 2003. **19**(14): p. 1849-51.
71. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
72. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
73. Julenius, K., et al., *Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites*. Glycobiology, 2005. **15**(2): p. 153-64.
74. Gupta, R., et al., *Scanning the available Dictyostelium discoideum proteome for O-linked GlcNAc glycosylation sites using neural networks*. Glycobiology, 1999. **9**(10): p. 1009-22.
75. Zhou, F.F., et al., *GPS: a novel group-based phosphorylation predicting and scoring method*. Biochem Biophys Res Commun, 2004. **325**(4): p. 1443-8.
76. Xue, Y., et al., *PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory*. BMC Bioinformatics, 2006. **7**: p. 163.
77. Blom, N., et al., *Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence*. Proteomics, 2004. **4**(6): p. 1633-49.
78. Kim, J.H., et al., *Prediction of phosphorylation sites using SVMs*. Bioinformatics, 2004. **20**(17): p. 3179-84.
79. Kiemer, L., J.D. Bendtsen, and N. Blom, *NetAcet: prediction of N-terminal acetylation sites*. Bioinformatics, 2005. **21**(7): p. 1269-70.
80. Chen, H., et al., *MeMo: a web tool for prediction of protein methylation modifications*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W249-53.
81. Li, A., et al., *Prediction of Nepsilon-acetylation on internal lysines implemented in Bayesian Discriminant Method*. Biochem Biophys Res Commun, 2006. **350**(4): p. 818-24.
82. Monigatti, F., et al., *The Sulfinator: predicting tyrosine sulfation sites in protein sequences*. Bioinformatics, 2002. **18**(5): p. 769-70.
83. Xue, Y., et al., *NBA-Palm: prediction of palmitoylation site implemented in Naive Bayes algorithm*. BMC Bioinformatics, 2006. **7**: p. 458.
84. Zhou, F., et al., *CSS-Palm: palmitoylation site prediction with a clustering and scoring strategy (CSS)*. Bioinformatics, 2006. **22**(7): p. 894-6.
85. Maurer-Stroh, S., B. Eisenhaber, and F. Eisenhaber, *N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence*. J Mol Biol, 2002. **317**(4): p. 541-57.
86. Podell, S. and M. Gribskov, *Predicting N-terminal myristoylation sites in plant proteins*. BMC Genomics, 2004. **5**(1): p. 37.
87. Bologna, G., et al., *N-Terminal myristoylation predictions by ensembles of neural networks*. Proteomics, 2004. **4**(6): p. 1626-32.
88. Xue, Y., et al., *SUMOSP: a web server for sumoylation site prediction*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W254-7.
89. Fankhauser, N. and P. Maser, *Identification of GPI anchor attachment signals by a Kohonen self-organizing map*. Bioinformatics, 2005. **21**(9): p. 1846-52.
90. Eisenhaber, B., P. Bork, and F. Eisenhaber, *Prediction of potential GPI-modification*

