Computational Methodologies and Resources for Discovery of Phosphorylation Regulation and Function in Cellular Networks

by

Kristen M Naegle

S.M. Biological Engineering, M.S. Electrical Engineering, B.S. Electrical Engineering

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

Author .....
Department of Biological Engineering
May 21, 2010

Certified by .....
Douglas A. Lauffenburger
Professor
Thesis Supervisor

Accepted by .....
Darrell Irvine
Chairman, Department Committee on Graduate Students

This Doctoral Thesis has been examined by the following Thesis Committee:

Douglas A. Lauffenburger, Ph.D.
Professor of Biological Engineering
Massachusetts Institute of Technology

Forest M. White, Ph.D.
Thesis Committee Chair
Associate Professor of Biological Engineering
Massachusetts Institute of Technology

Michael B. Yaffe, Ph.D.
Professor of Biological Engineering and Biology
Massachusetts Institute of Technology

Computational Methodologies and Resources for Discovery of Phosphorylation Regulation and Function in Cellular Networks

by

Kristen M Naegle

Submitted to the Department of Biological Engineering
on May 21, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Engineering

Abstract

Post-translational modications (PTMs) regulate cellular signaling networks by modifying activity, localization, turnover and other characteristics of proteins in the cell. For example, signaling in receptor tyrosine kinase (RTK) networks, such as those downstream of epidermal growth factor receptor (EGFR) and insulin receptor, is initiated by binding of cytokines or growth factors, and is generally propagated by phosphorylation of signaling molecules. The rate of discovery of PTM sites is increasing rapidly and is signicantly outpacing our biological understanding of the function and regulation of those modications. The ten-fold increase in known phosphorylation sites over a five year time span can primarily be attributed to mass spectrometry (MS) measurement methods, which are capable of identifying and monitoring hundreds to thousands of phosphorylation sites across multiple biological samples. There is significant interest in the field in understanding these modifications, due to their important role in basic physiology as well as their implication in disease. In this thesis, we develop algorithms and tools to aid in analysis and organization of these immense datasets, which fundamentally seek to generate novel insights and testable hypotheses regarding the function and regulation of phosphorylation in RTK networks. We have developed a web-accessible analysis and repository resource for high-throughput quantitative measurements of post-translational modifications, called PTMScout. Additionally, we have developed a semi-automatic, high-throughput screen for unsupervised learning parameters based on their relative ability to partition datasets into functionally related and biologically meaningful clusters. We developed methods for comparing the variability and robustness of these clustering solutions and discovered that phosphopeptide co-clustering robustness can recapitulate known protein interaction networks, and extend them. Both of these tools take advantage of a new linear motif discovery algorithm, which we additionally used to find a putative regulatory sequence downstream of the highly tumorigenic EGFRvIII mutation that indicates casein kinase II (CK2) activity may be increased in glioblastoma.

Thesis Supervisor: Douglas A. Lauffenburger
Title: Professor

Acknowledgments

I have had the great fortune of being part of a wonderful and collaborative research environment at MIT. In particular, the Lauffenburger and White labs have provided a respectful and fantastic place to work. Every individual within these labs has been generous with their time and ideas, which in various ways, has contributed a great deal to the success of this work. In particular, I would like to thank the following people in these labs for their contributions to my education and research: Megan Palmer, Arthur Goldsipe, Brian Joughin, Melissa Gymrek, Pam Kreeger, Shannon Alford, Hyung Do Kim, Ben Cosgrove, Dan Kirouac, Joel Wagner, Justin Pritchard, Stacey Pawson, Emily Miraldi, Alejandro Wolf-Yadlin, Josh Apgar, Paul Huang, Bracken King, Carol Huang, and Mark Fleury.

A special thanks to each of my committee members, Douglas Lauffenburger, Forest White, and Michael Yaffe; I had the joy of being closely advised by all of them throughout my Ph.D. work. I would like to especially thank Doug Lauffenburger for always believing in my work and in me as a researcher. He is an ideal advisor and I wish all graduate students could have the experience I have had. I had the fortune of collaborating with Roy Welsch, which was essential to my understanding of multiple hypothesis correction. Brian Joughin has been a great source of both scientific advice and technical support as well as a pleasure to work with. Without Melissa Gymrek, PTMScout would not have been possible; she made a dream come true when she developed the web interface to the database I had developed.

I would like to thank Medtronic for funding my first year of work, the Siebel Foundation for funding the last year of work, and ICBP and CDP for funding the years in between.

I would like to thank my parents for always supporting me. This milestone could not have been possible without John Naegle, my wonderful husband of eight years. He has supported me, unconditionally, in everything I have attempted. He is a wonderful person, husband and father.

Dedicated to John Naegle

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 A role for post-translational modification in the cell

Cellular organisms encode their genetic information in the form of four nucleic acids. During transcription, this information is transcribed into another set of four nucleic acids. During translation this information is then translated into yet another set of chemical information: amino acids. Proteins, the fundamental functional unit of the cell, are composed of these twenty amino acids. It turns out that these twenty amino acids represent only a small fraction of the possible chemical entities that can compose a protein, since the cell has one more mechanism of control available, post-translational modification. Post-translational modifications (PTM) all entail the covalent transfer of a biochemical entity to a particular amino acid residue within the target protein following, or in tandem with, translation. These biochemical entities can range from small molecules, such as a phosphate group, to large protein segments, such as ubiquitin. The effect of this covalent modification on protein function is as diverse as the range of possible modifications. Table 1.1 demonstrates a sampling of the various types of post-translational modifications, the residues they modify, and examples of their resulting cellular effects. In March of 2010, Uniprot, a repository for protein information, contained controlled vocabularies for 405 post-translational

modifications representing roughly 200 different types of functional groups.

Table 1.1: Examples of post-translational modifications. Post-translational modifications capitalize on a variety of different functional groups, yielding a diverse repertoire of molecular functional control.

PTM	**Functional Group**	**Protein Side Chain**	**Example Effects**
Phosphorylation	phosphate	Y,S,T,H	enzymatic activity protein-protein interactions
Acetylation	acetyl group	N-terminal, K	histone code binding
Palmitoylation	fatty acid	C	membrane association protein-protein interactions
Glycosylation	saccharides	N, S, T, OH-K	protein folding
Ubiquitination	ubiquitin (8.5kDa protein)	K	protein stability degradation
SUMOylation	SUMO (12kDa protein)	K	binding nuclear import

The covalent modification of a protein requires the assistance of an enzyme and in some cases, such as ubiquitination, the assistance of multiple enzymes [61]. In the case of reversible modifications, complementary enzymatic processes exist for the removal of the modification. The use of post-translational modification allows the cell to quickly alter the function and regulation of proteins within the cell. This is a marked difference between transcriptional control circuitry, which can require thirty minutes to several hours to effect change and post-translational control, which can take only seconds. This use of cellular control is not only fast, but can also be tightly regulated.

Phosphorylation, a post-translational modification that involves the transfer of a phosphate group from a donor ATP molecule, is highly utilized in the cell. It has been thought that at least 30% of human proteins undergo phosphorylation [16]. Although there are several residues capable of phosphate addition, the majority of stable eukaryotic phosphorylations exist on serine, threonine and tyrosine amino acid

side chains. Protein kinases and phosphatases are the enzymes responsible for phosphorylation and de-phosphorylation, respectively, of target proteins. Kinases are one of the largest family of proteins in the human proteome consisting of 518 catalogued members, which constitutes roughly 1.7% of all human genes [64]. Identification of phosphatases reveals a smaller family of proteins, however the number of tyrosine phosphatases roughly matches the number of tyrosine kinases in the human proteome [3]. Both protein tyrosine kinases and phosphatases have been implicated in a variety of diseases [52, 73, 113]

A primary function of phosphorylation is the alteration of a protein's enzymatic activity. For example, phosphorylation within the activation loop segment of a kinase catalytic domain increases its enzymatic activity through conformational alterations [75]. In addition to altering enzymatic activities, another primary role of phosphorylation is to induce protein-protein interactions and protein localization changes. Several phosphopeptide binding domains exist which recognize phosphorylated forms of an amino acid sequence [128] such as Src homology region 2 (SH2) domains, phosphotyrosine binding domains (PTB), 14-3-3 domains, and WW domains. Given the specific nature of phosphopeptide domain recognition, a protein containing such a domain and a protein containing a cognate phosphorylation sequence can be made to interact in a temporal- or condition-specific manner. These protein-protein interactions can lead directly to controlled localization of proteins. For example, the phosphorylation of a transcription factor, STAT, induces dimerization, the product of which can then be imported into the nucleus [18].

The linear amino acid sequence directly surrounding the site of phosphorylation plays an important role in the recognition of the sequence by kinases and binding domains. Kinase-target specificity is a combination of additional factors, including the proximity of the kinase and target as a result of adaptor or scaffolding proteins [11]. Information regarding phosphatases has, and remains, limited in scope. In the past, it has been assumed regulation of receptor tyrosine kinase (RTK) networks is primarily controlled by regulation of kinases and that phosphatase activity and quantity is uniform in time. However, phosphatase mutations have been implicated

in cancer [73, 113] indicating an important role for the regulation of phosphatases in cellular signaling networks as well. It is now thought that the balance of both positive and negative regulators control duration and amplitude of RTK stimulated responses. It is believed that protein tyrosine phosphatases, like protein tyrosine kinases, recognize their targets in part by the specific linear sequence surrounding a phosphorylated residue [117].

De novo prediction of the function and regulation of phosphorylation modifications is a difficult problem given the complicated functional role of phosphorylation and its control by two separate enzymatic processes. A study by Kumar et. al. [55] showed that even "canonical" phosphorylation functions are situation dependent. There is no blanket statement regarding the general activity increase or decrease in signaling networks due to phosphorylation. For example, phosphorylation of Src family kinase domains on the activation loop (Y415) increases enzymatic activity. However, phosphorylation of its cytoplasmic tail (Y527) by the kinase Csk produces a Src-SH2 binding recognition site, which inhibits kinase catalytic activity through protein conformational changes [6]. Phosphorylation of the phosphatase Shp-2 on Y580 increases the activity of the phosphatase [62], thereby generally decreasing the phosphorylation on targets of Shp-2.

A resource for the repository of known phosphorylations in proteomes of many species, Phospho.ELM, included 1,703 known phosphorylation sites in 2004 when it was first published [23]. By 2009, that number had increased more than ten-fold to almost 20,000 documented phosphorylation sites [22]. This indicates a rapid expansion in the characterization of new sites of phosphorylation within cells across many species. In combination with the complexity of the regulation and function of phosphorylation within cells and signaling networks, this expansion represents an active area of ongoing research.

1.2 Receptor tyrosine kinase networks and the epidermal growth factor receptor

A specific class of proteins in the cell, receptor tyrosine kinases (RTK), monitor extracellular cues and translate these cues into phenotypic outcomes, primarily through the utilization of phosphorylation as a regulatory mechanism. Most receptor tyrosine kinases contain an extracellular binding region, a transmembrane spanning region, and a cytoplasmic kinase catalytic domain. The typical first steps in RTK activation involve two receptors binding their extracellular ligands followed by receptor dimerization, activation, and then cross-autophosphorylation. Autophosphorylation events on the receptors are numerous and occur extensively on the cytoplasmic tail of the receptors. These phosphorylation sites can then recruit other signaling proteins. Once recruited to the receptors, those proteins are then phosphorylated and go on to effect a cascade of changes in the cell that will eventually lead to a discernible phenotypic outcome, such as migration, proliferation, differentiation, structural changes, or basic functional increases and decreases, such as glucose uptake. A second class of receptors rely on the same network principles, but have no native kinase function and rely on the recruitment of cytosolic kinases in place of autophosphorylation ability. An important example of this is the interleukin family of receptors, which are essential components of immune system functionality.

Despite the wide diversity of RTK families, such as the insulin receptor, fibroblast growth factor receptors, epidermal growth factor receptors, Met, and neuronal growth factor receptors, there are a relatively small number of common downstream pathways utilized across these receptor systems. These pathways include the PI(3)K/Akt, Jak/Stat, Plcγ/PKC, and Ras/MAPK pathways [34]. The control of these and other pathways in RTK networks can result in transcriptional changes, metabolic and cytoskeletal changes, which drive phenotypic alterations such as proliferation, differentiation, and migration. Another common feature of these networks is their negative regulation, which is driven primarily by two mechanisms: (1) dephsophorylation on network components following signaling activation and (2) downregulation of the re-

ceptor by endocytosis. It is thought that recycling and degradation of the receptor works to desensitize the immediate future response of the cell to the extracellular cue, as well as to establish a differential signaling mode, compared to signaling from the plasma membrane [12]. Differential EGFR signaling from the plasma membrane, versus from the endocytic compartment, is just one example of the spatial localization of RTK signaling networks. Lipid rafts, focal adhesions, endosomes and the nucleus are just a few of the cellular compartments that can function as important and specific locations for cellular signaling.

The epidermal growth factor receptors, known as either HER or ErbB receptors, are an important receptor tyrosine kinase family involved in a variety of cellular and tissue functions. They play an especially important role in the differentiation and development of epithelial tissues such as lung, heart, brain, and breast tissues. There are four members of the ErbB family, which recognize a variety of ligands, see Figure 1-1 [131]. ErbB family members have been implicated in the progression and severity of a number of cancers including glioblastoma [81] and breast [15], lung [68], and ovarian cancer [100]. Several veins of therapeutic development have focused on the EGFR family and its network components. For example, gefitinib (Iressa, AstraZeneca) and erlotinib (Tarceva, Genentech) are ATP analogs specific to EGFR, which are effective in the treatment of cancers expressing EGFR mutants with increased catalytic activity [63,80]. ErbB family antibodies are another avenue of successful therapeutic development, such as trastuzumab for breast cancer (Herceptin, Genentech) and cetuximab (Brystol Meyers Squibb BMS/ImClone) for colorectal cancer and neck and squamous cell carcinoma.

Scientists have employed a variety of modeling techniques as they have sought to understand the underlying biochemical structure of RTK networks. These models serve as a framework for testing our current understanding of the network as well as allowing for *in silico* perturbation experiments, such as inhibition of a network component by a drug. Just as experimental measurements in the ErbB network have been substantial, so has the development of various models, from fully mechanistic [49,94], to logic-based [91], to probabilistic [89]. These models typically include either

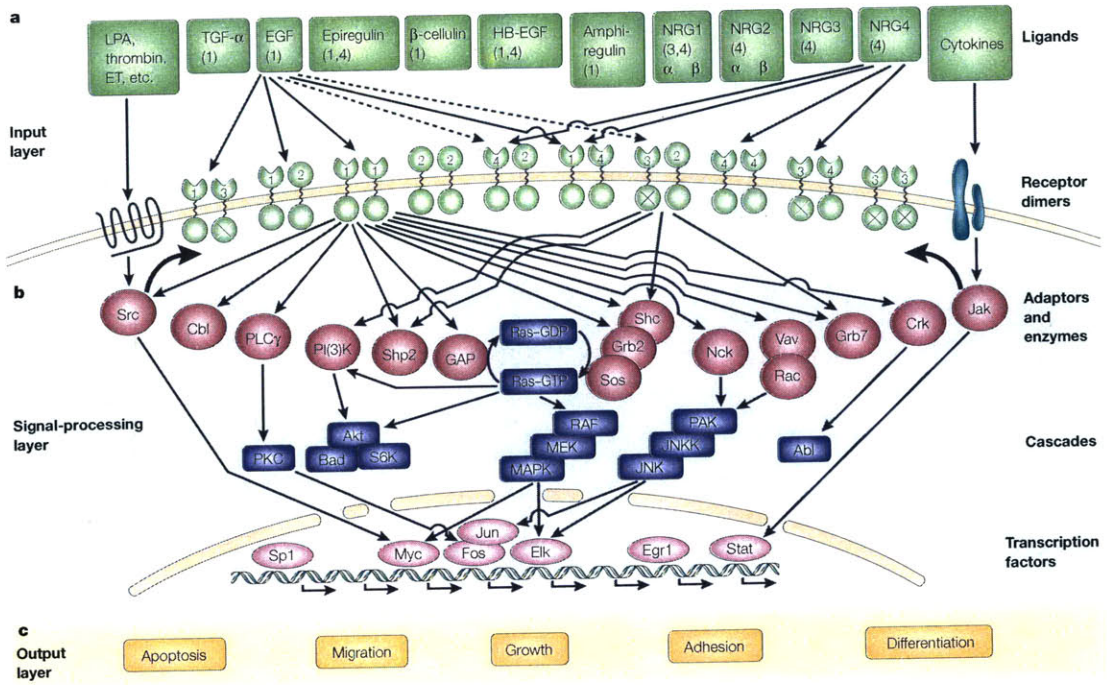


Figure 1-1: a — Ligands and the ten dimeric receptor combinations comprise the input layer. Numbers in each ligand block indicate the respective high-affinity ErbB receptors. For simplicity, specificities of receptor binding are shown only for epidermal growth factor (EGF) and neuregulin 4 (NRG4). ErbB2 binds no ligand with high affinity, and ErbB3 homodimers are catalytically inactive (crossed kinase domains). Trans-regulation by G-protein-coupled receptors (such as those for lysophosphatidic acid (LPA), thrombin and endothelin (ET)), and cytokine receptors is shown by wide arrows. b — Signalling to the adaptor/enzyme layer is shown only for two receptor dimers: the weakly mitogenic ErbB1 homodimer, and the relatively potent ErbB2-ErbB3 heterodimer. Only some of the pathways and transcription factors are represented in this layer. c — How they are translated to specific types of output is poorly understood at present. (Abl, a proto-oncogenic tyrosine kinase whose targets are poorly understood; Akt, a serine/threonine kinase that phosphorylates the anti-apoptotic protein Bad and the ribosomal S6 kinase (S6K); GAP, GTPase activating protein; HB-EGF, heparin-binding EGF; Jak, janus kinase; PKC, protein kinase C; PLCgamma, phospholipase Cgamma; Shp2, Src homology domain-2-containing protein tyrosine phosphatase 2; Stat, signal transducer and activator of transcription; RAF-MEK-MAPK and PAK-JNKK-JNK, two cascades of serine/threonine kinases that regulate the activity of a number of transcription factors.) Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology **2**: 127-137. Copyright 2001. [131]

a small number of specific phosphorylation sites, or the representation of lumped phosphospecies components. This limitation is due primarily to the tradeoff between the number of species and mathematical tractability as well as the methods in which validation data may be generated, for example by the use of single phosphospecies antibodies. In 2005, Oda and Kitano used the standardized form of Systems Biology Markup Language (SBML) to represent the highest resolution of the ErbB network to date [77].

1.3 Global phosphorylation measurements of RTK networks

Monitoring the global state of RTK networks in response to cues, therapeutics, and other factors is fundamental to expanding our understanding of the basic processes at work in the cell as well as our ability to design interventions in diseased states. One successful method for measuring the system under various conditions is the combination of phospho-specific antibodies and high-throughput platforms such as multicolor flow cytometry [89] and bead-based technologies, such as Luminex [91]. Although these platforms are capable of multiplexed measurements across many cellular states and conditions, they are subject to the following limitations: (1) our current understanding of the phosphorylations occurring in the signaling network of interest, (2) the existence of specific antibodies to those phosphorylation sites, and (3) possible antibody interference due to competing protein-protein binding events or protein conformational changes. The first two assumptions are clearly problematic at this relatively early stage in our knowledge of RTK networks. The explosive growth in the documented phosphoproteome over the last decade is a testament to our limited knowledge of the number of phosphorylation sites and their role in the cellular environment.

One of the fundamental reasons our knowledge of the phosphoproteome has increased so drastically in recent years is due to the use of mass spectrometry (MS).

MS is able to exquisitely differentiate small changes in mass due to modifications on residues and therefore it is an excellent measurement tool for discovering those modifications and their location on the protein. Additionally, multiple methods have been developed in order to quantify the relative differences between phosphorylated protein states in different conditions by peptide and protein labeling strategies [1]. One class of labeling strategies is stable isotope labeling with amino acids, SILAC [79]. Alternatively, iTRAQ labeling involves the addition of a label by incorporation of an isobaric tag [88]. Enrichment for phosphorylation is the key step required to measure phosphopeptides given their relatively low abundance compared to nonphosphorylated forms of peptides, and has included techniques such as strong cation exchange (SCX) [116], immobilized metal affinity chromatography (IMAC) [116], and phosphospecific antibody immunoprecipitation [132]. This general strategy, as well as chemical exchanges, have also been used to measure other types of modifications, including lysine acetylation [14], glycosylation [121], and ubiquitination [82].

A number of research groups, studying a variety of different biological problems, have utilized mass spectrometry as a means to discover, as well as quantify, changes in phosphorylation within cells, such as the profiling the yeast phosphoproteome [120] and phosphoproteomic profiling of a variety of lung carcinoma cell lines [86]. The size and complexity of these datasets vary considerably, from tens of sites [104] to thousands [78]. Between strategies for sample separation, phosphorylation enrichment, sample labeling, and instrumentation, there are an overwhelming number of options for MS discovery of phosphorylation in cells and tissue samples. Amidst all of these options is a general trend that discovery, quantitative measurement and reproducibility of phosphorylation measurement is improving continuously. Figure 1-2 shows how coverage of the ErbB phosphotyrosine network has increased roughly two orders of magnitude since the first MS measurements in 2002 by Steen et. al. [104].

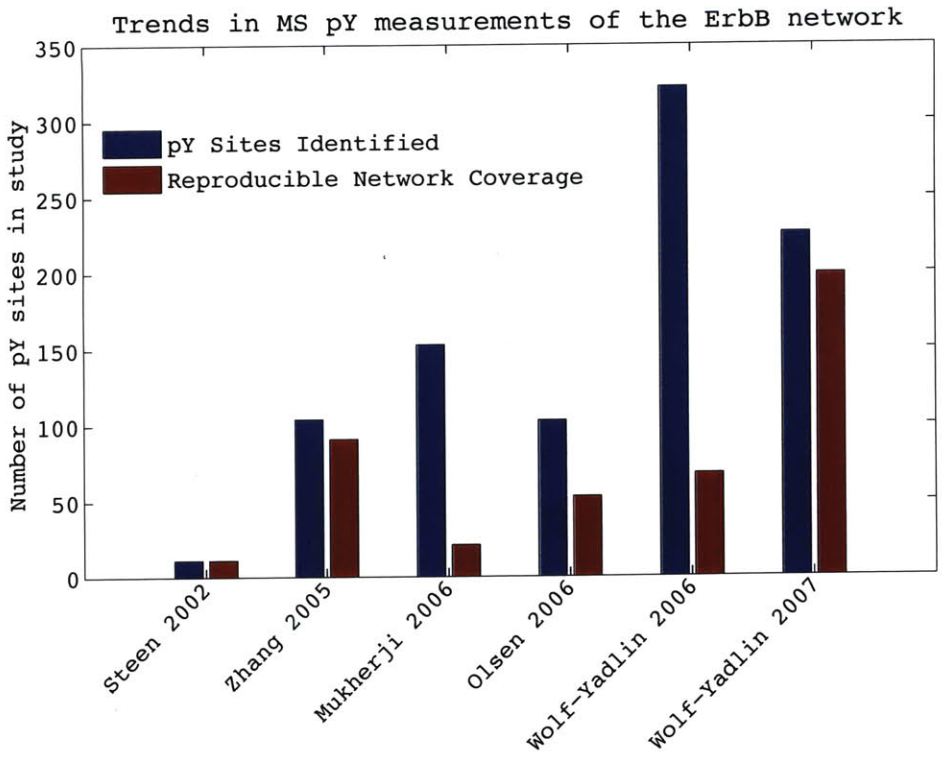


Figure 1-2: Trends in network coverage of the ErbB system. There has been a drastic improvement in the coverage of the network in MS measurement of the ErbB system as well as reproducible measurements of the same nodes across multiple conditions or times. The following studies were used: Steen 2002 [104], Zhang 2005 [132], Mukherji 2006 [69], Olsen 2006 [78], Wolf-Yadlin 2006 [123], Wolf-Yadlin 2007 [122] and the evaluation of reproducibility is based on the number of sites reproduced in all experiments within the study.

1.4 Tools and repositories for high-throughput phosphoproteomic measurements

Given the immensity of the data being generated in phosphoproteomic MS experiments, from the various species, dataset sizes, and the degree of relative quantification, the field has been faced with two fundamental problems. The first is how does one store and make available known measurements? Secondly, how does one generate an understanding about the wealth of this information, from function to regulation of all of the emerging phosphorylation sites in all of these species, cells, and conditions? In

response to these demands, a number of repositories and tools have emerged. Chapter 4 of Liu et. al. [60] gives a nice overview of the current state of phosphoproteomic tools.

Currently, high-throughput phosphoproteomic experimental data is cataloged in repository sources of three different varieties: (1) proteome repositories that catalog the presence of a known modification, such as Uniprot [114], (2) specialized repositories of phosphorylation including Phospho.ELM [22], Phosphosite [39] and dbPTM [57], (3) experiment-specific repositories. There are two categories of experiment-specific repositories, those that have catalog the MS spectra [9] and those that catalog the modification and its quantification within the experiment. At the time of this work only one example of experimental quantification storage existed, PHOSIDA, which is limited to data generated by the lab that developed it [33].

A wealth of tools has been developed for predicting kinase-substrate and phosphopeptide binding domain-substrate relationships. Most capitalize on the fact that recognition by kinases and binding domains is conveyed, in part, by the linear amino acid sequence surrounding the target residue, such as KinasePhos [124], PPSP [126], and Scansite [76]. Scansite [76], for example, uses degenerate peptide library screens [102] to build position-specific scoring matrices (PSSM) representing the specificity of a kinase or binding domain targets. A protein sequence can then be scored against each PSSM and a likelihood of recognition can be quantified. NetworKIN [59] combines Scansite predictions with protein-protein interaction networks to further specify possible kinase-substrate relationships. In addition to tools for predicting kinase-substrate interactions and domain-substrate interactions, other phosphoproteomic tools have focused on analysis of phosphoprotein conservation across species, such as SysPTM [58] and PhosphoBlast [118].

1.5 Motivation of global phosphoproteomic measurement

The explosion in our knowledge of the phosphoproteome (a ten-fold increase in a four year span), due in a large part to improvements of measurement methodologies, is evidence of the interest in understanding the role of phosphorylation in regulating normal cellular function and its role in the genesis and progression of human disease. It is thought that tyrosine phosphorylation represents only 1% of all protein phosphorylation, whereas phosphoserine is much more abundant, representing the majority of protein phosphorylations [42]. Despite the low abundance of tyrosine phosphorylation, according to data in PhosphoSite [39] at the time of this writing, tyrosines actually represent 23% of known phosphorylations in the human proteome. The disparity in these numbers is most likely due to the intense and concerted effort of the field to measure tyrosine phosphorylation. The driving force behind this effort is the important and fundamental role tyrosine phosphorylation plays in signaling transduction, in particular the signaling networks of receptor tyrosine kinases. These experimental studies seek to expand our knowledge regarding the possible phosphorylation states within the cell, how they are regulated, what role they play in the network, and how their dysregulation leads to disease. Understanding the underlying biochemical control and function is pivotal to our understanding of abnormalities and the development of successful therapeutics for the treatment of disease.

Knowing only whether a phosphorylation site is present is insufficient information to discern the function and regulation of a phosphorylation and so MS experiments have sought to measure quantitative differences in the network across various states to help elucidate a functional role for phosphorylation sites. One example is to measure a signaling network in time, following stimulation. For example, EGF stimulation of EGFR and ErbB containing cell lines will kick off a dynamic series of phosphorylation and signaling events that can then be captured by MS. These dynamics can then be parsed to yield a variety of information regarding potential upstream regulatory events and co-regulation among phosphorylations. Alternatively, measurement of dif-

ferential cellular states, such as diseased tissue versus normal tissue, can give insights regarding the mode and function of the differences in disease. Additionally, global phosphoproteomic measurements of a network before and after inhibition by a drug can yield insight into the cellular effects of that drug, indicating potential efficacy and mode of action.

At this point in time, global phosphoproteomic measurements are capable of greatly expanding our current knowledge of cellular networks. For example, a dynamic measurement of the ErbB network in human mammary epithelial cells, the latest experiment in Figure 1-2, shares only a 20% overlap with those modifications depicted in the most complete ErbB system network model by Oda et. al. [77]. In order to incorporate this wealth of knowledge we must first understand the complex regulation of each site and the functional role each site plays in the signaling network.

1.6 The present work

In this thesis we develop tools and algorithms for the analysis of global phosphoproteomic experiments, which also serve as general frameworks for inference and handling of other large-scale quantitative biological measurements. The motivation of this work is to enhance and improve the biological information that can be garnered from quantitative phosphoproteomic measurements, such as hypotheses concerning regulation, function, and interactions of phosphorylation sites within RTK networks, or isolation of key network components responsible for dysregulation. The key concept used in this work relies on the idea that a large dataset, too large to be evaluated as a whole, can be broken into components based on some common feature. These subsets can then be searched in other feature dimensions for enrichment. In addition to linking the two feature dimensions, this method can hypothesize information about those components in the group with unknown function, a "guilty-by-association" method of inference. This thesis demonstrates this concept in a layered manner. In Chapter 2, we show that simple rules of shared regulation downstream of a mutated receptor yield insight regarding a common controlling component. In Chapter 3, we expand the

dimensionalities of both subset selection and subsequent shared information searches to include metadata annotations, such as shared molecular function and cellular localization. In Chapter 4, we take a look at full dataset partitioning through the use of unsupervised learning, a method that has proven useful in the field of gene expression analysis.

In Chapter 2 the feature we focus on is enrichment of the linear amino acid sequence surrounding similarly regulated phosphorylation sites. In order to do this we develop a greedy motif algorithm. When a subset consists of highly co-regulated phosphopeptides, the enriched sequence may yield insight regarding the regulating enzymes or binding partners. This is an important extension, because although tools like KinasePhos [124] or Scansite [76] can predict some of this information, they do not have the capability to discover regulatory motifs for unchartered enzymes and binding partners, in particular phosphatase motifs.

Chapter 3 establishes a web-based resource, PTMScout, now available to the phosphoproteomic community at large. PTMScout provides a flexible interface for arbitrary subset generation, as well as providing the only repository of its kind that allows for the full scientific community to store and analyze experimental datasets regarding phosphorylation. The large degree of metadata present in PTMScout, annotations of the biological molecules within experimental datasets, enables the high-throughput framework developed in Chapter 4 for the analysis of unsupervised learning parameters. The framework developed in Chapter 4, like PTMScout developed in Chapter 3, focuses on enabling scientists in the community to generate the maximum amount of biological hypotheses from quantitative experiments by decreasing the barrier to utilization of specialized computational tools.

In addition to the methods and tools developed, this thesis also includes biological inference concerning the ErbB network. For example, in Chapter 2 we pose a direct link between a variant of EGFR correlated with poor prognosis of patients with glioblastoma, EGFRvIII, and increased activity of CK2, indicating a potential intervention point for treatment of a currently untreatable disease. In Chapters 3 and 4, a variety of biology is posed, including implications of components involved

in the crosstalk of the EGF receptor and focal adhesions, extension of our current knowledge regarding phosphoproteins involved in endocytosis of the receptor, and hypotheses regarding roles for several of EGFR phosphorylation sites. The biological relationships shared in each of the chapters represents only a fraction of those available, but ideally it establishes the usefulness of each of the methods posed in generating relevant biological hypotheses.

Chapter 2

An Integrated Comparative Phosphoproteomic and Bioinformatic Approach Reveals a Novel Class of MPM-2 Motifs Upregulated in EGFRvIII-Expressing Glioblastoma Cells

2.1 Summary

Glioblastoma (GBM, WHO grade IV) is an aggressively proliferative and invasive brain tumor that carries a poor clinical prognosis with a median survival of 9 to 12 months. In a prior phosphoproteomic study performed in the U87MG glioblastoma cell line, we identified tyrosine phosphorylation events that are regulated as a result of titrating EGFRvIII, a constitutively active mutant of the epidermal growth factor

receptor (EGFR) associated with poor prognosis in GBM patients. In the present study, we have used the phosphoserine/phosphothreonine-specific antibody MPM-2 (mitotic protein monoclonal #2) to quantify serine/threonine phosphorylation events in the same cell lines. By employing a bioinformatic tool to identify amino acid sequence motifs regulated in response to increasing oncogene levels, a set of previously undescribed MPM-2 epitope sequence motifs orthogonal to the canonical "pS/pT-P" motif was identified. These motifs contain acidic amino acids in combinations of the -5, -2, +1, +3, and +5 positions relative to the phosphorylated amino acid. Phosphopeptides containing these motifs are upregulated in cells expressing EGFRvIII, raising the possibility of a general role for a previously unrecognized acidophilic kinase (e.g. casein kinase II (CK2)) in cell proliferation downstream of EGFR signaling.

2.2 Introduction

Glioblastoma (GBM, WHO grade IV) is a complex disease driven by a number of genetic aberrations that dysregulate normal cellular processes such as proliferation, apoptosis and cell cycle control [31]. In particular, expression of EGFRvIII, a constitutively active mutant of the epidermal growth factor receptor (EGFR), promotes GBM cell proliferation and survival by preventing cell cycle arrest upon serum withdrawal [72]. This loss in serum dependency has been attributed to a downregulation of the cyclin-dependent kinase (CDK) inhibitor p27 as a result of phosphatidylinositol 3-kinase (PI3K) activation by EGFRvIII [72]. Improved characterization of the regulatory network by which EGFRvIII alters mitotic processes in GBM would not only provide further insight into its mitogenic signaling networks but also generate a broader inventory of candidate target genes that may serve as points of therapeutic intervention.

While proximal signals downstream of receptor tyrosine kinases (RTKs) such as EGFR are largely propagated by tyrosine phosphorylation, distal cellular processes are often the consequence of serine/threonine phosphorylation events, which comprise more than 99% of the phosphoproteome. This large background makes the enrich-

ment of interesting phosphoproteomic subsets, such as mitogenic signaling proteins, particularly challenging [42]. This problem is highlighted by a recent global phosphoproteomic study of EGF-mediated signaling in HeLa cells where fewer than 10% of the identified phosphorylation sites were found to be responsive to EGF stimulation [78]. In order to overcome this limitation in global phosphoproteomic analysis, we have devised a sequential immunoprecipitation (IP) strategy coupled to mass spectrometry (MS) that builds on a previously described phosphotyrosine-enrichment approach to quantify the mitotic phosphoproteome downstream of EGFRvIII, Figure 2-1 [41].

To access the subset of phosphoserine and phosphothreonine modifications in the mitotic compartment, we have employed MPM-2, a monoclonal antibody, derived from mitotic HeLa cell lysates, that recognizes a wide variety of mitotic phosphorylated antigens [19]. Despite its widespread use in the literature as a marker of serine/threonine phosphorylation in mitotic cells, only a small number of the substrates recognized by MPM-2 have been identified [106, 125]. Furthermore, only limited characterization of the *in vivo* phosphorylation sites of these substrate proteins has been performed. However, *in vitro* peptide library screens have shown that the binding specificity of MPM-2 is dominated by the "pS/pT-P" motif commonly propagated by the cyclin-dependent kinases (CDKs) and mitogen-activated protein kinases (MAPKs) [87, 130].

