Characterizing the evolutionary dynamics of protein phosphorylation sites for functional phospho-proteomics

by

Soon Heng Tan

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Molecular Genetics
University of Toronto

Abstract

Characterizing the evolutionary dynamics of protein phosphorylation sites for functional phospho-proteomics

Soon Heng Tan

Doctor of Philosophy

Graduate Department of Molecular Genetics

University of Toronto

2012

Protein phosphorylation is a prevalent reversible post-translational modification that influences protein functions. The advent of phospho-proteomic technologies now enables proteome-wide quantitative detection of residues phosphorylated under different physiological conditions. The functional consequences of the majority of these phosphorylation events are unknown. This calls for endeavors to characterize their molecular functions and cellular effects. This can be facilitated by systematic approaches to categorize phosphorylation events, interpret their importance and infer their functions. I carried out comparative, evolutionary and integrative analyses on *in vivo* phosphorylation events to address these challenges. First, I performed cross-species comparative phospho-proteomic analysis to identify evolutionarily conserved phosphorylation events in human. A sequence alignment approach was used to identify phosphorylation events conserved at similar sequence positions across orthologous proteins and a network alignment approach was applied to identify potential evolutionarily conserved kinase-substrate interactions. Conserved human phosphoproteins identified are found enriched for proteins encoded by known cancer- and disease-associated genes. Next, I developed a new approach to analyze the sequence conservation of known phosphorylated residues on human, mouse and yeast proteins that factored in the background mutational rates of protein and phosphorylatable residue. Furthermore, sites were analyzed according to (i) characterized functions, (ii) prevalence, (iii) stoichiometry, their occurrence in (iv) structurally disordered/ordered protein regions, in (v) proteins of various abundance and in (vi) proteins with different protein interaction propensity to identify the factors influencing sequence conservation of phosphorylated residues. Importantly, my analysis suggests that false positives and randomly phosphorylated residues are present in existing phosphorylation datasets and they are more common on high abundance proteins. Lastly, I characterized the theoretical maximum phosphorylation capacity in terms of phosphorylatable residues and discovered that genomic tyrosine frequency correlates negatively and significantly with tyrosine kinase frequency and cell type in metazoan. This observation suggests that fidelity of phosphotyrosine signaling occurred partially through global tyrosine depletion.

Acknowledgements

My life has been blessed with the unfailing love of my wife, Judice Koh, who wholeheartedly supported my decision to pursue PhD study about five years ago although it means moving to a foreign land halfway around the globe, away from her family and friends. She gave me the courage to embark the voyage when I hesitated, and gave me the security to complete it. Whatever I had achieved, if any, is very much hers also. I am also blessed by the selfless love of my father who did everything he can so that I will not hold off my decision. I am grateful to my mother-in-law for her care and to my brother-in-law for running errands at home.

What started as a PhD study turned up to be a journey of self-discovery and -actualization. For this, I am indebted to many people. I like to thank both my supervisors, Dr. Gary Bader and Dr. Tony Pawson, for giving me the opportunity to undertake my PhD study under their care. Gary is a very patient and dedicated supervisor who is there for his students when they need him. Tony is kind and wise. I am grateful for their unconditional support of my proposed research although it deviates substantially from what they originally intended for me.

I thank Dr Rune Linding who introduce me to the topics of *protein phosphorylation* and for his scientific guidance on my proposed research. He is a co-worker, an informal mentor and a friend who never hesitate to share his knowledge and opinion. My gratitude also goes to Dr. Anne-Claude Gingras and Dr. John Parkinson, my thesis advisors, for their helpful critical feedbacks on my research, particularly in the beginning. Their strict and high expectation have improved my research and made me a better scientist. I am also thankful to Claus Jørgensen for pointing out fallacies in my scientific thinking and to Adrian Pasculescu who I have learnt much from.

I have grown to love and is proud to be associated with the foreign land that I had embarked my voyage many years ago. My PhD and stay in Canada have been made more easy and memorable by many people I meet in this great country. For this, many thanks go to the all members of Bader lab, in particular to Shirley Hui, Vuk Pavlovic, Shobhit Jain, Daniele Merico, Oliver Stueker, David Gfeller, Moyez Dharsee, Michael Matan and Sylva Donaldson. The lunch chat, cottage stay, camping trips, conference trips, picnic, pub nights with them among many others are parts of my lovely memory from Toronto.

Members of Pawson lab have contributed greatly to my knowledge in *signaling* particularly Bernard Liu, Jerry Dish, Jing Jin, Greg Findlay, Ruijun Tian, Yong Zheng, Chen Chen, Nicolas Bisson and Frederic Fellouse. They are great people and I thank them for sharing their scientific knowledge and insight with me. My time in Pawson lab has been made more pleasant by the companionship of Bernard Liu, Marina Olhovsky, Jonathan So, Evangelia Petsalakis, Mohamed Soliman, Cunjie Zhao, Anna Yue Dai, Cristina Virag, Kelly Williton, Elizabeth Gray and Dorothy Zhao.

I am also thankful of the companionship of Xiaojian Shao, Dong Dong, Jingjing Li, Ke Jin, Xiao Li and Lee Zamparo. It has been great fun going out together for chinese food and hotpot.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Section 1.3.1, 1.4 and 1.5 were published in:

C. S. Tan and R. Linding. Experimental and computational tools useful for (re)construction of dynamic kinase-substrate networks. *Proteomics*, 9(23):5233-42., 2009 Dec.

1.1 Protein phosphorylation

Protein phosphorylation is arguably the most prevalent reversible post-translational modification (PTM) as it is estimated to occur on approximately one-third of all human proteins [30]. Phosphorylation is known to modulate the enzymatic activity, three-dimensional structure, degradation, subcellular localization, and biomolecular interaction of phosphorylated proteins. In eukaryotes, the process, catalyzed by a protein kinase, involves the attachment of a phosphate group from adenosine-5'-triphosphate (ATP) to serine, threonine or tyrosine typically but aspartate, histidine and arginine can also be phosphorylated. On the other hand, dephosphorylation (the removal of phosphate from a phosphorylated residue) is catalyzed by protein phosphatases. Protein phosphorylation allows cell to dynamically modulate protein functions [160] in response to extra- and intracellular cues, so as to elicit appropriate cellular behaviors (*e.g.* proliferation, differentiation, migration and apoptosis) needed for survival and proper morphological development.

The dynamic balance between protein phosphorylation (by protein kinases) and dephosphorylation (*e.g.* by protein phosphatases or through protein degradation) in response to combinatorial intra- and inter-cellular cues imposes dynamic structures in signaling networks which function as molecular switches or logic gates to regulate cellular activities [37]. Errors in protein phosphorylation or dephosphorylation often result in dysfunctional cellular processes leading to cancer and complex regulatory diseases [205]. Hyper-phosphorylated retinoblastoma protein, for instance, is linked to multiple cancers [87, 28, 22] while hypophosphorylated retinoblastoma protein is implicated in adult T-cell lymphoma progression [121]. In brief, identifying phosphorylated proteins and their sites of phosphorylation provides important clues for understanding cell biology.

Phosphoproteins are routinely detected by isotopic labeling using inorganic phosphate isotopes ^{32}P or ^{33}P. The isotopes, in the form of ATP, are often added in *in vitro* kinase assays to detect phosphorylation on purified proteins of interest. Antibodies

(Ab) that target phosphoresidues can be used in place of radioactive isotopes to detect phosphorylated proteins. However, these methods are typically not scalable for simultaneous detection of thousands of phosphoproteins needed for a system-level understanding of how protein kinases and phosphatases influence cellular behavior. A protein can be phosphorylated at multiple sites, and each site individually and in combination may have different functional consequences. Knowing the site of phosphorylation on a protein can help in deciphering the molecular effect of phosphorylation. It is, therefore, important to identify the exact site of phosphorylation on proteins. This can be achieved by Edman sequencing [40] and mutagenesis genetic experiments, but the techniques are generally time-consuming.

Various proteome-wide techniques have been developed to identify phosphoproteins and their phosphorylated sites. Among them, mass spectrometry has recently emerged as a popular choice for at least two reasons [138, 155]. First, it is a high-throughput (HTP) identification method that can detect thousands of phosphorylation sites in a single experiment [155]. Second, coupled with good separation techniques, it can detect phosphoproteins and their phosphorylated sites from complex samples such as cell lysate. The underlying principle behind phosphorylation site detection by mass spectrometry is that phosphorylated peptides produce unique tandem mass spectra (MS/MS) that help in their identification. First, proteins are broken down into peptides, typically by trypsin, which are then separated and ionized prior to m/z (mass-to-charge) measurement by mass spectrometer. The ionized peptides are subsequently isolated and broken down into ion fragments at amide bonds. From each ionized peptide, an ensemble of ion fragments each with different M/Z value is generated that collectively produce a unique mass spectrum useful for determining the sequence of the peptide [171]. A phosphorylated peptide will have a detectable 80 Da mass shift per phosphate in its mass spectrum over its unphosphorylated form due to the additional phosphate group. Hence, a phosphorylated peptide can be identified from its MS/MS spectrum. The exact site of phosphorylation on a phosphorylated peptide can also be determined from its mass spectrum albeit at a higher degree of error. Another common proteome-wide technique useful for identifying substrates of individual kinases is protein/peptide array (described in later section) [137] .

As a result of these proteome-wide techniques, there has been an explosion of known phosphorylation sites in the last five years or so. The first proteome-wide screen was carried out in *S. cerevisae* that identified 383 phosphosites [42]. Subsequently, proteome-wide screens for phosphorylation sites using phosphoproteome technologies were deployed on various human cell lines [129, 83], mouse [187], fly [203], plant [91] and bacteria [105]. To date, human-focused phosphoproteome studies have detected more than 50,000 phosphorylation sites of which at least half have not been detected in directed experiments (personal analysis). The identification of these phosphorylation sites provide new information that can enhance our understanding of how protein kinases and phosphatases regulate cellular processes and dictate cell behavior. In this chapter, I survey the various proteome-wide techniques used for profiling phosphoproteins and phosphorylation sites, and their applied biological studies. I also survey various reported evolutionary analysis of protein phosphorylation sites. Finally, in the last section, I highlight the challenges in exploiting the new information for understanding phosphoregulation of cellular activities, outline my research objectives and summarize my research work.

1.2 Technologies for high-throughput protein phosphoprofiling

1.2.1 Phosphopeptide enrichment techniques

Mass spectrometry has emerged as a popular method for identifying phosphoproteins and their sites of phosphorylation. The key steps in the identification process can be sum-

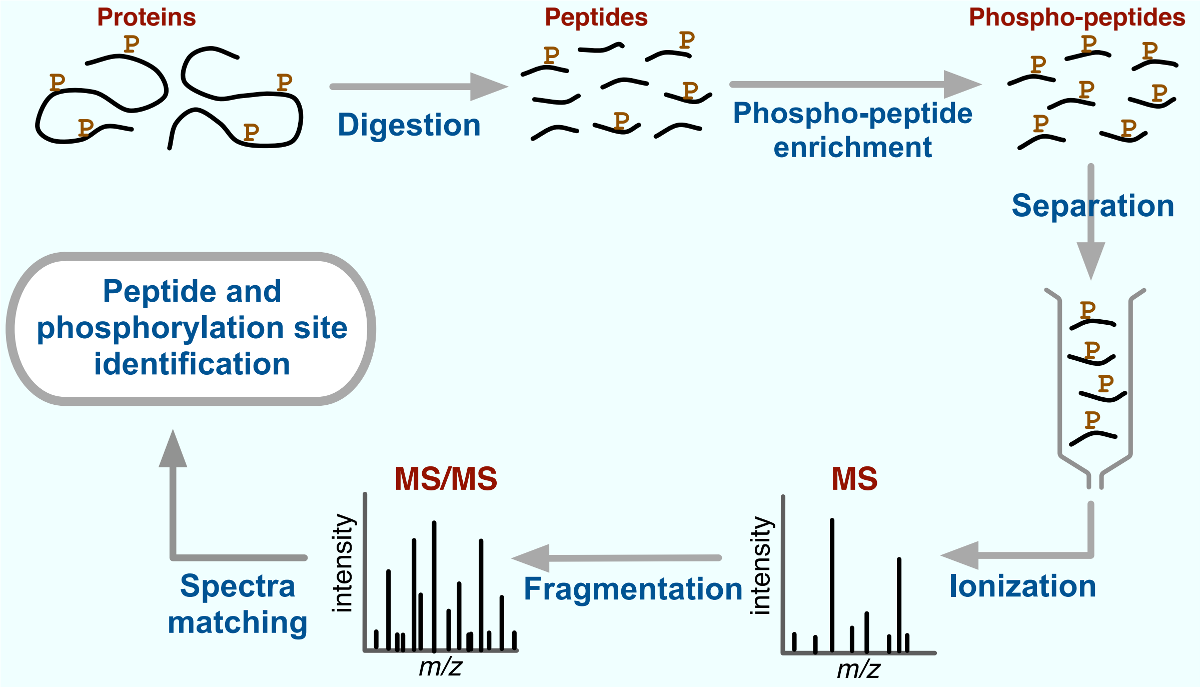

Figure 1.1: Main steps in the identification of phosphorylated peptides and phosphorylation sites by tandem mass spectrometry analysis.

marized as follows (Figure 1.1): (i) breaking down of proteins in a sample into peptides, typically by trypsin, (ii) enrichment of phosphorylated peptides, (iii) purification of peptides using separation techniques like high-performance liquid chromatography (HPLC), (iv) generation of mass spectra for separated peptides using mass spectrometer and finally, (v) matching of mass spectra to a spectra database to identify peptide sequences and its corresponding proteins.

A key advancement in recent years that greatly facilitates phosphopeptide detection by mass spectrometry is the development of phosphopeptide enrichment techniques. As many phosphoproteins are of low abundance, their detection by mass spectrometry is a challenge. In addition, only a few copies of a cellular protein may be phosphorylated at one time (phosphorylation stoichiometry) especially for phosphotyrosines. This can result in the signal from phosphopeptides being completely masked by its more abundant unphosphorylated forms. In many situations, the ability to identify minute phosphoproteins and phosphopeptides in a complex mixture (such as cell lysate) is desirable. Thus, in recent years, techniques for phosphopeptide enrichment prior to identification by mass spectrometry have emerged that greatly facilitate the identification of phosphorylation sites. Some of these techniques are:

Immobilized Metal Affinity Chromatography (IMAC)

One of the first phosphopeptide enrichment techniques applied in genome-wide detection of phosphorylation events is immobilized metal affinity chromatography (IMAC). The technique was first described in 1999 [136], and was subsequently adapted for large-scale detection of phosphorylation events in *S. cerevisiae* [42]. The technique exploits the affinity of phosphate for immobilized metal. However, unphosphorylated peptides that contain acidic residues (like glutamic and aspartic acids) can also bind to the immobilized metals. To prevent this, all peptides in a sample may first be converted to corresponding peptide methyl esters prior to phosphopeptide enrichment by IMAC. The enrichment technique has been coupled with liquid chromatography (LC)-tandem mass spectrometry (MS/MS) to detect phosphorylation events in cancerous cells [83, 118].

Strong Cation Exchange Chromatography (SCX)

Another chromatographic method used for phosphopeptide enrichment is strong cation exchange chromatography (SCX). The technique was first applied in a proteome-wide manner to detect phosphorylation events in a HeLa cell line [10] where proteins in a cell lysate were first separated by SDS-PAGE followed by in-gel trypsin digestion. This technique produces mostly peptides with +2 net charge at low pH. Singly-phosphorylated peptides, on the other hand, have a +1 net charge due to the presence of the negatively charged phosphate group. Phosphopeptides were then separated from non-phosphorylated peptides by SCX on the basis of charge difference. A total of 2002 sites from 976 proteins were identified in the experiment. It should be noted that the approach cannot detect doubly-phosphorylated peptides which have a net charge of 0 at low pH. It addition, it was found that phosphorylation sites with proline and acidiphilic amino acids in its flanking sequence make up a large fraction of the identified sites, suggesting a possible bias in the technique [10]. Like IMAC, the SCX had been applied on lysate from whole tissue, e.g. [9].

Antibody-based Enrichment Techniques

Antibodies (Ab) against phosphoresidues have also been used to enrich for phosphoproteins and phosphopeptides. In Grønborg *et al.* [54], proteins were immunoprecipitated with anti-*p*Ser Ab and anti-*p*Thr Ab followed by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) and nanoelectrospray tandem mass spectrometry to detect serine and threonine phosphorylation events. Other works used anti-*p*Tyr Ab to profile tyrosine phosphorylation in epidermal growth factor receptor signaling pathway [170] and in Jurkat leukemia T-cell line [151]. As tyrosine phosphorylation is a rare event compared to serine/threonine phosphorylation and that anti-*p*Tyr Ab generally has higher efficiency than anti-*p*Ser Ab and anti-*p*Thr Ab, enrichment techniques using anti-*p*Tyr Ab has emerged as a common experimental strategy to study phosphotyrosine signaling pathways [156]. In addition to antibodies targeting phosphoserine, phosphothreonine and phosphotyrosine generically, antibodies that target phosphoresidues with specific consensus sequence or motif had been used to identify proteins potentially phosphorylated by specific kinases or specific classes of kinases [54, 110].

Chemical Modification

Phosphorylated peptides can also be chemically modified to facilitate their enrichment. One approach involves replacing the phosphate moiety on phosphorylated peptides with biotinylated moieties [128]. The phosphorylated biotinylated peptides are then enriched with immobilized avidin and characterized by mass spectrometry. The shortcoming of this technique is it is not applicable to phosphotyrosine [53]. Another technique involves converting phosphopeptides into covalent tethers for attachment to a polyamine dendrimer [183]. The phosphopeptides with its high molecular mass dendrimer are then separated from non-phosphorylated peptides by size exclusion filtering.

Combinatorial Approach

Various phosphopeptide enrichment techniques have been combined to improve enrichment. For example, IMAC was used in conjunction with immunoprecipitation to profile tyrosine phosphorylation events in Jurkat cells [153] and in the interferon-α signaling pathway [206] using mass spectrometry. In another work, IMAC, SCX and immunoprecipitation were used in combination to detect phosphorylation events occurring in mouse liver [187]. Phosphopeptide modification by dendrimer coupled with immunoprecipitation has also been used to characterize tyrosine phosphorylation in T-cells [183].

1.2.2 Quantitative mass spectrometry techniques

The ability to quantify the amount of phosphoproteins and phosphorylation sites in addition to their detection provides further insight into how protein kinases and protein phosphatases regulate cellular activities. While there are methods to measure the absolute quantity of peptides (e.g AQUA [46]), they are presently not commonly used for quantifying phosphopeptides. Instead, techniques for relative quantification of phosphopeptides in one sample compared to the same phosphopeptides in another sample are more commonly used. The key step is often to differentially label proteins/peptides from different samples with isotopes or special chemical reagents such that the same phosphopeptide from different samples are of different mass that can be differentiated and quantified. Once samples are differentially labelled, they are mixed together before identification by mass spectrometry. While the term “quantitative phosphoproteome” is commonly used in literature, the quantitative measurements were often relative in nature.

When Forest White and his colleagues first applied the IMAC technique for detecting phosphorylation events [42], they also incorporated a quantitative approach in the technique. They showed that peptides of same sequence converted to methyl esters using methanol and deuterated methanol separately produces different mass spectra that is detectable by mass spectrometry. Another approach uses whole-cell stable isotope labeling by incorporating ^{15}N isotope into cell culture for quantification [127]. Here, I surveyed the more commonly used labeling methods.

1.3 Technologies for high-throughput protein phosphoprofiling

Stable Isotope Labeling by Amino acids in Cell culture (SILAC)

The SILAC technique, first described in 2002, involves growing cells in media with an isotopically-labeled form of an essential amino acid until the amino acid is incorporated into the proteome of the cell ([131], see Figure 1.2 for an overview). Isotope-labeled peptides produce different mass spectra compared to normal peptides. The method was adopted to elucidate the tyrosine phosphorylation state in HeLa cells at different time points after epidermal growth factor (EGF) stimulation [14]. To enable comparison of phosphorylation state across three different time points after EGF stimulation, three different isotope-labeled arginines ($^{12}C_6{}^{14}N_4$-Arg, $^{13}C_6{}^{14}N_4$-Arg, $^{13}C_6{}^{15}N_4$-Arg) are used. Cells grown in media separately with each of the isotope-labeled arginines were mixed and lysed. Tyrosine-phosphorylated proteins were then purified with anti-*p*Tyr Ab, digested and then identified/quantified with LC-MS.

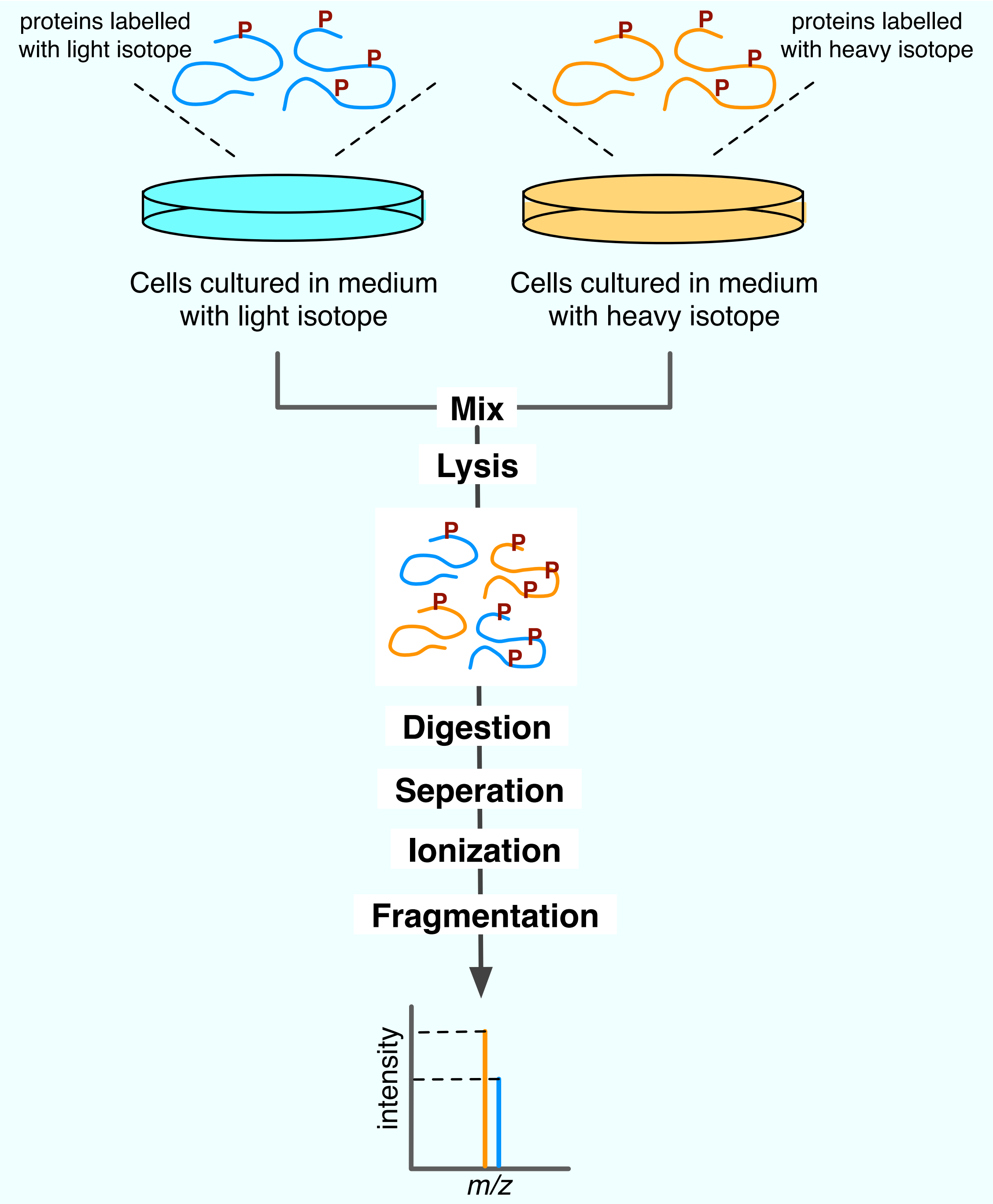

Figure 1.2: Stable Isotope Labeling by Amino acids in Cell culture (SILAC). Main steps in quantifying relative phosphorylation states of protein/peptide/site from two samples using SILAC. In isobaric tagging, peptides from different samples are labelled with different isobaric tags like iTRAQ reagents before they are mixed together.

Isobaric Tag for Relative and Absolute Quantitation (iTRAQ)

Rather than culturing cells to incorporate isotopes into their proteome, peptides can be directly labeled using iTRAQ reagents. Each iTRAQ reagent consists of a protein-reactive group that attaches to protein/peptides and a reporter group that fragments to produce different m/z value, allowing identification and quantification directly from the MS^2 spectrum [149]. There are up to eight different iTRAQ reagents, each producing different mass spectra for peptides of the same sequence. Thus, up to eight samples can be simultaneously assessed using iTRAQ. The technique has been applied to compare phosphorylation profiles of parental human mammary epithelial cells after 0, 5, 10 and 30 minutes of EGF simulation [204]. In this example, phosphopeptides were enriched with IMAC, and the temporal profiles of 104 phosphotyrosine sites on 76 proteins were generated.

1.3.1 Protein and peptides microarrays

Multiple putative protein substrates of protein kinases can be immobilized on a solid support, such as glass slide or streptavidin-coated membrane, as miniature protein array or proteome chip [104]. They can then be overlaid with isotope-labeled ATP and a protein kinase of interest to perform an *in vitro* kinase reaction assay [208, 207] . Phosphorylated substrates can be subsequently detected using high-resolution phosphorimaging [76]. This approach allows rapid screening for putative substrates of a specific kinase using a small amount of reagents and can be scaled up for proteome-wide assays [104]. For example, using microarrays containing 4400 unique *S. cerevisiae* proteins [104], Ptacek and colleagues tested 82 *S. cerevisiae* protein kinases and identified 4200 phosphorylation events on 1325 proteins [137]. In another study, potential substrates of Abl and Abl-related gene (Arg) tyrosine kinases were assessed using a microarray containing 2400 different human proteins [20].

Kinase-substrate interactions detected using protein microarray as described above may not occur physiologically due to lack of biological context, such as cellular colocalization and/or protein coexpression between kinases and their detected substrates. Moreover, high kinase concentration, which does not reflect physiological or cellular levels, is often used to increase sensitivity of these assays. In addition, many physiological kinase-substrate interactions can be missed because contextual factors like adaptor proteins or coactivators (e.g. cyclins) and priming phosphorylation sites or kinases are often not present in the arrays. Nevertheless, a protein microarray assay serves to identify potential kinase-substrate relations that can either be validated through downstream biochemical and genetic experiments or corroborated with biological data from other studies.

Peptides with a phosphoacceptor residue at a fixed position can be immobilized on a chip just like full-length proteins, and subsequently incubated with isotope-labeled ATP and kinase of interest, followed by phosphorimaging to identify phosphorylated peptides. Peptides spotted on microarrays can be random sequences [152], from degenerate oriented peptide libraries [167] or are subsequences found in proteins [52, 94, 89]. If peptides corresponding to subsequences in proteins are used, the precise sites of phosphorylation on substrates can be determined. However, both false positive and false negative phosphorylation sites can be detected by this approach, as the 3-D structural context of the phosphoacceptor residues which can affect phosphorylation are not represented in the assays. On the other hand, if peptides of random sequence or degenerate oriented peptide libraries are used, the observed phosphorylated peptides can be used to derive a position-specific scoring matrix (PSSM), as a statistical model, to quantify the phosphorylation propensity of a phosphoacceptor residue based on amino acids flanking the residue [114, 200]. The PSSM can then be used to scan a proteome to identify putative substrates for the kinase assayed [114, 200]. Exact sites of phosphorylation are predicted using this bioinformatics approach although the approach is known to have high false positive rate presumably because non-naturally occurring peptides devoid from the content/context of whole protein sequences were used.

A challenge for peptide chips based on random sequence peptide is the huge surveyable peptide sequence space. An alternate approach is to incubate the kinase of interest and isotope-labeled ATP in solution with a large set of fixed-length peptides of different sequences. To facilitate the identification of sequence patterns needed for phosphorylation by a protein kinase of interest, peptides can be divided into pools such that peptides in each pool match a consensus pattern. Such an approach was taken by Cantley, Yaffe, Turk and coworkers to determine the sequence specificities of serine/threonine kinases [200, 68, 117]. Unique peptide pools were generated such that in each pool, all peptides have a common amino acid at one of the residue position while amino acids in other positions are degenerated. In Turk's approach, a total of 198 unique peptide pools were generated as phosphothreonine, phosphotyrosine and the 20 naturally occurring unmodified amino acids were individually fixed at each of the nine residue positions flanking a central phosphoacceptor residue. Each peptide in the pools is biotin-tagged, allowing the peptide to be spotted onto a streptavidin-coated membrane. Phosphorylated peptides can subsequently be detected by autoradiography or phosphorimaging. Amino acids preferred by the kinase of interest at each position flanking the phosphoacceptor residue can then be determined. In addition, the fixation of a phosphorylated residue at one of the positions flanking a phosphorylated phosphoacceptor residue allows the detection of phosphorylation that requires priming phosphorylation sites [143].

1.4 Applied proteome-wide phosphoprofiling of biological systems

The experimental techniques and technologies described in the previous section had been adapted to study signal transduction, in particular to identify the phosphorylation sites targeted by different kinases. Here, we surveyed some of the strategies adopted toward this goal.

1.4.1 Perturbation-based assays

Protein or peptide chips can be applied to identify putative substrates of kinases. However, they are *in vitro* experimental techniques that do not necessarily capture the physiological/cellular concentration and co-localization factors of protein kinases and their substrates [98]. Instead, Smolka *et al.* [166] combined quantitative MS techniques with perturbation studies to identify cellular phosphorylation sites and substrates of yeast DNA damage checkpoint kinases Mec1/Tel1 and Rad53 upon induction of DNA damage. Quantitative MS techniques can detect sites that are differentially phosphorylated across two or more cellular conditions/perturbations but do not directly identify their effector kinases. However, by detecting phosphorylation sites that were specifically altered between kinase-null (Mec1/ Tel1 and Rad53) and wild-type *S. cerevisiae*, Smolka *et al.* identified 62 putative target sites of Mec1/Tel1 and Rad53 on 55 proteins. These differentially phosphorylated sites were enriched in the known phosphorylation motifs (linear motifs) of Mec1/Tel1 and Rad53, which further suggest that many identified targets are physiological substrates of the kinases.

A quantitative MS approach was also adopted to identify proteins in zebrafish Fyn/Yes morpholino knockdown embryos that were differentially phosphorylated compared to those in wild-type embryos [93]. Using similar approaches, Matsuoka *et al.* identified putative ATR and ATM phosphorylation sites that were altered upon DNA damage in human embryonic kidney 293T cells, and thus, should correspond to physiological targets of the two kinases [110]. Putative phosphorylation sites of ATR and ATM were identified using antibodies against *p*SQ or *p*TQ sites that are known to be targeted by ATM and ATR kinases. Among the phosphorylation sites detected, 905 sites from among 700 proteins were found up-regulated fourfold after induction of DNA damage, of which

55 sites were found on 31 ATR and ATM substrates known to be implicated in DNA damage signaling. Hence, the 700 proteins are possible physiological substrates of ATM and ATR. The accuracy and scalability of this approach depend on the availability of suitable antibodies and their qualities. In addition, the specificity of the antibody is an important consideration in such assays, as many protein kinases are known to have similar specificities.

A potential pitfall in the above-mentioned approaches is that not all perturbed sites identified are direct targets of the deleted kinase. Instead, they could be targets of other kinases that are activated downstream of the deleted kinases in signaling cascades. For example, in Smolka *et al.* [166] , about half of the sites down-phosphorylated in Mec1/ Tel1 mutant are also down-phosphorylated in a Rad53 mutant. As these sites express Rad53 phosphorylation motifs and given that Rad53 acted downstream of Mec1/ Tel1, they are likely targets of Rad53.

1.4.2 Chemical-genetics approaches

As mentioned above, a challenge in perturbation approaches is that one cannot always be certain which kinase(s) phosphorylated the observed altered sites as many protein kinases could share similar consensus motifs or targets. By structural alteration of a kinase (through mutagenesis) such that it can incorporate a specific modified form of ATP, a detected phosphorylated protein containing the modified ATP is most likely targeted by the mutant kinase [162, 102, 39]. Cells (NIH 3T3) with such analogue-sensitive (AS) mutants of v-Src kinase were generated, lysed and incubated with analog ATP to identify putative substrates of v-Src with *in vivo* concentration of proteins [163]. An AS mutant of Pho85 kinase was generated similarly and assayed for putative substrates in whole-cell extracts of *S. cerevisiae* [34]. A similar approach was combined with computational search by Ubersax *et al.* to identify *in vivo* substrates of Cdk1 [70]. The phosphorylation detection was performed on potential substrates expressing a Cdk1 consensus phosphorylation motif (S/T-P-x-K/R) using cell lysate incubated with purified AS mutant Cdk1 with analog ATP to identify 181 proteins that were efficiently phosphorylated by the AS mutant Cdk1. To validate some of the identified substrates *in vivo*, small molecules that inhibit the AS mutant Cdk1 were added to cultures of *S. cerevisiae* AS mutant Cdk1 strains to detect proteins with decreased phosphorylation as likely targets of Cdk1 *in vivo*. A total of 12 high confidence *in vivo* substrates of Cdk1 were identified in this manner. Similar approach was adopted by Holt *et al* [62] to identify phosphorylation sites targeted by Cdk1 during cell cycle. Two issues with these approaches are that it can not be ruled out that the ATP analog may be picked up and utilized by other distantly related kinases, and that the structural change in the AS kinase may alter its *in vivo* specificity. Another concern is the scalability of the assay for proteome-wide studies as it is unclear presently whether AS mutants can be created for most kinases.

1.5 Computational analysis of phosphoproteomic data

Recent advancement in technologies described now permits the identification and quantification of thousand of phosphorylation sites in a single experiment. The relative differences in phosphorylation level of multiple sites between cell samples subjected to different treatments, or at different time points after treatment can now be surveyed in a high throughput manner. Computational data analysis and modeling approaches are needed to organize and interpret the large datasets of site- and context-specific *in vivo* phosphorylation events assembled in various HTP phosphoproteomic studies. One of the key challenges is to delineate detected phosphorylation sites to their effector kinases. This is important for inferring the kinase-substrate interaction networks that are essential for mechanistic understanding of cell behavior and for therapeutic intervention [78, 133]. Here, I survey some of the computational data analysis and modeling approaches that have been used to analyze large-scale phosphorylation data sets (see Figure 1.3 , and

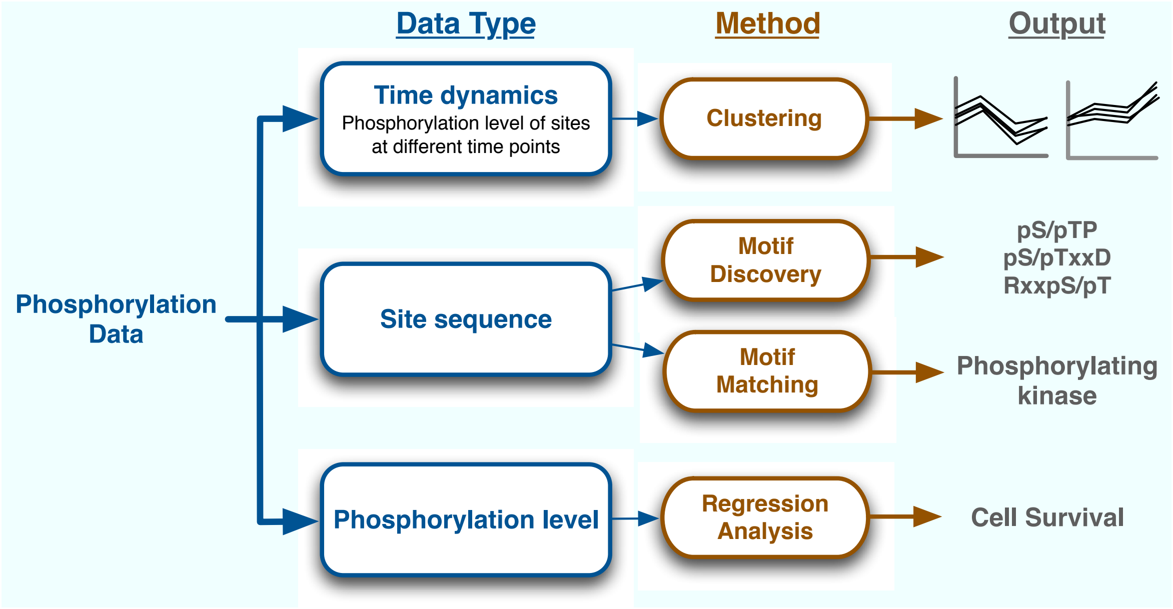

Figure 1.3: An overview of computational analysis performed on large scale phosphorylation data.

the computational tools for infering transient kinase-substrate interaction networks.

1.5.1 Clustering of phosphorylation sites with similar temporal profiles

Surface receptor kinases belong to a class of protein kinases that assimilate extracellular signals to initiate appropriate cell behaviors. Proteome-wide studies of phosphorylation events initiated by one such receptor kinase, epidermal growth factor (EGF) receptor, have been conducted where sites differentially phosphorylated at time points 1, 5, 10 and 20 min after EGF activation in HeLa cells were probed using quantitative phosphoproteomic techniques [129, 13]. These studies revealed the dynamic temporal nature of protein phosphorylation in which many sites are either up- or down-phosphorylated at different times after EGF stimulation. Phosphorylation sites with similar temporal profiles were grouped using Fuzzy c-means (FCM) clustering to facilitate biological analysis and interpretation [129]. Unlike other hard-partition methods like *k*-means and self-organizing map (SOM), FCM allows an instance to belong to different clusters with different scores that add up to 1.

In a similar work, the temporal dynamics of a large number of tyrosine phosphorylation sites were analyzed at four time points (0, 5, 10 and 30 min) after 5, 10 and 30 min incubation of human mammary epithelial cell line with 25nM EGF using untreated cells as a control [204]. SOM with U-matrix method [186] of visualizing was used to identify coregulated phosphorylation sites. In essence, SOM is a technique for mapping high-dimensional data (in this case, each phosphorylation site has 16 features corresponding to the four time points under the four different treatments) to lower dimension, often 2-D, that facilitate manual grouping of clusters by visual inspection. An advantage of SOM is it allows an overview of similarity between clusters.

1.5.2 Regression analysis of phosphorylation data to predict cell fate

Partial least-square regression (PLSR) is a class of regression technique that combines a data compression technique (through principal component analysis) with regression to predict dependent variables using input variables from limited samples. In situations where the number of input variables exceeds the number of observations, or when the input variables exhibit multi colinearity (meaning some variables are highly correlated), or when there are missing data, PLSR is an adequate choice for regression analysis over other conventional regression techniques. In a landmark paper, PLSR was used to predict cell fate of individual HT-29 cell with high accuracy after stimulation by combination of three cytokines (tumor necrosis factor, EGF and insulin). The inputs to the PLSR are the quantitative experimental readouts of 11 signaling proteins at multiple time points after cytokine stimulation [71]. The work highlights the value of including the dynamic response of signaling molecules upon cytokine stimulation in regression analysis, as using only the input cytokine concentration failed to correctly predict cell fate. PLSR was also applied to correlate the temporal dynamics of phosphotyrosine sites with migration and proliferation cell behaviors mediated by the ErbB2 family of tyrosine receptors [195, 88]. Other applications of PLRS include correlating multiple signaling events to observed cell behaviors [115, 52]. Most importantly, these studies showed that apparently genetically identical cells can react differently to the same stimulus depending on the physiological state of the proteins in the cells.

1.5.3 Site prediction and motif discovery from phosphorylation sites

Another common data analysis strategy is the identification of over-represented sequence patterns among detected phosphorylation sites that may correspond to phosphorylation motifs of some kinases. This strategy can be used to detect novel phosphorylation motifs of uncharacterized kinases [80] or novel binding motifs of phosphoresidue-binding domains like SH2 and PTB [160]. Although phosphorylation or binding motifs could be determined by *in vitro* methods such as protein microarray and degenerate oriented peptide libraries (see below), these experiments are conducted *in vitro* and thus may not fully reflect *bona fide* motifs. Motif discovery tools such as Gibbs motif samplers [92], MEME [7], PRATT [77], TEIRESIAS [146] and D-STAR [181] can be used to discover motifs [123] from sets of phosphorylation sites determined in phosphoproteomic experiments. For example, PRATT was previously used to identify phosphorylation motifs of kinases from the sequences of substrates detected in protein chip experiments [137].

Many generic motif discovery algorithms do not explicitly correct for unbalanced distribution of amino acids found in proteins that contribute to spurious motifs. To address this shortcoming, MotifX, a recent motif discovery algorithm, incorporated background frequencies of amino acids in proteins to improve the extraction of phosphorylation motifs from phosphoproteomic data [158]. MotifX was applied in various studies to identify known and novel motifs in mammalian species [165, 132, 8] and in Arabidopsis [175].

Motif patterns extracted by MotifX are restricted with either all amino acids or a single amino acid at each position, thus motifs with degenerate positions like the *p*Y-x-x-[LIV] motif of JAK2 kinase [4] were excluded. Moreover, the greedy iterative nature of the algorithm could potentially exclude the discovery of some motifs. MoDL is a motif discovery algorithm created to extract degenerated motifs found in phosphorylation data [148] using the principle of minimum description length.

Motif extraction from phosphorylation sites detected in HTP phosphoproteomic studies coupled with downstream experimental validation could lead to discovery of novel *in vivo* phosphorylation motifs of protein kinases and phosphoresidue-binding domains. This is exemplified in Miller *et al.* [113] where a novel binding motif for a SH2 domain in inositol 5-phosphatase 2 (SHIP2) was discovered. In the work, 481 unique tyrosine-phosphorylated peptides detected by tandem MS experiments in mammalian cell lines were grouped into 20 clusters, followed by motif extraction using TEIRESIAS [146]. A novel N-terminal hydrophobic motif [DE]-x-xx-[ILV]-[ILV]-pY was extracted from one of the clusters, in which three out of the four peptides expressing the motifs were validated to bind SHIP2 in pull-down assays. Mutational analysis on two amino acid positions immediately N-terminal to the phosphotyrosine confirmed the generality of the motif. Interestingly, proteins expressing the motif are enriched with cell surface receptor linked signal transduction function, in agreement with known association of SH2-containing protein with receptor-linked signaling. The work is probably the first system-wide approach that combined both bioinformatics analyses and experimental validation to discover novel motifs.

1.5.4 Computational identification of protein kinases targeting MS-derived phosphorylation sites

One of the key challenges is to delineate MS-derived phosphorylation sites to their effector kinases. Here, I survey the computational methods and tools that have been developed or conceptually can be used for this purpose. Simple consensus motif searching using known phosphorylation motifs of kinases can be used to associate MS-identified phosphorylated sites or proteins to their effector kinases [166, 185, 110]. However, relying on simple regular expression search can be highly unspecific [114]. Thus, several computational methods had been developed to better identify potential targeting kinases (or kinase family) of MS-identified phosphorylation sites.

Machine-learning approaches

A subset of these tools deployed machine-learning algorithms to predict novel phosphorylation sites. The basic methodology involves training machine-learning models using known positive and negative examples of sites phosphorylated by kinases of interest, and then testing the capability of the models to differentiate both positive and negative samples in separate data sets. The resulting computational models can subsequently be applied to new data to predict potential phosphorylation sites of specific kinases. Support vector machines, a statistical machine learning method, have been used in KinasePhos [196] and PredPhospho [84] for predicting kinase-specific phosphorylation sites. Similarly, artificial neural networks and Bayesian Decision theory were employed in NetPhosK [60, 15], GANNPhos [182] and PPSP [198] to predict kinase-specific phosphorylation sites. MetaPredPS [192], a meta-predictor, combined predictions from GPS [199], KinasePhos, PPSP, PredPhospho and Scansite [200] through a generalized weighted voting strategy to improve prediction for phosphorylation sites targeted by four protein kinase families (CDK, CK2, PKA and PKC).

Similarity based approaches

Alternative approaches have been adopted to predict kinase-specific phosphorylation sites: GPS 2.0 predicts kinase-specific phosphorylation sites in a query sequence based on the sequence similarity to known sites of kinases [199]. To improve prediction performance, a derivative of the BLOSUM 62 substitution matrix was derived for each kinase group to optimize similarity comparisons between the sites. Predikin employs sequence-structure analysis of protein kinases to infers phosphorylation motifs for uncharacterized serine/threonine kinase sequences submitted by user [21]. The pkaPS method uses a simplified analytical model to score physical and chemical requirements at amino acid positions from 18 to 123 of phosphoacceptor residues to predict putative phosphorylation sites of protein kinase A (cAMP-dependent kinase, PKA) [126].

Contextual modeling of kinase specificity – NetworKIN

Computational and *in vitro* experimental detection of kinase substrates and their phosphorylation sites often omit contextual factors like subcellular compartmentalization and differentiated protein expression that can prevent phosphorylation. In addition, positive factors that coregulate phosphorylation such as colocalization via anchoring proteins, scaffolds and substrate capture by non-catalytic interaction domains and docking motifs are typically not captured in these experiments. These factors, in combination with the challenges of mapping transient and context-dependent kinase-substrate interactions using current protein-interaction assays, have in part led to a large gap between the understanding of *in vivo* phosphorylation sites and the kinases that modulate them. Currently, in the Phospho.ELM database [36], there are thousands of annotated *in vivo* phosphorylation sites, of which only about 25% have been linked to at least one *in vivo* kinase [99]. To address this problem, the NetworKIN algorithm was developed to predict *in vivo* kinases for identified phosphorylation sites [98]. The principle behind this algorithm is to model kinase specificity using contextual information for phosphoproteins and kinases in combination with sequence models of kinase consensus motifs [114]. By combining probabilistic modeling of network context with the linear motifs recognized by the catalytic kinase domain, it has been shown that NetworKIN can assign a specific kinase to an observed *in vivo* phosphorylation site with a 2.5-fold higher accuracy than previous methods such as Scansite and NetphosK.

The human kinome specificity atlas – NetPhorest

NetPhorest is a database containing specificity motifs of protein kinases and phosphoresidue-binding protein domains derived using peptide arrays. The database currently contains consensus motifs for 179 human protein kinases and 104 human SH2 and PTB domains. It also consists of an ensemble of probabilistic classifiers for inferring which protein kinase or phosphoresidue-binding protein domain likely targeted experimentally observed phosphorylation sites. Hence, predictors in NetPhorest are unlike existing predictors that were developed to predict novel phosphorylation sites of kinases or novel binding sites of phosphoresidue-binding protein domains. NetPhorest has a framework to automate data set construction and training of sequence models for linear motifs involved in phosphorylation mediated signaling.

1.6 Evolutionary and functional analysis of phosphoproteomic data

As I seek to study the evolutionary dynamics of protein phosphorylation for interpreting the importance and functions of newly discovered protein phosphorylation events, here, I survey related work on the conservation of phosphorylation sites.

At the level of phosphoprotein conservation, Mann and colleagues [50] reported that phosphoproteins are more likely to have homologs in other eukaryotes than proteins not known to be phosphorylated, based on the phosphoproteomes identified by mass spectrometry in 5 eukaryotes (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus*). Higher conservation of phosphoproteins over non-phosphorylated proteins are also reported for prokaryotes for MS-identified phosphoproteins in *Escherichia coli* and *Bacillus subtilis* [107]. However, in another study, Mann and colleagues reported that phosphoproteins identified in 4 prokaryotes (*Escherichia coli*, *Bacillus subtilis*, *Lactococcus lactis* and *Halobacterium salinarium*) are rarely homologous to each other and are relatively sparse compared to eukaryotes [50]. Conservation analysis of phosphoproteins in *Saccharomyces cerevisiae* also revealed that protein phosphorylation may have been a factor influencing the retention of duplicated genes after WGD (Whole Genome Duplication) in yeast [82].