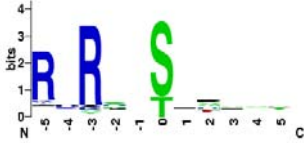



- sites in proprotein sequences*. J Mol Biol, 1999. **292**(3): p. 741-58.
91. Wong, Y.H., et al., *KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns*. Nucleic Acids Res, 2007. **35**(Web Server issue): p. W588-94.
 92. Seet, B.T., et al., *Reading protein modifications with interaction domains*. Nat Rev Mol Cell Biol, 2006. **7**(7): p. 473-83.
 93. Ng, S.K., et al., *InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes*. Nucleic Acids Res, 2003. **31**(1): p. 251-4.
 94. Stock, A.M., V.L. Robinson, and P.N. Goudreau, *Two-component signal transduction*. Annu Rev Biochem, 2000. **69**: p. 183-215.
 95. Manning, G., et al., *The protein kinase complement of the human genome*. Science, 2002. **298**(5600): p. 1912-34.
 96. Arthur, J.W., A. Sanchez-Perez, and D.I. Cook, *Scoring of predicted GRK2 phosphorylation sites in Nedd4-2*. Bioinformatics, 2006. **22**(18): p. 2192-5.
 97. Duda RO, H.P., Stork DG, *Pattern classification*. 2nd edition. Vol. 680. 2004, Beijing: China Maching Press.
 98. Haykin, S., *Neural Networks: A comprehensive foundation, 2nd Ed.* 1999: Prentice-Hall.
 99. Hoekstra, A., Kraaijveld, M.A., Ridder, D. de and Schmidt, W.F. , *The Complete SPRLIB & ANNLIB*. April 1996: Pattern Recognition Group, Delft University of Technolog.
 100. Cortes, C.a.V.V., *Support-vector networks*. Machine Learning, 1995. **20**: p. 273-297.
 101. Yu, C.S., et al., *Prediction of protein subcellular localization*. Proteins, 2006. **64**(3): p. 643-51.
 102. Nguyen, M.N. and J.C. Rajapakse, *Two-stage multi-class support vector machines to protein secondary structure prediction*. Pac Symp Biocomput, 2005: p. 346-57.
 103. Williams, R.D., et al., *Prognostic classification of relapsing favorable histology Wilms tumor using cDNA microarray expression profiling and support vector machines*. Genes Chromosomes Cancer, 2004. **41**(1): p. 65-79.
 104. Lin, C.-C.C.a.C.-J., *LIBSVM : a library for support vector machines*. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2001.
 105. Obenauer, J.C., L.C. Cantley, and M.B. Yaffe, *Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs*. Nucleic Acids Res, 2003. **31**(13): p. 3635-41.
 106. Xue, Y., et al., *GPS: a comprehensive www server for phosphorylation sites prediction*. Nucleic Acids Res, 2005. **33**(Web Server issue): p. W184-7.
 107. Wan, J., et al., *Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection*. Nucleic Acids Res, 2008. **36**(4): p. e22.
 108. Liang, H.K., et al., *Amino acid coupling patterns in thermophilic proteins*. Proteins, 2005. **59**(1): p. 58-63.
 109. Bryson, K., et al., *Protein structure prediction servers at University College London*. Nucleic Acids Res, 2005. **33**(Web Server issue): p. W36-8.
 110. Chang, C.-C. and C.-J. Lin, *LIBSVM : a library for support vector machines*. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2001.
 111. Johnson, S.A. and T. Hunter, *Kinomics: methods for deciphering the kinome*. Nat Methods, 2005. **2**(1): p. 17-25.
 112. Linding, R., et al., *Systematic discovery of in vivo phosphorylation networks*. Cell, 2007. **129**(7): p. 1415-26.
 113. Hjerrild, M., et al., *Identification of phosphorylation sites in protein kinase A*

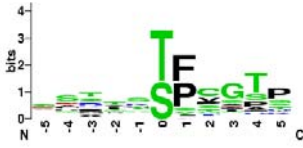
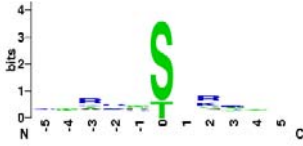
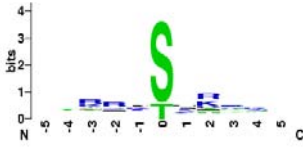
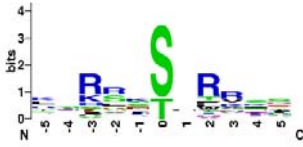