Quantitative phosphoproteomic mass spectrometry offers the ability to analyze the effects of different conditions, treatments, and cell lines on the global phosphorylation-mediated state of intracellular signaling [93, 132]. In order to obtain mechanistic insight into how changes in phosphorylation affect cell phenotype it is necessary to combine the data from quantitative phosphoproteomics with additional information, including protein sequence surrounding the phosphorylation site. Kinases that generate phosphosites, phosphopeptide-binding domains that use phosphosites as signals to prompt a response, and phosphatases that remove phosphosites are all regulated in part by the amino acid sequence surrounding the phosphorylated residue [47,127,129]. There is a great deal of literature and a number of online resources linking linear amino acid sequence motifs to associated kinases and binding domains [4, 76]. Here

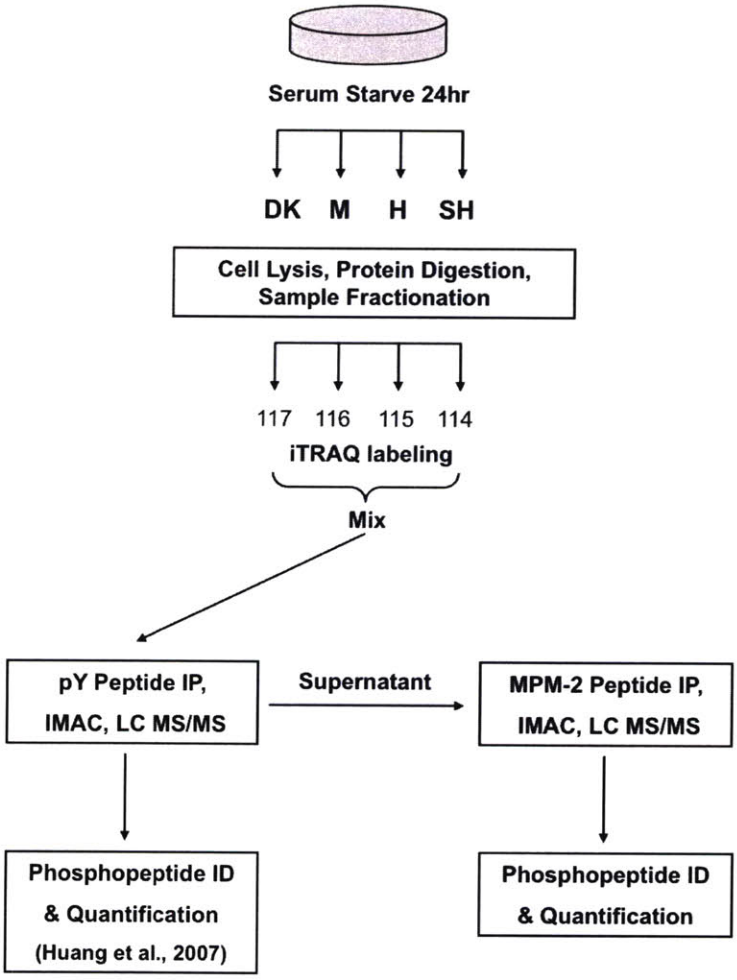


Figure 2-1: Experimental workflow for MS discovery of phosphorylation. U87MG sublines (U87-M, 1.5 x 10^6 copies/cell; U87-H, 2.0 x 10^6 copies/cell; U87-SH, 3.0 x 10^6 copies/cell; U87-DK, 2.0 x 10^6 inactive copies/cell) were serum starved for 24 hours prior to cell lysis and protein digestion. Digested peptides were stable-isotope labeled with the isobaric iTRAQ reagent, mixed and subjected to phosphotyrosine immunoprecipitation (IP) using a pan-specific phosphotyrosine antibody [41]. Mitotic phosphopeptides were then immunoprecipitated from the supernatant with the MPM-2 antibody. Eluted phosphopeptides were further enriched with immobilized metal affinity chromatography (IMAC) prior to liquid chromatography tandem mass spectrometry analysis (LC-MS/MS). Phosphopeptide identification (ID) and quantification was performed as described in the methods.

we describe a bioinformatics tool to identify amino acid sequence motifs significantly enriched among the phosphopeptides associated most strongly with various expression levels of EGFRvIII. We anticipate that this new motif information will lead to enhanced mechanistic biological insight by connecting the probed processes to sequence motifs associated with known molecules and molecular functions and by revealing motifs of unknown biological function that can be explored further. We also expect that our new method will prove useful in many other problems of interest in basic cellular biochemistry and in therapeutics discovery applications.

2.3 Results

To characterize the effect of EGFRvIII on the mitotic cellular signaling networks, we have utilized the MPM-2 antibody to enrich for peptides containing sites of serine and threonine phosphorylation from U87MG glioblastoma cell lines with titrated levels of the EGFRvIII. A previous phosphoproteomic study of EGFRvIII receptor-mediated signaling has determined the effect of titrating EGFRvIII receptor levels on phosphotyrosine-driven networks [41]. We now build on those foundational findings by investigating a key subset of serine/threonine substrate phosphorylation sites up-regulated by EGFRvIII expression in this same battery of cell lines. Since this study is focused on signaling downstream of the EGFRvIII receptor, cells were subjected to serum starvation prior to analysis to minimize any confounding signaling events that may arise from components in serum and cell culture media. After depleting phosphotyrosine-containing peptides using the pan-specific phosphotyrosine antibody PY100, the iTRAQ-labeled supernatant was subjected to a subsequent immunoprecipitation using the MPM-2 monoclonal antibody, Figure 2-1. Peptides eluted from the MPM-2 IP were further enriched for phosphopeptides using immobilized metal affinity chromatography (IMAC) prior to liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis. Two biological replicates were performed, resulting in the identification and quantification of 87 unique sites of phosphorylation on 68 phosphopeptides (58 proteins), Table 2.1. Of these sites, 11 were found to be novel

with respect to the resources Phospho.ELM [22], PHOSIDA [33], PhosphoSitePlus (www.phosphosite.org), and a recent study of mitotic phosphoproteins [20]. Three of the sites have not been detected previously in humans, but only in homologous proteins.

This phosphoproteomic analysis is, to our knowledge, the most extensive characterization of MPM-2 substrates to date. Our present study is also distinct from previous MPM-2 proteomic analyses in that our MS analysis provides quantitative information on *in vivo* MPM-2 substrates with site-specific resolution. A previous IVEC screen to identify MPM-2 substrates in Xenopus embryo extracts was performed by Stukenberg et al. and identified 20 candidate proteins that underwent mitotic phosphorylation [106]. More recently, a proteomic study of MPM-2 substrates performed using 2D gel electrophoresis identified [101] MPM-2 candidate substrate proteins [125]. Strikingly, there is no overlap between the proteins identified in these two studies and our current analysis. The lack of similarity in the datasets is likely due to context-dependent variation, including the use of different cell lines and chemically-induced cell cycle synchronization or mitotic activation in previous studies [106, 125] compared to asynchronously cycling EGFRvIII-expressing cells in the current study. Additionally, we have performed substrate isolation using solution-based peptide IP coupled to mass spectrometry, an approach that may yield different substrates from the cDNA screens and 2D gel electrophoresis analysis carried out in the two prior studies.

Ectopic expression of EGFRvIII in U87MG cells results in an increased proliferation rate and a larger G2-M cell population under serum depravation conditions [40, 72]. Consistent with the well-recognized binding affinity of MPM-2 to phosphoproteins in mitotic cells, we observe that phosphorylation of the established proliferation markers Ki-67 and MCM3 (minichromosome maintenance protein 3) were upregulated 1.4-fold and 2.7-fold respectively in the U87-H subline, which expresses a high level of EGFRvIII, compared to the U87-DK kinase-dead control cells [37, 101]. It has previously been demonstrated that EGFRvIII downregulates p27 expression via activation of the PI3K pathway, resulting in an increase in CDK2-cyclin activity

Table 2.1: MS detected MPM-2 substrates in U87 EGFRvIII expressing cells. There were 68 phosphopeptides measured, covering 58 proteins, following enrichment of U87 cells expressing a kinase dead (DK) EGFRvIII and medium (M), high (H), and super high (SH) levels of EGFRvIII.

gene name	trypsinized phosphopeptide	site	DK	M	H	SH
AFF4	MFsPMEEK	S694;	0.48	0.67	1	0.89
AKAP11	SSAFsPLGGCTPAECFCQTDIGGDR	S456;	0.73	1.09	1	1.27
ARFGAP1	EWSLESSPAQNWtPPQPR	T135;	0.53	0.79	1	1
ATF2	MPLDLsPLATPIIR	S112;	0.62	0.67	1	1.15
BCLAF1	AEGEWEDQEALDYFsDKESGK	S385;	0.54	0.74	1	0.87
C14orf106	EFLEQLPKDDHDDFFSTtPLQHQR	T993;	0.63	0.94	1	1.18
CHD8	HFSTLKDDDLVEFsDLEsEDDERPR	S1420;S1424;	0.39	1.43	1	1
CTAGE5	EHsPYGPsPLGWPSSETR	S442;S447;	0.13	1.4	1	1.5
CTAGE5	EHSPYGPsPLGWPSSETR	S447;	0.2	1.36	1	1.73
EHD1	DKPTYDEIFYTLsPVNGK	S456;	0.54	0.68	1	1.23
EHD4	DKPVYDELFYTLsPINGK	S459;	0.68	0.73	1	1.33
EIF3C	QPLLLsEDEEDTKR	S39;	0.74	0.79	1	0.86
EIF4EBP1	VVLGDGVQLPPGDYSTtPGGTLFSTtPGGTR	T37;T46;	0.56	0.62	1	0.89
EIF4EBP1	VVLGDGVQLPPGDYSTTPGGTLFSTtPGGTR	T45;T46;	0.59	0.76	1	1.02
EIF4EBP2	TVAISDAAQLPHDYCTtPGGTLFSTtPGGTR	T37;T46;	0.64	1.11	1	1.04
ERCC6	KVPVQEIDDDFFPssGEEAEAASVGEGGGGGRK	S429;S430;	0.69	0.98	1	1.1
FAM33A	QTDLELsPLTKEEK	S101;	0.48	0.6	1	1
FAM40B	RYDRPQDSEFsPVDNCLQSVLGQR	S788;	0.64	0.7	1	0.92
FASN	ADEASELACPtPKEDGLAQQQTQLNLR	T2204;	0.71	1.16	1	1.02
FLJ20297	QLPDCIVGEDGLILtPLGR	T640;	0.58	1.1	1	1
FZR1	RSsPDDGNDVSPYSLsPVSNK	S138;S151;	0.44	0.82	1	0.88
FZR1	SSPDDGNDVSPYSLsPVSNK	S151;	0.6	0.81	1	0.91
HERC1	DRWIsENQDSADVDPQEHSFTR	S1328;	0.42	2.35	1	1.12
HNRPF	ATENDIYNFFsPLNPVR	S310;	0.55	0.66	1	0.79
KIAA0460	DVEDMELsDVEDDGSKIIVEDRK	S337;	1	0.85	1	0.96
KIAA1458	RGTFsDQELDAQSLDDEDDNMHHAVYPAVNRFsPsPR	S315;S345;S343;	0.57	1.06	1	1.04
LOC439961	LTDEDFsPFGSGGGLFSGGK	S104;	0.45	0.6	1	1.14
LOC440991	DEILPTtPISEQK	T221;	0.68	0.73	1	0.84
MAP1A	ELVLssPEDLTQDFEEMKR	S526;S527;	0.85	1.08	1	1
MAP1B	SVNFSLtPNEIK	T1156;	0.72	0.69	1	1.23
MAP1B	SDISPLtPRESsPLYsPTFSDSTSAVK	T1662;S1667;S1671	0.67	0.62	1	0.72
MAP1B	AAEAGGAEEQYGFLTtPTK	T941;	0.78	0.64	1	1.23
MCM3	DGDSYDPYDFsDTEEEMPQVHTPK	S711;	0.37	0.63	1	0.88
MCM3	DGDSYDPYDFsDTEEEMPQVHtPKTADSQETK	S711;T722;	0.37	0.92	1	0.99
MEF2A	GCDsPDPDTSYVLtPHTEEK	S98;T108;	0.48	0.67	1	0.71
MKI67	AAVGEEKDINTFVGtPVEK	T1923;	0.72	0.98	1	0.95
MPHOSPH10	SDLRKsPVFsDEDsDLDFDISKLEQQSK	S163;S171;S167;	0.4	0.57	1	0.92
MPHOSPH6	DHANYEEDENGDItPIK	T147;	0.05	0.71	1	1.23
NKAP	IGELGAPEVWGLsPK	S149;	0.6	0.85	1	1.03
NUMA1	LPPKVEsLESLYFtPIPAR	S1755;T1762;	0.51	0.78	1	1.46
NUP98	NLNNSNLFsPVNRDSENLAsPSEYPENGER	S595;S606;	0.51	0.77	1	0.74
PDE5A	EQMPLtPPRFDHDEGDQCSR	T137;	4.67	1.36	1	1.32
PGRMC1	LLKEGEEPTVYsDEEEPKDESAR	S181;	0.55	0.54	1	0.68
PGRMC2	LLKPGEEPSEYtDEEDTKDHNKQD	T211;	0.6	0.86	1	1.1
PRPF31	SSGTAsSVAFtPLQGLEIVNPQAAEK	S450;T455;	0.72	1.2	1	1.01
RANBP2	KKPEDSPSDDDVLIVYELtPTAEQK	T2639;	0.54	0.74	1	1.04
RB1	DREGPTDHLESACPLNLPLQNNHTAADMYLsPVRsPK	S608;S612;	0.44	0.63	1	0.99
RBL1	EKEAVItPVASATQSVSR	T385;	0.47	0.75	1	1.13
RCAN1	QFLIsPPAsPPVGWK	S163;S167;	0.7	0.74	1	1.22
RIF1	NYTEDIFPVtPPELEETIRDEK	T702;	0.61	0.97	1	1.25
RRM2	VPLAPITDPQQLQLsPLK	S20;	0.52	0.57	1	1.01
SCD	GSTLDLsDLEAEK	S203;	0.36	0.91	1	1.13
SDCCAG1	NPYLLsEEEDDDVDGDVNVEKNETEPPKGK	S417;	0.5	0.95	1	1.1
SMARCAD1	RNDDIsELEDLSELEDLKDAK	S146;	0.47	0.73	1	1.13
SMARCAD1	RNDDIsELEDLsELEDLKDAK	S146;S152;	0.78	0.82	1	1.03
SON	SFsIsPVR	S2011;S2013;	0.75	0.72	1	0.86
SQSTM1	SRLtPVSPESSSTEEK	T269;	1.08	0.54	1	1.2
SQSTM1	SRLtPVsPESSSTEEK	T269;S272;	0.95	0.62	1	1.1
SRRM2	GEFSAsPMLK	S1124;	0.69	0.89	1	1.03
SRRM2	ELSNsPLRENSFGsPLEFR	S1320;S1329;	0.64	0.92	1	1.13
SSB	FAsDDEHDEHDENGATGPVKR	S366;	0.52	0.7	1	0.64
SURF2	DLGSTEDGDGtDDFLtDKEDEKAKPPR	T190;T195;	0.48	0.85	1	1.03
SURF2	DLGSTEDGDGTDDFLtDKEDEK	T195;	0.45	0.85	1	1.26
THRAP3	NREEEWDPEYtPK	T874;	0.69	0.94	1	1.23
TMEM51	YYVPsYEEVMNTNYSEAR	S133;	0.7	1.08	1	1.15
TOP2B	ASPITNDGEDEFVPsDGLDKDEYTFSPGK	S1408;	0.93	1.25	1	1.1
YTHDC2	STDSSSYPsPCAsPSPPSSGK	S1263;S1267;	1	0.54	1	0.76
ZC3H13	GNIETTSEDGQVFsPK	S993;	0.57	0.57	1	0.92

and Rb (retinoblastoma protein) hyperphosphorylation, allowing cells to enter the cell cycle [72]. In line with this result, we also observe that phosphorylation of Rb (S608 & S612) and the Rb family member p107 (T385) increase more than 2-fold in the U87-H cells compared to the control U87-DK cells. These phosphorylation sites directly precede proline residues, a characteristic motif recognized by proline-directed kinases such as the CDKs [59].

Of the 58 proteins identified in this study, only 8 are annotated in the Gene Ontology database as having a role in the cell cycle. It is surprising that only 15% of the phosphopeptides immunoprecipitated by MPM-2 have a previous association with cell cycle, especially given that MPM-2 is considered to specifically recognize substrates in proliferating cells and mitotic cell lysates [19]. Nonetheless, those proteins that are labeled as having the GO process annotation term cell cycle are enriched (p-value of 0.01) in the subgroup of peptides whose phosphorylation level is upregulated is in the top quartile in the U87-H cell line as compared to the control U87-DK cells.

Intriguingly, only 59 of the 87 phosphorylation events identified in this study were on a serine or threonine residue followed by a proline, Figure 2-2A. Of the 28 remaining sites, 20 had an aspartic or glutamic acid in the +1 position, directly to the C-terminal side of the phosphorylated residue. Moreover, 16 of the 68 phosphopeptides identified in the MS study contained at least one "pS/pT-D/E" site, and no "pS/pT-P" site, demonstrating that a large fraction of the acid-directed sites were specifically recognized by the MPM-2 antibody, and were not merely neighbors to proline-directed sites on the same peptide. To ensure that this surprising departure from the canonical MPM-2 epitope was not a byproduct of non-specific binding, a degenerate peptide library experiment was performed to determine whether non-"pS-P" containing motifs could be directly recognized by the MPM-2 antibody. Peptide libraries were synthesized on a cellulose membrane and immunoblotted with MPM-2 to discover the *in vitro* affinity of MPM-2 for positional dependence and amino acid composition of favorable motifs, Figure2-2. Importantly, due to the prevalence of the "pS-P" motif in MPM-2 literature, all libraries degenerate at the +1 position excluded proline in the +1 position, in an attempt to minimize the effect of what might

be a dominant interaction. The results show that MPM-2 binds directly to peptides containing acidic residues in the +1 position, as well as to peptides containing proline, at levels significantly above the background. In fact, the *in vitro* binding of MPM-2 to the "pS-E" and "pS-D" libraries is higher than to "pS-P".

The largest positional variance occurs in the -1 and +2 positions, where the aliphatic and aromatic amino acids, I, L, F, and V increase the affinity for MPM-2 significantly over pS alone. These results are in very good agreement with two previous degenerate peptide studies that also found aliphatic and aromatic dependence in the -1 and +2 positions [87, 130]. In comparison with these previous studies, the most significant difference seen in this study is the preference for acidic residues in the +1 position in addition to the canonical pS-P. This acidic motif was obscured in one of the previous studies because Yaffe et al. chose to fix both the pS and a +1 proline in response to an initial screen that showed heavy +1 proline selectivity [130], although significant acidic residue preference was also detected in this initial screen (data not shown). The experiment by Rodriguez et al. did allow for degeneracy in the +1 position [87], yet only detected a preference for glycine and proline in this position. It is possible that the discrepancy between this study and our current results may be due to a mixture of pS and pT in the oriented position in the Rodriguez et al. screen, especially given that, as shown in Figure 2-2C, phosphothreonine, but not phosphoserine, alone in many positions is sufficient to bind to MPM-2 in a negative control library with a fixed, oriented non-phosphorylated serine residue.

The selectivity of MPM-2 for pT is also evident in our MS phosphoproteomics data set, where 33% of the phosphorylated sites discovered in this study are phosphothreonines, a 3-fold increase over the previously reported pT:pS ratio in the phosphoproteome [42]. This data also demonstrates that MPM-2 strongly favors pT followed by proline in the +1 position, as 87% of the pT sites enriched by MPM-2 match this "pT-P" motif. To gauge the specificity of MPM-2 for pT and surrounding amino acids, we compared our results to the general composition of the human phosphoproteome to date, as represented by the Phospho.ELM [22] database of identified protein phosphorylation sites. By comparison, pT represents approximately 14% of

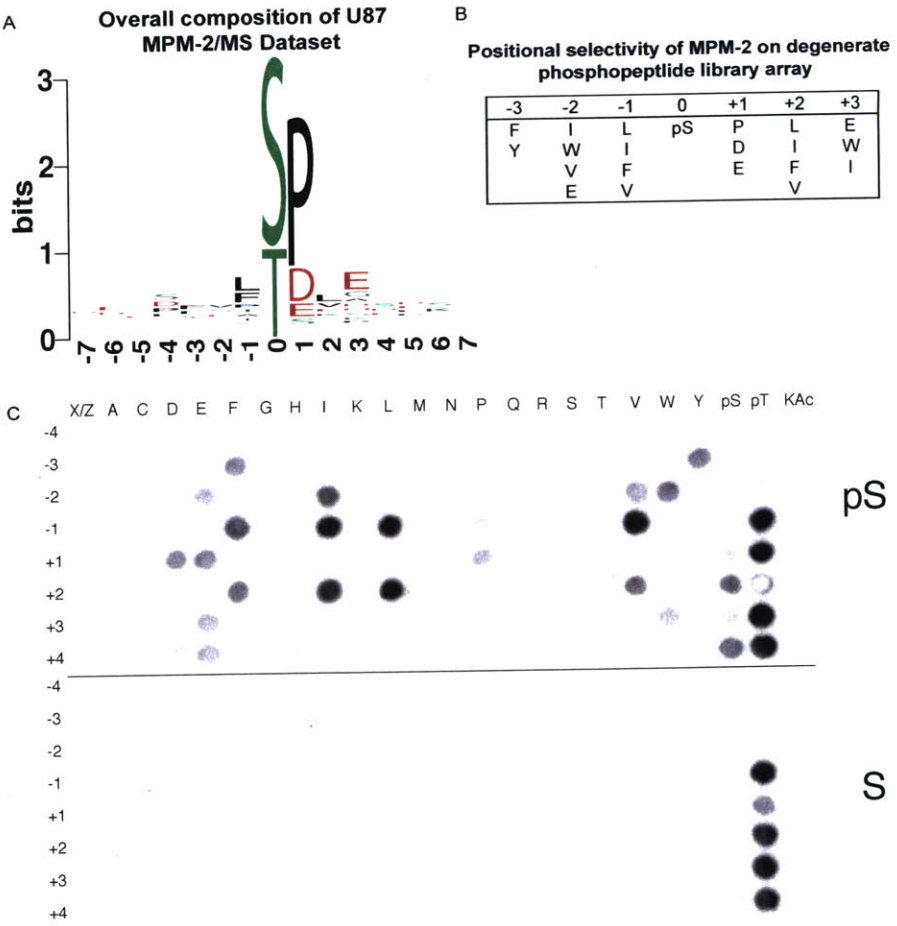


Figure 2-2: Degenerate library screen of MPM-2 specificity. (A) Motif logo of mass spec dataset [17]. The height of each amino acid represents its frequency at that position, and the total stack height of a position represents total conservation. (B) Positional selectivity of two-fold or greater (see methods) of MPM-2 for residues surrounding a phosphorylated residue according to the results of a degenerate peptide library screen. (C) A degenerate library screen of MPM-2 selectivity. The blot is composed of two sections, the top-half contains a phosphoserine-oriented library and the bottom half is a serine-oriented control library. On each spot is an entire degenerate library oriented on the central serine residue, i.e. X-X-X-X-pS/S-Z-X-X-X, where X represents all naturally occurring amino acids except cysteine and Z additionally excludes proline. A second position, indicated by the row numbering position with respect to the orienting pSer or Ser residue, is fixed to a particular residue, indicated by column-wise position. All natural amino acids, as well as phosphoserine, phosphothreonine, and acetylated lysine, were tested for their contribution to MPM-2 recognition. Phosphoblot quantitation can be found in Table A.3.

the known human phosphoproteome, and 44% of these sites are have a proline in the +1 position. Since the phosphoproteome is still largely uncharacterized, it is difficult to predict whether this percentage is reflective of the true biological composition of pT sites, or whether it is the result of study bias. However, it is clear that there is enrichment for "pT-P" phosphorylation sites in our data set beyond that which can be accounted for in the known human phosphoproteome, indicating good agreement for phosphothreonine with the canonical "pS/pT-P" MPM-2 epitope.

To highlight the effect of EGFRvIII on mitotic regulatory networks, we performed a motif enrichment analysis (see Methods) on the sequence surrounding the mapped phosphorylation sites in the peptides captured by the MPM-2 antibody and upregulated in the top quartile of all detected phosphosites in cells expressing either a medium, high, or super-high level of EGFRvIII relative to the kinase-dead negative control. Surprisingly, the motifs enriched in the EGFR expressing cells did not contain the C-terminal proline corresponding to the generally accepted specificity of the MPM-2 antibody [106, 119, 130]. Though the motif "pS/pT-P" was present in about half of the phosphopeptides in the top quartile for each of the U87-M, U87-H, and U87-SH cell lines, it was present in a higher fraction (over two-thirds) of the total data, and was therefore not enriched among sites upregulated downstream of EGFRvIII signaling and captured by MPM-2. Instead, a number of acid-directed motifs were found significantly enriched among EGFRvIII-regulated phosphosites (Table 2.2 for U87-H cells and Tables A.1 and A.2 for U87-M and U87-SH cells). Motifs were found containing aspartic acid, glutamic acid, or both, at positions -5, -2, +1, +3, and +5 relative to the phosphorylated residue.

The rate of false positive motif discovery must be considered because the statistical significance of a large number of amino acid sequence motifs has been calculated in this study. Familywise error rate control using traditional Bonferroni correction would be overly penalizing considering the extremely large search space. Therefore, to approach the question of whether the sequence motif enrichments we observed might be spurious false positives, we took an empirical approach. We generated 1000 random foregrounds of 25 phosphosites (corresponding in size to our foregrounds of

Table 2.2: Motifs significantly enriched among top quartile of MPM-2 antigen peptides upregulated in U87-H cells vs. U87-DK controls.

Motif[1]	Motif in Foreground	Motif in Background	Foreground Size	Background Size	Statistical Significance
D.x	8	10	25	95	2.73×10^{-4}
-.x	12	20	25	95	3.27×10^{-4}
-.s	10	15	25	95	3.98×10^{-4}
-.x-	8	11	25	95	8.24×10^{-4}
-.s....E	5	5	25	95	9.17×10^{-4}
-..D.x	5	5	25	95	9.17×10^{-4}
-.s-.-	7	9	25	95	1.05×10^{-3}
-.s....-	6	7	25	95	1.17×10^{-3}
D.x-	6	7	25	95	1.17×10^{-3}
x-	12	23	25	95	2.09×10^{-3}
D.s	6	8	25	95	3.80×10^{-3}
-.s-.E	6	8	25	95	3.80×10^{-3}
-..-.x	6	8	25	95	3.80×10^{-3}
-.xD	6	8	25	95	3.80×10^{-3}
-.sD.E.E	4	4	25	95	3.97×10^{-3}
-.s.L..-	4	4	25	95	3.97×10^{-3}
sD.-.-O	4	4	25	95	3.97×10^{-3}
-..D.x-	4	4	25	95	3.97×10^{-3}
D.xD	4	4	25	95	3.97×10^{-3}
-...xP..S	4	4	25	95	3.97×10^{-3}
D.s-.E	5	6	25	95	4.48×10^{-3}
-.sD.-.-	5	6	25	95	4.48×10^{-3}
-.s-L-	5	6	25	95	4.48×10^{-3}
xD.E.E	5	6	25	95	4.48×10^{-3}
s-	10	19	25	95	5.81×10^{-3}
xD.-.-	7	11	25	95	6.47×10^{-3}
-.s.L	6	9	25	95	9.30×10^{-3}
s-.-	9	17	25	95	9.30×10^{-3}
sD.-.-	6	9	25	95	9.30×10^{-3}
-..x	6	9	25	95	9.30×10^{-3}
x-.-	10	20	25	95	9.60×10^{-3}
xP	13	64	25	95	0.983

[1] "s" = pS, "x" = pS/pT, "." = Any amino acid, "-" = D/E, "O" = M/I/L/V

interest), and tabulated the number of detected enriched motifs in each, as well as the statistical significance of the most significant discovered motif. Only 6.8%, 1%, and 2%, respectively, of random foregrounds have as many statistically significant ($p < 0.01$) motifs identified as the foregrounds built from the top quartile of sites in U87-M,

U87-H, or U87-SH cells relative to the U87-DK control, Figure 2-3A. Moreover, only 5.6%, 6%, and 6% of random foregrounds have a motif with a statistical significance as significant as the most significant motif found among the top quartile of phosphosites in U87-M, U87-H, or U87-SH cells, respectively, as compared to U87-DK controls Figure 2-3B. Finally, 2.6%, 0.8%, and 1.5% of random foreground datasets have both as many motifs and as significant a strongest motif as foregrounds generated from each of the three EGFRvIII-expressing cell lines U87-M, U87H, and U87-SH. Taken together, these empirical metrics indicate that our motif analyses are identifying a biologically relevant phenomenon.

2.4 Methods

2.4.1 Cell Culture and Retrovirus Infection

For this study, U87MG glioblastoma cells lines were transfected with EGFRvIII and sorted to generate sublines that express the following mean receptor levels, U87-M (1.5 x 10^6 copies/cell), U87-H (2.0 x 10^6 copies/cell) and U87-SH (3.0 x 10^6 copies/cell). In addition, we have included the U87-DK subline that expresses 2.0 x 10^6 copies/cell of a kinase-dead version of the EGFRvIII receptor as a negative control for its activity [40,41]. Cell lines were cultured in DMEM with 10% fetal bovine serum, 2 mM glutamine, 100 units/ml penicillin, and 100 mg/ml streptomycin in 95% air/5% CO2 atmosphere at 37°C. U87MG cells expressing EGFRvIII or DK receptors were selected in 400 μg/ml G418. For U87MG cells expressing titrated levels of EGFRvIII, a bulk population of cells was prepared by retroviral transduction with pLERNL and stained as previously described [41] with an anti-EGFR monoclonal antibody Ab-1 (clone 528; Oncogene Science, Cambridge, MA), followed by fluorescein isothiocyanate-conjugated goat anti-mouse Ig antibody (PharMingen, Minneapolis, MN) and sorted for medium (1.5 x 10^6 receptors, U87-M), high (2.0 x 10^6 receptors, U87-H), and superhigh (3.0 x 10^6 receptors, U87-SH) receptor amounts. For this procedure, U87-EGFRvIII cells engineered previously and determined to express 2 x

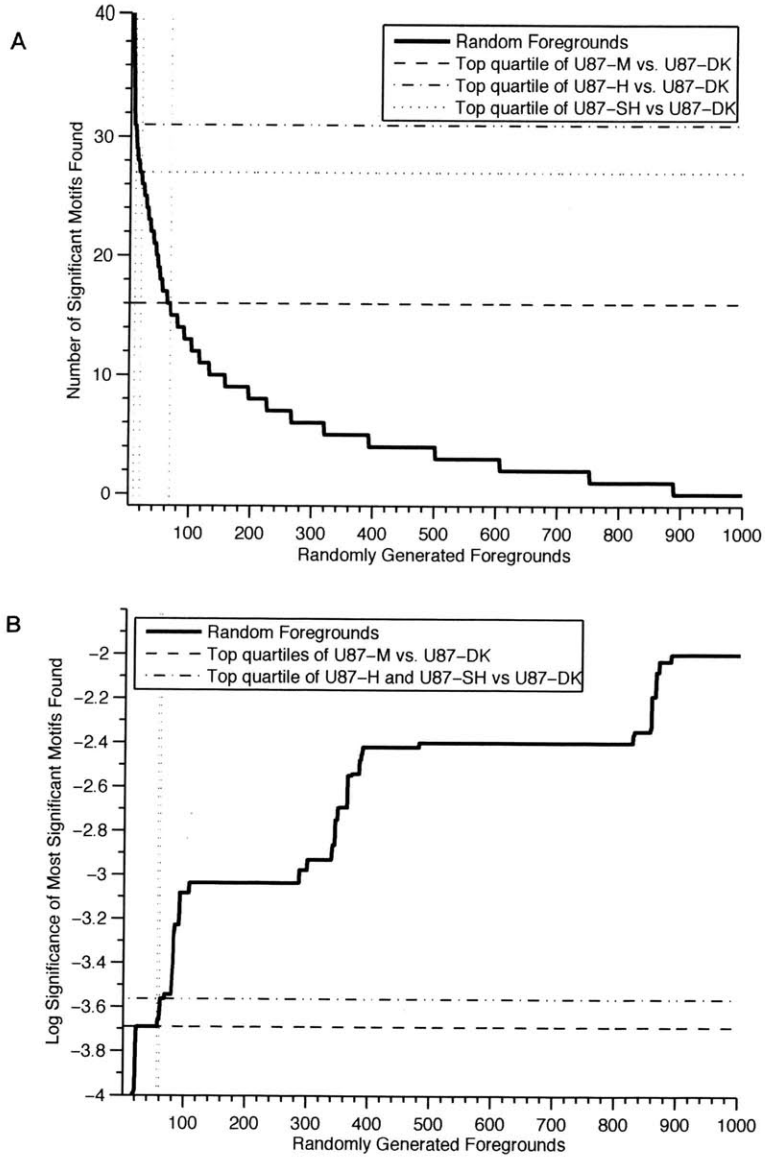


Figure 2-3: Comparison of motif enrichment analyses of regulated data with motif enrichment analyses of randomly selected phosphosites. Motif enrichment analyses of the top quartile of upregulated phosphosites in cells expressing medium, high, and super-high amounts of EGFRvIII versus a kinase dead control were compared to 100 analyses of a randomly selected quarter of phosphosites. (A) Comparison of the statistical significance of the most significant motif found in each analysis. (B) Comparison of the number of motifs found in each analysis. Vertical lines indicate the number of random foregrounds with a metric value the same as, or more extreme than, the values for the MS-motivated foregrounds.

10^6 receptors per cell were used as a gating control. The sorted cells were then maintained in culture and receptor levels were analyzed again by flow cytometry prior to experimental use.

2.4.2 Cell lysis, Protein digestion and Peptide fractionation

U87MG cells were maintained in DMEM medium supplemented with 1% FBS. 1.5x10^6 cells per 10cm plate were seeded for 24 hours, then washed with PBS and incubated for 24 hrs in serum-free media. Cells were lysed in 1 ml of 8 M urea. For each of the two biological replicates performed, lysate from three 10 cm plates were pooled together. Cells were reduced with 10 mM DTT for 1 hr at 56°C, alkylated with 55mM iodoacetamide for 45 min at room temperature, and diluted to 12 ml with 100mM ammonium acetate, pH 8.9, prior to digestion with 40 μg of trypsin (Promega). The lysates were digested overnight at room temperature. Digested lysate were acidified to pH 3 with acetic acid and loaded onto a C18 Sep-Pak Plus cartridge (Waters). The peptides were desalted (10ml 0.1% acetic acid) and eluted with 10 ml of a solution of 25% acetonitrile and 0.1% acetic acid. Each sample was divided into 5 aliquots and lyophilized to dryness.

2.4.3 iTRAQ labeling of peptides and immunoprecipitation

Lyophilized peptides were subjected to labeling with the iTRAQ 4-plex reagent (Applied Biosystems). Each aliquot of peptides was dissolved in 30 μl of 0.5 M triethylammonium bicarbonate, pH 8.5 and reacted with two tubes of iTRAQ reagent (dissolved in 70 μl of ethanol each). The reagents for each of the conditions used were, iTRAQ-114 (U87-DK), iTRAQ-115 (U87-M), iTRAQ-116 (U87-H) and iTRAQ-117 (U87-SH). The mixture was incubated at room temperature for 50 min and then concentrated to 30 μl. The four different isotopically labeled samples were combined and acidified with 360 μl of 0.1% acetic acid and then reduced to dryness.

The combined sample was reconstituted with 150 μl of IP buffer (100 mM Tris, 100 mM NaCl, 1% NP-40, pH 7.4), 300 μl of water and the pH was adjusted to 7.4.

After immunoprecipitation with pTyr100 (Cell Signaling Technology, Beverly, MA) for the phosphotyrosine-containing peptides, which were used in a prior study [41], the supernatant was incubated with 10 μg of protein G Plus-agarose beads (Calbiochem) and 12 μg of MPM-2 antibody (Upstate) for 8 hrs at 4°C. Phosphopeptides were washed and eluted as previously described [41].

2.4.4 Immobilized metal affinity chromatography (IMAC) and Mass Spectrometry

Immobilized metal affinity chromatography (IMAC) was performed to enrich for phosphorylated peptides and remove non-specifically retained non-phosphorylated peptides. Eluted peptides were loaded onto a 10 cm self-packed IMAC (20MC, Applied Biosystems) capillary column (200 m ID, 360 μm OD), and rinsed with organic rinse solution (25% MeCN, 1% HOAc, 100 mM NaCl) for 10 min at 10μl/min. The column was then equilibrated with 0.1% HOAc for 10 min at 10 μl/min and then eluted onto a 10 cm self-packed C18 (YMC-Waters 10 5 μm) precolumn (100 μm ID, 360 μm OD) with 50μl of 250mM Na2HPO4, pH 8.0. After a 10 min rinse with 0.1% HOAc, the precolumn was connected to a 10 cm self-packed C18 (YMC-Waters 5 μm ODS-AQ) analytical capillary column (50 μm ID, 360 μm OD) with an integrated electrospray tip (1 μm orifice). Peptides were eluted with a 125 minute gradient with solvents A (1% HOAC) and B (70% MeCN in 1% OHAc): 10 min from 0% to 13%, 95 min from 13% to 42%, 10 min from 42% to 60% and 10 min from 60% to 100%. Eluted peptides were directly electrosprayed into a QqTof mass spectrometer (QSTAR XL Pro, Applied Biosystems). MS/MS spectra of the five most intense peaks with 2 - 5 charge states in the full MS scan were automatically acquired in information-dependent acquisition mode with previously selected peaks excluded for 40 sec.

2.4.5 Phosphopeptide sequencing, quantification and clustering

MS/MS spectra were extracted and searched using MASCOT (Matrix Science). For MASCOT, data was searched against the human non-redundant protein database with trypsin specificity, 2 missed cleavages, precursor mass tolerance of 2.2 amu for the precursor ion and 0.15 for the fragment ion tolerance. Phosphorylation sites and peptide sequence assignments were validated and quantified by manual confirmation of raw MS/MS data (raw MS/MS data available at http://web.mit.edu/fwhitelab/data/index.html). Peak areas of iTRAQ marker ions (m/z 114, 115, 116 and 117) were obtained and corrected according to manufacturers instructions to account for isotopic overlap. The quantified data was then normalized with values from the iTRAQ marker ion peak areas of non-phosphorylated peptides in the supernatant of the immunoprecipitation (used as a loading control to account for possible variation in the starting amount of sample for each condition). Each condition was normalized against the U87-H cell line to obtain fold changes across all 4 conditions.

2.4.6 Phosphopeptide library array

Methods used here are similar to Elia et al. [26]. An ABIMED peptide arrayer was used to synthesize degenerate libraries on an amino-PEG cellulose membrane. The libraries consisted of four degenerate positions each on the N- and C-terminal sides of the central phosphoserine or serine positions, i.e. X-X-X-X-pS/S-Z-X-X-X, where X represents all naturally occurring amino acids except cysteine and Z additionally excludes proline. On most spots, one of the degenerate positions was fixed as a specific amino acid in addition to the orienting phosphoserine or serine. The cellulose membrane was blocked for 1.5 hr at room temperature in 3% milk and 1% TBS-T at pH 7.4. It was then incubated with the primary antibody, MPM-2 (Upstate), at room temperature for 1.5 hr at 0.5μg/mL in 1% TBS-T. The membrane was washed, blocked, and then probed with a secondary anti-mouse HRP-conjugated antibody (GE Healthcare) overnight at 4°C. MPM-2 library binding was detected using enhanced

chemiluminescence, imaged and quantified using a Kodak Image Station.