In addition to higher conservation of phosphoproteins reported for both prokaryotic and eukaryotic species, phosphorylation sites have been reported to be more conserved than their non-phosphorylated counterparts. Gnad *et al.* reported higher conservation of fly's phosphoserines in human over other serines while Malik *et al.* [107] observed overall higher conservation of human phosphorylation sites (without distinguishing between *p*S, *p*T and *p*Y) in rat, mouse, cow, chicken, zebrafish and Xenopus for 1744 phosphorylation sites identified on mitotic spindles isolated from cultured human cells. Similarly, phosphorylation sites in *Escherichia coli* and *Bacillus subtilis* are reported to be more conserved than other phosphorylatable residues in various species, although statistical significance could not be established due to the small number of phosphorylation sites. However, conservation rate of phosphorylation sites has been reported to be similar to other solvent accessible phosphorylatable residues [75]. It should be noted that in this particular study, conservation of phosphorylation sites and phosphorylatable residues were not computed between homologous sequences and protein structures from across diverse species. Observations from such a conservational analysis approach can therefore be skewed by varying divergence rates of different lineages. In addition, occurrence of most known phosphorylation sites in unstructured regions in proteins [75, 90, 62, 159], which in general are evolving faster than structured protein regions [23], can contribute to the lower conservation rate observed for phosphorylation sites over other phosphorylatable residues [90]. Boekhorst *et al.* [18] compared phosphorylation sites from six eukaryotes to identify conserved phosphorylation events occurring at similar positions across homologous proteins, as determined from sequence alignments, and found the overlap to be statistically significant.

While many studies analyzed the conservation of large sets of MS-identified phosphorylation sites, a few focused on the conservation of phosphorylation sites with characterized functions. Conservational analysis of a set of 249 functionally characterized phosphorylation sites in *Saccharomyces cerevisiae* across *Saccharomyces bayanus*, *Saccharomyces paradoxus*, and *Saccharomyces mikatae* revealed that phosphorylation sites are generally more conserved than the residues flanking them on primary sequences [6]. Looking specifically at phosphorylation sites demonstrated experimentally to be targeted by CDK1 (cyclin-dependent kinase 1) [185], Ba and Moses further observed that the residues flanking the phosphorylated residues that are known to influence phosphorylation by CDK1 are more conserved than other flanking residues [6]. Comparing functionally characterized to uncharacterized MS-identified phosphorylation sites on mitotic spindles, Malik *at al.* [107] found that the former are significantly more conserved than the latter but noted that potential preferences by experimentalists to study well-conserved phosphorylation sites may have biased the observation. Similarly, Landry *et al.* [90] observed that phosphorylation sites characterized to have functional roles as annotated in HPRD (Human Protein Reference Database) are more conserved than phosphorylation sites identified in large scale phosphoproteomic studies.

Many phosphoserines/threonines expressing similar consensus motifs were found located in close proximity to each other as cluster on primary protein sequences, a phenomenon that is not observed for phosphotyrosines [159]. Many proteins targeted by CDK1 are known to contain clusters of CDK1 consensus phosphorylation motifs that are observed across orthologs in different numbers and at different positions [120, 62]. Such phenomena have been exploited to improve identification of substrates targeted by specific kinases [119, 6, 25]. In a systematic analysis on conservation degree and phosphorylation likelihood, Budovskaya *et al.* observed that proteins in which PKA (Protein

Kinase A) consensus motifs are conserved over longer evolutionary time are more likely to be targeted by PKA based on *in vitro* assays [25].

Tracing CDK1 consensus motifs in known CDK1 targets and non-targets with their inferred ancestral sequences across *Saccharomyces cerevisiae*, *Saccharomyces bayanus*, *Saccharomyces paradoxus*, and *Saccharomyces mikatae*, Ba and Moses concluded that the CDK1 consensus motifs are evolutionary conserved on *bona fide* targets of CDK1 [6] which is not unexpected. However, they also observed constrained appearance of CDK1 consensus motifs in CDK1 *bona fide* targets compared to non-targets. The authors reasoned that this can arise if new CDK1 sites disrupt functions of CDK1 targets while the appearance of CDK1 consensus motifs on non-targets are not evolutionarily constrained because CDK1 does not target these sites[6]. Correlating microarray expression data across *Homo sapiens*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Arabidopsis thaliana* with phosphorylation data, Jensen *et al.* observed that although periodically expressed and constitutively expressed subunits in evolutionary conserved cell cycle protein complexes differ considerably among the four species, protein phosphorylation occurs preferentially on periodically expressed proteins in each species[73]. In a comparative analysis of phosphoproteomes across three yeast species (*Saccharomyces cerevisae*, *Candida albicans* and *Schizosaccharomyces pombe*), it was observed that the intensity level of phosphorylation is highly conserved within different cellular activities although the intensity can vary considerably among individual proteins within each functional group across the three species [11].

1.7 Thesis summary

Mass spectrometry and related technologies have unveiled many novel phosphorylation sites that can potentially provide insight to the regulation of cellular activities. However, many phosphorylation sites will need further experimental characterization to elucidate their functional roles. Following the rationalization behind the term "functional genomics", a field that seeks to characterize the function of each gene, I termed this research endeavor broadly as "functional phosphoproteomics". Some key issues can slow down the discovery process. First, there could be many false positives in the present set of phosphorylation sites from large-scale mass spectrometric screens [125, 35]. Second, some phosphorylation events may be silent (as coined by Philip Cohen [31]) that, although genuine, have no or little functional consequences. Lastly, the identification of promising targets for studies from the large list of phosphorylation site is a daunting task for biologists.

Hence, for my thesis research, I investigated the evolution of experimentally determined protein phosphorylation sites with the objective to assess the utility of sequence conservation profiling for interpreting the importance and function of uncharacterized phosphorylation sites. Specifically, I asked whether sequence conservation can be used to identify functional phosphorylation sites. To begin, I identified a set of human phosphorylation sites in which similar positions in orthologous proteins, as determined by sequence alignments, are phosphorylated in fly, worm or budding yeast. I subsequently investigated the sequence conservation profiles of residues with such conserved phosphorylation events (Chapter 2). Some phosphorylation sites may have been identified with the help of sequence conservation analysis. To exclude these sites in our analysis, we specifically obtained phosphorylation sites identified in untargeted proteomic-wide screens for our analysis. In the same work, I also explored and devised a non-alignment approach to identify phosphorylation events on orthologous proteins that are likely mediated by orthologous protein kinases but may not occur at similar sequence positions across orthologous proteins.

Next, I extended my conservational analysis to a larger set of human phosphorylation sites (Chapter 3) for which I do not have data to validate that similar positions of these sites in orthologous proteins are also phosphorylated. I reasoned that the larger dataset and rich annotation available for some of these sites could provide insight into the evolution of phosphorylation sites. Specifically, for this part of my work, I seek to understand how various factors, such as site prevalence (as gauged by its detection frequency across multiple studies), stoichiometry and occurrence in disordered/ordered protein regions influence its sequence conservation. I also seek to know whether sites with characterized functions exhibit unique sequence conservation patterns. In the last part of my work, I analyzed the maximum phosphorylation propensity (as gauged by frequency of phosphorylatable residues) encoded in the proteomes of various metazoan species with the aim of understanding how it might have shaped the frequency and conservation patterns observed for the various phosphorylatable residues (Chapter 4). In the last chapter, I summarized my findings and perspectives, and proposed future research endeavors that may help identify more important phosphorylation events and uncover their functions.

Chapter 2

Comparative phosphoproteomics to identify evolutionary conserved phosphorylation events

Work presented in this chapter was published as:

C. S. Tan*, B. Bodenmiller*, A. Pasculescu, M. Jovanovic, M. O. Hengartner, C. Jørgensen, G. D. Bader, R. Aebersold, T. Pawson, R. Linding, *Sci Signal.*, 2(81):ra39, 2009 Jul.

I performed all the computational analysis in the paper except for 1) Section 2.2.10 and the associated analysis presented in Figure 2.12, which were performed by A. Pasculescu, 2) NetworKIN and NetPhorest prediction in Section 2.2.8, which were generated by R. Linding, and 3) the generation of phosphorylation site data described in Section 2.2.2, which was carried out by B. Bodenmiller and M. Jovanovic.

* denotes co-first authors

2.1 Introduction

Protein kinases recognize and phosphorylate linear motifs in proteins [114]. These molecular events can directly control the activities of other proteins and the dynamic assembly of directional protein-protein interaction networks. In combination with phosphatases, kinases regulate the phosphorylation dependent binding of linear motifs to modular protein domains, such as the Src homology 2 (SH2) domain that recognizes phosphorylated tyrosine motifs and the BRCA1 C-terminal (BRCT) domain that recognizes phosphorylated serine and threonine motifs, and thereby create logic gates [38, 12] that enable the cell to swiftly and precisely respond to both internal and external perturbations [160, 115]. Although interaction maps [150, 45, 172, 191] provide useful information, it is the network dynamics and utilization that mediate cellular processing of environmental cues [65, 72]. Quantitative mass spectrometry (MS) measurements of phosphorylation networks and their dynamics are now rapidly unraveling thousands of cellular phosphorylation sites [194, 10, 17, 110, 147, 129]. With the functional and phenotypic characterization of previously unknown sites lagging behind their detection, a systematic way to highlight and prioritize important phosphorylation events is needed to guide functional experimental studies.

In addition, the conservation and evolutionary trace of most sites remain largely unknown. Unlike protein domains, which are conserved over long evolutionary distances, phosphorylation motifs are short and often reside in disordered fast-evolving regions [100, 139, 124, 120, 73]. These properties render phosphorylation sites difficult to align and trace evolutionarily [75, 107, 105, 18]. Here, I assembled human phosphorylation sites previously identified in both large scale MS [high throughput (HTP)] and low-throughput (LTP) targeted experiments [36, 64] and explored their conservation with the phosphorylated proteins (phosphoproteomes) of three target model organisms (fly, worm, and yeast) that were measured with a similar experimental and computational pipeline. Through a combination of sequence-alignment and reconstructive, network-alignment approaches, I investigated the conservation of protein phosphorylation events at two distinct levels: sites that are conserved at similar positions (termed positionally conserved) in orthologous proteins between human and at least one target species (such sites are termed “core sites” for the purpose of communication in this work) and those that are involved in conserved kinase-substrate regulatory networks but that are not necessarily constrained to the same location within phosphoproteins from humans and the model organisms (such proteins are termed ”core net proteins” in this work).

To identify human sites that are conserved in distantly related model organisms and thereby likely to be important for fundamental cellular activities, I first identified positionally conserved sites with a full-length (global) sequence-alignment algorithm [81] to map the experimentally identified phosphorylation sites from the target species (*Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*) to orthologous human phosphoproteins (Figure 2.1). This approach led to a conservative assessment of conserved sites because it requires the position of a site to be fixed within a multiple-sequence alignment. However, kinases can regulate cellular activities in ways that do not require their sites to occur at precise positions in protein sequences [123, 120, 73, 62], as is the case in the threshold dependent regulation of the Sic1 protein [122], for which phosphorylation at each of several sites promotes binding to Cdc4. Similarly, the ultrasensitive inactivation of Wee1 kinase is mediated by cyclin-dependent kinase 1 (Cdk1) decoy sites in both Wee1 and other proteins that distract CDK1 away from the causal sites in Wee1 [85]. Therefore, we aimed to identify conserved human phosphorylation events that are not necessarily conserved at the same sites between orthologous kinases and substrates in the target species by deploying the NetworKIN [98] algorithm in combination with NetPhorest [114] to infer the relevant protein kinases for substrates identified in the phosphosphoproteomes of human and each target species. The computationally reconstructed human kinase-substrate network was subsequently overlaid with that of the target species to identify conserved kinase-substrate relationships.

By taking two distinct approaches to assess phosphorylation conservation, we provide insight into the evolution of phosphorylation-based regulation with potential impact for our understanding of normal biological processes and complex diseases.

2.2 Materials and methods

2.2.1 Assembly of non-redundant human phosphorylaton data

Human phosphorylation sites were collected from the two major online databases PhosphoSite (release 2.0 [64]) and Phospho.ELM (release 7.0 [36]). As the two databases use protein sequences from different releases of SwissProt to track the positions of phosphorylation sites, all data were mapped into a reference sequence set from Ensembl (release 46, 2007 [43]). This helped to resolve cases where identical sites had different positions due to revisions of the SwissProt sequence referenced and to identify and remove redundant sites. The mapping between SwissProt primary accessions and its corresponding Ensembl human protein identifiers (release 46) was obtained from Ensembl through its BioMart interface. Finally, the positions of the phosphorylation sites in the Ensembl protein sequences were identified by exact string matching (using the peptide from -7 to +7 surrounding the phosphorylated central residue as defined in the Phospho.ELMor PhosphoSite database). This procedure resulted in 23,977 nonredundant (at 100% identity level) human phosphorylation sites for the comparative analysis.

2.2.2 Generation of phosphorylation data in fly, worm and yeast

Phosphorylation sites in *D. melanogaster*, *C. elegans* and *S. cerevisiae* were identified using mass spectrometry by our collaborators, Bernd Bodenmiller and Marko Jovanovic, from the University of Zurich, Switzerland. Here, I described the experimental procedures they adopted.

Generation of peptide samples

The *D. melanogaster* phosphorylation data was generated as follows: Kc167 cells were grown in Schneiders Drosophila medium (Invitrogen) supplemented with 10% fetal calf serum, 100 U penicillin (Invitrogen) and 100 g/ml streptomycin (Invitrogen, Auckland, New Zealand) in an incubator at 25°C. To increase the number of mapped phosphorylation sites, different batches of cells were pooled. Cells were either: 1) grown in rich medium, 2) serum-starved, 3) treated for 30 min with 100 nM Rapamycin (LClabs, Woburn, MA, USA) in rich medium, 4) treated for 30 min with 100 nM insulin (serum starved), or 5) treated for 30 min with 100 nM Calyculin A (rich medium). Then the cells were washed with ice-cold phosphate-buffered saline and resuspended in ice-cold lysis buffer containing 10 mM HEPES, pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM dithiothreitol and a protease inhibitor mix (Roche, Basel, Switzerland). To preserve protein phosphorylation, several phosphatase inhibitors were added to a final concentration of 20 nM calyculin A, 200 nM okadaic acid, 4.8 μm cypermethrin (all bought from Merck KGaA, Darmstadt, Germany), 2 mM vanadate, 10 mM sodium pyrophosphate, 10 mM NaF and 5 mM EDTA. After 10 min incubation on ice, cells were lysed by douncing. Cell debris and nuclei were removed by centrifugation for 10 min at 4°C at 5500 g. Then the cytoplasmic and membrane fraction were separated by ultracentrifugation at 100000 g for 60 min at 4° C. The proteins of the cytosolic fraction (supernatant) were subjected to acetone precipitation. The protein pellets were resolubilized in 3 mM EDTA, 20 mM Tris-HCl, pH 8.3, and 8 M urea. The disulfide bonds of the proteins were reduced with tris (2-carboxyethyl) phosphine at a final concentration of 12.5 mM at 37°C for 1 h. The produced free thiols were alkylated with 40 mM iodoacetamide at room temperature for 1 h. The solution was diluted with 20 mM Tris-HCl (pH 8.3) to a final concentration of 1.0 M urea and digested with sequencing-grade modified trypsin (Promega, Madison, WI) at 20 μg per mg of protein overnight at 37°C. Peptides were desalted on a C18 Sep-Pak cartridge (Waters, Milford, MA) and dried in a speedvac.

The *S. cerevisiae* phosphorylation data was generated as follows: Wild type (BY7092: can1Δ his3Δ leu2Δ ura3Δ met15Δ) were grown to OD ∼0.8 at 30°C in synthetic dened (SD) medium (per liter 1.7 g YNB, 5 g ammonium sulfate, 2% glucose, 0.03 g isoleucine, 0.15 g valine, 0.04 g adenine, 0.02 g arginine, 0.1 g leucine, 0.03 g lysine, 0.02 g methionine, 0.05 g phenylalanine, 0.2 g threonine, 0.02 g histidine, 0.02 g tryptophane, 0.03 g tyrosine, 0.02 g uracil, 0.1 g glutamic acid and 0.1 g aspartic acid). Cells were harvested at 30°C by centrifugation at 850 g and then washed once in SD medium. Finally, they were collected by centrifugation and shock-frozen in liquid nitrogen. Pellets were thawed in ice cold lysis buffer (20 mM Tris/HCl pH 8.0, 100 mM KCl, 10 mM EDTA, 0.1% NP40, 20 nM calyculin A, 200 nM okadaic acid, 4.8 μM cypermethrin (all obtained from Merck KGaA, Darmstadt, Germany), 2 mM vanadate, 10 mM sodium pyrophosphate and 10 mM NaF) using 1 mL of lysis buffer per gram of yeast. The cells were lysed by glass-bead beating (using acid-washed glass beads), the protein supernatant was precipitated using ice-cold acetone, and the pellet was resuspended in 8 M urea, 20 mM Tris/HCl at pH 8.3. After dilution to $<$ 1.5 M urea with 20 mM Tris/HCl at pH 8.3, proteins were digested using trypsin in a w/w ratio of 1:125 and puried using C18 reverse phase chromatography (Sep-Pack, Waters).

The *C. elegans* phosphorylation data was generated as follows: Wild-type *C. elegans* strain N2 (Bristol) was grown on 9-cm nematode growth medium (NGM) agar plates seeded with a lawn of the *E. coli* strain OP50 or in 100-ml liquid cultures in S-basal buffer in bevelled flasks (with concentrated E. coli NA22 as a food source). Worms were harvested from plates or liquid culture and separated from the bacteria by washing with M9 buffer three times. The worms were harvested by centrifugation at 500 g and shock frozen using liquid nitrogen. Subsequently, the worms were homogenized with glass beads (diameter of 212-300 μm, Sigma-Aldrich, St Louis, MO, USA) in the ratio of 1:1:2 (worms:beads:buffer) in a cell disrupter (FastPrep FP120, Thermo Savant, Qbiogene Inc., Carlsbad, CA, USA) at 4°C three times for 45 s at level 6. The buffer used was

50 mM Tris/HCl, pH 8.3, 5 mM EDTA, 8 M urea. After glass bead beating treatment, 0.125% SDS was added and the homogenate was incubated for 1 h at room temperature to solubilize proteins. Cell debris was removed by centrifugation. The peptides were produced from the proteins in the supernatant as described above.

Peptide separation by isoelectric focusing

All peptides were separated according to their isoelectric point. For the *D. melanogaster* this was performed using an free-flow electrophoresis instrument, type prometheus from FFE Weber Inc. (now BD-Diagnostics, PAS) and FFE-Weber reagent basic kit (Prolyte 1, Prolyte 2, Prolyte 3 and Prolyte 4-7 and pI markers) (BDDiagnostics, NJ, USA). The digested peptides were diluted in separation media containing 8 M Urea, 250 mM Mannitol and 20% ProLyte solution at a concentration of 10 mg/ml. This sample was loaded continuously for 1 h at 1 ml/h. Total collection time was 24 h and the volume of each collected fraction was about 25 to 50 ml. A Thermo Orion needle tip micro pH electrode (Thermo Electron Corporation, Beverly, MA) was used to measure the pH value of each fraction. Peptides from the FFE fractions 18-60 were purified on a C18 Sep- Pak cartridge (Waters Corporation, Milford, MA, USA) (1). For *C. elegans* and *S. cerevisiae* the dried-down peptide samples (15 mg and 20 mg, respectively) were seperated with an Offgel fractionator and therefore resolubilized to a final concentration of 1 mg/ml in off-gel electrophoresis buffer containing 6.25% glycerol and 1.25% IPG buffer (GE Healthcare). The peptides were separated on pH 3-10 IPG strips (GE Healthcare) with a 3100 OFFGEL fractionator (Agilent) as previously described (4; 5). We performed a 1-hour rehydration at maximum 500 V, 50 mA, and 200 mW followed by the separation at maximum 8000V, 100 mA, and 300 mW until 50 kVh were reached. Following isoelectric focussing, the fractions were concentrated and cleaned up by C18 reversed-phase spin columns according to the manufacturers instructions (Sep-Pack, Waters).

Phosphopeptide isolation

Phosphopeptides were isolated using a titanium dioxide resin as follows: 1-3 mg of dried peptides were reconstituted in 280 μl of a washing solution (WS), containing 80% acetonitrile and 3.5% TFA, which is saturated with phtalic acid (~100g phtalic acid per ml). Then 1.25 mg TiO_2 (GL Science, Saitama, Japan) resin was placed into a 1-ml Mobicol spin column (MoBiTec, Gottingen, Germany) and was subsequently washed with 280 μl water, 280 μl methanol, and finally was equilibrated with 280 μl WS for at least 10 minutes. After removal of the WS by centrifugation using 500 x g, the peptide solution was added to the equilibrated TiO_2 in the blocked Mobicol spin column and was incubated for > 30 min with end-over-end rotation. After this step, the peptide solution was removed by centrifugation, and the resin was thoroughly washed two times each with 280 μl of the WS, with a 80% acetonitrile, 0.1% TFA solution, and finally with 0.1% TFA. In the final step, phosphopeptides were eluted from the TiO_2 resin using two times 150 μl of a 0.3 M NH_4OH solution (pH ~10.5). After elution, the pH of the pooled eluents was rapidly adjusted to 2.7 with 10% TFA, and the phosphopeptides were purified with an appropriate reverse-phase column suitable for up to 20 μg peptide. Besides the separated peptides, this procedure was also performed on yeast and worm whole-cell or whole-organism lysates.

Alternatively, phosphopeptides were also isolated with immobilization by metal affinity chromatography (IMAC). In detail, 1-3 mg of peptides were reconstituted in 280 μl of a WS, consisting of 250 mM acetic acid with 30% acetonitrile at pH 2.7. Then 60 μl of uniformly suspended PHOS-Select iron affinity gel (Sigma Aldrich), corresponding to ~30 μl resin, was placed into a 1-ml Mobicol spin column. The resin was equilibrated three times with 280 μl of the WS. After removal of the WS by centrifugation at 500 x g, the peptide solution was added to the equilibrated IMAC resin in the blocked Mobicol spin column. To obtain reproducible results, it is crucial that the pH in all replicate samples is maintained at ~2.5. The affinity gel was then incubated with the peptide solution for 120 min with end-over-end rotation. After the incubation, the liquid was removed by centrifugation and the resin was thoroughly washed two times with 280 μl of the WS, and once with ultra pure water. In the final step, phosphopeptides were eluted once with 150 μl of a 50 mM phosphate buffer (pH 8.9) and once with 150 μl of a 100 mM phosphate buffer (pH 8.9), each time incubating the resin $<$ 3 min with the elution buffer. Both elutes were pooled, the pH were rapidly adjusted to 2.7 using 10% TFA, and the phosphopeptides were purified with an appropriate reverse-phase column. This procedure was performed on separated fly and worm peptide samples and yeast whole-cell lysates.

Finally, phosphoramidate chemistry (PAC) was used in the case of D. melanogaster for phosphopeptide isolation on the peptide samples after isoelectric focusing (1; 2). Phosphopeptides were isolated with phosphoramidate chemistry as follows: 1 mg of dried peptide was reconstituted in 750 μl of methanolic HCl, which was prepared by slowly adding 120 μl of acetyl chloride to 750 μl of anhydrous methanol. The methyl esterification was then allowed to proceed at 12°C for 120 min. The solvent was quickly removed in a cool vacuum concentrator and peptide methyl esters were dissolved in 40 μl methanol, 40 μl water, and 40 μl acetonitrile. Then 500 μl of a solution containing 50 mM N-(3-Dimethylaminopropyl)-N ethylcarbodiimide (EDC), 100 mM imidazole pH 5.6, 100 mM 2-(N-Morpholino)ethanesulfonic acid (MES) pH 5.6, and 2 M cystamine was added to the peptide solution. The reaction was allowed to proceed at room temperature with vigorous shaking for 8 hours. The solution was then loaded onto an appropriate reverse-phase column and the derivatized peptides were subsequently: First, washed with 0.1% TFA; second, treated with 10 mM TCEP (pH should be adjusted to $\sim$3 using sodium hydroxide (NaOH)) for 8 minutes, in order to produce free thiol groups; third, washed again with 0.1% TFA to remove residual TCEP. Finally, the derivatized peptides were eluted with 80% acetonitrile, 0.1% TFA and the pH was adjusted to 6.0 with phosphate buffer. Then acetonitrile was partially removed in the vacuum concentrator to yield a final concentration of ∼30%, and the derivatized phosphopeptides were incubated with 5 mg maleimide functionalized-glass beads for 1 h at pH 6.2 in a Mobicol column. (The beads were synthesized by dissolving 120 μmol hydroxybenzotriazole, 120 μmol of 3-maleimidopropionic acid, and 120 μmol diisopropylcarbodiimide in 1 ml of dry dimethylformamide, completely. After 30 minutes of incubation, 100 mg CPG beads (Proligo Biochemie, Hamburg, Germany) corresponding to 40 μmol free amino groups were added for 90 minutes. After the reaction, beads were washed with dimethylformamide and dried with a vacuum concentrator. Beads were stored dry at 4°C. The derivatized beads were washed two times sequentially with 300 μl 3 M NaCl, water, methanol, and, finally, with 80% acetonitrile to remove nonspecifically bound peptides. In the last step, the beads were incubated with 5% TFA, 30% acetonitrile for 1 h to recover the phosphopeptides. The recovered sample was dried in the vacuum concentrator. This procedure was also performed on yeast whole-cell lysates.

Mass spectrometry data analysis and sampling depths

The liquid chromatography-tandem mass spectrometry (LC-MS, on a Thermo Fisher Scientic LTQ ORBITRAP XL) analysis and database searches were performed as described in [17]. The *S. cerevisiae* and *C. elegans* MS spectra were searched against the SGD (release October 10th 2007) and WormBase (release WS183) databases and the *D. melanogaster* data were searched against the FlyBase database v4.3. In addition to the data stored in the PhoshoPep database, we added in the case of *S. cerevisiae* electron transfer dissociation (ETD) fragmentation data from [29], although these data only constitute 19% of the total dataset. In this study, the *D. melanogaster* phosphopeptide isolates were most extensively analyzed in terms of the total number of LC-MS/MS runs employed and consequently larger coverage was achieved than for the other target species. Finally, the expected sizes of the phosphoproteomes of yeast, worm, and fly strongly differ, simply due to differences in their genome sizes and repertoires of kinases

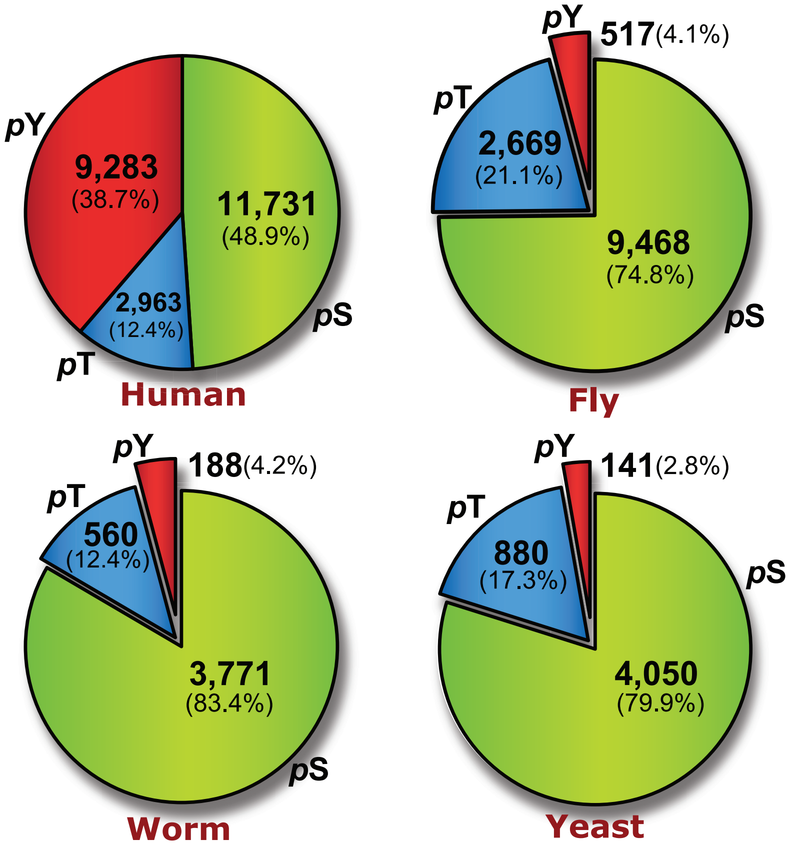

Figure 2.1: Number of phosphorylation sites assembled.

and phosphatases.

2.2.3 Identification of phosphorylated orthologs for human phosphoproteins in the three query species

Ortholog information of human phosphoproteins inferred by Ensembl (release 46) ortholog detection pipeline was obtained from Ensembls BioMart interface. Specifically, Ensembl identifiers of genes orthologous to human genes together with identifiers of their translated protein products were retrieved. The details of the ortholog detection pipeline are described at http://aug2007.archive.ensembl.org/info/data/compara/homology_method.html. Briefly, gene families are identified from all sequences in the database by WU-Blastp and Smith-Waterman searches, followed by construction of a phylogenetic tree for each gene family to identify orthology and paralogy relationships between gene pairs. Finally, we used this information to identify human phosphoproteins with orthologs that were phosphorylated in at least one of the target species (we termed these phospho-orthologs). Subsequently, the sequences of these human phosphoproteins were aligned with those of their target species phosphoorthologs to identify positionally conserved phosphorylation events (such phosphorylation sites are termed “core” sites).

2.2.4 Identification of core sites

The phosphorylation core sites were detected from multiple sequence alignments (MSAs) of each human phosphoproteinwith all its detected phosphoorthologs (as described above). To improve each MSA, we included the protein sequence of the longest splice variant (or an arbitrarily chosen longest if several exist with identical length) of one-to-one orthologous genes from 19 eukaryotic species spanning the evolution between *Homo sapiens* and *D. melanogaster* (*Aedes aegypti, Anopheles gambiae, Bos taurus, Canis familiaris, Ciona intestinalis, Ciona savignyi, Danio rerio, Gallus gallus, Gasterosteus aculeatus, Macaca mulatta, Monodelphis domestica, Mus musculus, Ornithorhynchus anatinus, Oryzias latipes, Pan troglodytes, Rattus norvegicus, Takigufu rubripes, Tetraodon nigroviridis*, and *Xenopus tropicalis*). For the sake of completeness, we also included the orthologous protein sequences for each target species that had no detected phosphorylation. Finally, these sequences were aligned using the MAFFT (v6.240, E-INS-i option with default parameters) algorithm on an IBM x366 running CentOS (LINUX). The resulting MSAs were subsequently processed by a Perl script to identify the human phosphoresidues that are aligned in the same column with a phosphoresidue observed in any target species (we termed these phosphorylation sites core sites). We did not require the aligned phosphoresidues to be identical amino acids to allow detecting cases where one phosphoresidue is converted to another during evolution (for example, pT to pS or pY).

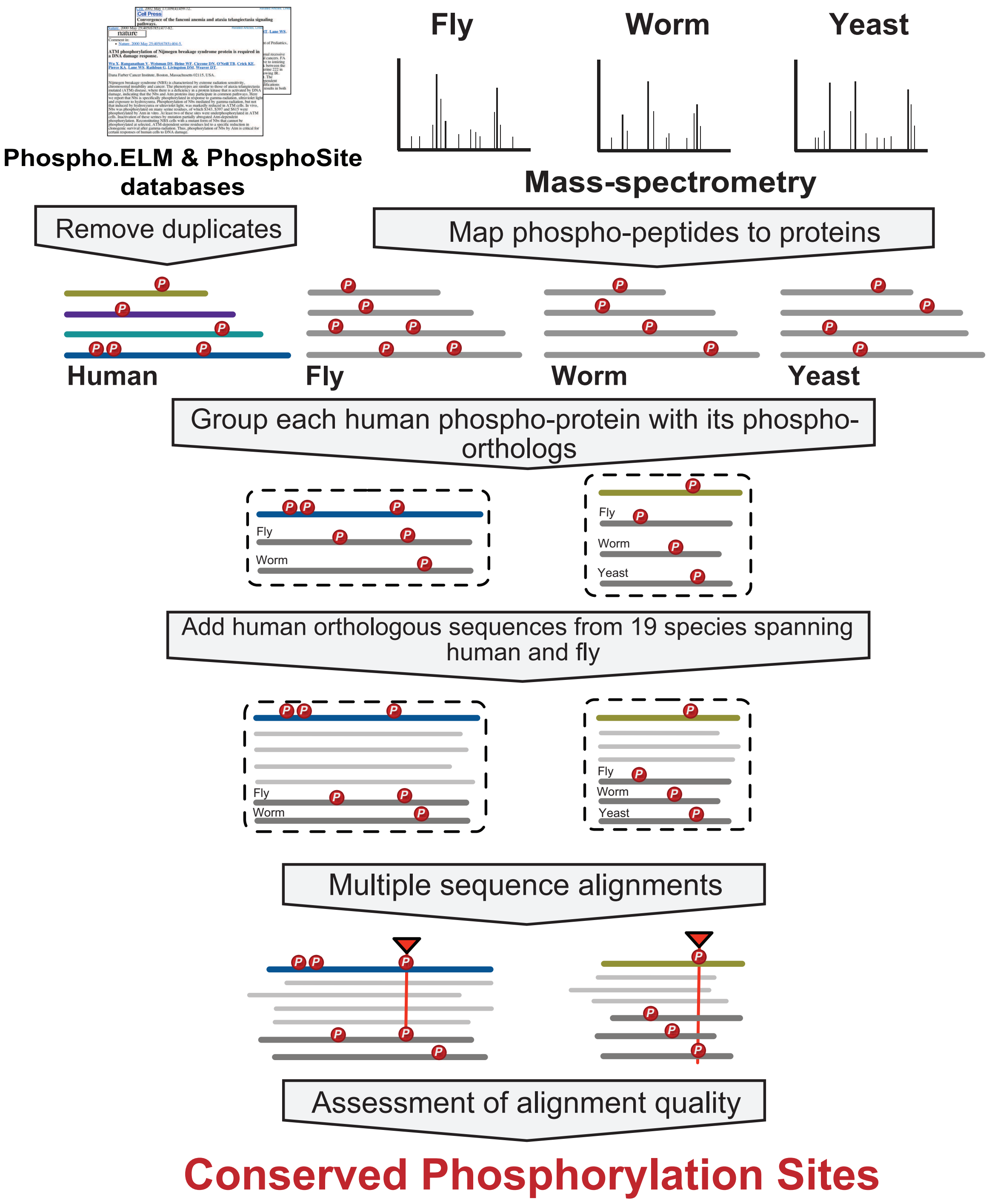
Phospho.ELM & PhosphoSite databases
Fly
Worm
Yeast
Mass-spectrometry
Remove duplicates
Map phospho-peptides to proteins
Human
Fly
Worm
Yeast
Group each human phospho-protein with its phospho-orthologs
Fly
Worm
Fly
Worm
Yeast
Add human orthologous sequences from 19 species spanning human and fly
Fly
Worm
Fly
Worm
Yeast
Multiple sequence alignments
Assessment of alignment quality
Conserved Phosphorylation Sites

Figure 2.2: Schematic overview of core site detection.

2.2.5 Assessing local alignment quality of core sites with shuffled phosphoortholog sequences

We repeated the MSA with shuffled sequences of phosphoorthologs to identify spurious core sites that could arise from poorly aligned regions in the sequence alignment by randomchance alone. First, we identified pairs of aligned phosphoresidues lying in potential poorly aligned regions, which we defined as those having less than 50% identity between human and the target species in the sequence region -5 to $+5$ (excluding position 0) relative to the human phosphoresidue. For each of these pairs of aligned phosphoresidues, we then computed the BLOSUM62 alignment score between human and target species of sequence region -5 to $+5$ relative to the human phosphoresidue, and repeated the MSA, as outlined above, 500 times but with the sequence of the phosphoortholog shuffled randomly each time. We then computed the empirical P value for the BLOSUM62 computed alignment score of the aligned phosphoresidues as the fraction of trials in which the shuffled phosphoortholog sequence aligned to the same region in the human phosphoprotein to a phosphorylatable residue (S, T or Y) with equal or better BLOSUM62 score than the actual phosphoortholog sequence. Finally, we used these values to only consider core sites that have an empirical P value < 0.05 resulting in 479 core sites.

2.2.6 Assessing the statistical significance on the number of observed aligned phosphoresidues

We adopted a simple probabilistic model to estimate the statistical significance of the number of observed aligned phosphoserine, phosphothreonine, and phosphotyrosine residues between human and each target species. First, we computed the number of aligned phosphoresidues expected by random chance between human and each target species in the nonshuffled MSA (separate analyses were performed for phosphoserine, phosphothreonine, and phosphotyrosine). Here, we illustrate, as an example, how the number of aligned phosphotyrosines expected by random chance between human and fly was derived: Let A be the set of human tyrosine-phosphorylated proteins whose orthologs in fly are tyrosine phosphorylated. Correspondingly, let B be the set of fly tyrosine-phosphorylated proteins that correspond to A. Next, let P_A and P_B be the proportion of tyrosines in protein set A and B, respectively, that are phosphorylated, and let N_{AB} be the total number of tyrosines in A that are aligned to tyrosines in B as observed in the MSA (described above). It then follows that the number of human phosphotyrosines aligning to phosphotyrosines in fly expected by random chance, assuming joint probability of two independent events, is computed as $P_A \times P_B \times N_{AB}$. Finally, we assessed the statistical significance of the difference between expected random occurrence and observed number of aligned phosphotyrosines by a X^2 test. Similar analyses were then performed between human and each target species for serine, threonine, and tyrosine separately.

2.2.7 Phosphorylation motif discovery from positionally conserved phosphorylation sites

For every pair of aligned phosphorylated residues, a consensus sequence of the local alignment from -5 to $+5$ of the aligned phosphorylated residues is first defined. For example, ..RK.SP..D. is the consensus pattern of GTRKG*p*SPLKDE aligned to NERKV*p*SPDEDM. Next, a consensus pattern S encoded as a vector set $V = (v_{-5}, v_{-4}, , v_4, v_5)$ is defined, where vector v_i is a vector of the 20 elements coding for number of specific amino acids appearing at position i among the consensus sequences. Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. The similarity between vector set V_x and V_y is computed as the sum of cosine similarities of all corresponding vectors across the two sets, as follows. This serves to quantify how similar are two set of consensus sequences based on frequency of amino acids observed at each position of the consensus sequences. First, a vector set is encoded for every consensus sequence. Next, the similarity between pairs of vector sets are computed, and the

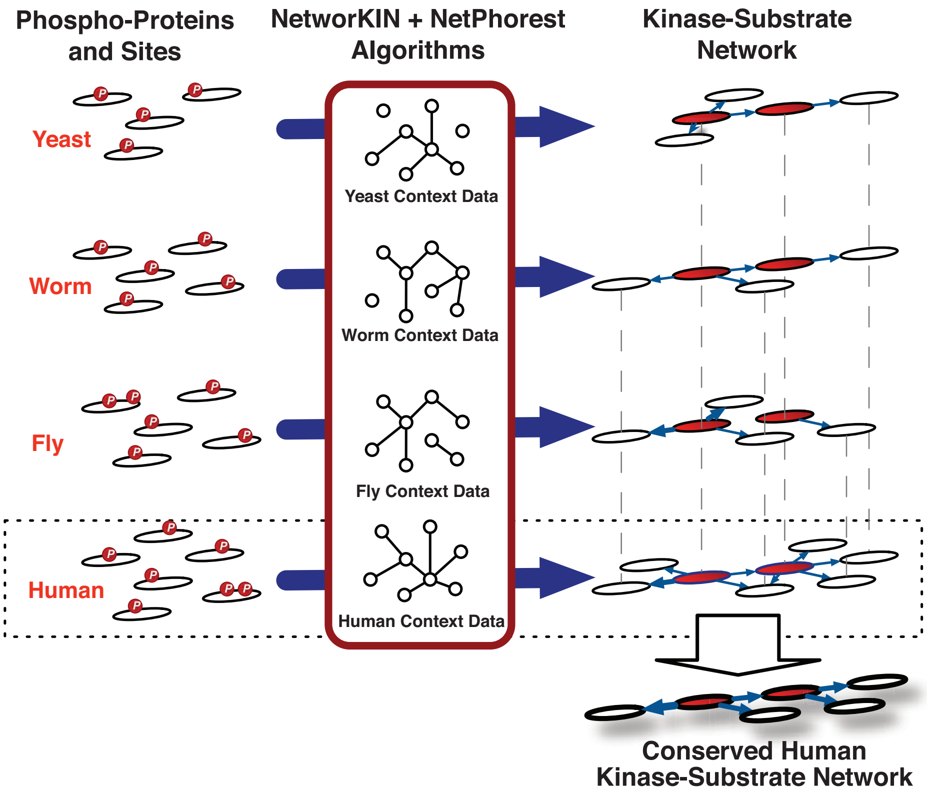

Figure 2.3: Identification of potential conserved kinase-substrate interaction network. NetworKIN and NetPhorest algorithms were applied to the target species phosphoproteomes to reconstruct kinase-substrate networks. Second, interactions within these networks were superimposed (or aligned) with each other. Finally, for each substrate, we defined a phosphorylation conservation propensity k of the number of phosphorylation events supported by orthologous kinase-substrate phosphorylation in the target species.

most similar pair is then merged into a new vector set by summing up the corresponding vectors across the two old sets. The previous step is iteratively performed, and if the two most similar vector sets at each iteration encode 10 or more core sites, they are output and removed from further computation. Lastly, core sites in human and target species represented by output vector sets are then visualized separately with sequence logos for manual inspection and classification.

2.2.8 Computational reconstruction of conserved human kinase-substrate networks

Many cellular processes and behaviors are mediated by protein kinases through phosphorylation of its substrate proteins. Identifying these kinase-substrate interactions will help to construct the signaling networks that relay extracellular signals to elucidate cellular responses. We seek to identify such kinase-substrate networks in human that are evolutionarily conserved in our query species. We used the NetworKIN algorithm (v2.0b [98, 99]) to predict the kinases that may phosphorylate the phosphorylation sites in the four species (*H. sapiens*, *D. melanogaster*, *C. elegans* and *S. cerevisiae*), resulting in four directed and weighted kinase-substrate networks. We used default parameters for NetworKIN, setting the ranking score cutoff to 0.7 for human and 0.5 for target species. This setting was an empirical decision made on the basis of the relatively weak association data in worm and fly compared to yeast and human. In addition, we expect conservation to reduce spurious protein-protein associations. Many predictions from NetworKIN are based on indirect probabilistic associations of proteins; thus, a direct physical interaction is not an absolute prerequisite for the algorithm to associate a substrate with a kinase. Because STRING [74] utilizes evidence transfer between the target species, our approach will be somewhat biased toward these associations. However, the systematic analysis of the phosphoproteomes of the target species and the use of linear motif from NetPhorest [114] serve as unbiased starting material for the NetworKIN prediction algorithm, minimizing this issue. NetPhorest database contain a set of probabilistic classifiers that infer which protein kinase (or kinase family) and/or phosphoresidue-binding protein domain most likely targeted a phosphorylation site given the known specificity of protein kinases.

Each edge in the networks represents a predicted kinase-substrate relationship. The weight of the edge is proportional to the total number of sites among spliced variants of the substrate gene product predicted to be phosphorylated by the kinase. The human kinase-substrate network is compared across the three target species to infer a network of evolutionary conserved kinase-substrate relationships in human. Each inferred evolutionary conserved kinase-substrate relationship in human is further scored (see Figure 2.4 for illustration). For each predicted human kinase-substrate relationship (a, b), kinase a and substrate b that are orthologous to kinase-encoding gene set A_x and substrate-encoding gene set B_x in target species x (fly, worm, or yeast), where A_x and B_x can be an empty set. Let n be the edge weight of (a, b) and m_x be the maximum edge weight among kinase-substrate pairs from A_x and B_x in xs weighted kinase-substrate network. The human kinase-substrate relationship (a, b) is considered conserved in target species x if $m_x > 0$ (kinase-substrate relationship between members of A_x and B_x is predicted by NetworKIN based on phosphorylation data in target species x). The conservation score C_x of kinase-substrate relationship (a, b) across target species x is then selected as the smaller number of n and m_x, essentially requiring every predicted kinase-substrate relationship inferred in human to be supported by similar one inferred in target species. The final conservation score C_{total} of kinase-substrate relationship (a, b) in human across the three target species is the sum of C_{fly} , C_{worm}, and C_{yeast} which sum up the support for the inferred kinase-substrate relationship from the 3 query species. Finally, the conserved phosphorylation propensity k of a substrate b is calculated as the sum of C_{total} of each conserved kinase-substrate relationship that b is implicated in. Finally, we chose not to compress multiple orthologous kinases into a single node, such as JNK1 and JNK2 into a JNK group [114], because it is possible for functional divergence to occur after duplication such that the initial set of substrates targeted by an ancient kinase become uniquely targeted among the duplicated kinases.

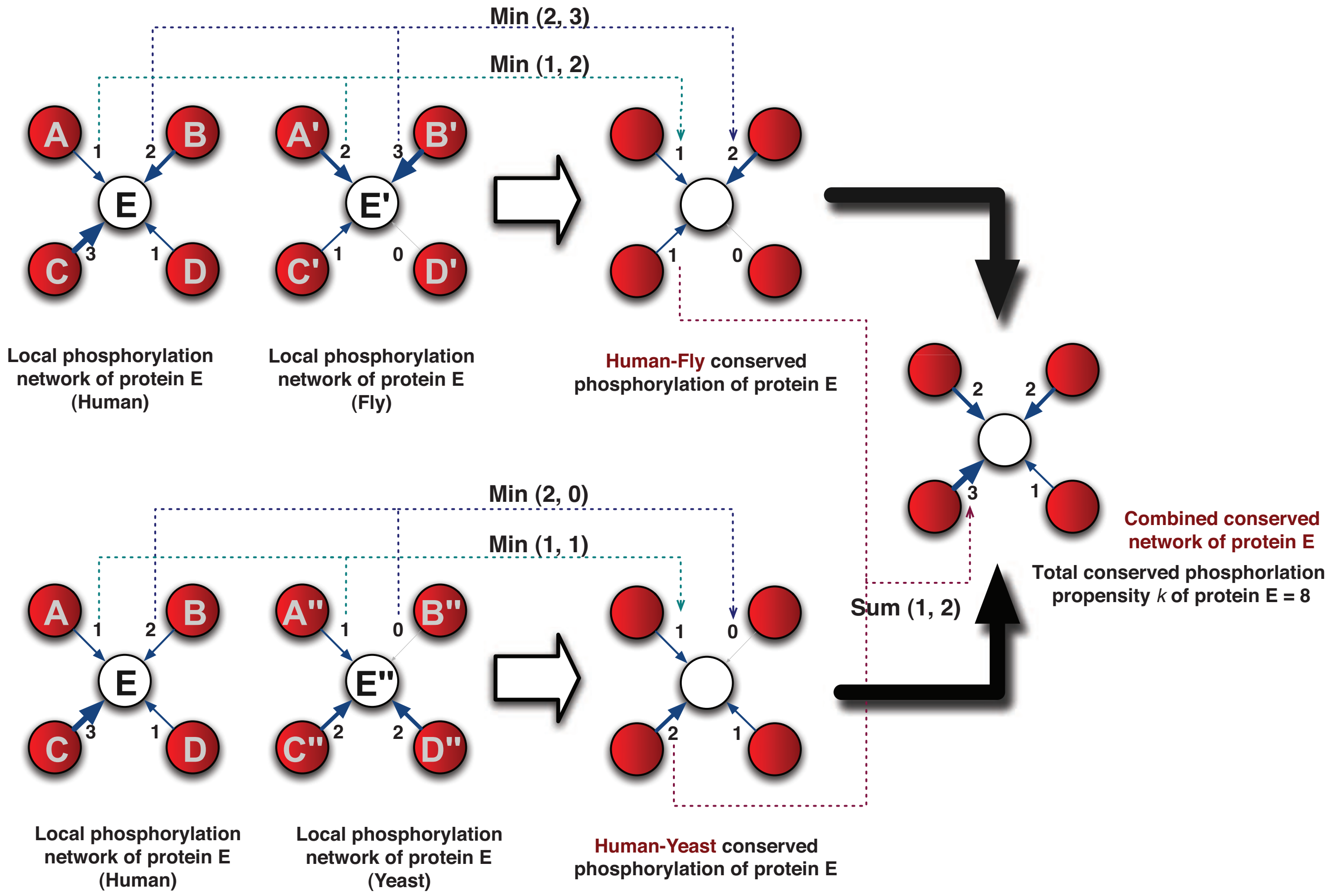
Min (2, 3)
Min (1, 2)
A
B
E
C
D
1
2
3
1
A'
B'
E'
C'
D'
2
3
1
0
1
2
1
0
Local phosphorylation network of protein E (Human)
Local phosphorylation network of protein E (Fly)
Human-Fly conserved phosphorylation of protein E
Min (2, 0)
Min (1, 1)
A
B
E
C
D
1
2
3
1
A''
B''
E''
C''
D''
1
0
2
2
1
0
2
1
Local phosphorylation network of protein E (Human)
Local phosphorylation network of protein E (Yeast)
Human-Yeast conserved phosphorylation of protein E
Sum (1, 2)
2
2
3
1
Combined conserved network of protein E
Total conserved phosphorlation propensity k of protein E = 8

Figure 2.4: Schematic diagram of how the conserved phosphorylation propensity k of each human substrate is computed.