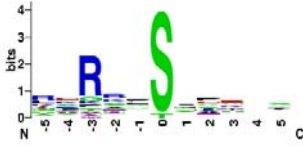

- substrates using artificial neural networks and mass spectrometry*. J Proteome Res, 2004. **3**(3): p. 426-33.
114. von Mering, C., et al., *STRING 7--recent developments in the integration and prediction of protein interactions*. Nucleic Acids Res, 2007. **35**(Database issue): p. D358-62.
 115. Chia-Ting Yang, C.-H.C., Ya-Ling Yu, Tsu-Chun Emma Lin, Sheng-An Lee, Chueh-Chuan Yen, Jinn-Moon Yang, Jin-Mei Lai, Yi-Ren Hong, Tzu-Ling Tseng, Kun-Mao Chao and Chi-Ying F. Huang, *PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database*. Bioinformatics, 2008.
 116. Neves, S.R. and R. Iyengar, *Modeling of signaling networks*. Bioessays, 2002. **24**(12): p. 1110-7.
 117. Choi, C., et al., *Consistent re-modeling of signaling pathways and its implementation in the TRANSPATH database*. Genome Inform, 2004. **15**(2): p. 244-54.
 118. Sachs, K., et al., *Causal protein-signaling networks derived from multiparameter single-cell data*. Science, 2005. **308**(5721): p. 523-9.
 119. Steffen, M., et al., *Automated modelling of signal transduction networks*. BMC Bioinformatics, 2002. **3**: p. 34.
 120. Roberts, C.J., et al., *Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles*. Science, 2000. **287**(5454): p. 873-80.
 121. Xenarios, I., et al., *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions*. Nucleic Acids Res, 2002. **30**(1): p. 303-5.
 122. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update*. Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
 123. Chatr-Aryamontri, A., et al., *MINT: the Molecular INTERaction database*. Nucleic Acids Res, 2006.
 124. Kerrien, S., et al., *IntAct--open source resource for molecular interaction data*. Nucleic Acids Res, 2007. **35**(Database issue): p. D561-5.
 125. Camon, E., et al., *The Gene Ontology Annotation (GOA) Database--an integrated resource of GO annotations to the UniProt Knowledgebase*. In Silico Biol, 2004. **4**(1): p. 5-6.
 126. Ashburner, M., et al., *Gene ontology: tool for the unification of biology*. The Gene Ontology Consortium. Nat Genet, 2000. **25**(1): p. 25-9.
 127. Sprenger, J., et al., *LOCATE: a mammalian protein subcellular localization database*. Nucleic Acids Res, 2008. **36**(Database issue): p. D230-3.
 128. Guo, T., et al., *DBSubLoc: database of protein subcellular localization*. Nucleic Acids Res, 2004. **32**(Database issue): p. D122-4.
 129. Wiwatwattana, N., et al., *Organelle DB: an updated resource of eukaryotic protein localization and function*. Nucleic Acids Res, 2007. **35**(Database issue): p. D810-4.
 130. Rey, S., et al., *PSORTdb: a protein subcellular localization database for bacteria*. Nucleic Acids Res, 2005. **33**(Database issue): p. D164-8.
 131. Davis, M.J., et al., *MemO: a consensus approach to the annotation of a protein's membrane organization*. In Silico Biol, 2006. **6**(5): p. 387-99.
 132. Bono, H., et al., *FANTOM DB: database of Functional Annotation of RIKEN Mouse cDNA Clones*. Nucleic Acids Res, 2002. **30**(1): p. 116-8.
 133. Barrett, T., et al., *NCBI GEO: mining tens of millions of expression profiles--database and tools update*. Nucleic Acids Res, 2007. **35**(Database issue): p. D760-5.
 134. Dhillon, I.S.a.M., D. S., *Concept decompositions for large sparse text data using clustering*. Machine Learning, 2001. **42**: p. 143-175.
 135. Bebek, G. and J. Yang, *PathFinder: mining signal transduction pathway segments*

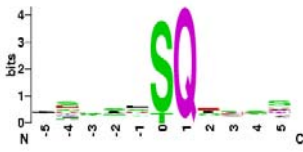

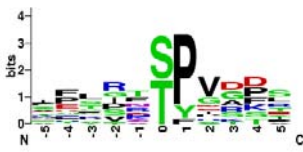





- from protein-protein interaction networks*. BMC Bioinformatics, 2007. **8**: p. 335.
136. Sharan, R., et al., *Conserved patterns of protein interaction in multiple species*. Proc Natl Acad Sci U S A, 2005. **102**(6): p. 1974-9.
 137. Wingender, E., H. Karas, and R. Knuppel, *TRANSFAC database as a bridge between sequence data libraries and biological function*. Pac Symp Biocomput, 1997: p. 477-85.
 138. Hochreiter, S., D.A. Clevert, and K. Obermayer, *A new summarization method for Affymetrix probe level data*. Bioinformatics, 2006. **22**(8): p. 943-9.
 139. Craparo, A., T.J. O'Neill, and T.A. Gustafson, *Non-SH2 domains within insulin receptor substrate-1 and SHC mediate their phosphotyrosine-dependent interaction with the NPEY motif of the insulin-like growth factor I receptor*. J Biol Chem, 1995. **270**(26): p. 15639-43.
 140. Sachdev, D. and D. Yee, *Disrupting insulin-like growth factor signaling as a potential cancer therapy*. Mol Cancer Ther, 2007. **6**(1): p. 1-12.
 141. Lehman, J.A. and J. Gomez-Cambronero, *Molecular crosstalk between p70S6k and MAPK cell signaling pathways*. Biochem Biophys Res Commun, 2002. **293**(1): p. 463-9.
 142. Forrest, A.R., et al., *Phosphoregulators: protein kinases and protein phosphatases of mouse*. Genome Res, 2003. **13**(6B): p. 1443-54.
 143. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions*. Nature, 2002. **417**(6887): p. 399-403.
 144. Deane, C.M., et al., *Protein interactions: two methods for assessment of the reliability of high throughput observations*. Mol Cell Proteomics, 2002. **1**(5): p. 349-56.
 145. Sprinzak, E., S. Sattath, and H. Margalit, *How reliable are experimental protein-protein interaction data?* J Mol Biol, 2003. **327**(5): p. 919-23.
 146. Scott, J., et al., *Efficient algorithms for detecting signaling pathways in protein interaction networks*. J Comput Biol, 2006. **13**(2): p. 133-44.
 147. Dan, I., N.M. Watanabe, and A. Kusumi, *The Ste20 group kinases as regulators of MAP kinase cascades*. Trends Cell Biol, 2001. **11**(5): p. 220-30.
 148. Theodosiou, A. and A. Ashworth, *MAP kinase phosphatases*. Genome Biol, 2002. **3**(7): p. REVIEWS3009.