2.4.7 Determination of MPM-2 selectivity

The quantified phosphoserine-oriented peptide library was normalized to the serine-oriented control library. Each fixed position was then normalized to the average of the completely degenerate column (column 1) to determine total selectivity of that fixed position over pS alone. Quantification and normalization are provided in Table A.3.

2.4.8 Preparation of data for motif enrichment analysis

Of the 68 phosphopeptides identified in this study, 44 are singly phosphorylated, 21 are doubly phosphorylated, and 3 are phosphorylated on three amino acids. We therefore expanded the dataset to include the full complement of individual phosphosites, each centered on a single phosphorylation. For sites quantified more than once in the context of different phosphopeptides with different partner residues simultaneously phosphorylated, we included each instance of the site, for a total of 95. In three instances, the exact residue of phosphorylation could not be determined, and a choice between two possible sites was made arbitrarily. Analysis was repeated with all 8 possible selections of the identities of these three sites, with no significant qualitative effect on the results (data not shown). We expanded each site to include the 7 amino acids N-terminal and 7 amino acids C-terminal of the phosphorylated residue using the Entrez Protein database. For each of the three cell lines expressing medium, high, or super-high levels of EGFRvIII, we identified the top quartile of phosphosites upregulated relative to kinase-dead U87-DK control cells. For U87-M, U87-H, and U87-SH cells, this corresponded to 1.67-fold, 2.10-fold, and 2.20-fold enrichment, respectively, and 25 sites. The 24th- and 25th- most upregulated sites in each cell line came from the same phosphopeptide and had identical quantification.

2.4.9 Enriched motif search

The space of possible motifs in a 15-mer peptide containing a central fixed phosphoresidue is enormous over $2.1 \text{x} 10^{19}$, if a few combinations of chemically similar amino acids are allowed for. A strategy must therefore be used for restricting the search to those motifs most likely to be significantly enriched in the regulated data of interest. For each of phosphoserine, phosphothreonine, and the combination of the two, the significance of enrichment in the regulated data relative to the full background was calculated (see section 2.4.10) for every motif that can be created by fixing any one or two of the seven positions on either side of the phosphorylated residue as any of the twenty amino acids, as the combination of the basic amino acids arginine and lysine, as the combination of the acidic amino acids glutamate and aspartate, or as the combination of the hydrophobic amino acids leucine, isoleucine, valine and methionine. For each motif identified with an enrichment significance of 0.01 or less that appeared 3 or more times in the EGFRvIII-regulated foreground data, the significance of all motifs that can be created by fixing the identities of any further one or two amino acids was calculated. This procedure was recursed until no further significant motifs were found.

2.4.10 Motif significance calculation

For each motif of potential interest, the statistical significance of enrichment of that motif in the EGFRvIII-regulated foreground subset of the total background data was calculated by summing the distribution of the hypergeometric distribution from the number of appearances, k, to the number of possible appearances:

$$p(k') = \sum_{k'=k}^{min(n,K)} \frac{\binom{K}{k'}\binom{N-K}{n-k'}}{\binom{N}{n}} \tag{2.1}$$

where N is the number of phosphosites in the full dataset, n is the number of sites in the EGFRvIII-regulated subset, m is the number of motif sites in the full data, and k is the number of motif sites in the regulated data subset. This corresponds exactly

to the probability of seeing as many instances or more of the motif as are seen in the EGFRvIII-regulated dataset by chance if drawing a dataset the same size as the regulated data randomly from the full dataset.

2.4.11 Empirical analysis of false positive rate

To characterize the rate of false positive motif discovery, 1000 foreground sets of 25 phosphosites (corresponding in size to the top-quartile datasets studied) were randomly generated, and motif enrichment analysis was performed on each. The number of motifs found and the significance of the most significantly enriched motif for each were tabulated and compared to the same statistics for the foreground datasets of interest.

2.5 Conclusions

Discovery of an antibody epitope through MS measurement revealed biologically relevant substrates and non-canonical selectivity for a well-studied biological probe. The previously-established belief that MPM-2 binds only to "pS/pT-P" sites was most likely derived from inadequate substrate-detection sensitivity, with earlier methods capable of only picking up the most abundant of MPM-2s substrates: "pS-P" sites. With contemporary state-of-art peptide IP and MS peptide identification, we have now been able to complement the traditional *in vitro* affinity assay with *in vivo* biological substrates for MPM-2, thereby providing positional selectivity for cellular substrates. For example, of the singly phosphorylated peptides, 98% conform to selectivity for P/E/D in the +1 position. On average, each of these phosphopeptides conformed to three positions of selectivity defined in Figure 2-2B, and all conformed to at least one. The range of conformity among the sequence positions varied greatly, with the -1 positional selectivity ranking second to the +1 position, whereas the -2 position contributed to only about 30% of the peptides identified. Of the peptides that conform to the -1 positional selectivity of MPM-2 for isoleucine, leucine, phenylalanine, and valine, 90% of the MS-identified singly-phosphorylated substrates

contain either a leucine or a phenylalanine, while no peptides contain valine, although there are known "V-pS/pT" and "V-pS/pT-P" sites listed in the Phospho.ELM [22] database. This observation demonstrates the difference between favorable *in vitro* degenerate library interactions and actual biological substrates. The most likely explanation for this difference is that proteins *in vivo* must be optimized for both kinase-dependent phosphorylation and for MPM-2 binding, while *in vitro* MPM-2 binding is not restricted by kinase-dependent phosphorylation.

The set of peptides identified by an MS experiment arises from a convolution of many factors. In order to appear in an IP/MS experiment, a peptide must be present in the cell, have sufficient affinity for the antibody, be compatible with any further purification, and be efficiently measured and sequenced via MS. This convolution of effects makes it difficult to perform motif-enrichment analysis of MS data. To address these issues, we have developed a tool to identify amino acid sequence motifs enriched in a regulated subset of a larger dataset. By comparing a biologically interesting subset of an MS dataset to the entire MS dataset, many of the confounding elements are cancelled. Most existing motif identification tools, such TEIRESIAS [85], EMOTIF [44], and PRATT [74], solve a fundamentally different problem: motif identification without regard to a biologically relevant background. In contrast, our method compares a list of phosphopeptides of interest to a proteomic background, and is therefore similar to the MOTIF-X method of Schwartz and Gygi [97]. Although similar, these methods differ in several respects important to the goals of this study. First, and most importantly, our method uses the full cohort of peptides identified in the MS study as a comparative background. This dataset provides the most specific set of peptides possible, thereby effectively minimizing all influences except the biological regulation of interest. Second, our method allows each peptide to be associated with any number of enriched motifs by searching over individual and pairs of amino acid positions, expanding the space of motifs that may be found at a manageable computational expense. Third, statistical significance of enrichment is calculated by using the exact hypergeometric probability function, which is more appropriate to the relatively small background data sets of

current interest than the binomial approximation. Finally, an empirical analysis of randomly selected data provides some indication that foregrounds generated from up-regulated phosphosites within this dataset have meaningful arrays of enriched amino acid sequence motifs.

A very recent study by Dephoure et al. provides a great number of novel phosphosites upregulated in mitotic cells relative to an asynchronous control [20]. MOTIF-X [97] was used to compare phosphoserine sites upregulated in mitotic cells to all serines in the human proteome. This comparison revealed four motifs that the authors associated with CDKs, two that the authors associated with the kinases Aurora A and Plk1, and two that the authors felt indicated uncharacterized mitotic kinase function. It is worth noting that the motifs we have discovered here do not correspond well to any of the motifs described by Dephoure et al. This disparity in findings is likely due to the targeted nature of our study: we have examined only phosphopeptides captured by the antibody MPM-2, and focused on those motifs that correspond explicitly to increased levels of EGFRvIII expression.

Although the accepted "pS/pT-P" epitope of MPM-2 is most often generated by the cell-cycle dependent kinases [102] and the MAP kinases [35], a number of other kinases including Plk1, NIMA, and MEK are known to generate MPM-2 epitopes in at least one substrate [54], but not necessarily in a manner that matches the motifs found in this study. The motifs we identify here as being associated with EGFRvIII upregulation in U87 glioblastoma cells are most reminiscent of the specificity of casein kinase II, with acidic amino acids prevalent in the -2, +1, +3 and +5 positions, corresponding well to the substrate specificity of CK2 for serine and threonine residues with nearby negative amino acids, with the +1, +2, and particularly the +3 residues most often acidic [65, 103]. The presence of acidic residues in the +1 position, in particular, seems to indicate that the known acidophilic mitotic kinase, Plk1, is not wholly responsible, as Plk1 substrates most often have an aliphatic residue in the +1 position [71]. In fact, one of the phosphorylated sites identified in this study, serine 366 on the RNA-binding protein La, is a known CK2 substrate [98]. We attempted to measure CK2 activity in H and DK cellular lysates, through the incorporation of

radio-labeled phosphate of an ideal CK2 peptide target sequence of RRDDDSDDD, see Appendix A.2. Unfortunately, CK2 inhibition by TBCA in these experiments indicated this sequence may not be specific enough for use in cellular lysate studies.

The idea of a role for CK2 in mitosis is not a new one. Though thought to be constitutively active, CK2 is known to be required in the G2/M portion of the S. cerevisiae cell cycle [38], as well as to promote the meiosis of Xenopus oocytes [70]. Specific roles have been found for CK2 in the activation of the cell-cycle regulatory phosphatase CDC25B [111], and at a mitosis-specific site on DNA topoisomerase IIα; in fact, a pair of studies has identified casein kinase II as being capable of generating an MPM-2 epitope on DNA topoisomerase IIα [27, 28]. The amino acid sequence surrounding the phosphorylated residue, "NRRKRKPpSTSDDSDS" contains the common +3 acid aspect of motifs found enriched in MPM-2-captured, EGFRvIII-regulated sites in our data, although it does not match directly to any of the motifs we have identified. CK2 activity has also been found to be upregulated in a large number of cancers (reviewed in [2] and [92]). While the similarity of the motifs we have identified to the known substrate specificity of CK2 is striking, it is similarly possible that there is another acidophilic kinase with an undescribed motif specificity and uncharacterized role in mitosis. In fact, the artificial neural network analysis available through the recently published NetPhorest algorithm provides the following consensus CK2 motif: *E-E-E-(E/D/S)-(E/D)*-**S/T**- (D/E/S/G)-(D/E/S)-**(E/D)**-*(E/D)-(E/D)-(E/D)-E*, with weak selectivity in italics, moderate selectivity in normal text, and strong selectivity in bold [66]. Although this motif does not exactly match to any of the motifs found in this study, the panel of motifs identified here is a combination of kinase phosphorylation specificity, antibody binding specificity, and regulation downstream of EGFRvIII. The motifs described here, therefore, likely represent only a subset of the kinase recognition motif. Regardless of the kinase responsible, we have identified a set of motifs that are upregulated in a manner coordinated with the expression level of EGFRvIII in glioblastoma cell lines. EGFRvIII expression itself correlates with poor prognosis in GBM patients [31], therefore understanding the generation and regulation of these motifs may lead to an improved

understanding of glioblastoma and to improvements in its treatment.

Similar analyses can be applied to other diseases and disorders. We hypothesize that identified motifs will be a signature for kinases, phosphopeptide-binding domains, and perhaps phosphatases, associated with the particular regulation being analyzed. Properly interpreted, these motifs may provide mechanistic insight into the origin and phenotype of the samples studied. This information may aid in determining systematic signaling differences among existing cell lines, thereby enabling mechanistic hypotheses with respect to treatment. Ultimately, it should be possible to apply these methods to characterize individual patient tumors, as differentiated from adjacent healthy tissue, with the goal of discovering hidden dysregulated signaling modules that could be effective therapeutic targets.

Chapter 3

PTMScout: A Web Resource For Analysis of High-Throughput Post-Translational Proteomic Studies

3.1 Summary

The rate of discovery of post-translational modification (PTM) sites is increasing rapidly, and is significantly outpacing our biological understanding of the function and regulation of those modifications. To help meet this challenge, we have created PTMScout, a web-based interface for viewing, manipulating, and analyzing high-throughput experimental measurements of PTMs in an effort to facilitate biological understanding of protein modifications in signaling networks. PTMScout is constructed around a custom database of PTM experiments and contains information from external protein and post-translational resources, including Gene Ontology annotations, Pfam domains, and Scansite predictions of kinase and phosphopeptide binding domain interactions. PTMScout functionality comprises dataset comparison tools, dataset summary views, and tools for protein assignments of peptides identified

by mass spectrometry (MS). Analysis tools in PTMScout focus on informed subset selection via common criteria, and on automated hypothesis generation through subset labeling derived from identification of statistically significant enrichment of other annotations in the experiment. Subset selection can be applied through PTMScout's flexible query interface, available for quantitative data measurements and data annotations, as well as an interface for importing dataset groupings by external means, such as unsupervised learning. We exemplify the various functions of PTMScout in application to datasets that contain relative quantitative measurements, as well as datasets lacking quantitative measurements, producing a set of interesting biological hypotheses. PTMScout is designed to be a widely accessible tool, enabling generation of multiple types of biological hypotheses from high-throughput PTM experiments and advancing functional assignment of novel PTM sites. PTMScout is available at http://ptmscout.mit.edu.

3.2 Introduction

Post-translational modifications (PTMs) regulate cellular signaling networks by modifying activity, localization, turnover and other characteristics of proteins in the cell. For example, signaling in receptor tyrosine kinase (RTK) networks, such as those downstream of epidermal growth factor receptor (EGFR) and insulin receptor, is initiated by binding of cytokines or growth factors, and is generally propagated by phosphorylation of signaling molecules. Additionally, receptor surface expression can be regulated by ubiquitination while gene expression can be regulated by acetylation of transcription factors and histones. With the increasing utilization of high-throughput mass spectrometry (MS) technologies and the ability to enrich for a particular modification from a biological sample, hundreds or even thousands of PTM sites can now be identified in a single experiment and relatively quantified across biological conditions [132]. This increase in the number of PTM sites identified in each analysis has led to a rapid and accelerating expansion of known post-translational modifications, as evidenced by the number of the entries in a knowledgebase of phosphorylation over

the past five years: in 2004, when Phospho.ELM was first published [23] it contained 1,703 known phosphorylation sites; in 2009 Phospho.ELM contained 19,649 known sites of phosphorylation, a more than ten-fold increase.

A number of database resources, including Phospho.ELM [22], PhosphoSite [39], PHOSIDA [33], and SysPTM [58], have emerged in response to the large production of phosphorylation data and are expanding to include other PTMs. For instance, PhosphoSite [39], originally established as a compendium of phosphorylation sites, has started to incorporate acetylation, methylation, glycosylation, ubiquitination, and other PTMs. Sites of modification can also be found in large protein compendia such as UniprotKB [114]. In addition to storing PTM data, SysPTM [58] brings bioinformatics resources to bear on PTMs in the context of the PTM compendium, such as mapping known PTMs onto signaling pathways [58]. Other bioinformatic tools specific to phosphorylation exist as well: for example, Phospho.ELM, PHOSIDA [33], Scansite [76], KinasePhos [124], and PPSP [126] contain predictions or annotations of the kinase responsible for the phosphorylation of particular sites. Unfortunately, there is no wide-ranging resource currently that allows users to browse the data from diverse high-throughput PTM experiments (with the exception of PHOSIDA, which is specific to experiments done by the research group of Matthias Mann [33]).

Despite the daunting volume of PTM measurement by MS, the lack of computational methods for deriving experimental hypotheses from these datasets has become a bottleneck limiting the contribution of high-throughput PTM study to biological understanding. To address this limitation, a few studies have implemented unsupervised learning techniques as a method of reducing data dimensionality to elucidate dynamic and functional patterns in phosphoproteomic measurements [78,132]. These methods were successful in highlighting functionality of novel protein modifications, but the vast majority of uncharacterized PTMs remain without putative biological function, even after unsupervised learning.

We have developed PTMScout in an effort to bring hypothesis generation tools into proteomic studies of PTMs, while leaving each peptide in the larger context of the phosphoproteome, acetylome, etc. PTMScout provides an interface to a novel

database of post-translational protein modifications, which incorporates functional annotations and mRNA expression profiles. Individual experiments and their quantitative measurements are treated as unique entities, uploaded by their producers for availability to the broader scientific community. These data may then be analyzed through PTMScout via subset selection by functional or dynamic annotation along one axis followed by identification of statistically significantly enriched annotations along all other axes. Additionally, by utilizing expression profiles and multiple existing datasets, PTMScout provides information regarding the assignment of individual peptides to particular proteins, in cases where there are a number of proteins or protein isoforms from which a peptide may have been cleaved.

The main functions of PTMScout are six-fold, see Figure 3-1 for a depiction. First, PTMScout allows browsing of experimental data alongside publically available annotations, such as Gene Ontology (GO) terms and protein domain structures. Second, PTMScout allows for direct comparison of a particular experiment to one or more other experiments, including highlighting novel sites of modification. Third, PTMScout highlights potentially contentious assignments of peptides to proteins and gives biologists tools, such as tissue and cellular mRNA expression data, to determine an optimum protein assignment. Fourth, PTMScout allows for dataset reduction by subset selection, either on the data itself or on its imported annotations. Fifth, PTMScout provides automated statistical significance testing of a number of metrics orthogonal to the selection criteria (including quantitation in dynamic data and external annotations) in selected subsets. Finally, PTMScout provides an interface to unsupervised learning and automatic partition labeling based on statistically significantly enriched information.

Example use cases of PTMScout for generating biological insights from a number of previously published datasets will be shown using the data described in Table 3.1, which includes a mix of datasets with quantitatively measured conditions in cells stimulated with EGF ligand (datasets *EGF4* [132], *EGF7* [122], and *HER2* [123]), as well as a discovery dataset from acetylated intracellular proteins lacking quantitative information (*AcK* [14]). Using the readily available tools within PTMScout, Figure

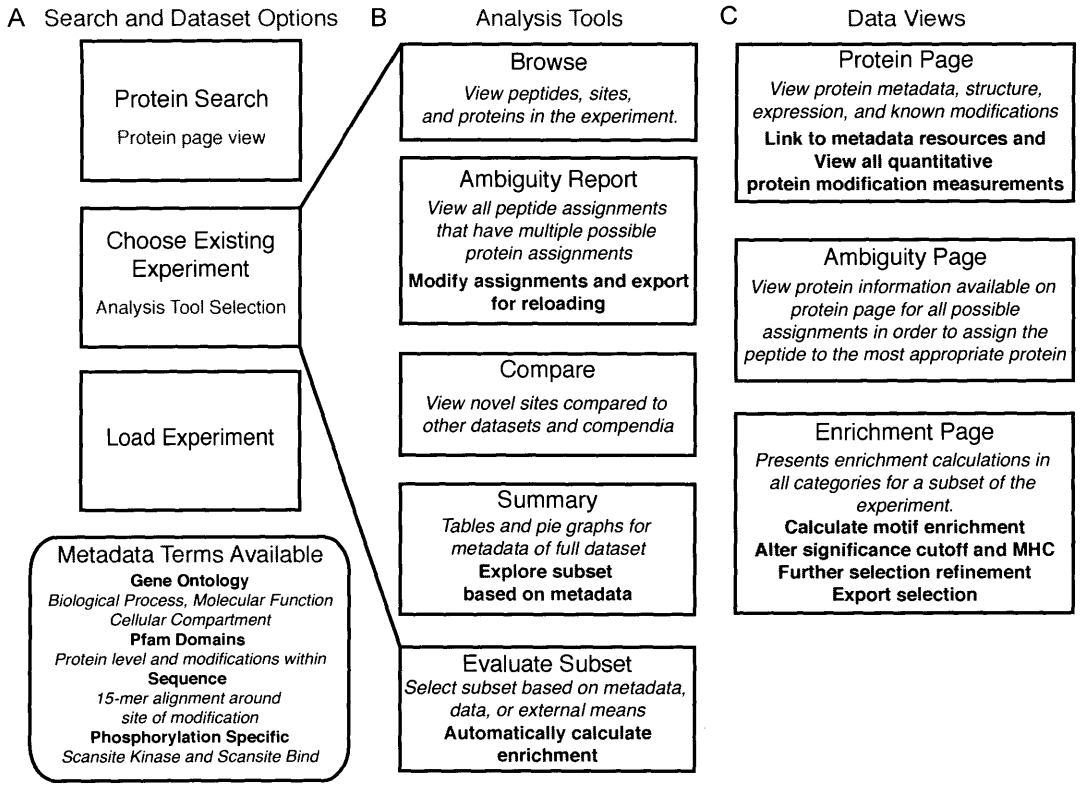


Figure 3-1: A depiction of the major features, analysis tools, and page view types available in PTMScout. (A) One can choose to search by protein, load a new dataset, or analyze an existing dataset. The metadata incorporated in PTMScout to annotate the biological molecules of interest is described. (B) Analysis tools for experimental datasets include: experiment browsing, full dataset ambiguity reports and assignment functions, dataset comparisons and novelty analysis, dataset summary features, which can be linked directly to subset evaluation, and subset selection and evaluation by metadata or data queries. (C) Most analysis and search tools rely on three fundamental data views: protein pages, ambiguity pages, and enrichment pages of subset selection features.

3-1, we were able to construct multiple biological hypotheses regarding the potential functional characterization of multiple PTM sites, including a role for phosphorylation of Y497 on FRK in EGFR proliferation by subset selection in the *EGF4* [132] dataset. We were also able to find potential signal integrators between focal adhesions and the EGFR pathway by using a mix of subset selection and enrichment based on dynamics, as well as metadata selection and the ability to view data on a protein across datasets. Moreover, peptide assignment ambiguity tools were used to indicate a preferable

protein assignment for the Src family kinase activation loop phosphorylation event. An interface to arbritrary dataset clustering was used to recapitulate unsuprevised learning results from the *EGF4* [132] dataset and subset selection by quantitative data was used to expand the endocytic signaling module in this dataset. Using dataset comparison tools the degree to which the *AcK* [14] dataset expands our current knowledge of acetylation is quantified. By using subset selection based on protein sequence and previously described acetyl transferase sequence recognition we demonstrate that CBP/p300 acetylates both non-histone and histone proteins, and in particular, it targets RNA binding proteins. Finally, the dataset summary view demonstrates there may be acetylation sites missed by using trypsin as the proteolytic enzyme prior to MS measurement.

3.3 Results

PTMScout is a web application that provides access and a computational interface to an underlying MySQL database. The PTMScout database contains data from high-throughput studies of protein modifications and existing PTM compendia, Figure 3-2A. Phosphorylation and acetylation experiments are currently included, but PTMScout has been designed to incorporate additional modification types as they become available experimentally. The database incorporates information at the protein level, such as Gene Ontology (GO) terms [5] and Pfam domain structures [29], Figure 3-2C, as well as information at the level of individual sites of modification, such as Scansite [76] predictions of the enzymes responsible for, or proteins interacting with, modification sites, short peptide sequences aligned around the site of modification, and the predicted Pfam domain in which a site falls, if any, Figure 3-2B. The information contained in the database, and the level of its specificity, is depicted in Figure 3-2 using an example peptide measured on JAK2 after stimulation by EGF [122]. In addition to high-throughput datasets of measured modifications, PTMScout incorporates larger, curated datasets of known post-translational modifications [22, 39, 114] for easy comparison of a new experiment to the current state of knowledge for a

Reference Name	Dataset Name	Cell Type	Stimulation	Measurements	PTM	Dataset Size (peptides)
EGF7 [122]	Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks	HMEC	EGF	0,1,2,4,8,16,32 minutes	pY	222
EGF4 [132]	Quantitative proteomic analysis of phosphotyrosine-mediated cellular signaling networks: SUPPLEMENTAL TABLE #1	HMEC	EGF	0,5,10,30 minutes	pY	77
HER2 [123]	Effects of HER2 overexpression on cell signaling networks governing proliferation and migration	HMEC 24H	EGF, HRG	0, 5, 10, 30 minutes EGF, HRG stimulation Parental, 24H cell lines	pY	68
AcK [14]	Lysine acetylation targets protein complexes and co-regulates major cellular functions	MV-411	X	X	AcK	3,286

Table 3.1: Datasets used to demonstrate the functionality of PTMScout. Reference Name is used for quick reference to the dataset of interest.

particular modification or modifications on a particular protein. PTMScout version 1.2, at the time of this writing, included 16 unique datasets, 11 experimental and 5 compendia, totaling 224,072 modifications across 72 species (133,440 phosphoserine, 38,906 phosphothreonine, 34,149 phosphotyrosine, and 17,577 acetyllysine).

The scale of high-throughput proteomic PTM data is approaching that of genomic data, leading to similar problems, as it is difficult to derive biological meaning from large datasets without a specific prior hypothesis. To address this challenge, PTMScout tools allow users to partition their data into a more comprehensible format using subset selection in one of four ways. First, users can select a subset of data that is annotated with a particular label from an imported data source (such as a GO

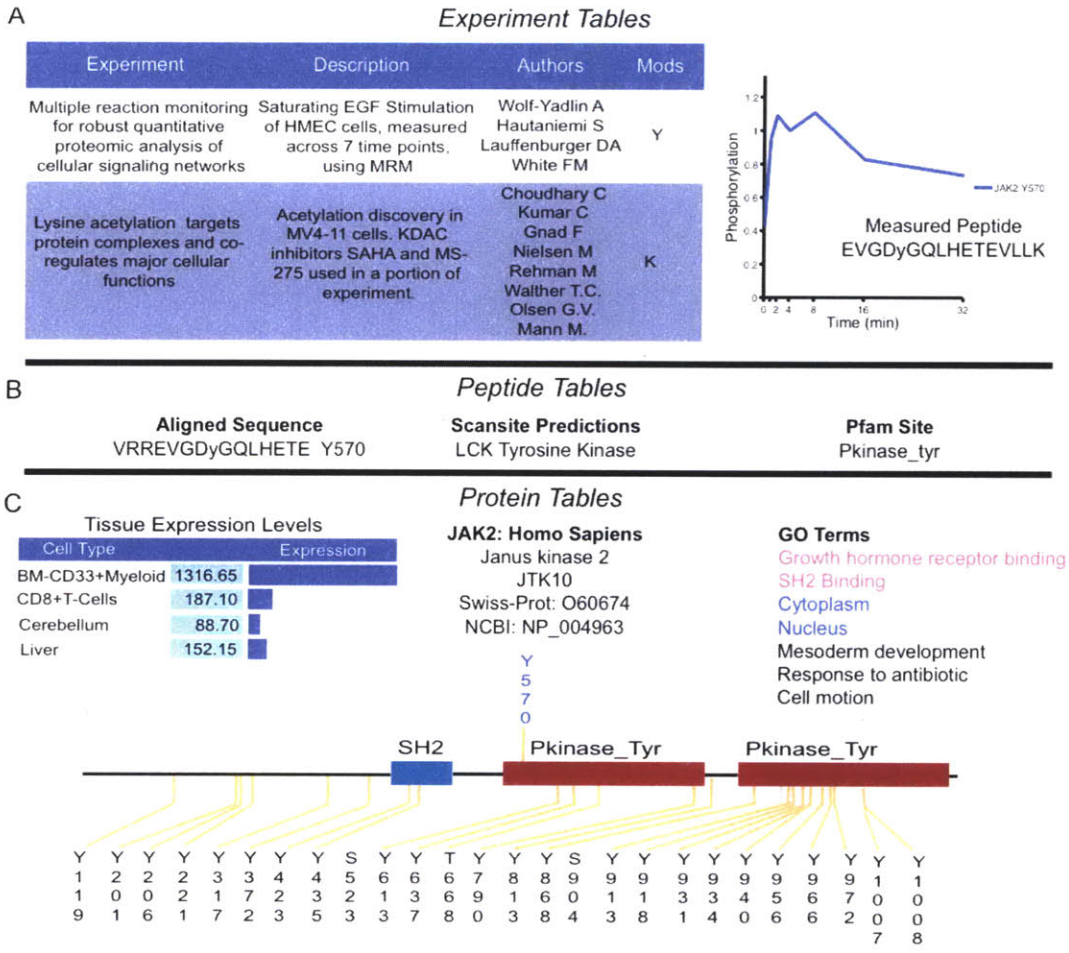


Figure 3-2: An abstracted database schema of PTMScout with an example peptide measured from JAK2 in the *EGF7* experiment [122]. The database consists of three major classes of information. (A) Experimental data are the basic data elements and consists of all measured modified peptides in an experiment and their associated quantification, when applicable. (B) PTMScout site-level information includes the 15-mer sequence of peptides centered on individual sites of modification, the domain within which the site falls in its protein (Pfam_site), and any predictions on the function or regulation of that individual modification. (C) Protein-level information includes Gene Ontology (GO) annotations, Pfam domain structures, GNF SymAtlas mRNA expression information, and a variety of database accessions, gene names and protein names. Peptide and experiment data in the database are connected through protein identifiers, allowing for direct comparison across datasets and compendia.

term). Second, users can select a subset of data based on quantitative characteristics of the experimentally measured data. Third, users can partition the experiment by an external means, such as unsupervised learning and import the partitioning scheme

to PTMScout. Finally, the user can combine any or all of the first three methods to create a subset of data. Once a data subset is selected by one of these means, the statistical significance of enrichment with respect to the full dataset is automatically calculated according to a hypergeometric distribution for all other metadata annotations, as well as some qualitative characteristics of data dynamics. The fundamental philosophy of this method is that subset selection, partitioning, or clustering in one feature dimension could produce a biological hypothesis, highlighted by a statistically significant enrichment of a term or feature in another dimension.

Features for subset selection and enrichment include GO terms, Pfam domains (at a protein level as well as site level), kinase and binding domain predictions from Scansite, local sequence features and measured quantitative features. Many features are categorical, and selection and enrichment significance calculations are straightforward, as detailed in Experimental Procedures. Selection and enrichment of quantitative data and sequence features require special mention due to their increased complexity. We allow for local sequence feature selection by regular expression queries. For example, to search for a common SH2 binding motif containing a phosphotyrosine with a hydrophobic amino acid in the +3 position, one can choose to search for 'y..[PLIVM] (phosphotyrosine followed by any two amino acids, followed by a proline, leucine, isoleucine, valine, or methionine). In the reverse problem, searching for meaningful linear amino acid sequences in a subset, e.g. to establish an enriched motif, requires an algorithm to reduce the search space to a feasible size. For this purpose, PTMScout uses a previously developed greedy search motif algorithm [46]. Additionally, PTMScout implements flexible search queries for quantitative data fields by allowing the user to create simple mathematical expressions as selection criteria. For example, one can search for an early response subset of tyrosine phosphorylation events in the EGFR signaling network (e.g. the *EGF7* experiment [122], see Table 3.1) by requiring a four-fold change in the first minute with the query "time(1min) $\div$ time(0min) ≥ 4". PTMScout searches for quantitative data enrichment in a subset by testing for specific qualitative descriptors of quantitative dynamic features: fold change, maximum modification, interval of peak upregulation, and interval of peak

downregulation amongst each of the quantitative data points. By rigorously defining a "dynamics feature space", we are able to calculate the statistical significance of enrichment of these labels in the same way we might calculate the significance of the representation of a GO term or kinase prediction in a subset. Finally, all query types can be combined, enabling the identification of, for example, a subset of sites that adhere to canonical SH2 binding motifs and are upregulated within the first time point.

3.3.1 Activating kinase events in the EGFR pathway

While *ab initio* prediction of the function of specific phosphorylation sites is difficult, typically phosphorylation within the activation loop of the catalytic domain of protein kinases can be expected to enhance activity of the kinase by driving a structural transition [75]. In order to predict whether a PTM falls within a kinase activation loop, PTMScout searches for the conserved flanking amino acid sequences 'DFG' on the N-terminal side and 'APE', on the C-terminal side of the site of modification [75]. This definition of the activation loop conservation was expanded by aligning the kinase catalytic domains of kinases within PTMScout using ClustalW2 [56] (see Experimental Procedures). Of the 2,089 tyrosine and serine/threonine kinase domains in PTMScout version 1.2, ~79% of the activation loops are confidently predicted, and with some certainty, another ~3.6% are predicted. Activation loops cannot be predicted for the remaining kinases based on searching for conserved flanking sequences.

We used PTMScout to explore the subset of kinase catalytic domain modifications in the *EGF4* experiment [132] by performing a metadata query requesting PTMs that fall into the Pfam domains 'Pkinase' (serine/threonine kinase domains) or 'Pkinase_tyr' (tyrosine kinase domains). Figure 3-3 illustrates the query results, which include modifications within the catalytic domains of the kinases CDC2, EPHA1, EPHA2, GSK3A, INSR/IGF1R, MAPK1, MAPK3, PRPF4B and PTK2. PTMScout predicts that all of these except the site on CDC2 fall into the activation loop of their respective kinase, and are therefore, potentially kinase activating events. Ad-

ditionally, phosphorylation levels for each of these sites increased, albeit slightly in some cases, upon stimulation of EGF (see Figure 3-3B). This subset is significantly enriched, based on a FDR corrected value of 0.01, for proteins that contain 'Pkinase' domains relative to the full *EGF4* dataset (see Figure 3-3C). It is interesting to note that while all modifications to serine/threonine kinases in the dataset occur within the catalytic domain and are represented in this subset, less than half of measured modifications to tyrosine kinases occur within the catalytic domain.

We examined the remaining phosphorylation events on tyrosine kinases by choosing a subset based on proteins with "Domains=Pkinase_tyr". In addition to the four phosphorylation sites that occur within tyrosine kinase catalytic domains, another sixteen sites are found on ten tyrosine kinases, including Y497 on Fyn-related kinase (FRK), a Src family kinase. This site falls on the c-terminal tail of the protein, which, based on relative proximity between the c-terminus of the protein and the kinase domain, is similar in location to the negative regulatory site (Y527) of Src [6]. Alignment of the eleven Src family kinases indicates all but one member, SRM, contain a tyrosine in this region of the protein, and there is evidence for phosphorylation on all of these sites, see Table 3.2. By extension, it is reasonable to predict that Y497 on FRK may bind FRK's SH2 binding domain, thereby inhibiting kinase activity of FRK. Amongst the ten Src family kinases with known phosphorylation sites on a tyrosine in the c-terminal tail of the protein, the sequence surrounding Y497 on FRK is the most dissimilar, potentially indicating that FRK Y497 may be phosphorylated by a kinase different from the phosphorylating kinase(s) for the analogous sites on other family members. Phosphorylation of FRK Y497 increases two-fold by 30 minutes after stimulation of EGF, which indicates that EGF stimulation may cause this particular Src family kinase to decrease in catalytic activity. Gene Ontology annotations for FRK indicate nuclear localization and involvement in the negative regulation of cell cycle progression. If suppression of cell cycle progression is dependent on its kinase activity, then this phosphorylation site may be one specific mechanism by which EGF stimulation enhances cell growth and proliferation.

A

Protein	Site	Activation Loop
CDC2	Y15	No
EPHA1	Y781	Yes
EPHA2	Y772	Yes
GSK3A	Y279	Yes
INSR	Y1189	Yes
MAPK1	Y187	Yes
MAPK3	Y204	Yes
PRPF4B	Y849	Yes
PTK2	Y576	Yes

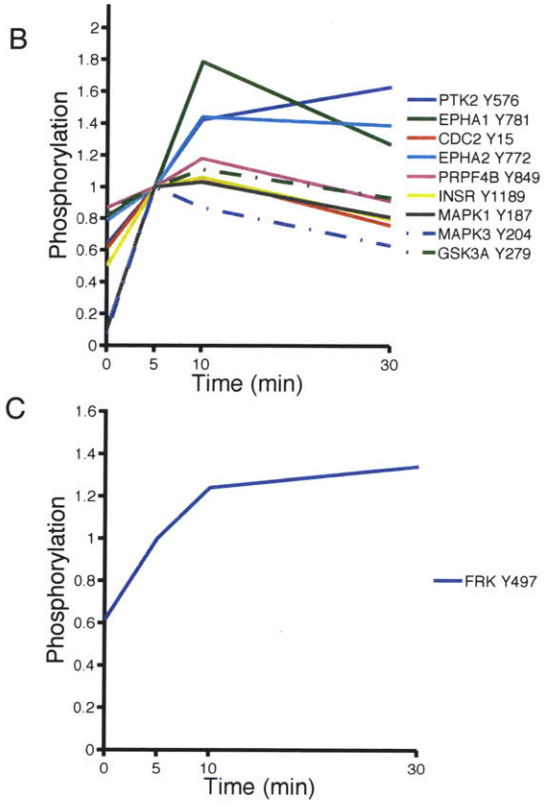


Figure 3-3: Subset selection of the kinase catalytic domain phosphorylation sites in the *EGF4* dataset [132]. (A) Nine sites were found to occur within kinase catalytic domains. With the exception of CDC2, all sites are predicted to fall within the activation loop of their respective kinases. (B) Quantitative measurement graph of the nine phosphorylation sites indicates they are all responsive to EGF stimulation to some extent. (C) Quantitative dynamic measurements of another tyrosine kinase phosphorylation site on a Src family kinase, FRK, which falls outside of the kinase catalytic domain, on the c-terminal portion of the protein. Due to its similarity to the negative regulation site of Src, this may indicate that negative regulation of FRK increases after stimulation by EGF.