2.2.9 Assessing statistical significance of inferred conserved kinase-substrate relationships

To assess the statistical significance of the human kinase-substrate relationships inferred to be conserved in the target species, we repeated the procedure described above in Section 2.2.8 2,000 times, using randomized kinase-substrate networks of the three target species with the predicted human kinase-substrate network. Each time, randomized kinase-substrate networks in target species are created by switching all originally predicted substrates of each kinase with that of another randomly selected kinase within the same species. The empirical P value is then computed as the fraction of trials that have the same or more inferred conserved human kinase-substrate relationships than the original analysis.

2.2.10 Prediction of intrinsic disordered regions in human phosphoproteins

We used the DISOPRED2 predictor (http://bioinf.cs.ucl.ac.uk/disopred/) [193] to predict disordered regions in human protein sequences by inputting these to the predictor. The nonredundant (NR) protein sequence database required for the predictor to run was obtained from the National Center for Biotechnology Information in November 2007. The NR database was filtered for transmembrane protein regions with the pfilt program provided with DISOPRED2. Subsequently, we analyzed the output with custom Perl scripts and SQL queries.

2.2.11 Assembly of disease-associated gene data set

We obtained a list of cancer-associated genes annotated in four peer reviewed publications [56, 189, 116, 59] from CancerGene (http://cbio.mskcc.org/cancergenes) [59]. The first two publications reviewed genes important in cancer development, maintenance, and metastasis, and the last two reported genes with mutations causally implicated in oncogenesis as observed in primary neoplasms. As the cancer-associated genes reported in Futreal *et al.* [59] form the basis of cancer-associated genes in Cancer Gene Census (www.sanger.ac.uk/genetics/CGP/Census/), we obtained the latest list from the database. Subsequently, the gene symbols and aliases obtained were mapped to Ensembl gene entries with the alias mapping file provided by the STRING database (http://string.embl.de) [190], resulting in a final set of 413 cancer-related genes. In addition we assembled a data set of genes involved in genetic diseases from the OMIM database (www.ncbi.nlm.nih.gov/omim/). These genes were obtained from OMIM and mapped to gene identifiers in the Ensembl database (release 46). This resulted in a total set of 2174 human genes associated with disease.

2.2.12 Statistical and function enrichment analysis

Gene Ontology (GO) term enrichment analyses were performed with the BiNGO (v2.00) [106] plugin for Cytoscape (v2.5.2) [164]. The GO annotations of human genes were retrieved from Ensembl (release 48) and the statistical significance of overrepresented GO terms was determined with hypergeometric distribution tests (corrected for multiple hypothesis testing with false discovery rate). The statistical significance of GO terms associated with core site genes was estimated by comparing the GO terms of two sets of human genes encoding phosphoproteins: those that have orthologs in at least one target species and its subset of genes that have phosphoorthologs in the target species. The statistical significance of GO terms associated with human core net genes (substrates with inferred conserved kinase-substrate relationships in target species) was estimated by comparing it to the entire set of human genes encoding phosphoproteins that have phosphoorthologs in at least one target species, and the phosphoorthologs that have kinase-substrate relationship predicted by NetworKIN.

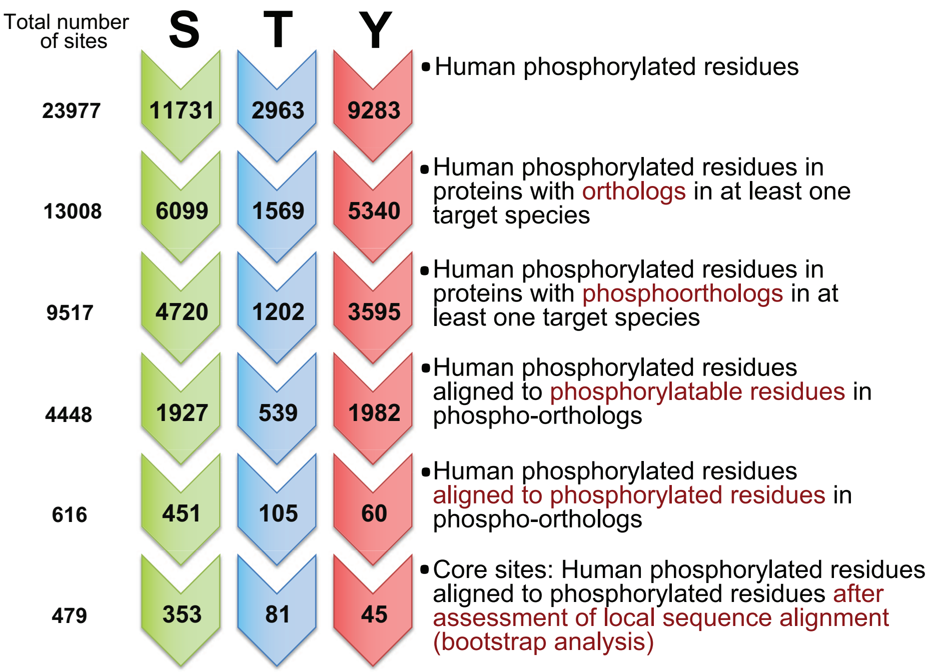
Total number of sites
S
T
Y
23977
11731
2963
9283
• Human phosphorylated residues
13008
6099
1569
5340
• Human phosphorylated residues in proteins with orthologs in at least one target species
9517
4720
1202
3595
• Human phosphorylated residues in proteins with phosphoorthologs in at least one target species
4448
1927
539
1982
• Human phosphorylated residues aligned to phosphorylatable residues in phospho-orthologs
616
451
105
60
• Human phosphorylated residues aligned to phosphorylated residues in phospho-orthologs
479
353
81
45
• Core sites: Human phosphorylated residues aligned to phosphorylated residues after assessment of local sequence alignment (bootstrap analysis)

Figure 2.5: Number of human phosphorylation sites at different stages of conservation detection.

2.3 Result

A total of 23,977 human phosphorylation sites found across 6456 phosphoproteins encoded by 6293 genes were assembled from the two primary online databases PhosphoSite (release 2.0) [64] and Phospho.ELM (release 7.0) [36]. For *D. melanogaster*, *C. elegans*, and *S. cerevisiae*, we used phosphorylation site data that were generated with a similar experimental and computational pipeline (see Methods and Supplementary Materials) and are available via the PhosphoPep database (www.phosphopep.org) [17, 16]. Our study included 12,654, 4519, and 5071 phosphorylation sites for *D. melanogaster*, *C. elegans*, and *S. cerevisiae*, respectively. We observed an exceptionally high fraction of phosphotyrosine sites in the assembled human phosphorylation data that can largely be attributed to HTP phosphotyrosine antibody-based studies [147, 151]. The portion of phosphoserine, phosphothreonine, and phosphotyrosine is shown in Figure 2.1.

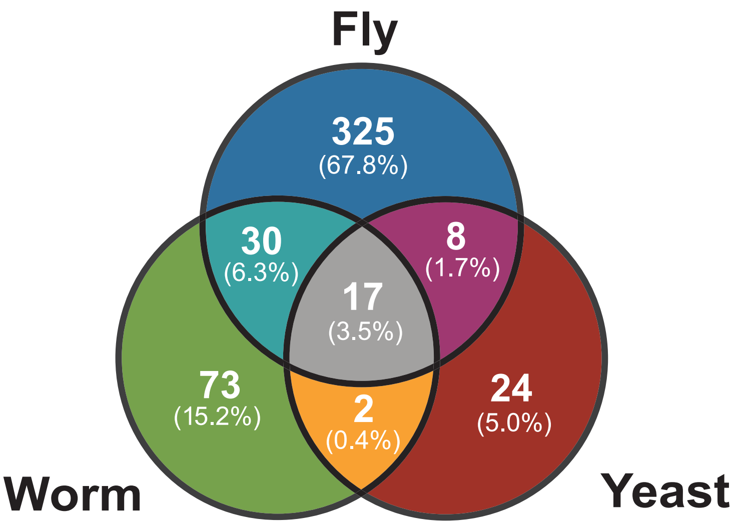

Figure 2.6: Target species source of the 479 core sites.

2.3.1 Positionally conserved phosphorylaton events in human and function enrichment analysis

Of all the human sites assembled, 39.7% were in found in proteins orthologous to phosphoproteins detected in at least one target species (Figure 2.5). Deploying a sequence-alignment protocol (Figure 2.2, see Methods) with the MAFFT program [81] on the three target phosphoproteomes and the human phosphorylation set (see Methods), we identified 479 sites (termed core sites) that were conserved between human and at least one target species in 344 proteins encoded by 337 human genes (termed core site genes, Figure 2.1). Of these core sites, 73.7% are phosphoserines, 16.9% are phosphothreonines, and 9.4% phosphotyrosines (Figure 2.5). These sites make up 10.8% of the 4448 human phosphoresidues that were aligned to phosphorylatable residues in at least one target species, and in most cases, these numbers are significantly higher than expected by random chance from observed alignments (table 2.1).

Among the 479 sites, 139 (≈29%) were found within 75 protein domain families (compared to the global average of ≈20% for all 29,977 human phosphorylation sites), 57 were conserved in at least two target species, and 17 were conserved in all three target

Table 2.1: Observed versus expected of core sites by chance.

Between human and fly				
Residue	Expected	Observed	Odd Ratio	*p*-value
pS	55.53	325	5.85	2.2e-16
pT	6.43	64	9.95	6.9e-12
pY	11.23	33	2.94	1.1e-3

Between human and worm				
Residue	Expected	Observed	Odd Ratio	*p*-value
pS	14.9	116	7.79	2.2e-16
pT	0.89	7	7.87	-
pY	2.99	13	4.35	1.2e-2

Between human and yeast				
Residue	Expected	Observed	Odd Ratio	*p*-value
pS	9.54	46	4.82	1.0e-07
pT	0.73	2	2.74	-
pY	1.48	5	3.38	-

species (Figure 2.6). We observed that core sites shared between humans and more than one target species have an increased tendency to be located within protein domains: 9 of the 17 omnipresent core sites occurred in domains from 6 families (dehydrogenase E1, phosphoglucomutase-phosphomannomutase, glycogen synthase, PhoX homologous, Cdc37 N-terminal kinase binding, 60S acidic ribosomal, and serine-threonine protein kinase catalytic domain), suggesting that the phosphorylation of these protein domains is of ancient origin. It should be noted that not all the core sites identified are phosphorylated by kinases; for example, phosphorylation of the core site Ser^{175} in the phosphoglucomutase domain of human glucose-1,6-bisphosphate synthase likely happens by self-catalysis.

To analyze the functional context of core site genes, we constructed a functional association network among these genes with the STRING resource (Figure 2.7). This network revealed a tight cluster of functionally associated core site genes that encode components of various protein complexes and signaling networks, as well as singleton genes that were not confidently associated to any other core site gene. The β-catenin destruction complex and clathrin coat proteins of coated pits appear to be heavily regulated by protein phosphorylation of ancient origin because they contain core sites in four out of four and four out of five of their conserved protein components, respectively (Table 2.3). Function enrichment analysis with Gene Ontology [5] annotation revealed that core site genes are involved in fundamental cellular processes. For example, amino acid phosphorylation, RNA splicing, cell division, and translation were statistically enriched over the super set of human phosphoproteins that have orthologs in target species ($P < 0.05$, hypergeometric test, Benjamini and Hochberg false discovery rate correction). Thus, the observed enrichment suggests that even processes not previously appreciated as regulated by phosphorylation, such as the phosphorylation-mediated regulation of many RNA splicing proteins observed in human cells, arose early during evolution before the last common ancestor of fly and human.

Tracing the experimental sources of the core sites, we found that 65.3% of the core

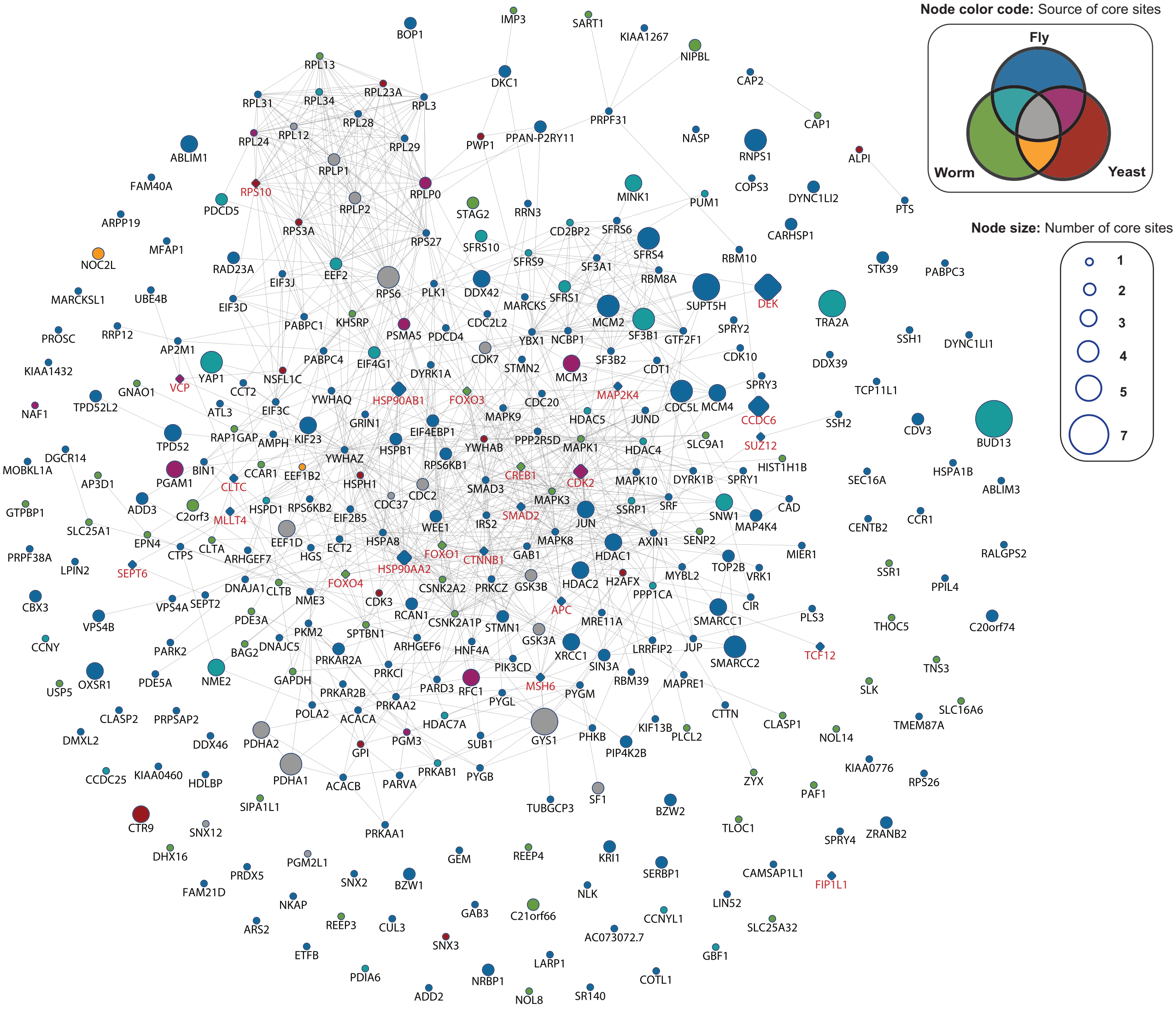

Figure 2.7: Functional association network of the human phosphoproteins containing core site(s) The comparative approach identified 479 phosphorylation sites in 344 proteins (mapping to 337 human genes). The STRING resource (66) was used to construct a functional association network of these proteins (using only high confidence probabilistic associations). Nodes represent genes andare colored according to which target species contains the conserved phosphorylation site(s) and node size indicates the number of core sites that the encoded proteins have. Known cancer-associated genes are highlighted by red text and diamond nodes (see Table 2.4).

Table 2.2: Core sites identified in components of beta-catenin destruction complex.

Gene Symbol (HGNC)	Ensembl Protein ID	Residue	Position	Target Species	Source Type
GSK3B	ENSP00000264235	Y	216	WORM FLY YEAST	HTP LTP
GSK3B	ENSP00000264235	S	9	FLY	HTP LTP
APC	ENSP00000257430	S	1279	FLY	HTP LTP
AXIN1	ENSP00000262320	T	79	FLY	HTP
CTNNB1	ENSP00000344456	S	675	FLY	HTP LTP

Table 2.3: Core sites identified in components of the clathrin coat of coated pits.

Gene Symbol (HGNC)	Ensembl Protein ID	Residue	Position	Target Species	Source Type
CLTA	ENSP00000242285	Y	94	WORM	HTP
CLTC	ENSP00000269122	Y	634	FLY	HTP
AP2M1	ENSP00000292807	T	156	FLY	HTP LTP
CLTB	ENSP00000309415	Y	87	WORM	HTP

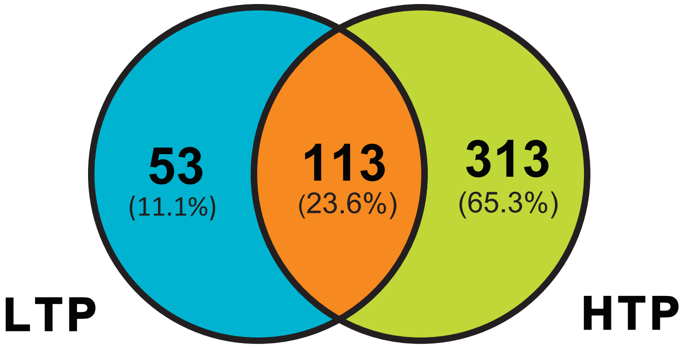

Figure 2.8: Experimental source of the 479 core sites identified in human.

sites were detected in HTP experiments reported in the past 5 years (Figure 2.8) [194, 10, 17, 110, 147, 36, 64]. Moreover, some of these newly discovered and highly conserved sites appear in extensively studied proteins. For example, Thr^{187} in human Wee1 (a major cell cycle regulator kinase) and Ser^{502} in human EEF2 (an essential factor for protein synthesis) with highly conserved flanking regions (defined as the -5 to $+5$ positions of a phosphorylated residue) of 80% and 100% identity, respectively, were conserved from human to fly. These observations suggest that our systematic and comparative approach reveals important clues for deciphering the functional phosphoregulatory events that occur in fundamental cellular processes.

The NetPhorest atlas, which currently consists of 179 probabilistic classifiers trained from known relationships between kinases and phosphorylation sites and *in vitro* proteomics experiments [114], matches experimentally validated phosphorylation sites to probabilistic sequence models of kinase consensus (specificity) motifs. To gain further insight into the regulation of core sites, we deployed the NetPhorest algorithm to delineate the kinases or kinase families that are likely to target human core sites. Although many phosphorylation sites can be targeted by multiple kinases or kinase families [114, 36], here, we restricted our analysis to the top three predictions from NetPhorest that exceeded previously calibrated thresholds [114].

We found that CDK2 and CDK3 kinase family and casein kinase 2 (CK2) were the most frequently predicted kinases, each matching ≈29% of the human core sites. In comparison, only ≈8% and ≈6% of all human phosphorylation sites were matched to CDK2 and CDK3 kinase family and CK2, respectively. The high proportion of core sites predicted to be targeted by CDK2 and CDK3 kinase family and CK2 is not unexpected, because these kinases are functionally pleiotropic [98] and are involved in several fundamental cell processes such as cell survival, proliferation, and differentiation. In addition, we found that kinases involved in the cellular response to stress, such as p38 and c-Jun N-terminal kinase (JNK) family members were predicted to phosphorylate ≈24% and ≈19% of the core sites (compared to ≈7% and ≈5% for all human phosphorylation sites), respectively. Although one might expect ancient kinase families to target the core sites, we did not find strong evidence supporting this. Highly conserved core sites (sites with at least 80% sequence similarity within the flanking region) were predicted to be targeted by kinases of different evolutionary origin, such as the insulin receptor (InsR), Eph family members EphA3 through 6, and the nonreceptor tyrosine kinase Src (all of metazoan origin), and phosphoinositide kinase 1 (PDK1), serum- and glucocorticoid-inducible kinase (SGK), and NEK3 [NIMA (never in mitosis gene a)-related kinase 3] through 5 and 11 (all of primordial origin) [108].

Tracing the conservation of the 479 core sites across 19 eukaryotic species spanning human and the target species in evolution confirmed that the core sites are highly conserved, implying that many core sites are under negative selection and are likely important for fundamental cellular processes. For example, we found that 92.3% of the human core site phosphoresidues were preserved in the distantly related *Xenopus tropicalis* compared to 73.6% of other phosphorylatable residues between the same species. When human and mouse (*Mus musculus*) were compared, these numbers were 97.8% and 90.4%, respectively (Figure 2.9 shows the conservation for the respective residue in selected species). Human tyrosine residues in general are highly conserved probably because of their roles in maintaining protein structure; thus, core site phosphotyrosines do not appear much more conserved than other tyrosines (Figure 2.9).

Although changes in the flanking regions could reveal diverged sequence specificities of kinases that have evolved from yeast to human, such analysis is confounded by the possibility that phosphorylation sites can evolve independently from their effector kinases. For example, the presence of multiple sites on a single substrate targeted by a single kinase could create functional redundancy that allows mutations to accumulate in the phosphorylation sites [61]. Thus, using NetPhorest, we instead analyzed the conservation of kinase (or kinase family) consensus motifs matching the core sites between human and the target species. We estimated the proportion of aligned core site pairs with sufficient conservation within the flanking region for the kinase (or kinase family) predicted for each site of an aligned pair to be identical. This revealed that 67.4% of the aligned site pairs had identical kinases (or families) assigned, and 70% of these sites were predicted to be targeted by the CDK2, CDK3, or CK2 kinases.

Relaxing the analysis to include the top two or three best predictions showed that 81.6% and 86.8% of the core site pairs shared the same kinases or kinase families, respectively. The kinases that regulate the remaining (≈13% to 18%) core sites may have changed during evolution. This potential rewiring of the core phosphorylation networks could enable cells to utilize the same core sites to relay signals from different kinases in response to new environmental cues or stimuli. However, we cannot conclusively argue this point because we do not have consensus motifs for all kinases and thus may miss pairs of aligned sites that match conserved but hitherto unknown phosphorylation (kinase consensus) motifs. To explore this further, we performed an orthogonal analysis by clustering core sites on the basis of sequence similarity within their flanking regions between human and target species to identify potential previously unknown phosphorylation (kinase consensus) motifs. First, we grouped aligned sites that shared similar conserved flanking residues (see Methods). Next, we visualized the grouped core sites as sequence motif logos [114] and manually organized them into proline-based, arginine-

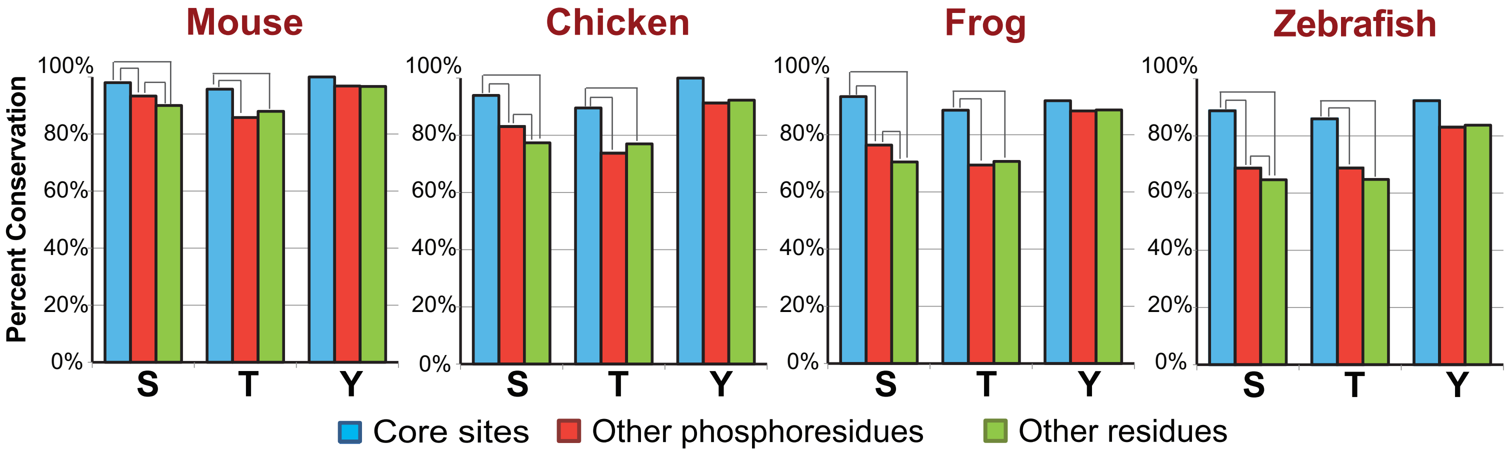

Figure 2.9: Sequence conservation of core sites. Proportion of conserved residues for different subsets of serine (S), threonine (T), and tyrosine (Y) in human core site proteins across orthologous proteins in selected species (*M. musculus*, *G. gallus*, *X. tropicalis*, and *D. rerio*) computed from MSAs (see Methods). Only human phosphoresidues with at least 20% identity from sequence position −5 to +5 of the residue (excluding position 0) to the orthologous sequence in the target species are included in the statistics. Other residues refers to those instances of the specified amino acid that are not known to be phosphorylated. Connectors linking two bars denote that the difference observed is statistically significant ($P < 0.05$, Fishers exact test, one-tailed).

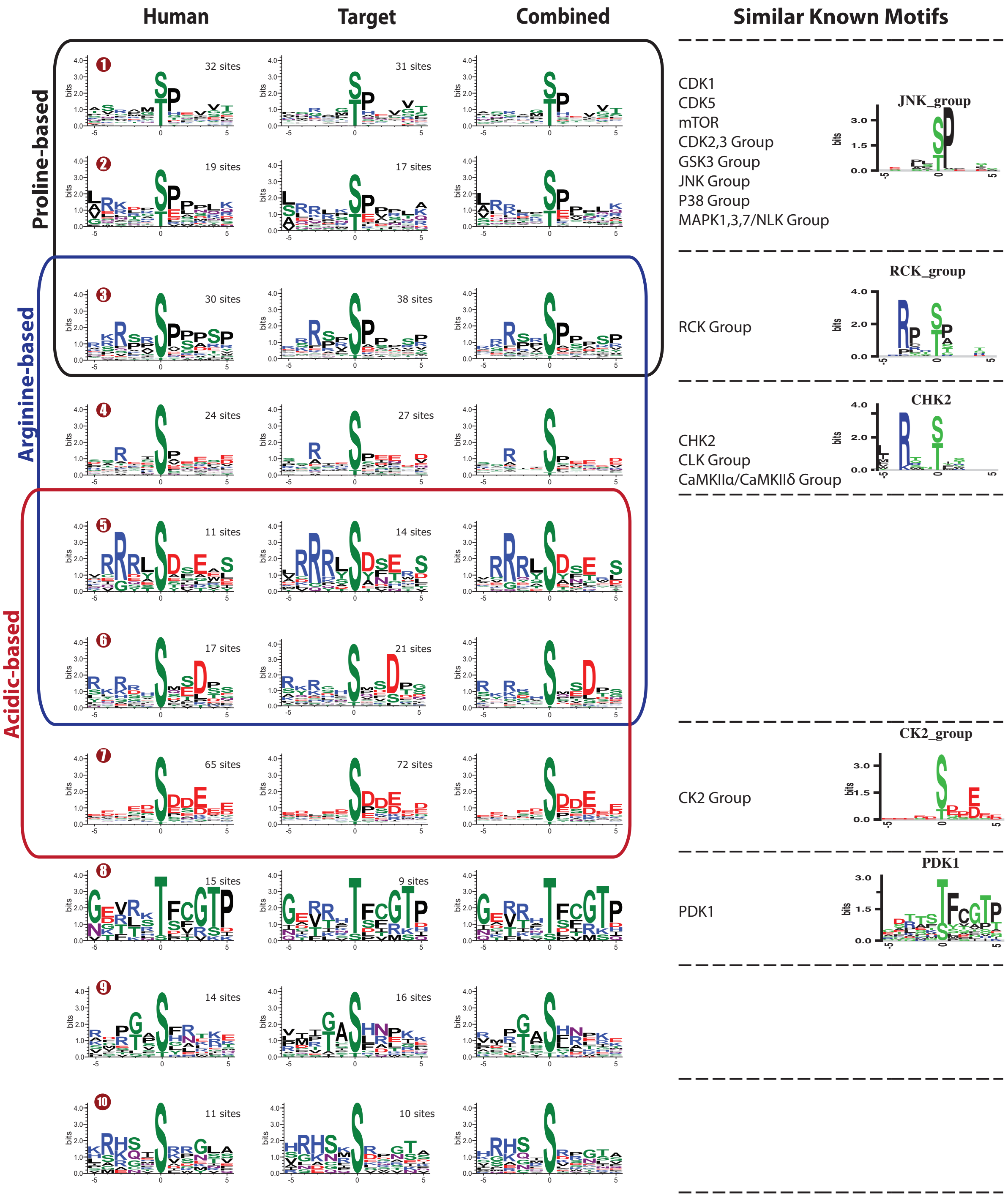

Figure 2.10: Conserved motif analysis of core sites.

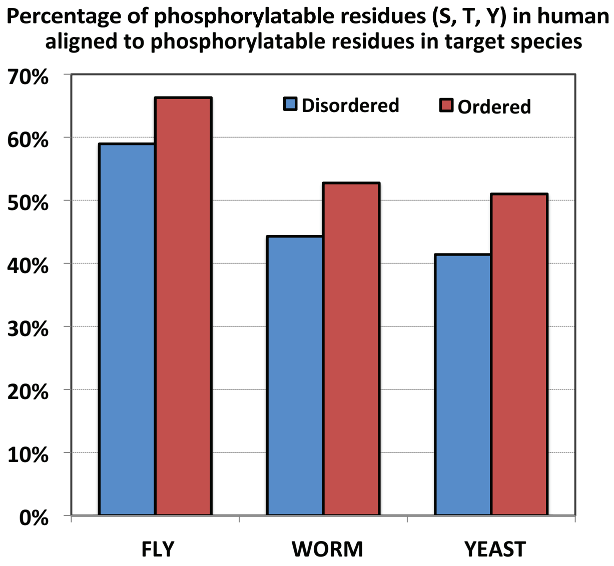

Figure 2.11: Conservation of phosphorylatable residues in structured/unstructured regions.. Intrinsic disordered regions in proteins are generally rapidly evolving and hard to align. However, human phosphorylatable residues (serine, theronine, and tyrosine) in disordered regions that are well-aligned to target species (defined as those with at least 50% identity in -5 to +5 of the central residue) are less conserved than phosphorylatable residues in ordered regions at the same identity threshold. Here, we deemed a phosphorylatable residue in human to be conserved in target species if it is aligned to another phosphorylatable residue. The statistics are computed only from human orthologs in target species that have one-to-one orthologous relationship to avoid relaxation of conservation caused by duplicate genes

based, and acidic-based phosphoserine or phosphothreonine motifs (Figure 2.10). Most of the revealed motifs resembled known human kinase consensus motifs [114], such as that of PDK1, suggesting the possibility of exploiting comparative phosphoproteomics to discover kinase consensus motifs.

2.3.2 Putative evolutionary conserved kinase-substrate phosphorylation network in human

Linear motifs, such as phosphorylation sites, often reside in disordered regions that can change rapidly or undergo convergent evolution [100, 139, 124, 120, 73]. We observed that ≈50% of human phosphorylation sites in proteins with orthologous phosphoproteins were not aligned to phosphorylatable residues in any of the target species (Figure 2.5) and that ≈64% of the sites in these proteins were located in intrinsically disordered regions, in agreement with previous reports [32, 90]. These observations suggest that many phosphorylation sites are fast evolving and, therefore, do not exhibit strong evolutionary conservation at the sequence position level in distantly related organisms. Even within well-aligned regions (as assessed by the overall sequence identity of residues flanking the phosphorylatable residues), we noticed that phosphorylatable residues in disordered regions are less conserved than phosphorylatable residues in ordered regions (Figure 2.11). These properties render phosphorylation sites (linear motifs) located in disordered regions difficult to align and trace during evolution [75, 107, 105, 18], which is further supported by the observation that core sites in disordered regions were underrepresented (Figure 2.12).

A key role of kinases is to modulate cellular signaling networks (for example, by creating binding sites for SH2 domains). Because these events may not require phosphorylation events to occur at precise positions in protein sequences [124, 120, 73, 61], we investigated the evolutionary conservation of phosphorylation at the level of protein networks rather than strictly focusing on the positionally conserved sites in individual proteins. Specifically, we sought to identify phosphorylation events on orthologous proteins that are mediated by orthologous kinases between human and the target species.

The NetworKIN algorithm can computationally reconstruct phosphorylation networks [98] by modeling kinase specificity from contextual information for phosphoproteins and kinases in tandem with sequence models of kinase consensus motifs. The kinome coverage of NetworKIN was extended by the NetPhorest atlas [114]. A potential concern in using these tools on nonhuman data relates to whether the orthologous kinases in yeast, worm, and fly have similar consensus motifs. NetworKIN made reliable predictions in yeast [98] and for several yeast kinases with known human orthologs, the motifs appear identical (B. Turk, personal communication), which is in agreement with the observations reported above for core sites. Furthermore, this conservation of kinase consensus motifs is expected

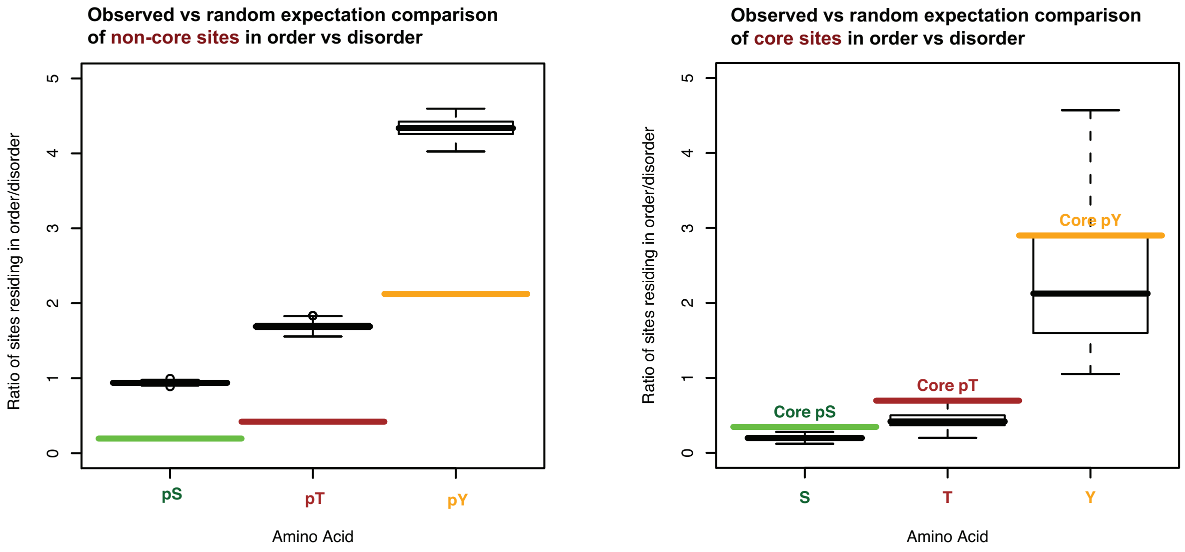

Figure 2.12: Occurrence of core sites in structured/unstructured regions. Known human phosphorylation sites occur more often in disordered regions than would be expected by random chance. The boxplots (in black) display the distributions of in-order/in disorder ratios for 100 randomly picked sets of serines, threonines, and tyrosines from human phosphoproteins. Phosphosphorylated serines, threonines, and tyrosines might fall by random chance into these sets as well. The size of each set equals that of the respective observed number of phosphorylated residues. The colored bars denote the observed in-order/in-disorder ratio of phosphorylated residues. Core sites occur more often in ordered or structured regions than do other phosphorylation sites. To assess the significance of this observation, we sampled the distribution of in-order/in-disorder ratio with sets of randomly picked phosphorylation sites from core site proteins. The distributions of the random sampling are shown as boxplots (in black) and the colored bars denote the in-order/in-disorder ratio of core sites. Separate analyses were performed for phosphoserine, -threonine, and -tyrosine.

from evolutionary principles: Consensus motifs of pleiotropic kinases [98] must be under strong selective pressure because a motif change could potentially affect the complete target and function space for that kinase. Finally, NetworKIN filters predictions on the basis of context; thus, even if a motif falsely matches a site in a target species, it is not likely that the context data would allow inclusion of this prediction.

By deploying the NetworKIN [98] algorithm in combination with NetPhorest [114], we predicted protein kinases for all phosphoproteins identified in human and the target species. We used the default parameters for NetworKIN (see Methods), which allows a single site to be phosphorylated by multiple kinases and then overlaid the human phosphorylation network with those of the target species (Figure 2.3) to obtain a human phosphorylation network limited to those phosphoproteins and kinases that were conserved in at least one target species (core net). We further quantified phosphorylation conservation by defining a propensity (denoted as k) for each human substrate, which represented the number of a human substrates phosphorylation events that were supported by orthologous kinase-substrate relationships in target species (Figure 2.4). Thus, k captures the phosphorylation events on a human protein that are supported by orthologous (conserved) kinase-substrate relationships predicted in the target species. Due to gene duplication that occurred along the lineages of human and target species, multiple kinase-substrate relationships in human may be supported by single kinase-substrate relationship in target species. Conversely, a single kinase-substrate relationship in human may be supported by multiple kinase-substrate relationships in target species.

The initial ($k \geq 0$) human phosphorylation network contained 25,563 interactions between 113 kinases and 5,515 substrates, whereas the human phosphorylation network resulting from overlaying the networks from the target species, for $k > 0$, had 1,255 interactions between 27 human kinases and 778 substrates (encoded by 759 genes, termed core net genes), of which 1,105 interactions (88%) and 698 substrates (encoded by 682 genes) were not attributed to core sites. Randomized network analysis (see Methods) revealed that this overlap was unlikely to occur by chance (empirical $P < 0.001$; Figure 2.13). Figure 2.14 shows the subset of the inferred conserved human phosphorylation network with $k > 6$.

2.3.3 Association of identified ancient phosphoproteins with cancers and OMIM diseases

The two methods yielded different but somewhat overlapping sets of genes. The alignment-based approach identified the 337 core site genes and the kinase-substrate, network-based approach identified the larger set of the 759 core net genes, which included 525 genes that

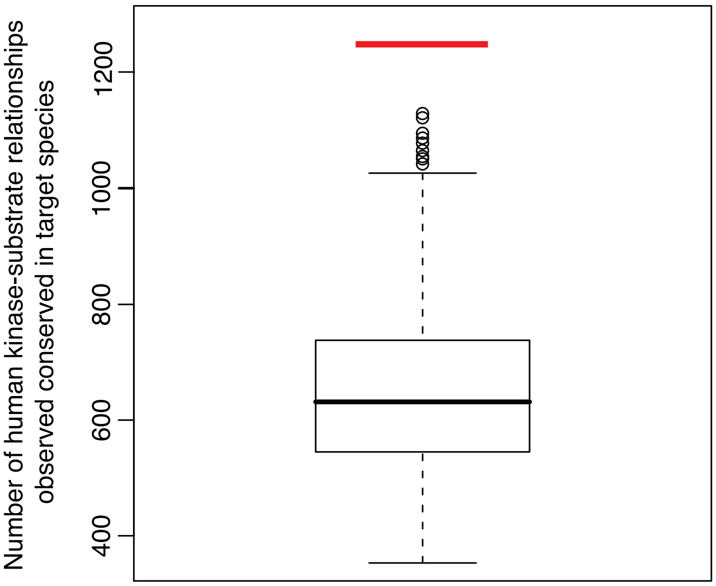

Figure 2.13: Distribution of number of human substrate relations observed as conserved in target species based on randomized trials. The distribution from 2,000 random trials (see Section 2.2.9) is visualized here using box plot. The top and bottom of the box are the 25th and 75th percentile of the observed distribution respectively while the horizontal line in the middle of the box indicates the 50th percentile. The highest bar and lowest bar perpendicular to the dotted line in the center indicate the highest and lowest data point are within 1.9 interquartile range (IQR) of the upper and lower quartile respectively. The red bar denotes the actual number of observed conserved human kinase-substrate relationships.

were not part of the core site gene set. We analyzed each of these gene sets to determine if they were enriched in genes associated with cancer.

First, human genes encoding phosphoproteins were statistically enriched in cancer-associated genes (see Methods) over the entire protein set (4.5% versus 1.8% background, $P < 0.05$, hypergeometric test; Figure 2.15). However, the core site gene set was more enriched in cancer-associated (see Methods) genes over the entire set of genes encoding phosphoproteins ($P = 0.05$, hypergeometric test; Figure 2.15). This enrichment occurred despite the fact that the subset of human genes encoding phosphoproteins with orthologous proteins in target species was not more enriched in cancer-associated genes than the entire set of human phosphoproteins, regardless of whether their orthologous proteins are phosphorylated (4.1% versus 4.5%) or not (3.8% versus 4.5%). We spec-

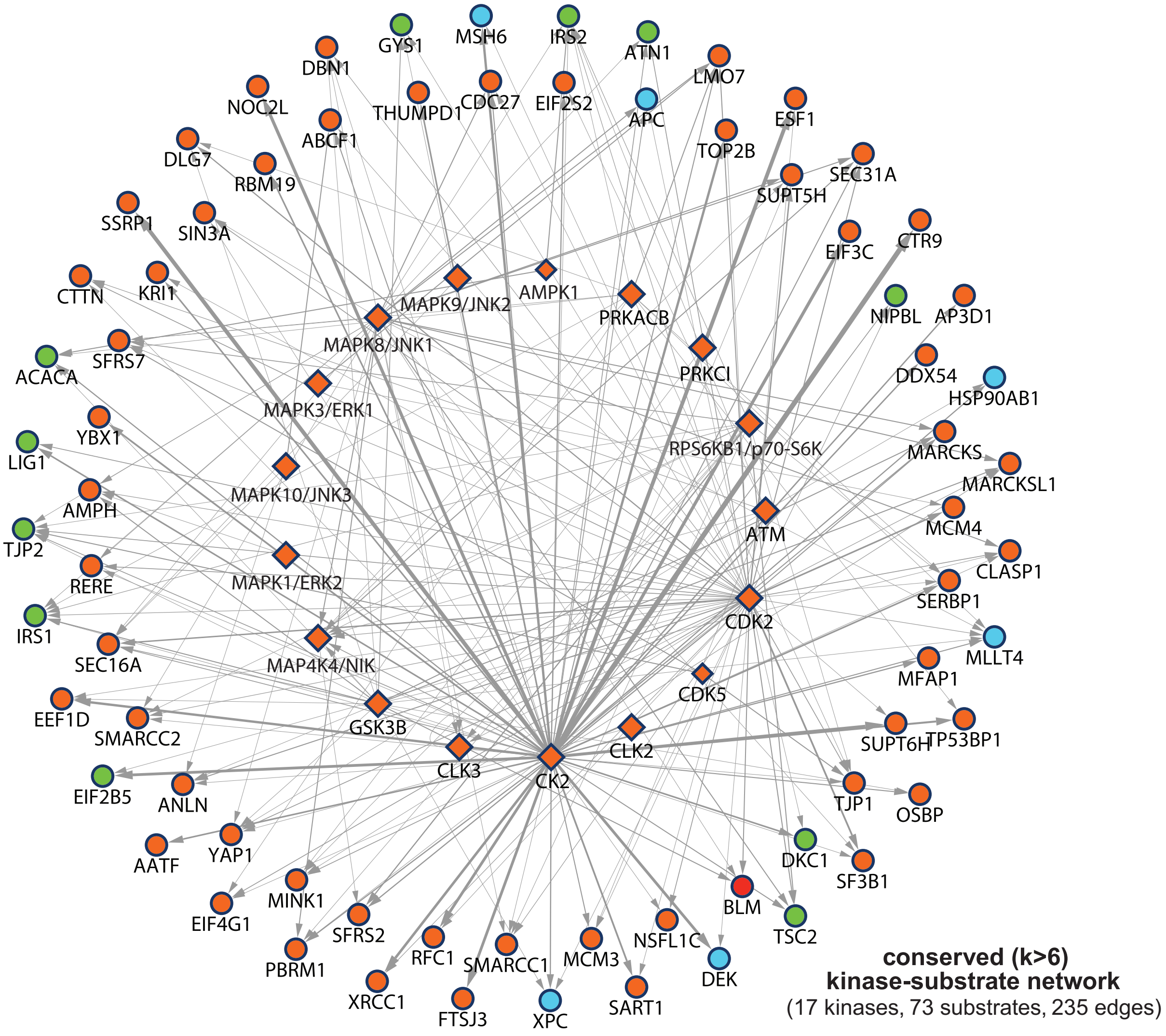

Figure 2.14: Identified kinase-substrate interaction network involving conserved phosphorylation protein hubs. Increasingly conserved human phosphorylation networks could be isolated on the basis of increasing k. Here, we show a conserved human phosphorylation network of $k > 6$. The thickness of the edges corresponds to the number of conserved interactions between the kinase and substrate across the target species. Diamond nodes represent kinases predicted to target the phosphoproteins. Proteins known to be implicated in cancer and other diseases are colored blue and green, respectively.