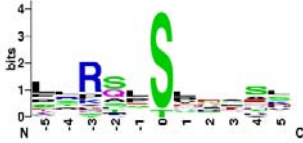





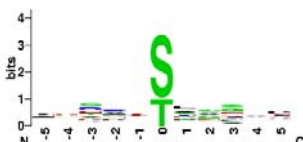
Appendix I – Human Kinase Families



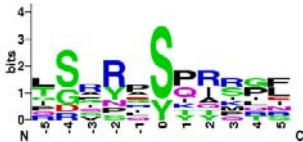
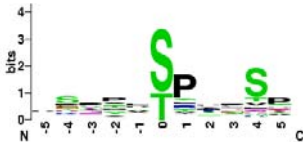


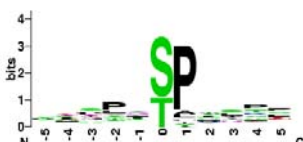


Table A1 221 human kinase families with sequence logos of amino acid surrounding substrate sites.


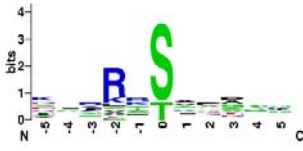
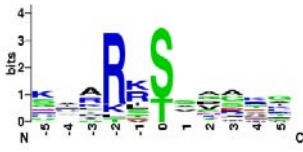
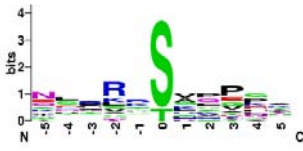



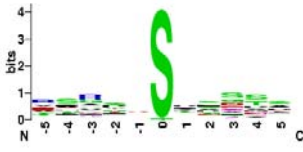
No.	Group	Family	Sub family	Description	Kinase member	Number of phosphosites in human	Sequence logo
1	AGC	PKB		Protein kinase B	AKT1 AKT2 AKT3	89(63)	
2	AGC	DMPK	ROCK	Rho Kinase	ROCK1 ROCK2	25(15)	
3	AGC	DMPK	GEK	Genghis Khan	DMPK1 MRCKa DMPK2 MRCKb	3(3)	
4	AGC	DMPK	CRIK	Citron Rho-interacting kinase	CRIK	0(0)	
5	AGC	GRK	GRK	G-protein coupled Receptor Kinase	GPRK7 RHOK GPRK6 GPRK5 GPRK4	64(19)	
6	AGC	GRK	BARK	Beta Adrenergic Receptor Kinase	BARK1 BARK2	32(14)	
7	AGC	MAST	MAST	Microtubule Associated Serine/Threonine Kinase	MAST1 MAST2 MAST3 MAST4	0(0)	
8	AGC	MAST	MASTL	MAST like	MASTL	0(0)	
9	AGC	NDR		Nuclear Dbf2-related kinases	NDR1 NDR2 LATS1 LATS2	2(2)	
10	AGC	PKA		Protein kinase A	PKACa PKACb PKACg	232(151)	

11	AGC	PDK1		Phosphoinositide-dependent protein kinase	PDK1	45(20)	
12	AGC	PKC		Protein kinase C	PKCh PKCa PKCb PKCd PKCe PKCg PKCi PKCt PKCz	280(168)	
13	AGC	PKC	Alpha	Protein kinase C, alpha	PKCa PKCb PKCg	121(74)	
14	AGC	PKC	Delta	Protein kinase C, delta	PKCd PKCt	30(22)	
15	AGC	PKC	Eta	Protein kinase C, eta	PKCe PKCh	14(12)	
16	AGC	PKC	Iota	Protein kinase C, iota	PKCi PKCz	13(11)	
17	AGC	PKG		Protein kinase G	PKG1 PKG2	25(15)	
18	AGC	PKN		Protein kinase N	PKN1 PKN2 PKN3	0(0)	
19	AGC	RSK	RSK	Ribosomal S6 Kinase	RSK1 RSK2 RSK3 RSK4	40(31)	
20	AGC	RSK	p70	p70 subfamily of Ribosomal Specific Kinase	p70S6K p70S6Kb	9(7)	