Table 3.2: The c-terminal tyrosine that is phosphorylated in each of the human Src family kinases and their relative positions corresponding to the given Swissprot accession number.

Name	Swissprot	Site	Peptide
YES	P07947	Y537	FTATEPQyQPGENL
SRC	P12931	Y530	FTSTEPQyQPGENL
FYN	P06241	Y531	FTATEPQyQPGENL
FGR	P09769	Y523	FTSAEPQyQPGDQT
HCK	P08631	Y521	YTATESQyQQQP
LYN	P07948	Y508	YTATEGQyQQQP
LCK	P06239	Y505	FTATEGQyQPQP
BLK	P51451	Y501	YTATERQyELQP
BRK	Q13882	Y447	RLSSFTSyENPT
FRK	P42685	Y497	YFETDSSySDANNFI
SRM	Q9H3Y6	NA	none

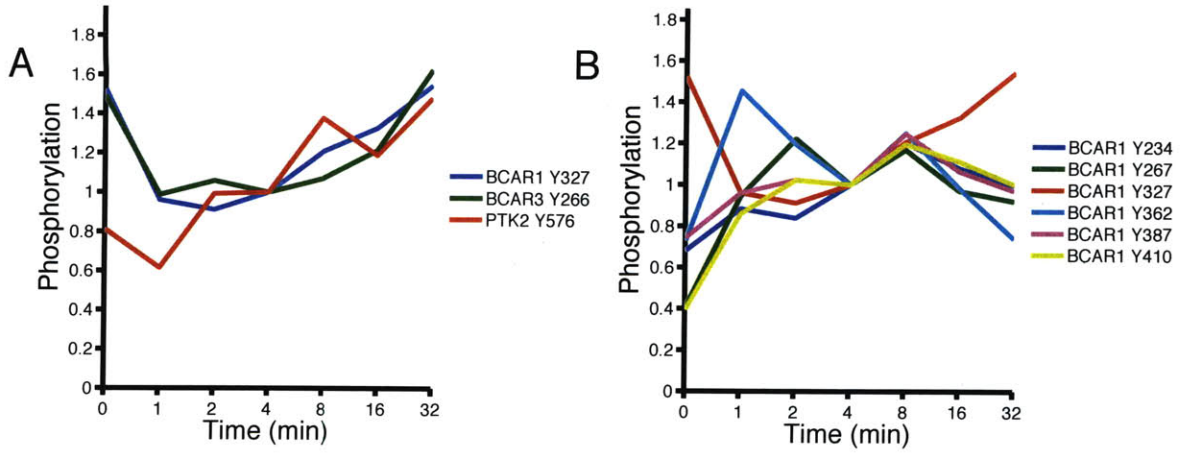
3.3.2 Focal adhesion signaling in response to EGF

PTMScouts flexible query interface allows users to apply intuitive rules for defining a subset of interest, based on the features inherent in any particular dataset. For example, using the data in the *EGF7* experiment [122], we selected a subset of phosphorylation sites that are immediately downregulated in response to EGF stimulation, a rare event. To generate this subset, we required that baseline phosphorylation be at least 30% higher than at one minute after stimulation: "time(0min) $\div$ time(1min) $\geq$ 1.3". This query resulted in the selection of only three phosphorylation sites on three proteins: BCAR1 Y327, BCAR3 Y266 and PTK2/FAK Y576. PTMScout shows that Y576 falls within the activation loop of the kinase domain of PTK2/FAK. Dynamics for these sites after EGF stimulation, shown in Figure 3-4A, indicate that these sites immediately decrease upon EGF introduction but recover within thirty-two minutes. The other functional annotations enriched in this cluster indicate a function in integrin-mediated signaling and localization at the focal adhesions (GO Biological Process term 'integrin-mediated signaling pathway' and GO Cellular Component term 'focal adhesion'). BCAR1, BCAR3 and FAK have been implicated as important signaling molecules with involvement in both EGFR and focal adhesion (FA) signal-

ing pathways [13, 95]. While the mechanism underlying decreased phosphorylation of these sites following EGFR activation is still unclear, our findings may indicate specific phosphorylation events involved in EGFR/FA crosstalk.

BCAR1 has five additional sites of phosphorylation which increase in response to EGF treatment in the *EGF7* experiment: Y234, Y267, Y362, Y387, and Y410. Phosphorylation events on BCAR1 were chosen as a subset by selecting "protein name=BCAR1", see Figure 3-4B for quantitative measurements of all six sites in response to EGF treatment. All of the phosphorylation sites measured on BCAR1 in this dataset have high sequence similarity, as shown by motif analysis: all have proline in the +3 position relative to the phosphotyrosine, and all but one site additionally contains an aspartic acid in the +1 position. CRK and NCK are known to bind to BCAR1 [36] and enrichment analysis indicates that the majority of CRK and NCK binding events that occur downstream of EGFR in this dataset occur on BCAR1, based on enrichment for CRK and NCK binding predictions by Scansite. Given the apparent redundancy of phosphospecific binding functionality of the six phosphorylation sites measured on BCAR1, it is interesting that one site has a completely opposing dynamic response to EGF stimulation. In integrin-mediated signaling complexes, BCAR1 is associated with at least three tyrosine kinases, Src, FAK and Abl [67]. Although specific kinase targets on BCAR1 are not clearly mapped, we see that the rare dynamic of BCAR1 Y327 correlates with a similar decrease in phosphorylation on the activating site of FAK, Y576 [67], possibly indicating a difference in enzymatic control of Y327 compared to the remaining five phosphorylation sites.

Physical aspects of EGF addition (e.g. shear stress from addition of solution containing EGF and swirling of media) could be responsible for mechanotransduction-related signaling events at the focal adhesions, versus a signaling response due to the growth factor itself. PTMScout has the ability to plot all quantitative measurements of modifications on a particular protein across all experiments contained in PTMScout. Across multiple experiments, Y327 is consistently downregulated in response to stimulation by EGF, see Figure 3-5; however, in the HER2 experiment [123], Y327 decreases in response to EGF, but increases in response to HRG. Addition of HRG



C

Motif	Subset*	Full Expt**	P-Value
.....y..PP...	4	4	1.23 x 10-7
.....yD.P....	**5**	**9**	**1.33 x 10-7**
.....yDOP....	4	6	1.82 x 10-6
.....-.yD.P....	4	7	4.22 x 10-6
.....y.OP....	5	16	4.49 x 10-6
.....-.yD.PP...	3	3	9.49 x 10-6
.....y.VPP..	3	3	9.49 x 10-6
.....yD.P...O	3	3	9.49 x 10-6
.....yDVP....	3	3	9.49 x 10-6
.....Vy..PP.A.	3	3	9.49 x 10-6
.....y..P....	**6**	**40**	**1.80 x 10-5**
.....y.VP....	4	10	2.48 x 10-5
.....OyD.P....	3	4	3.76 x 10-5

* out of 6
** out of 234

Figure 3-4: BCAR1 subset selection and enrichment in the *EGF7* dataset [122]. Dynamics are plotted along a log2 transformed axis for early time-point clarity. (A) Dynamics of the only three phosphorylation sites in the *EGF7* dataset that decrease in the first minute by 30% or more. These sites are enriched for 'integrin-mediated signaling pathway' annotation in GO Biological Process annotations. (B) Dynamics of all six phosphorylation sites on BCAR1, all of which increase in response to EGF, except Y327. (C) Motif enrichment of all six phosphorylation sites compared to the full dataset. Number of sites matching the motif in the subset and the full dataset are given along with the significance of that enrichment. All sites have a +3 proline and all but one have an aspartic acid in the +1 position.

should produce similar mechanical cues (see above) compared to the introduction of EGF, so these results indicate that the decrease in phosphorylation on these focal

adhesion signaling molecules is probably EGF-specific rather than a consequence of mechanical handling.

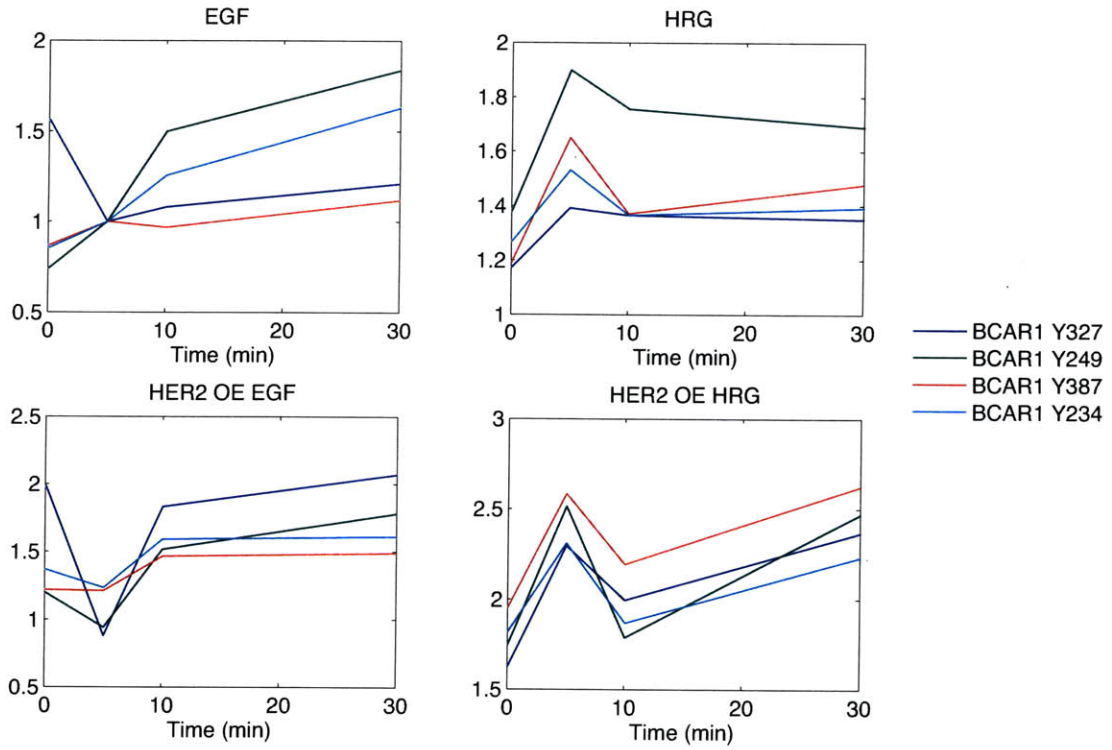


Figure 3-5: Quantitative measurements of phosphorylation sites on BCAR1 in the *HER2* [123] dataset. As in the *EGF7* [122] experiment, Y327 decreases in response to addition of EGF, however, it increases in response to addition of HRG indicating a ligand-specific, versus mechanotransduction, response.

3.3.3 Assignment of Src family kinase activation loop phosphorylation sites

Proteolytic peptide fragments can present an assignment problem as peptides can often match multiple proteins within a proteome. This ambiguity typically occurs when there are multiple isoforms or multiple gene products with a high degree of similarity surrounding sites of modification. For quantitative data, there is no clear way to deconvolute the degree to which each protein contributed to a particular peptide measurement without further intensive experimentation. To address the issue

of ambiguous peptide/protein assignments, PTMScout generates an automated report that allows users to immediately see all peptides within an experimental dataset that could have been assigned to multiple proteins. Additionally, while viewing any peptide assignment in the dataset throughout PTMScout, potentially ambiguous assignments are highlighted for the user, and information is presented that may help with selecting a particular protein among many choices. Specifically, for every protein that may have contributed to the peptide measurement, PTMScout illustrates: 1) the degree to which other datasets and compendia have annotated a protein, 2) the extent of GO annotations, 3) protein domain structure, and 4) tissue expression, available for Mouse and Human tissues, as well as NCI60 cell lines, incorporated from GNF SymAtlas [107]. New protein assignments can be made using a web form, exported as a new dataset, and then reloaded as a child experiment of the original. This process allows scientists to explore a dataset using the assignments they prefer, while faithfully maintaining the assignments chosen in the initial load of the dataset to PTMScout.

To demonstrate the usefulness of PTMScout's peptide assignment tools, we examined the possible protein assignments of the trypsinized fragment representing activation loop phosphorylation of several Src family kinases from the *EGF7* dataset [122]. Although the initial assignment of trypsinized peptide 'LIEDNEyTAR' was to the proto-oncogenic tyrosine kinase LCK, based on its sequence the measured peptide could belong to any of the proteins LCK, YES1, FYN, or SRC. All of these proteins are Src family kinases, but closer examination of their individual characteristics can help make a more informed protein choice. While FYN, SRC, and YES1 are expressed ubiquitously across all cell types, LCK is only highly expressed in leukocytes and T-cells according to data imported to PTMScout from SymAtlas [107]. Since the *EGF7* experiment was performed on human mammary epithelial cells (HMECs), which express an extremely low level of LCK mRNA, it is unlikely that the peptide measured resulted from the cleavage of LCK. Among FYN, SRC, and YES1, based on relatively similar mRNA expression in epithelial type cells, FYN is the protein with the most GO annotations. There are three possible isoforms of FYN (FYN-1, FYN-2,

FYN-3) that match the given sequence. The majority of experiments and compendia have preferentially chosen isoform B, the canonical sequence of FYN. In the absence of any external confirmatory experiments, the combination of all relevant information indicates that the most informative selection for the peptide 'LIEDNEyTAR' in the *EGF7* dataset is FYN isoform B.

3.3.4 Unsupervised learning highlights roles for proteins in endocytosis of EGFR

PTMScout can be used to explore the characteristics enriched in subsets created by unsupervised learning algorithms. The intent of unsupervised learning is to partition the members of a dataset into clusters based on similar quantitative measurements. This dataset reduction may then highlight interesting points of biology based on shared functionality in the cell. In addition, novel pathway components with unknown function can be hypothesized to share similar function or similar pathway effectors. An example of unsupervised learning applied to a quantitative PTM dataset appears in the experimental study of the *EGF4* experiment by Zhang et al [132]. Using a self-organizing map (SOM) [51], nine potential signaling modules were created based on similar dynamics. Two of these clusters were explored in depth: one cluster included EGFR Y1173 as well as several proteins known to bind directly to the receptor, while the second cluster included several proteins known to be involved in endocytosis. The 'endocytosis' cluster contained sites whose phosphorylation reached maximum levels relatively late, at ten minutes, and were strongly dephosphorylated at thirty minutes. Two members of this cluster had no known function in the EGFR network, proteins known at the time as Ymer and Chr20 ORF18 , also known as CCDC50 and RBCK1, respectively. Based on their grouping with phosphorylation sites on endocytic proteins STAM1, STAM2, EPS15, ACK1 and ANXA2 the authors proposed a role for Ymer/CCDC50 and Chr20 ORF18/RBCK1 in endocytosis and trafficking of EGFR. Since that time RBCK1 was found to be involved in endocytic pathways following cytokine stimulation [112], and Ymer/CCDC50 was found to suppress ligand-mediated

down-regulation of EGFR [109], thereby validating the original hypothesis derived from unsupervised learning.

The type of observation made in Zhang et al. [132], regarding a common endocytic functionality of several proteins in a cluster, involves extensive familiarity with the proteins and intensive manual curation. To determine if PTMScouts enrichment analysis of clusters could help bypass the onerous task of determining similarity, we evaluated the clustering solution given in Zhang et. al. through the arbitrary grouping interface of PTMScout. The interface to arbitrary data partitioning currently involves: exporting a dataset, appending it with cluster assignments, and then importing the appended file for enrichment analysis. The endocytic cluster highlighted in Zhang et al. was enriched for proteins containing the domains 'UIM', ubiquitin interaction motif, and 'VHS', which are both indications of endocytosis according to Pfam [29]. Additionally, this cluster was enriched for dynamic terms: "peak phosphorylation" at ten minutes and "peak down regulation" between ten and thirty minutes.

To further investigate the functional assignments of phosphorylation sites with temporal dynamics featuring peak activation at ten minutes followed by a quick dephosphorylation between ten and thirty minutes, we created a subset of data with these features using the data-driven subset selection interface. Specifically, we required phosphorylation at ten minutes to be 10% greater than phosphorylation at five minutes and 50% greater than phosphorylation at thirty minutes. This produced the subset shown in Figure 3-6, which includes the seven phosphorylation sites from the SOM endocytic cluster as well as phosphorylation sites on PTPN18, TFRC and CAV1. TFRC and CAV1 are both known to be involved in endocytosis, but functional characterization for PTPN18 in the EGFR pathway has not yet been elucidated. Based on the inclusion in this cluster, PTPN18 may also participate in the endocytic pathway. Another possibility is that phosphorylation of Y389 on PTPN18 may play a role in the negative regulation of EGFR since the temporal profile for this site follows so closely with the phosphorylation dynamics of the negative regulation machinery of endocytosis.

A

Protein	Site
ANXA2	Y30
CAV1	Y14; Y6; Y25
CCDC50	Y145
EPS15	Y849
PTPN18	Y389
RBCK1	Y288
STAM	Y381
STAM2	Y374
TFRC	Y20
TNK2	Y859; Y860

B

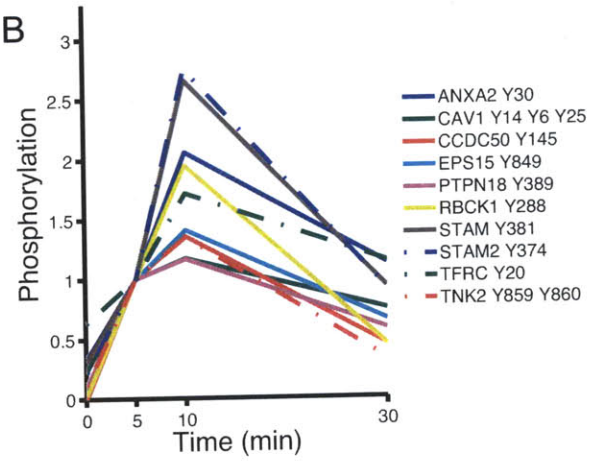


Figure 3-6: Extended endocytic subset from the EGF4 dataset. (A) The subset of the EGF4 dataset where phosphorylation at ten minutes is at least 10% higher than at five minutes and phosphorylation at ten minutes is at least 50% higher than at thirty minutes. This subset includes all members of the SOM trafficking cluster from [132] as well as sites on PTPN18, CAV1, and TFRC. (B) Dynamics of the extended endocytic subset.

3.3.5 Trypsin is potentially limiting in measurement of acetylation and '[GS]k' is an acetylation motif specific to RNA binding proteins

The *AcK* dataset [14], published in August 2009, is one of the most recently loaded experiments in PTMScout. This single dataset was larger than all other large-scale measurements of acetylation recorded in Uniprot (version 15.8, released in September 2009). Upon comparing it with curated acetylation datasets using the comparison tool implemented in PTMScout, we found that only approximately 2.5% of the approximately 3,200 acetylated peptides in the *AcK* dataset have been previously detected. If we include Uniprot annotation records that extend acetylation knowledge by predicted similarity among species, protein families, and other non-strict annotations, this fraction increases only to 5%. Although the *AcK* dataset as originally published [14] contains 3,885 acetylated peptides, PTMScout contains only 3,286 acetylated peptides, since the remaining peptides were given nucleotide record identifiers. PTMScout handles only proteomic accession types.

PTMScout allows users to view a breakdown of their dataset by annotation terms incorporated in the database (such as GO annotations, domain structures, kinase predictions, etc.), through the *Experiment Summary* functionality of PTMScout. Table 3.3 represents the top terms for GO Molecular Function (MF) and predicted Pfam domains of the AcK dataset [14]. One of the top MF annotation terms is 'None' (i.e. no MF annotation), indicating that many of the proteins acetylated in this dataset are not yet annotated with regards to function. The domain information in Table 3.3 represents the number of proteins containing the indicated domains in the dataset. As can be seen from this table, acetylation frequently occurs on proteins containing 'RRM_1' domains, which are thought to be an indication of an RNA binding protein [29]. This information is consistent with the prevalence of the 'RNA binding' term in the GO MF breakdown. Acetylation of histone proteins is present, as expected, but there is also a significant degree of acetylation on signaling proteins, as indicated by domains such as 'Pkinase', 'SH3_1', 'PH' and 'SH2'. Figure 3-7 illustrates a motif logo [17] for the entire *AcK* dataset, indicating the amino acid frequencies surrounding the site of acetylation. Surprisingly, there is an abundance of lysines surrounding the central modified lysine, with the exception of those positions most proximal (-3 to +2) to the central residue. This systematic enrichment of lysine in the vicinity of acetyllysine indicates that trypsin, which cleaves peptides to the C-terminal side of lysine and arginine residues, may not be the most efficient protease for high-throughput analysis of the acetylome, since it may be producing peptides too small to be analyzed and sequenced by reverse-phase liquid chromatography mass spectrometry (LC-MS). Clearly this technique was successful in identifying thousands of sites; however, the motif analysis would suggest that an improved approach might be to combine several samples processed using different proteolytic enzymes to achieve a more comprehensive coverage of the acetylome.

Acetyl transferases, like kinases, are thought to recognize linear amino acid sequences surrounding the site of modification [53]. The motif '[GS]k' (an acetyllysine preceded by a glycine or serine) was found to be a consensus sequence for two related acetyl transferases, CBP and p300, using direct substrate identification with recom-

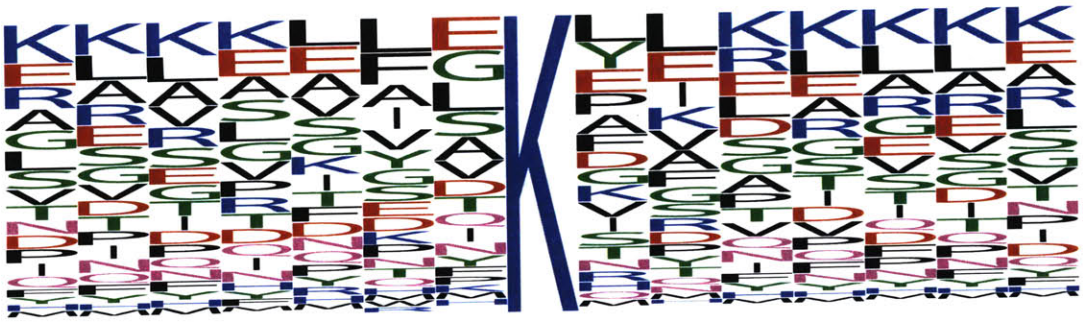


Figure 3-7: Summary: at-a-glance feature for the AcK dataset [14] includes a frequency motif logo [17] representation of all singly acetylated sites aligned on the central modified residue. There is a high frequency of lysines in all positions except those immediately proximal to the central residue.

binant acetyl transferase [7]. In addition to identification of a motif for these acetyl tranferases, the authors looked for an expanded role for acetyl tranferases beyond the canonical roles of histone and transcription factor modification, and found proteins such as Rch1, a nuclear importin, to be acetylated by CBP. We chose to look at the subset of acetylated sites in the large *AcK* experiment [14] that matched the consensus motif of CBP/p300 by searching for '[GS]k' sequences. This subset selection returned 656 acetyllysine sites, approximately 20% of the entire dataset, of which 292 were 'Sk' sites and 364 were 'Gk'. Histone proteins, as identified by the presence of a histone domain, are not significantly enriched in this subset, in agreement with the hypothesis that CBP/p300 can acetylate both histones and non-histone proteins. However, proteins responsible for acetylation of histone proteins are enriched in the subset of acetylation sites possessing the '[GS]k' motif, as indicated by enrichment of GO Biological Process (BP) terms histone H3/H4-K5/H4-K8/H4-K12 acetylation and GO Cellular Compartment (CC) term 'histone acetyltransferase complex'. Additionally, acetylation of 'RRM_1' domain-containing proteins is enriched within this subset, specifically for acetylation within the domain itself. If '[GS]k' is indeed specific to CBP/p300 recognition, our PTMScout results indicate that CBP/p300 is responsible for acetylation of RNA binding proteins and proteins indicated to be involved in histone acetylation.

Table 3.3: Terms for GO Molecular Function and Pfam domains for the *AcK* experiment [14] with the highest incidence. The number of terms present in the dataset represents the total number of proteins with that GO term or at least one of the indicated domains. There are 1,662 unique proteins in the dataset.

GO:MF	No. Prot.	Pfam Domains	No. Prot.
protein binding	664	None	56
None	401	RRM_1	55
RNA binding	62	WD40	38
DNA binding	52	Pkinase	37
ATP binding	50	Helicase_C	37
protein homodimerization activity	44	PHD	28
molecular_function	44	DEAD	26
transcription factor activity	41	AAA	20
identical protein binding	40	SH3_1	20
structural constituent of ribosome	39	PH	19
transcription coactivator activity	31	Bromodomain	17
ATPase activity	29	zf-C2H2	15
transcription factor binding	27	TPR_1	15
unfolded protein binding	26	CH	15
protein N-terminus binding	23	efhand	15
protein C-terminus binding	23	SH2	14
zinc ion binding	22	SAP	13
transcription activator activity	20	Histone	13
calcium ion binding	19	Myb_DNA-binding	11
enzyme binding	18	UQ_con	11
transcription corepressor activity	18	HEAT	11
GTPase activity	17	Ank	11
protein heterodimerization activity	16	zf-C3HC4	10
ubiquitin-protein ligase activity	16	PCI	10
actin binding	16	Filament	10
translation initiation factor activity	16	Cpn60_TCP1	10
single-stranded DNA binding	16	PWWP	9
protein serine/threonine kinase activity	15	Mito_carr	9

3.4 Methods

3.4.1 Database and data resources

The master database underlying PTMScout was built using MySQL. The database schema is outlined in Figure B-1. External, publically accessible protein information, including sequence, alternate accessions, gene names, and species is retrieved from

NCBI GenPept, Refseq [84], IPI [48], or Swiss-Prot [114], depending on the accession type given in a new dataset. Gene ontology terms are from The Gene Ontology consortium [5]. Species-specific annotation files and the current ontology file are downloaded from the Gene Ontology website and GO programming packages were used to parse annotation and ontology files. GO terms based on inferred electronic annotations (IEA), which have not undergone further curation, are not stored in the PTMScout database. Results in this paper were produced using downloaded files from GO v1.2 and annotation files retrieved December 12, 2009. Protein domain information comes from Pfam [29]. When possible, Swiss-Prot identifiers are used to parse domains from the current Pfam release. When lookup in the stand-alone Pfam release is not possible, the Pfam-A HMM library and the BioPerl Hmmpfam package are used to predict domains in a protein sequence. Predictions with scores less than 1e-5 are considered, and when there is overlap between domains, the domain with the most stringent score is kept. Results in this paper were produced from Pfam release 23. When a phosphorylation site falls into a predicted structural domain, we include this as a separate annotation of enrichment, denoted 'Pfam site'. Gene expression information comes from the Genomic Institute of the Novartis Research Foundation (GNF) SymAtlas project [107]. Expression information, analyzed by gcRMA, for human and mouse tissue types and the NCI60 cell lines were downloaded and placed in PTMScout as expression tables. GNF Symatlas annotation tables are used to link PTMScout proteins with appropriate tissue/NCI60 mRNA expression.

For phosphorylation sites, PTMScout currently includes predictions of the responsible kinases and binding partners from Scansite [76], when available. Scansite predictions for an input peptide sequence are automatically retrieved, parsed and then stored in prediction tables of PTMScout. PTMScout Scansite prediction stringencies correspond with suggested scores from Scansite. Ambiguous peptide-protein assignments are identified by exact match of the full-length peptide sequence identified by MS among all of the protein data sources imported to PTMScout, which is also expanded to include proteins within the relevant species by searching the RefSeq [84] database for a peptide match.

Curated datasets of phosphorylation sites were obtained from Phospho.ELM [22] and PhosphoSite [39], by request. Automatic curation of Uniprot [114] for phosphorylation and acetylation is performed by searching for both large-scale analysis terms (example search: [PHOSPHORYLATION LARGE SCALE ANALYSIS AT]) as well as modified residues (search: [MOD_RES]). Uniprot search results were then parsed and placed into a PTMScout loadable format.

Kinase activation loop predictions are based on finding the conserved amino acid sequence 'DFG' to the N-terminal side of the modification and a flanking 'APE' to the C-terminal side [75]. This exact requirement matched 58% of all kinase domains in the PTMScout database version 1.1 as of January, 2009. We used ClustalW2 [56] to align all kinase catalytic domains within PTMScout, which at the time included 306 domains. We found that 180 of them had both the conserved 'DFG' and 'APE', while 70 had only the conserved 'DFG' sequence. Those that do not match the motif exactly usually have partially conserved flanking sequences, such as 'DYG' or 'SLE'. On average the two surrounding motifs were within 25 amino acids of each other, and 83% of proteins contained the motifs within 22 to 27 amino acids of each other. Based on the resulting ClustalW2 alignment, we developed a set of rules for identifying activation loop modifications. First, if the amino acid sequence surrounding the modification site contains a 'DFG' and an 'APE' motif, or degenerate sequences 'D[FPLY]G' and '[ASP][PILW][ED]' spanning less than 35 amino acids, it is marked confidently as being within the activation loop. If degenerate matches are made and are more than 35 amino acids apart, then it is marked as potentially being within the activation loop.

3.4.2 Calculations

Selection of foregrounds occurs at the level of proteins, the level of experimentally measured peptides, or the level of individual sites of post-translational modification, depending on the category of data or annotation being used for the selection. For example, Gene Ontology terms and structural domain criteria will select subsets at the protein level, whereas quantitative data will select at the measured peptide level,

and enzyme specificity predictions and sequence motif features will select at the PTM level. We define the p-value for enrichment of a characteristic in a foreground, relative to a background, as the probability that a characteristic would be as enriched, or more enriched, if the foreground were randomly selected from the full data. This quantity can be calculated exactly using the hypergeometric distribution. The probability of having k or more labels in the foreground occurring by random chance when we choose any n objects from the background, size N, having a total of K objects with that same label, is calculated as:

$$p(k') = \sum_{k'=k}^{min(n,K)} \frac{\binom{K}{k'}\binom{N-K}{n-k'}}{\binom{N}{n}} \tag{3.1}$$

In order to determine k, K, n, and N for a label, translation from the selection criteria specificity to the label specificity of interest is performed. Not all mappings of selection specificity to label specificity are 1:1. For example, quantitative measurement selection may lead to redundant selection at a protein level. A search for significantly enriched amino acid sequence motifs was performed using a previously published greedy search algorithm with a search index of +/-7 amino acids surround the site of modification and a branch cutoff term of 0.01 [46].

Categorical multiple hypothesis correction can be user-corrected through the PTMScout interface. Bonferroni [24] is the most stringent correction method, where the corrected alpha is the desired p-value divided by the number of labels tested. False discovery rate is implemented according to the method of Benjamini and Hochberg [10].

PTMScout can be found online at http://ptmscout.mit.edu. A tutorial for using PTMScout to obtain the results presented in this paper can be found in the PTMScout documentation. Unless otherwise noted, results are from version 1.2 of PTMScout, which includes Pfam Release 23, Gene Ontology annotations from version 1.2 downloaded on December 12, 2009, and Uniprot compendium results from Release 15.11. All protein records in PTMScout version 1.2 have been retrieved after December 11th, 2009. All terms considered enriched in the results have an FDR adjusted p-value of 0.05 or better, unless noted otherwise.

3.5 Conclusions

PTMScout provides uniform, web-based access to MS measured PTM data and automates much of the feature selection and information extraction that is currently performed manually following MS analysis of biological samples. For example, residue position assignment and comparison to PTM data compendia for discovery of novel PTM measurements are intensive manual operations that are performed automatically in PTMScout. Additionally, programmatic access of protein databases by PTMScout during dataset loading allows for protein assignment error checking, thereby correcting typical errors in protein assignment including redirected records, updated records with significantly changed sequence information, and species assignment errors. While PTMScout can automatically handle most protein record redirections, protein errors causing terminal failures due to sequence mismatch between the peptide and the assigned protein are reported in an error log; erroneous species assignments can be seen easily in the dataset summary function of PTMScout. Furthermore, PTMScout allows for user-defined uploading of their own mass spectrometry datasets.

Data analysis by subset selection and subsequent enrichment have proven to be useful tools for deriving biological hypotheses. However, hypothesis generation is currently limited by metadata annotations. For example, the endocytic cluster found in the *EGF4* experiment [132] had only a few GO annotations indicating a role in endocytosis despite several reports demonstrating that the majority of proteins in the cluster participate in the endocytic pathway. Despite these limitations, the cluster featured enrichment of UIM and VHS domain containing proteins, thereby enabling hypotheses regarding phosphorylation of specific sites and regulation of endocytosis. An expanded endocytic cluster was generated in PTMScout through use of relative quantitative dynamics, leading to identification of another protein which may be involved in EGFR endocytosis. Although the richness of hypotheses and observations is expanded by inclusion of relative quantitation across multiple conditions, PTMScout is also successful at deriving insight from datasets without quantitation, as demonstrated with the *AcK* dataset [14].

By considering the composition of the entire dataset, enrichment testing provides a way to uniquely label a dataset partition. An interesting cluster highlighted in the *EGF4* study was an 'early response cluster', which was composed of several known EGFR binding proteins. Interestingly, enrichment for quantitative dynamic features failed to corroborate this feature as being specific to that cluster alone. Despite the fact that all members of the cluster experienced a large increase in phosphorylation within the first five minutes of stimulation, this 'early response' label is applicable to more than two-thirds of the entire dataset and therefore, although this label is correct, it is not a unique feature of that cluster compared to the remaining dataset. By performing enrichment of subsets compared to the background of the dataset itself, versus the entire phosphoproteome or acetylome, experimental biases, such as antibody specificity or MS fragmentation patterns, are eliminated.

PTMScout is a widely and readily accessible, user-friendly, web-based PTM database with multiple bioinformatic tools to enable automated feature selection and subset generation. Here we have demonstrated the application of PTMScout to multiple published phosphorylation and acetylation datasets, leading to multiple hypotheses regarding the potential functionality of various proteins and PTM sites. As more experimental datasets are loaded into PTMScout additional biological insight, not currently available from individual datasets, will emerge, as we are able to compare the regulation and response of individual phosphorylation sites under a variety of conditions. Application of PTMScout to quantitative PTM datasets will facilitate the main data analysis challenge facing high-throughput PTM proteomics, and will provide putative functional assignments to a greater percentage of previously uncharacterized sites.

Chapter 4

High-Throughput Quantitative Phosphoproteomic Dataset Analysis Using Combinatorial Parametric Unsupervised Learning

4.1 Summary

Many cellular receptor systems utilize phosphorylation as a means to transduce extracellular signals into phenotypic responses, such as changes in migration or proliferation. The extent of phosphorylation modifications in these signaling cascades is widespread, however the biological function of the vast majority of modification sites is currently unknown. Quantitative mass spectrometry (MS) is capable of measuring the relative phosphorylation of hundreds to thousands of individual peptides under various signaling conditions, which may be useful in understanding the normal function and regulation of these sites, as well as their role in disease. Unsupervised learning methods have proven useful in highlighting interesting biology within MS datasets, but its implementation has been limited to a few datasets and the use of single clustering methods, without systematic exploration of the space of possible methods.

We have employed unsupervised learning in such datasets in order to gain an understanding of phosphorylation events in the epidermal growth factor receptor (EGFR) network. We hypothesize that the parameters of unsupervised learning, including data pre-processing, distance metric, number of clusters, and choice of algorithms will play an important role in deriving specific types of biological understanding. Therefore, we apply a combinatorial parametric approach to unsupervised learning, producing hundreds of cluster sets using comprehensive combinations of a select set of data transforms, distance metrics, algorithms, and targeted number of clusters. To evaluate the biological information represented in this large number of cluster sets, we employ automatic enrichment analysis of biological labels, such as predictions of the responsible kinase and Gene Ontology terms. We have found that in some cases, a set of parameters can optimally derive clusters with enrichment in a particular type of biological labels. In other cases, a combination of several parameters is better suited for deriving the full spectrum of possible enrichments in a category of biological information in a given dataset. These dependencies of parameter and information enrichment are dataset-dependent. Additionally, the integration of clustering analyses produces a metric for robustness of phosphopeptide co-clustering that may pinpoint phospho-specific signaling interactions within the network. This method correctly associates EGFR phosphorylation sites with their known immediate effector proteins; for example, EGFR Y1172 is closely associated with phosphorylation of Shc at sites Y427, Y349, and Y350; whereas EGFR Y1069 is associated with phosphorylation of Cbl on three tyrosine residues, but most strongly with the phosphorylation of Cbl at Y552. Correspondingly, the SH2 domain of Shc has been shown to bind EGFR Y1172, while EGFR Y1069 phosphorylation binds the E3-ubiquitin ligase Cbl, which is associated with internalization of the receptor. Site-specific characterizations of known relationships, as well as generation of novel relationships, may be generated using this method, which may aid in our ability to construct phosphorylation-specific network models.

4.2 Introduction

In receptor tyrosine kinase (RTK) networks, such as the epidermal growth factor receptor (EGFR), phosphorylation plays a central role in the translation of extracellular cues into phenotypic changes, such as differentiation, proliferation, and migration. Phosphorylation on proteins in the RTK network induce a variety of signaling events including protein-protein interactions, enzymatic activation and inactivation, and cellular localization changes, such as translocation to the nucleus or recruitment to the plasma membrane. Understanding RTK networks, and the phosphorylation that occurs within them, will be essential for developing representative signaling models. These models are helpful for representing both typical and dysregulated networks. Dysregulated RTK networks play an important role in disease progression. For example, EGFR mutations and amplification are associated with a variety of cancers.