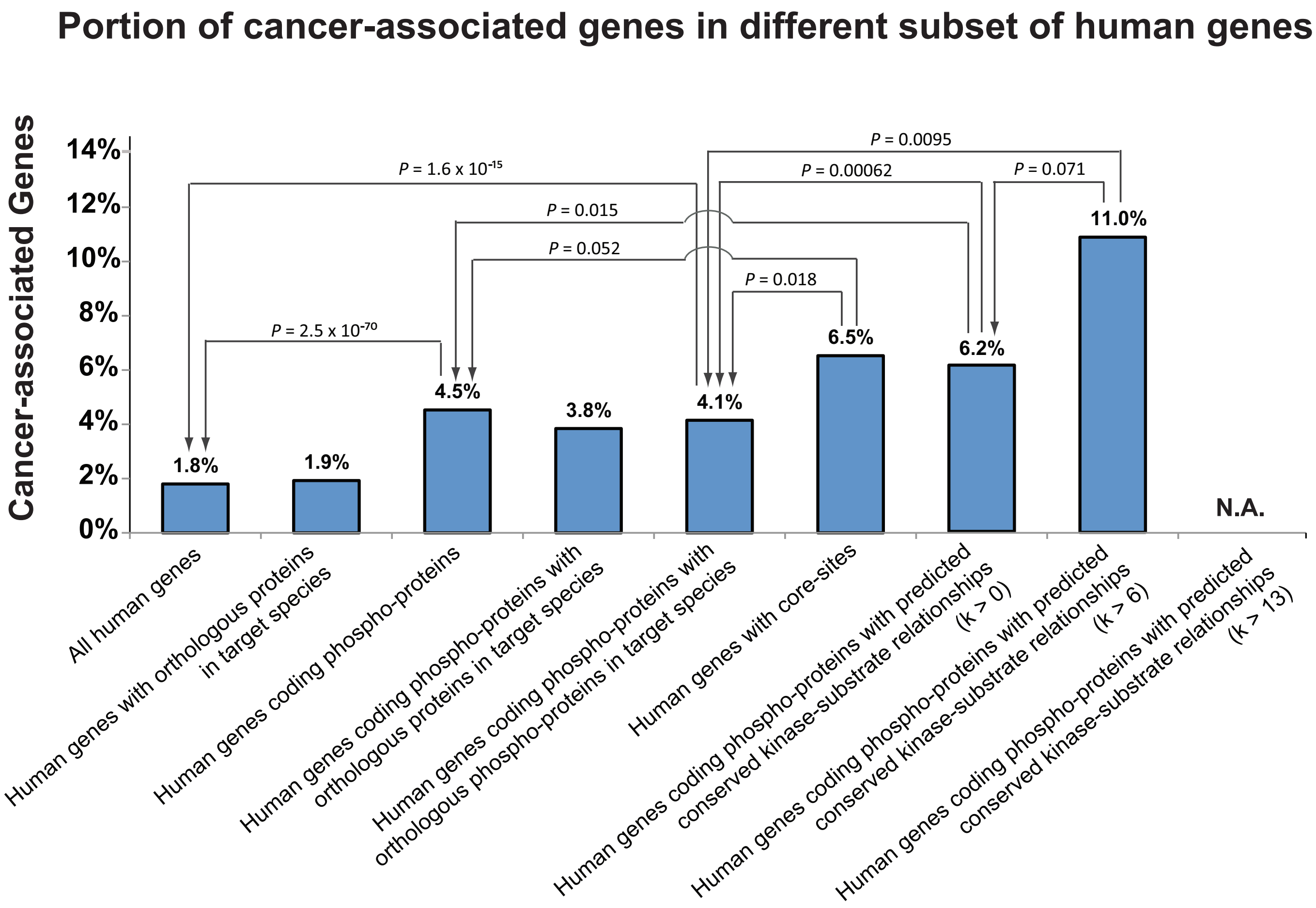

Figure 2.15: Evolutionary conserved phosphorylation proteins are significantly encoded by cancer-associated genes. Human phosphoproteins are enriched in cancer-related genes but both core site genes and core net genes ($k > 0$) are statistically more enriched in cancer-associated genes than background phosphoproteins (top: hypergeometric test, with the protein group of the arrow target used as background). In addition, we observed that core net genes with a higher k are more enriched in cancer-associated genes.

ulate that some core sites in the products of these genes may be aberrantly regulated in transformed cells. For example, phosphorylation of the core site Ser^{315} in FOXO3A by SGK1 prevents FOXO3A from inducing cell cycle arrest and apoptosis [24], thereby promoting cell proliferation. Hence, it is plausible that deregulated phosphorylation of Ser^{315} in FOXO3A could contribute to neoplastic growth. Intriguingly, 15 core sites in these cancer-associated genes were only recently detected in large-scale MS experiments (Table 2.4) which suggests that investigation of these sites may provide clues to further understand the functional role of these proteins in normal and malignant cells.

Similarly, core net genes were statistically enriched for cancer-associated genes ($P = 1.5 \times 10^{-2}$ when compared to all human genes encoding phosphoproteins and $P = 6.2 \times 10^{-4}$ when compared to human genes encoding phosphoproteins with orthologous phosphoproteins in the target species, hypergeometric test; Figure 2.15), identifying approximately one fold more cancer-associated genes than did the alignment-based method (47 versus 22) with a slight drop in specificity (6.5% versus 6.2%; Figure 2.15). This suggests that the network comparison approach can identify potentially important phosphorylation events occurring in less conserved protein regions. Note that the predicted conserved effector kinases of phosphoproteins from the 759 genes were not included in the enrichment analysis unless they were among the 759 genes. In total, 52 unique cancer-associated genes were identified in the combined set of core site and core net genes.

Analysis of the topological features of predicted human phosphorylation networks revealed that the number of kinase-substrate relationships of human phosphoproteins correlated positively with enrichment in cancer-associated genes in the entire human phosphoproteome, as well as to a lesser degree in its HTP subset (Figure 2.17, top graphs). A weaker positive correlation was observed for other diseases in general, as defined in Online Mendelian Inheritance in Man (OMIM) (Figure 2.17, bottom graphs). Hence, a highly phosphorylated regulatory hub protein is more likely to be encoded by a gene implicated in disease. Moreover, there seems to be a strong linear correlation

Table 2.4: Core sites identified in 22 cancer-associated genes.

Gene Symbol (HGNC)	Ensembl Protein ID	Residue	Position	Target Species	Source Type
APC	ENSP00000257430	S	1279	FLY	LTP
CCDC6	ENSP00000263102	S	367	FLY	HTP
CCDC6	ENSP00000263102	S	240	FLY	HTP
CCDC6	ENSP00000263102	S	323	FLY	HTP
CDK2	ENSP00000266970	T	160	FLY	HTP LTP
CDK2	ENSP00000266970	Y	15	YEAST	HTP LTP
CLTC	ENSP00000269122	Y	634	FLY	HTP
CREB1	ENSP00000236996	S	117	WORM	LTP
CTNNB1	ENSP00000344456	S	675	FLY	HTP LTP
DEK	ENSP00000244776	S	230	FLY	HTP
DEK	ENSP00000244776	S	231	FLY	HTP
DEK	ENSP00000244776	S	306	FLY	HTP
DEK	ENSP00000244776	S	232	FLY	HTP
FIP1L1	ENSP00000351383	S	85	FLY	HTP
FOXO1	ENSP00000368880	S	319	WORM	LTP
FOXO3	ENSP00000339527	S	315	WORM	LTP
HSP90AA2	ENSP00000216281	S	231	FLY	HTP LTP
HSP90AA2	ENSP00000216281	Y	627	FLY	HTP
HSP90AB1	ENSP00000360709	Y	619	FLY	HTP
HSP90AB1	ENSP00000360709	S	226	FLY	HTP LTP
MAP2K4	ENSP00000262445	S	257	FLY	HTP LTP
MLLT4	ENSP00000345834	S	1090	FLY	HTP
FOX04	ENSP00000363377	S	262	WORM	LTP
MSH6	ENSP00000234420	S	14	FLY	HTP
RPS10	ENSP00000347271	T	101	YEAST	HTP
SEPT6	ENSP00000341524	T	418	FLY	HTP
SMAD2	ENSP00000262160	T	8	FLY	HTP LTP
SUZ12	ENSP00000316578	S	583	FLY	HTP
TCF12	ENSP00000267811	S	559	FLY	HTP
VCP	ENSP00000367954	S	748	YEAST	LTP

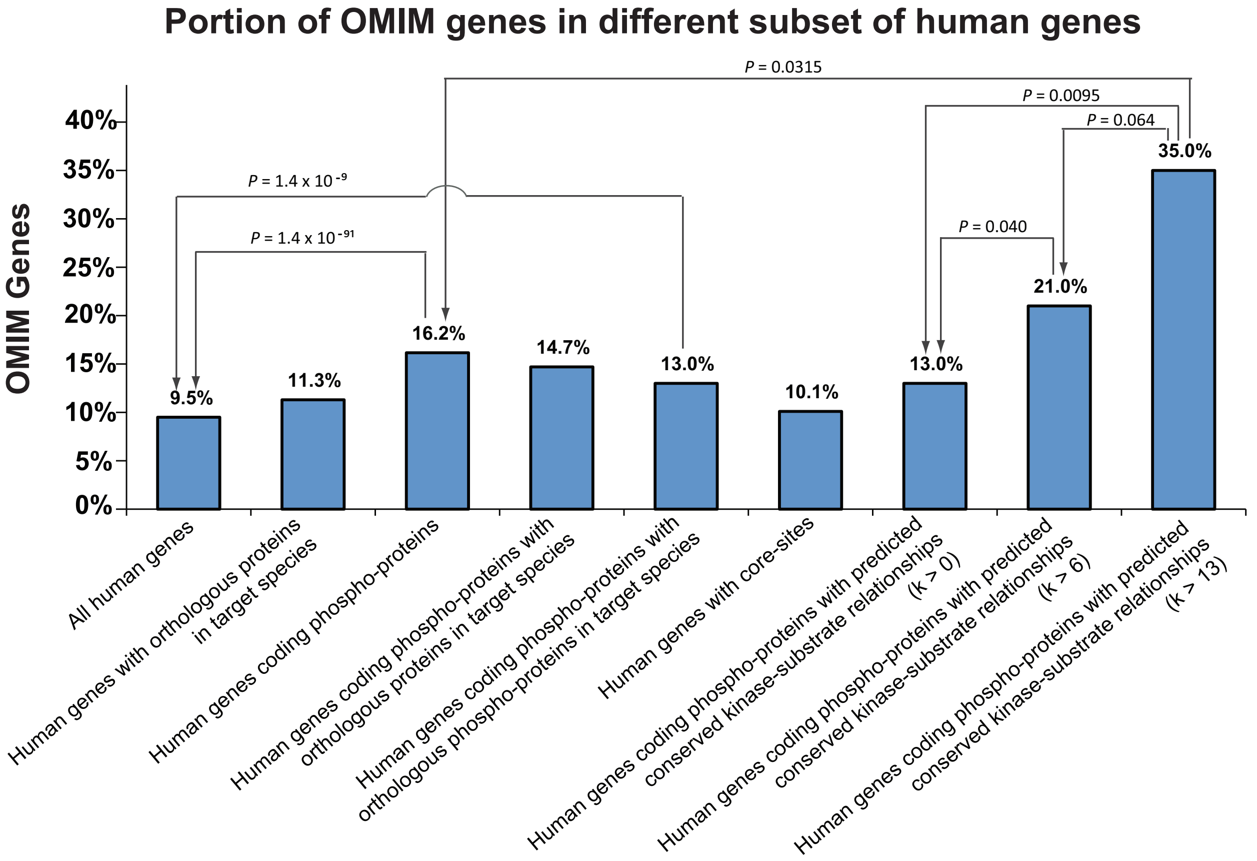

Figure 2.16: Evolutionary conserved phosphorylation proteins are significantly encoded by disease genes annotated in OMIM. Human phosphoproteins are enriched in cancer-related genes but both core site genes and core net genes (k > 0) are statistically more enriched in cancer-associated genes than background phosphoproteins (top: hypergeometric test, with the protein group of the arrow target used as background). In addition, we observed that core net genes with a higher k are more enriched in cancer-associated genes.

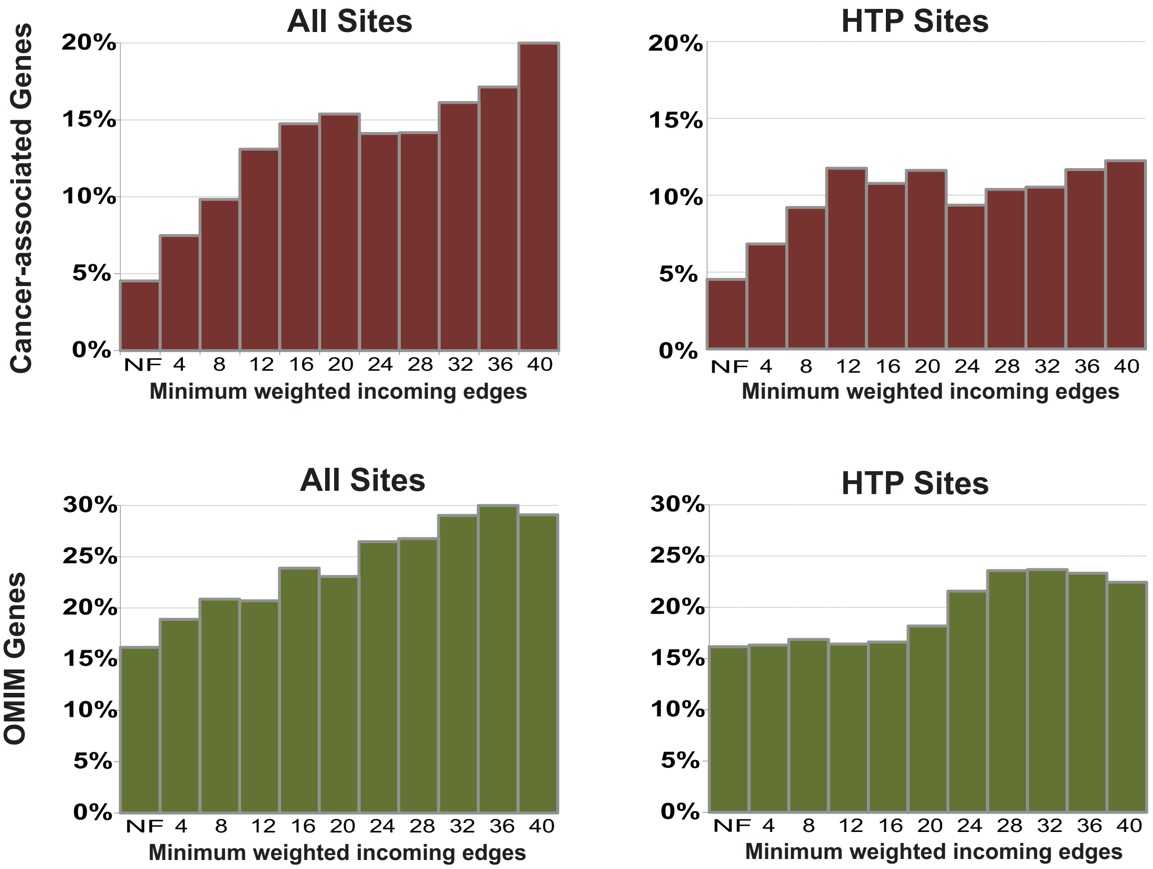

Figure 2.17: Increasing disease enrichment of phosphorylation protein hubs. Genes encoding human signaling hub proteins are enriched in disease genes. A directed kinase-substrate regulatory network is first inferred from assembled human phosphorylation data by NetworKIN. A phosphorylation propensity score n is computed for each gene, which is the sum of weighted incoming edges of kinases phosphorylating the genes products. The weight of an incoming edge from each kinase to a gene is defined as the number of sites in the genes products inferred to be targeted by a kinase. Human genes are then filtered by this score n to assess association with cancer-associated and disease genes from OMIM. n is computed from the entire set of human phosphorylation, as well as its subset from HTP studies. **NF**: not filtered.

between this likelihood and the signal integration properties of the proteins. A concern was that this observation could stem from ascertainment bias because disease genes are extensively studied. We therefore interrogated the human phosphoproteome with the conservation phosphorylation propensity (k) associated with each human protein, given that the measure is computed with target phosphoproteomes from unbiased systematic studies. We found that k correlated positively with cancer-associated genes and OMIM disease genes (Figure 2.15 and 2.16). Thus, it appears that genes encoding proteins that receive and integrate many signaling events have an increased tendency to be implicated in disease, which agrees with similar suggestions [184, 141], and that their signal integration properties are conserved in the target species. Another possibility is that these genes encode products that need to be tightly regulated by protein phosphorylation in human and target species and that are vulnerable to deregulation likely caused by mutations or changes in protein abundance.

Accordingly, we identified in the core net proteins that are involved in several complex diseases, which may be suitable for experimental and therapeutic studies. We identified proteins related to Alzheimers disease, SEPT1 ($k = 4$) and DBN1 ($k = 7$), which are supported by evidence that misregulation of phosphorylation is important in neurological disorders (44). We identified proteins related to viral infection, the human immunodeficiency virus 1 (HIV-1) infection-related proteins SFRS2, SFRS5, and SFRS7 ($k = 13$, 6, and 13, respectively). We identified proteins associated with the cell polarity, TJP1, TJP2 and MINK1 ($k = 10$, 17, and 7, respectively). We identified proteins implicated in controlling cell and organ size, the Hippo-associated protein YAP1 ($k = 12$), and metabolism, the insulin receptor substrate proteins IRS1 and IRS2 ($k = 16$ and 14, respectively). All these proteins are predicted substrates for the following kinases that are involved in the same set of diseases: CDK2 (cancer and HIV infection), MAP4K4 (cancer and insulin resistance), ATM (cancer), PRKACA and GSK3 (diabetes, cancer, Alzheimers disease, and HIV), MAPK8 (HIV infection and Alzheimers disease), and

Table 2.5: Correlation of cancer-associated genes with conserved phosphorylation propensity k computed for individual query species.

	Fly + Worm + Yeast			Fly			Worm			Yeast		
k	Cancer Genes	Total Genes	%	Cancer Genes	Total Genes	%	Cancer Genes	Total Genes	%	Cancer Genes	Total Genes	%
≥ 1	47	759	6.2%	44	631	7.0%	3	96	3.1%	6	215	2.8%
≥ 2	25	450	5.6%	21	344	6.1%	1	49	2.0%	3	101	3.0%
≥ 3	15	271	5.5%	15	209	7.2%	0	17	0.0%	1	41	2.4%
≥ 4	12	192	6.2%	12	138	8.7%	0	14	0.0%	1	25	3.6%
≥ 5	11	143	7.7%	10	93	10.8%	0	6	0.0%	1	15	6.7%
≥ 6	10	103	9.7%	8	70	11.4%	0	3	0.0%	1	9	11.1%
≥ 7	8	73	11.0%	7	47	14.9%	0	3	0.0%	0	7	0.0%
≥ 8	4	54	7.4%	4	31	12.9%	0	3	0.0%	0	5	0.0%

Table 2.6: Correlation of OMIM disease genes with conserved phosphorylation propensity k computed for different query species.

	Fly + Worm + Yeast			Fly			Worm			Yeast		
k	Cancer Genes	Total Genes	%	Cancer Genes	Total Genes	%	Cancer Genes	Total Genes	%	Cancer Genes	Total Genes	%
≥ 1	99	759	13.0%	85	631	13.5%	8	96	8.3%	27	215	12.6%
≥ 2	55	450	12.2%	42	344	12.2%	5	49	10.2%	14	101	13.9%
≥ 3	34	271	12.5%	26	209	12.4%	1	17	5.9%	7	41	17.1%
≥ 4	24	192	12.5%	20	138	14.5%	1	14	7.1%	4	25	16.0%
≥ 5	23	143	16.1%	19	93	20.4%	0	6	0.0%	3	15	20.0%
≥ 6	18	103	17.5%	16	70	22.9%	0	3	0.0%	3	9	33.3%
≥ 7	15	73	20.5%	14	47	29.8%	0	3	0.0%	2	7	28.6%
≥ 8	13	54	24.1%	10	31	32.3%	0	3	0.0%	1	5	20.0%
≥ 9	11	41	26.8%	5	20	25.0%	0	2	0.0%	0	4	0.0%
≥ 10	10	35	28.6%	4	16	25.0%	0	2	0.0%	0	4	0.0%
≥ 11	8	27	29.6%	4	11	36.4%	0	2	0.0%	0	1	0.0%
≥ 12	7	24	29.2%	4	11	36.4%	0	1	0.0%	0	1	0.0%
≥ 13	7	22	31.8%	4	9	44.4%	0	1	0.0%	0	0	0.0%
≥ 14	7	20	35.0%	4	6	66.7%	0	1	0.0%	0	0	0.0%
≥ 15	4	16	25.0%	3	5	60.0%	0	1	0.0%	0	0	0.0%

RPS6KB1 (RNA splicing and HIV infection).

2.3.4 Function enrichment analysis

Given the disease associations observed in both core site and core net gene sets, we investigated the prevalence of phosphorylation conservation at both the site and the network levels across different cellular functions. That is, we aimed to identify cellular processes in which phosphorylation events are preferentially positionally conserved across orthologs or preferentially mediated by orthologous kinases (conserved kinase-substrate relationship). Specifically, we compared functions of core site genes and core net genes against the complete set of human phosphoproteins that are orthologous to known phosphoproteins in the target species. There are a total of 337 core site genes and 758 core net genes with 233 common genes between the two gene sets. We find that core site genes are statistically enriched (hypergeometric test, Benjamini and Hochberg false discovery rate correction) in genes encoding proteins involved in amino acid phosphorylation ($P = 8.0 \times 10^{-5}$) and RNA splicing ($P = 1.9 \times 10^{-3}$) and encoding cytosolic ribosomal proteins ($P = 2.0 \times 10^{-2}$). Manual inspection revealed that of the core sites present in protein kinases, 26 are located within activation loop regions, which are important for the regulation of kinase activity (Figure 2.18). Hence, some core sites are structurally constrained for allosteric regulation, suggesting why this particular subset is positionally conserved.

In contrast, core net genes were enriched (hypergeometric test, Benjamini and Hochberg false discovery rate correction) in genes associated with the cell cycle ($P = 1.4 \times 10^{-4}$), chromosome organization and biogenesis ($P = 5.4 \times 10^{-4}$), DNA-dependent regulation of transcription ($P = 3.3 \times 10^{-3}$), macromolecular complex assembly ($P = 2.4 \times 10^{-3}$), and protein targeting ($P = 1.6 \times 10^{-2}$). In 403 out of 688 core net genes with localization annotation, core net genes were strongly enriched in genes encoding proteins that localize to the nucleus ($P = 1.7 \times 10^{-15}$), which correlates with the finding that core net genes are

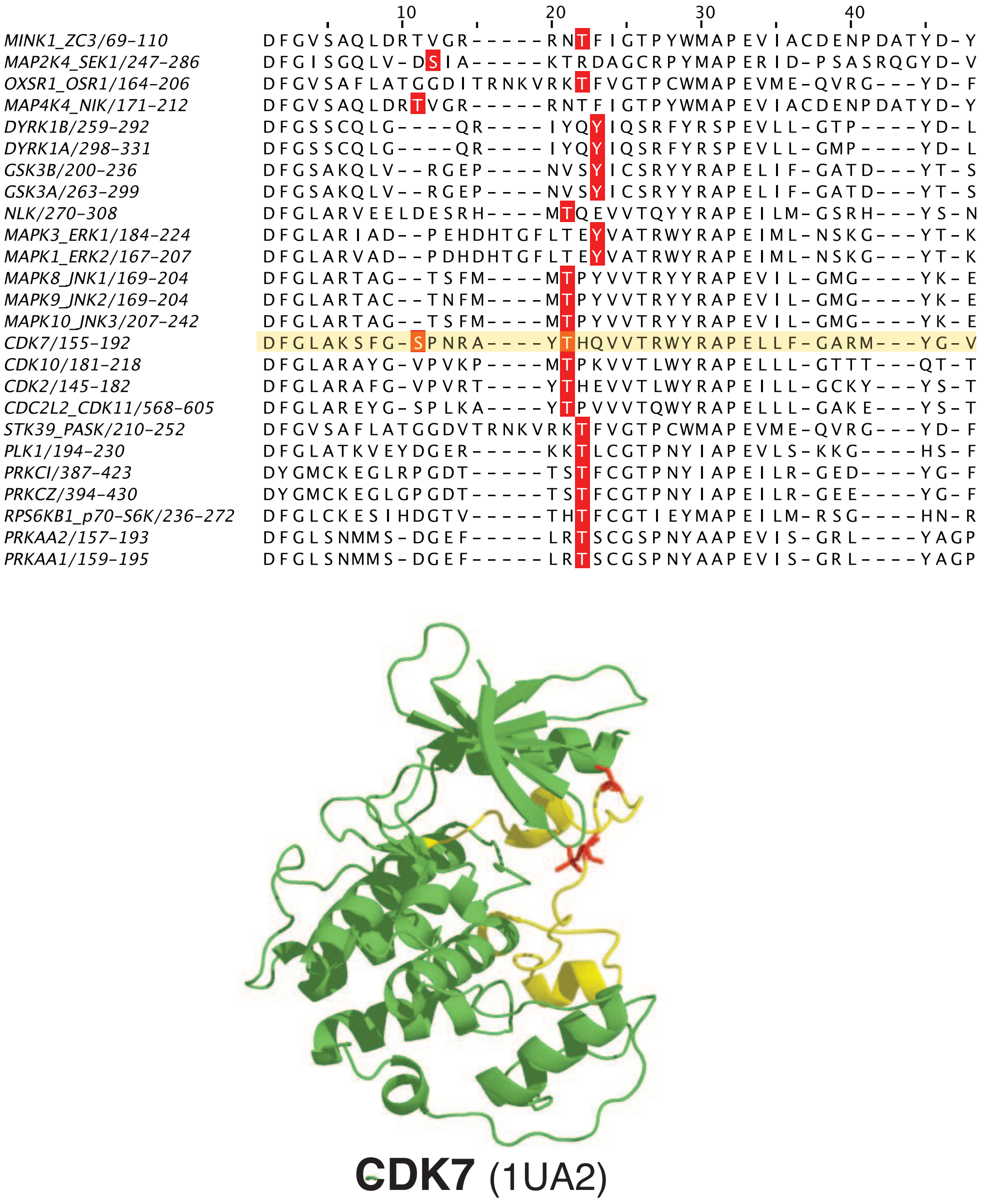
10
20
30
40
MINK1_ZC3/69-110 DFGVSAQLDRTVGR-----RNTFIGTPYWMAPEVIACDENPDATYD-Y
MAP2K4_SEK1/247-286 DFGISGQLV-DSIA-----KTRDAGCRPYMAPERID-PSASRQGYD-V
OXSR1_OSR1/164-206 DFGVSAFLATGGDITRNKVRKTFVGTPCWMAPEVME-QVRG---YD-F
MAP4K4_NIK/171-212 DFGVSAQLDRTVGR-----RNTFIGTPYWMAPEVIACDENPDATYD-Y
DYRK1B/259-292 DFGSSCQLG----QR----IYQYIQSRFYRSPEVLL-GTP----YD-L
DYRK1A/298-331 DFGSSCQLG----QR----IYQYIQSRFYRSPEVLL-GMP----YD-L
GSK3B/200-236 DFGSAKQLV--RGEP----NVSYICSRYYRAPELIF-GATD---YT-S
GSK3A/263-299 DFGSAKQLV--RGEP----NVSYICSRYYRAPELIF-GATD---YT-S
NLK/270-308 DFGLARVEELDESRH----MTQEVVTQYYRAPEILM-GSRH---YS-N
MAPK3_ERK1/184-224 DFGLARIAD--PEHDHTGFLTEYVATRWYRAPEIML-NSKG---YT-K
MAPK1_ERK2/167-207 DFGLARVAD--PDHDHTGFLTEYVATRWYRAPEIML-NSKG---YT-K
MAPK8_JNK1/169-204 DFGLARTAG--TSFM----MTPYVVTRYYRAPEVIL-GMG----YK-E
MAPK9_JNK2/169-204 DFGLARTAC--TNFM----MTPYVVTRYYRAPEVIL-GMG----YK-E
MAPK10_JNK3/207-242 DFGLARTAG--TSFM----MTPYVVTRYYRAPEVIL-GMG----YK-E
CDK7/155-192 DFGLAKSFG-SPNRA----YTHQVVTRWYRAPELLF-GARM---YG-V
CDK10/181-218 DFGLARAYG-VPVKP----MTPKVVTLWYRAPELLL-GTTT---QT-T
CDK2/145-182 DFGLARAFG-VPVRT----YTHEVVTLWYRAPEILL-GCKY---YS-T
CDC2L2_CDK11/568-605 DFGLAREYG-SPLKA----YTPVVVTQWYRAPELLL-GAKE---YS-T
STK39_PASK/210-252 DFGVSAFLATGGDVTRNKVRKTFVGTPCWMAPEVME-QVRG---YD-F
PLK1/194-230 DFGLATKVEYDGER-----KKTLCGTPNYIAPEVLS-KKG----HS-F
PRKCI/387-423 DYGMCKEGLRPGDT-----TSTFCGTPNYIAPEILR-GED----YG-F
PRKCZ/394-430 DYGMCKEGLGPGDT-----TSTFCGTPNYIAPEILR-GEE----YG-F
RPS6KB1_p70-S6K/236-272 DFGLCKESIHDGTV-----THTFCGTIEYMAPEILM-RSG----HN-R
PRKAA2/157-193 DFGLSNMMS-DGEF-----LRTSCGSPNYAAPEVIS-GRL----YAGP
PRKAA1/159-195 DFGLSNMMS-DGEF-----LRTSCGSPNYAAPEVIS-GRL----YAGP
CDK7 (1UA2)

Figure 2.18: Core sites identified in activation loops of protein kinases. Top: A multiple sequence alignment of activation loop regions in protein kinases with core sites marked in red. Gene symbols, kinase names, and the sequence coordinates for each activation loop is shown on the left. Bottom: The corresponding region from panel A is highlighted (yellow) on the protein structure of CDK7.

strongly associated with chromosome organization and biogenesis and DNA-dependent regulation of transcription. Correspondingly, our results support the notion that functional conservation of phosphorylation does not necessitate positional conservation: For example, protein phosphorylation in cell cycle associated proteins in yeast can be conserved, yet dynamic as a result of site relocation [73]. Correspondingly, our analysis of the core net sites has identified more cellular activities that may be subject to a similar mode of evolution in phosphorylation regulation.

2.4 Discussion

To assess the evolutionary history of phosphorylation sites, it is essential to appreciate that the lack of evidence for a phosphorylation event does not infer a nonphosphorylated site. Rather, the site could be phosphorylated but below the limits currently detectable; alternatively, phosphorylation may depend on a missing environmental cue or the site may become dephosphosphorylated under the experimental conditions used. In addition, some sites are only phosphorylated in specific cell types, rendering their detection even more difficult. Thus, phosphorylation events are highly context dependent [98, 78] and dynamic [194]. Indeed, Gygi and co-workers derived a phosphoproteome from fly embryos and compared it to the one used here (derived from the Kc-167 cell line) [203] and found about 25% overlap despite the fact that the same species was analyzed. Although this difference can partly be explained by different experimental and computational pipelines, it undoubtedly also reflects differences between the biological systems studied (for example, complete embryos contain many specialized cells in contrast to the defined Kc-167 cell line). This highlights that a large number of additional phosphorylation sites are likely to be discovered by continued and improved phosphoproteomic analysis. In particular, studies of the utilization, dynamics, and functional roles of the sites will be important [88, 154], because these reflect cellular information processing much more directly than the number of sites itself.

Although we began with over 40,000 combined phosphorylation sites in both human and the target species (yeast,worm, and fly), we only identified 479 positionally conserved phosphorylation events in human. About 45% (10,969) of the non-core sites were found in human phosphoproteins with no detectable orthologous proteins in any of the target species. Of the remaining 13,008 human phosphoresidues, only 4448 aligned to phosphorylatable residues in at least one target species, of which 479 are aligned to phosphorylated residues (Figure 2.5). This limited overlap is presumably due to the large evolutionary distance (more than 600 million years [57]) and actual physiological differences between human and the target species and the incomplete coverage of the phosphorylation mapping data sets due to mass spectrometer limitations (for example, sensitivity) and the limited number of experimental conditions or biological contexts analyzed (for example, developmental stages).

The incompleteness of the data is illustrated by the composition of the phosphorylated residues of the human phosphoproteome analyzed here (Figure 2.1), which is biased toward *p*Y [39% observed versus an average of about 4% observed in the target species phosphorylation data and in other large-scale phosphoproteomic studies [129, 203]]. This overrepresentation of *p*Y can be attributed to the use of *p*Y-antibodies in several HTP studies [for example, [147] and to the notion that phosphotyrosine peptides are more easily detected with MS than are phosphoserine- or phosphothreonine containing peptides [10]. Therefore, this *p*Y-overrepresentation can be used to estimate a lower bound on the coverage of the human phosphorylation data, by computing how many additional (*p*S and *p*T) sites are needed to dilute the fraction of *p*Y down to the large-scale average of 4% phosphotyrosine (which is likely an overestimate). Thus, we estimate that there are at least 200,000 more *p*S and *p*T sites yet to be discovered in the human phosphoproteome. This estimate will rise with additional discovery of *p*Y sites. A caveat here is that many phosphorylation events are detected in transformed cells, or cells exposed to growth factors or other stimuli [10, 129], which is likely to change the relative amounts of pY, pT, and pS compared to those of nonstimulated cells, as observed by Hunter and co-workers [67]. Nonetheless, this estimate illustrates the incompleteness of current phosphoproteomic data. In addition, 160 core sites were only found in one target species, although phosphorylatable (S, T, or Y) residues are present at orthologous positions in at least one other target species.

Our analysis supports the notion that many phosphorylation sites evolve quickly [90, 11] and, therefore, lack strong conservation at the sequence and position levels – 65% of human phosphorylated residues in proteins with orthologous proteins in target species were not aligned to phosphorylatable residues. Some of the missed phosphorylation sites could be phosphorylation events that need not be positionally conserved [124, 120, 73, 122, 85]. To this end, we investigated the conservation of sites involved in regulatory networks by overlaying predicted kinase-substrate relationships. The two complementary approaches (network versus site alignment) highlight phosphorylation events that are conserved across species spanning long evolutionary distances and, hence, are likely functionally important for fundamental cellular activities. The utility of these approaches is highlighted by the identification of multiple low-abundance signaling proteins and disease-related genes. Consequently, we identify genes encoding products that need to be tightly regulated by protein phosphorylation in human and target species and that are vulnerable to deregulation likely caused by mutations or changes in protein abundance. We did not see any enrichment in disease association in the subset of human phosphoproteins with orthologs that are phosphorylated in our phosphoproteomes over the background set of all human phosphoproteins. In contrast, both the network and the site alignment approaches identified a subset of genes encoding phosphoproteins that were significantly more enriched in disease-associated genes over the entire set of human phosphoproteins. We also noticed that core site genes were not enriched in OMIM disease genes compared to the global set of genes encoding phosphoproteins with orthologous proteins in the target species, regardless of whether the orthologous proteins are phosphorylated or not. On the contrary, cross-species signaling hubs among the core net genes (those with high k) had an increased tendency to be implicated in both cancer and other diseases. This suggests that core site genes are implicated in a narrower set of diseases than are core net genes.

Whereas earlier work [18] has studied the conservation of phosphorylation sites across diverse species where the data have come from diverse experimental approaches, our work focused on querying human phosphorylation sites with phosphoproteomes from three model organisms generated on a similar experimental and computational platform. This resulted in a higher coverage of conserved phosphorylation events, as exemplified by the identification of substantially more positionally conserved phosphorylation sites identified between human and yeast than the previous study (51 versus 1) [18]. Our comparison is still relatively rough as we, for example, compare a full multicellular organism (worm), a single-cell organism (yeast), and various cell lines (fly and human). Therefore, we expect the numbers of conserved phosphorylation sites to increase as comparative phosphoproteomics develops in the future. Reports on the investigation of the sequence conservation of phosphorylation sites have reached conflicting conclusions: Gnad *et al.* [51] and Malik *et al.* [107] reported that experimentally validated phosphorylated residues were more conserved than other phosphorylatable residues; whereas Jimenez *et al.* [75] suggested the opposite. Furthermore, it has been suggested that sites identified with large-scale MS are less likely to be functionally important unless they display conservation at the sequence level [90, 51]. However, we argue that such strategies will filter away many biologically important phosphorylation sites that need not be positionally conserved. Our network-alignment approach enables studies of phosphorylation events that are not necessarily positionally conserved and underlines the importance of assessing phosphorylation conservation at both site and network level.

2.5 Conclusion

In summary, we have systematically investigated the conservation of phosphorylation sites in human regulatory networks by comparison to distantly related model organisms. We identified cross-species phosphorylation events that occur on proteins that have an increased tendency to be implicated in diseases caused by mutations. This result suggests that a similar approach could be taken to identify networks misregulated in cancer, diabetes, or mental illnesses. We note that multiple diseases seem to converge on the conserved regulatory network (core net). Therefore, we argue that it is important to consider conserved kinase-substrate relationships rather than just conservation of phosphoproteins when searching for disease-related genes. Furthermore, these results suggest that multiple diseases might be targeted using common therapeutic agents [41]. This idea is supported by a recent study in mice indicating that type 1 diabetes can be suppressed by imatinib [103], a small-molecule tyrosine kinase inhibitor developed as a cancer drug. Similar supportive evidence is emerging related to the role of the kinase AMPK (adenosine monophosphate-activated protein kinase) in cancer and diabetes. Therefore, we envisage human regulatory network analysis similar to those used here may be useful for identifying signaling networks for therapeutic intervention.[133].

Chapter 3

Evolutionary dynamics of phosphorylation sites with different functions and structural features

3.1 Introduction

Previously, I identified a set of human phosphorylation sites which corresponding positions on orthologous proteins in fly, worm and yeast are also phosphorylated. I then characterized the conservation of these phosphorylated residues in human across selected vertebrate species. In this chapter, I extend my analysis to encompass human phosphorylation sites for which I do not have data to validate that their corresponding positions (as determined by sequence alignment) on orthologous proteins are also phosphorylated. I reasoned that residue conservation analysis of large sets of phosphorylation sites, grouped based on their features, could provide further insight into the evolution of cellular regulation by protein phosphorylation. In addition, I seek to delineate some of the factors that could influence the evolvability of phosphorylated residues and confound the interpretation of functionally important phosphorylation sites by sequence conservation. In this chapter, I report analysis on the conservation of phosphorylated residues grouped according to sites' i) characterized functions, ii) prevalence (as gauged by detection frequency across multiple HTP studies), iii) stoichiometry, iv) occurrence in structurally disordered/ordered protein regions, v) occurrence in proteins of various abundance and vi) occurrence in proteins with different protein interaction propensity. The conservation of human phosphorylated residues across 19 vertebrate species with sequenced genomes are analyzed. To complement this analysis and to identify potential universal trends in the evolution of protein phosphorylation, I also assembled phosphorylation sites reported in *S. cerevisiae* (budding yeast) and characterized their residue conservation across 31 fungal species. Lastly, I analyzed the sequence conservation of published phosphorylated residues in *M. musculus* (mouse) proteins detected from 9 mouse tissues to investigate how their conservation could potentially differ from phosphorylated residues detected in free-living single cells of *S. cerevisiae* and human cell lines.

Each set of grouped phosphorylated residues can have a different composition of serine, threonine and tyrosine, distributed differently among structurally disordered/ordered regions, and occur in different proteins, all of which are likely under different evolutionary constraints unrelated to protein phosphorylation. Hence, comparing the absolute conservation rate across sets of phosphorylated residues can be misleading. For example, some sets of phosphorylated residues may appear to be more conserved than others because they occur largely in highly conserved proteins. It is therefore important to ”normalize” for these confounding factors to facilitate comparative analysis of different sets of phosphorylated residues. In this chapter, I devise a method that considers the above-mentioned confounding factors for comparing the conservation of phosphorylated residues across different sets.

3.2 Materials and Methods

3.2.1 Assembly of non-redundant human phosphorylaton data

Human phosphorylation sites annotated in PhosphoSitePlus and Phospho.ELM were downloaded from the two online databases in September 2009. As the two databases use protein sequences from different releases of the SwissProt database to track the positions of phosphorylation sites, all data were mapped to a reference human sequence set from the Ensembl online database (release 55, 2009). Phosphorylation sites were mapped to the longest translation of each human gene in the Ensembl database (release 55) whenever possible. This helped to resolve cases where identical sites had different positions due to revisions of the SwissProt sequence referenced and to remove redundant sites. The mapping between SwissProt primary accessions and its corresponding Ensembl human protein identifiers (release 55) was obtained from Ensembl through its BioMart interface. Finally, the positions of the phosphorylation sites on protein sequences from Ensembl were identified by exact string matching (using the peptide sequence spanning from -7 to +7 position of phosphorylated residue that was extracted from protein sequence in the Phospho.ELM or PhosphoSite database). This procedure resulted in 51,448 non-redundant human phosphorylation sites consisting of 32,014 phosphorylated serines (62.2%), 9,292 phosphorylated threonines (18.1%) and 10,142 phosphorylated tyrosines (19.7%).

3.2.2 Assembly of human phosphorylation sites with annotated functions

Human phosphorylation sites with characterized molecular or cellular functions are annotated in the PhosphoSitePlus database. A custom perl script was written to search for such information from web pages describing each human phosphorylation site on the PhosphoSitePlus website. The list of characterized functions annotated on 50 or more phosphorylation sites are listed in Table 3.1. Many phosphorylation sites are annotated with multiple function terms that may be related. To investigate the extent of this, we computed the Dice's coefficient (D) for the overlap of phosphorylation sites between function terms listed in Table 3.1.

The Dice's coefficient (D) is defined as

$$D = \frac{2\left|X \bigcap Y\right|}{|X| + |Y|}$$

where X and Y are the set of sites annotated with molecular/cellular function x and y respectively.

3.2.3 Assembly of human phosphorylation sites with stoichiometry information

Phosphorylation sites in HeLa S3 cells at different cell cycle stages were detected by Olsen *et al.* [130]. Information on these sites was retrieved from the online supplementary data of the publication. The phosphorylation sites were reported on protein sequences from the International Protein Index (IPI) database (release 3.37). To facilitate comparison

Table 3.1: List of function terms that have been each annotated on 50 or more human phosphorylation sites from the PhosphoSitePlus database. The number of genes encoding proteins with each set of phosphorylation sites are listed in bracket. Gene count is used as a phosphorylation site can be found on different spliced variants.

Function	No. of sites (genes)	Function	No. of sites (genes)
regulates molecular association	1260 (569)	cytoskeletal reorganization	206 (118)
altered intracellular location	617 (317)	protein stabilization	161 (83)
regulates transcription	575 (239)	regulates cell motility	160 (81)
activation	452 (229)	enzymatic inhibition	149 (89)
enzymatic activation	448 (208)	altered conformation	125 (79)
regulates cell growth	307 (162)	receptor internalization	92 (37)
phosphorylation	306 (160)	altered receptor desensitization	84 (35)
regulates cell cycle	295 (136)	regulates cell adhesion	80 (59)
inhibition	265 (143)	ubiquitination	70 (41)
regulates apoptosis	256 (129)	regulates cell differentiation	61 (34)
protein degradation	236 (122)		

with dataset assembled from PhosphoSitePlus and Phospho.ELM, the phosphorylation sites detected by Olsen *et al.* were mapped to human protein sequences from the Ensembl database (release 55) using exact string matching of phosphorylated peptides given in the supplementary data of the publication. The protein mapping between IPI primary accessions and its corresponding Ensembl human protein identifiers (release 55) was extracted from the IPI data file obtained at ftp://ftp.ebi.ac.uk/pub/databases/IPI/old/HUMAN/. A total of 22,311 out of 24,714 phosphorylation sites identified by Olsen *et al.* were mapped to the reference protein sequence set from Ensembl. Of the phosphorylation sites that are not mapped, many are because the Ensembl protein identifiers provided by IPI database are not present in release 55 of the Ensembl database. Some sites are also not mapped because sequence of the same protein varies across the two databases. Using a combination of quantitative mass spectrometry techniques, Olsen *et al.* were also able to quantify the average stoichiometry during mitosis for a subset of identified phosphorylated sites. Among the mapped phosphorylation sites, 4,324 have site stoichiometry information which was used for residue conservation analysis.

3.2.4 Preprocessing of *M. musculus* phosphorylaton sites

Published phosphorylation sites identified in nine mouse tissues using MS-based methods were obtained from the authors of the study [69]. The nine tissues are brain, brown fat, heart, liver, lung, kidney, pancreas, spleen and testis. The phosphorylation sites were reported on mouse protein sequences from the IPI database (version 3.6). In the IPI database, mapping of protein sequences to corresponding gene accession ID from Ensembl are provided for most proteins. For each phosphorylation site detected, I extracted subsequences containing the phosphorylated residue with 10 amino acids (less when the site is located near N- or C-terminal) flanking the phosphorylated residue at each side. This subsequence was then mapped to the longest protein translation (Ensembl release 55, 2009) of the Ensembl gene provided in the IPI database. This identifies the location of the same site on protein sequence from the Ensembl database. A total of 30,494 out of the 35,965 sites were mapped onto corresponding mouse protein sequences in Ensembl database (release 55, 2009)

3.2.5 Assembly of non-redundant *S. cerevisiae* phosphorylaton sites

Phosphorylation sites in *S. cerevisiae* identified in 11 recent high-throughput (HTP) phosphoproteomic studies were assembled for residue conservation analysis (Table 3.2). These phosphorylation sites were extracted from supplementary data hosted on the journal website of each publication. Phosphorylation sites annotated in the PhosphoGRID online database [169] were also retrieved from the database website in February 2010 to supplement the data assembled from HTP studies. As there might be variation in the protein sequences used in various HTP studies to map the positions of detected phosphorylation sites, all phosphorylation sites assembled were mapped to *S. cerevisiae* protein sequences housed in the Ensembl database (release 55). Phosphorylation sites were mapped to the sequences in the reference set with exact string matching using the peptides with localized phosphorylated residues provided by each HTP studies. Table 3.2 lists the number of phosphorylated serines, phosphorylated threonines and phosphorylated tyrosines mapped for PhosphoGRID and each HTP studies. In total, 21,355 phosphorylation sites consisting of 16,055 phosphorylated serines (75.2%), 4,545 phosphorylated threonines (21.3%) and 755 phosphorylated tyrosines (3.5%) were assembled.

3.2.6 Collection of protein interaction and abundance data of *S. cerevisiae* proteins

To understand how the protein interaction propensity of phosphorylated proteins influences the conservation of phosphorylated residues on them, the high quality pairwise

Table 3.2: List of phosphoproteomic studies reporting *S. cerevisiae* phosphorylation sites assembled. Phosphorylation sites annotated in PhosphoGRID online database, which are mostly from conventional directed biological studies, were used to supplement data assembled from high-throughput (HTP) studies. The number of phosphorylated residues identified by each studies are listed except for PhosphoGRID where the actual number of non-redundant sites used to supplement HTP data are listed.

Ref	Title	pS	pT	pY	Total
[62]	Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. (2009)	8090	1976	342	10408
[168]	Global analysis of the yeast osmotic stress response by quantitative proteomics. (2009)	4608	1334	164	6106
[180]	Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. (2009)	4833	963	208	6004
[2]	A multidimensional chromatography technology for in-depth phosphoproteome analysis. (2008)	4922	893	64	5879
[49]	High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. (2009)	3025	549	39	3613
[166]	Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases. (2007)	2100	429	0	2529
[96]	Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. (2007)	1218	243	14	1475
[11]	Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. (2009)	929	317	19	1265
[29]	Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. (2007)	659	119	1	779
[55]	Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. (2005)	557	101	4	662
[42]	Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. (2002)	221	50	3	274
[169]	PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast Saccharomyces cerevisiae. (2010)	2355	577	30	2962

interaction among *S. cerevisiae* proteins detected by Yu *et al.* [201] using yeast two-hybrid (Y2H) method were used. The dataset, referred to as CCSB-YI1 in the publication, consisting of 1809 pairwise protein-protein interactions among 1278 proteins was obtained at http://interactome.dfci.harvard.edu/S_cerevisiae/download/CCSB-Y2H.txt. To understand how protein abundance affects protein phosphorylation and the conservation of phosphorylated residues, abundance information derived by Ghaemmaghami *et al.* [47] on *S. cerevisiae* proteins at log-phase growth was downloaded from the online supplementary data of the publication. In the work, *S. cerevisiae* proteins are tagged with high-affinity epitope, expressed under the control of their natural promoters and their absolute abundances measured using quantitative western blot analyses.