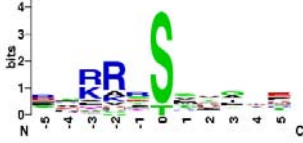



				main Associated Protein			
38	Atypical	PIKK	SMG1		SMG1	0(0)	
39	Atypical	PIKK	ATM	Ataxia telangiectasia mutated	ATM ATR	67(34)	
40	Atypical	PIKK	DNAPK	DNA Protein Kinase	DNAPK	25(13)	
41	Atypical	PIKK	FRAP	FKBP12-rapamycin-associated protein	FRAP	14(4)	
42	Atypical	RIO			RIOK1 RIOK2 RIOK3	0(0)	
43	Atypical	TAF1			TAF1 TAF1L	5(2)	
44	Atypical	TIF1		Transcriptional Intermediary Factor	TIF1a TIF1b	0(0)	
45	CAMK	CAMK-U nique			VACAMKL STK33	0(0)	
46	CAMK	CAMK1		CAMK family 1	CaMK1a CaMK1b CaMK1g CaMK1d CaMK4	18(14)	
47	CAMK	CAMK2		CAMK family 2	CaMK2a CaMK2b CaMK2g CaMK2d	56(36)	
48	CAMK	CAMKL	NIM1		NIM1	0(0)	
49	CAMK	CAMKL	NuaK	Novel (Nua) Kinase family	NuaK1 NuaK2	4(4)	
50	CAMK	CAMKL	PASK	PAS domain Kinase	PASK	2(1)	



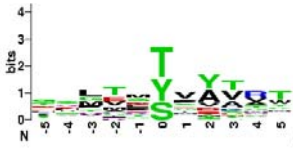
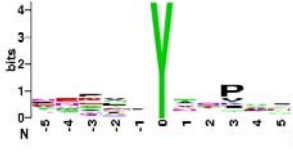
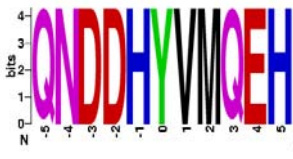
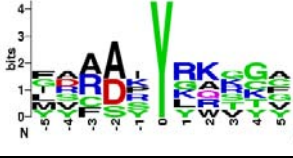
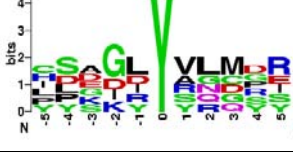
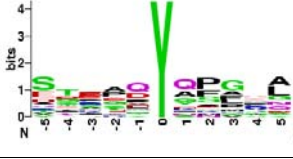

63	CAMK	DCAMK L		Doublecortin and CaMK-Like	DCLK1 DCLK2 DCLK3	0(0)	
64	CAMK	MAPKAP K	MAPKA PK	MAP Kinase Associated Protein Kinase	MAPKAP2 MAPKAP3 MAPKAP5	25(16)	
65	CAMK	MAPKAP K	MNK	MAPK iNtegrating or iNteracting Kinase	MNK1 MNK2	1(1)	
66	CAMK	MLCK		Myosin Light Chain Kinases	smMLCK TTN caMLCK skMLCK Sgk085	3(2)	
67	CAMK	PHK		Phosphorylase Kinase	PHKg1 PHKg2	9(5)	
68	CAMK	PIM			PIM1 PIM2 PIM3	1(1)	
69	CAMK	PKD		Protein Kinase D	PKD1 PKD2 PKD3	7(6)	
70	CAMK	PSK		Protein Serine Kinase	PSKH1 PSKH2	0(0)	
71	CAMK	SgK495			SgK495	0(0)	
72	CAMK	Trbl		tribbles	Trb1 Trb2 Trb3	0(0)	
73	CAMK	Trio			Trad Trio SPEG Obscn	0(0)	
74	CAMK	TSSK		Testis Specific Serine Kinase	TSSK1 TSSK2 TSSK3 TSSK4 SSTK	0(0)	
75	CK1	CK1		Cell Kinase 1	CK1a CK1d CK1e CK1g2 CK1g3 CK1a2 CK1g1	63(33)	

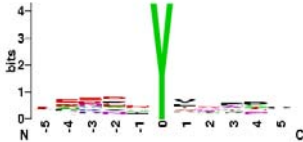
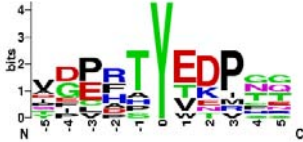