Mass spectrometry measurement of phosphorylation events in cellular signaling networks is providing immense insight regarding the resolution of protein modifications, greatly increasing our understanding of the specific modifications occurring in the cell as well as their relative changes in response to network perturbations, such as ligand stimulation or kinase inhibition. The success of MS methodologies in quantifying increasingly larger degrees of measurements, in both conditional and dynamic space as well as more sites on more proteins, means that the datasets are beyond the scope of manual analysis. A few groups have turned towards unsupervised learning in attempts to reduce the dimensionality of their datasets into comprehensible biological network components for a better understanding of the underlying biology [78, 132], motivated by the success of such applications in the field of microarray expression data and other types of high throughput biological data analysis. Phosphoproteomic data represents a new challenge in unsupervised learning and to date no extensive analysis of unsupervised learning has been applied to a dataset of this type.

Application of unsupervised learning to biological datasets is extensive and includes a seemingly endless option of algorithms, such as Kmeans [110], hierarchical clustering [25], self organizing maps [108], and affinity propagation [30]. Unsupervised

learning algorithms seek to group a multidimensional dataset into clusters where intra-cluster differences are minimized and inter-cluster differences are maximized. Therefore, the criteria used to judge cluster fitness, i.e. the distance metric, is an important factor in determining the final clustering solution [21, 43, 83]. Also, transformations of the data can similarly effect the relationships between the data vectors, and will effect the final result [21, 115]. In addition to choosing an appropriate algorithm, distance metric, and data transformation, scientists are faced with also having to determine a suitable number of clusters (K) in which to partition their dataset since few algorithms incorporate concurrent optimization of K. A variety of methods for determining the natural cluster structure of a dataset have been proposed, see [32] for an example of their comparative performances using a set of microarray experiments. Taken as a whole, the historical application of unsupervised learning to microarray data, as well as other large biological datasets, paints an overwhelming picture of variability and possibility.

In this work we propose a semi-automatic framework that capitalizes on protein annotation information, such as Gene Ontology terms, kinase predictions and linear amino acid sequences. Enrichment of these terms allows for a high-throughput evaluation technique for measuring the performance for a set of algorithms, distance metrics, data transformations, and clustering set sizes (K). We apply this framework to dynamic phosphoproteomic measurements of the EGFR network in human mammary epithelial cells. We find, as has been seen in microarray analysis [21], that the exact choice of all the possible variables in clustering analysis is heavily dependent on the type of information desired from such an analysis as well as the specific dataset in question. Additionally, we find utility in taking the vast array of clustering solutions as a whole, both in the robust biological hypotheses and stories that result and their direct overlap with known protein network information. We find that highly co-regulated phosphorylation dynamics, as viewed by the robustness of clustering solutions, recapitulates known protein interactions as well as known phosphorylation site specific interactions. We can recapitulate Cbl recruitment to EGFR Y1069 as well as the shared functionality of Y1172 and Y1197 in recruiting Shc and Grb2 to the

receptor. The information derived in the various applications of this process will help inform our understanding of the role and regulation of individual phosphorylation sites in RTK networks.

4.3 Results

In this work we will focus on measurements of the epidermal growth factor receptor (EGFR) network. We evaluate a seven time point measurement of human mammary epithelial cells (HMEC) stimulated with a saturating concentration of EGF, where measurements were taken before stimulation (0 min) and then subsequently at 1, 2, 4, 8, 16, and 32 minutes following EGF addition [122]. Enrichment and fractionation steps focused on capturing proximal EGFR tyrosine phosphorylation signaling events. This dataset represents extensive measurement of the phosphotyrosine EGFR signaling network, with 204 unique phosphopeptide measurements and high resolution at early time points. Most phosphopeptides represent singly phosphorylated residues, but eleven peptides are doubly phosphorylated. Throughout this work we will refer to this dataset as *EGF7* for brevity.

It has been shown in unsupervised learning of expression datasets that varying individual components of an unsupervised learning system, such as the transform or the distance metric will perturb the measured relationships in various ways, impacting the final clustering solution [21, 43]. To demonstrate this holds true in phosphoproteomic datasets, we clustered the *EGF7* dataset into nine clusters using Kmeans and self organizing map (SOM) algorithms. Figure 4-1 illustrates the clustering solutions of a single parameter change relative to a parent set using Kmeans, a "normMax" transformation (each peptide vector is normalized to its own maximum value) and Euclidean distance. Since the solution of Kmeans and SOM is dependent on the randomized staring point, the random seed was set as the same value in every implementation. Visually, one can see that changing the algorithm, the distance metric, or the transformation heavily effects the final solution. Mutual information can be used to quantify the total difference between any two sets. In this example, the distance

metric change from Euclidean to Cityblock has the smallest impact, whereas changing the algorithm has the largest impact on the final solution.

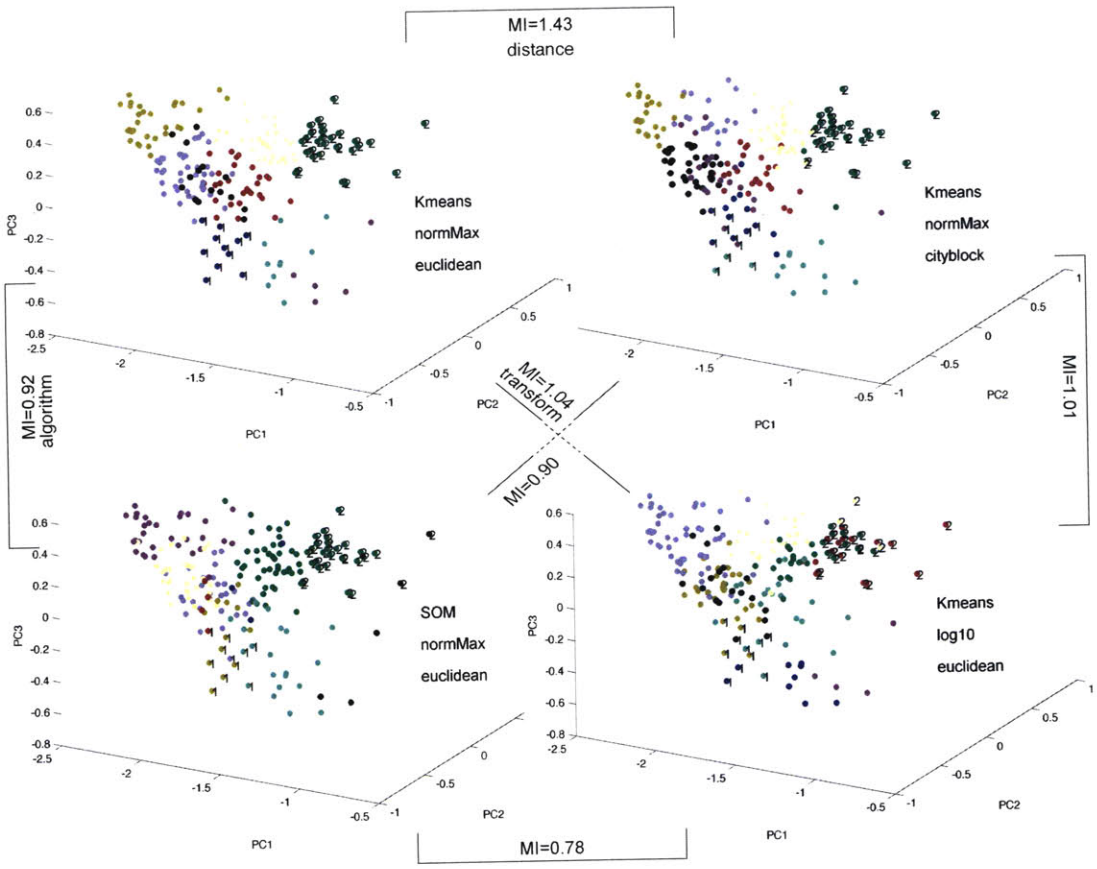


Figure 4-1: Changes in parameters during unsupervised learning impact the final clustering solution. Each figure represents the partitioning of the *EGF7* dataset into nine clusters by the indicated algorithm, transform, and distance metric. The random seed, and therefore the algorithms staring point, is fixed for all cases. For visualization purposes the 204x7 dataset has been projected onto the first three principal components and all members of a cluster are represented by the same color. The mutual information between each set is given to quantify the absolute differences. For reference, two clusters in the top left implementation have been numbered and those numbers have been carried over in the other sets. SOM refers to self organizing map, normMax refers to taking the normalization of each time vector according to its own maximum value, and log10 refers to taking the base-10 log transform of each vector.

Since an optimal dataset partitioning solution is not known *a priori*, we must

determine a method in which to judge the fitness of a particular clustering solution. We choose to judge fitness using known annotations of the biological molecules in the dataset, specifically information about the structure and function of the proteins as well as information about the regulation and function of the phosphorylation sites. We also explore the resulting dynamic partition labels based on relative levels of phosphorylation at each time point. This fitness test is performed in an automatic fashion; labels in each biological and dynamic metric are tested for statistically significant overrepresentation within every cluster of a set. We control for type I error, or an excess of false positives due to multiple hypothesis testing, by applying the false discovery rate (FDR) procedure [10] at the level of each set within each metric. See Figure 4-2B for a depiction of an MCA set, enrichment, and correction. Table 4.1 defines the biological metrics used in this study as well as metric abbreviations that will be used throughout the remaining work.

Table 4.1: Description and categorization of metrics. There are a total of 13 metrics analyzed. The metrics tested describe information regarding the protein a peptide arises from, the particular site of phosphorylation, or the quantitative data. The abbreviated labels used throughout this work are given.

Metric Level	**Metric**	**Short Name**
Protein Metrics	Gene Ontology: Molecular Function	F
	Gene Ontology: Biological Process	P
	Gene Ontology: Cellular Compartment	C
	Pfam Domains	Pfam
Site Metrics	Domain Phosphorylation	Pfam Site
	PhosphoELM Kinase Annotations	PELM Kinase
	Linear sequences	Motifs
	Scansite Kinase Predictions	Scansite Kinase
	Scansite Binding Predictions	Scansite Bind
Dynamic/ Quantitative Metrics	Minimum Phosphorylation	MinValue
	Maximum Phosphorylation	MaxValue
	Maximum Positive Change	MaxPosChange
	Maximum Negative Change	MaxNegChange

Motivated by the hypothesis that altering the parameters of unsupervised learning, thereby perturbing the relationships between phosphopeptide measurements, will yield different but meaningful biological enrichment, we designed a semi-automatic

framework for testing the application of a variety of parameter combinations. Figure 4-2 illustrates the workflow for applying combinatorial analysis to a biological dataset of interest. The data is first subjected to a set of transformations before being clustered using a combination of algorithms, distance metrics, and target set sizes. Each of these parameters is combined with the others producing M sets of sets, on the order of 500 sets total, which we will refer to as a "Multiple Clustering Analysis", MCA. The MCA size, and therefore the number of parameters chosen in each round, is limited in order to perform all subsequent computational steps on an average computer. Biological label enrichment in each cluster of all sets is calculated using a hypergeometric test and deemed significant if it falls below an FDR corrected alpha value of 0.05. Parameters that perform poorly with respect to production of label enrichment can be removed in subsequent rounds. Iterations of parameter selection, clustering, enrichment and parameter rating can continue until improvement ceases, and then the loop is exited with a final MCA. This final MCA will then be used in further analyses with the aim of generating testable hypotheses concerning interacting and tightly regulated phosphorylation sites.

4.3.1 Evaluation of unsupervised learning parameters in the *EGF7* dataset

A number of MCA iterations were applied to the *EGF7* dataset. Each round consisted of a selection of the full parameter set listed in Table 4.2, where the final MCA parameters are listed in the 'Final' column of Table 4.2. The relationships between the parameters of unsupervised learning and particular biological information were evaluated by searching for parameter enrichment in rank-ordered lists of the MCA sets based on total number of labels enriched in each of the thirteen categories, Figure 4-3A. Kmeans appears to be particularly useful at producing enrichment in Pfam domains, GO BP and GO MF terms. GO CC, another metric describing protein level information, is not composed significantly of any particular type of algorithm. There are a few terms that consistently appear in the bottom 25% of rankings, such

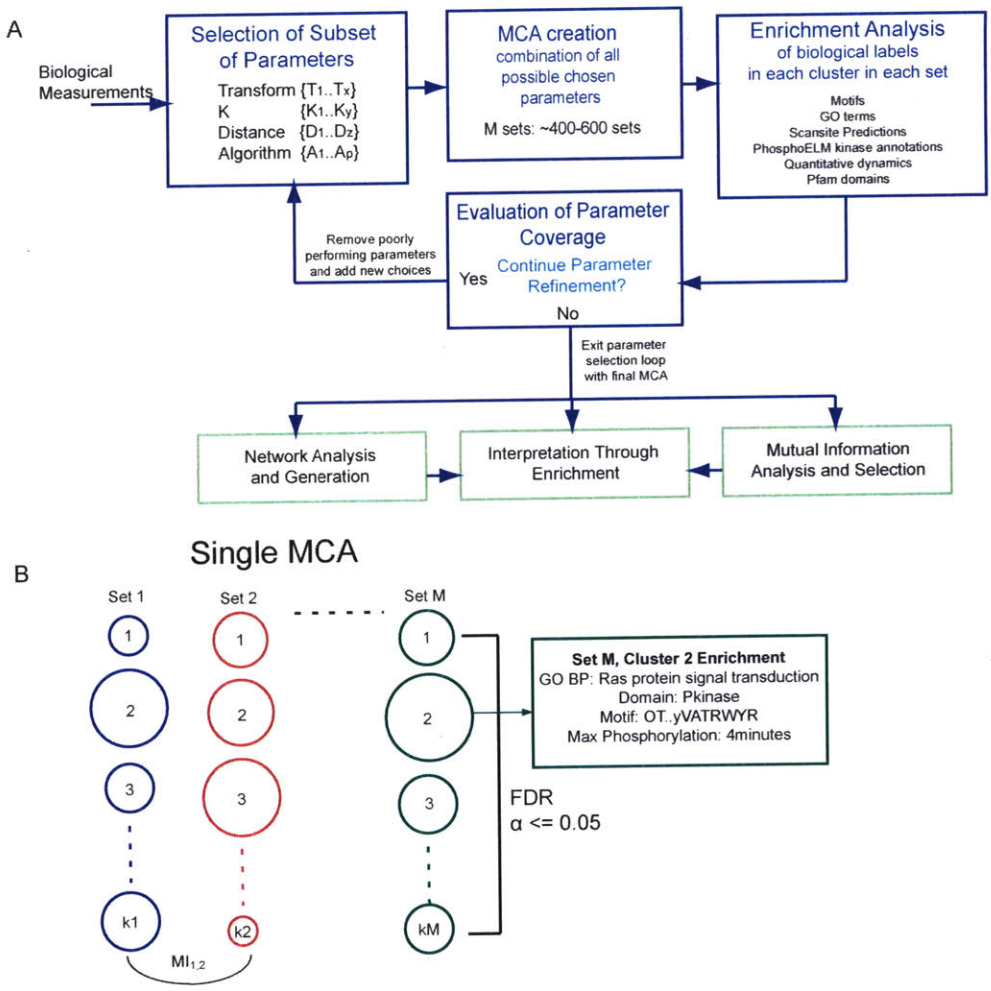


Figure 4-2: The workflow and terminology of parametric combinatorial analysis of biological datasets. A) Quantitative, high-throughput biological measurements are subjected to a battery of clustering analyses, which are combinatorial applications of a chosen subset of parameters in unsupervised learning. A "Multiple Comparison Analysis", MCA, is a set of clustering solutions produced in the iteration of one loop. Enrichment analysis aids in rating the performance of each set of parameters. Once a refined MCA has been developed it can be evaluated through three main modes of analysis: network generation, enrichment interpretation and mutual information. B) Depiction of an example MCA composed of M sets, each with a different number of clusters. Enrichment is calculated for each cluster for all labels in a set of biologi cal metrics. Type I error is controlled for each metric type by using the False Discovery Rate (FDR) procedure across all clusters in a set.

as K=5 and K=10. When sets with parameters of K=5 and K=10 are removed, the average enrichment per set in all categories generously improved. Conversely, removal of affinity propagation (AP) clustering, which appears in the top 25% of several categories, decreases enrichment per set in all categories. This evaluation technique was employed in each round of MCA generation to create a final, satisfactory parameter set.

Table 4.2: The parameters considered across multiple iterations of MCA creation and the parameter subset used for MCA_{final}. See Methods for exceptions in algorithms with regard to distance metric and K limitations.

Parameter	**Full Set**	**Final Set**
Transforms	None zscore normMax center pareto rangeScale log10 rangeScale_differential FFT zscore_FFT normMax_log10	None zscore normMax rangeScale log10 power(0.5) differential rangeScale_differential zscore_FFT normMax_log10
K	5,10,15,18,20,24,25,28,30,35,40	15,20,25,30
Distances	Euclidean Correlation CityBlock Cosine Chebychev Minkowski Spearman 1/MI	Euclidean Correlation CityBlock Cosine Chebychev
Algorithms	Ncut Affinity Propagation Kmeans SOM Hierarchical	Ncut Affinity Propagation Kmeans SOM

In order to find those sets that produce the best enrichment in all categories, the intersection of the best 25% of sets was taken. Surprisingly, the intersection produced

Data Type	Parameter	F		C		P		Pfam			
		Best	Worst	Best	Worst	Best	Worst	Best	Worst		
Protein	Transform K Distance Algorithm	normMax 23, 25, 27, 30 city, euclid AP, Kmeans	5, 10 corr Ncut	pow cheby, city	5, 10 corr, eucl Kmeans	normMax 23, 25, 27, 30 euclid AP, Kmeans	5,10 corr Ncut	normMax 25, 30 city, euclid Kmeans	log10 5 corr		
		Pfam Site		**Motifs**		**Scansite Kinase**		**Scansite Bind**		**PELM Kinase**	
		Best	Worst	Best	Worst	Best	Worst	Best	Worst	Best	Worst
Site	Transform K Distance Algorithm	25, 27, 30 city AP	5,10 eucl Kmeans	log10 25, 30 Kmeans	FFT, zscore 5	log10, pow eucl SOM	5,10 city, eucl Kmeans	log10, pow eucl SOM	5,10 eucl Kmeans	diff, log10, pow 23 city AP	zscore 5,10 corr, eucl Kmeans, Ncut
		MinValue		**MaxValue**		**MaxPosChange**		**MaxNegChange**			
		Best	Worst	Best	Worst	Best	Worst	Best	Worst		
Dynamics	Transform K Distance Algorithm	diff, zscore corr, cosine Hierarchical	FFT 5 city Kmeans, Ncut	normMax, zscore 17, 25, 30 Kmeans	diff, FFT 5,10 Hierarchical	zscore 17,25,30 AP, Kmeans	FFT 5	zscore 17, 25, 30 corr AP, Kmeans	diff, FFT 5 city		

Figure 4-3: Unsupervised learning parameters and biological enrichment dependencies. Each set in an early iteration of MCA, comprised of 665 sets, was ranked according to performance for enrichment of labels in each biological metric. Overrepresentation of parameters in the best and worst quartile is listed, if it exists. Overrepresentation is judged by a hypergeometric test value less than 1e-4. Abbreviations: Distances: (city)block, (euclid)ean, (cheby)chev, (corr)elation; Transforms: normMax (each vector normalized to its own maximum), (diff)erential, (pow)er-square root, log10 - log base 10, FFT-fast frequency transform.

a null set; no single set appears in the best 25% of all categories. In order to further evaluate the relationship between parameters and particular types of biological information, we did a pairwise comparison of the best and worst ranking quartile of sets across the thirteen metrics. Figure 4-4 indicates that for both protein- and dynamic-level information, there is a good deal of overlap within the same category type. Generally, when there is a high degree of protein-level enrichment, there tends to be a large degree of enrichment in other types as well. Additionally, we observe clustersets with the best enrichment in site-level metrics, specifically Scansite predictions and Phospho.ELM kinase annotations, have significantly decreased overlap with sets that perform the best in dynamic-level enrichment.

We utilized mutual information (MI) as another method for studying the similarity of clustering solutions and the parameters that gave rise to the solutions. Specifically, we use an MI estimation algorithm, MIST [50], as the metric for set similarity. The MI value for every pairwise comparison of sets within MCA_{final} is shown in matrix form in Figure 4-5A. Hierarchical clustering was used to group the MI values to determine if any particular aspect of clustering drives similarity and dissimilarity of clustering solutions. Based on groups generated in the hierarchical clustering of the MI matrix, we found the highest similarity was typically determined by the transform and then secondarily by the algorithm. Only at very small groupings of the clustergram are relations sorted by distance metric. Examples of groups that were all described by a single parameter set are shown as labels on the heatmap, which demonstrates the differential, FFT, log10, and rangeScale transforms mark the largest separation between set architectures in the MCA. Next, the algorithms SOM and Ncut appear to be distinguishable, whereas the remaining unlabeled sections are composed of a indistinguishable mix of algorithms, distances and transforms.

There are many methods that may be used to reconcile the currently popular application of a single algorithm, distance metric, set size, and transform with our proposed Mutliple Comparison Analysis, based on selection of some subset of the MCA sets for individual analysis of a more traditional kind. One method is to simply choose the top performers in each of the categories for direct analysis. Here, we

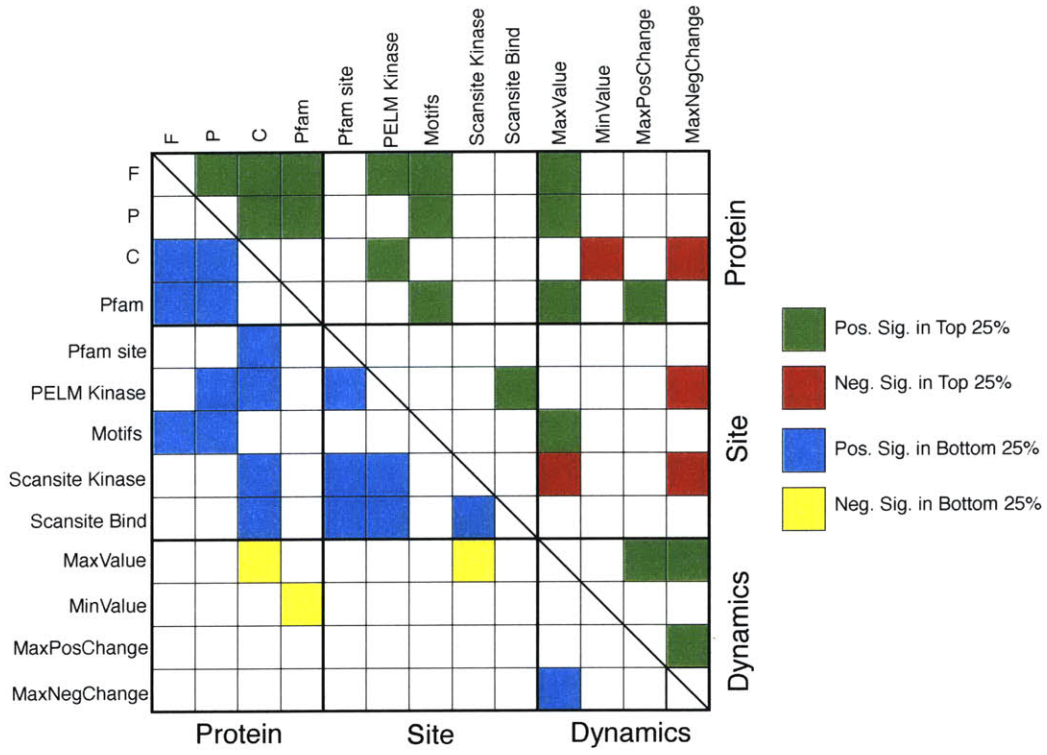
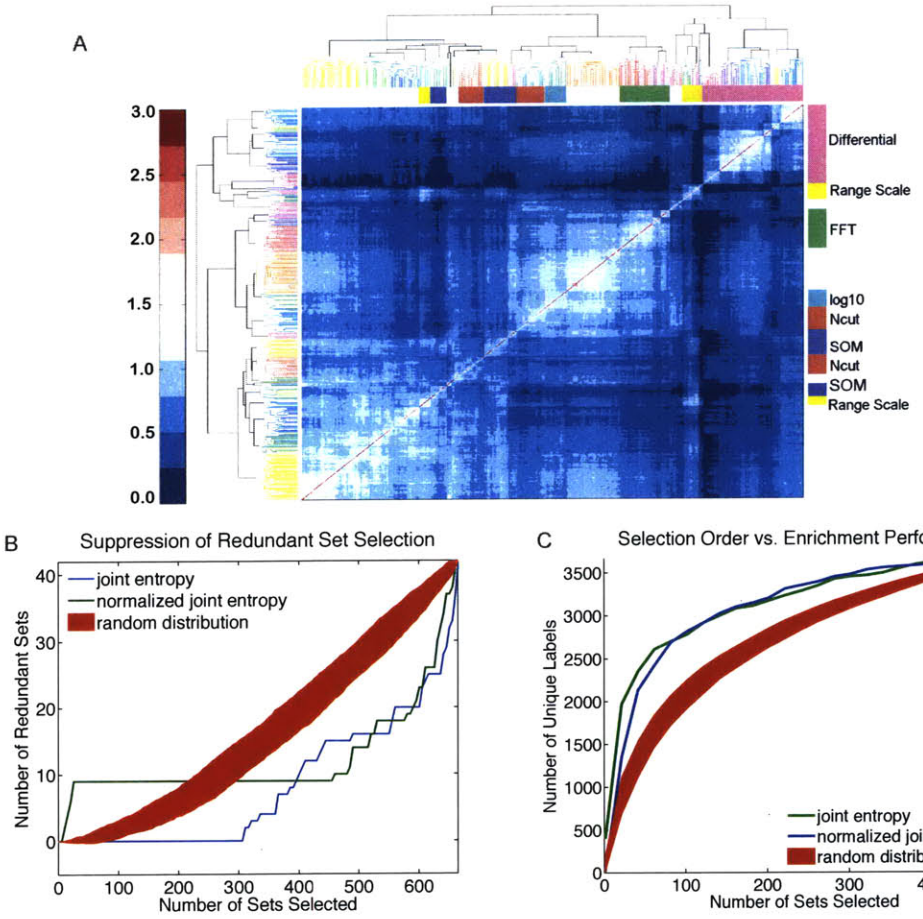


Figure 4-4: Pairwise comparison of the overlap in the best and worst 25% of sets based on each metric in MCA_{final}. We performed 1000 random selections of two sets of the same size to generate a normal distribution whose mean represents the expected overlap value between any two sets pulled from that background size. We then evaluated whether pairwise overlap was significantly higher ('Pos. Sig') or lower ('Neg. Sig.') than expected by random. Significance cutoff was set at a Bonferroni corrected alpha value of 0.05. The top right represents the pairwise comparison of the best performing 25% and the bottom left is the comparison of the worst performing 25% of sets in each category.

propose selection of sets that represent the highest degree of dissimilarity in order to understand the spectrum of possible results and biological enrichment, using joint entropy as a selection criterion. The sets in an MCA are ordered based on their relative impact on joint entropy when removed from the set. We found that selection of sets by this order resulted in two important effects. First, selection by entropy suppresses the selection of completely redundant sets within the selection, which agrees with the goal of this method of set selection, i.e. to guarantee the most diverse set architectures, Figure 4-5B. The second observation is that selection by joint entropy improves selection for a variety of resulting biological enrichment, indicating that selection of the most variable architectures will also yield the most diverse biological hypotheses highlighted by enrichment. An example of the first five sets within MCA_{final} are shown in Figure 4-5D.

An alternative to traditional analysis of individual sets is to evaluate the sets *en masse* by looking at the biological enrichments that occur throughout a majority of all sets in a MCA. Figure 4-6A indicates the top six biological labels enriched in every category and the number of times they occur in MCA_{final}. In some cases, the number of times a label is enriched is larger than the number of sets in an MCA, indicating that multiple clusters within some sets are annotated with that label. For example, the label indicating maximum positive change occurring between 0 and 1 minute is enriched, on average, in three clusters of every set. Specific biological stories emerge from these redundant enrichments, for example several of these labels denote the clustering of MAPK signaling components denoted by: MAPK activation loop phosphorylation (motif 'HTGFLTEyVATRWYR'), biological process term 'Ras protein signal transduction', and phosphorylation occurring within serine/threonine kinase domains (Pfam Site: 'Pkinase'). Phosphorylation within tyrosine kinase catalytic domains is also consistently seen (Pfam Site: 'Pkinase_tyr'), indicating that several tyrosine kinases are potentially activated in a similar dynamic manner. We see consistent grouping of proteins with similar domains, such as 'SH3_1' and 'UIM', where 'UIM' is short for ubiquitin interacting motif. 'UIM' enrichment may indicate the robust creation of a cluster involved in ubiquitin-mediated endocytosis of the



D

Joint Entropy	rangeScale 30 cityblock Kmeans	(rangeScale)differential 30 correlation Kmeans	log10 30 correlation Kmeans	pow 30 cityblock Kmeans	normMax 30 cityblock Kmeans
Normalized Joint Entropy	(zscore)FFT 30 euclidean SOM	(rangeScale)differential 30 correlation Kmeans	rangeScale 30 euclidean SOM	differential 27 euclidean SOM	rangeScale 30 cosine Ncut

Figure 4-5: Mutual information as a set theoretic and selection criteria. A) A hierarchically clustered heat map of all pairwise MI calculations of MCA_{final}. It is denoted where groups are composed entirely of a particular parameter. B) Performance of selection by joint entropy to suppress selection of redundant set architectures. A randomized control of 100 selections is plotted, the mean +/- a standard deviation, is given for comparison. The first round of MCA is used since it contains a larger number of redundant sets. C) Selection by joint entropy maximizes selection for unique biological enrichments in MCA_{final}. A control of 100 random selections is given, plotted as the mean +/- the standard deviation. D) The parameters in the top five clustersets in MCA_{final} chosen by entropy and normalized entropy algorithms.

receptor.

A randomized version of the data matrix was generated and then clustered using the same set of parameters as MCA_{final} in order to validate enrichment results and measure the false positive rate. The rates of enrichment per cluster and the rate of rejected hypotheses per test performed are shown in Figure 4-6B and C. As expected, a dramatic increase in rate of enrichment of dynamic terms, i.e. null hypothesis rejection, in the results of real data versus random data is seen. Specifically, it appears that clustering is differentiating when increases in phosphorylation and maximum phosphorylation occur. If we assume that all null hypotheses rejected in randomized data are false positives, then Figure 4-6C depicts the false positive rate in three random controls. The false positive rate for all categories is controlled below the target alpha value of 0.05, sometimes 10- to 100-fold below. The only exception of this is PELM Kinase annotations.

4.3.2 Inferring phosphosite-specific signaling layers through robust co-clustering

Another layer of information that can be derived from the creation of an MCA is the robustness of phosphopeptide co-clustering. We evaluate this by creating a "co-occurrence" matrix, which consists of the value of the number of times any two phosphopeptides are seen in the same cluster within all sets of the MCA, Figure 4-7A. Given the high variability of results in various parameters, this metric demonstrates how stable the association is between any pair of measurements. We found there was extremely significant overlap, on the order of 10^{-12}, between the co-occurrence matrix values and known protein interactions (based on GeneGO network information), Figure 4-7B. This method performs better than random when used as a selection criteria for protein interactions, Figure 4-7C. Figure 4-7D indicates the total number of interactions deemed significant at a particular cutoff of co-occurrences. We evaluated our randomized matrices used in Figure 4-6, which were also subjected to the MCA_{final} parameters, as randomized controls for network information. The distribu-

A

	F		P		C		Pfam	
Protein	protein binding	757	signal transduction	333	cytosol	131	Pkinase_Tyr	436
	protein serine/threonine kinase	248	Ras protein signal transduction	306	nucleus	121	SH3_1	363
	---	228	---	291	cytoskeleton	53	Pkinase	354
	cadherin binding	166	cell proliferation	227	basolateral plasma membrane	36	Pfam-B_10374	197
	transcription factor binding	152	nervous system development	226	intracellular	32	UIM	196
	hematopoietin/ interferon-class	145	insulin receptor signaling pathway	193	plasma membrane	32	PDZ	179

	Motifs		Pfam site		Scansite Kinase		Scansite Bind		PELM Kinase	
Site	HTGFLTEyVATRWYR	264	Pkinase	219	PDGFRB_Y_kin	48	CRK_SH2	70	---	330
	....EE.y...P..A	162	Pkinase_Tyr	165	FGR_Y_kin	29	SHC1_PTB	69	MAP2K1	174
	K..S...y.L.....	149	Annexin	54	ABL1_Y_kin	21	NCK1_SH2	67	EGFR	99
	.....yO.SR.YR	116	~~~	15	---	16	---	56	ZAP70	80
	.....y...EI.O	108	Pfam-B_6279	12	EGFR_Y_kin	11	ITK_SH2	52	FLT4	77
	...PE.Vy.....O	106	Pfam-B_7012	10	SRC_Y_kin	7	FYN_SH2	21	Fyn	45

	Max Value		Min Value		MaxPosChange		MaxNegChange	
Dynamics	8 min	1516	1 min	236	0 min->1 min	1549	8 min->16 min	721
	2 min	1436	32 min	211	1 min->2 min	849	2 min->4 min	639
	32 min	872	2 min	92	8 min->16 min	766	16 min->32 min	621
	16 min	750	0 min	91	4 min->8 min	732	0 min->1 min	280
	1 min	590	4 min	37	16 min->32 min	617	4 min->8 min	182
	4 min	288			2 min->4 min	426	~~~	178

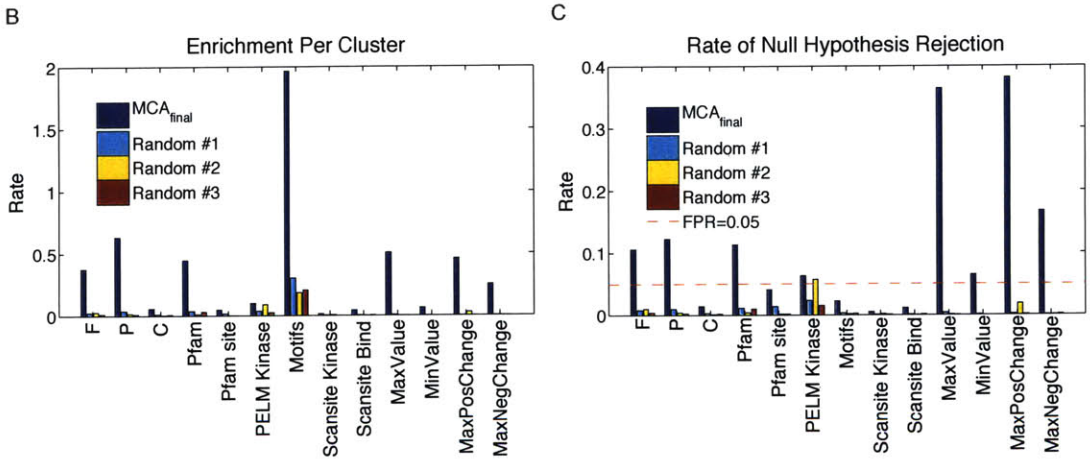


Figure 4-6: Enrichment results in MCA_{final} for the *EGF7* dataset. A) Table of the top six enrichment labels in all metrics and the number of times they are enriched in the full MCA_{final}, which includes 490 sets. B) Average enrichment per cluster in MCA_{final} versus random controls. Enrichment for each metric was aggregated and normalized by the total number of clusters. C) Rates of rejection of the null hypothesis test. In random controls, this ratio indicates the False Positive Rate, which was controlled using FDR with alpha=0.05.

tion of co-occurrences for real and random data are roughly the same, indicating that all implementations of unsupervised learning will produce robust co-clustering of randomized data as well as real data. Figure 4-7D shows the number of co-occurrences at a given threshold is roughly the same, however, the number of peptide pairs that occur at values greater than 50 (or 10% of the time) drop off faster in random data than in clustering of real data. Importantly, the randomized data control performs no better at classifying known protein interactions than expected by random, Figure 4-7B and C.