3.2.7 Collection of protein sequences of selected metazoan and fungal species

Sequences of known and inferred proteins of 20 vertebrate species, including *Homo sapiens*, with 6X or more genome coverage were retrieved from the Ensembl online database (release 55) at their FTP website (http://jul2009.archive.ensembl.org/info/data/ftp/ index.html). These 20 metazoan species are *Homo sapiens* (Human), *Pan troglodytes* (Chimpanzee), *Pongo pygmaeus* (Orangutan), *Cavia porcellus* (Guinea Pig), *Rattus norvegicus* (Rat), *Mus musculus* (Mouse), *Monodelphis domestica* (Opossum), *Canis familiaris* (Dog), *Bos taurus* (Cow), *Equus caballus* (Horse), *Ornithorhynchus anatinus* (Platypus), *Gallus gallus* (Chicken), *Taeniopygia guttata* (Zebra Finch), *Anolis carolinensis* (Anole Lizard), *Xenopus tropicalis* (Frog), *Oryzias latipes* (Medaka), *Gasterosteus aculeatus* (Stickleback), *Tetraodon nigroviridis* (Tetraodon), *Takifugu rubripes* (Fugu) and *Danio rerio* (Zebrafish).

Sequences of known and inferred proteins in 32 fungal species, including *S. cerevisiae*, were obtained from various online sources. Table 3.3 lists the 32 fungal species and the source of their proteome sequences.

Table 3.3: List of fungal species used to analyze the conservation of phosphorated residues in *S. cerevisiae*. The source of proteome sequences of each species is also listed.

Species	Database	Species	Database
Ashbya gossypii	EMBL-EBI	*Kluyveromyces waltii*	YGOB
Aspergillus nidulans	BROAD	*Lodderomyces elongisporus*	BROAD
Aspergillus niger	BROAD	*Saccharomyces bayanus*	SGD
Candida albicans	BROAD	*Saccharomyces castelli*	SGD
Candida glabrata	Gnolevures	*Saccharomyces cerevisiae*	SGD
Candida guilliermondii	BROAD	*Saccharomyces kluyveri*	Gnolevures
Candida lusitaniae	BROAD	*Saccharomyces kudriavzevii*	SGD
Candida parapsilosis	BROAD	*Saccharomyces mikatae*	SGD
Candida tropicalis	BROAD	*Saccharomyces paradoxus*	SGD
Coccidioides immitis	BROAD	*Schizosaccharomyces japonicus*	BROAD
Coccidioides posadasii	BROAD	*Schizosaccharomyces octosporus*	BROAD
Debaryomyces hansenii	Gnolevures	*Schizosaccharomyces pombe*	BROAD
Fusarium graminearum	BROAD	*Uncinocarpus reesii*	BROAD
Kluyveromyces lactis	Gnolevures	*Verticillium dahliae*	BROAD
Kluyveromyces polysporus	YGOB	*Yarrowia lipolytica*	Gnolevures
Kluyveromyces thermotolerans	Gnolevures	*Zygosaccharomyces rouxii*	Gnolevures

3.2.8 Identification of orthologous sequences

The InParanoid algorithm [142] (version 2.0) was used to infer orthologous sequences of human and mouse proteins across 19 vertebrate species with sequenced genomes using the BLOSUM80 scoring matrix with default parameters. Similarly, the InParanoid algorithm was also used to infer protein sequences in 31 fungal species orthologous to *Saccharomyces cerevisiae* proteins but the BLOSUM62 scoring matrix was used instead with default parameters. BLAST software (version 2.2.12) [3] needed by Inparanoid was obtained from NCBI's FTP website. In all cases, only the longest translation of each known/inferred genes were input into InParanoid for ortholog prediction. BLOSUM80 and BLOSUM62 were used as recommended in InParanoid for vertebrate species and eukaryotic species respectively.

3.2.9 Prediction of intrinsic disordered regions in proteins

The DISOPRED2 predictor [http://bioinf.cs.ucl.ac.uk/disopred/] [193] was used to identify structurally (intrinsic) disordered and ordered regions in known phosphorylated human and *S. cerevisiae* proteins. The non-redundant (NR) protein sequence database required for the predictor was obtained from the National Center for Biotechnology Information in November 2007. The NR database was filtered for transmembrane protein regions with the *pfilt* program provided with DISOPRED2. Subsequently, the output was extracted with custom Perl scripts.

3.2.10 Computing the normalized relative divergence rate of phosphorylated residues

Multiple sequence alignments were carried out to determine the conservation of serine, threonine, tyrosine and their phosphorylated subsets. The sequence of each phosphorylated protein in human, mouse or *S. cerevisiae* was grouped with their orthologous protein sequences in other species (vertebrate species for human proteins and fungal species for *S. cerevisiae* proteins). The MAFFT (v6.240, E-INS-i option with default parameters) [81] multiple sequence alignment algorithm was used to align each set of orthologous protein sequences.

Conservation of serine, threonine, tyrosine and their phosphorylated subsets on human, mouse or *S. cerevisiae* proteins across other species were then computed from the resulting sequence alignments using custom Perl scripts. To avoid possible accelerated sequence divergence arising from gene duplication which could confound conservation analysis, only protein sequences with one-to-one orthology relationship to phosphorylated proteins, as inferred by the InParanoid algorithm, were included for sequence alignment. Between a phosphorylated protein and each orthologous protein, residues on phosphorylated protein were omitted from the conservation analysis when less than three of its ten adjacent flanking residues are aligned to identical amino acids on the other protein sequence. When computing the conservation of residues across orthologous proteins, two scenarios are considered in this work. One scenario considers serine and threonine equivalent in terms of phosphorylation propensity and effects. Hence, observed substitution of serine by threonine or vice versa is considered conserved. The other scenario considers that serine, threonine and tyrosine as different from each other such that one cannot be substituted by the other two phosphorylatable residues in terms of phosphorylation propensity and effect. Different scenarios are considered for each set of phosphorylated residues.

A measure termed relative divergence rate (RD) is defined in this work to quantify the relative conservation (or evolvability) of phosphorylated residues compared to background phosphorylatable residues. A RD of 0.5, for example, tells us that for every 100 background phosphorylatable residues that have changed, 50 phosphorylated residues would also have changed, while a RD greater of 2 tells us that for every 100 background phosphorylatable residues that have changed, 200 phosphorylated residues would also have changed. Hence, a RD above one indicates phosphorylated residues are changing faster than background phosphorylatable residues while a RD below one means phosphorylated residues are more conserved than background phosphorylatable residues. The RD for a set of residues phosphorylated in species x across species y can be computed as:

$$RD = \frac{D_{phos}}{D_{background}}$$

where D_{phos} and $D_{background}$ are the portion of phosphorylated residues and background phosphorylatable residues in species x not conserved in species y. Note that residues on phosphorylated proteins were omitted from computation when less than 3 of its 10 adjacent flanking residues are aligned to identical amino acids on the other protein sequence. Instead of computing $D_{background}$ from all phosphorylatable residues,

$D_{background}$ is computed from randomly selected phosphorylatable residues in phosphorylated proteins, but maintaining the same number of serines, threonines and tyrosines, and the same distribution across structurally disordered/ordered protein regions as was observed for the phosphorylated residues in each phosphorylated protein. $D_{background}$ is averaged over 1,000 random selection trials. The RD computed using this random sampling approach is termed *normalized RD* in this work.

While the normalized RD of phosphorylated residues can be computed for each pair of species, the D_{phos} and $D_{background}$ obtained for each pair of species were used to fit a linear model of $Y = BX$ where Y and X is the divergence rate of phosphorylation residues (D_{phos}) and background phosphorylatable residues ($D_{background}$), and B is the coefficient or the gradient of the best fit line. We used B as our estimate for the most prevalent normalized RD across the various species analyzed. The linear model of $Y = BX$ is fitted using the *glm* function under the Generalized Linear Model package in the R statistical tool.

To find out to what extent the normalized RD, derived using the linear model, can vary due to the random selection of background phosphorylatable residues, I computed the normalized RD for the entire set of functionally annotated human phosphorylation site but repeated the procedure 20 times. The observed standard deviation is 0.00018, indicating that our computed measures are stable. There are 2911 human phosphorylation sites with annotated functions. I repeated this analysis on smaller datasets of 236 and 61 human phosphorylation sites implicated in ”protein degradation” and ”regulates cell differentiation” respectively. The observed standard deviation for the two data sets are 0.00037 and 0.00177 respectively. In this work, I restricted the minimum number of sites for any dataset to 50. Hence, the normalized RD values computed in this work are robust to randomization to at least 2 decimal places.

3.3 Results

3.3.1 Human phosphorylated residues with characterized functions are well-conserved

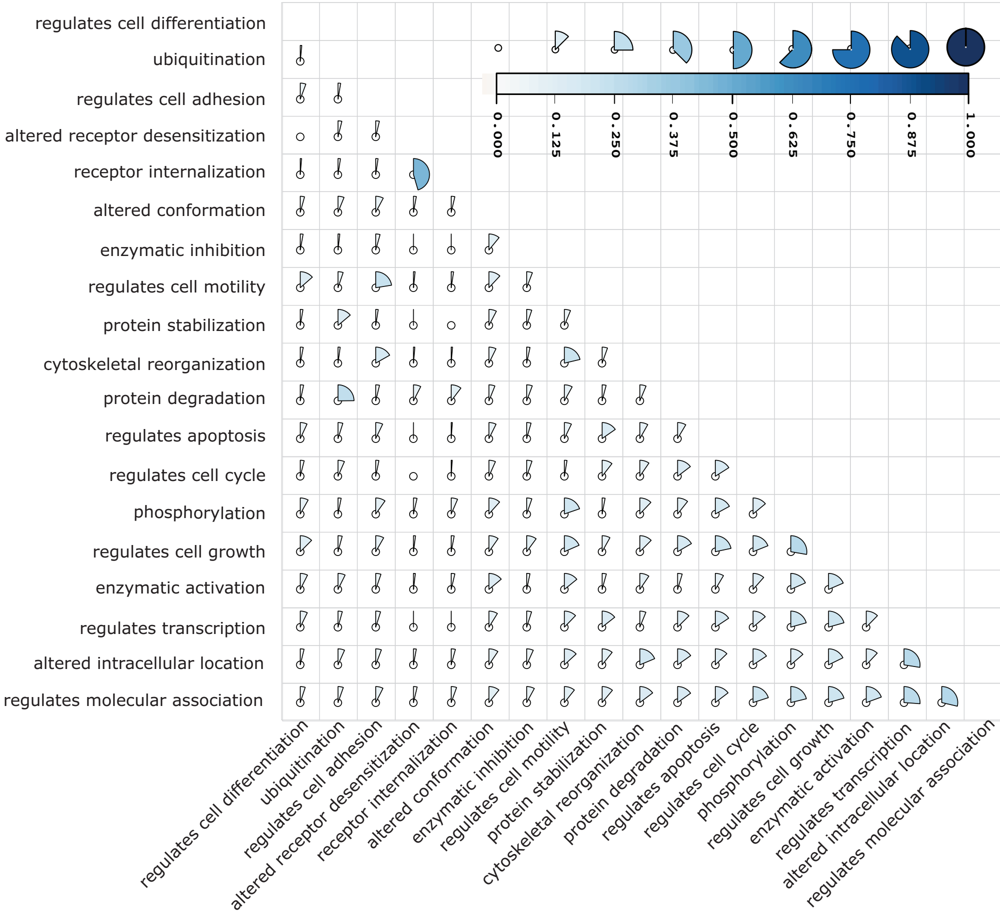

Figure 3.1: Extent of site overlap between different functions quantified using Dice's coefficient. The functions are ordered according to their number of sites with *regulate cell differentiation* and *regulate molecular association* having the least and most sites respectively.

Sequence conservation analysis of phosphorylated residues could be used to identify physiologically important phosphorylation events. This is based on the premise that if residues which phosphorylation arises early in evolution that is beneficial for survival (or propagation), their removal by mutation is likely not favored during evolution. Hence, these residues are more likely to be conserved in other species. Here, I first carried out conservation analysis on the set of human phosphorylated residues for which functional effects of phosphorylation have been experimentally characterized. This is to quantify the extent to which functional roles of phosphorylation affect the conservation of phosphorylated residues in other species. Information on human phosphorylation sites with characterized functions as annotated in PhosphoSite was retrieved from the online database for this purpose.

A total of 2,910 non-redundant human phosphorylation sites of known function were obtained and 21 function terms were found annotated for 50 or more phosphorylation sites each (see Table 3.1). To determine which molecular or cellular effect of protein phosphorylation are possibly constrained to the similar protein positions across orthologous sequences, I analyzed subsets of these phosphorylation sites organized by functions listed in Table 3.1 except for terms ”activation” and ”inhibition” which I personally, as a biologist, deemed too vague about the functions of the phosphorylation sites for meaningful interpretation. Many phosphorylation sites are annotated under multiple function terms possibly due to semantic overlap of the terms, and because phosphorylation sites can have multiple functional effects. The extent of site overlap between pairs of function terms listed in Table 3.1 is quantified using Dice's coefficient and visualized in Figure 3.1. Average Dice's coefficient of site overlap between all pairs of functions listed in Table 3.1 is 0.091 ($std.\ dev. = 0.071$). For pairs of function terms with high site overlap (arbitrarily defined as those with Dice's coefficient of 0.25 or more), many are not unexpected based on our present understanding on the function of protein phosphorylation. For example, kinases are known to “regulate transcription” by “alter[ing] intracellular location” of transcription factor through phosphorylation, and phosphorylation is known to promote “ubiquitination” that tag proteins for “protein degradation” that can explain the high site overlap observed between these two pairs of functions.

In this work, I define a measure termed relative divergence rate (RD) to quantify the conservation of phosphorylated residues compared to background phosphorylatable residues. Briefly, the RD for a set of residues phosphorylated in species x across species y can be computed as $RD = \frac{D_{phos}}{D_{nonphos}}$ where D_{phos} and $D_{nonphos}$ are the portions of phosphorylated and other phosphorylatable residues in species x respectively that are not conserved in species y. In this text, the D of a set of residues is referred to as its divergence rate. Hence, RD is the ratio between divergence rate of phosphorylated residues and background phosphorylatable residues between 2 species. A RD above one indicates that phosphorylated residues are changing faster than background phosphorylatable residues while a RD below one means phosphorylated residues are more conserved than background phosphorylatable residues.

The conservation of human phosphorylated residues across 19 vertebrate species with sequenced genomes are analyzed. However, instead of computing the normalized RD of residues phosphorylated in human for each of the 19 species of which some may vary substantially, a linear regression approach is used to estimate the normalized RD most prevalent (or "average" for intuitive understanding) among the 19 species (see section 3.2.10 for more details). This linear regression approach is more robust to outliers than taking the average of normalized RD across the 19 species.

For this analysis, I considered the scenario that serine and threonine are equivalent in terms of phosphorylation potential and effect. Hence, a serine substituted by threonine and vice versa at the same position on an orthologous protein is considered conserved. The computed normalized RD for all phosphorylated residues with characterized function is 0.56 (see top left-hand plot in Figure 3.2), indicating that residues in which phosphorylation have functional effects are generally diverging at half the rate of background phosphorylatable residues. This indicates that beneficial functional consequences of phosphorylation promote the conservation of phosphorylated residues.

All analyzed subsets of phosphorylated residues under different function have normalized RD of less than one, indicating that all are more conserved than background

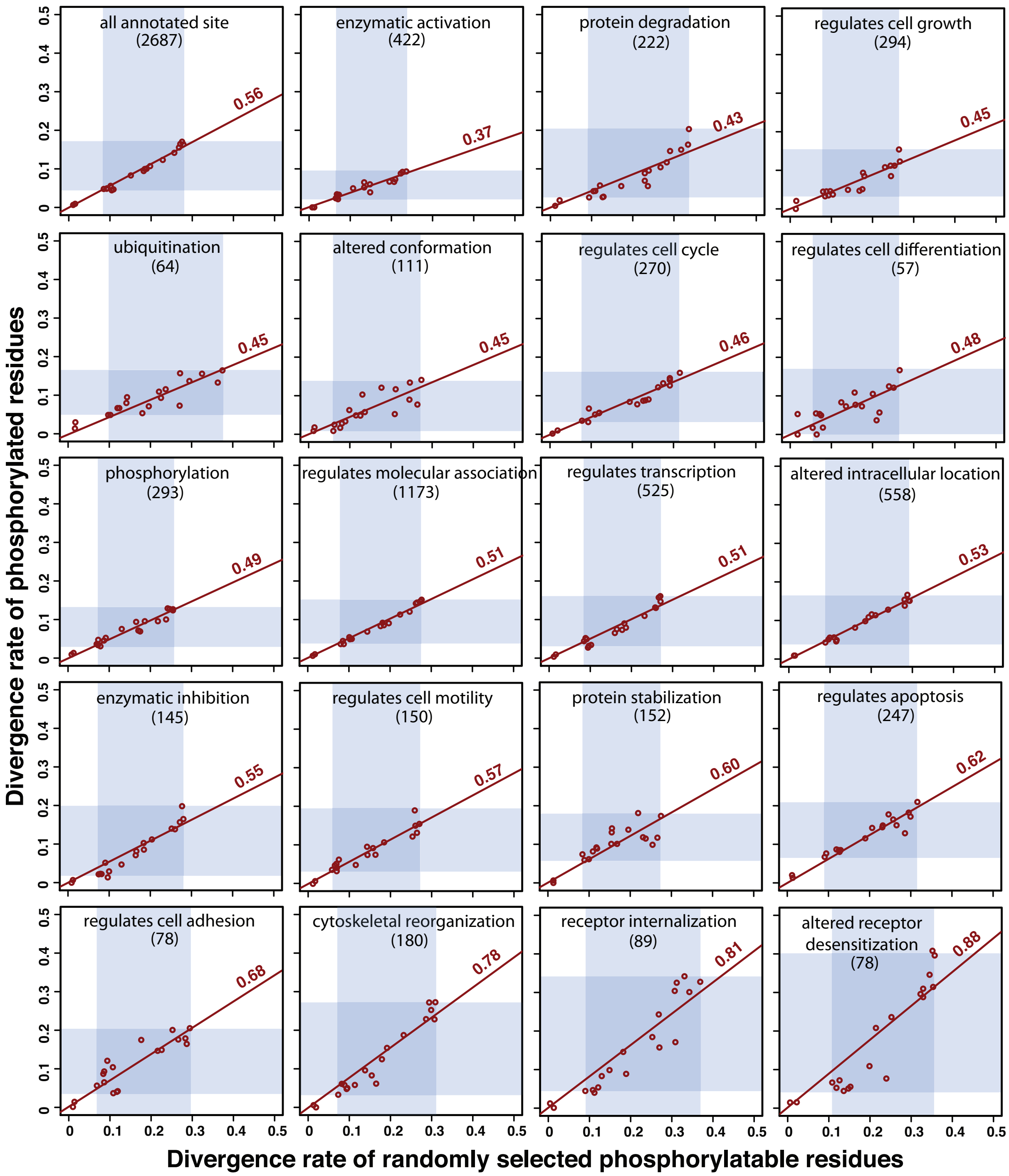

Figure 3.2: Normalized relative divergence rate (*RD*) of human phosphorylated residues of different characterized function across 19 vertebrate species. The divergence rate of phosphorylated residues and background phosphorylatable residues in human across each of the 19 vertebrate species are computed and plotted for each function. The divergence rate is the portion of residues in human not conserved in each species as determined from multiple sequence alignment of sets of orthologous protein sequences. A best fit line was determined for each plot using linear regression. The most prevalent normalized *RD* across the 19 species is the coefficient shown on the line. Excluding the two phylogenetically nearest species, the blue shadings highlight the maximum and minimum divergence rate at each axis. As the number of residues aligned differs among species due to variations in orthologs inferred and sequence alignment quality, the number of phosphorylated residues aligned to chimpanzee's protein sequences is used to indicate the size of each data subset. This number is denoted at the bottom right corner of each plot and color-coded accordingly.

phosphorylatable residues with median *RD* of 0.51 (Figure 3.2) which can be interpreted that if 100 background phosphorylatable residues are found mutated, only 51 phosphorylated residues will be mutated. Among the different functions, phosphorylated residues that are implicated in enzymatic activation are most conserved with a normalized *RD* of 0.37. These phosphorylated residues are found on proteins encoded by a total of 208 human genes, of which 125 encode protein kinases. As many protein kinases are known to play important roles in regulating diverse cellular activities, it is not unexpected that phosphorylated residues on them that are implicated in their activation are strongly conserved across other species. In addition, as discussed in chapter 2, many activation sites on protein kinases need to be positionally conserved for allosteric regulation of protein kinases. However, it is possible that the function of some of these phosphorylation sites are inferred based on homology or sequence conservation, and thus contribute to higher sequence conservation observed. Among the least conserved subsets of phosphorylated residues, but nevertheless still more conserved than background phosphorylatable residues, are those implicated in "altered receptor desensitization", "receptor internalization" and "cytoskeletal reorganization" with normalized *RD* of 0.88, 0.81 and 0.78 respectively. The rest of the functions have sites with normalized *RD* between 0.43 to 0.68. Closer scrutiny confirmed that many sites under "altered receptor desensitization" and "receptor internalization" are observed on receptor proteins. A number of non-exclusive reasons can explain the general weaker conservation of phosphorylated residues implicated in these functions. It is possible that phosphorylation regulation of these functions are less positionally constrained to similar positions across orthologous proteins or more phosphorylated residues implicated in these functions have appeared uniquely along the human lineage.

3.3.2 Frequently detected human phosphorylated residues are more conserved

Here, I assessed the conservation of phosphorylated residues detected by high throughput (HTP) experimental studies using determined normalized *RD* of functionally characterized phosphorylated residues as a base for comparison. I also assessed whether phosphorylated residues detected more frequently in HTP studies are more conserved. This is based on the reasoning that some phosphorylated residues detected by multiple studies are less likely to be false positive sites or spurious phosphorylation events. I only considered data from HTP studies to minimize residues selected by experimentalists for phosphorylation characterization due to their strong conservation. I first considered the scenario that serine and threonine are equivalent in terms of phosphorylation potential and effect, and assessed conservation of phosphorylated serines and threonines collectively as a group. Phosphorylated residues are progressively filtered by the minimum number of HTP studies that reported them. I considered a publication to be a HTP study if the publication reported more than 500 phosphorylation sites in our assembled list of phosphorylation sites. Similar to previous analyses, we computed the divergence rate of randomly selected phosphorylatable residues maintaining the same number of serine and threonine, and the same distribution in structurally disordered/ordered protein regions as the phosphorylation sites observed for each phosphorylated protein.

I found that phosphorylated serines and threonines detected by HTP studies have a normalized *RD* of 0.85 (see Figure 3.3) indicating that, overall, phosphorylated serines and threonines are diverging slower than background serine and threonine but are not as conserved as functionally characterized phosphorylation sites. The normalized *RD* progressively dropped to 0.72, 0.64 and 0.58 for phosphorylated residues reported in at least two, at least three and at least four HTP studies respectively. A noteworthy observation is phosphorylated residues detected in at least four HTP studies have similar normalized

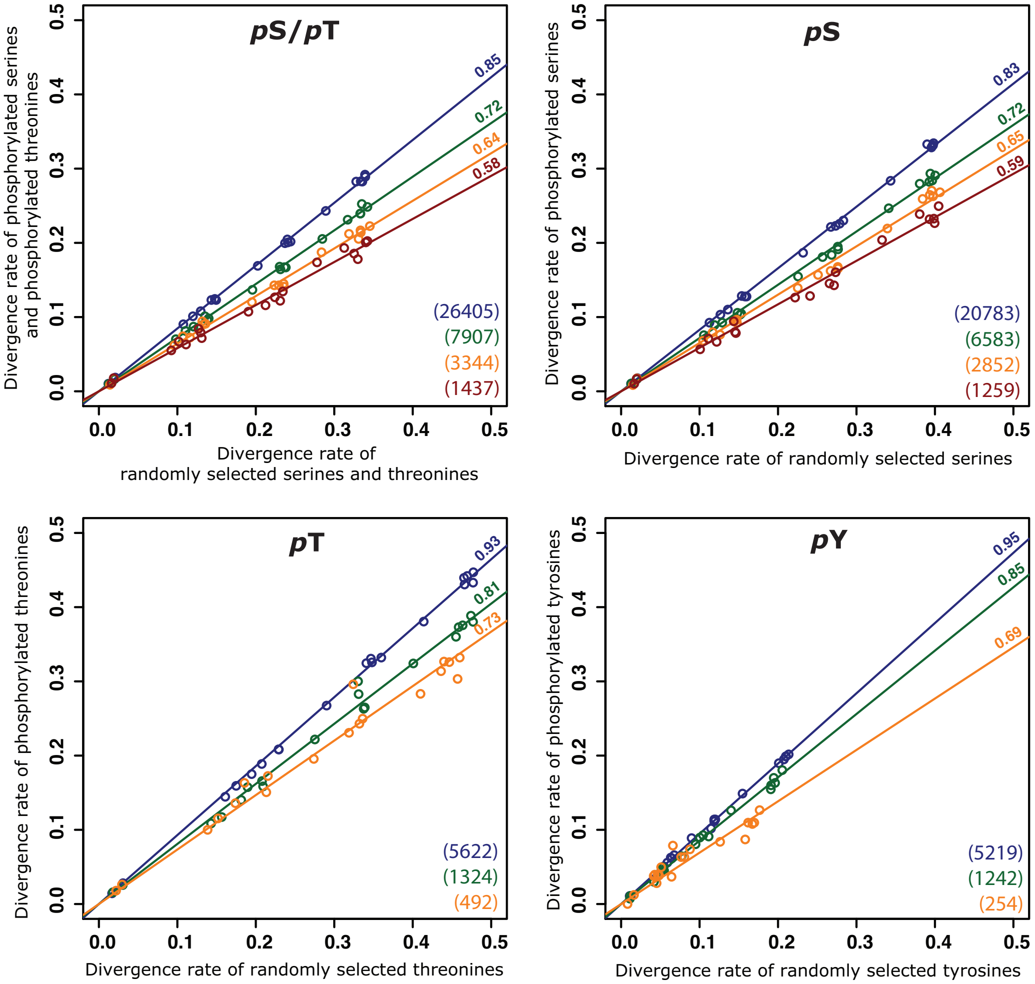

Figure 3.3: Conservation of human phosphorylated residues filtered by their detection frequency. Phosphorylated residues are progressively filtered according to the number of HTP studies reporting them. Blue, green, orange and red circles are for phosphorylated residues detected in ≥1, ≥2, ≥3, ≥4 HTP studies respectively. The best fit regression line and its coefficient is shown for each subset of phosphorylated residues. The coefficient is the normalized *RD* across computed for across the 19 vertebrate species. For the top left-hand plot, serine-threonine substitution are considered conserved and phosphorylated serines and phosphorylated threonines are collectively analyzed. The remaining three plots consider each amino acid individually. Filtered data sets that with less than 50 phosphorylated residues are omitted from analysis. As the number of residues aligned differs among species due to variations in orthologs inferred and sequence alignment quality, the number of phosphorylated residues aligned to chimpanzee's protein sequences is used to indicate the size of each data subset. This number is denoted at the bottom right corner of each plot and color-coded accordingly.

RD with functionally characterized phosphorylation sites. I repeated the above analysis considering the scenario that serine, threonine and tyrosine cannot substitute for each other in term of phosphorylation potential and functional effects. This complementary analysis also revealed that more frequently detected phosphorylated residues are more conserved (Figure 3.3).

There are a number of non-exclusive explanations for this observation. It is possible that less frequently detected phosphorylation sites are more likely to result from spurious phosphorylation events that do not have important functional consequences. A related explanation is that less frequently detected phosphorylation sites are more likely to be false positives identified by individual labs using different spectra analysis algorithms while real phosphorylation sites are identified more consistently. Another plausible explanation is that more frequently detected phosphorylation sites are effecting stronger or more persistent functional consequences that are favored across species.

3.3.3 Higher stoichiometry sites in human observed during mitosis are not more conserved

A recent HTP study by Olsen *et al.* [130] on HeLa S3 cells using quantitative mass spectrometry techniques managed to obtain site stoichiometry information for about 4,500 phosphorylation sites on human proteins [130]. Using this set of phosphorylation sites, I investigated whether phosphorylated residues of sites with observed higher stoichiometry are more conserved than those of lower stoichiometry. I considered serine-threonine substitution as conserved for this analysis. The computed normalized *RD* of phosphorylated residues with >20%, >40% and >60% site stoichiometry are 0.97, 0.98 and 0.93 respectively (Figure 3.4). Hence, this set of phosphorylated residues with site stoichiometry information are only slightly more conserved than background phosphorylatable residues and, unexpectedly, phosphorylated residues of higher site stoichiometry are only marginally more conserved than phosphorylated residues of lower site stoichiometry. To

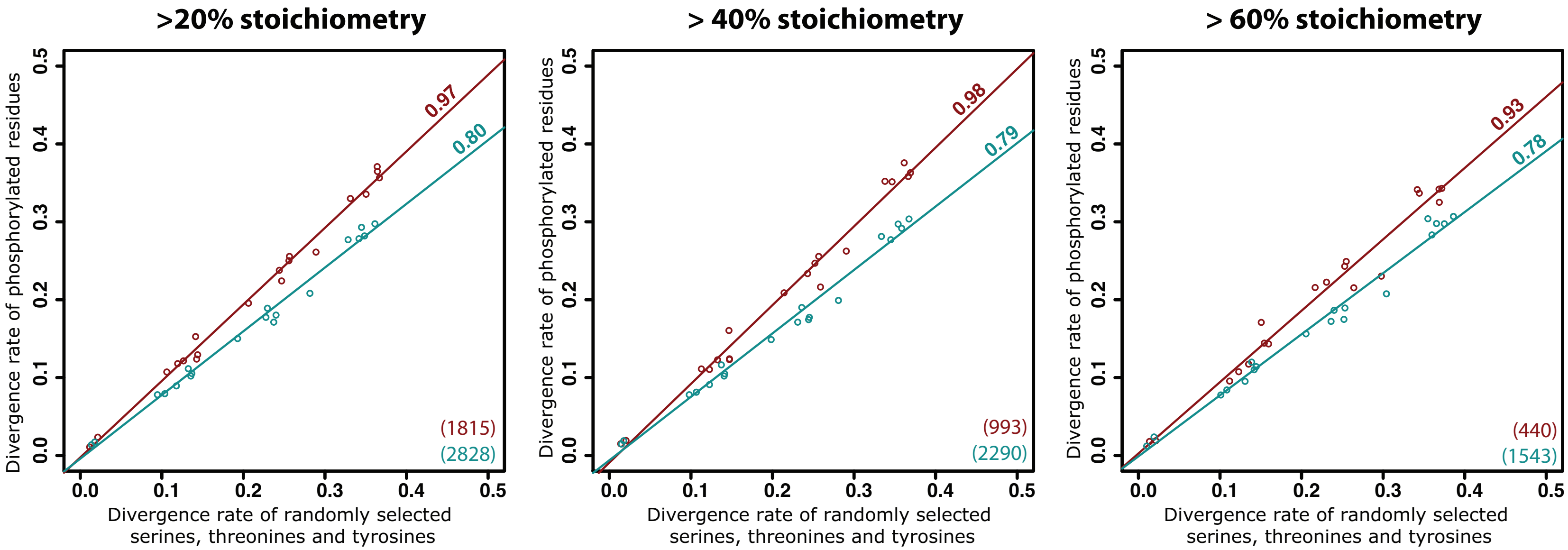

Figure 3.4: Conservation of human phosphorylated residues filtered by site stoichiometry. Red circles represent phosphorylated residues with site stoichiometry information obtained by Olsen *et al.* [130]. Cyan circles represent phosphorylated residues detected by ≥2 HTP studies (excluding Olsen *et al.*) on the same set of phosphorylated proteins at each site stoichiometry threshold. A best fit line was determined for each set of phosphorylated residues using linear regression and the most prevalent normalized *RD* across the 19 species is the coefficient shown on the line. As the number of residues aligned differs among species due to variations in orthologs inferred and sequence alignment quality, the number of phosphorylated residues aligned to chimpanzee's protein sequences is used to indicate the size of each data set. This number is denoted at the bottom right corner of each plot and color-coded accordingly.

investigate whether phosphorylated residues on the phosphorylated proteins are generally weakly conserved regardless of site stoichiometry, I analyzed the conservation of phosphorylated residues on the same set of phosphoproteins at each site stoichiometry threshold that are detected in 2 or more HTP studies (excluding Olsen *et al.* [130]). I found that the sets of phosphorylated residues detected by multiple HTP studies are more conserved even when compared to phosphorylated residues with >60% site stoichiometry.

It is possible that the detection method adopted in the HTP study had identified many false phosphorylation sites leading to the weak conservation observed. About 24,000 phosphorylation sites were actually detected in the HTP study but site stoichiometry information was only obtainable for a subset of them. To investigate whether the phosphorylation sites detected in the study are in general not well-conserved regardless of site stoichiometry, I analyzed the residue conservation of all the phosphorylation sites detected in the study including with site stoichiometry information. I found that they have a normalized RD of 0.86 which is very similar to the combined set of phosphorylated residues detected in other HTP studies (≈0.85). This result suggests that the portion of false positive and spurious phosphorylation sites detected in the study is unlikely to be higher than other studies. Hence, it seems that high stoichiometry phosphorylation sites are not necessarily more conserved or the experimental methods used for quantifying site stoichiometry may not be robust.

3.3.4 Phosphorylated residues are relatively more conserved in structurally disordered than in structurally ordered protein regions

Here, I analyzed the conservation of phosphorylated residues within structurally disordered (*a.k.a* intrinsic disordered) regions and structurally ordered regions of proteins as predicted by the DISOPRED algorithm [193]. I first performed the analysis on the set of

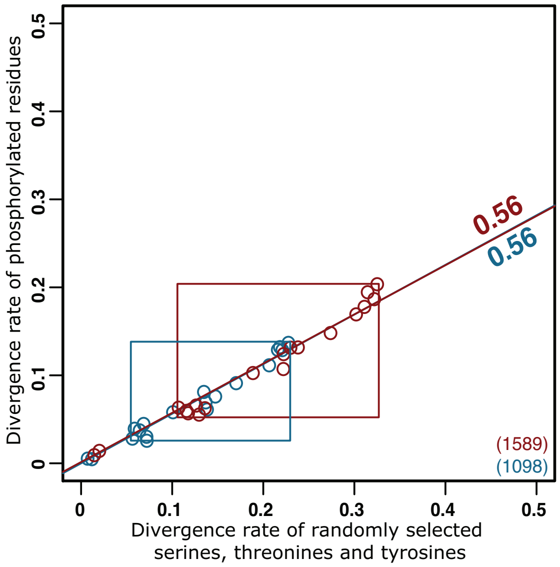

Figure 3.5: Conservation of human phosphorylated residues with characterized functions in structurally disordered and ordered protein regions. The conservation of phosphorylated residues of sites with experimentally characterized functions are analyzed according the protein regions they are embedded in. Red circles and cyan circles represent phosphorylated residues occurring in inferred structurally disordered and ordered protein regions, respectively. A best fit line was determined for each set of phosphorylated residues using linear regression and the most prevalent normalized RD across the 19 species is the coefficient shown on each line. Excluding the two most phylogenetically closest species, the red and cyan rectangles respectively bound the maximum and minimum divergence rate at each axis for the two sets of phosphorylated residues. As the number of residues aligned differs among species due to variations in orthologs inferred and sequence alignment quality, the number of phosphorylated residues aligned to chimpanzee's protein sequences is used to indicate the size of each data set. This number is denoted at the bottom right corner of the plot.

phosphorylation sites with experimentally characterized functions that were previously analyzed in Section 3.3.1 and considered serine-threonine substitution to be conserved. Overall, phosphorylated residues of functionally characterized sites in structurally disordered protein regions seem less conserved than that of functionally characterized sites in structurally ordered protein regions (judging by the distribution of red and cyan circles on the x-axis in Figure 3.6). However, both sets of phosphorylated residues have normalized RD of 0.56 suggesting that important functional consequences of phosphorylation impose a similar degree of evolutionary constraint on phosphorylated residues across the two protein regions (see Figure 3.5). This is somewhat not unexpected as the functional

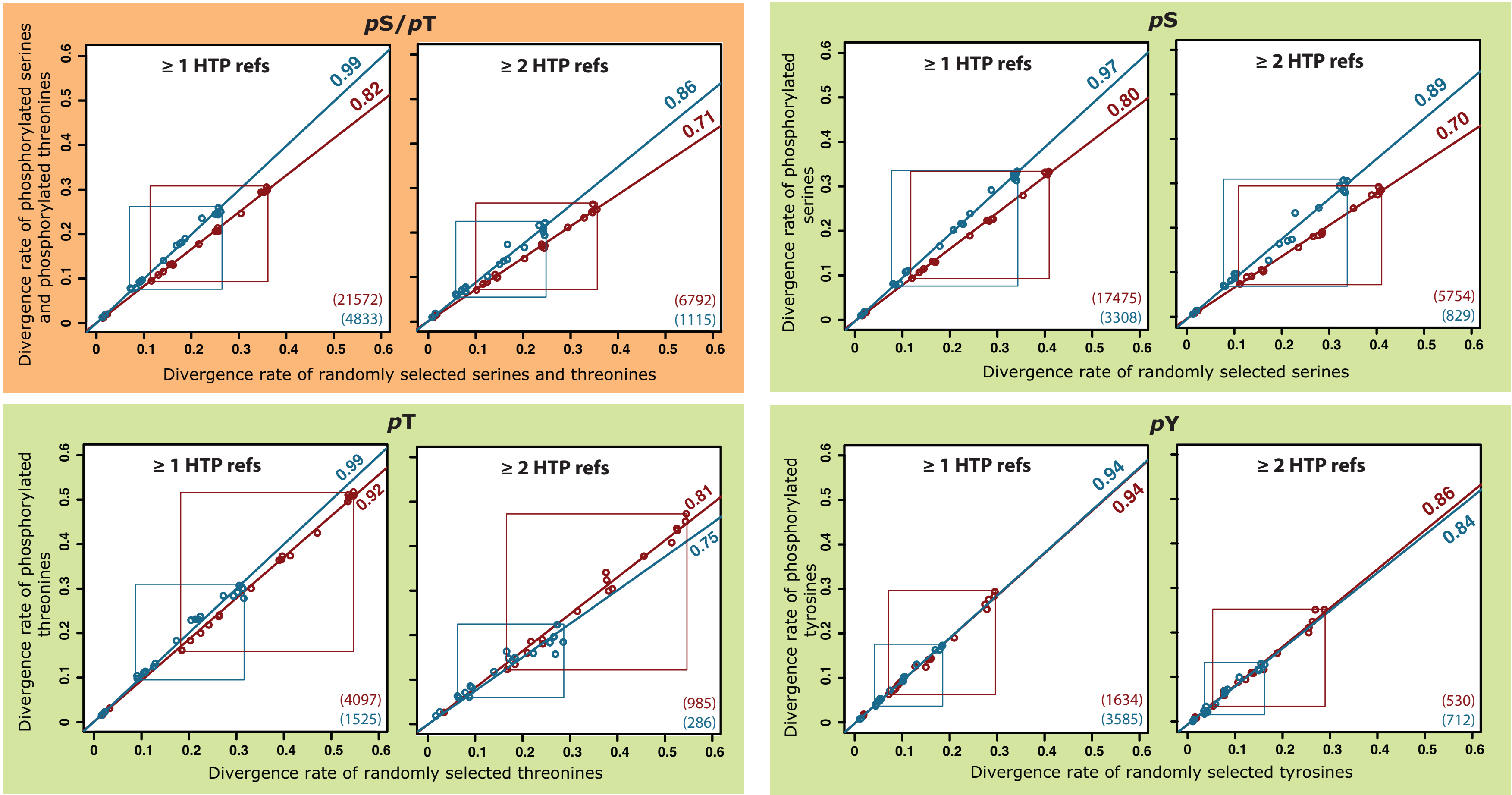

Figure 3.6: Conservation of human phosphorylated residues in structurally disordered and ordered protein regions. Plots in the orange panel consider serine-threonine substitution as conserved, and analyzed phosphorylated serines and phosphorylated threonines collectively as a group. Plots in the green panels consider the conservation of phosphorylated serines, phosphorylated threonines and phosphorylated tyrosines individually and that each cannot be substituted by the other phosphorylatable residues in terms of phosphorylation propensity and effect. Red circles and cyan circles represent phosphorylated residues in structurally disordered and structurally ordered protein regions, respectively. The normalized *RD* most prevalent across the 19 species as inferred using linear regression is the coefficient shown on each line. Excluding the two most phylogenetically closest species, the red and cyan rectangles respectively bound the maximum and minimum divergence rate at each axis for the two sets of phosphorylated residues. As the number of residues aligned differs among species due to variations in orthologs inferred and sequence alignment quality, the number of phosphorylated residues aligned to chimpanzee's protein sequences is used to indicate the size of each data set. This number is denoted at the bottom right corner of each plot and color-coded accordingly.

constraint imposed on these residues, resulting from their phosphorylation, should be very similar overall regardless of the protein regions they are embedded in.

Next, I examined the conservation of phosphorylated residues across intrinsic disordered protein regions and structured protein regions using all human phosphorylation sites assembled from the PhosphoSitePlus and Phospho.ELM databases,. I first considered the scenario where serine and threonine is equivalent in term of phosphorylation potential and effect (yellow panel in Figure 3.6), and analyzed conserved phosphorylated serines and phosphorylated threonines collectively as a group. I also filtered the phosphorylated residues by the number of HTP studies that reported them. I observed that phosphorylated serines and threonines in structured regions are overall not more conserved than other serines and threonines from the same type of protein regions (normalized $RD = 0.99$). However, the subset of phosphorylated residues detected in 2 or more HTP studies are more conserved with a normalized RD of 0.86. In comparison, phosphorylated serines and threonines in structurally disordered regions are well-conserved with normalized RD of 0.82 and 0.71 for the entire set and the subset detected in 2 or more HTP studies. This analysis was repeated for phosphorylated serine, phosphorylated threonine and phosphorylated tyrosine individually, and a similar phenomenon is observed (green panels in Figure 3.6).

Overall, the result suggests that more spurious phosphorylation events were occurring in structurally ordered protein regions. One possible explanation is that serines and threonines in structurally ordered regions are typically less accessible (located within the protein core, for example) to protein kinases but are increasingly phosphorylated by protein kinases when they are denatured or broken down into peptides under artificial conditions prevalent in HTP studies. For example, denaturing reagents like urea and sodium dodecylsulfate (SDS) are commonly used in protein assays. Another reagent that can possibly increase phosphorylation on residues that are typically surface inaccessible is dithiothreitol (DTT). This chemical is frequently used in lysis buffers and restriction enzyme buffers to preserve the activities of enzymatic proteins. It is also used in MS-based proteomic analysis to improve protein fragmentation by reducing disulfide bonds in proteins. However, breakage of disulfide bonds can disrupt the tertiary structure of a protein and expose typically surface inaccessible residues to protein kinases unnaturally. Phosphorylated tyrosines, on the other hand, are conserved consistently across the two protein regions with normalized RD of ≈ 0.86. Interestingly, I did not observe stronger conservation of more frequently detected phosphorylated tyrosines, for unknown reasons.

3.3.5 *S. cerevisiae* phosphorylation residues are only well-conserved in phylogenetically close species

In the previous section, I observed that more frequently detected phosphorylation sites on human proteins are generally more conserved. To verify whether this is a possible universal trend, I assembled phosphorylation sites on *S. cerevisiae* proteins reported by HTP studies. Again, we consider studies reporting more than 500 sites as HTP. For *S. cerevisiae*, phosphorylated tyrosines are excluded from analysis because the number of phosphorylated tyrosines is small. Based on this criteria, a total 20,172 phosphorylated serines and threonines are identified in HTP studies. The conservation of these sites across 31 fungal species (3 *Taphrinomycotina*, 7 *Pezizomycotina* and 21 *Saccharomycotina* species) were analyzed. I progressively filtered phosphorylation sites based on their detection frequency. Based on 31 species, phosphorylated residues detected on *S. cerevisiae* proteins are not conserved with normalized RD of 0.99, meaning phosphorylated residues are lost at the same pace as background serines and threonines (Figure 3.7, top left plot). Moreover, more frequently detected phosphorylation sites are only marginally more conserved. For example, phosphorylation sites detected in 4 or more HTP studies only have a normalized RD of 0.95 (Figure 3.7, top left plot). Closer visual inspection revealed that phosphorylation residues are conserved in phylogenetically close species but not in distantly related species. To validate this observation, I analyzed

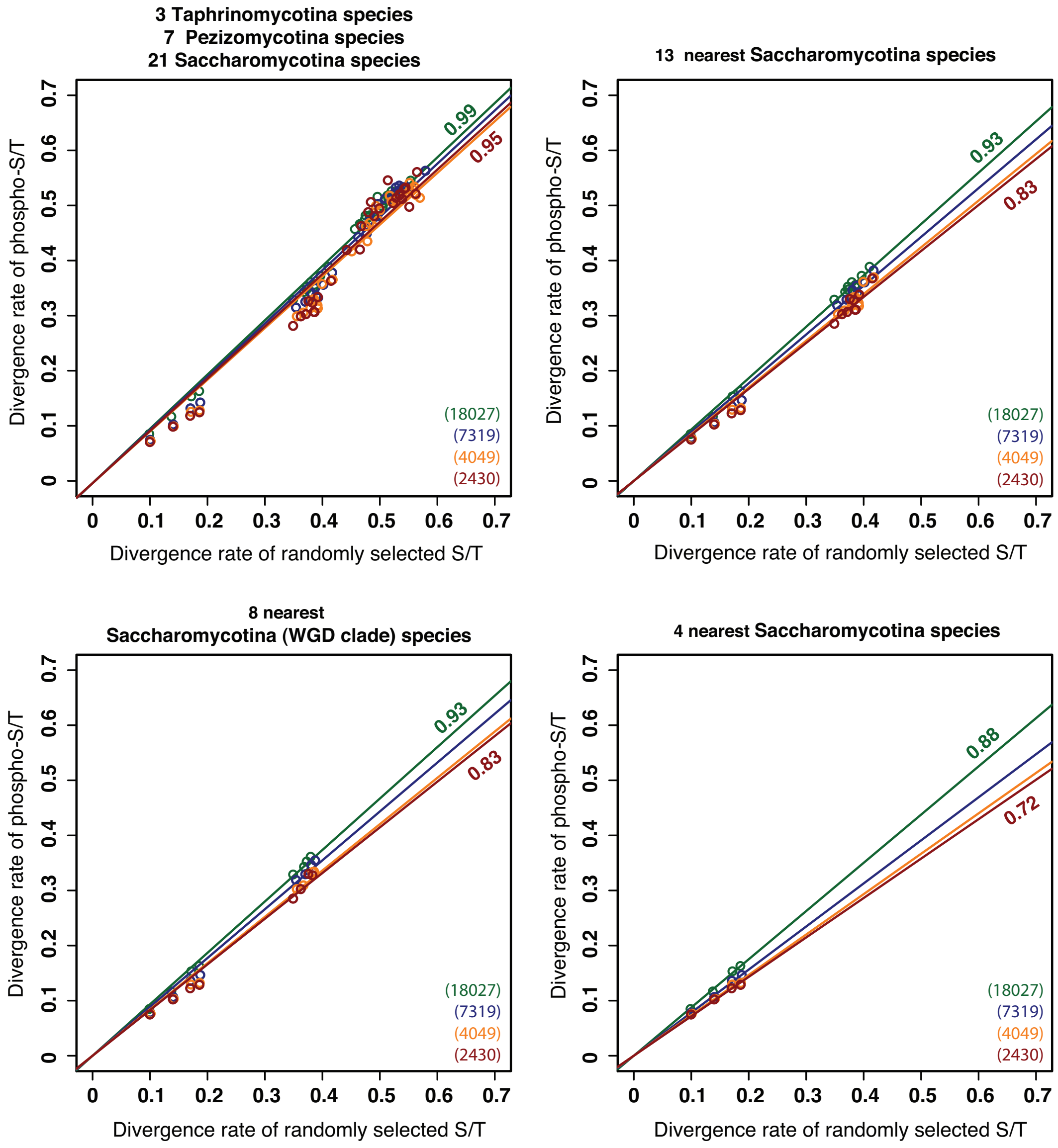

Figure 3.7: Conservation of *S. cerevisiae* phosphorylated residues filtered by their detection frequency across different groups of fungal species. Phosphorylated serine and phosphorylated threonine are analyzed collectively as a group and serine-threonine substitution is deemed conserved. Green, blue, orange and red circles are for phosphorylated serines and phosphorylated threonines detected in ≥1, ≥2, ≥3, ≥4 HTP studies respectively. The best fit regression line is shown for each subset of phosphorylated residues while the coefficient of line are shown for ≥1 and ≥4 HTP studies only. The top left plot considers data points from 31 fungal species for regression analysis. Subsequently, only subsets of the species of decreasing phylogenetic distance to *S. cerevisiae* are analyzed. As the number of residues aligned differs among species due to variations in orthologs inferred and sequence alignment quality, the number of phosphorylated residues aligned to protein sequences from *S. paradoxus* is used to indicate the size of each data set. This number is denoted at the bottom right corner of each plot and color-coded accordingly.

the conservation of *S. cerevisiae* phosphorylation sites in phylogenetically closer species. Considering only species in the *Saccharomycotina* clade, the normalized *RD* is 0.93 and 0.83 for sites detected in at least one and in at least four HTP studies respectively. Further restricting analysis to the four nearest species in the *Saccharomyces* genus improved normalized *RD* to 0.88 and 0.72 for sites detected in at least one and in at least four HTP studies respectively. Hence, it seems that *S. cerevisiae* sites are only well-conserved in close species. The conservation signal is weaker than that for phosphorylated residues observed in human. For example, the set of phosphorylated serines and threonines detected in four or more HTP studies on human proteins has a normalized *RD* of 0.58 across 19 vertebrate species, spanning an estimated 450 million years compared to 0.72 for the same class of sites in yeast across the 4 nearest species in the *Saccharomyces* genus. The fishes (*O. latipes*, *G. aculeatus*, *T. nigroviridis*, *T. rubripes* and *D. rerio*) are the phylogenetically most distant species from human among the 19 vertebrates that are estimated to have diverged from the human lineage $\approx$ 450 million years ago [57]. In contrast, *S. cerevisiae* has diverged from the 4 nearest *Saccharomyces* species less than 150 million years ago [44]. Regardless, the trend that more frequently detected sites are more conserved is still observed for *S. cerevisiae* sites, although we observed that *S. cerevisiae* sites are weakly conserved compared to those detected in human.