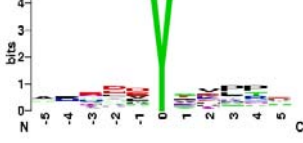
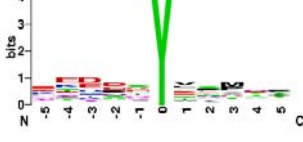
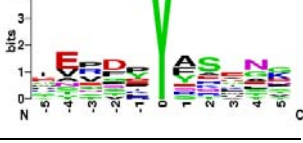
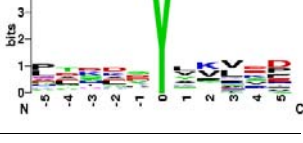
88	CMGC	DYRK	PRP4		PRP4	1(1)	
89	CMGC	DYRK	HIPK	Homeodomain Interacting Protein Kinases	HIPK1 HIPK2 HIPK3 HIPK4	3(3)	
90	CMGC	DYRK	DYRK	Dual-specificity Tyrosine Regulated Kinase	DYRK1A DYRK1B DYRK2 DYRK3 DYRK4	9(7)	
91	CMGC	GSK		Glycogen Synthase 3 Kinase	GSK3A GSK3B	56(34)	
92	CMGC	MAPK	MAPK	Mitogen Activated Protein Kinase	Erk1(MAPK3) Erk2(MAPK1) Erk3(MAPK6) Erk4(MAPK4) Erk5(MAPK7) Erk7(MAPK15) JNK1(MAPK8) JNK2(MAPK9) JNK3(MAPK10) NLK p38a(MAPK14) p38b(MAPK11) p38g(MAPK12) p38d(MAPK13)	248(140)	
93	CMGC	MAPK	JNK	JNK subfamily of MAPK	JNK1(MAPK8) JNK2(MAPK9) JNK3(MAPK10)	47(27)	
94	CMGC	MAPK	p38	p38 subfamily of MAPK	p38a(MAPK14) p38b(MAPK11) p38g(MAPK12) p38d(MAPK13)	62(35)	
95	CMGC	MAPK	ERK	Extracellular signal-Regulated protein Kinase	Erk1(MAPK3) Erk2(MAPK1) Erk3(MAPK6) Erk4(MAPK4) Erk5(MAPK7) Erk7(MAPK15)	138(88)	
96	CMGC	MAPK	nmo	nemo	NLK	2(1)	

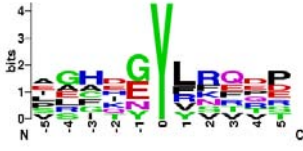
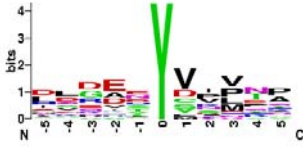
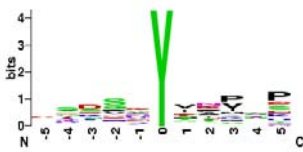



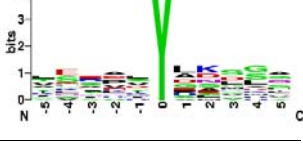
97	CMGC	RCK			MAK MOK ICK	1(1)	
98	CMGC	SRPK		SR Protein Kinase; phosphorylates SR splicing factors	MSSK1 SRPK1 SRPK2	0(0)	
99	Other	AUR	Aur	Aurora Kinase	AurA AurB AurC	42(19)	
100	Other	AUR	AurB	Aurora Kinase B	AurB	21(12)	
101	Other	AUR	AurA	Aurora Kinase B	AurA	23(9)	
102	Other	BUB			BUB1 BUBR1	0(0)	
103	Other	Bud32			PRPK	1(1)	
104	Other	CAMKK	Meta	Metazoan-specific family of CAMK (Calcium/Calmodulin Regulated Kinase) Kinase	CaMKK1 CaMKK2	3(3)	
105	Other	CDC7		Cell Division Control 7	CDC7	2(1)	
106	Other	Haspin			Haspin	0(0)	
107	Other	IKK		I kappa Kinase	IKKa IKKb IKKe TBK1	43(12)	
108	Other	IRE		Inositol REquiring	IRE1 IRE2	0(0)	
109	Other	MOS			MOS	0(0)	