An important difference between hypotheses derived from robust co-clustering of phosphopeptide information and protein interaction information is that the resolution of the interaction, i.e. site-specific interactions in the case of clustered data. To determine whether predictions via this method make biological sense, we evaluated the layer of the network for which the most resolution is known, that of the EGF receptor itself. Figure 4-8 shows the table of all peptide measurements occurring on the EGF receptor in the *EGF7* dataset and the ten phosphopeptides that most robustly co-cluster with each receptor peptide. EGFR Y1172 and Y1197 are known Shc (SHC1) docking sites. EGFR Y1172 has been shown to recruit the PTB domain of Shc wherease 1197 the SH2 domain [8]. In turn, Shc can recruit Grb2 [8], thereby recruiting GAB1, to the receptor layer. We see this functionality in the co-clustering rankings; two phosphopeptides on Shc most closely co-cluster with Y1172 and also highly co-cluster with Y1197, although to a lesser extent. The top ranking co-clustered site of Y1197 is on a protein called PDLIM1/CLIM1, a cytoskeletal adaptor protein. Another site with known functionality on EGFR is Y1069, which is responsible for recruiting the E3 ubiquitin ligase Cbl. EGFR Y1069 co-clusters more often with Y552 on Cbl than any other site in the dataset. The dynamic plots demonstrate that there is a large similarity between the upregulation of phosphorylation on sites Y1172, Y1197 and Y1069 and the marked difference between them is the relative sustainment of Y1172 and Y1197, where as Y1069 undergoes a rapid and more complete downregulation within the time points measured, Figure 4-8A. This difference between EGFR site regulations is visible by eye and by inspection of the

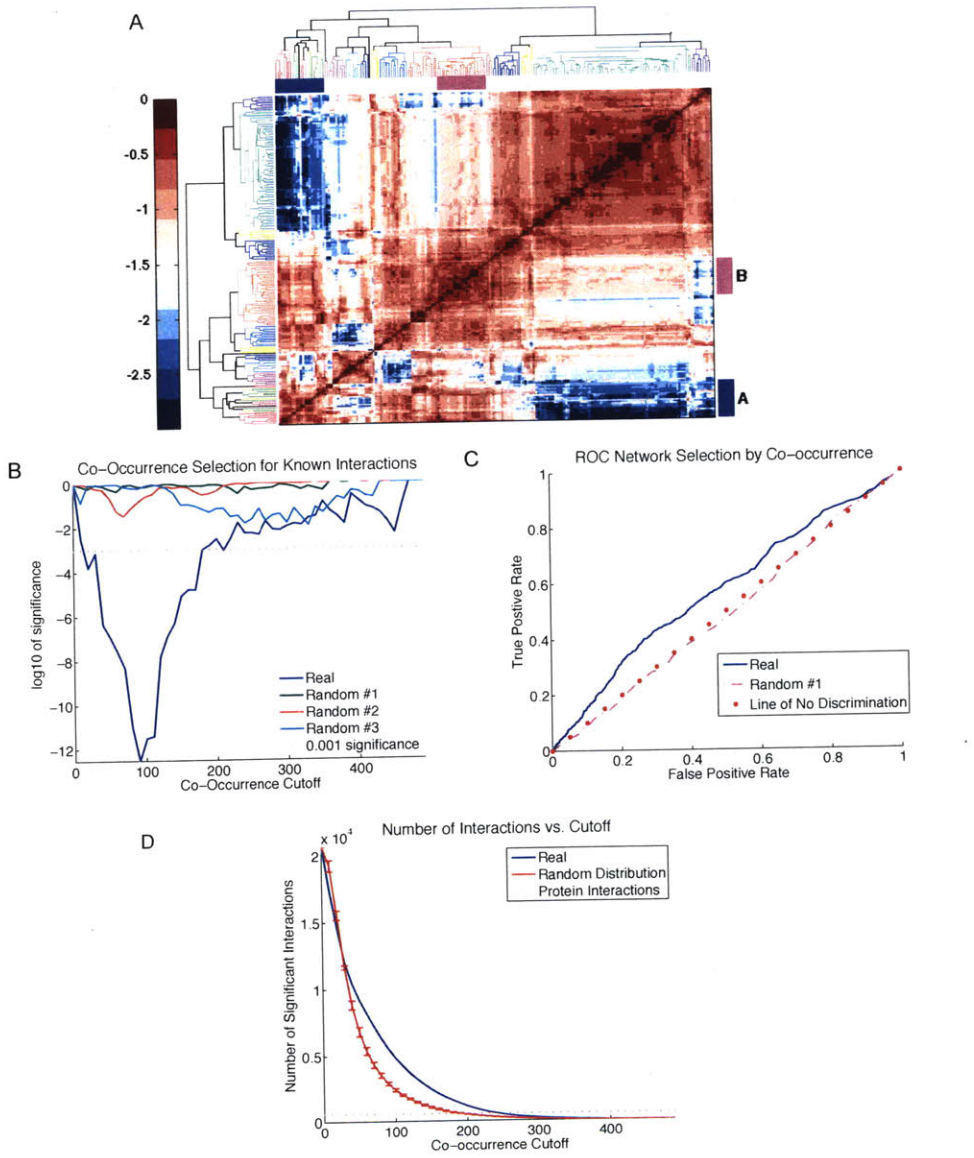


Figure 4-7: Robust co-clustering as a method of network inference. A) A hierarchically clustered heatmap of the MCA$_{final}$ co-occurrence matrix. The log transform of the matrix, normalized by the total number of sets in the MCA$_{final}$ was used (490 sets). Label A indicates a group of phosphopeptides that occur a large number of times with most other phosphopeptides. Label B indicates a second group type of phosphpeptides that co-occur with only a subset of phosphopeptides. B) The significance of selection for known network interactions when selecting by a binary co-occurrence matrix, thresholded by the number of co-occurrences indicated by 'cutoff'. C) The corresponding receiver operator curve for selection of network interactions by a thresholded co-occurrence matrix. D) The number of significant interactions given a co-occurrence cutoff and the number of annotated protein interactions.

terms often enriched along with Y1069. Roughly 25% of the time, Y1069 appears in a cluster whose enrichment includes maximum negative change in phosphorylation between two and four minutes, a distinct difference between Y1069 and the other phosphorylation sites on EGFR.

There are two instances of doubly phosphorylated forms of a peptide on the receptor, S991/Y998 and S1166/Y1172. From observations of the rank ordered lists of co-clustered phosphopeptides, it appears both of these forms participate in distinctly different groups compared to their singly tyrosine phosphorylated forms, Figure 4-8A. It appears that Y998 moves from co-clustering with phosphorylations on proteins involved in cytoskeletal and phospholipid signaling, as indicated by phosphorylations on tensin-3 (TNS3), catenin delta-1 (CTNND1), PLCγ-1 (PLCG1), and SHIP-2/INPPL, to correlations with late-stage EGFR tyrosine signaling as indicated by transcription factor phosphorylation, STAT5B/STAT5 (the peptide could be from either or both STAT gene products), as well as MAPK and PI3K signaling. The second ranked phosphorylation association with S991/Y998 is the negative regulation site on a Src family kinase, PTK6/Brk Y447. The dynamics of this group of phosphorylations are marked by delayed onset and sustained through the remainder of the time course. EGFR phospho-Y1172 moves from interactions with SHC1 to syntaxin-4 (STX4), a known SNARE, a translation initiation factor, EIF4B, and ITSN2, which is thought to be a link between endocytic membrane traffic and actin assembly, when doubly phosphorylated on S1166 and Y1172.

Given that EGFR Y1069 associates very robustly with only one site of three sites measured on Cbl, we evaluated the top rankings for all Cbl sites, Figure 4-9. Although Cbl Y552 is the highest-ranking site for Y1069 on EGFR based on co-clustering, the converse is not true. However, EGFR Y1069 does rank in the ten co-clustering phosphopeptides with Cbl Y552, but a variety of other sites rank higher than that of EGFR Y1069, including AHNAK a highly phosphorylated protein known as neuroblast differentiation-associated protein, but with no apparent role known in mature epithelial cells and the EGFR network. The remaining two sites on Cbl share many top ranking sites, including phosphorylations on INPPL/SHIP-2, PLCG1,

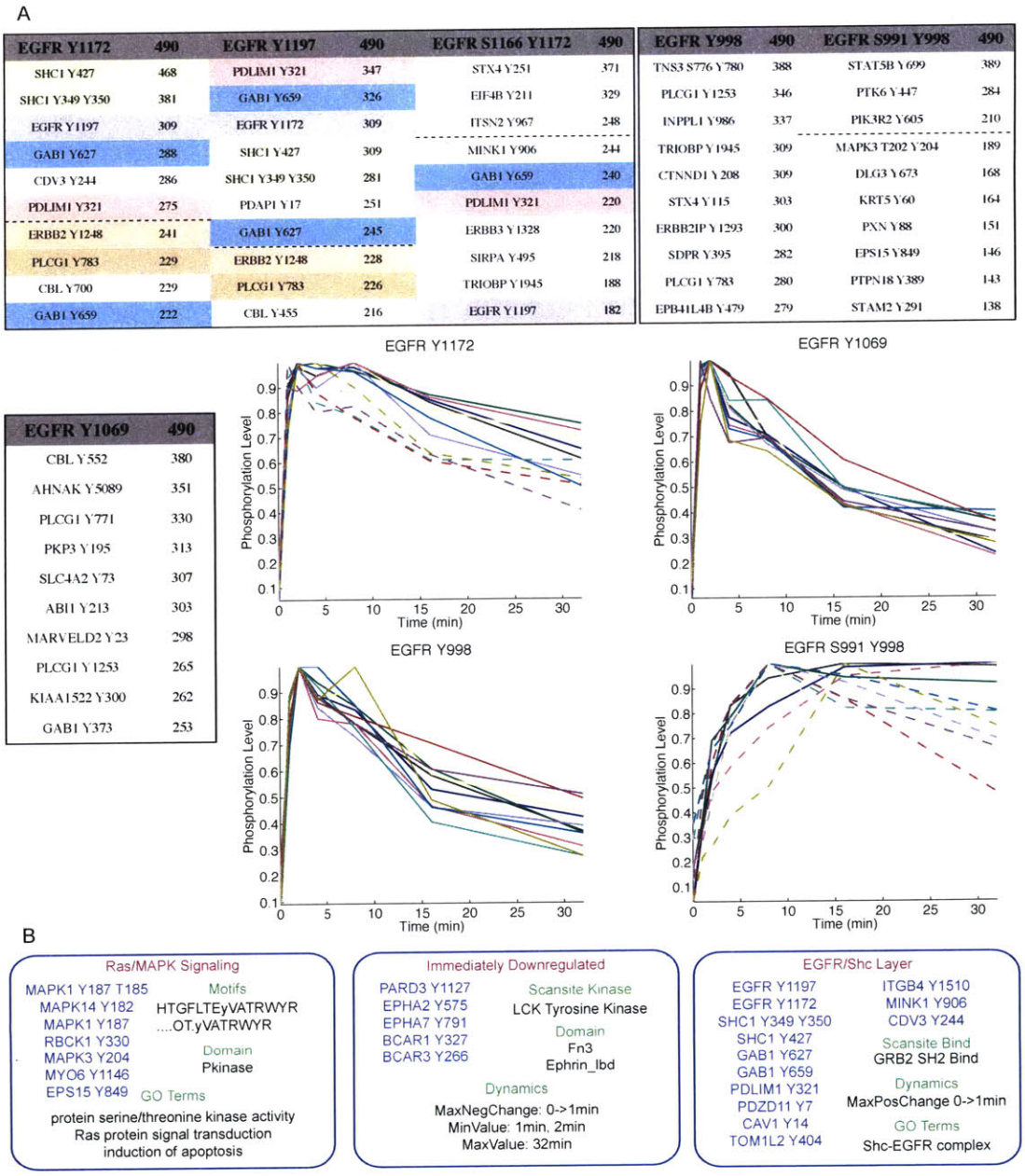


Figure 4-8: Robust co-clustering recapitulates known EGFR interactions and can generate supergroups of partitioned phopshopeptides. A) The top ten sites that co-cluster, and the number of times they co-cluster, is listed for each phosphеptide measured on the EGF receptor. The total possible number of co-occurrences is 490, the total number of sets in MCA_{final} and values falling below half of this are separated by a dashed line. The dynamic plots for example groups with EGFR peptides, which are normalized by their maximum phosphorylation level. Dynamics representing co-occurrences below 50% are shown with dashed lines. B) Three example supergroups derived with a co-occurrence cutoff of 120 and their corresponding enrichment labels.

and LSR, i.e. lipolysis-stimulated lipoprotein receptor, all proteins involved in lipid signaling. Cbl Y552 also associates strongly with a site on PLC-γ-1, however its site, Y771, is distinct from the sites strongly associated with Cbl Y455 and Y700, which include PLCγ-1 sites Y783 and Y1253.

The top ranking co-occurrences for the two measured phosphopeptides on Shc are seen in Figure 4-9B, which are as similar in overlap with each other as Cbl Y455 and Y700 and EGFR Y1172 and Y1197 are. Given the high similarity, but also variability, between lists of sites, we sought a method in which to find concurrent robustness of multiple phosphopeptides based on a particular co-occurrence cutoff value. This method seeks to define supergroups of phosphorylation sites. In order to visualize the resulting supergroups, we used the Systems Biology Markup Language (SBML) in order to depict group memberships and their interconnectivity. Examples of supergroups derived at four different thresholds are available at the end of the chapter, Figures 4-12 through 4-16. As the cutoff value of co-occurrence is varied the total number of groups, the average group size, and the connectivity of groups via shared phosphopeptides varies, Figure 4-10. At the extremes, with a cutoff of zero co-occurrences there is one supergroup with number of members equal to the total number of phosphopeptides (204 in the case of the *EGF7* dataset) and when the cutoff is equal to the total number of clustersets in the MCA, M, there are typically M groups with one member each. In no iteration of the MCA generation did we observe two phosphopeptides that co-clustered M times.

Setting a cutoff value for the number of co-occurrences can be based on a variety of methods. One method is to base it on the expected value of co-occurrences in a randomized trial of reshuffling set members in the MCA. For the set architecture in MCA_{final}, this has an expected value of roughly 66 co-occurrences and standard deviation of 7, Figure 4-11. A cutoff of 80 co-occurrences translates to the number of co-occurrences that would rarely be seen by random, being two standard deviations above the random mean. The choice of cutoff could instead be based on maximizing overlap with known protein interaction information (Figure 4-7B) or the tradeoff between the average size of groups and the extent of group interconnectivity, Figure

A

CBL Y552	490	CBL Y700	490	CBL Y455	490
AHNAK Y5089	441	INPPL1 Y1135	347	PLCG1 Y783	397
PLCG1 Y771	405	ERBB2 Y1248	341	PLCG1 Y1253	296
PKP3 Y195	393	ARHGEF5 Y1100	300	INPPL1 Y1135	290
ABI1 Y213	385	PLCG1 Y783	289	CBL Y700	283
EGFR Y1069	380	CBL Y455	283	FLOT2 Y241	281
SLC4A2 Y73	363	KIAA1522 Y300	231	TNS3 S776 Y780	274
KIAA1522 Y300	322	LSR Y551	231	MARVELD2 Y23	271
MARVELD2 Y23	318	EGFR Y1172	229	INPPL1 Y986	263
CTNND1 Y174	292	SHC1 Y427	222	EGFR Y998	262
GAB1 Y373	286	PLCG1 Y1253	209	SDPR Y395	259
				ERBB2 Y1248	259
				LSR Y586	259

B

SHC1 Y427	490	SHC1 Y349 Y350	490
EGFR Y1172	468	SHC1 Y427	402
SHC1 Y349 Y350	402	EGFR Y1172	381
EGFR Y1197	309	GAB1 Y627	344
GAB1 Y627	294	CDV3 Y244	317
CDV3 Y244	281	EGFR Y1197	281
PDLIM1 Y321	276	PDLIM1 Y321	243
PLCG1 Y783	239	ERBB2 Y1248	218
ERBB2 Y1248	239	PTK2B Y580	213
GAB1 Y659	223	GAB1 Y659	204
CBL Y700	222	BCAR1 Y267	196

Figure 4-9: Top rankings for multiply phosphorylated docking proteins of EGFR. A) Top ten ranking site co-occurrences with Cbl phosphorylation sites. B) Top ten ranking site co-occurrences with Shc phosphorylation sites.

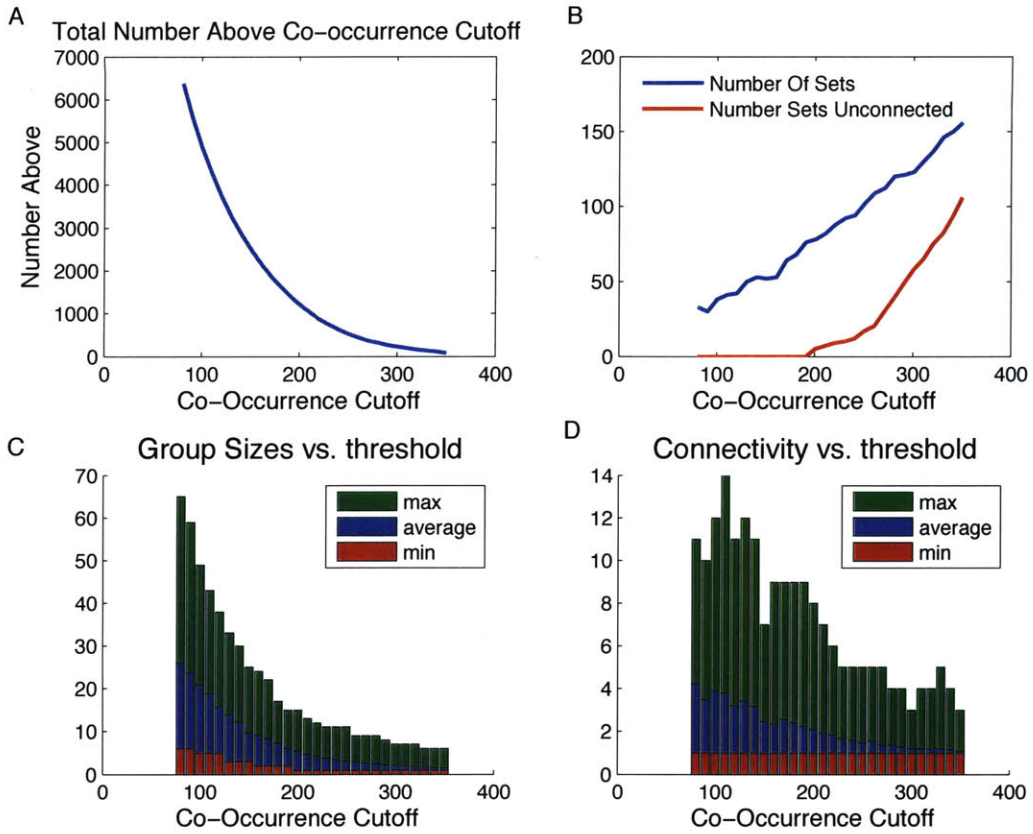


Figure 4-10: Group network statistics based on the co-occurrence cutoff. The minimum co-occurrence chosen for consideration is the value of the random expectation plus the standard deviation of the random distribution. A) The number of interactions considered significant versus cutoff. There are 20,706 total possible interactions for the 204 phosphopeptides of the *EGF7* dataset. B) The total number of groups in a network structure at a cutoff and the corresponding number of groups that are not connected to any other group through a joint phosphopeptide. C) The maximum, average, and minimum group size for every cutoff. D) The maximum, average, and minimum number of connections any one phosphopeptide has versus cutoff. For example, using a cutoff of 80, the average peptide belongs to four supergroups. As the cutoff increases phosphopeptides progressively belong to fewer supergroups.

4-10. We have included images of the SBML generated network figures of supergroups for co-occurrence cutoffs of 80, 120, 200, 240 and 300 Figures 4-12 through 4-16. We found that a cutoff value of 120 yielded manageable group sizes and reasonable interpretability. This cutoff value overlaps significantly with known network

information and is significant when compared to random expectation. Examples of three supergroups resulting from this cutoff are shown in Figure 4-8B. In the same manner that we evaluated biological enrichment for clusters in an MCA, we evaluated enrichment in each supergroup and overlaid this information with the supergroup members. What results is a robust grouping of phosphopeptides based on dynamics with informative partition labels. One example group represents components of Ras/MAPK signaling, whose labels overlap with what we saw repetitively throughout most sets of MCA_{final}. Second is a set unique in its regulation within this dataset, immediate downregulation following stimulation by EGF. Components of this group include sites on two proteins known to be involved in focal adhesion signaling and cooperativity with EGF response, BCAR1 and BCAR3 [95]. This cluster additionally contains two peptides from EPHA receptors, EPHA2 Y575 and EPHA7 Y791, the activating phosphorylation event on EPHA7. Finally, Figure 4-8B includes a supergroup representing the full EGFR-Shc layer components, which includes members from the individually ranked lists of SHC1, EGFR Y1172, and EGFR Y1197. This supergroup is enriched for predictions of Grb2 SH2 binding sites and for immediate upregulation, in the first minute, following EGF addition.

4.4 Methods

4.4.1 Dataset preparation and biological term annotation

The *EGF7* dataset was downloaded from the supplementary information of PNAS. There were originally 222 peptide measurements; some of these phosphopeptides represented the redundant measurement of an identical phosphorylation site but with a trypsin miscleavage. In most cases, the miscleaved form of the peptide often clusters with the perfectly cleaved form and skewed the results of enrichment analysis. Therefore, we systematically removed the miscleaved form of all redundant peptides reducing the dataset to 204 unique phosphopeptide measurements on 141 proteins representing 215 total phosphorylations. The dataset was then loaded into PTM-

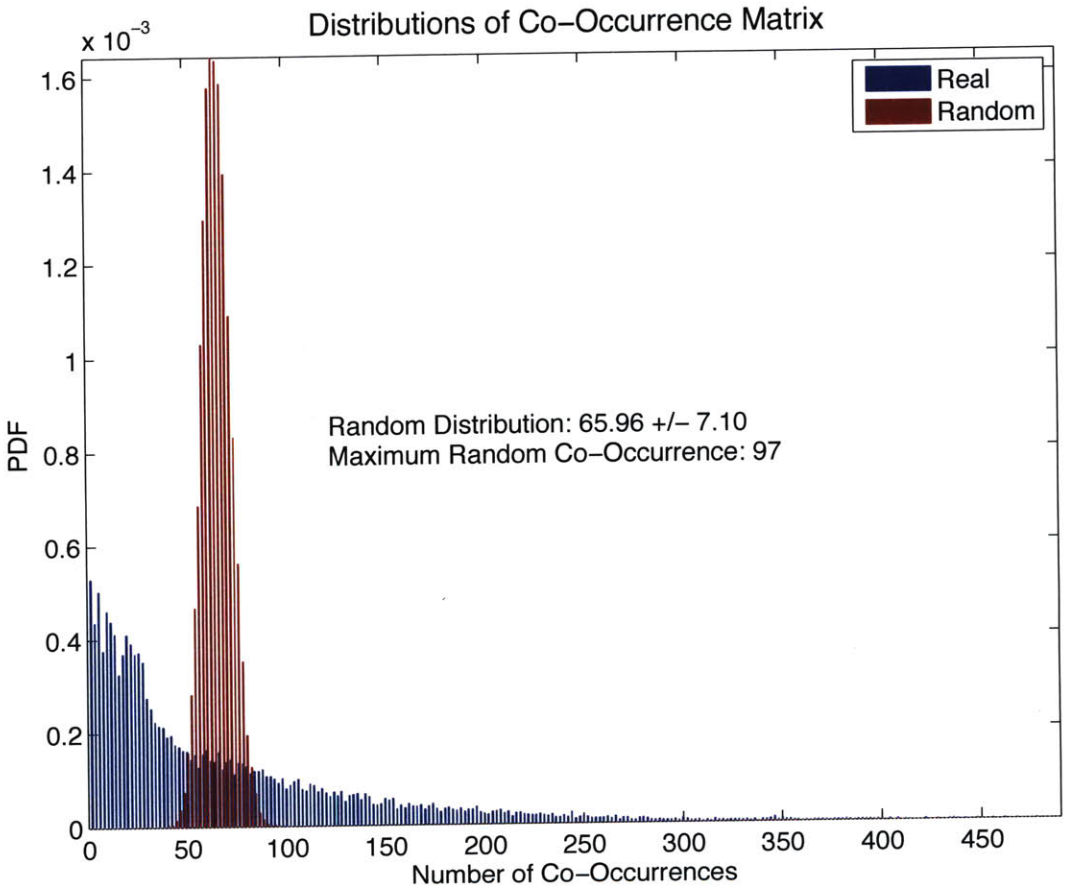


Figure 4-11: The probability distribution of the co-occurrence matrix. The upper triangular portion of MCA_{final}, not including the diagonal. A random expected distribution is created by randomizing the cluster assignments in the set architecture of MCA_{final}. By random assignment in the given architecture, we have an expectation of 66 co-occurrences. In this random test, a maximum of 97 co-occurrences was observed.

Scout, Chapter 3, and default selections in the ambiguity tool were used to assign peptides to proteins with the largest number of annotations, with the exception of the Src family kinase activation loop peptide, which was manually assigned to the most well annotated Fyn isoform, where the default assignment in PTMScout and the original dataset assignment is an isoform of Lck. In order to append Gene Ontology terms, domains, Scansite predictions, we imported and exported the dataset through PTMScout, using a new database slice export feature. For clustering, a flat form of the dataset was exported from PTMScouts export for clustering feature.

4.4.2 Clustering

The flat text file of the dataset was imported into Matlab based on DataRail object structures [90]. Transforms, distance metrics, and algorithms are from the Matlab environment and its toolboxes, downloaded from other resources, or developed for our purposes. Ncut code for Matlab was obtained from http://www.cis.upenn.edu/~jshi/software/ based on the algorithm description in [99], affinity propagation (AP) clustering code was downloaded from http://www.psi.toronto.edu/affinitypropagation/ based on the algorithm described in [30]. A self organizing map (SOM) Matlab toolbox was downloaded from http://www.cis.hut.fi/somtoolbox/ and is based on the algorithm described in [51]. Affinity propagation clustering was modified to accept an arbitrary distance metric, but does not accept an argument for K. SOMs only utilize the Euclidean distance metric. Average linkage distance is used in hierarchical clustering. Kmeans uses the squared value of the Euclidean distance and does not accept the Chebychev distance metric. An additional SOM parameter is used to indicate the length direction in a rectangular grid pattern; if the division of K into two dimensions is non-square, two solutions are produced, one solution where the width dimension is largest and one with the height dimension set to the largest value. The largest value of K is bounded by a number that would produce roughly 5 phosphopeptides per cluster, assuming a solution were to equally distribute all phosphopeptides, which in this case is Kmax=40. For non-deterministic algorithms, such as SOMs and Kmeans, we store the random seed so that we could exactly reproduce the result, but allow the

random seed to vary between individual implementations so as to ensure we do not force all implementations of the algorithm into a poorly performing local minima.

4.4.3 Enrichment, multiple hypothesis correction and parameter refinement

Enrichment calculations are performed using the database slice and was programmed in Perl and MySQL. Enrichment is calculated as in PTMScouts subset selection enrichment analysis, using a hypergeometric distribution calculation, Chapter 3. The motif algorithm (Chapter 2 and [46]) is set using a branch cutoff of 0.01 and search space of +/-7 amino acids surrounding the site of phosphorylation. Scansite prediction levels of three and better are considered, based on an empirical analysis of the tradeoff between the false positive rate and total hypothesis rejection. When domains were calculated *de novo* in PTMScout, predictions of 1e-5 and better are used. False discovery rate correction (FDR) was performed at the metric and set levels as following: p-value calculations were accumulated for all tests within a category and the p-value satisfying an FDR alpha value of 0.05 was used to determine final enrichment for that metric.

Relationships between parameters and metrics are calculated by rank-ordering the sets within each of the thirteen categories according to the total number of enriched labels within a set. Each parameter within the four categories is then tested for overrepresentation in the best and worst performing 25% of the rank orderings. Overrepresentation is calculated according to the hypergeometric distribution and considered significant when below a Bonferroni corrected p-value cutoff of 0.05. To determine parameter candidates for removal in future iterations, the average power of enrichment for a parameter is calculated, i.e. the number of labels enriched per set, before and after removal. Those removals that significantly improve all categories are dropped from future iterations of the process. Those parameters that improve most categories and only slightly decrease a small number of categories are also removed. When removal of candidate parameters begins to have only a slight impact on en-

richment per cluster, around a 1% change or less, the parameter refinement loop is exited with a final MCA. The randomized control process subjected a randomized version of the data matrix to the same MCA final parameters. This random MCA was evaluated in exactly the same way as the real MCA.

4.4.4 Mutual information calculation and selection

Mutual information was calculated as in the MIST algorithm for compensation for under sampled distribution spaces [50]. Selection order for maximum variability is performed as follows: the total joint entropy for all sets in the MCA is calculated before and after the removal of a test set from the MCA. The set that affects the joint entropy, or normalized joint entropy, the least is removed. This process is repeated until all sets in the MCA have been ordered according to this procedure. The resulting order represents those sets with the most mutual information with the largest number of other sets. This order is then taken in reverse, which instead indicates the sets with the lowest total mutual information. We performed the selection operation in this manner, requiring reversal, in order to prevent biases for selection of sets with larger cluster numbers. In order to calculate redundant sets within an MCA, the standard mutual information value between all sets is calculated and those sets with both identical self-MI and pairwise-MI are considered to be redundant.

4.4.5 Co-Occurrence calculations and network analysis

The co-occurrence matrix is calculated by adding up all pairwise incidences of two phosphopeptides in a cluster, within all sets. Network information is obtained from the MetaCore software suite (www.genego.com). We created a binary matrix from the network interaction information from MetaCore in the same order as the co-occurrence matrix, assigning all interactions between two proteins a value of 1 for every phosphopeptide measured arising from that protein. For a given co-occurrence cutoff, values in the co-occurrence matrix are set to 1 if above and 0 if below the cutoff. Taking only the upper triangular portion of the two matrices and ignor-

ing the diagonal, since co-occurrence values of all self-phosphopeptide interactions is meaningless via the co-occurrence matrix metric, we calculate the significance of co-occurrence values above the cutoff overlapping with the binary values of the network matrix using a hypergeometric distribution calculation, as in the methodology of label enrichment. The same information is used to build the receiver operator characteristics. SBML (www.sbml.org), which can be viewed using freeware packages such as Cell Designer (www.celldesigner.org), is generated in order to visualize the co-occurrence network at a given a threshold or cutoff. Groupings are made based on finding all phosphopeptides that co-occur at the defined cutoff and above and are placed in a compartment. Those phosphopeptides participating in multiple groups are then copied into the global compartment and reaction arrows are drawn between this central phosphopeptide and all of its counterparts within compartments.

4.5 Conclusions

One function of this semi-automatic framework is to provide a high-throughput method for unsupervised learning parameter screening, a way in which to understand how the parameters of clustering influence the resulting biological relationships, in the context of known annotations. The result of these parameter screens is that there are a variety of possible and informative clustering solutions, none of which are maximally informative on their own. This implies that in seeking representations of the data in a reduced dimensionality through clustering, the desired types of biological information will influence the application of unsupervised learning or that a variety of solutions should be taken into consideration. Alternatively, the combination of these variable results may indicate those biological stories that are robust to perturbations of the relationship between the quantitative measurements. When we looked at these robust biological enrichment terms, there were repetitive formations of particular components of the network, for example components of the Ras/MAPK signaling pathway. This robustness indicates that relative to the rest of the dataset, these dynamics are uniquely similar and our confidence in those components implicated for the first time

with such a process increases with the robustness of their relationships with such a cluster.

Phosphorylation changes in the EGFR network, proximal to the receptor, occur very quickly following EGF addition. Roughly 42% of the measured phosphopeptides reach a maximum value within two minutes. Even the most downstream events, such as MAPK and STAT5 phosphorylation, reach a maximum value within eight minutes following stimulation. These highly similar trajectories are highlighted in the apparently poor separability of the dataset in the first three principal components, as shown in Figure 1. Despite these apparent limitations, we are encouraged that unsupervised learning is able to discern subtle differences in the regulation of phosphorylation based on its ability to produce groups with related function. This ability is highlighted by the results of MCA analysis of randomized data; this process does not produce meaningful biological relationships among clusters, unlike clustering of real data. These random relationships, or false positives, are controlled by the FDR procedure. In fact our multiple hypothesis correction empirically controls false positives at a rate that is roughly 10-fold more stringent than the level we set during testing, while leaving a large number of biological and dynamic features intact for the real data. Although we may then be able to loosen our significance requirement, or use an alternate hypothesis correction technique with more power, such as the pFDR [105], the level and method of correction used here is sufficient to yield a multitude of enrichment labels with high confidence.

The results and interpretations of this framework depend entirely on known annotations and sources of biological information, which is a major limitation. Parameter performance as judged by these metadata terms will reflect the natural hierarchy and structure of the metadata terms and not necessarily reflect the inherent structure of the underlying data. Future inclusion of sources of information would require reevaluation of parameter performance. Despite limitations of current biological metrics used and non-inclusion of other possible resources, the use of such information is a vast improvement to the alternative of manual human evaluation, which is not only biased and limited, but impossible to perform in such a high-throughput manner as

to allow for evaluation of a set of possible algorithms and their various parameters. Perhaps the most important advantage of this semi-automatic framework is that scientists who are not experts in the field of data mining can empirically evaluate the application of a variety of algorithms or mathematical data transformations within the context of their biological problem and measurements. This separates scientists from having to rely on the erroneous assumption that what has worked well in an unrelated dataset would now work well in their current evaluation. In our evaluations of a variety of phosphoproteomic datasets we find the best performing parameters vary depending on the dataset evaluated.

An additional utility of this framework is the generation of robust phosphopeptide-phosphopeptide relationships, which not only recapitulates known protein network information, but also appears to indicate network interactions occurring specifically at the resolution of the modifications themselves. Using the combination of all sets in an MCA, we are able to differentiate the known functions of the EGFR tyrosine sites 1172, 1197, and 1069. Their robust co-clustering with each other as well as similarity of top ranking partners highlights the shared functionality of Y1172 and Y1197 for recruiting Shc and subsequently Grb2. The higher ranking of GAB1 phosphorylation with Y1197 may indicate a tighter relationship between GAB1 recruitment to the receptor through Y1197, or perhaps a higher conformational availability of Y659 on GAB1 through Y1197 recruitment versus Y627. Additionally, both Y1172 and Y1197 co-cluster with ErbB2/HER2 Y1248. The sequence surrounding Y1248 on ErbB2 is highly similar to EGFR Y1197 [96] and has been shown to be capable of binding SHC by two different *in vitro* studies [45,96]. The *in vitro* result, the sequence information, and the network predictions by robust co-clustering suggest a strong possibility that ErbB2 Y1248 recruits SHC *in vivo*, in the same manner as EGFR Y1197.

strengthens the predicted *in vivo* binding of ErbB2 Y1248 and SHC.

EGFR Y998 phosphorylation is strongly correlated with phosphorylations on proteins involved in phospholipid signaling. However, when S991 and Y998 are phosphorylated at the same time, we see a shift in the top-ranked co-occurring sites from phospholipid proteins to STAT5 Y699 phosphorylation. A study by Schulze et.

al [96] showed that a phosphorylated bait peptide sequence surrounding Y998 is able to bind STAT5 *in vitro*. Phosphorylation of STAT in this work appears to be most similarly correlated with concurrent phosphorylation of Y998 and S991, indicating that if phospho-Y998 is responsible for recruitment of STAT, subsequent phosphorylation is enhanced by the presence of phospho-S991. Serine-991 phosphorylation may play a role in enhancing STAT5 binding by conformational changes in the receptor or direct binding. Alternatively phospho-S991 may recruit effecter enzymes responsible for STAT Y699 phosphorylation, or S991 phosphorylation may simply be a marker for a change in localization, for example endocytic vesicular localization, of the receptor and therefore a change in signaling. Without further experimental testing it is not possible to know for sure which mechanism may be at work for the dynamic relationship between S991/Y998 and STAT5 Y699 phosphorylation. However, this method might play an important role in the first steps of highlighting functional relationships between phosphorylation sites in RTK networks generating testable hypotheses that can then inform a site-specific network model of RTK networks. This method may be most useful in application to protein modification by phosphorylation since a primary function of phosphorylation is to provide a highly dynamic and controllable mechanism for protein binding and recruitment. As proteins are recruited into complexes with the enzymes that will phosphorylate and dephosphorylate particular residues, the shared regulation and function is observed through tightly correlated dynamic changes. Future work will explore the utility of network generation via this method using non-dynamic measurements of the network and modifications of other sorts, such as ubiquitination, acetylation, or glycosylation.