3.3.6 *S. cerevisiae* phosphorylated residues are more conserved in more promiscuous interacting proteins

Next, I obtained data on pairwise protein-protein interaction among *S. cerevisiae* proteins detected by Yu *et al.* [201] using a yeast two-hybrid (Y2H) method to determine how protein interaction promiscuity of proteins influence the conservation of phosphorylated residues on them. Yu *et al.*'s dataset is used to estimate the protein interaction promiscuity because they explicitly tested every bait protein against every protein in a prey protein set, hence the data is minimally influenced by ascertainment bias. I reasoned

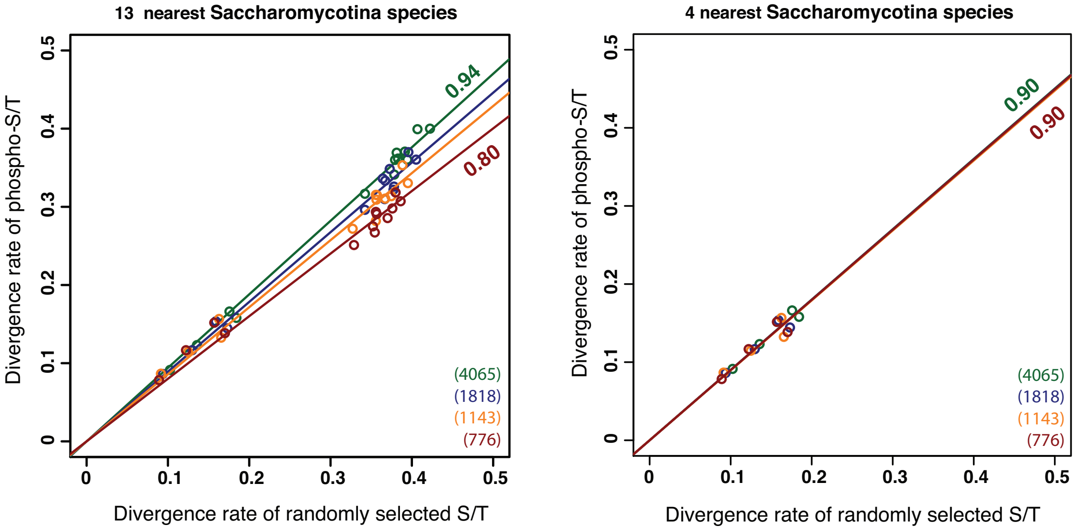

Figure 3.8: Conservation of *S. cerevisiae* phosphorylated residues on proteins filtered by their number of protein interaction partners. Phosphorylated serine and phosphorylated threonine are analyzed collectively as a group and serine-threonine substitution is deemed conserved. Green, blue, orange and red circles are for phosphorylated serines and phosphorylated threonines detected on proteins with ≥1, ≥2, ≥3, ≥4 protein interaction partners respectively as determined in [201]. As the number of residues aligned differs among species due to variations in orthologs inferred and sequence alignment quality, the number of phosphorylated residues aligned to protein sequences from *S. paradoxus* is used to indicate the size of each data set. This number is denoted at the bottom right corner of each plot and color-coded accordingly.

that phosphorylated proteins with more interacting protein partners are more likely to have more functional phosphorylation events implicated in protein-protein interactions, hence more phosphorylated residues on them should be conserved. Therefore, I progressively filtered phosphorylated proteins based on a minimum number of interacting protein partners and analyzed the conservation of phosphorylated residues on them. I considered the scenario that serine and threonine are equivalent in phosphorylation propensity and effect. As previous conservation analysis on the entire set of assembled phosphorylated residues in *S. cerevisiae* revealed that 4 and 13 closest *Saccharomycotina* species provided the most unique information, the normalized divergence rate across 4 and 13 closest *Saccharomycotina* species were computed for this analysis. Comparing conservation of phosphorylated residues in *S. cerevisiae* across 13 nearest *Saccharomycotina* species revealed that phosphorylated proteins that interact with more protein partners progressively have lower normalized *RD* (see Figure 3.8). However, this is not observed for the 4 nearest *Saccharomycotina* species. It is unclear why stronger conservation is only observed for more distantly related species for phosphorylated residues occurring in proteins that interact with more proteins.

3.3.7 *S. cerevisiae* phosphorylated residues on less abundant proteins are more conserved

Intuitively, proteins of higher abundance are more likely to be spuriously phosphorylated due to increased random encounters with protein kinases. Consequently, proteins of higher abundance can have more phosphorylated trypsinized peptides that facilitate their detection by mass spectrometry. Hence, proteins of higher abundance could have more non-functional phosphorylation sites. To find out whether this is a possibility, I collected published information on protein abundance of *S. cerevisiae* proteins determined at log growth phase [47] and analyzed conservation of phosphorylated residues on proteins progressively filtered based on their abundance. Phosphorylation sites on proteins of higher abundance are progressively analyzed at >100, >1,000, >10,000 and >100,000 molecules per cell. I found that for phosphorylated residues on proteins with >100 copies per cell, they are weakly conserved across the 13 nearest *Saccharomycotina* species with a normalized *RD* of 0.94 (Figure 3.9). I observed that phosphorylated residues on proteins with higher abundance showed weaker conservation with normalized *RD* greater than 1 for those on proteins with >10,000 copies per cell. This phenomenon is more pronounced for conservation across the 4 nearest *Saccharomycotina* species, starting from normalized *RD* of 0.90 at >100 copies per cell that progressively increases with higher protein abundance to normalized *RD* of 1.24 at >100,000 copies per cell.

To further validate the above observation, I repeated the analysis across proteins with >10,000 molecules per cell and across those with <1,000 molecules per cell (Figure 3.9,

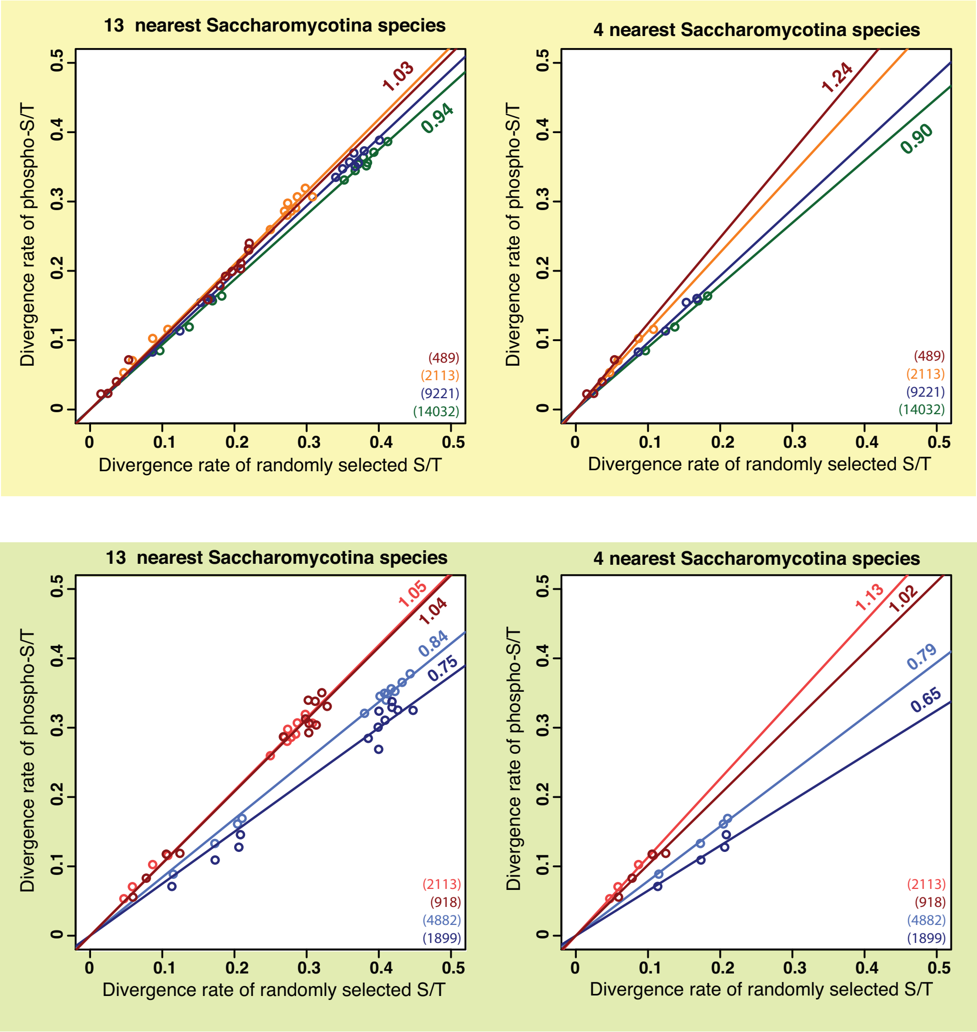

Figure 3.9: Conservation of *S. cerevisiae* phosphorylated residues on proteins filtered by protein abundance. Phosphorylated serine and phosphorylated threonine are analyzed collectively as a group and serine-threonine substitution is deemed conserved. For plots on yellow background, green, blue, orange and red circles are for phosphorylated serines and phosphorylated threonines detected on proteins with >100, >1,000, >10,000, >100,000 molecules per cell. For plots on green background, orange circles are for phosphorylated residues detected in ≥1 HTP studies on proteins with >10,000 molecules per cell, red circles are for those detected in ≥2 HTP studies on proteins with >10,000 molecules per cell, light blue circles are for those detected in ≥1 HTP studies on proteins with <1,000 molecules per cell, dark blue circles are for those detected in ≥2 HTP studies on proteins with <1,000 molecules per cell. As the number of residues aligned differs among species due to variations in orthologs inferred and sequence alignment quality, the number of phosphorylated residues aligned to protein sequences from *S. paradoxus* is used to indicate the size of each data set. This number is denoted at the bottom right corner of each plot and color-coded accordingly.

green panel). I also analyzed the subsets of phosphorylated residues on these proteins that have been detected by 2 or more HTP studies. Across both the 13 nearest and the 4 nearest *Saccharomycotina* species, phosphorylated residues on low abundance proteins (those less than 1,000 molecules per cell) show strong conservation, with a stronger signal for the subset of phosphorylated residues detected in multiple HTP studies. However, phosphorylated residues on high abundance proteins (those greater than 10,000 molecules per cell) are not well-conserved with normalized *RD* greater than 1, even for the subset of phosphorylated residues detected in multiple HTP studies. Hence, it is plausible that highly abundant proteins have proportionally more spurious phosphorylation sites than lower abundant proteins. Until now, all sets of phosphorylated residues analyzed are changing either at the same pace or slower than background phosphorylatable residues. Hence, the observation that the set of phosphorylated residues on highly abundant yeast proteins are potentially changing faster than background phosphorylatable residues is interesting. However, more tests and analysis are needed to assess whether these sites are under positive selection.

3.3.8 *M. musculus* phosphorylated residues detected from tissues are highly conserved

So far, the HTP phosphorylation sites analyzed were identified in free-living single cells from the unicellular organism *S. cerevisiae* and human cell lines. Using recently published phosphorylation sites detected in nine mouse tissues using MS-based experimental methods, I investigated the potential difference in conservation for phosphorylated residues detected in tissues and free-living single cells of unicellular organism *S. cerevisiae* and human cell lines. I computed the normalized *RD* for phosphorylation sites detected in each mouse tissue across 19 vertebrate species which include human. Interestingly, *M. musculus* phosphorylated residues across the nine tissue are conserved at similar rates with *RD* ranging from 0.64 to 0.67. These RD rates are lower (more conserved) than that observed for global HTP phosphorylation sites from human cell lines ($RD \approx 0.85$, Figure 3.3) and *S. cerevisiae* ($RD = 0.88$ across the four nearest species, Figure 3.7). Most importantly, the conservation rate of functionally characterized human phosphorylation sites ($RD = 0.56$) are more similar to that of *M. musculus* phosphorylation sites ($RD \approx 0.66$) than to HTP human phosphorylation from cell lines ($RD \approx 0.85$). This result suggests that a large portion of the phosphorylated residues detected from *M. musculus* tissues are under evolutionary constraint and hence more likely to be functionally important.

The underlying experimental procedure and site localization algorithm(s) adopted in the mouse project might have produced cleaner data (less false positive sites) leading to the observed low *RD*. In comparison, the set of human phosphorylated residues detected in cell lines by two or more HTP studies has a higher *RD* (see Figure 3.3). Assuming that most of the false positive sites had been removed in this set of phosphorylated residues, the result suggests other factors may be at play. One other plausible and non-exclusive explanation is that phosphorylation events in tissues are more tightly regulated and hence less likely to be sporadic compared to those detected in free-living single cells. For phosphorylation sites detected in human cell lines, it is possible that a portion of them arose from deregulated protein kinases and impaired phosphatases. However, frequently detected sites in *S. cerevisiae*, such as those detected in 4 or more HTP studies, are weakly conserved across very closely related species (see Figure 3.7, bottom right plot) compared to the conservation rates of phosphorylated residues in mouse and human. Not withstanding that unicellular organisms like *S. cerevisiae* are possibly diverging faster than multicellular organisms, this comparison among sites from mouse tissues, human cell lines and *S. cerevisiae* overall suggests that stochastic (non-determisitic) phosphorylation events could be prevalent in free-living single cells.

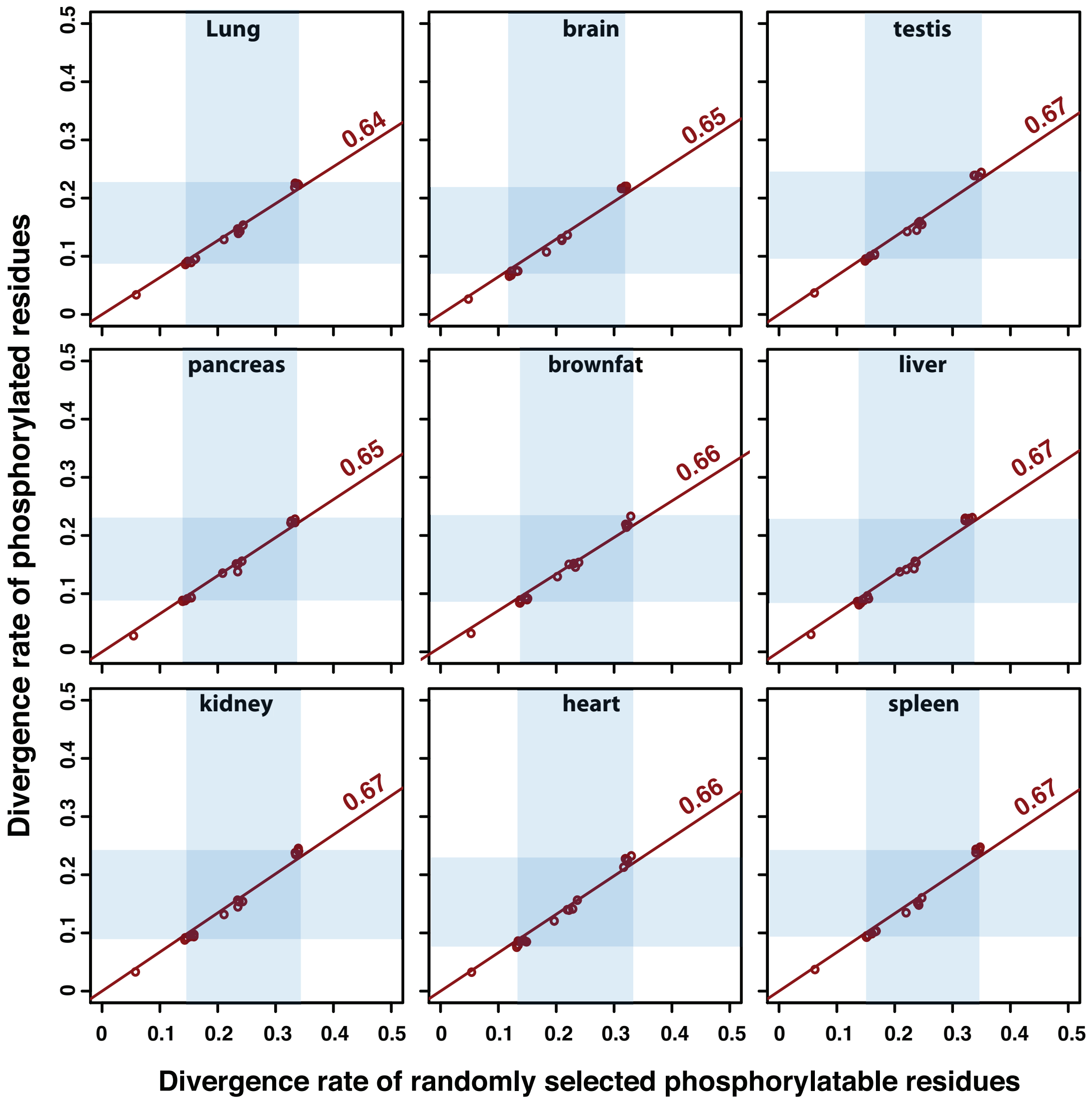

Figure 3.10: Normalized relative divergence rate of *M. musculus* phosphorylated residues detected in different tissues.The divergence rate of phosphorylated residues and background phosphorylatable residues in *M. musculus* across each of the 19 vertebrate species are computed and plotted for sites detected in each tissue. The divergence rate is the portion of residues in *M. musculus* not conserved in each species as determined from multiple sequence alignment of sets of orthologous protein sequences. A best-fit line was determined for each plot using linear regression. The most prevalent normalized *RD* across the 19 species is the coefficient shown on the line. Excluding the phylogenetically nearest species, the blue shadings highlight the maximum and minimum divergence rate at each axis.

3.4 Discussion

As most of the phosphorylation sites assembled for this study are identified in HTP studies using mass spectrometry, it helps for result interpretation to understand the basic techniques and procedures behind the detection method. The basic steps in the method are: (1) the breaking down of proteins in a sample into peptides, often by trypsin, (2) the enrichment of phosphorylated peptides from non-phosphorylated peptides, (3) the fractionation of phosphorylated peptides using separation techniques like high-performance liquid chromatography (HPLC), (4) fragmentation of separated peptides to produce mass spectra and (5) finally, the matching of mass spectra produced uniquely by different peptides to a spectra ”fingerprint” database to identify peptide sequences and the localization of phosphorylation sites on peptides.

Phosphorylation sites on a peptide can be easily falsely localized if multiple phosphorylatable residues are present on the peptide. Hence, falsely identified phosphorylation sites can contribute to lower conservation observed for phosphorylation sites identified in mass spectrometric screens particularly in HTP studies where localized sites are often not validated by manual inspection of spectra. This hypothesis seems to be supported by the observation that phosphorylated residues of sites detected in multiple HTP studies are more conserved, based on the reasoning that phosphorylation sites detected with different site-localization algorithms across multiple HTP studies are more likely to be true sites. However, another possible and non-exclusive explanation for the lower conservation observed for HTP phosphorylation sites is spurious or random phosphorylation events. It is plausible that phosphorylation sites identified by multiple HTP studies possibly under different physiological conditions are less likely to be the result of spurious phosphorylation events. The observation that more commonly detected phosphorylation sites are more conserved is consistent for human and *S. cerevisiae* sites although for *S. cerevisiae*, it is only observed for phylogenetically very close yeast species.

I observed that functionally characterized phosphorylation sites on human proteins are equally conserved across structurally disordered and ordered regions which suggest similar evolutionary constraints imposed by functional phosphorylation sites across the two types of protein regions. However, phosphorylated serines and phosphorylated threonines detected in HTP studies are more conserved in structurally disordered regions than in structurally ordered regions in proteins, even for more frequently detected subset of phosphorylation sites. I expected proportionally more spurious phosphorylation events to occur on structurally disordered regions which, in general, should be more accessible to the catalytic site of protein kinases. However, the result suggests that proportionally more phosphorylation sites detected within structurally ordered protein region are spurious. It is plausible that phosphorylatable residues in structurally ordered region are typically less accessible to protein kinases than those from structurally disordered regions, but are spuriously targeted by protein kinases when proteins are broken down into peptides by trypsin in typical mass spectrometric screens. On the other hand, phosphorylatable residues in structurally disordered regions, being more readily accessible by protein kinases, may have evolved mechanisms to reduce spurious phosphorylation, such as changing the amino acids flanking phosphorylatable residues that discourage contact with the catalytic site of protein kinases. For reasons that are unclear, phosphorylated tyrosines across structurally disordered and ordered protein regions seem to diverge at the same rate even for the more frequently detected subset of sites.

One of the more interesting observations is that phosphorylation sites on proteins of high abundance are seemingly less conserved than those on proteins of low abundance if we compared their conservation rate relative to background phosphorylatable residues. Assuming the probability of spurious phosphorylation occurring on any phosphorylatable residue on any protein is $X\%$, highly abundant proteins will have, in absolute number, more peptides with spurious phosphorylation sites. More phosphorylated peptides should translate into a higher probability of detection in mass spectrometric screens. This can explain the lower conservation of phosphorylation sites observed on proteins of high abundance. In fact, phosphorylation sites on proteins of high abundance seem to be changing faster than background phosphorylatable residues suggesting that some of them are under positive selection, hence may be involved in cell biology unique in *S. cerevisiae*.

I reasoned that phosphorylation sites of higher stoichiometry should be less likely to result from spurious phosphorylation events. However, I observed that phosphorylated residues of higher site stoichiometry are only marginally more conserved than those of lower site stoichiometry. One plausible explanation is the site stoichiometry information is not derived accurately. As the phosphorylation sites are identified in HeLa cells, another explanation is that site stoichiometry had changed uniquely in cancerous cells like HeLa as a result of deregulated protein kinases. Hence, phosphorylation sites of high stoichiometry identified from a cancerous cell line may not be important for physiological and morphological development of a multicellular organism. A somewhat related observation is that *S. cerevisiae* phosphorylation sites are only more conserved than background phosphorylatable residues in phylogenetically close species (Figure 3.7). In general, human phosphorylation sites seems to be more conserved than phosphorylation sites in *S. cerevisiae* perhaps because phosphorylation sites in multicellular organisms are subjected to stronger evolutionary constraints needed to prevent cell anarchy in a multicellular context.

3.5 Conclusion

In this analysis, we have identified many factors that influence the conservation of phosphorylated residues detected in one species across other species. A portion of the non-conserved phosphorylated residues could have been falsely identified or a result of spurious phosphorylation. Some of the spurious phosphorylation events could have been introduced by experimental treatments. Also, some of the non-conserved phosphorylation sites could be unique in cancerous cells as a result of deregulated protein kinases.

Although spurious phosphorylation sites or those occurring uniquely in cancerous cells are less conserved than background phosphorylatable residues, their non-conservation does not necessarily mean they do not affect cellular activities. Conservation analysis can be used to identify phosphorylation events that are potentially evolutionarily conserved for fundamental or conserved cellular activities. The converse, that is non-conserved phosphorylation events do not have functional roles or are not functionally important, is not necessarily correct as new phosphorylation events may be implicated in the unique biology of a species.

Chapter 4

Evolutionary dynamics of phosphotyrosine signaling systems in metazoan, yeast and choanoflagellate

Half of the work presented in this chapter (Section 4.3.1 and associated Materials and Methods) was published in:

C. S. Tan, A. Pasculescu, W. A. Lim, T. Pawson, G. D Bader, R. Linding. Positive selection of tyrosine loss in metazoan evolution. *Science*, 325(5948):1686-8, 2009 Sep.

The other half of the work presented in this chapter (Section 4.3.2 and associated Materials and Methods) will be published in *Science* magazine as a response to a technical comment raised for above publication. All work presented in this chapter were carried out by me except 1) domain prediction described in Section 4.2.4 which was carried out by R. Linding but I processed the result, and 2) NetPhorest prediction described in Section 4.2.6 and analysis presented in Figure 4.11 which was carried out by E. Schoof and P. Creixell but I conceived the experiment and interpreted the result.

4.1 Introduction

It is a biological paradox that organism complexity shows limited correlation with gene repertoire size for metazoan species [178]. For example, there are ≈ 22,000, 20,000 and 14,000 inferred genes in human, *C. elegans* and *D. melanogaster* respectively based on release 51 of the Ensembl database although human seemingly looks much more complex than *C. elegans* in term of size and morphology. However, it has been observed that the occurrence of some protein domain families do correlate positively with organism complexity in metazoan as approximated by the number of cell types in each species [188, 179], especially those involved in regulation of cellular process. Of interest relating to protein phosphorylation are the protein domain families that make up the phosphotyrosine signaling machinery in metazoan. One of the domain families consists of tyrosine kinases which catalyze the transfer of a phosphate from ATP to hydroxyl group on tyrosine residues. Tyrosine kinases are believed to facilitate multicellularity in metazoan because their known roles in cell-cell communication and tissue boundary formation in conjunction with their almost unique presence in metazoan. While tyrosine kinases transfer phosphate from ATP to tyrosine, tyrosine phosphatases catalyze the reverse to keep the level of tyrosine phosphorylation in check. Some phosphorylated tyrosines become temporal binding sites for phosphotyrosine binding protein domains such as SH2 (Src homology) and PTB (phosphotyrosine binding) domains. The evolution of phosphotyrosine signaling machinery has been studied in terms of occurrences and associated domain architecture of tyrosine kinase, tyrosine phosphatase and phosphotyrosine-binding protein domains as a whole system. However, there has been no systematic attempts to understand the evolution of phosphotyrosine signaling system contributed by the changes in tyrosine frequency.

For the last part of my dissertation work, I analyzed the maximum phosphorylation potential encoded in the proteomes of various metazoan species, *S. cerevisiae*, *S. pombe* and the choanoflagellate *M. brevicollis*. This is in contrast to my previous works where

I analyzed known phosphorylation sites. Here, I analyzed how the number of protein kinases could influence the maximum phosphorylation potential in various species as approximated by the frequencies of various phosphorylatable residues.

4.2 Materials & Methods

4.2.1 Collection of protein sequences and coding DNA sequences

Known and inferred protein sequences of budding yeast (*S. cerevisiae*), fission yeast (*S. pombe*), worm, (*C. elegans*), sea squirt (*C. intestinalis*), fly (*D. melanogaster*), mosquito (*A. gambiae*), zebrafish (*D. rerio*), tetraodon pufferfish (*T. nigroviridis*), Japanese pufferfish (*T. rubripes*), frog (*X. tropicalis*), chicken (*G. gallus*), dog (*C. familiaris*), cow (*B. taurus*), mouse (*M. musculus*), rat (*R. norvegicus*), chimpanzee (*P. troglodytes*), and human (*H. sapiens*) were obtained from Ensembl online database (release 51). The known and inferred protein sequences of *M brevicolliss* under the *Filtered Models (”best”)* were obtained from http://genome.jgi-psf.org/Monbr1/Monbr1.home.html at the DOE Joint Genome Institute. Sequence sets are processed to retain the longest protein translation of each gene for each species. For data sets obtained from Ensembl online database, the coding DNA sequence for the longest protein translation of each gene is extracted from corresponding cDNA sequence given in Ensembl using custom perl script. Similarly, the coding DNA sequence for each *M brevicollis* protein is extracted from corresponding DNA transcript obtained from http://genome.jgi-psf.org/Monbr1/Monbr1.home.html.

4.2.2 Identification of tyrosine-phosphorylated human proteins

Experimentally determined phosphotyrosine sites in human proteins were obtained from the Phospho.ELM [36] and PhosphoSitePlus [64] databases in November 2008 and mapped to obtained human protein sequences. In total, the dataset contains 12,659 phosphotyrosines in 6450 proteins. A human protein is classified as tyrosine-phosphorylated (*p*Tyr) if any of its tyrosines is phosphorylated in our assembled phosphorylation data, or otherwise classified as Non-*p*Tyr protein.

For each human gene product, all known and predicted splice variants were aligned using the AMAP multiple sequence alignment software [157], and non-redundant amino acid residue counts were computed from the alignments. We found no substantial difference in tyrosine residue counts using an alternative approach considering only the longest translation of each human gene. Genes coding 200 amino acids or less are excluded from computation to reduce sizable, but non-significant percentage changes in tyrosine content due to small protein size.

4.2.3 Identification of human-yeast orthologous sequence protein pairs

All known and predicted human, *S. cerevisiae* and *S. pombe* protein sequences retrieved from Ensembl (release 51) are processed to retain only the longest protein translation of each gene. Human-yeast orthologous proteins were then inferred using the Inparanoid algorithm [142] (version 2.0) using BLOSUM62 as recommended for eukaryotic species and the downloaded sequences, based on stringent bi-directional best BLAST [3] hits with the processed human and yeast protein sequences.

4.2.4 Computing occurrence of protein domains

First, all HMM models from SMART protein domain database [95] were input into the SMART text-mode pipeline of HMMER to scan the set of longest protein translation of each gene in each species. The occurrence of tyrosine kinase (SMART ID: TyrKc), serine/threonine kinase (SMART ID: S_TKc, S_TK_X) , SH2 (SMART ID: SH2) and PTB (SMART ID: PTB) protein domains in each species were then computed based on a E-value cutoff of 1.0E-3 for observed sequence similarity score. Inferred orthologs of human

MEK and MLK kinases across the 16 species were retrieved from Ensembl/Compara for the analysis.

4.2.5 Computing amino acid substitution rate between human-yeast orthologous protein pairs

Pairwise sequence alignments between orthologous human-*S.cerevisiae* and human-*S.pombe* protein pairs are performed using MAFFT with default parameters [81]. To avoid accelerated sequence divergence due to functional redundancy of paralogs, only protein pairs with an inferred one-to-one orthologous relationship between human and *S. cerevisiae/S. pombe* are aligned. We then computed the frequency of mutated phenylalanine and tryptophan from yeast to human and vice versa being substituted by tyrosine. To reduce error due to faulty alignments, residues with less than 5 identical aligned flanking residues out of 10 positions (5 on each side) are excluded.

4.2.6 Predicting phosphorylation propensity of tyrosine conserved between human and yeast

For tyrosines that are conserved at the same positions on orthologous human-*S. cerevisiae* protein pairs as identified from pairwise sequence alignments, they are input to NetPhorest algorithm to investigate their phosphorylation propensity by human tyrosine kinases. The phosphorylation propensity of aligned tyrosines in human and *S. cerevisiae* as determined by the amino acids flanking them as inferred by NetPhorest is compared.

4.2.7 Statistical analysis

All statistical tests were performed using the R statistical package. Differences in tyrosine, phenylalanine, tryptophan quantity (either as frequency or as absolute count) of humanyeast orthologous protein pairs were computed and distributions of the computed differences for the pY and Non-pTyr proteins are assessed using the Mann-Whitney test.

4.3 Result and Discussion

4.3.1 Positive selection of tyrosine loss in metazoan evolution

As phosphotyrosine signaling system is implicated in cell-cell communication and tissue boundary formation in metazoan, it likely facilitates the emergence of multicellular metazoan and contributed to organism complexity among metazoan species. Studies have focused on the co-expansion of components in phosphotyrosine signal machinery, such as tyrosine kinase and phosphotyrosine binding domains, across different species. However, the tyrosine phosphorylation potential of a biological is partially determined by frequency of tyrosine present in the proteome, in addition to tyrosine kinases. To gain insight into the relationship between maximum tyrosine phosphorylation potential of a biological system and organism complexity, I analyzed the correlation of genomic tyrosine frequency using the number of cell type in species as a proxy for organism complexity.

I observed a striking negative correlation of genomic tyrosine frequency with the number of distinct cell types in 15 metazoan species [188] and *S. cerevisiae* (Spearman's r = –0.89, $P \approx 3.0 \times 10^{-6}$; Pearson's r = –0.89, $P \approx 4.0 \times 10^{-6}$ Figure 4.1). I included *S. cerevisiae* as a unicellular eukaryote for comparison. The genomic tyrosine frequency is the portion of amino acids of all known and inferred proteins in each species that are tyrosine residues. Thus, metazoans with more cell types have proportionally less potential phosphotyrosine. Similarly, we observed that the number of tyrosine kinase domains correlates negatively with genomic tyrosine frequency (Spearman's r = -0.68, $P \approx 3.7 \times 10^{-3}$; Pearson's r = –0.81, $P \approx 1.3 \times 10^{-4}$, Figure 4.2). Including dual-specificity mixed lineage kinases (MLKs) and mitogen-activated protein kinase kinases (MEKs) revealed a similar pattern (Figure 4.2).

I also observed statistically significant negative correlation of the number of distinct

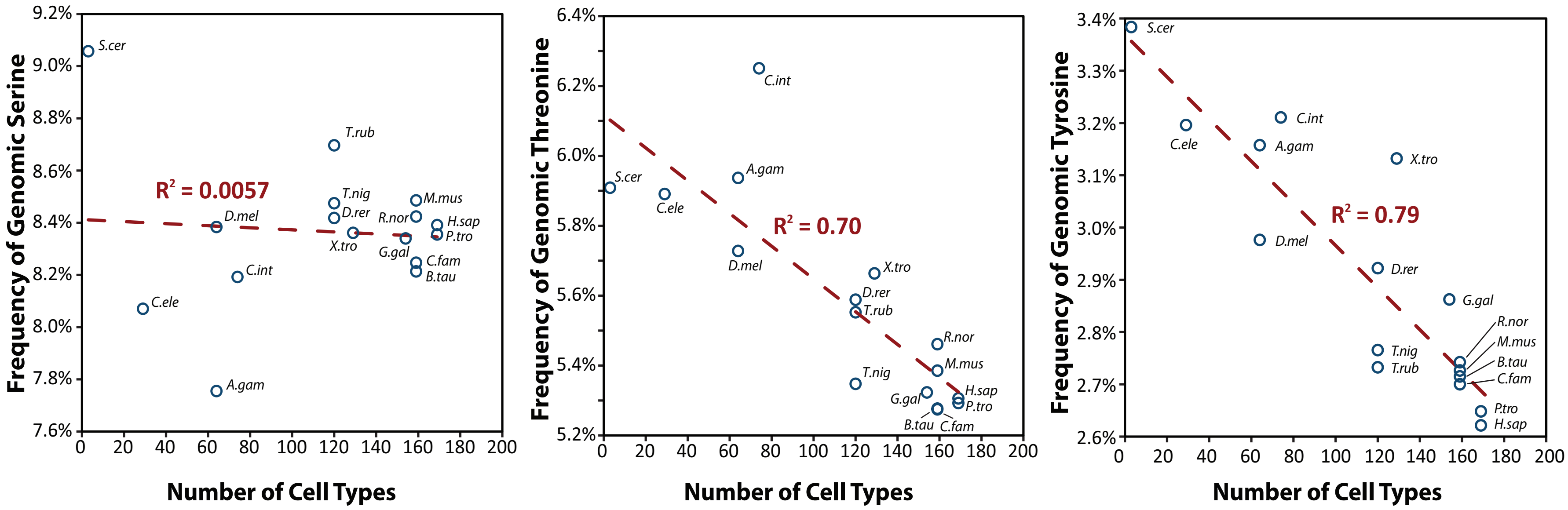

Figure 4.1: Correlation of cell type number with genomic serine, threonine and tyrosine frequencies across 14 metazoan species and *S. cerevisiae*. The budding yeast (S. cerevisiae) is included as a unicellular eukaryote for comparison. The species analyzed are yeast (***S. cerevisiae***), worm, (***C. elegans***), sea squirt (***C. intestinalis***), fly (***D. melanogaster***), mosquito (***A. gambiae***), zebrafish (***D. rerio***), tetraodon pufferfish (***T. nigroviridis***), Japanese pufferfish (***T. rubripes***), frog (***X. tropicalis***), chicken (***G. gallus***), dog (***C. familiaris***), cow (***B. taurus***), mouse (***M. musculus***), rat (***R. norvegicus***), chimpanzee (***P. troglodytes***), and human (***H. sapiens***).

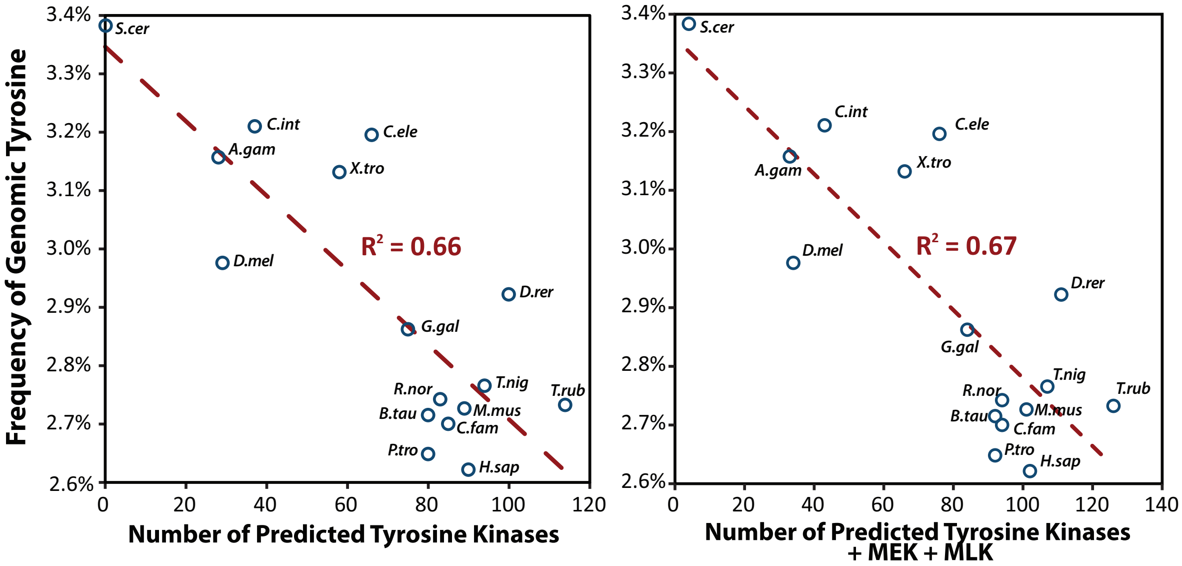

Figure 4.2: Correlation of genomic tyrosine frequencies with the number of predicted tyrosone kinase domain in metazoan and yeast species. Although dual-specificity mixed lineage kinases (MLKs) and mitogen-activated protein kinase kinases (MEKs) are not dedicated tyrosine kinases, they are also analyzed because they can phosphorylate tyrosine albeit less efficient than dedicated tyrosine kinase.

cell types with genomic threonine frequency (Spearman's r = –0.85, $P \approx 3.7 \times 10^{-5}$; Pearson's r = –0.84, $P \approx 5.3 \times 10^{-5}$) but not with genomic serine frequency (Figure 4.1). The number of inferred serine/threonine kinase domains also seems to correlate negatively with genomic threonine frequency (Spearman's r = –0.51, $P \approx 4.5 \times 10^{-2}$; Pearson's r = –0.70, $P \approx 2.4 \times 10^{-3}$, Figure 4.3) and not with genomic serine frequency (Spearman's r = –0.46, $P \approx 7.5 \times 10^{-2}$; Pearson's r = –0.26, $P \approx 3.4 \times 10^{-1}$, Figure 4.3). As the negative correlation observed for genomic threonine frequency with number of cell type and with the number of serine/threonine kinase domain is relatively weaker than the trend observed for genomic tyrosine, I focus on characterizing the evolutionary dynamics of tyrosine and how it is associated with phosphotyrosine signaling.

The expression of Src tyrosine kinase in the unicellular *S. pombe* and *S. cerevisiae*, whose genomes encode no tyrosine kinase, is found to be toxic due to deleterious tyrosine phosphorylation [176, 177]. This together with the observed negative correlation of tyro-

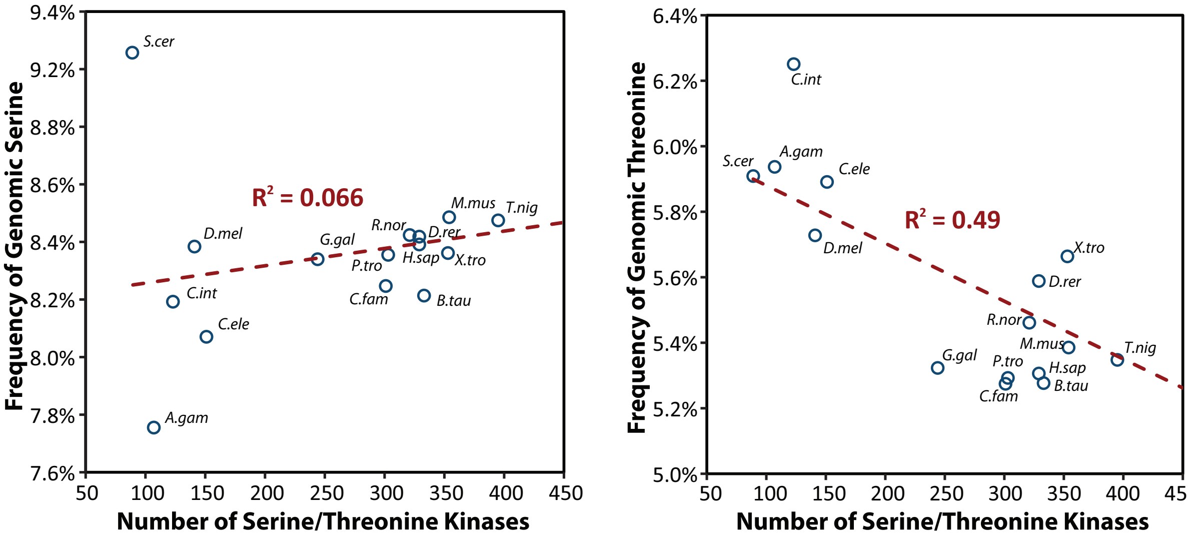

Figure 4.3: Correlation of genomic serine and threonine frequencies with the number of predicted serine/threonine kinase domain in metazoan and yeast species.

sine kinase number with tyrosine frequency suggest an evolutionary model in which the acquisition of a tyrosine kinase results in systems-level adaptation to remove deleterious phosphorylation events that cause aberrant cellular behavior and diseases [66]. Assuming that a cell begins with a single tyrosine kinase, which is subsequently duplicated, it follows that the kinases may functionally diverge, as a result of relaxation in evolutionary constraints, to phosphorylate new substrates. Emerging kinase specificities could be retained if new substrates confer selection advantage. However, it is unlikely that every new phosphorylation event is beneficial. I hypothesize that optimization of newly emerged signaling networks would follow [202] through the elimination of detrimental phosphorylation events by tyrosine-removing mutations. Even if many new phosphorylation sites are not deleterious, an organism with minimized noisy signaling systems is likely to have a fitness advantage. This scenario is repeated with the subsequent duplication of tyrosine kinases leading to more tyrosine residues lose (7).

Despite several recent systematic phosphoproteomic studies [78], many human proteins have no observed phosphotyrosines. Our model suggests that tyrosine loss has occurred predominantly in these proteins in order to minimize tyrosine phosphorylation.

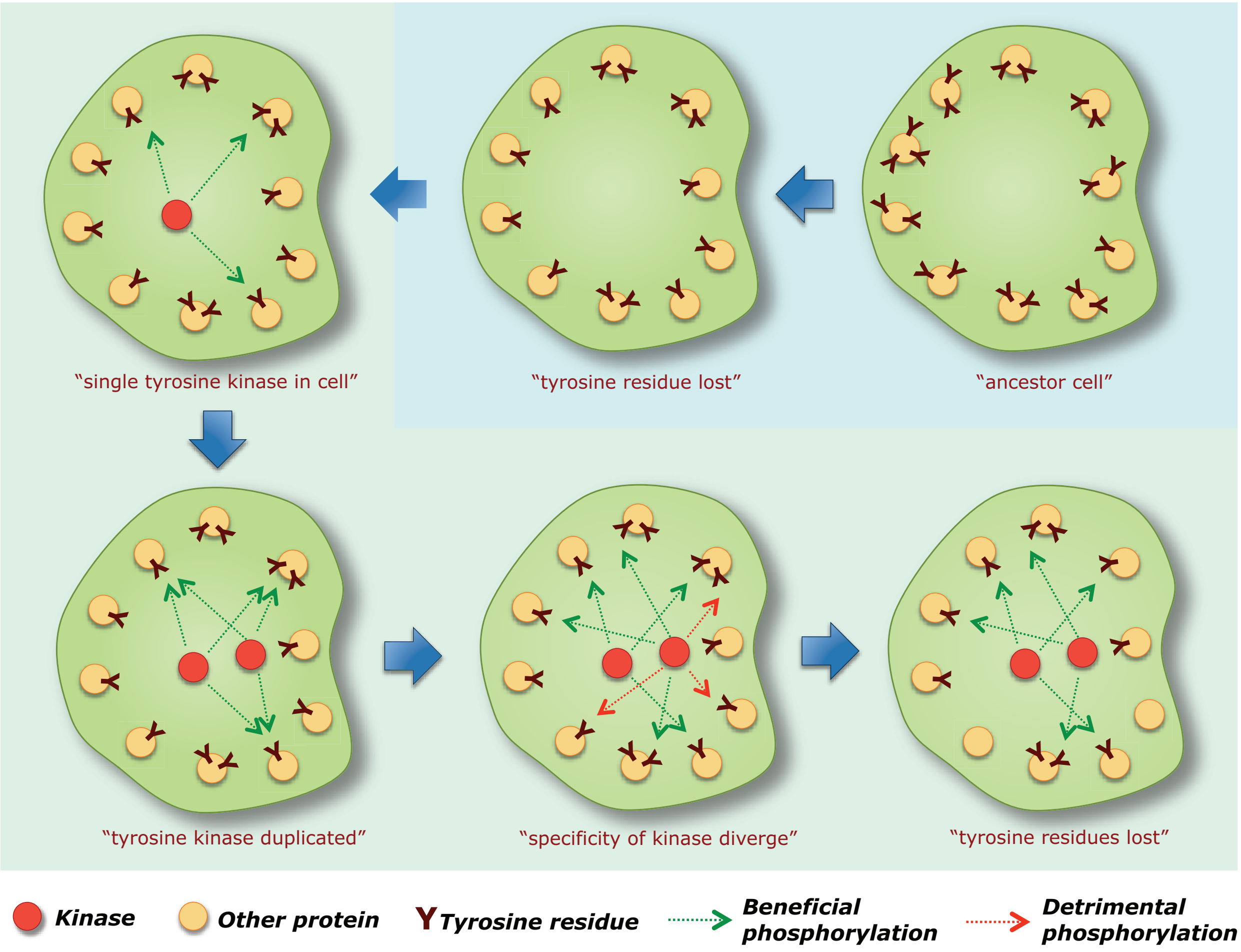
"single tyrosine kinase in cell"
"tyrosine residue lost"
"ancestor cell"
"tyrosine kinase duplicated"
"specificity of kinase diverge"
"tyrosine residues lost"
Kinase
Other protein
Y Tyrosine residue
Beneficial phosphorylation
Detrimental phosphorylation

Figure 4.4: Proposed evolutionary model leading to observed tyrosine depletion in metazoan. The blue-green panel is extension to the original evolutionary model proposed (green panel). As expression of Src tyrosine kinase in the unicellular *S. cerevisiae* and *S. pombe*, whose natural genomes encode no tyrosine kinase, is toxic due to deleterious tyrosine phosphorylation, it is plausible an initial tyrosine depletion occurred to permit the appearance of the first tyrosine kinase. When a tyrosine kinase is duplicated, and evolved new specificity to target different tyrosine that is overall beneficial to survival and propagation, the new kinase will be retained during evolution. However, it is plausible some new tyrosine phosphorylation events might be deleterious which can be eliminated through removal of specific tyrosines while retaining the beneficial phosphotyrosine sites. With more tyrosine kinase, more tyrosine are removed and same time exert a stronger constraint on the appearance of new tyrosine.