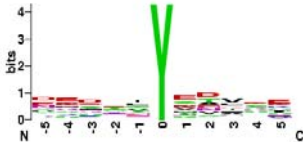

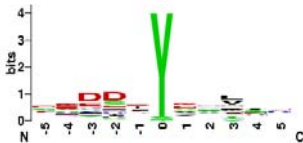
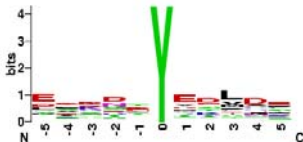
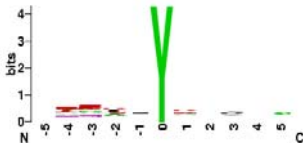

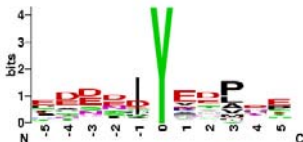
125	Other	SCY1		SCYL1 SCYL2 SCYL3	0(0)	
126	Other	SgK071		SgK071	0(0)	
127	Other	SgK493		SgK493	0(0)	
128	Other	SgK496		SgK496	0(0)	
129	Other	Slob		Named after the first member, Drosophila SLOW Borders Slob	0(0)	
130	Other	TBCK		TBC domain-containing Kinase TBCK	0(0)	
131	Other	TLK		Tousled-Like Kinase TLK1 TLK2	0(0)	
132	Other	TOPK		T-cell Originated Kinase, also known as PBK (PDZ domain-binding kinase) PBK	0(0)	
133	Other	TTK		TTK	0(0)	
134	Other	ULK	ULK	Unc-51 Like Kinase ULK1 ULK2 ULK3 ULK4	0(0)	
135	Other	ULK	Fused	Fused	0(0)	
136	Other	VPS15		PIK3R4	1(1)	
137	Other	WEE		Homologs of yeast Wee1; cell cycle kinase Wee1 Wee1B MYT1	2(2)	
138	Other	WNK		With No Lysine kinases. Missing the canonical catalytic lysine, instead using a nearby lysine for catalysis Wnk1 Wnk2 Wnk3 Wnk4	4(2)	
139	Other	CDPK		CDPK	3(3)	
140	RGC	RGC		Receptor Guanylate Cyclases ANPa ANPb CYGD CYGF HSER	0(0)	

141	STE	STE-Unique		COT NIK	3(2)	
142	STE	STE11	MAP3K	MAP3K (MAP kinase kinase kinase) genes, homologous to yeast Ste 11 MAP3K1 MAP3K2 MAP3K3 MAP3K4 MAP3K5 MAP3K6 MAP3K7 MAP3K8	17(10)	
143	STE	STE20	PAK	p21-Activated protein Kinase PAK1 PAK2 PAK3 PAK4 PAK5 PAK6	34(25)	
144	STE	STE20	PAKB	Class B p21-Activated protein Kinase PAK4 PAK5 PAK6	1(1)	
145	STE	STE20	PAKA	Class A p21-Activated protein Kinase PAK1 PAK2 PAK3	29(23)	
146	STE	STE20	SLK	Ste20-Like Kinase LOK SLK	1(1)	
147	STE	STE20	STLK	Serine Threonine Like Kinase STLK5 STLK6 STLK6-rs	0(0)	
148	STE	STE20	TAO	Thousand and One Kinase, whose sequence was 1001 AA long. Also known as the SULU family TAO1 TAO2 TAO3	0(0)	
149	STE	STE20	YSK	Yeast SPS1/STE20-like kinase YSK1 MST3 MST4	1(1)	
150	STE	STE20	KHS	Kinase Homologous to STE20 KHS1 KHS2 GCK HPK1	3(3)	
151	STE	STE20	FRAY	STLK3 OSR1	0(0)	

152	STE	STE20	MSN	misshapen	HGK TNIK NRK MINK	1(1)	
153	STE	STE20	MST	Mammalian STE20-like kinase	MST1 MST2	21(20)	
154	STE	STE20	NinaC		MYO3A MYO3B	0(0)	
155	STE	STE7	MAP2K	MAP2K (MAP kinase kinase) genes, homologous to yeast Ste 7	MAP2K1 MAP2K2 MAP2K3 MAP2K4 MAP2K5 MAP2K6 MAP2K7	29(13)	
156	TK	Abl		Abelson murine leukemia homolog	ABL1(Abl) ABL2(ARG)	36(26)	
157	TK	Ack		Activated Cdc42-associated tyrosine kinase	ACK TNK1	1(1)	
158	TK	ALK		Anaplastic Lymphoma Kinase	ALK LTK	6(3)	
159	TK	Axl			AXL MER TYRO3	6(2)	
160	TK	CCK4		Colon Carcinoma Kinase 4	CCK4	0(0)	
161	TK	Csk		C-SRC kinase	CSK CTK	19(8)	
162	TK	DDR		Discoidin Domain Receptor kinase	DDR1 DDR2	1(1)	