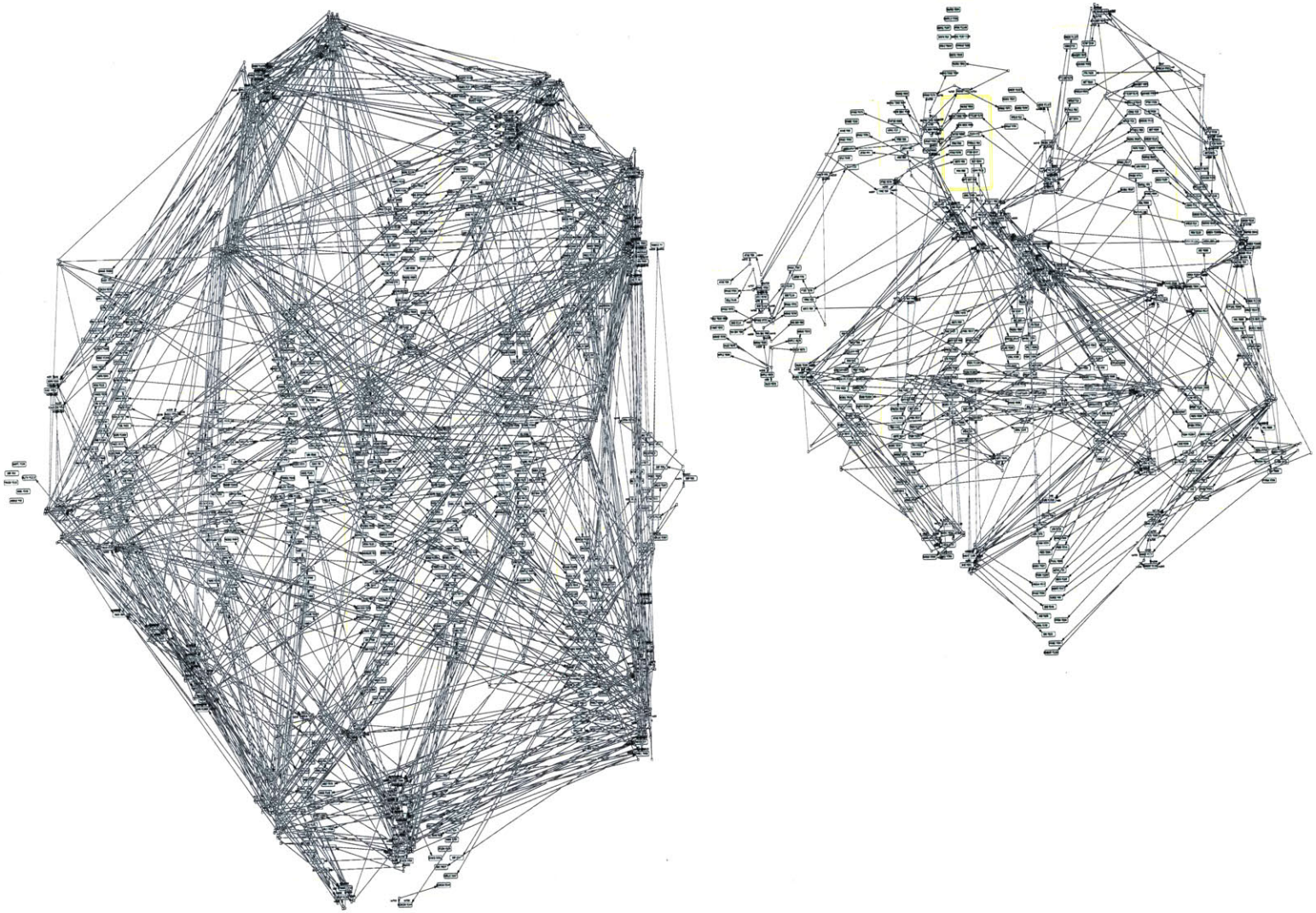


Figure 4-12: Supergroup architecture based on a co-occurrence cutoff of 80.

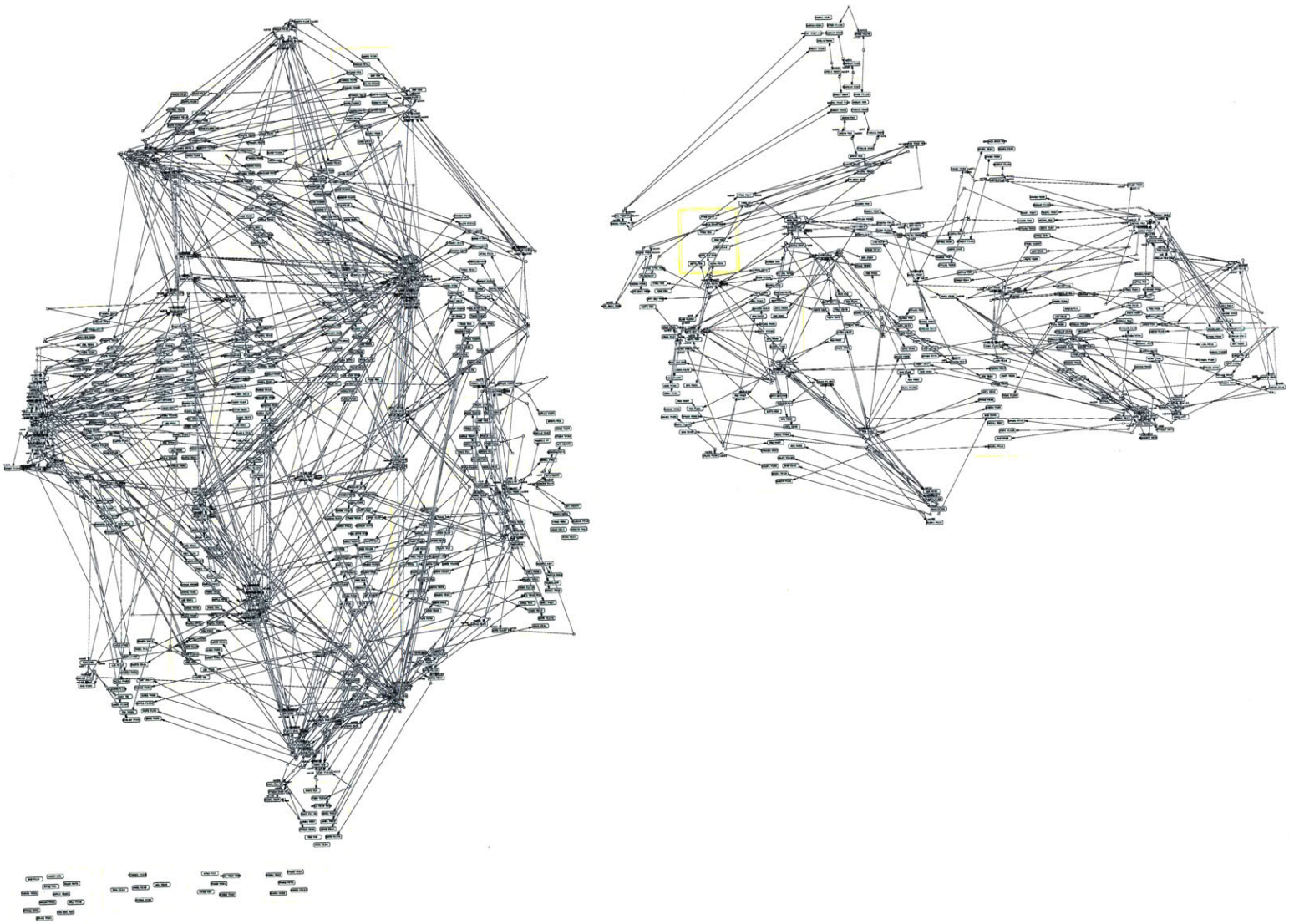


Figure 4-13: Supergroup architecture based on a co-occurrence cutoff of 120.

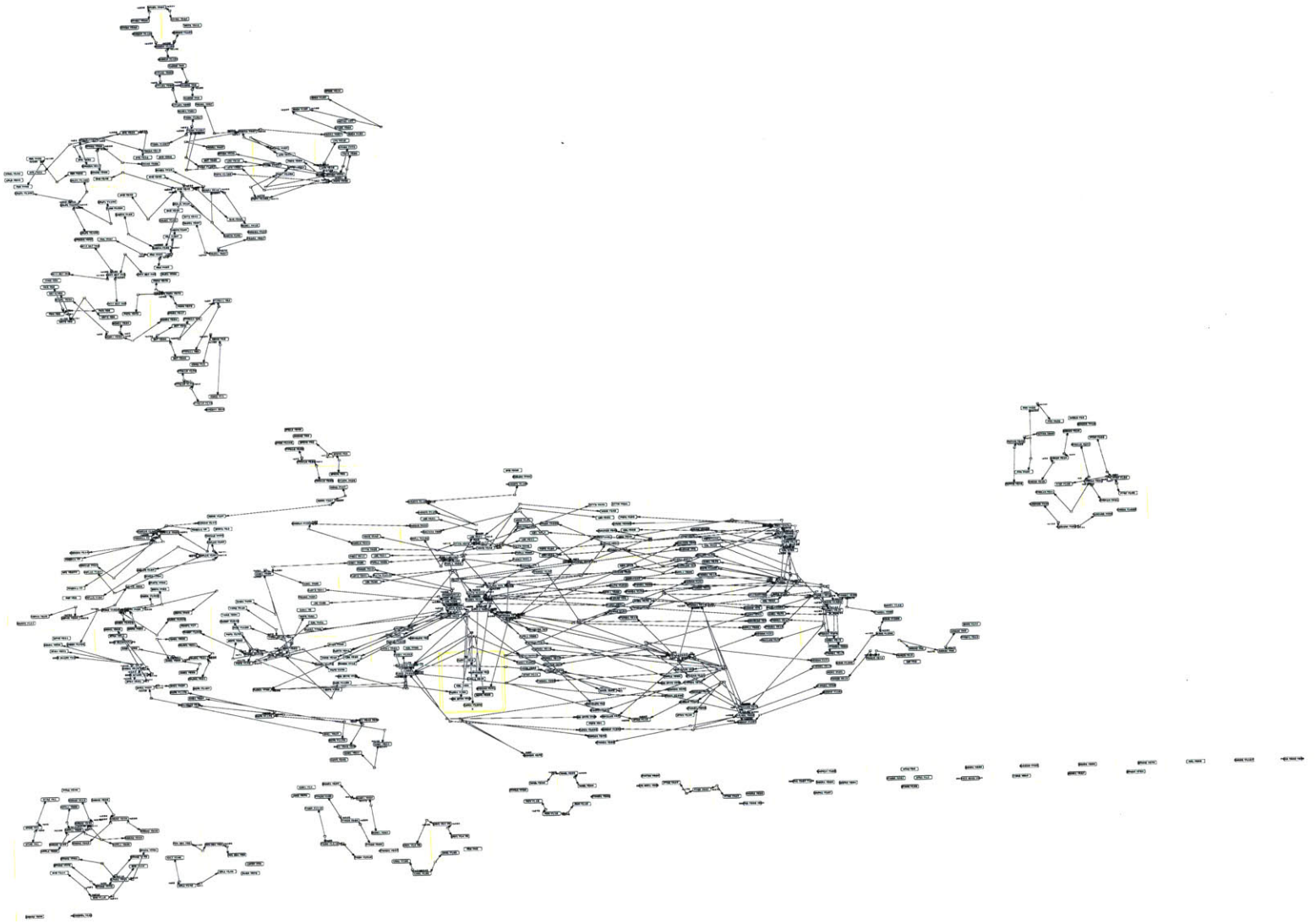


Figure 4-14: Supergroup architecture based on a co-occurrence cutoff of 200.

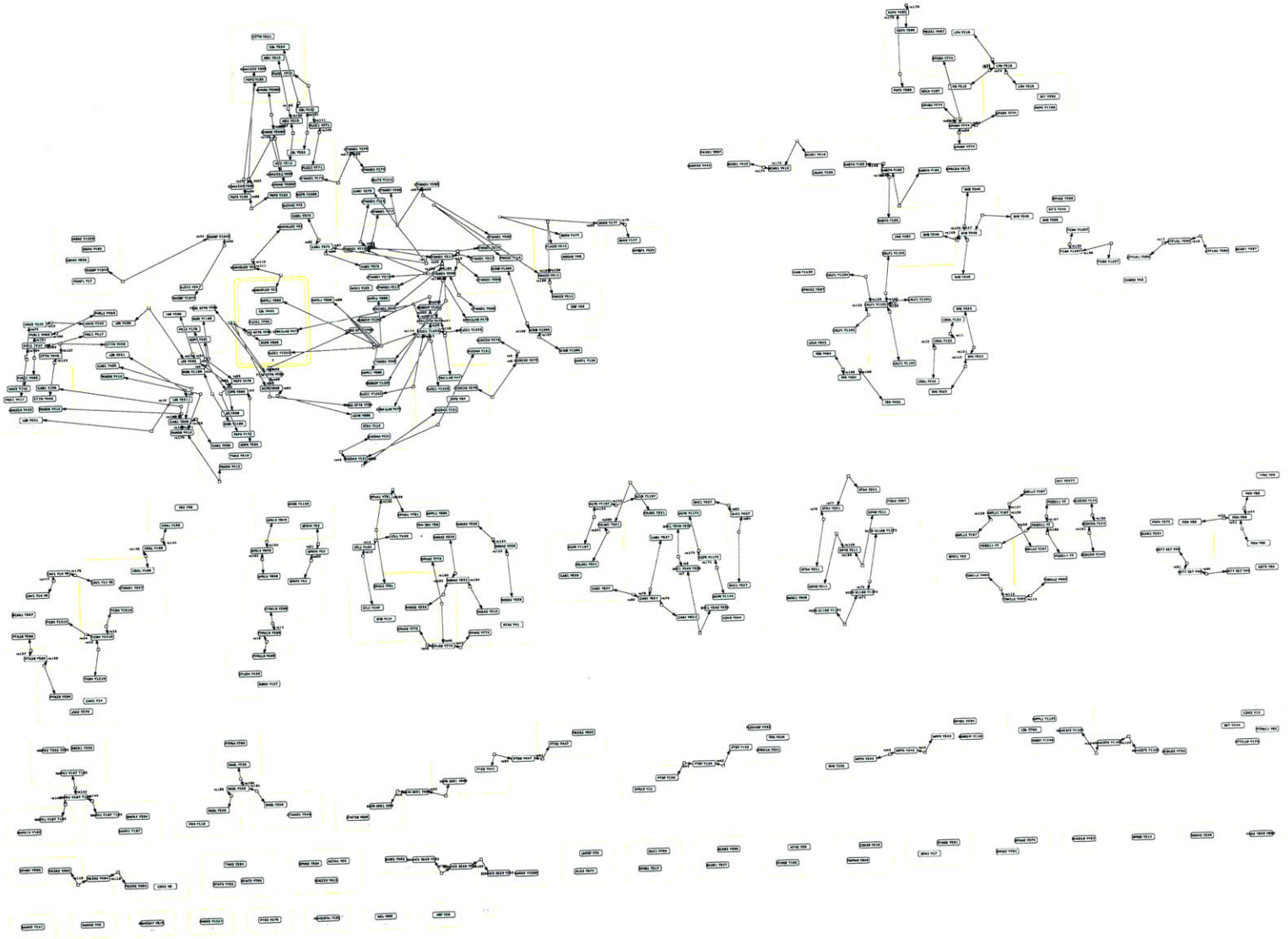


Figure 4-15: Supergroup architecture based on a co-occurrence cutoff of 240.

Figure 4-16: Supergroup architecture based on a co-occurrence cutoff of 300.

Chapter 5

Concluding Remarks and Future Directions

5.1 Experimental support for derived hypotheses

The methodologies developed in this thesis have sought to generate interesting and pertinent biological relationships, which may help improve our understanding of biological networks. Although we can show that some of the hypotheses derived from these frameworks are supported by literature evidence, many have yet to be explored. This means that the immediate future direction of this work should include seeking experimental support, or disagreement, with the derived hypotheses. In Chapter 2 we hypothesized the motif upregulated in EGFRvIII expressing cells is a consequence of increased CK2 activity. The kinase activity experiments performed using whole cell lysates of the EGFRvIII U87 cells was not specific and therefore could not indicate a whether CK2 activity differed based on EGFRvIII expression, Appendix A.2. Future experiments for CK2 activity testing will require isolation of CK2 from the cell lysate mixtures. Additionally, we may need to evaluate if the differential phosphorylation of possible CK2 substrates in EGFRvIII expressing cells is due to higher nuclear localization of CK2.

Specific mechanisms surrounding how BCAR1 might be a signal integrator of the EGFR and focal adhesion pathways are another interesting hypothesis to pursue.

Does dephosphorylation of BCAR1 Y327, BCAR3 Y266 and PTK2 Y576 indicate movement of these components away from focal adhesions? Does the co-regulated decrease in the activation loop of PTK2/FAK indicate the differential regulation of BCAR1 Y327 and the other five phosphorylation sites measured in the *EGF7* dataset? If BCAR1 does move from the focal adhesions to elsewhere in the cell, is it taking its Crk-binding capacity to a new location? We might begin to understand the answers to these questions by monitoring the location of these components before and after EGF stimulation.

Robust co-clustering analysis of Chapter 4 indicates that this method is useful for predicting protein-protein interactions. Therefore, undertaking to look at whether these newly hypothesized interactions take place *in vivo* will be important. Additionally, determining the dynamics of these protein interactions will be important to understanding the mechanism underlying the interaction. Two studies that looked at *in vitro* recruitment capability of EGFR phosphorylation sites produced a wide range of possible interactions, and in some cases contradictory information [45, 96]. This method might help narrow in on the important question of physiologically relevant interactions.

5.2 The phosphoproteome

In a historical perspective of phosphorylation by Philip Cohen [16], he states the complexity of the regulation and function of phosphorylation as follows "If a third of the 30,000 proteins encoded by the human genome contain covalently bound phosphate, an 'average' protein kinase (on the basis of the probable number of protein kinases) would be expected to phosphorylate about 20 different proteins in vivo, and an 'average' protein phosphatase would be expected to dephosphorylate 60 proteins. ". Since that time the knowledge of the phosphoproteome has greatly expanded and if we now look at the requirement for kinase- and phosphatase-substrate recognition on a phosphorylation site level, versus protein substrate level, we find a drastic increase in the predicted complexity. Figure 5-1 shows the breakdown of the number

of phosphorylations per protein for the 11,832 phosphorylated proteins in the human proteome that currently reside in PTMScout. By taking the arithmetic average of the documented phosphorylation sites for human proteins in PTMScout, we see that there are seven phosphorylation sites for every phosphorylated human protein. If we still assume that 30% of human proteins are phosphorylated then the complexity indicated by Cohen is at least seven-times greater than originally projected; that a kinase would need to phopshorylate roughly 140 **sites** and phosphatases would need to, on average, act on 420 phosphorylation **sites**.

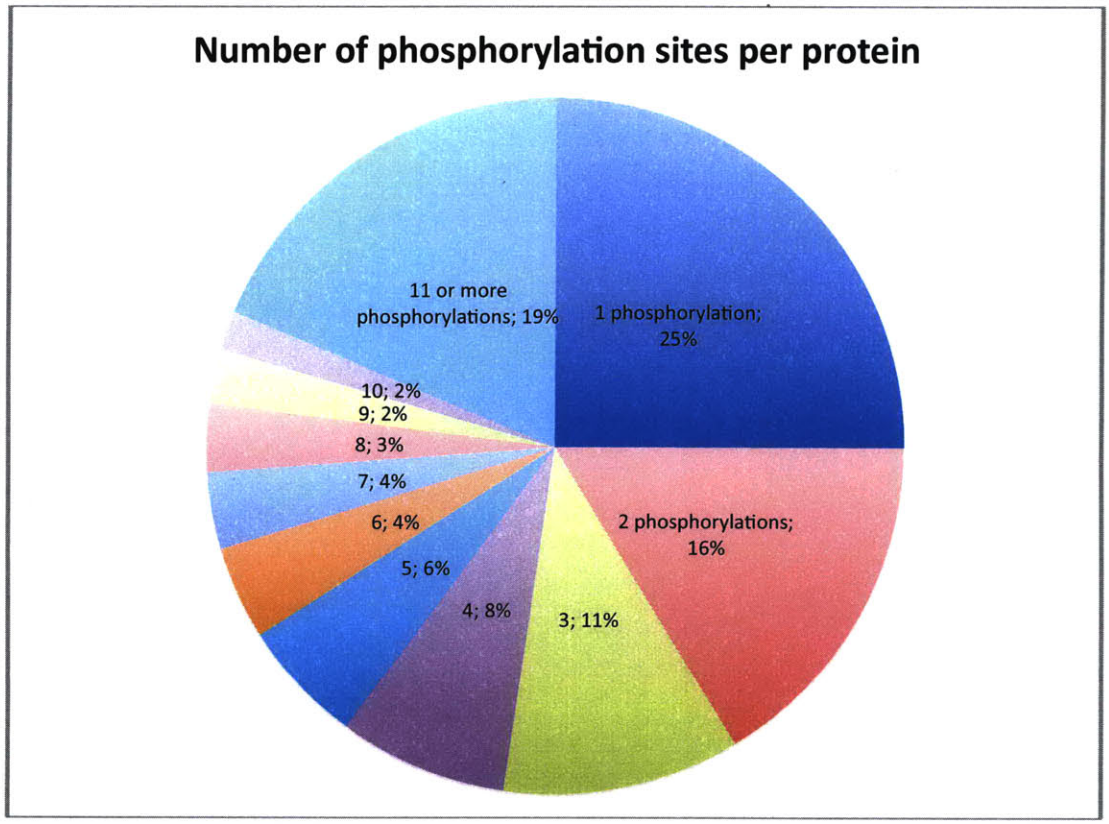


Figure 5-1: Breakdown of the number of phosphorylations (serine, threonine, and tyrosine) documented for human proteins. There were 11,832 phosphorylated human proteins in PTMScout as of April 2010 (PTMScout v1.2).

An important question in the field is that of biological relevance and importance of phosphorylation. Given the abundance of phosphorylation sites, are all of them biologically active and relevant, or are some a consequence of bystander phospho-

rylation, i.e. phosphorylation of non-functional sites due to sequence similarity of functional kinase targets? One major effort in the area has focused on conservation of phosphorylation in order to gain an understanding of a site's biological importance based on evolutionary conservation. However, kinases themselves are highly conserved, so presumably off-target as well as target molecules may also be conserved, independent of their relevance to basic physiological processes. Another method to determine relevance of a phosphorylation site within a measured biological system is to gauge it based on relative fold-change across conditions. Under this method, for example, downstream events of the EGFR network would be considered more likely to be involved in the specific network response if they illustrate a two-fold change or more when stimulated with EGF. Although this method does demonstrate the relative change of a pool of phosphorylation sites, it still says nothing about their biological relevance. For example, a 1% change in the phosphorylation of a whole cell lysate measurement of Akt would be considered an irrelevant change, based on most fold-changes suggested. However, if that 1% change represented a 75% change in a particular population of the cell, for example the membrane associated population, which is the important component for driving a particular downstream effect, then this would be extremely important. Therefore, one of the best methods for understanding the importance of a phosphorylation event is to understand why and how it is regulated and its subsequent function. Given the immensity of this undertaking, computational methodologies, such as those proposed in this thesis, will be vital to generating predictions regarding regulation and function in a manner that matches the speed of data generation.

5.3 Limitations of MS phosphoproteomic measurement

There are many advantages to mass spectrometry as a measurement methodology for post-translational modifications. In particular, the sensitivity of MS far surpasses

most other techniques, given a correctly chosen enrichment strategy. This is one of the reasons we believe that we were able to detect non-proline directed substrate recognition by MPM-2 for the first time, Chapter 2. Additionally, resolution for very specific residues is mostly guaranteed, unlike techniques that rely on detection using antibodies. However, there are some important limitations to consider when deriving biological hypotheses from MS measured data. In particular, scientists are turning towards automatic peak detection algorithms in order to handle the increasingly larger number of detectable peptides. These algorithms are accompanied by some acceptable false positive rate, typically 5%. A 5% false positive rate for a dataset that includes the assignment of 1000 phosphorylation sites translates to an erroneous assignment of roughly 50 phosphorylation sites. As these sites are gathered into resources such as Uniprot [114], PhosphoSite [39], and Phospho.ELM [22], the occurrence of potentially incorrect assignments are lost to the end-user of that information. Experimental evidence for post-translational modifications in PTMScout is clearly denoted, allowing the user to evaluate the extent of agreement between multiple, independent experiments. However, the end-user must still take potential for erroneous data into account. One other limitation of site-assignment for post-translational modifications within a measured peptide occurs when using MS methodologies, independent of the nature of the assignment method (manual or automatic). Modification sites lying within the middle of large peptide sequences, where more than one possible candidate amino acid exists, cannot be determined when non-complete parent-ion fractionation occurs. This error must be properly handled during data presentation and subsequent addition to post-translational modification resources. The additional complication introduced in data handling of MS datasets by ambiguous site assignment within peptides has not yet had a standard solution accepted by the field. Unfortunately, this means that handling of ambiguous site assignments varies from lab to lab and dataset to dataset.

Another important limitation of many MS experiments, which is not limited to only this measurement method, is the evaluation of whole cell lysates. As mentioned in Chapter 1, the location of signaling components within RTK networks can

be an important factor in the determining downstream effects. This is also one of the reasons that evaluating biological relevance based on quantitative fold-changes is non-ideal, as outlined in the hypothetical example of Akt phosphorylation. One method for determining the specific signaling differences among compartments is to use cellular fractionation prior to quantitation, for example Olsen et. al. [78] measure phosphorylation in a nuclear and cytosolic fraction of human cells. However, this fraction technique is not standard for most MS measurement pipelines and when used still presents the same problems regarding fraction impurities that other biological measurements face. Unfortunately, this means that deriving spatial information from these datasets is extremely difficult given the nature of the measurements. The incorporation of Gene Ontology cellular compartment information may be one way in which the tools in this thesis could aid in making the distinction between specific signaling locations.

5.4 Expansion of PTMScout

PTMScout was named with the intention of it eventually becoming a general resource for multiple post-translational modifications. We have taken the first steps in order to expand PTMScout to include more than just phosphorylation by incorporating acetylation measurements. The field of MS measurement will continue to rapidly expand the repertoire of available experimental evidence of PTMs, and PTMScout will need to be expanded to meet that need. A mundane, but important obstacle to this incorporation will be codification of a structured vocabulary, both machine- and human-readable, for representation of the vast number of possible modifications on a limited number of amino acids. Uniprot's [114] controlled vocabulary of modifications can serve as a starting point for the solution to this problem.

Another major need in the field is the ability to analyze data prior to publication. Unfortunately, the solution to this need is more complicated than what could be reasonably addressed in the initial implementation of PTMScout. Ideally, a resource will be established that has a shared and up-to-date view of the modification states of pro-

teins, but that can allow for stand-alone databases of local and private data. Other desirable extensions of the tool include expansion of metadata terms, such as pathway annotations, and incorporation of unsupervised learning algorithms to further automate analysis. Additionally, the framework of experimental datasets, metadata annotations of the participating biological molecules, and subset selection with automatic enrichment is extensible to other types of large-scale biological datasets, for example, RNAi, gene expression, and protease experiments.

5.5 The next steps in high-throughput unsupervised learning analysis

Chapter 4 utilized a single experimental dataset of the EGFR network and showed we could produce interaction style maps through the analysis of robust co-clustering. We have additionally analyzed several datasets, both time-dynamic measurements of the system as well as steady-state experiments with conditional perturbations, such as overexpression of EGFRvIII. We are encouraged that all analyses of time-dynamic measurements of the system performed so far have recapitulated known network information. However, conditional experiments have not recapitulated known network information. This may be a function of the inherent nature of the data, that steady-state co-regulation of phosphopeptides cannot distinguish the temporal nature of the way in which network interactions were derived. Also, it may be a function of "network rewiring" that may occur upon the introduction of a transforming event, such as overexpression of an oncogenic mutation. In this last case, the disagreement between derived network topologies from experimental measurement and "canonical" network models may in fact indicate the fragility and specificity of the derived network models. These differences will be important to explore in order to understand whether the derivation capability of experimental information is limited in certain cases or our basic understanding and representation of the network is at odds with the real biology of the system.

Another important future step is to compare the basic network components that are derived using the same system measured under different experimental conditions. For example, when we compare the grouping of BCAR1 phosphorylation sites between the *EGF7* [122] and *HER2* [123] datasets, we see that the immediate dephosphorylation of BCAR1 Y327 is EGF specific. When the data driving the relationships includes both EGF and HRG stimulation, all sites on BCAR1 consistently cluster together and Y327 no long clusters with focal adhesion proteins like BCAR3 and FAK/PTK2. Therefore, we gain more understanding of the function and regulation of these sites as we expand the repertoire of conditions under which they are measured. These types of comparisons will help us to understand the condition- and time-specific modes in which phosphorylation-specific interactions occur.

5.6 Bringing it all together: modification codes

The treatment of modifications, for the most part, has been performed for individual sites of phosphorylation. For example, tens of phosphorylation sites exist in the cytoplasmic tail of the EGF receptor, yet we do not fully understand how each site interacts with other sites. For example, although multiply phosphorylated, can a receptor bind multiple signaling components through each of its sites, or will steric hindrance prevent multiple binding events? If so, how will determination of the binding event take place? Alternatively, is there a phosphorylation "state" that exists on the receptor based on the dimerization partner or the ligand of stimulation? Additionally, how do different PTMs interact with each other? We know, for example, that phosphorylation of Y1069/Y1045 on the EGF receptor recruits Cbl, thereby causing ubiquitination of EGFR. This indicates a model of sequential modification events in the network. We are starting to see now, with the advent of high-throughput acetylation experiments, that the kinase activation loop is actetylated as well as phosphorylated. What does this mean for the activity of the kinase? Does the addition of an acetylation change the activity or does it serve another purpose? Are the modifications mutually exclusive, sequential, or independent? A host of questions and

increased complexity will emerge in the future as data becomes available. To adequately address this complexity, both experimental and computational strategies will be required.

Appendix A

Information and Materials for Chapter 2

A.1 Motif enrichment tables for EGFRvIII vs. DK

Table A.1: Motifs significantly enriched among top quartile of MPM-2 antigen peptides upregulated in U87-M cells vs. U87-DK controls.

Motif[1]	Motif in Foreground	Motif in Background	Foreground Size	Background Size	Statistical Significance
xD.E.E	6	6	25	95	2.04×10^{-4}
x-.E.-	8	11	25	95	8.24×10^{-4}
S.....**x**-	5	5	25	95	9.17×10^{-4}
xD.E.-	7	9	25	95	1.05×10^{-3}
PY..**s**	6	7	25	95	1.17×10^{-3}
S....**x**O	6	8	25	95	3.80×10^{-3}
S.....**s**..-	4	4	25	95	3.97×10^{-3}
S.....**s**-	4	4	25	95	3.97×10^{-3}
-.sD.E.E	4	4	25	95	3.97×10^{-3}
S....**x**.L	4	4	25	95	3.97×10^{-3}
S....**x**.....O	4	4	25	95	3.97×10^{-3}
S.....**x**D	4	4	25	95	3.97×10^{-3}
D.**x**D	4	4	25	95	3.97×10^{-3}
F**x**D.E.-	4	4	25	95	3.97×10^{-3}
xD.E-E	4	4	25	95	3.97×10^{-3}
x-	11	23	25	95	9.36×10^{-3}
xP	13	64	25	95	0.983

[1] "s" = pS, "x" = pS/pT, "." = Any amino acid, "-" = D/E, "O" = M/I/L/V

Table A.2: Motifs significantly enriched among top quartile of MPM-2 antigen peptides upregulated in U87-SH cells vs. U87-DK controls.

Motif[1]	Motif in Foreground	Motif in Background	Foreground Size	Background Size	Statistical Significance
D.x	8	10	25	95	2.73×10^{-4}
-.x-	8	11	25	95	8.24×10^{-4}
-..D.x	5	5	25	95	9.17×10^{-4}
-.s-.-	7	9	25	95	1.05×10^{-3}
D.x-	6	7	25	95	1.17×10^{-3}
-.x	11	20	25	95	2.02×10^{-3}
x-	12	23	25	95	2.09×10^{-3}
-.s	9	15	25	95	2.83×10^{-3}
D.s	6	8	25	95	3.80×10^{-3}
-.s-.E	6	8	25	95	3.80×10^{-3}
-..-.x	6	8	25	95	3.80×10^{-3}
-.xD	6	8	25	95	3.80×10^{-3}
-.sD.E.E	4	4	25	95	3.97×10^{-3}
sD.-.-O	4	4	25	95	3.97×10^{-3}
-..D.x-	4	4	25	95	3.97×10^{-3}
D.xD	4	4	25	95	3.97×10^{-3}
-...XsP..S	4	4	25	95	3.97×10^{-3}
D.s-.E	5	6	25	95	4.48×10^{-3}
-.sD.-.-	5	6	25	95	4.48×10^{-3}
-.s-L-	5	6	25	95	4.48×10^{-3}
xD.E.E	5	6	25	95	4.48×10^{-3}
s-	10	19	25	95	5.81×10^{-3}
xD.-.-	7	11	25	95	6.47×10^{-3}
s-.-	9	17	25	95	9.30×10^{-3}
sD.-.-	6	9	25	95	9.30×10^{-3}
-..x	6	9	25	95	9.30×10^{-3}
x-.-	10	20	25	95	9.60×10^{-3}
xP	12	63	25	95	0.995

[1] "s" = pS, "x" = pS/pT, "." = Any amino acid, "-" = D/E, "O" = M/I/L/V

A.2 CK2 activity measurements in EGFRvIII expressing cells

A.2.1 Protocol

A Casein Kinase 2 assay kit was purchased from Millipore, Catalog #17-132. [γ-^{32}]ATP (3000 Ci/mmol) was purchased from Perkin Elmer. U87-H and U87-DK cells were lysed (using Chapter 2 Cell lysis protocol) and mixed with a 90μl cold /10μl hot ATP mix, 10μl PKA inhibitor, and CK2 substrate peptide (RRRDDDSDDD) for 10 minutes at 30°C. The reaction was stopped with 20μl of 40% trichloroacetic acid (TCA). P81 paper was spotted with 25μl of the reaction mix and allowed to dry for 30 seconds before being transferred to a 0.75% phosphoric acid containing conical tube. The P81 paper was washed 6 times with 0.75% phosphoric acid for 1 minute then followed by a 1 minute wash with acetone. After drying the paper was immersed in scintillation fluid and read using a scintillation counter. CK2 inhibition controls

were performed by pre-incubating lystes with either a DMSO control or 2.5, 5 and 10 μM TBCA for one hour on ice.

A.2.2 TBCA inhibitor control

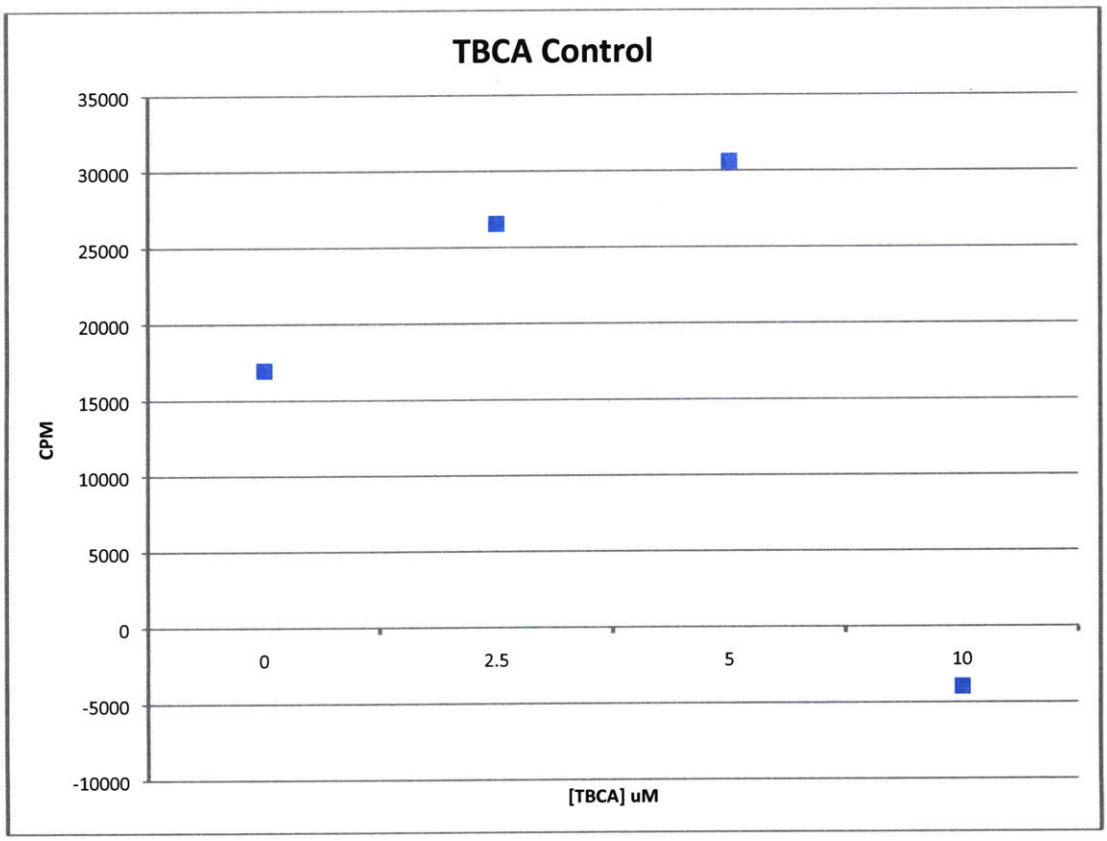


Figure A-1: The scintillation count of U87-DK cell lysates with background scintillation removed and incubated with either DMSO (0) or 2.5, 5, or 10μl TBCA, a CK2 inhibitor.