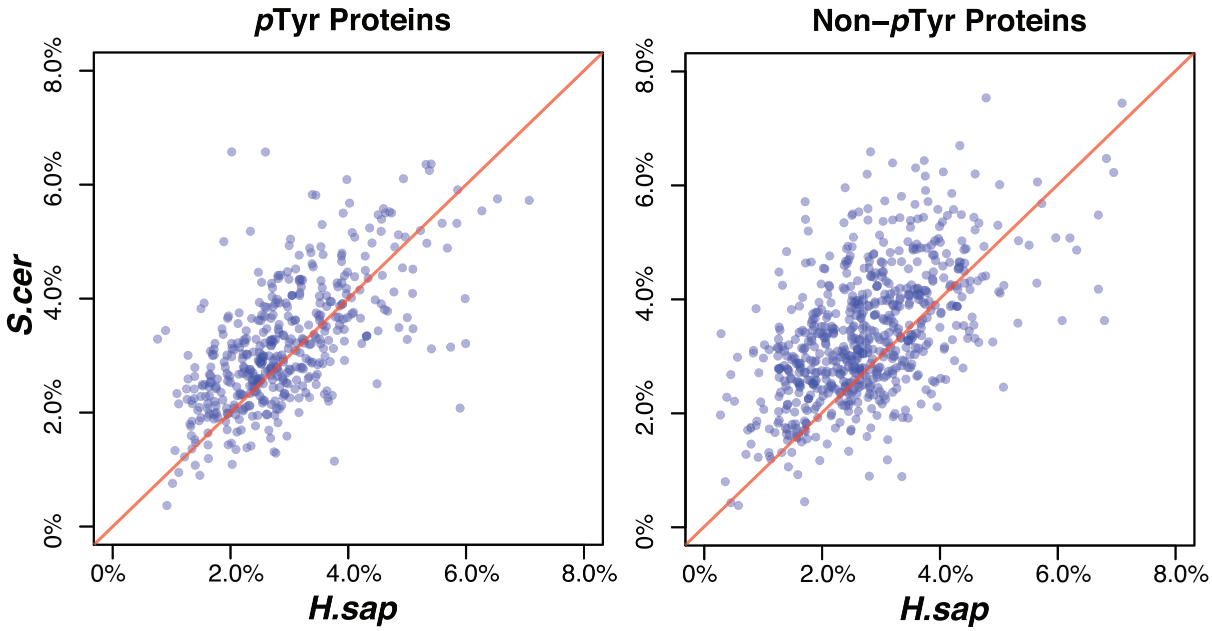

Figure 4.5: The tyrosine frequency in human-yeast ortholog protein pairs. Every point in the scatter plot represents a human-yeast ortholog protein pair where the (x, y) values denote the tyrosine content in human and yeast proteins, respectively. For simplicity, only proteins with an inferred one-to-one orthologous relationship between human and yeast are analyzed (for example, to avoid accelerated sequence divergence due to functional redundancy of paralogs). Orthologous protein pairs lying above the red diagonal ($x = y$) lines have higher tyrosine composition in yeast than in human. The left scatter plot is for 437 human proteins conserved in yeast and known to be tyrosine phosphorylated, and the right plot is for 647 human proteins conserved in yeast not known to be tyrosine phosphorylated.

To test this hypothesis, we investigated differences in tyrosine loss between these proteins (Non-pTyr) and those that are tyrosine phosphorylated (pTyr). Comparing members of these two groups to their orthologous proteins in *S. cerevisiae*, which lack conventional tyrosine kinases, enabled us to assess the degree of tyrosine loss that may be triggered by the onset of phosphotyrosine signaling in metazoans.

A significantly smaller fraction of amino acids are tyrosines in human proteins than in their yeast orthologs ($P \approx 3.5 \times 10^{-4}$, paired Wilcoxon signed rank test, Figure 4.5). However, this phenomenon was statistically more pronounced in non-pTyr proteins than in pTyr proteins ($P \approx 5.1 \times 10^{-9}$, Mann-Whitney test, Figure 4.5). A similar trend was observed on the basis of absolute tyrosine residue counts ($P \approx 2.0 \times 10^{-7}$, Mann-

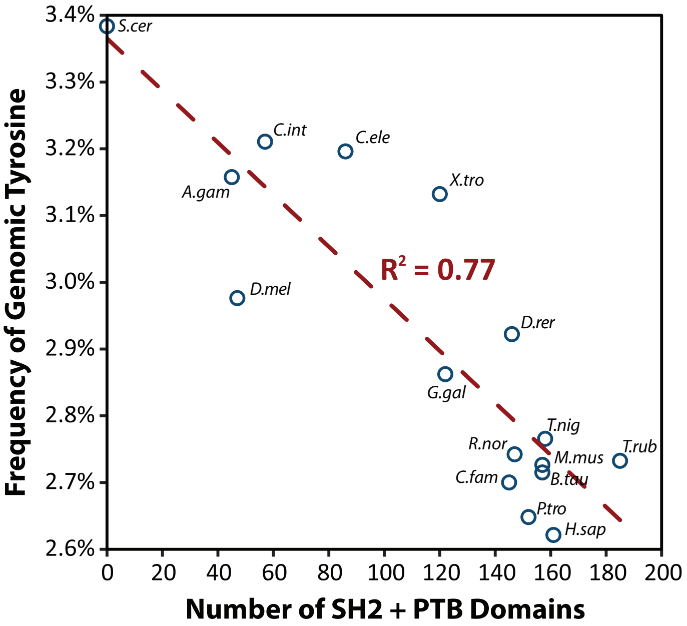

Figure 4.6: Correlation of genomic tyrosine frequencies with the number of predicted SH2 and PTB domains in metazoan and yeast species.

Whitney test) and on a higher confidence subset of pTyr proteins that either have multiple phosphotyrosines or have sites observed in multiple studies ($P \approx 1.3 \times 10^{-7}$, Mann-Whitney test).

Thus, tyrosine loss was strongly favored in human protein evolution, most notably in protein subsets that are not known to be tyrosine-phosphorylated. Genetic drift [197] is unlikely to account for these differences observed in a large number of evolutionarily distant human-yeast protein orthologs. Because tyrosine is an essential and the most expensive amino acid to biosynthesize [140] after tryptophan and phenylalanine, essentiality and biosynthetic cost could be major factors in the observed loss. However, this is unlikely because we observed a strong positive correlation of number of cell types with tryptophan and a weaker negative correlation for phenylalanine.

Thus, I propose that positive selection of tyrosine-removing mutations occurred in the metazoan lineage to reduce adventitious tyrosine phosphorylation, at least in part. This optimization process probably shaped signaling networks crucial for the development of multicellular animals. Additionally, this could provide a mechanism to prevent unspecific phosphorylation events that operates with the evolution of domains, consensus motifs [200], and contextual factors to colocalize kinases with their substrates [200, 98, 114]. We observed a slightly stronger negative correlation of genomically encoded tyrosine content with the number of inferred phosphotyrosinebinding domains than tyrosine kinase domain count (Spearman's r = -0.81, Pearson's r = -0.88, Figure 4.6), which is in agreement with the notion that tyrosine phosphorylation exerts parts of its functional effects through creating binding sites for phosphobinding domains like Src homology 2 (SH2) and phosphotyrosine binding domain (PTB) [160].

The choanoflagellate Monosiga brevicollis, which is a member of the only known unicellular lineage with canonical tyrosine kinases [86], is an outlier in the cell-type correlation studied above. This observation is consistent with the emerging picture that choanoflagellates represent a distinct evolutionary branch from metazoans in which phosphotyrosine-signaling systems have been used for divergent functions [135, 109]. Nevertheless, the Monosiga analysis is still consistent with optimization of phosphotyrosine signaling in this lineage; compared with the metazoans analyzed here, Monosiga has higher numbers of tyrosine kinases (127) and lower genomically encoded tyrosine content (2.3%).

4.3.2 Effect of GC directional force on tyrosine frequency

Elevated GC content (G+C) has been observed in some metazoan species, especially among the warm-blood mammals. Biased A/T → G/C nucleotide substitution due to various factors is postulated to cause this phenomenon. As tyrosine is encoded by two AT-rich codons (TAT, TAC), the observed tyrosine depletion in metazoan species might have been facilitated by biased A/T → G/C nucleotide substitution. Here, I investigated whether the observed tyrosine depletion has been passively driven, in the absence of natural selection for amino acid changes, by 1) the mutational forces behind high GC

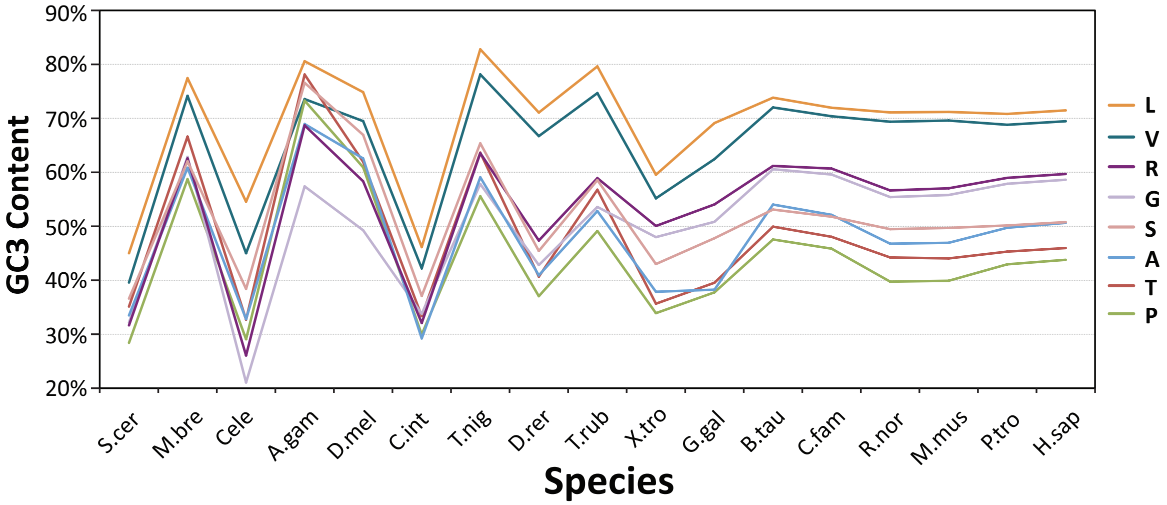

Figure 4.7: GC content at the 3rd codon position (GC3) for each set of four-fold degenerate codons in different species. The four-fold degenerate codon set of each amino acid are GCN (Alanine), CGN (Arginine), GGN (Glycine), CTN (Lecuine), CCN (Proline), TCN (Serine), ACN (Threonine) and GTN (Valine). The species are sorted with decreasing evolutionary distance from human. The GC3 content for these eight amino acids are highly correlated with each other (average Pearson's $r = 0.85$, standard deviation = 0.11). GC3 content for all sets of four-fold degenerate codons are summed up to derive the GC4 content for each species. The GC4 content is strongly correlated with GC3 content of these eight amino acids across the species analyzed (average Pearson's $r = 0.96$, standard deviation = 0.036).

content, and 2) the selection for their effects at the nucleotide level (e.g. transcription and translation) that I collectively termed GC directional force.

I first computed a GC content measure minimally influenced by selection for amino acid change that is the GC content at the 3rd position of all four-fold degenerate codons (conventionally referred to as GC4) which nucleotide substitutions are synonymous. I then correlated tyrosine frequency with the computed GC4 content in each species to quantify how much the observed tyrosine depletion could have been passively driven by GC directional force on protein-coding regions that can directly affect amino acid changes. There are eight amino acids encoded by four-fold degenerate codons (Alanine, Arginine, Glycine, Leucine, Proline, Serine, Threonine and Valine). I found that the GC content at the 3rd position (conventionally referred to as GC3) for each of these amino acids showed strong pair-wise correlation with each other (average Pearson's R

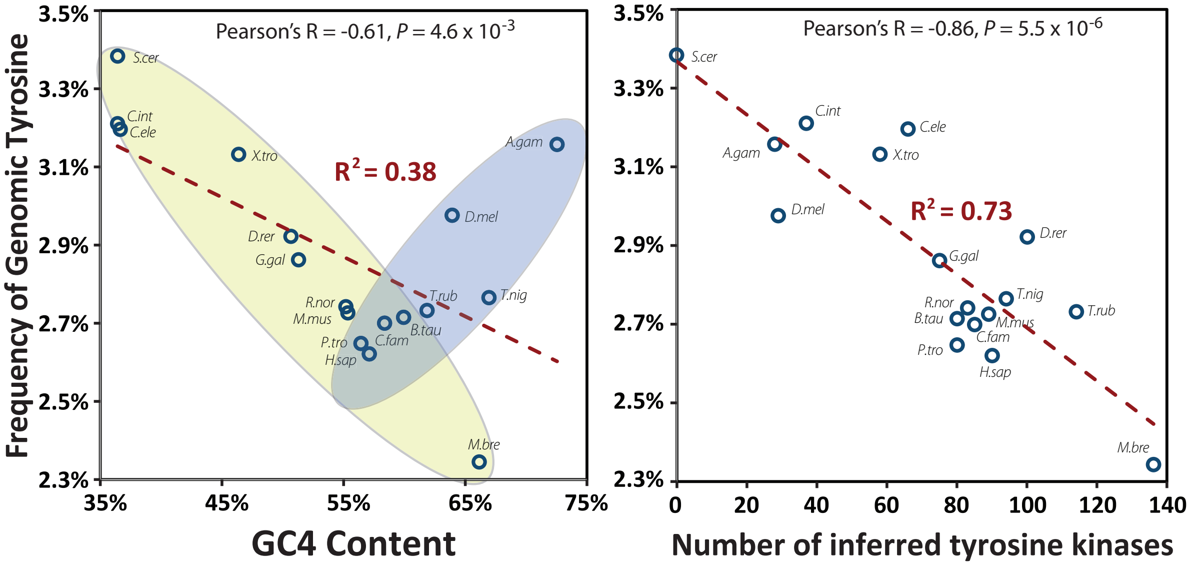

Figure 4.8: Correlation of genomic tyrosine frequency with GC4 content and number of predicted tyrosine kinases across multiple metazoan species, *S. cerevisiae* and *M. brevicollis* . The number of tyrosine kinase is predicted as previously described in *Material and Method* except for *M. brevicollis* which is based on [13].

= 0.85, std dev. = 0.11, see Figure 4.7) which indicate the presence of a global and uniform GC directional force acting on coding regions within each species. The species analyzed is the same set analyzed in the previous section but I added choanoflagellate *M. brevicollis* as it is the fully sequenced species currently known to have the most number of tyrosine kinases. As *M. brevicollis* is from an unicellular lineage that branched off before the metazoan lineage, including *M. brevicollis* in my analysis could also allow me to assess the generality of any observed correlation across multiple lineages. The computed GC4 content showed strong pairwise correlation with the GC3 content of each of the eight amino acids (average Pearson's $R = 0.96$, std dev. $= 0.036$). Thus, I conclude that computed GC4 content is a robust readout of global GC directional force that is minimally influenced by natural selection for amino acid change. I also computed GC content at all codon positions (conventionally referred to as GC123) for comparison.

I found that variation in GC4, GC123 and tyrosine kinase number can individually and maximally account for up to 37.6%, 56.2% and 73.4% of variation in tyrosine frequency respectively (Figure 4.8, R^2 analysis). I observed that the GC4 content of a subset of species does correlate negatively with tyrosine frequency (Figure 4.8, left plot, yellow ellipse) but the trend is reversed for many species (Figure 4.8, left plot, blue ellipse). *A. gambiae*, for example, has the highest GC4 content but the 4th highest tyrosine frequency. In contrast, the frequency of tyrosine co-varies more consistently with the number of tyrosine kinases (Figure 4.8, right plot). I note that this analysis cannot exclude the possibility that GC directional force might indeed be passively contributing to the observed tyrosine depletion in a subset of the species analyzed but, in general, the observed tyrosine frequency co-varies more consistently with the tyrosine kinase number than with GC directional force.

While I previously acknowledged that other mechanisms may have contributed to the observed tyrosine depletion, I have supported my proposal with empirical data that human proteins with no detectable phosphotyrosines (non-pTyr proteins) have lost considerably more tyrosines than known tyrosine-phosphorylated proteins (pTyr proteins) as compared to their *S. cerevisiae* orthologs. Here, I extended the analysis to *S. pombe* and reached the same conclusion (Figure 4.9). The two yeast species are compared because they are the known eukaryotes phylogenetically closest to human that lack a dedicated phosphotyrosine-signaling system, and we implicitly assume both species are informative of the ancestral tyrosine frequency.

In addition, I also performed similar analysis for phenylalanine and tryptophan as both amino acids are physicochemically similar to tyrosine. While I observed proportionally more tyrosine is depleted in human non-pTyr proteins than in known human pTyr proteins as compared to their *S. cerevisiae* ($P \approx 5.1 \times 10^{-9}$, Mann-Whitney test, Figure 4.9) and *S. pombe* orthologs ($P \approx 2.2 \times 10^{-9}$, Mann-Whitney test), neither phenylalanine nor tryptophan is lost preferentially in either protein groups (Figure 4.9). As phenylalanine and tyrosine are structurally identical except for a phosphorylatable hydroxyl group on tyrosine, and are likely subjected to a similar degree of GC directional

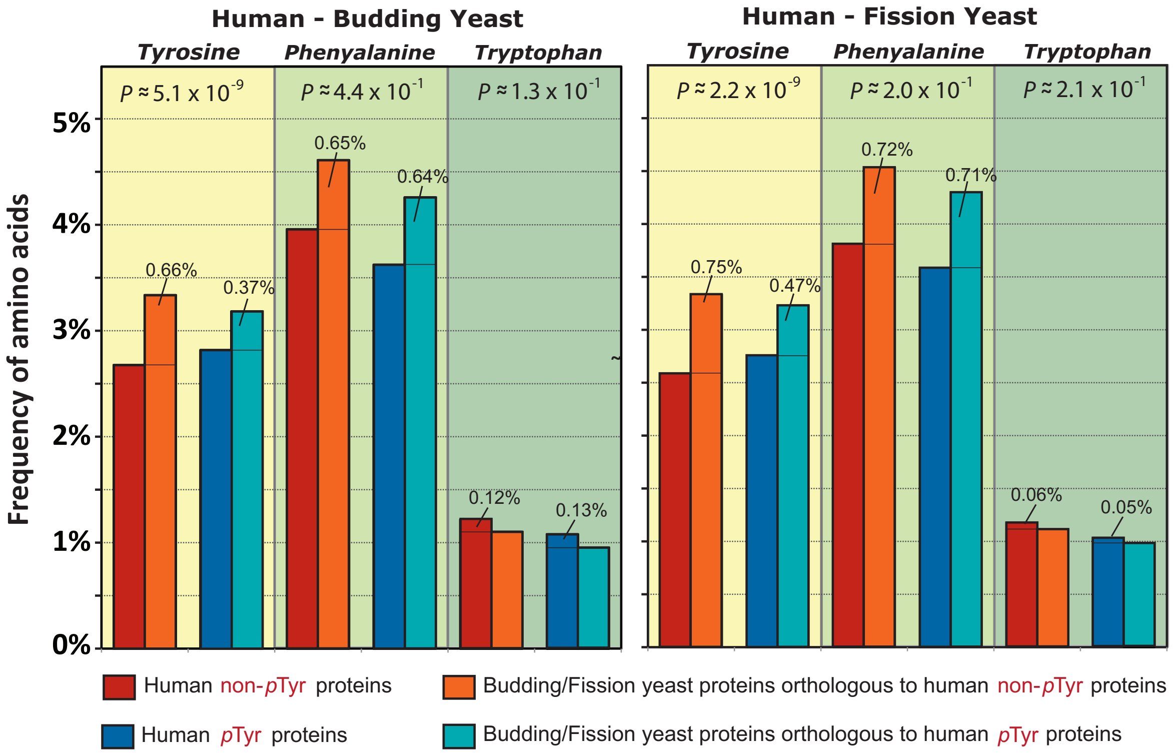

Figure 4.9: Frequency of tyrosine, phenylalanine and tryptophan in human proteins and their orthologues in *S. cerevisiae* and *S. pombe*. Human proteins are divided into known tyrosine-phosphorylated (pTyr Protein) and those not known to be tyrosine-phosphorylated (non-pTyr Protein). Only proteins with inferred one-to-one orthologous relationship are analyzed. Statistical significance indicated are for differences in the distribution of observed differences in amino acid frequency of human-yeast orthologous protein pairs between pTyr and Non-pTyr protein sets computed with Mann-Whitney test (one-tailed) using *R* statistical software.

force given both are encoded by two AT-rich codons each (Phe: TTT, TTC; Tyr: TAT, TAC), the observed preferential loss of tyrosine is due to the presence of its phosphorylatable hydroxyl group. This observation strongly supports my proposal that signaling fidelity is a driving force behind the observed tyrosine depletion. I included tryptophan in this analysis because tryptophan is physicochemically somewhat similar to phenylalanine and tyrosine.

During evolution, phenylalanine and tyrosine are commonly substituted by each other due to their similar physicochemical properties and encoding codons (see BLOSUM matrices, for example [58]). I investigated their substitution pattern from yeast to human.

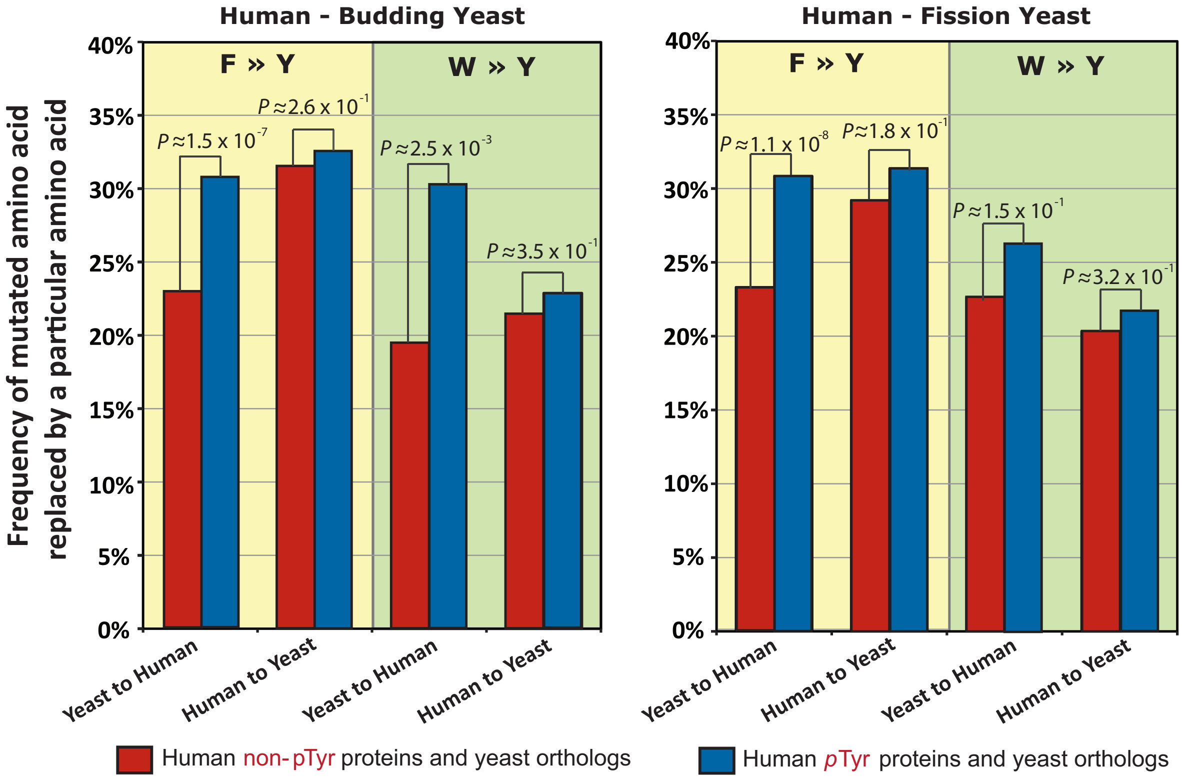

Figure 4.10: Observed substitution of phenylalanine and tryptophan by tyrosine between human and yeasts. Pairwise sequence alignments between orthologous protein pairs are performed using Mafft with default parameters [15]. To reduce error due from faulty alignments, residues with less than 5 identical aligned flanking residues out of 10 positions (5 on each side) are excluded. Statistical significance of observed differences are computed with Fisher's exact test (one-tailed) using *R* statistical software

As expected, tyrosines in yeast are most frequently substituted by phenylalanine (35%) in human, whereas phenylalanine and tryptophan in yeast are frequently substituted by tyrosine in human (Figure 4.10). We observed the substitution rate of phenylalanine by tyrosine from yeast to human is similar to the rate for human to yeast for *p*Tyr proteins. However, we observe that the substitution of phenylalanine and tryptophan in yeast by tyrosine in human is significantly under-represented in non-*p*Tyr proteins compared to *p*Tyr proteins ($P \approx 1.5 \times 10^{-7}$, Fisher's exact test, one-tailed, Figure 4.10). This phenomenon is minimally influenced by GC directional force as only T ↔ A substitution is required to directly switch between phenylalanine and tyrosine. We detected no statistical difference ($P < 0.01$, Fisher's exact test, two-tailed) in the substitution of tyrosine in yeast by phenylalanine and tryptophan in human between the two protein groups. However, the rate is significantly lower for non-pTyr from yeast to human. Thus, constrained substitution of phenylalanine and tryptophan by tyrosine is possibly another mechanism contributing to observed tyrosine depletion, and support my observation that there is selection pressure to remove tyrosines for signaling fidelity.

Next, for the set of tyrosines that are conserved between human and S. cerevisiae as identified from sequence alignments, we applied the NetPhorest algorithm [114] to investigate their propensity to be phosphorylated by human tyrosine kinases that are influenced by residues flanking the tyrosines on primary sequences. In general, we found that tyrosines from human are less phosphorylatable by human tyrosine kinases than the corresponding tyrosines from *S. cerevisiae* ($P \approx 5.3 \times 10^{-5}$, Wilcoxon Test, Figure 4.11). This suggests there is selection force(s) favoring mutations flanking tyrosines that reduce tyrosine phosphorylation.

4.4 Conclusion

Through investigating the maximum phosphorylation potential encoded in the proteomes of various metazoan species and budding yeast, I observed a strong negative correlation of tyrosine frequency with the number of cell type and the number of inferred tyrosine kinases found in each species. Thus, it seems that species that have more tyrosine kinases have proportionally less tyrosines encoded in their genome. I proposed an evolutionary model to explain the observed negative correlation of tyrosine frequency with tyrosine kinase number. Based on this model, I hypothesized that human proteins presently not known to be tyrosine-phosphorylated (non-pTyr) have experienced higher tyrosine loss during evolution than known tyrosine-phosphorylated human proteins (pTyr). To test this hypothesis, I compared members of these two groups to their orthologous proteins in *S. cerevisiae* and *S. pombe*, unicellular organisms with no known dedicated tyrosine

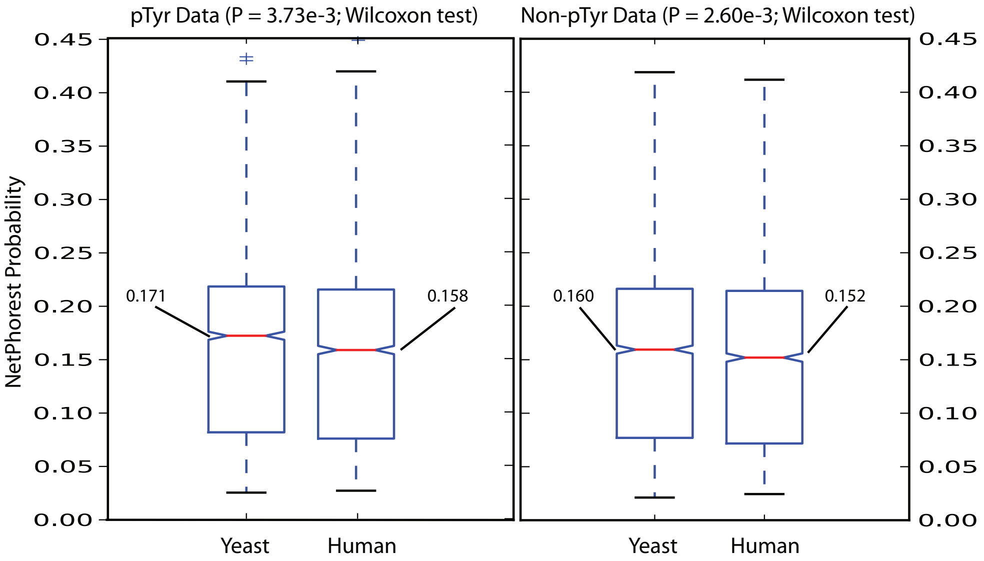

Figure 4.11: Lower phosphorylation propensity of tyrosines in human compared to *S. cerevisiae*. Tyrosines conserved between human and S. cerevisiae , as identified from pairwise sequence alignments, are tested for their phosphorylation propensity by human tyrosine kinases using the NetPhorest algorithm. Proteins with experimentally observed phospho-tyrosines (pTyr) are tested separately from proteins that have no experimentally observed phospho-tyrosines (non-pTyr). The median probabilities are labeled for each dataset, and the indicated p-values were calculated using the Wilcoxon test.

kinase, to assess the degree of tyrosine loss that may be triggered by the onset of phosphotyrosine signaling in metazoans. I observed proportionally less tyrosines in human proteins over orthologous proteins in yeast, but the observed tyrosine depletion is statistically more prominent in non-pTyr over pTyr human proteins which supports our hypothesis. I also observed preferential loss of tyrosine over phenylalanine in non-pTyr proteins than pTyr proteins and amino acid substitution flanking tyrosine in non-pTyr proteins that generally disfavor tyrosine phosphorylation.

A factor that could contribute to the observed tyrosine depletion is that the amino acid cannot be synthesized *de novo* in mammals; it is synthesized from phenylalanine which is obtained from the diet. In bacteria, fungi and plants, the three amino acids phenylalanine, tyrosine and tryptophan are synthesized *de novo* from chorismate produced by the shikimate pathway [144]. Hence, tyrosine and phenylalanine are essential in most, if not all, of the metazoan species analyzed but are not essential in *S. cerevisiae* and *S. pombe.* This difference in essential amino acids may play a part in the tyrosine depletion observed between human and the two yeast species. However, it cannot readily explain the differences in tyrosine frequency within the metazoan species analyzed particularly for the more closely related species. It may be plausible that more complex organisms, as assessed based on the number of unique cell types, evolved to reduce reliance on essential amino acids. However, this is not likely the case as tryptophan frequency, another essential amino acid, positively correlates with the number of unique cell types among the species analyzed in this work. In addition, it also cannot explain the difference in tyrosine depletion observed between textitpTyr and non-*p*Tyr human proteins.

As the observed tyrosine depletion might have been passively driven by biased A/T → G/C nucleotide substitution in the absence of natural selection for amino acid change, I analyzed the GC4 content of coding DNA sequences to quantify how much of the observed tyrosine depletion could have been contributed by passive A/T → G/C nucleotide substitution. We observed a weaker correlation of GC4 content with tyrosine frequency than with tyrosine kinase frequency, indicating that the observed tyrosine depletion is not predominantly caused by passive A/T → G/C nucleotide substitution. As genetic mutations is required to generate phenotypic variation for selection forces to act upon, biased A/T → G/C nucleotide substitution during metazoan evolution likely facilitate tyrosine depletion needed for the expansion of tyrosine kinase and the optimization of phosphotyrosine signaling. Our tyrosine-phenylalanine substitution analysis also suggests that the presence of tyrosine kinases could impose constraints on the appearance of new tyrosines that favor change in tyrosine frequency in one direction (like a ratchet that turns in one direction).

Taken together, our findings reveal there is a positive selection of tyrosine loss in metazoan to reduce deleterious tyrosine phosphorylation. At a higher lever, the observation suggests that phosphotyrosine signaling by dedicated tyrosine kinases, as a biological innovation that probably assisted in the development of multi-cellular animals, required system-level adaptive mutations. This phenomenon highlights a general principle of adaptive evolution pertaining to the introduction of new components into a complex system and serves as an important framework when considering the evolution of complex biological systems. In addition, this revelation provides additional insight into why tyrosine phosphorylation is a relatively rare event in vivo compared to serine phosphorylation. Furthermore, it can account for, at least partially, why human tyrosine residues, whether phosphorylated or not, are generally very conserved in other species, as the appearance of new tyrosines is expected to be suppressed during evolution in metazoa.

Other factors, such as tyrosine sulfation, could have contributed to the observed tyrosine depletion, which raises the question of whether other post-translational modifications and regulatory mechanisms are under similar evolutionary selection that could also explain genomic GC conversion. The numbers of genomically-encoded threonines showed strong negative correlations with serine/threonine kinases and cell-type numbers, although these trends were not observed with serines (Figure 4.1 and 4.3.

Chapter 5

Summary and future directions

Perspectives presented in Section 5.2.2 and 5.2.3 were published in:

C. S. Tan, C. Jørgensen, R. Linding. Roles of junk phosphorylation in modulating biomolecular association of phosphorylated proteins? *Cell Cycle*, 9(7):1276-80, 2010 Apr.

As part of the future work described in Section 5.2.7, I initiated and worked with Xiaojian Shao from the Bader lab to develop computational methods for inferring interaction changes between protein domains and their peptide ligands given their primary sequences, with the intention to extend it to protein kinases and phosphoresidue-binding domains. A pilot project toward this goal was published in:

X. Shao*, **C. S. Tan***, C. Voss, S. S. Li, N. Deng, G. D. Bader. A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain-peptide interaction from primary sequence. *Bioinformatics*, 27(3):383-90, 2011 Feb.

* denotes co-first authors

5.1 Summary

Protein phosphorylation is a prevalent reversible post-translational modification that influences structural conformation, enzymatic activities, molecular association and subcellular localization of proteins. In eukaryotic cells, protein kinases transfer a phosphate group from ATP to the side-chain hydroxyl group of a specific set of serine, threonine and tyrosine residues in the proteome. Identifying these phosphorylated residues and the protein kinases that targeted them are crucial for understanding the dynamic regulation of cellular activities by protein phosphorylation. Phosphoproteomic technologies allow proteome-wide quantitative detection of proteins and residues phosphorylated under different physiological conditions [110, 173, 195, 79], and have been applied to unveil the phosphoproteomes of several model organisms [16, 203]. The functional consequence of the majority of these phosphorylation events are unknown; this calls for endeavors to characterize their molecular functions and effects on cellular decision processes. Systematic approaches to categorize phosphorylation events, pinpoint their potential functions and generate testable hypothesis will help prioritize phosphorylation sites for experimental characterization. For my research, I analyzed the evolutionary dynamics of protein phosphorylation sites and assessed the utility of conservation analysis to interpret functionally important phosphorylation events.

5.1.1 Comparative phosphoproteomics

Phosphorylation events that are conserved across orthologous proteins, especially those that are across distantly related species, are arguably under evolutionary constraint due to their involvement in fundamental cellular processes. Based on this assumption, I identified phosphorylation events on human proteins that are conserved at similar sequence positions on orthologous proteins from fly, worm and yeast (my query species). Phosphorylation sites on human proteins were assembled from two major phosphorylation site databases (PhosphoSite and PhosphoELM) while phosphorylation sites on fly, worm and yeast proteins were obtained by our collaborators using untargeted MS-based phosphoproteomic screens. In total, I obtained around 24,000 human phosphorylation sites and around 22,000 sites from my query species but managed to identify only 479 phosphorylation events on 344 human proteins that are positionally conserved on orthologous proteins from at least one query species. For ease of communication, the 344 human proteins are termed core site proteins.

About 45% of the human phosphorylation sites assembled occurred on proteins that have no detectable ortholog in our query species as based on Ensembl's ortholog inference algorithm. Another 15% of the assembled phosphorylation sites occurred on human proteins which putative orthologous sequences in the query species are not found phosphorylated in our MS-based phosphoproteomic screens. The remaining 40% of the assembled phosphorylation sites are found on human proteins which orthologs are phosphorylated in at least one query species. About half of these phosphorylated residues ($\approx$ 19% globally) are aligned to phosphorylatable residues in our query species of which less than 14% are found phosphorylated in our MS-based screens; the last number suggests that our assembled phosphorylation data for the query species are likely not comprehensive. Hence, more positionally conserved phosphorylation sites will be uncovered when more phosphorylation sites in query species are identified.

We are aware of cases where a protein kinase regulates protein function in ways that the precise location of phosphorylation on a substrate is not crucial, but targeting of the substrate by the protein kinase is nevertheless evolutionarily conserved. I identified 778 phosphorylation events on 698 human proteins that are potentially conserved in such a manner in my query species. For ease of communication, the 698 human proteins are termed core net proteins. I found both core site and core net protein sets were enriched in proteins encoded by disease-associated genes. Unlike what I expected, I found that human phosphoproteins with an identified phosphorylated ortholog in any query species are not enriched in proteins encoded by disease-associated genes. This may be because there are many spurious phosphorylation sites in the assembled data sets. An interesting finding is human proteins with more phosphorylation sites are more frequently encoded by disease-associated genes and this phenomenon is observed for evolutionary conserved phosphorylation hubs identified in my study.

I found that cytosolic ribosomal proteins and proteins involved in amino acid phosphorylation and RNA splicing are over-represented in the core site protein set. We speculated that some phosphorylation events are conserved across orthologous proteins at the same position to induce conformational changes that probably required precisely coupled interactions between added phosphate and surrounding residues. This is evident in the set of positionally conserved phosphorylation events observed on the activation loops of protein kinases that are known to induce conformational changes on the loops. On the other hand, the core net protein set is over-represented in proteins associated with the cell cycle, chromosome organization and biogenesis, DNA-dependent regulation of transcription, macromolecular complex assembly and protein targeting. In particular, core net proteins are also statistically enriched with protein- and DNA-binding annotation compared to the superset of human phosphoproteins with phosphorylated orthologs in query species. This suggests that many non-positionally conserved phosphorylation events may be regulating the molecular association of phosphorylated proteins with DNA and other proteins.

5.1.2 Factors affecting conservation of phosphorylated residues

For the next part of my research, I performed sequence conservation analysis of human phosphorylated residues across vertebrates that have genomes fully sequenced. I also performed similar analysis across various fungal species for phosphorylation sites identified in *S. cerevisiae*. A measure is derived to quantify how conserved are phosphorylated residues compared to randomly selected phosphorylatable residues in the same set of proteins. This measure accounts for the intrinsic conservation rate of proteins and phosphorylatable residues minimally influenced by protein phosphorylation. I further grouped phosphorylated residues based on some meaningful features for such analysis in an attempt to delineate factors that could influence the conservation of phosphorylated residues, and conversely how these factors affect the interpretation of functional sites through sequence conservation analysis.

I found that human phosphorylated residues detected by multiple HTP studies are overall more conserved than those detected by one HTP study. This trend is also observed for phosphorylated residues in *S. cerevisiae* across closely related yeast species. This might be because frequently detected phosphorylation sites are less likely to be stochastic phosphorylation events or less likely to be falsely identified by peptide identification or site localization computer algorithms used to interpret MS spectra. It is also likely that many frequently detected phosphorylation sites are highly conserved because they are targeted by multiple kinases and hence the phosphorylated residues are under stronger functional constraint. Unlike what I expected, I found that supposedly high stoichiometry sites observed in HeLa cells during mitosis are not more conserved than lower stoichiometry sites. One explanation is that the method used to quantify site stoichiometry is inadequate. Another plausible explanation is that many high stoichiometry phosphorylation sites observed in HeLa cells are by phosphorylated dysfunctional protein kinases that do not occur typically in normal cells.

My conservation analysis suggests that a larger percentage of the assembled HTP phosphorylation sites in structured protein regions are spurious than the set of phosphorylation sites in intrinsic disordered protein regions. One possible explanation for this is that the experimental procedures or conditions employed in some HTP studies might have unnaturally exposed phosphorylatable residues in the protein interior to protein kinase, such as when proteins are denatured or broken down into peptides. For examples, denaturing reagents like urea and sodium dodecylsulfate (SDS) are commonly used in protein assays. Dithiothreitol (DTT) is frequently used in lysis and restriction enzyme buffers to reduce oxidation of a protein sample, and preserve the activities of enzymatic proteins. In proteomic analysis by mass spectrometry, DTT is often used to reduce disulfide bonds in protein for better protein fragmentation. Since the chemical reduces disulfide bonds, it can disrupt the tertiary structure of protein thereby exposing typically surface inaccessible residues unnaturally to protein kinases. The effects of denaturing reagents and DTT on detecting non-natural phosphorylated residues could be systematically investigated by varying their concentration in experiment protocols and analyzing the resulting relative divergence rates of phosphorylated residues detected. If the relative divergence rate of phosphorylated residues detected is lower (i.e. more conserved) when less denaturing reagents and/or DTT are used, then it is definitely plausible that many surface inaccessible phosphorylated residues had been unnaturally exposed to protein kinases in some phosphoproteomic experiments.

A very important finding, in my opinion, is that background phosphorylatable residues are as conserved as their phosphorylated residues on the most abundant yeast proteins. This observation implies that without normalizing for the intrinsic conservation rate of phosphorylatable residues on high abundant proteins, phosphorylated residues from these proteins will be interpreted as highly conserved but this can be because the sequences of high abundant proteins are overall highly conserved. This observation also suggests that stochastic phosphorylation events on highly abundant proteins are more readily detected by mass spectrometers because more randomly phosphorylated peptides from these proteins are available for detection. A caveat in this argument is that functionally important phosphorylation sites on highly abundant proteins are also expected to be more frequently detected unless there are many more random phosphorylation sites than genuine functional sites on high abundant proteins.

Another very interesting finding is that HTP phosphorylated residues detected in mouse tissues are as conserved as functionally characterized human phosphorylated residues while the HTP phosphorylated residues from human cell lines and from unicellular *S. cerevisiae* are less conserved. This observation suggests kinase activity is more tightly regulated in normal cells of multicellular organisms than in free-living single cells of unicellular organism or cell line. Correspondingly, this implies that phosphorylation sites detected in unicellular organisms or from cell lines are more likely to be spurious than those detected from normal cells in multicellular organisms. Another possibility is that the experimental procedures and site identification analysis employed for the mouse tissue project produce less spurious phosphorylation sites. One way to assess whether more spurious phosphorylation occurred in a unicellular context is to apply the same site-detection protocols and mass spectrometric analysis on fly (or its embryo) and fly cell lines. The relative divergence rate of phosphorylation sites detected from the different samples can then be compared. Also, mutant kinases such fusion and constitutively active kinases could be introduced into cells to assess whether newly appeared phosphorylation sites are less conserved than sites identified before the introduction.

5.1.3 Selection against spurious tyrosine phosphorylation in metazoan

For the last part of my research, I computed the frequencies of phosphorylatable residues (serine, threonine and tyrosine seperately) in *S. cerevisiae* and across different metazoan species to determine the theoretical maximum phosphorylation capacity of each species. I then analyzed the relationship between the theoretical maximum phosphorylation capacity and the number of unique cell types of each species as a proxy for organismal complexity. An interesting result is observed for the theoretical maximum capacity of tyrosine phosphorylation. I expected higher tyrosine frequency in proteomes of species with more cell types and and with more genes encoding tyrosine kinase (TK), reasoning that a greater portion of tyrosine are phosphorylated and hence under evolutionary constraint in these species. However, a negative correlation of tyrosine frequency with the number of cell type and the number of tyrosine kinase is observed. This does not imply that functional phosphotyrosines are not under evolutionary constraint but suggests that selection against phosphotyrosine is stronger or more pervasive than the selection for functional phosphotyrosines. I proposed that the observed depletion of tyrosine helps to abate deleterious phosphotyrosine that could otherwise be brought about by the expansion of TK-encoding genes observed in metazoan species. An experimental support of this hypothesis is that the expression of Src tyrosine kinase in the unicellular *S. cerevisiae* and *S. pombe*, whose genomes encode no tyrosine kinase, is toxic due to aberrant tyrosine phosphorylation [176, 177]. Hence, in addition to tyrosine phosphatase and the spatio-temporal and contextual regulation of tyrosine kinases, deleterious tyrosine phosphorylation is probably also minimized by the observed tyrosine depletion in metazoan proteins. Although the negative correlation of tyrosine frequency with cell type observed for metazoan species broke down in unicellular choanoflagellate *M brevicollis*, my proposed TK-tyrosine evolutionary model is nevertheless consistent; the species has the lowest tyrosine frequency and largest number of gene encoding tyrosine kinase among the species analyzed.

I also investigated how much A/T → G/C nucleotide substitution observed in some metazoan species could have contributed to the observed tyrosine depletion in the absence of selection pressure. Based on the species analyzed, I found that GC content variation, in the absence of selection for amino acid changes, can only maximally account for 38% of the observed tyrosine frequency variation while tyrosine kinase content variation can account up to 73% of the observed tyrosine frequency variation. In addition to biased A/T → G/C nucleotide substitution that can facilitate tyrosine depletion, I also identified other mechanisms that can reduce tyrosine phosphorylation. This includes the restrained substitution of phenylalanine by tyrosine that is independent of biased A/T → G/C nucleotide substitution, and the removal of phosphorylation-promoting residues flanking tyrosines on primary sequence. It has been reported that proteins can evolve to disfavor interactions [145, 202, 1, 101]. The observed negative selection of tyrosine phosphorylation suggests that selection against promiscuous interaction of a specific protein or a family of proteins can occur at the system level involving many other proteins. Conversely, this implies that natural selection acted to reduce promiscuous interaction as much as to promote specific interaction, if not more, and this can occur *in cis* on an interacting protein or *in trans* on other proteins. My observation also suggests genome-wide adaptive evolution may be required to optimize an initial overall beneficial genetic perturbation such as the expansion of tyrosine kinase.