163	TK	EGFR		Epidermal Growth Factor Receptor	EGFR ErbB2 ErbB3 ErbB4	48(22)	
164	TK	Eph		Ephrin receptors	EphA1 EphA2 EphA3 EphA4 EphA5 EphA6 EphA7 EphA8 EphA10 EphB1 EphB2 EphB3 EphB4 EphB5 EphB6	11(7)	
165	TK	FAK		Focal Adhesion Kinase	FAK PYK2	9(6)	
166	TK	Fer	Fer	Proto-oncogene tyrosine-protein kinase FER	FER	0(0)	
167	TK	Fer	Fes	Proto-oncogene tyrosine-protein kinase Fes/Fps	FES	7(2)	
168	TK	FGFR		Fibroblast Growth Factor Receptor	FGFR1 FGFR2 FGFR3 FGFR4	30(8)	
169	TK	InsR		Insulin Receptor and associated Kinases	INSR IRR	30(9)	
170	TK	InsR	IGF1R	Insulin-like growth factor I receptor	IGF1R	11(3)	
171	TK	JAK	JakA	First (active) kinase domain of the dual-domain Janus Kinases	JAK1 JAK2 JAK3	29(17)	

172	TK	TYK	TYK2	Non-receptor tyrosine-protein kinase	TYK2	7(4)	
173	TK	Lmr		Lemur Kinase	LMR1 LMR2 LMR3	0(0)	
174	TK	Met	Met	Met proto-oncogene tyrosine kinase	MET RON	19(4)	
175	TK	Musk		Muscle-Specific Kinase	MUSK	0(0)	
176	TK	PDGFR	PDGFR	Platelet-derived growth factor	PDGFRa PDGFRb	30(10)	
177	TK	PDGFR	Fms	Platelet-derived growth factor, FMS	FMS	2(1)	
178	TK	PDGFR	FLT3	Fms-Like Tyrosine kinase	FLT3	4(1)	
179	TK	PDGFR	Kit	Platelet-derived growth factor	KIT	4(1)	
180	TK	Ret		Proto-oncogene tyrosine-protein kinase receptor ret	RET	14(3)	
181	TK	Ror		Regeneron Orphan Receptors	ROR1 ROR2	0(0)	
182	TK	Ryk		Receptor Tyrosine Kinase	RYK	0(0)	
183	TK	Sev		Named after Drosophila sevenless, a receptor tyrosine kinase involved in eye cell fate determination	ROS	0(0)	
184	TK	Src	SRM	Proto-oncogene tyrosine-protein kinase SRM	SRM	0(0)	

185	TK	Src	BLK	Proto-oncogene tyrosine-protein kinase BLK	BLK	4(2)	
186	TK	Src	Brk	Proto-oncogene tyrosine-protein kinase BRK	BRK	4(2)	
187	TK	Src	Fgr	Proto-oncogene tyrosine-protein kinase FGR	FGR	4(3)	
188	TK	Src	Fyn	Proto-oncogene tyrosine-protein kinase FYN	FYN	37(21)	
189	TK	Src	HCK	Proto-oncogene tyrosine-protein kinase HCK	HCK	9(7)	
190	TK	Src	Lck	Proto-oncogene tyrosine-protein kinase Lck	LCK	48(25)	
191	TK	Src	LYN	Tyrosine-protein kinase LYN	LYN	33(20)	
192	TK	Src	Src	Proto-oncogene tyrosine-protein kinase Src	SRC	108(68)	
193	TK	Src	YES	Proto-oncogene tyrosine-protein kinase YES	YES	3(2)	
194	TK	Src	Frk	Fyn-related kinase	FRK	0(0)	
195	TK	Syk	SYK	Spleen tyrosine kinase	SYK	38(17)	

196	TK	Syk	ZAP70	70 kDa zeta-associated protein, Syk-related tyrosine kinase	ZAP70	16(8)	
197	TK	Tec		Tec protein tyrosine kinase family	TXK TEC ITK BMX BTK	26(13)	
198	TK	Tec	BTK	Bruton tyrosine kinase	BTK	14(7)	

220	TKL	STKR	TGFbR	Serine/Threonine Kinase Receptors; receptors for TGFb ligands	TGFbR1 TGFbR2	12(3)	
221	TKL	TKL-Unique			MLKL	0(0)	