A.3 MPM-2 degenerate peptide library quantitation

Table A.3: Raw and processed values for the MPM-2 solid-phase library screen

Serine																									
	x	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	pS	pT	KAc
-4	57.68	56.64	61.94	90.94	102.5	83.58	62.62	72.84	71.12	56	65.03	80.37	91	59.66	76.23	53.86	70.26	69.42	66.47	87.79	82.52	67.09	64.61	58.97
-3	62.57	58.43	85.89	81.6	78.94	247.56	72.77	69.04	77.48	60.68	70.53	62.22	66.25	58.54	60.75	57.02	65.52	62.87	65.32	109.2	280.73	75.91	90.22	59.3
-2	70.41	63.12	78.25	102.7	188.96	130.59	63.21	67.34	367.09	63.98	96.01	122.61	59.75	65.75	94.21	57	58.11	66.15	222.48	288.54	96.48	93	146.58	85.05
-1	70.36	81.92	72.98	61.74	68	412.35	70.49	71.22	534.65	74.83	523.81	66.46	62.25	147.84	58.31	58.75	58.46	87.86	538.74	79.93	101.99	117.94	587.22	61.96
1	67.83	80.74	90.13	243.85	266.34	117.06	111.72	88.83	85.62	66.77	91.83	68.96	71.2	182.79	64.14	67.93	67.57	69.91	79.7	111.08	107.65	169.37	552.53	71.83
2	68.33	60.71	97.46	76.8	73.81	311.16	90.76	81.42	500.27	76.22	529.49	78.86	64.94	83.22	59.59	66.97	59.35	74.77	275.22	92.96	109.84	301.43	216.61	79.49
3	72.89	62.37	74.48	132.68	198.11	124.47	71.88	139.18	153.79	67.14	139.95	80.97	74.1	65.68	76	69.77	65	64.38	88.48	176.51	90	177.68	629.31	75.84
4	71.31	63.47	72.64	140.09	204.02	115.35	75.86	136.27	92.12	65.89	112.34	74.53	69.11	73.76	73.83	68.73	71.52	72.88	90.42	96.46	131.71	327.38	589.23	75.33
phosphoSerine Control																									
	x	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	pS	pT	KAc
-4	50.84	53.27	51.35	54.43	56.76	56.86	59.42	65.3	62.01	59.9	62.34	59.2	57.4	59.38	56.85	55.82	57.07	56.05	60.69	64.34	75.17	64.72	63.96	65.18
-3	51.53	51.41	75.49	53.54	52.77	56.02	57.13	68.89	58.39	56.81	56.46	56.07	57.37	57.86	59.08	56.4	57.14	56.19	57.06	58.26	63.12	64.12	85.14	69.9
-2	50.38	51.45	68.19	50.93	51.75	53.71	54.7	66.9	55.6	53.67	54.41	54.69	55.45	55.01	55.35	54.59	55.37	55.24	54.88	57.76	62.22	65.65	93.64	63.06
-1	49.68	51.17	64.73	51.68	52.24	52.4	54.2	64.61	53.16	52.31	52.52	53.81	54.52	53.86	57.72	53.41	54.79	54.76	54.26	57.9	60.53	83.98	533.74	65.04
1	49.09	50.37	55.95	49.55	49.93	52.38	52.8	56.82	53.09	51.55	51.25	52.66	53.15	54.66	56.59	54.97	53.91	55.28	54.07	56.84	60.93	68.09	277.56	65.2
2	47.76	68.47	60.51	51.57	50.86	52.18	53.25	55.01	52.43	52.16	51.99	52.36	54.11	53.08	56.05	52.7	53.79	54.04	54.25	57.34	60.76	81.38	492.4	66.04
3	45.42	49.09	58.28	50.29	50.3	51.77	51.86	52.74	50.75	49.87	50.8	50.52	52.02	50.91	54.07	50.15	51.9	52.11	53.19	56	58.26	87.3	552.55	64.03
4	45.01	47.77	52.9	49.04	50.09	48.72	50.27	50.79	47.75	46.73	47.93	47.14	48.89	48.39	49.34	48.18	48.07	49.83	50.05	51.44	54.08	132.51	483.78	59.78

Normalized Values																							
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	pS	pT	KAc
-4	0.76	0.86	1.2	1.29	1.05	0.76	0.8	0.82	0.67	0.75	0.97	1.14	0.72	0.96	0.69	0.88	0.89	0.78	0.98	0.79	0.74	0.72	0.65
-3	0.81	0.82	1.09	1.07	3.17	0.91	0.72	0.95	0.77	0.9	0.8	0.83	0.72	0.74	0.72	0.82	0.8	0.82	0.34	3.19	0.85	0.76	0.61
-2	0.88	0.82	1.45	2.62	1.74	1.74	0.72	4.73	0.85	1.26	1.61	0.77	0.86	1.22	0.75	0.75	0.86	2.9	3.58	1.11	1.02	1.12	0.97
-1	1.15	0.81	0.86	0.93	5.64	0.93	0.79	7.21	1.03	7.15	0.89	0.82	1.97	0.72	0.79	0.76	1.15	7.11	0.99	1.21	1.01	0.79	0.68
1	1.15	1.15	3.53	3.82	1.6	1.52	1.12	1.16	0.93	1.28	0.94	0.96	2.4	0.81	0.89	0.9	0.91	1.06	1.4	1.27	1.78	1.43	0.79
2	0.64	1.15	1.07	1.04	4.27	1.22	1.06	6.84	1.05	7.3	1.08	0.86	1.12	0.76	0.91	0.79	0.99	3.63	1.16	1.3	2.65	0.32	0.86
3	0.91	0.92	1.89	2.82	1.72	0.99	1.89	2.17	0.96	1.97	1.15	1.02	0.92	1.01	1	0.9	0.89	1.19	2.26	1.11	1.46	0.82	0.85
4	0.95	0.98	2.05	2.92	1.7	1.08	1.92	1.38	1.01	1.68	1.13	1.01	1.09	1.07	1.02	1.07	1.05	1.29	1.34	1.75	1.77	0.87	0.9

Appendix B

PTMScout Database Schema

id - each table has a unique internal id.

Experiment Tables

experiment Contains pertinent information about the experiment, including all the information regarding a published report. The *ambiguity* field, binary, indicates whether an experiment contains ambiguous peptide:protein assignments, if 1 it will cause ambiguity searches and storage upon dataset loading. This field is 0 for compendia.

MS MS holds the peptide that was measured, with lower cases indicating the site of modification. Links to data, protein, experiment and MS_pep tables.

data There are many data:MS mappings, so priority allows us to reconstruct the vector of data. Type is the type of data, for example stddev or time, and label the specific value of that data point, e.g. type=minute and label=1 means a data has *value* at 1 minute. When multiple runs exist based on multiple line entries for an MS peptide, run is incremented. "NA" is binary, 1 if that value did not exist for that particular measurement. This serves as a placeholder for reconstructing a vector.

Peptide Tables

pep The peptide table contains the singly modified alignments of a measured peptide (so there can be many pep to one MS if an MS measurement contains multiple modifications). For increased query speed, the Pfam domain a site falls in is stored in the field *pfam_site*.

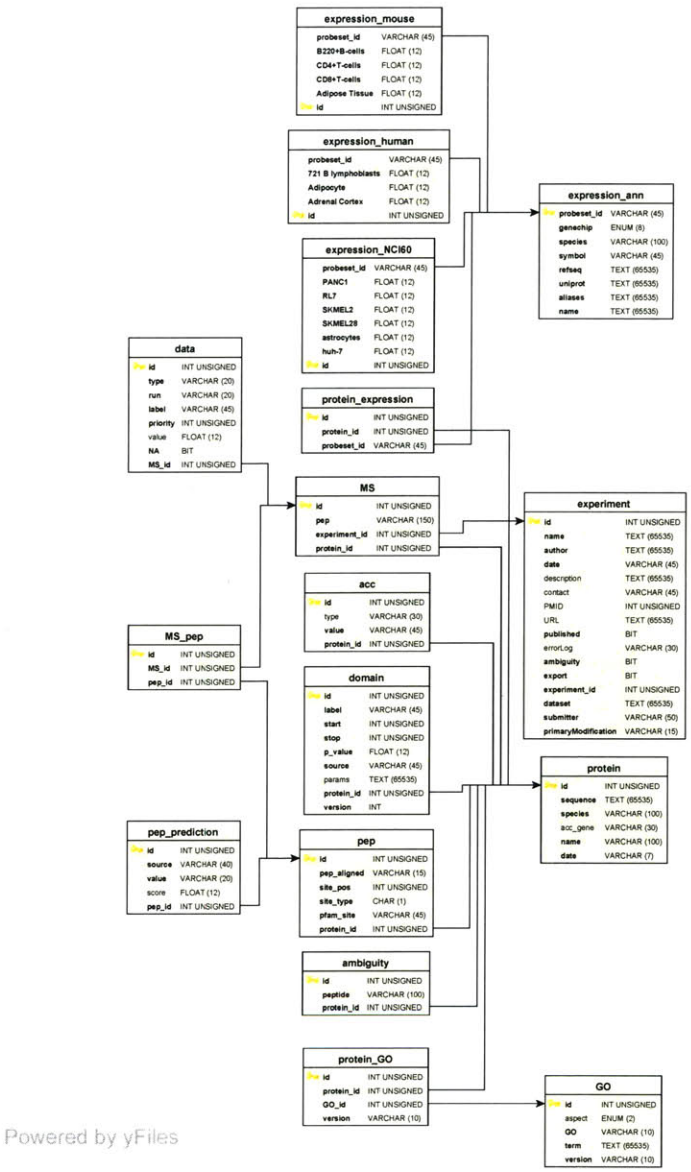


Figure B-1: Database schema for PTMScout. Expression tables have been condensed

pep_prediction There are many predictions per peptide possible. *Source* indicates the source of the prediction or annotation, such as Scansite_kinase and *score*

holds the value for those sources that are predictions.

ambiguity A table for storage and reference to peptide sequences and the proteins they can be assigned to through protein_id.

Protein Tables

protein Protein information including the full sequence and a representative gene name is stored here.

acc Protein accessions, based on type (such as swissprot, refseq, etc.) are stored here, many to one linkage with protein table.

domain There are potentially many to one domain entries per protein for a given source. Currently, the only source of domain information is Pfam. Predictions score stored as *p_value*.

protein_GO and GO protein_GO links the many to many relationships between GO terms and proteins. *aspect* contains the code for GO type, MF, BP, or CC. *version* is important for knowing and then updating the GO terms appropriately.

expression Expression tables from GNF symatlas project are imported as-is, and protein_expression links PTMScout internal *protein_id* with expression information via the *probeset_id*.

Appendix C

Code Statistics

Table C.1: Count of the code used to develop tools and algorithms for the work presented in this thesis. This count does not include Matlab code automatically generated by Perl.

Language	Lines of Code	Lines of Comments
Perl	37,106	11,315
Python	4,921	1,014
Javascript	3,627	617
Matlab	338,727	10,831

Appendix D

Licenses

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Apr 12, 2010

This is a License Agreement between Kristen M Naegle ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2406551474291
License date	Apr 12, 2010
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Molecular Cell Biology
Licensed content title	Untangling the ErbB signalling network
Licensed content author	Yosef Yarden, Mark X. Sliwkowski
Volume number	
Issue number	
Pages	
Year of publication	2001
Portion used	Figures / tables
Number of figures / tables	1
Requestor type	Student
Type of Use	Thesis / Dissertation
Billing Type	Invoice
Company	Kristen M Naegle
Billing Address	77 Massachusetts Ave Cambridge, MA 02139 United States
Customer reference info	
Total	0.00 USD
Terms and Conditions	

Bibliography

[1] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, —2003—.

[2] K. Ahmed, D. A. Gerber, and C. Cochet. Joining the cell survival squad: an emerging role for protein kinase ck2. *Trends Cell Biol*, 12(5):226–30, —2002—.

[3] A. Alonso, J. Sasin, N. Bottini, I. Friedberg, A. Osterman, A. Godzik, T. Hunter, J. Dixon, and T. Mustelin. Protein tyrosine phosphatases in the human genome. *Cell*, 117(6):699–711, —2004—.

[4] R. Amanchy, B. Periaswamy, S. Mathivanan, R. Reddy, S. G. Tattikota, and A. Pandey. A curated compendium of phosphorylation motifs. *Nat Biotechnol*, 25(3):285–6, —2007—.

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, —2000—.

[6] M. K. Ayrapetov, Y. H. Wang, X. Lin, X. Gu, K. Parang, and G. Sun. Conformational basis for SH2-Tyr(P)527 binding in Src inactivation. *J Biol Chem*, 281(33):23776–84, —2006—.

[7] Andrew J. Bannister, Eric A. Miska, Dirk Grlich, and Tony Kouzarides. Acetylation of importin-[alpha] nuclear import factors by CBP/p300. *Current Biology*, 10(8):467–470, —2000—.

[8] A. G. Batzer, D. Rotin, J. M. Urena, E. Y. Skolnik, and J. Schlessinger. Hierarchy of binding sites for Grb2 and Shc on the epidermal growth factor receptor. *Mol Cell Biol*, 14(8):5192–201, —1994—.

[9] R. C. Beavis. Using the global proteome machine for protein identification. *Methods Mol Biol*, 328:217–28, —2006—.

[10] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57(1):289–300, —1995—.

[11] R. Bose, M. A. Holbert, K. A. Pickin, and P. A. Cole. Protein tyrosine kinase-substrate interactions. *Curr Opin Struct Biol*, 16(6):668–75, —2006—.

[12] P. Burke, K. Schooler, and H. S. Wiley. Regulation of epidermal growth factor receptor signaling by endocytosis and intracellular trafficking. *Mol Biol Cell*, 12(6):1897–910, —2001—.

[13] MB Calalb, TR Polte, and SK Hanks. Tyrosine phosphorylation of focal adhesion kinase at sites in the catalytic domain regulates kinase activity: a role for src family kinases. *Mol. Cell. Biol.*, 15(2):954–963, —1995—.

[14] C. Choudhary, C. Kumar, F. Gnad, M. L. Nielsen, M. Rehman, T. C. Walther, J. V. Olsen, and M. Mann. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, 325(5942):834–40, —2009—.

[15] A. Citri, K. B. Skaria, and Y. Yarden. The deaf and the dumb: the biology of erbb-2 and erbb-3. *Exp Cell Res*, 284(1):54–65, —2003—.

[16] P. Cohen. The origins of protein phosphorylation. *Nat Cell Biol*, 4(5):E127–30, —2002—.

[17] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res*, 14(6):1188–90, —2004—.

[18] Jr. Darnell, J. E. Stats and gene regulation. *Science*, 277(5332):1630–5, —1997—.

[19] F. M. Davis, T. Y. Tsao, S. K. Fowler, and P. N. Rao. Monoclonal antibodies to mitotic cells. *Proc Natl Acad Sci U S A*, 80(10):2926–30, —1983—.

[20] N. Dephoure, C. Zhou, J. Villen, S. A. Beausoleil, C. E. Bakalarski, S. J. Elledge, and S. P. Gygi. A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci U S A*, —2008—.

[21] P. D'Haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–26, —2000—.

[22] F. Diella, C. M. Gould, C. Chica, A. Via, and T. J. Gibson. Phospho.ELM: a database of phosphorylation sites–update 2008. *Nucleic Acids Res*, 36(Database issue):D240–4, —2008—.

[23] Francesca Diella, Scott Cameron, Christine Gemund, Rune Linding, Allegra Via, Bernhard Kuster, Thomas Sicheritz-Ponten, Nikolaj Blom, and Toby Gibson. Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5(1):79, —2004—.

[24] Sandrine Dudoit and M. J. van der Laan. *Multiple testing procedures with applications to genomics*. Springer, New York ; London, —2008—.

[25] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8, —1998—.

[26] A. E. Elia, P. Rellos, L. F. Haire, J. W. Chao, F. J. Ivins, K. Hoepker, D. Mohammad, L. C. Cantley, S. J. Smerdon, and M. B. Yaffe. The molecular basis for phosphodependent substrate targeting and regulation of Plks by the Polo-box domain. *Cell*, 115(1):83–95, —2003—.

[27] A. E. Escargueil and A. K. Larsen. Mitosis-specific MPM-2 phosphorylation of DNA topoisomerase IIalpha is regulated directly by protein phosphatase 2A. *Biochem J*, 403(2):235–42, —2007—.

[28] A. E. Escargueil, S. Y. Plisov, O. Filhol, C. Cochet, and A. K. Larsen. Mitotic phosphorylation of DNA topoisomerase II alpha by protein kinase CK2 creates the MPM-2 phosphoepitope on Ser-1469. *J Biol Chem*, 275(44):34710–8, —2000—.

[29] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res*, 36(Database issue):D281–8, —2008—.

[30] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–6, —2007—.

[31] F. B. Furnari, T. Fenton, R. M. Bachoo, A. Mukasa, J. M. Stommel, A. Stegh, W. C. Hahn, K. L. Ligon, D. N. Louis, C. Brennan, L. Chin, R. A. DePinho, and W. K. Cavenee. Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev*, 21(21):2683–710, —2007—.

[32] R. Giancarlo, D. Scaturro, and F. Utro. Computational cluster validation for microarray data analysis: experimental assessment of Clest, Consensus Clustering, Figure of Merit, Gap Statistics and Model Explorer. *BMC Bioinformatics*, 9:462, —2008—.

[33] Florian Gnad, Shubin Ren, Juergen Cox, Jesper Olsen, Boris Macek, Mario Oroshi, and Matthias Mann. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biology*, 8(11):R250, —2007—.

[34] B. D. Gomperts, Peter E. R. Tatham, and Ijsbrand M. Kramer. *Signal transduction*. Elsevier/Academic Press, Amsterdam ; Boston, 2nd edition, —2009—.

[35] F. A. Gonzalez, D. L. Raden, and R. J. Davis. Identification of substrate recognition determinants for human ERK1 and ERK2 protein kinases. *J Biol Chem*, 266(33):22159–63, —1991—.

[36] Jun-Lin Guan. Focal adhesion kinase in integrin signaling. *Matrix Biology*, 16(4):195–200, —1997—.

[37] S. A. Ha, S. M. Shin, H. Namkoong, H. Lee, G. W. Cho, S. Y. Hur, T. E. Kim, and J. W. Kim. Cancer-associated expression of minichromosome maintenance 3 gene in several human cancers and its involvement in tumorigenesis. *Clin Cancer Res*, 10(24):8386–95, —2004—.

[38] D. E. Hanna, A. Rethinaswamy, and C. V. Glover. Casein kinase II is required for cell cycle progression during G1 and G2/M in Saccharomyces cerevisiae. *J Biol Chem*, 270(43):25905–14, —1995—.

[39] Peter V. Hornbeck, Indy Chabra, Jon M. Kornhauser, Elzbieta Skrzypek, and Bin Zhang. Phosphosite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6):1551–1561, —2004—.

[40] H. S. Huang, M. Nagane, C. K. Klingbeil, H. Lin, R. Nishikawa, X. D. Ji, C. M. Huang, G. N. Gill, H. S. Wiley, and W. K. Cavenee. The enhanced tumorigenic activity of a mutant epidermal growth factor receptor common in human cancers is mediated by threshold levels of constitutive tyrosine phosphorylation and unattenuated signaling. *J Biol Chem*, 272(5):2927–35, —1997—.

[41] P. H. Huang, A. Mukasa, R. Bonavia, R. A. Flynn, Z. E. Brewer, W. K. Cavenee, F. B. Furnari, and F. M. White. Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. *Proc Natl Acad Sci U S A*, 104(31):12867–72, —2007—.

[42] T. Hunter and B. M. Sefton. Transforming gene product of rous sarcoma virus phosphorylates tyrosine. *Proc Natl Acad Sci U S A*, 77(3):1311–5, —1980—.

[43] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, —1999—.

[44] I. Jonassen, J. F. Collins, and D. G. Higgins. Finding flexible patterns in unaligned protein sequences. *Protein Sci*, 4(8):1587–95, —1995—.

[45] R. B. Jones, A. Gordus, J. A. Krall, and G. MacBeath. A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature*, 439(7073):168–74, —2006—.

[46] B. A. Joughin, K. M. Naegle, P. H. Huang, M. B. Yaffe, D. A. Lauffenburger, and F. M. White. An integrated comparative phosphoproteomic and bioinformatic approach reveals a novel class of MPM-2 motifs upregulated in EGFRvIII-expressing glioblastoma cells. *Mol Biosyst*, 5(1):59–67, —2009—.

[47] P. J. Kennelly and E. G. Krebs. Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. *J Biol Chem*, 266(24):15555–8, —1991—.

[48] P. J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler. The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985–8, —2004—.

[49] B. N. Kholodenko, O. V. Demin, G. Moehren, and J. B. Hoek. Quantification of short term signaling by the epidermal growth factor receptor. *J Biol Chem*, 274(42):30169–81, —1999—.

[50] B. M. King and B. Tidor. MIST: Maximum information spanning trees for dimension reduction of biological data sets. *Bioinformatics*, 25(9):1165–72, —2009—.

[51] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, —1990—.

[52] K. S. Kolibaba and B. J. Druker. Protein tyrosine kinases and cancer. *Biochim Biophys Acta*, 1333(3):F217–48, —1997—.

[53] T. Kouzarides. Acetylation: a regulatory modification to rival phosphorylation? *EMBO J*, 19(6):1176–9, —2000—.

[54] A. Kumagai and W. G. Dunphy. Purification and molecular cloning of Plx1, a Cdc25-regulatory kinase from xenopus egg extracts. *Science*, 273(5280):1377–80, —1996—.

[55] N. Kumar, R. Afeyan, S. Sheppard, B. Harms, and D. A. Lauffenburger. Quantitative analysis of Akt phosphorylation and activity in response to EGF and insulin treatment. *Biochem Biophys Res Commun*, 354(1):14–20, —2007—.

[56] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–8, —2007—.

[57] T. Y. Lee, H. D. Huang, J. H. Hung, H. Y. Huang, Y. S. Yang, and T. H. Wang. dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res*, 34(Database issue):D622–7, —2006—.

[58] Hong Li, Xiaobin Xing, Guohui Ding, Qingrun Li, Chuan Wang, Lu Xie, Rong Zeng, and Yixue Li. SysPTM: A systematic resource for proteomic research on post-translational modifications. *Mol Cell Proteomics*, 8(8):1839–1849, —2009—.

[59] R. Linding, L. J. Jensen, G. J. Ostheimer, M. A. van Vugt, C. Jorgensen, I. M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J. G. Park, L. D. Samson, J. R. Woodgett, R. B. Russell, P. Bork, M. B. Yaffe, and T. Pawson. Systematic discovery of in vivo phosphorylation networks. *Cell*, 129(7):1415–26, —2007—.

[60] T.E. Liu and A. D. Lauffenburger. *Systems Biomedicine: Concepts and Perspectives*. Elsevier, —2009—.

[61] Harvey F. Lodish, Paul T. Matsudaira, Chris Kaiser, and Monty Krieger. *Molecular cell biology.* W.H. Freeman and Company, New York, 5th edition, —2004—.

[62] W. Lu, D. Gong, D. Bar-Sagi, and P. A. Cole. Site-specific incorporation of a phosphotyrosine mimetic reveals a role for tyrosine phosphorylation of SHP-2 in cell signaling. *Mol Cell*, 8(4):759–69, —2001—.

[63] T. J. Lynch, D. W. Bell, R. Sordella, S. Gurubhagavatula, R. A. Okimoto, B. W. Brannigan, P. L. Harris, S. M. Haserlat, J. G. Supko, F. G. Haluska, D. N. Louis, D. C. Christiani, J. Settleman, and D. A. Haber. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med*, 350(21):2129–39, —2004—.

[64] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–34, —2002—.

[65] F. Meggio and L. A. Pinna. One-thousand-and-one substrates of protein kinase CK2? *Faseb J*, 17(3):349–68, —2003—.

[66] M. L. Miller, L. J. Jensen, F. Diella, C. Jorgensen, M. Tinti, L. Li, M. Hsiung, S. A. Parker, J. Bordeaux, T. Sicheritz-Ponten, M. Olhovsky, A. Pasculescu, J. Alexander, S. Knapp, N. Blom, P. Bork, S. Li, G. Cesareni, T. Pawson, B. E. Turk, M. B. Yaffe, S. Brunak, and R. Linding. Linear motif atlas for phosphorylation-dependent signaling. *Sci Signal*, 1(35):ra2, —2008—.

[67] S. K. Mitra and D. D. Schlaepfer. Integrin-regulated FAK-Src signaling in normal and cancer cells. *Curr Opin Cell Biol*, 18(5):516–23, —2006—.

[68] D. K. Moscatello, M. Holgado-Madruga, A. K. Godwin, G. Ramirez, G. Gunn, P. W. Zoltick, J. A. Biegel, R. L. Hayes, and A. J. Wong. Frequent expression of a mutant epidermal growth factor receptor in multiple human tumors. *Cancer Res*, 55(23):5536–9, —1995—.

[69] M. Mukherji, L. M. Brill, S. B. Ficarro, G. M. Hampton, and P. G. Schultz. A phosphoproteomic analysis of the ErbB2 receptor tyrosine kinase signaling pathways. *Biochemistry*, 45(51):15529–40, —2006—.

[70] O. Mulner-Lorillon, J. Marot, X. Cayla, R. Pouhle, and R. Belle. Purification and characterization of a casein-kinase-II-type enzyme from Xenopus laevis ovary. Biological effects on the meiotic cell division of full-grown oocyte. *Eur J Biochem*, 171(1-2):107–17, —1988—.

[71] H. Nakajima, F. Toyoshima-Morimoto, E. Taniguchi, and E. Nishida. Identification of a consensus motif for Plk (Polo-like kinase) phosphorylation reveals Myt1 as a Plk1 substrate. *J Biol Chem*, 278(28):25277–80, —2003—.

[72] Y. Narita, M. Nagane, K. Mishima, H. J. Huang, F. B. Furnari, and W. K. Cavenee. Mutant epidermal growth factor receptor signaling down-regulates p27 through activation of the phosphatidylinositol 3-kinase/akt pathway in glioblastomas. *Cancer Res*, 62(22):6764–9, —2002—.

[73] P. M. Navolanic, L. S. Steelman, and J. A. McCubrey. EGFR family signaling and its association with breast cancer development and resistance to chemotherapy (Review). *Int J Oncol*, 22(2):237–52, —2003—.

[74] C. G. Nevill-Manning, T. D. Wu, and D. L. Brutlag. Highly specific protein sequence motifs for genome analysis. *Proc Natl Acad Sci U S A*, 95(11):5865–71, —1998—.

[75] Brad Nolen, Susan Taylor, and Gourisankar Ghosh. Regulation of protein kinases: Controlling activity through activation segment conformation. *Molecular Cell*, 15(5):661–675, —2004—.

[76] J. C. Obenauer, L. C. Cantley, and M. B. Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 31(13):3635–41, —2003—.

[77] K. Oda, Y. Matsuoka, A. Funahashi, and H. Kitano. A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol*, 1:2005 0010, —2005—.

[78] J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, and M. Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3):635–48, —2006—.

[79] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1(5):376–86, —2002—.

[80] J. G. Paez, P. A. Janne, J. C. Lee, S. Tracy, H. Greulich, S. Gabriel, P. Herman, F. J. Kaye, N. Lindeman, T. J. Boggon, K. Naoki, H. Sasaki, Y. Fujii, M. J. Eck, W. R. Sellers, B. E. Johnson, and M. Meyerson. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 304(5676):1497–500, —2004—.

[81] M. W. Pedersen, M. Meltorn, L. Damstrup, and H. S. Poulsen. The type III epidermal growth factor receptor mutation. biological significance and potential target for anti-cancer therapy. *Ann Oncol*, 12(6):745–60, —2001—.

[82] J. Peng, D. Schwartz, J. E. Elias, C. C. Thoreen, D. Cheng, G. Marsischky, J. Roelofs, D. Finley, and S. P. Gygi. A proteomics approach to understanding protein ubiquitination. *Nat Biotechnol*, 21(8):921–6, —2003—.

[83] I. Priness, O. Maimon, and I. Ben-Gal. Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, 8:111, —2007—.

[84] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue):D61–5, —2007—.

[85] I. Rigoutsos and A. Floratos. Combinatorial pattern discovery in biological sequences: The teiresias algorithm. *Bioinformatics*, 14(1):55–67, —1998—.

[86] K. Rikova, A. Guo, Q. Zeng, A. Possemato, J. Yu, H. Haack, J. Nardone, K. Lee, C. Reeves, Y. Li, Y. Hu, Z. Tan, M. Stokes, L. Sullivan, J. Mitchell, R. Wetzel, J. Macneill, J. M. Ren, J. Yuan, C. E. Bakalarski, J. Villen, J. M. Kornhauser, B. Smith, D. Li, X. Zhou, S. P. Gygi, T. L. Gu, R. D. Polakiewicz, J. Rush, and M. J. Comb. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell*, 131(6):1190–203, —2007—.

[87] M. Rodriguez, S. S. Li, J. W. Harper, and Z. Songyang. An oriented peptide array library (OPAL) strategy to study protein-protein interactions. *J Biol Chem*, 279(10):8802–7, —2004—.

[88] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*, 3(12):1154–69, —2004—.

[89] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–9, —2005—.

[90] J. Saez-Rodriguez, A. Goldsipe, J. Muhlich, L. G. Alexopoulos, B. Millard, D. A. Lauffenburger, and P. K. Sorger. Flexible informatics for linking experimental data to mathematical models via DataRail. *Bioinformatics*, 24(6):840–7, —2008—.

[91] R. Samaga, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, and S. Klamt. The logic of EGFR/ErbB signaling: theoretical properties and analysis of high-throughput data. *PLoS Comput Biol*, 5(8):e1000438, —2009—.

[92] S. Sarno and L. A. Pinna. Protein kinase CK2 as a druggable target. *Mol Biosyst*, 4(9):889–94, —2008—.

[93] K. Schmelzle and F. M. White. Phosphoproteomic approaches to elucidate cellular signaling networks. *Curr Opin Biotechnol*, 17(4):406–14, —2006—.

[94] B. Schoeberl, C. Eichler-Jonsson, E. D. Gilles, and G. Muller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat Biotechnol*, 20(4):370–5, —2002—.

[95] Randy S. Schrecengost, Rebecca B. Riggins, Keena S. Thomas, Michael S. Guerrero, and Amy H. Bouton. Breast cancer antiestrogen resistance-3 expression regulates breast cancer cell migration through promotion of p130Cas membrane localization and membrane ruffling. *Cancer Res*, 67(13):6174–6182, —2007—.

[96] W. X. Schulze, L. Deng, and M. Mann. Phosphotyrosine interactome of the ErbB-receptor kinase family. *Mol Syst Biol*, 1:2005 0008, —2005—.

[97] D. Schwartz and S. P. Gygi. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol*, 23(11):1391–8, —2005—.

[98] E. I. Schwartz, R. V. Intine, and R. J. Maraia. CK2 is responsible for phosphorylation of human La protein serine-366 and can modulate rpL37 5'-terminal oligopyrimidine mRNA metabolism. *Mol Cell Biol*, 24(21):9580–91, —2004—.

[99] J. Shi and J. Mailik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, —2000—.

[100] B. J. Simpson, J. Weatherill, E. P. Miller, A. M. Lessells, S. P. Langdon, and W. R. Miller. c-erbb-3 protein expression in ovarian tumours. *Br J Cancer*, 71(4):758–62, —1995—.

[101] A. Soling, M. Sackewitz, M. Volkmar, D. Schaarschmidt, R. Jacob, H. J. Holzhausen, and N. G. Rainov. Minichromosome maintenance protein 3 elicits a cancer-restricted immune response in patients with brain malignancies and is a strong independent predictor of survival in patients with anaplastic astrocytoma. *Clin Cancer Res*, 11(1):249–58, —2005—.

[102] Z. Songyang, S. Blechner, N. Hoagland, M. F. Hoekstra, H. Piwnica-Worms, and L. C. Cantley. Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr Biol*, 4(11):973–82, —1994—.

[103] Z. Songyang, K. P. Lu, Y. T. Kwon, L. H. Tsai, O. Filhol, C. Cochet, D. A. Brickey, T. R. Soderling, C. Bartleson, D. J. Graves, A. J. DeMaggio, M. F. Hoekstra, J. Blenis, T. Hunter, and L. C. Cantley. A structural basis for substrate specificities of protein ser/thr kinases: primary sequence preference of casein kinases i and ii, nima, phosphorylase kinase, calmodulin-dependent kinase ii, cdk5, and erk1. *Mol Cell Biol*, 16(11):6486–93, —1996—.

[104] H. Steen, B. Kuster, M. Fernandez, A. Pandey, and M. Mann. Tyrosine phosphorylation mapping of the epidermal growth factor receptor signaling pathway. *J Biol Chem*, 277(2):1031–9, —2002—.

[105] J.D. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, —2003—.

[106] P. T. Stukenberg, K. D. Lustig, T. J. McGarry, R. W. King, J. Kuang, and M. W. Kirschner. Systematic identification of mitotic phosphoproteins. *Curr Biol*, 7(5):338–48, —1997—.

[107] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–7, —2004—.

[108] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6):2907–12, —1999—.

[109] K. Tashiro, H. Konishi, E. Sano, H. Nabeshi, E. Yamauchi, and H. Taniguchi. Suppression of the ligand-mediated down-regulation of epidermal growth factor receptor by Ymer, a novel tyrosine-phosphorylated and ubiquitinated protein. *J Biol Chem*, 281(34):24612–22, —2006—.

[110] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nat Genet*, 22(3):281–5, —1999—.

[111] N. Theis-Febvre, O. Filhol, C. Froment, M. Cazales, C. Cochet, B. Monsarrat, B. Ducommun, and V. Baldin. Protein kinase CK2 regulates CDC25B phosphatase activity. *Oncogene*, 22(2):220–32, —2003—.

[112] Y. Tian, Y. Zhang, B. Zhong, Y. Y. Wang, F. C. Diao, R. P. Wang, M. Zhang, D. Y. Chen, Z. H. Zhai, and H. B. Shu. RBCK1 negatively regulates tumor necrosis factor- and interleukin-1-triggered NF-kappaB activation by targeting TAB2/3 for degradation. *J Biol Chem*, 282(23):16776–82, —2007—.

[113] N. K. Tonks. Protein tyrosine phosphatases: from genes, to function, to disease. *Nat Rev Mol Cell Biol*, 7(11):833–46, —2006—.

[114] Uniprot. The universal protein resource (UniProt) 2009. *Nucleic Acids Res*, 37(Database issue):D169–74, —2009—.

[115] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7:142, —2006—.

[116] J. Villen and S.P. Gygi. The scx/imac enrichment approach for global phosphorylation analysis by mass spectrometry. *Nature Protocols*, 3(10):1630–1638, —2008—.

[117] S. Walchli, X. Espanel, A. Harrenga, M. Rossi, G. Cesareni, and R. Hooft van Huijsduijnen. Probing protein-tyrosine phosphatase substrate specificity using a phosphotyrosine-containing phage library. *J Biol Chem*, 279(1):311–8, —2004—.

[118] Y. Wang and R. L. Klemke. PhosphoBlast, a computational tool for comparing phosphoprotein signatures among large datasets. *Mol Cell Proteomics*, 7(1):145–62, —2008—.

[119] J. M. Westendorf, P. N. Rao, and L. Gerace. Cloning of cDNAs for M-phase phosphoproteins recognized by the MPM2 monoclonal antibody and determination of the phosphorylated epitope. *Proc Natl Acad Sci U S A*, 91(2):714–8, —1994—.

[120] J. T. Wilson-Grady, J. Villen, and S. P. Gygi. Phosphoproteome analysis of fission yeast. *J Proteome Res*, 7(3):1088–97, —2008—.

[121] J. Wohlgemuth, M. Karas, T. Eichhorn, R. Hendriks, and S. Andrecht. Quantitative site-specific analysis of protein glycosylation by LC-MS using different glycopeptide-enrichment strategies. *Anal Biochem*, 395(2):178–88, —2009—.

[122] A. Wolf-Yadlin, S. Hautaniemi, D. A. Lauffenburger, and F. M. White. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc Natl Acad Sci U S A*, 104(14):5860–5, —2007—.

[123] A. Wolf-Yadlin, N. Kumar, Y. Zhang, S. Hautaniemi, M. Zaman, H. D. Kim, V. Grantcharova, D. A. Lauffenburger, and F. M. White. Effects of HER2 overexpression on cell signaling networks governing proliferation and migration. *Mol Syst Biol*, 2:54, —2006—.

[124] Y. H. Wong, T. Y. Lee, H. K. Liang, C. M. Huang, T. Y. Wang, Y. H. Yang, C. H. Chu, H. D. Huang, M. T. Ko, and J. K. Hwang. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res*, 35:W588–94, —2007—. 1362-4962 (Electronic).

[125] M. Xiang, C. Xue, L. Huicai, L. Jin, L. Hong, and H. Dacheng. Large-scale identification of novel mitosis-specific phosphoproteins. *Biochim Biophys Acta*, 1784(6):882–90, —2008—.

[126] Y. Xue, A. Li, L. Wang, H. Feng, and X. Yao. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, 7:163, —2006—.

[127] M. B. Yaffe. Phosphotyrosine-binding domains in signal transduction. *Nat Rev Mol Cell Biol*, 3(3):177–86, —2002—.

[128] M. B. Yaffe and L. C. Cantley. Signal transduction. grabbing phosphoproteins. *Nature*, 402(6757):30–1, —1999—.

[129] M. B. Yaffe and A. E. Elia. Phosphoserine/threonine-binding domains. *Curr Opin Cell Biol*, 13(2):131–8, —2001—.

[130] M. B. Yaffe, M. Schutkowski, M. Shen, X. Z. Zhou, P. T. Stukenberg, J. U. Rahfeld, J. Xu, J. Kuang, M. W. Kirschner, G. Fischer, L. C. Cantley, and K. P. Lu. Sequence-specific and phosphorylation-dependent proline isomerization: a potential mitotic regulatory mechanism. *Science*, 278(5345):1957–60, —1997—.

[131] Y. Yarden and M. X. Sliwkowski. Untangling the ErbB signalling network. *Nat Rev Mol Cell Biol*, 2(2):127–37, —2001—.

[132] Y. Zhang, A. Wolf-Yadlin, P. L. Ross, D. J. Pappin, J. Rush, D. A. Lauffenburger, and F. M. White. Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Mol Cell Proteomics*, 4(9):1240–50, —2005—.