5.2 Perspectives and Future Directions

5.2.1 False positive, technical and biological noise in existing phosphorylation site data

Current phosphorylation datasets likely contain many spurious phosphorylation sites falsely identified by computer algorithms that identify peptide sequence and infer phosphorylation site from MS/MS spectra. This problem is likely to abate in the future with better spectrum-analysis computer algorithms. If source spectra of existing MS-identified phosphorylation sites are available, new and improved spectrum-analysis computer algorithms can be redeployed on them to filter out potentially false sites. In addition, continued improvement of mass spectrometers with higher mass accuracy should generate better quality spectra that facilitate better peptide identification and site localization.

On other hand, genuine but random *in vivo* phosphorylation events can be increasingly detected in the future MS-based phosphoproteomic studies with more sensitive MS instrumentation. It can be challenging to differentiate such phosphorylation events from deterministic and functional ones. These two types of phosphorylation events are hard to define precisely themselves. Here, I define a genuine *in vivo* phosphorylation event to be stochastic if it occurred inconsistently and only in the minority of cells at similar physiological state. This definition itself is hard to assess as it requires single-cell assays which is presently not possible for many phosphorylation sites and proteins. I consider this type of phosphorylation event as biological noise and parallel intrinsic noise observed in gene expression. It is, however, presently feeble without more data to conclude that such biological phosphorylation noise does not have any effect on protein function or cellular activities at large. It is plausible that noisy phosphorylation facilitates phenotypic variation which can increase the probability that some individuals of a species possess the appropriate variation to survive or strive in a new environment. This characteristic can be very beneficial for unicellular organisms and cancerous cells which have limited control over their environment and are experiencing frequent environmental changes.

In my analysis, I found that HTP phosphorylation sites on inferred structured regions are not more conserved than randomly selected phosphorylatable residues from similar protein regions. However, conservation is only observed if I restrict the HTP phosphorylation sites to those detected in multiple studies. An explanation that I conceived is that buried phosphorylatable residues that are typically not accessible to protein kinases can become so when proteins are denatured or broken down under synthetic conditions. Future works could involve identifying phosphorylation residues that are surface inaccessible based either on known protein structures or using surface accessible prediction software, and assess whether they are conserved at similar rate with other phosphorylatable residues from the same protein regions. If this is indeed so, it implies that there is a possibility that some phosphorylation sites in existing dataset are genuine phosphorylation events which arise from inappropriate technical treatment. I termed such phosphorylation events as technical noise. However, such noisy phosphorylation events detected in structurally ordered protein regions, if existed, likely make up only a small portion of currently known phosphorylation events as only around 20% of known phosphorylation sites occurred in structured protein regions. Nevertheless, better sample preparation and appropriate experimental treatments will certainly help minimize technical phosphorylation noise.

For the purpose of studying cellular processes regulated by protein kinases and phosphatases, differentiating genuine deterministic phosphorylation events from technical and biological noises is crucial. It is plausible that some genuine and deterministic phosphorylation events may have no effect on cellular processes. Presumably, phosphorylated residues from such phosphorylation events will be less conserved compared to deterministic functional phosphorylation events. However, my conservation analysis of phosphorylated residues on proteins in *S. cerevisiae* reveals that using sequence conservation to interpret functional important sites can be limited for sites on high abundant proteins. This is because sequences of high abundant proteins are generally very conserved. It is, therefore, crucial to normalize the observed conservation rate of phosphorylated residues with that of other phosphorylatable residues from the same proteins when using sequence conservation to identify functionally important sites.

I found that HTP phosphorylated residues from mouse tissues are as conserved as known functional phosphorylated residues in human while HTP phosphorylated residues detected in human cell lines and in unicellular yeast are less conserved overall. I speculate that biological phosphorylation noise is more tolerated in unicellular organisms and in cancerous cells than in the normal cells of multicellular organisms. This also means that protein phosphorylation is more tightly regulated in the cells of multicellular organisms. Many human cell lines, the cancerous ones in particular, are known to have dysfunctional kinases. Hence, it is not unexpected that there are more biological phosphorylation noises in cancerous cells than in the physiologically normal cells of multicellular organisms.

5.2.2 Phosphorylation of non-conserved residues can be functional

It has been estimated that as much as 65% of HTP phosphorylation sites are non-functional based on the difference in position-dependent conservation between HTP phosphorylated serines/threonines and randomly selected serines/threonines [90]. The underlying basis is phosphorylation of non-conserved residues is non-functional. This is based on the assumption that protein kinases have to regulate protein functions through position-specific (on protein) phosphorylation. I argue, based on known cases of how protein kinases regulate cellular activities and the observation from my comparative phosphoproteomic work, that phosphorylation of non-conserved residues can have functional consequences, and advocate for complementary approaches to identify evolutionary conserved phosphorylation events.

Undoubtedly, a portion of the non-conserved phosphorylation sites are spuriously localized particularly for MS-identified phosphopeptides that have multiple phosphorylatable residues, despite that this is somewhat proclaimed to be statistically controlled in MS experiments. In addition, the sequences of some phosphopeptides might have been falsely identified. Furthermore, the phosphopeptide enrichment techniques conventionally applied in MS-based phosphoproteomic studies might have identified random low stoichiometry phosphorylation sites that may or may not have cellular effect. Otherwise, phosphorylation of non-conserved residues could be implicated in lineage- or species-specific cellular functions such as phosphorylation sites involved in cell-cell communication that are not expected to be conserved in unicellular organisms. Examples of species-specific phosphorylation site are the CDK phosphorylation sites on the Mcm3 protein that were gained in *S. cerevisiae* lineage after divergence from *C. albicans*. These phosphorylation sites are involved in mediating nuclear export of the MCM complex that is unique to *S. cerevisiae* [120, 97].

Identifying sites that only appear in a particular lineage and are selectively retained among its species can give insight to the unique cellular activities or development pertaining to that particular lineage. However, a phosphorylatable residue can appear conserved among species of a relative new lineage simply because insufficient divergence time has lapsed for a mutation to occur at that site. Thus, more species need to be sampled for lineages with short divergence times since the last common ancestor. Monitoring the dynamics of phosphorylation sites (increased- or decreased-phosphorylation) under different physiological conditions or stimulus is an alternative to evolutionary approaches for interpreting the importance of lineage- or species-specific phosphorylation sites. Importantly, lineage- or species-specific sites can be falsely identified easily if they lie in intrinsic disordered regions that, in general, are fast evolving [23, 100, 180], and hence, easily missed by multiple-sequence-alignment (MSA) algorithms that had been optimized for conserved globular domains [134]. An archtype example of a functionally important but seemingly lineage specific phosphorylation site is Ser46 on the tumor suppressor human p53 which has been implicated in regulating apoptosis, cell growth and transcription by numerous studies. Yes, the phosphorylated residue was not conserved in mouse p53 based on sequence alignment [26]. However, non-alignment-based computational analysis, biochemical and functional assays suggest that Ser58 in mouse p53 is functionally equivalent to Ser46 in human p53 [26]. Hence, the development of specialized alignment algorithms, alternative computational approaches [26, 180] and benchmarking datasets [134] will be crucial for minimizing identification of spurious non-conserved phosphorylation sites.

5.2.3 Phosphoregulation of protein interaction can be position-independent

An emerging view is that protein kinases can serve to fine-tune the bulk electrostatic charge of targeted proteins through phosphorylation. Such effects may inhibit the physical association of phosphorylated proteins with other negatively charged biomolecules such as phospholipid membranes [174] and polynucleotides [209] by electrostatic interference/repulsion [112, 161]. This mechanism does not require phosphorylation to occur at precise location but a general region on substrates to create negatively charged protein surfaces to modulate molecular association [161]. Protein phosphorylation is also known to instigate protein-protein interaction by bulk electrostatics as observed for Sic1 in *S. cerevisiae* which needs to be phosphorylated on any six out of the nine poorly conserved CDK1 phosphorylation sites for binding to a single binding site on Cdc24 component of SCF ubiquitin ligase [122]. In addition, if a phosphorylated residue on a protein interacting interface serves to attenuate protein-protein interaction through steric or electrostatic interference, phosphorylation of other serine, threonine or tyrosine on the same protein interacting interface presumably will have similar effect.

It is well established that a large portion of protein phosphorylation events dynamically promote protein-protein interaction by creating temporal binding sites for phosphoresidue-binding protein domains such as SH2, PTB, 14-3-3, WD40 and FHA for which the phosphorylated residues need not be precisely located on the substrates. Many linear motifs bound by modular interaction domains, which include phosphoresidue-binding domains, need not be conserved at specific positions but a general region across orthologous proteins [124, 48] like those involved in mediating protein subcellular localization. In particular, co-operative binding has been observed where multiple linear motifs on a protein have additive effects on binding [48, 63]. Similar effects have been observed for interactions modulated by protein phosphorylation through bulk electrostatic charge in disordered regions where each additional phosphorylation site progressively decrease or increase molecular associations [19].

The presence of functionally redundant phosphorylation sites for attenuating or promoting physical interaction of phosphorylated proteins implies some can be lost during evolution in some lineages with minimum functional consequences, and can contribute to phenotypic diversity observed in a population [11]. This can provide the evolutionary plasticity needed to fine-tune conserved cellular activities for unique developmental and physiological needs of individual species. Given that the majority of known phosphorylation sites are located in intrinsically disordered protein regions, it is likely that a portion of these sites is regulating biomolecular associations of proteins [111, 62] through electrostatic interference. In addition, non-conserved phosphorylation sites can serve as decoys to buffer against spurious phosphorylation of other sites. For example, Wee1 in *Xenopus* contains a set of poorly conserved CDK1 (Cyclin-dependent kinase 1) sites that soak up stochastic activation of CDK1 [85]. These CDK sites vary in number and location in Wee1 across different species but nevertheless are important for the timely inactivation of Wee1 during mitosis [85]. Interestingly, many proteins are multi-phosphorylated by CDK on residues located close to each other on primary sequences, and this has been exploited to improve prediction of *bona fide* CDK substrates [119, 27]. Hence, evolutionary modeling and analysis of phosphorylated residues that incorporate these scenarios can minimize spurious identification of non-functional sites based on sequence conservation analysis.

5.2.4 Protein kinases can regulate cellular activities at different molecular levels

Other than affecting individual proteins, protein kinases can target higher order molecular machineries at the level of protein complexes to regulate cellular activities [73, 33]. In the case where protein phosphorylation serves to disable the activities of a protein complex by targeting its subunits for ubiquitination and subsequently degradation, phosphorylation of any subunit is presumably sufficient for the purpose, and hence need not be evolutionary conserved on orthologous protein [33]. Similarly, if phosphorylation of a residue on a protein interacting interface serves to attenuate protein complex assembly, phosphorylation at interacting interface on any subunit presumably will have similar effect. Analysis of gene expression had revealed that periodically expressed and constitutively expressed subunits in evolutionary conserved cell cycle protein complexes differ among the four species. Interestingly, combined experimental and computational data analysis revealed that the periodically expressed subunits are preferentially phosphorylated compared to constitutively expressed subunits in each species [73] although the periodically expressed subunits differ in each species. This observation indicates dynamic interplay between gene expression and protein phosphorylation for regulating cellular activities, and at a higher level, suggests that changes in another regulatory mechanism can relax the evolutionary constraint on a functional site. In support, a recent comparative analysis of phosphoproteomes across three yeast species (*S. cerevisae*, *C. albicans* and *S. pombe*) revealed that the intensity of phosphorylation is highly conserved among different cellular activities although the intensity may vary considerably among individual proteins within each functional group across the three species. This indicates prevalent global switching of proteins targeted by kinases during evolution [11]. Future works can assess whether some and what kind of protein complexes are preferentially targeted by orthologous kinases but on different subunits across different species. Novel computational methods can then be developed to identify phosphorylation events conserved in such a manner.

5.2.5 Finding colocalizing phosphorylation sites and protein binding sites

It has been observed that many protein phosphorylation hubs are frequently also protein interaction hubs [184]. These hubs are enriched in intrinsic disordered regions that contain many currently known phosphorylation sites and potential binding sites of many peptide-binding protein domains like SH3, WW and PDZ[184]. The colocalization of protein phosphorylation sites and the binding sites of peptide-binding domain presumably permit coregulation of both interactions. Hence, I postulate phosphorylation of some proteins serves to attenuate their physical interaction with peptide-binding proteins by blocking or interfering with binding sites on the phosphorylated proteins. Future work can be directed to identify co-occurring phosphorylation sites and putative binding sites of peptide-binding protein domains in intrinsic disordered protein regions and to test these cases experimentally to assess the correctness of the prediction. If this is indeed so to some extent, it can then be tested whether predicted binding sites of peptide-binding protein domains located near *in vivo* phosphrylation sites are more frequently genuine. Evolutionarily conserved co-localization of predicted protein interaction site and phosphorylation site near each other across orthologous proteins might then be identified to help differentiate *bona fide* binding sites of peptide-binding protein domains from spurious ones.

5.2.6 Multivariate modeling to identify deterministic phosphorylation events

In Chapter 3, I observed that more frequently detected phosphorylation sites are under stronger evolutionary constraints presumably because they are less likely to be random phosphorylation events and/or are less likely to be spuriously identified by peptide identification and site localization algorithms. This observation suggests that site detection frequency can be used to filter phosphorylation data for deterministic phosphorylation events. However, this simple approach is likely to miss many deterministic phosphorylation events that have not been detected frequently enough.

A somewhat related observation from chapter 3 is that phosphorylated residues detected on higher abundance proteins are overall less conserved than the background phosphorylatable residues from the same protein set. I speculate that this is partially because randomly phosphorylated residues on high abundance proteins are more readily detectable by MS-based analysis. This also implies that randomly phosphorylated residues on high abundance proteins can be more frequently detected than randomly phosphorylated residues on low abundance proteins. In other words, phosphorylation sites detected on high abundance proteins are more likely to be random phosphorylation events than phosphorylation sites of same detection frequency detected on low abundance proteins. This is supported by the observation that, for the set of sites that have been detected in *S. cerevisiae* by multiple HTP studies, phosphorylated residues on low abundance proteins are much more conserved than phosphorylated residues detected on high abundance proteins (see Figure 3.9). This may partially explain why only the most frequently detected human phosphorylation sites in HTP studies have comparable sequence conservation rates to known functional phosphorylation sites in human (see Figure 3.3 and Figure 3.2). This suggests that the site frequency filtering approach can be biased against deterministic phosphorylation events detected on low abundance proteins and will increasingly include randomly phosphorylated residues on more abundant proteins. Hence, for the purpose of identifying deterministic phosphorylation sites, it will be useful to statistically model the effects of protein abundance on site detection frequency and incorporate the effects in site frequency filtering approaches. However, this will require prior knowledge on protein abundance which may not be available. One way to bypass this restriction is to use relative protein quantification measures derived from MS analysis, like spectra count of peptides, as a proxy for protein abundance. Using spectra count of peptides derived in MS analysis has the additional advantage of incorporating known and unknown intrinsic factors along MS-based proteomic protocols that influence peptide identification and quantification which presumably also affects the identification of phosphopeptides and phosphorylation sites. A more sophisticated approach is to incorporate effects of protein abundance, spectral count, site frequency and peptide identification score through multivariate modeling, such as multivariate regression analysis and artificial neural networks, to better filter spurious phosphorylation sites and random phosphorylation events. Conservation and phylogenetic analysis of phosphorylated residues can be subsequently applied to differentiate sites likely involved in fundamental cellular activities from those involved in unique biology of species studied.

5.2.7 Identifying phosphorylation motifs under positive or negative selection

False positive sites can be removed with better peptide identification and site localization algorithms, and random phosphorylation events can be filtered out by approaches outlined in the previous section. However, it may be useful for practical reasons to identify evolutionarily conserved phosphorylation events presumably implicated in some conserved cellular activities. In some cases, it may be useful to identify phosphorylation sites that contribute to divergence of protein function. Hence, another potential future direction is to develop computational or statistical approaches to identify phosphorylation sites that are under positive or negative selection. One way is to assess the conservation of observed motifs flanking phosphorylated residues. The motif assessed may be that of some kinase protein (or kinase family) or phosphoresidue-binding domain (or domain family). Phosphorylation sites with such motifs presumably are more likely to be functional. Another avenue is to develop computational tools to predict how observed mutation of residues flanking phosphorylated residues affect phosphorylation kinetics and binding affinity. There is likely sufficient data available now, with more expected in the near future, on specificity of protein kinases and phosphoresidue-binding protein domains (from large-scale phage display, protein/peptide array and fluorescence polarization experiments) for this purpose. Conversely, the same tools can be deployed to identify phosphorylation and binding motifs that are under negative selection either at the proteome-level or in a subset of proteins. Intuitively, if the motif of a kinase (or kinase family) or a phosphoresidue-binding domain (or domain family) are overall under negative selection, proteins that expressed motifs that evolved against this trend could be more likely *bona fide* targets of the kinase or phosphoresidue-binding domain.

5.3 Conclusion

Identifying functionally important phosphorylation events and inferring their molecular effects help prioritize phosphorylation sites for experimental studies. Furthermore, current phosphorylation data likely contain some technical and biological noise as well as non-existent sites spuriously identified by spectrum-analysis algorithms. I demonstrated that comparing phosphoproteomics of distantly related species can identify evolutionarily conserved human phosphoproteins that are notably associated with cancer and other genetic diseases. Sequence conservation of phosphorylated residues, especially when across distantly relates species, can complement comparative phosphoproteomics approaches to identify more functionally important phosphorylation events. However, identifying definite non-functional phosphorylation events based on non-conserved phosphorylated residues at similar positions across orthologous proteins can be challenging. This is because some molecular effects of protein phosphorylation are less position-specific such as the attenuation of physical association of phosphorylated proteins with phospholipid, DNA and other proteins by steric/electrostatic interference, and the creation of temporal binding sites for phosphoresidue-binding protein domains. Such phosphorylation events are under less evolutionary constraints to localize to similar sequence positions across orthologous proteins. This is supported by the observation that core net human proteins are enriched in protein- and DNA-binding function. In addition, sequence conservation of phosphorylated residues to interpret important phosphorylation sites can be compromised if the conservation metric does not normalize for the general factors such as protein abundance that influence residue and protein conservation. Hence, complementary approaches to identify phosphorylation events which functional consequences are evolutionary conserved across protein, protein complexes and pathways are needed in addition to conventional sequence analysis.

Bibliography

[1] E. Akiva, Z. Itzhaki, and H. Margalit. Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. *Proc. Natl. Acad. Sci. U.S.A.*, 105:13292–13297, Sep 2008.

[2] C. P. Albuquerque, M. B. Smolka, S. H. Payne, V. Bafna, J. Eng, and H. Zhou. A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Mol. Cell Proteomics*, 7:1389–1396, Jul 2008.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, Oct 1990.

[4] L. S. Argetsinger, J. L. Kouadio, H. Steen, A. Stensballe, O. N. Jensen, and C. Carter-Su. Autophosphorylation of JAK2 on tyrosines 221 and 570 regulates its activity. *Mol. Cell. Biol.*, 24:4955–4967, Jun 2004.

[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.*, 25:25–29, May 2000.

[6] A. N. Ba and A. M. Moses. Evolution of characterized phosphorylation sites in budding yeast. *Mol. Biol. Evol.*, 27:2027–2037, Sep 2010.

[7] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.

[8] B. A. Ballif, G. R. Carey, S. R. Sunyaev, and S. P. Gygi. Large-scale identification and evolution indexing of tyrosine phosphorylation sites from murine brain. *J. Proteome Res.*, 7:311–318, Jan 2008.

[9] B. A. Ballif, J. Villen, S. A. Beausoleil, D. Schwartz, and S. P. Gygi. Phosphoproteomic analysis of the developing mouse brain. *Mol. Cell Proteomics*, 3:1093–1101, Nov 2004.

[10] S. A. Beausoleil, M. Jedrychowski, D. Schwartz, J. E. Elias, J. Villen, J. Li, M. A. Cohn, L. C. Cantley, and S. P. Gygi. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U.S.A.*, 101:12130–12135, Aug 2004.

[11] P. Beltrao, J. C. Trinidad, D. Fiedler, A. Roguev, W. A. Lim, K. M. Shokat, A. L. Burlingame, and N. J. Krogan. Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol.*, 7:e1000134, Jun 2009.

[12] R. P. Bhattacharyya, A. Remenyi, B. J. Yeh, and W. A. Lim. Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu. Rev. Biochem.*, 75:655–680, 2006.

[13] B. Blagoev, I. Kratchmarova, S. E. Ong, M. Nielsen, L. J. Foster, and M. Mann. A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat. Biotechnol.*, 21:315–318, Mar 2003.

[14] B. Blagoev, S. E. Ong, I. Kratchmarova, and M. Mann. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat. Biotechnol.*, 22:1139–1145, Sep 2004.

[15] N. Blom, T. Sicheritz-Ponten, R. Gupta, S. Gammeltoft, and S. Brunak. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, 4:1633–1649, Jun 2004.

[16] B. Bodenmiller, D. Campbell, B. Gerrits, H. Lam, M. Jovanovic, P. Picotti, R. Schlapbach, and R. Aebersold. PhosphoPep–a database of protein phosphorylation sites in model organisms. *Nat. Biotechnol.*, 26:1339–1340, Dec 2008.

[17] B. Bodenmiller, J. Malmstrom, B. Gerrits, D. Campbell, H. Lam, A. Schmidt, O. Rinner, L. N. Mueller, P. T. Shannon, P. G. Pedrioli, C. Panse, H. K. Lee, R. Schlapbach, and R. Aebersold. PhosphoPep–a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. *Mol. Syst. Biol.*, 3:139, 2007.

[18] J. Boekhorst, B. van Breukelen, A. Heck, and B. Snel. Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol.*, 9:R144, 2008.

[19] M. Borg, T. Mittag, T. Pawson, M. Tyers, J. D. Forman-Kay, and H. S. Chan. Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc. Natl. Acad. Sci. U.S.A.*, 104:9650–9655, Jun 2007.

[20] S. N. Boyle, G. A. Michaud, B. Schweitzer, P. F. Predki, and A. J. Koleske. A critical role for cortactin phosphorylation by Abl-family kinases in PDGF-induced dorsal-wave formation. *Curr. Biol.*, 17:445–451, Mar 2007.

[21] R. I. Brinkworth, R. A. Breinl, and B. Kobe. Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl. Acad. Sci. U.S.A.*, 100:74–79, Jan 2003.

[22] C. Broceno, S. Wilkie, and S. Mittnacht. RB activation defect in tumor cell lines. *Proc. Natl. Acad. Sci. U.S.A.*, 99:14200–14205, Oct 2002.

[23] C. J. Brown, S. Takayama, A. M. Campen, P. Vise, T. W. Marshall, C. J. Oldfield, C. J. Williams, and A. K. Dunker. Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.*, 55:104–110, Jul 2002.

[24] A. Brunet, J. Park, H. Tran, L. S. Hu, B. A. Hemmings, and M. E. Greenberg. Protein kinase SGK mediates survival signals by phosphorylating the forkhead transcription factor FKHRL1 (FOXO3a). *Mol. Cell. Biol.*, 21:952–965, Feb 2001.

[25] Y. V. Budovskaya, J. S. Stephan, S. J. Deminoff, and P. K. Herman. An evolutionary proteomics approach identifies substrates of the cAMP-dependent protein kinase. *Proc. Natl. Acad. Sci. U.S.A.*, 102:13933–13938, Sep 2005.

[26] B. Cecchinelli, A. Porrello, C. Lazzari, A. Gradi, G. Bossi, M. D'Angelo, A. Sacchi, and S. Soddu. Ser58 of mouse p53 is the homologue of human Ser46 and is phosphorylated by HIPK2 in apoptosis. *Cell Death Differ.*, 13:1994–1997, Nov 2006.

[27] E. J. Chang, R. Begum, B. T. Chait, and T. Gaasterland. Prediction of cyclin-dependent kinase phosphorylation substrates. *PLoS ONE*, 2:e656, 2007.

[28] S. J. Chatterjee, B. George, P. J. Goebell, M. Alavi-Tafreshi, S. R. Shi, Y. K. Fung, P. A. Jones, C. Cordon-Cardo, R. H. Datar, and R. J. Cote. Hyperphosphorylation of pRb: a mechanism for RB tumour suppressor pathway inactivation in bladder cancer. *J. Pathol.*, 203:762–770, Jul 2004.

[29] A. Chi, C. Huttenhower, L. Y. Geer, J. J. Coon, J. E. Syka, D. L. Bai, J. Shabanowitz, D. J. Burke, O. G. Troyanskaya, and D. F. Hunt. Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.*, 104:2193–2198, Feb 2007.

[30] P. Cohen. The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur. J. Biochem.*, 268:5001–5010, Oct 2001.

[31] P. Cohen. The origins of protein phosphorylation. *Nat. Cell Biol.*, 4:E127–130, May 2002.

[32] M. O. Collins, L. Yu, I. Campuzano, S. G. Grant, and J. S. Choudhary. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. *Mol. Cell Proteomics*, 7:1331–1348, Jul 2008.

[33] U. de Lichtenberg, T. S. Jensen, S. Brunak, P. Bork, and L. J. Jensen. Evolution of cell cycle control: same molecular machines, different regulation. *Cell Cycle*, 6:1819–1825, Aug 2007.

[34] N. Dephoure, R. W. Howson, J. D. Blethrow, K. M. Shokat, and E. K. O'Shea. Combining chemical genetics and proteomics to identify protein kinase substrates. *Proc. Natl. Acad. Sci. U.S.A.*, 102:17940–17945, Dec 2005.

[35] E. W. Deutsch, H. Lam, and R. Aebersold. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics*, 33:18–25, Mar 2008.

[36] 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:D240–244, Jan 2008.

[37] J. E. Dueber, E. A. Mirsky, and W. A. Lim. Engineering synthetic signaling proteins with ultrasensitive input/output control. *Nat. Biotechnol.*, 25:660–662, Jun 2007.

[38] J. E. Dueber, B. J. Yeh, R. P. Bhattacharyya, and W. A. Lim. Rewiring cell signaling: the logic and plasticity of eukaryotic protein circuitry. *Curr. Opin. Struct. Biol.*, 14:690–699, Dec 2004.

[39] S. T. Eblen, N. V. Kumar, K. Shah, M. J. Henderson, C. K. Watts, K. M. Shokat, and M. J. Weber. Identification of novel ERK2 substrates through use of an engineered kinase and ATP analogs. *J. Biol. Chem.*, 278:14926–14935, Apr 2003.

[40] P. Edman. A method for the determination of amino acid sequence in peptides. *Arch Biochem*, 22:475, Jul 1949.

[41] O. Fedorov, B. Marsden, V. Pogacic, P. Rellos, S. Muller, A. N. Bullock, J. Schwaller, M. Sundstrom, and S. Knapp. A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. *Proc. Natl. Acad. Sci. U.S.A.*, 104:20523–20528, Dec 2007.

[42] S. B. Ficarro, M. L. McCleland, P. T. Stukenberg, D. J. Burke, M. M. Ross, J. Shabanowitz, D. F. Hunt, and F. M. White. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat. Biotechnol.*, 20:301–305, Mar 2002.

[43] P. Flicek, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. Hubbard,

A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal, and S. Searle. Ensembl 2008. *Nucleic Acids Res.*, 36:D707–714, Jan 2008.

[44] J. E. Galagan, M. R. Henn, L. J. Ma, C. A. Cuomo, and B. Birren. Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res.*, 15:1620–1631, Dec 2005.

[45] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, Mar 2006.

[46] S. A. Gerber, J. Rush, O. Stemman, M. W. Kirschner, and S. P. Gygi. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U.S.A.*, 100:6940–6945, Jun 2003.

[47] S. Ghaemmaghami, W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425:737–741, Oct 2003.

[48] T. J. Gibson. Cell regulation: determined to signal discrete cooperation. *Trends Biochem. Sci.*, 34:471–482, Oct 2009.

[49] F. Gnad, L. M. de Godoy, J. Cox, N. Neuhauser, S. Ren, J. V. Olsen, and M. Mann. High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. *Proteomics*, 9:4642–4652, Oct 2009.

[50] F. Gnad, F. Forner, D. F. Zielinska, E. Birney, J. Gunawardena, and M. Mann. Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes, and mitochondria. *Mol. Cell Proteomics*, 9:2642–2653, Dec 2010.

[51] F. Gnad, S. Ren, J. Cox, J. V. Olsen, B. Macek, M. Oroshi, and M. Mann. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.*, 8:R250, 2007.

[52] A. Gordus, J. A. Krall, E. M. Beyer, A. Kaushansky, A. Wolf-Yadlin, M. Sevecka, B. H. Chang, J. Rush, and G. MacBeath. Linear combinations of docking affinities explain quantitative differences in RTK signaling. *Mol. Syst. Biol.*, 5:235, 2009.

[53] M. B. Goshe. Characterizing phosphoproteins and phosphoproteomes using mass spectrometry. *Brief Funct Genomic Proteomic*, 4:363–376, Feb 2006.

[54] M. Grønborg, T. Z. Kristiansen, A. Stensballe, J. S. Andersen, O. Ohara, M. Mann, O. N. Jensen, and A. Pandey. A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies: identification of a novel protein, Frigg, as a protein kinase A substrate. *Mol. Cell Proteomics*, 1:517–527, Jul 2002.

[55] A. Gruhler, J. V. Olsen, S. Mohammed, P. Mortensen, N. J. Faergeman, M. Mann, and O. N. Jensen. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell Proteomics*, 4:310–327, Mar 2005.

[56] W. C. Hahn and R. A. Weinberg. Rules for making human tumor cells. *N. Engl. J. Med.*, 347:1593–1603, Nov 2002.

[57] S. B. Hedges. The origin and evolution of model organisms. *Nat. Rev. Genet.*, 3:838–849, Nov 2002.

[58] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, 89:10915–10919, Nov 1992.

[59] M. E. Higgins, M. Claremont, J. E. Major, C. Sander, and A. E. Lash. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res.*, 35:D721–726, Jan 2007.

[60] M. Hjerrild, A. Stensballe, T. E. Rasmussen, C. B. Kofoed, N. Blom, T. Sicheritz-Ponten, M. R. Larsen, S. Brunak, O. N. Jensen, and S. Gammeltoft. Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. *J. Proteome Res.*, 3:426–433, 2004.

[61] L. J. Holt, J. E. Hutti, L. C. Cantley, and D. O. Morgan. Evolution of Ime2 phosphorylation sites on Cdk1 substrates provides a mechanism to limit the effects of the phosphatase Cdc14 in meiosis. *Mol. Cell*, 25:689–702, Mar 2007.

[62] L. J. Holt, B. B. Tuch, J. Villen, A. D. Johnson, S. P. Gygi, and D. O. Morgan. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science*, 325:1682–1686, Sep 2009.

[63] S. Honnappa, S. M. Gouveia, A. Weisbrich, F. F. Damberger, N. S. Bhavesh, H. Jawhari, I. Grigoriev, F. J. van Rijssel, R. M. Buey, A. Lawera, I. Jelesarov, F. K. Winkler, K. Wuthrich, A. Akhmanova, and M. O. Steinmetz. An EB1-binding motif acts as a microtubule tip localization signal. *Cell*, 138:366–376, Jul 2009.

[64] P. V. Hornbeck, I. Chabra, J. M. Kornhauser, E. Skrzypek, and B. Zhang. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4:1551–1561, Jun 2004.

[65] 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 signal-

ing networks reveals a combinatorial therapeutic strategy for glioblastoma. *Proc. Natl. Acad. Sci. U.S.A.*, 104:12867–12872, Jul 2007.

[66] T. Hunter. Tyrosine phosphorylation: thirty years and counting. *Curr. Opin. Cell Biol.*, 21:140–146, Apr 2009.

[67] T. Hunter and B. M. Sefton. Transforming gene product of Rous sarcoma virus phosphorylates tyrosine. *Proc. Natl. Acad. Sci. U.S.A.*, 77:1311–1315, Mar 1980.

[68] J. E. Hutti, E. T. Jarrell, J. D. Chang, D. W. Abbott, P. Storz, A. Toker, L. C. Cantley, and B. E. Turk. A rapid method for determining protein kinase phosphorylation specificity. *Nat. Methods*, 1:27–29, Oct 2004.

[69] E. L. Huttlin, M. P. Jedrychowski, J. E. Elias, T. Goswami, R. Rad, S. A. Beausoleil, J. Villen, W. Haas, M. E. Sowa, and S. P. Gygi. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, 143:1174–1189, Dec 2010.

[70] J. M. Irish, N. Kotecha, and G. P. Nolan. Mapping normal and cancer cell signalling networks: towards single-cell proteomics. *Nat. Rev. Cancer*, 6:146–155, Feb 2006.

[71] K. A. Janes, J. G. Albeck, S. Gaudet, P. K. Sorger, D. A. Lauffenburger, and M. B. Yaffe. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science*, 310:1646–1653, Dec 2005.

[72] K. A. Janes, S. Gaudet, J. G. Albeck, U. B. Nielsen, D. A. Lauffenburger, and P. K. Sorger. The response of human epithelial cells to TNF involves an inducible autocrine cascade. *Cell*, 124:1225–1239, Mar 2006.

[73] L. J. Jensen, T. S. Jensen, U. de Lichtenberg, S. Brunak, and P. Bork. Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature*, 443:594–597, Oct 2006.

[74] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, 37:D412–416, Jan 2009.

[75] J. L. Jimnez, B. Hegemann, J. R. Hutchins, J. M. Peters, and R. Durbin. A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. *Genome Biol.*, 8:R90, 2007.

[76] R. F. Johnston, S. C. Pickett, and D. L. Barker. Autoradiography using storage phosphor technology. *Electrophoresis*, 11:355–360, May 1990.

[77] I. Jonassen. Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.*, 13:509–522, Oct 1997.

[78] C. Jørgensen and R. Linding. Directional and quantitative phosphorylation networks. *Brief Funct Genomic Proteomic*, 7:17–26, Jan 2008.

[79] C. Jørgensen, A. Sherman, G. I. Chen, A. Pasculescu, A. Poliakov, M. Hsiung, B. Larsen, D. G. Wilkinson, R. Linding, and T. Pawson. Cell-specific information processing in segregating populations of Eph receptor ephrin-expressing cells. *Science*, 326:1502–1509, Dec 2009.

[80] 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*, 5:59–67, Jan 2009.

[81] K. Katoh and H. Toh. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics*, 9:286–298, Jul 2008.

[82] M. Kellis, B. W. Birren, and E. S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, 428:617–624, Apr 2004.

[83] J. E. Kim, S. R. Tannenbaum, and F. M. White. Global phosphoproteome of HT-29 human colon adenocarcinoma cells. *J. Proteome Res.*, 4:1339–1346, 2005.

[84] J. H. Kim, J. Lee, B. Oh, K. Kimm, and I. Koh. Prediction of phosphorylation sites using SVMs. *Bioinformatics*, 20:3179–3184, Nov 2004.

[85] S. Y. Kim and J. E. Ferrell. Substrate competition as a source of ultrasensitivity in the inactivation of Wee1. *Cell*, 128:1133–1145, Mar 2007.

[86] N. King, M. J. Westbrook, S. L. Young, A. Kuo, M. Abedin, J. Chapman, S. Fairclough, U. Hellsten, Y. Isogai, I. Letunic, M. Marr, D. Pincus, N. Putnam, A. Rokas, K. J. Wright, R. Zuzow, W. Dirks, M. Good, D. Goodstein, D. Lemons, W. Li, J. B. Lyons, A. Morris, S. Nichols, D. J. Richter, A. Salamov, J. G. Sequencing, P. Bork, W. A. Lim, G. Manning, W. T. Miller, W. McGinnis, H. Shapiro, R. Tjian, I. V. Grigoriev, and D. Rokhsar. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature*, 451:783–788, Feb 2008.

[87] M. Kitagawa, H. Higashi, I. S. Takahashi, T. Okabe, H. Ogino, Y. Taya, S. Hishimura, and A. Okuyama. A cyclin-dependent kinase inhibitor, butyrolactone I, inhibits phosphorylation of RB protein and cell cycle progression. *Oncogene*, 9:2549–2557, Sep 1994.

[88] N. Kumar, A. Wolf-Yadlin, F. M. White, and D. A. Lauffenburger. Modeling HER2 effects on cell behavior from mass spectrometry phosphotyrosine data. *PLoS Comput. Biol.*, 3:e4, Jan 2007.

[89] C. Landgraf, S. Panni, L. Montecchi-Palazzi, L. Castagnoli, J. Schneider-Mergener, R. Volkmer-Engert, and G. Cesareni. Protein interaction networks by proteome peptide scanning. *PLoS Biol.*, 2:E14, Jan 2004.

[90] C. R. Landry, E. D. Levy, and S. W. Michnick. Weak functional constraints on phosphoproteomes. *Trends Genet.*, 25:193–197, May 2009.

[91] S. Laugesen, E. Messinese, S. Hem, C. Pichereaux, S. Grat, R. Ranjeva, M. Rossignol, and J. J. Bono. Phosphoproteins analysis in plants: a proteomic approach. *Phytochemistry*, 67:2208–2214, Oct 2006.

[92] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, Oct 1993.

[93] S. Lemeer, C. Jopling, J. Gouw, S. Mohammed, A. J. Heck, M. Slijper, and J. den Hertog. Comparative phosphoproteomics of zebrafish Fyn/Yes morpholino knock-down embryos. *Mol. Cell Proteomics*, 7:2176–2187, Nov 2008.

[94] S. Lemeer, C. Jopling, F. Naji, R. Ruijtenbeek, M. Slijper, A. J. Heck, and J. den Hertog. Protein-tyrosine kinase activity profiling in knock down zebrafish embryos. *PLoS ONE*, 2:e581, 2007.

[95] I. Letunic, R. R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, 34:D257–260, Jan 2006.

[96] X. Li, S. A. Gerber, A. D. Rudner, S. A. Beausoleil, W. Haas, J. Villen, J. E. Elias, and S. P. Gygi. Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. *J. Proteome Res.*, 6:1190–1197, Mar 2007.

[97] M. E. Liku, V. Q. Nguyen, A. W. Rosales, K. Irie, and J. J. Li. CDK phosphorylation of a novel NLS-NES module distributed between two subunits of the Mcm2-7 complex prevents chromosomal rereplication. *Mol. Biol. Cell*, 16:5026–5039, Oct 2005.

[98] R. Linding, L. J. Jensen, G. J. Ostheimer, M. A. van Vugt, C. Jørgensen, 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*, 129:1415–1426, Jun 2007.

[99] R. Linding, L. J. Jensen, A. Pasculescu, M. Olhovsky, K. Colwill, P. Bork, M. B. Yaffe, and T. Pawson. NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.*, 36:D695–699, Jan 2008.

[100] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, 31:3701–3708, Jul 2003.

[101] B. A. Liu, K. Jablonowski, E. E. Shah, B. W. Engelmann, R. B. Jones, and P. D. Nash. SH2 domains recognize contextual peptide sequence information to determine selectivity. *Mol. Cell Proteomics*, 9:2391–2404, Nov 2010.

[102] Y. Liu, K. Shah, F. Yang, L. Witucki, and K. M. Shokat. Engineering Src family protein kinases with unnatural nucleotide specificity. *Chem. Biol.*, 5:91–101, Feb 1998.

[103] C. Louvet, G. L. Szot, J. Lang, M. R. Lee, N. Martinier, G. Bollag, S. Zhu, A. Weiss, and J. A. Bluestone. Tyrosine kinase inhibitors reverse type 1 diabetes in nonobese diabetic mice. *Proc. Natl. Acad. Sci. U.S.A.*, 105:18895–18900, Dec 2008.

[104] G. MacBeath and S. L. Schreiber. Printing proteins as microarrays for high-throughput function determination. *Science*, 289:1760–1763, Sep 2000.

[105] B. Macek, F. Gnad, B. Soufi, C. Kumar, J. V. Olsen, I. Mijakovic, and M. Mann. Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol. Cell Proteomics*, 7:299–307, Feb 2008.

[106] S. Maere, K. Heymans, and M. Kuiper. BiNGO: a Cytoscape plugin to assess over-representation of gene ontology categories in biological networks. *Bioinformatics*, 21:3448–3449, Aug 2005.

[107] R. Malik, E. A. Nigg, and R. Korner. Comparative conservation analysis of the human mitotic phosphoproteome. *Bioinformatics*, 24:1426–1432, Jun 2008.

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

[109] G. Manning, S. L. Young, W. T. Miller, and Y. Zhai. The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *Proc. Natl. Acad. Sci. U.S.A.*, 105:9674–9679, Jul 2008.

[110] S. Matsuoka, B. A. Ballif, A. Smogorzewska, E. R. McDonald, K. E. Hurov, J. Luo, C. E. Bakalarski, Z. Zhao, N. Solimini, Y. Lerenthal, Y. Shiloh, S. P. Gygi, and S. J. Elledge. ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science*, 316:1160–1166, May 2007.

[111] V. Mayya, D. H. Lundgren, S. I. Hwang, K. Rezaul, L. Wu, J. K. Eng, V. Rodionov, and D. K. Han. Quantitative phosphoproteomic analysis of T cell receptor signaling reveals system-wide modulation of protein-protein interactions. *Sci Signal*, 2:ra46, 2009.

[112] S. McLaughlin and A. Aderem. The myristoyl-electrostatic switch: a modulator of reversible protein-membrane interactions. *Trends Biochem. Sci.*, 20:272–276, Jul 1995.

[113] M. L. Miller, S. Hanke, A. M. Hinsby, C. Friis, S. Brunak, M. Mann, and N. Blom. Motif decomposition of the phosphotyrosine proteome reveals a new N-terminal binding motif for SHIP2. *Mol. Cell Proteomics*, 7:181–192, Jan 2008.

[114] M. L. Miller, L. J. Jensen, F. Diella, C. Jørgensen, 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:ra2, 2008.

[115] K. Miller-Jensen, K. A. Janes, J. S. Brugge, and D. A. Lauffenburger. Common effector processing mediates cell-specific responses to stimuli. *Nature*, 448:604–608, Aug 2007.

[116] F. Mitelman. Recurrent chromosome aberrations in cancer. *Mutat. Res.*, 462:247–253, Apr 2000.

[117] J. Mok, P. M. Kim, H. Y. Lam, S. Piccirillo, X. Zhou, G. R. Jeschke, D. L. Sheridan, S. A. Parker, V. Desai, M. Jwa, E. Cameroni, H. Niu, M. Good, A. Remenyi, J. L. Ma, Y. J. Sheu, H. E. Sassi, R. Sopko, C. S. Chan, C. De Virgilio, N. M. Hollingsworth, W. A. Lim, D. F. Stern, B. Stillman, B. J. Andrews, M. B. Gerstein, M. Snyder, and B. E. Turk. Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Sci Signal*, 3:ra12, 2010.

[118] K. Moser and F. M. White. Phosphoproteomic analysis of rat liver by high capacity IMAC and LC-MS/MS. *J. Proteome Res.*, 5:98–104, Jan 2006.

[119] A. M. Moses, J. K. Heriche, and R. Durbin. Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. *Genome Biol.*, 8:R23, 2007.

[120] A. M. Moses, M. E. Liku, J. J. Li, and R. Durbin. Regulatory evolution in proteins by turnover and lineage-specific changes of cyclin-dependent kinase consensus sites. *Proc. Natl. Acad. Sci. U.S.A.*, 104:17713–17718, Nov 2007.

[121] K. Nakayama, Y. Yamada, T. Koji, T. Hayashi, M. Tomonaga, and S. Kamihira. Expression and phosphorylation status of retinoblastoma protein in adult T-cell leukemia/lymphoma. *Leuk. Res.*, 24:299–305, Apr 2000.

[122] P. Nash, X. Tang, S. Orlicky, Q. Chen, F. B. Gertler, M. D. Mendenhall, F. Sicheri, T. Pawson, and M. Tyers. Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. *Nature*, 414:514–521, Nov 2001.

[123] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. J. Gibson, J. Lewis, L. Serrano, and R. B. Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.*, 3:e405, Dec 2005.

[124] V. Neduva and R. B. Russell. Linear motifs: evolutionary interaction switches. *FEBS Lett.*, 579:3342–3345, Jun 2005.

[125] A. I. Nesvizhskii, O. Vitek, and R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods*, 4:787–797, Oct 2007.

[126] G. Neuberger, G. Schneider, and F. Eisenhaber. pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. *Biol. Direct*, 2:1, 2007.

[127] Y. Oda, K. Huang, F. R. Cross, D. Cowburn, and B. T. Chait. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. U.S.A.*, 96:6591–6596, Jun 1999.

[128] Y. Oda, T. Nagasu, and B. T. Chait. Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat. Biotechnol.*, 19:379–382, Apr 2001.

[129] 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*, 127:635–648, Nov 2006.

[130] Jesper V Olsen, Michiel Vermeulen, Anna Santamaria, Chanchal Kumar, Martin L Miller, Lars J Jensen, Florian Gnad, Jurgen Cox, Thomas S Jensen, Erich A Nigg, Soren Brunak, and Matthias Mann. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal*, 3(104):ra3, 2010.

[131] 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*, 1:376–386, May 2002.

[132] C. Pan, F. Gnad, J. V. Olsen, and M. Mann. Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors. *Proteomics*, 8:4534–4546, Nov 2008.

[133] T. Pawson and R. Linding. Network medicine. *FEBS Lett.*, 582:1266–1270, Apr 2008.

[134] E. Perrodou, C. Chica, O. Poch, T. J. Gibson, and J. D. Thompson. A new protein linear motif benchmark for multiple sequence alignment software. *BMC Bioinformatics*, 9:213, 2008.

[135] D. Pincus, I. Letunic, P. Bork, and W. A. Lim. Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. *Proc. Natl. Acad. Sci. U.S.A.*, 105:9680–9684, Jul 2008.

[136] M. C. Posewitz and P. Tempst. Immobilized gallium(III) affinity chromatography of phosphopeptides. *Anal. Chem.*, 71:2883–2892, Jul 1999.

[137] J. Ptacek, G. Devgan, G. Michaud, H. Zhu, X. Zhu, J. Fasolo, H. Guo, G. Jona, A. Breitkreutz, R. Sopko, R. R. McCartney, M. C. Schmidt, N. Rachidi, S. J. Lee, A. S. Mah, L. Meng, M. J. Stark, D. F. Stern, C. De Virgilio, M. Tyers, B. Andrews, M. Gerstein, B. Schweitzer, P. F. Predki, and M. Snyder. Global analysis of protein phosphorylation in yeast. *Nature*, 438:679–684, Dec 2005.

[138] J. Ptacek and M. Snyder. Charging it up: global analysis of protein phosphorylation. *Trends Genet.*, 22:545–554, Oct 2006.

[139] P. Puntervoll, R. Linding, C. Gemund, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferre, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Kuster, M. Helmer-Citterich, W. N. Hunter, R. Aasland, and T. J. Gibson. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, 31:3625–3630, Jul 2003.

[140] D. W. Raiford, E. M. Heizer, R. V. Miller, H. Akashi, M. L. Raymer, and D. E. Krane. Do amino acid biosynthetic costs constrain protein evolution in Saccharomyces cerevisiae? *J. Mol. Evol.*, 67:621–630, Dec 2008.

[141] D. Rambaldi, F. M. Giorgi, F. Capuani, A. Ciliberto, and F. D. Ciccarelli. Low duplicability and network fragility of cancer genes. *Trends Genet.*, 24:427–430, Sep 2008.

[142] M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, 314:1041–1052, Dec 2001.

[143] G. Rena, Y. L. Woods, A. R. Prescott, M. Peggie, T. G. Unterman, M. R. Williams, and P. Cohen. Two novel phosphorylation sites on FKHR that are critical for its nuclear exclusion. *EMBO J.*, 21:2263–2271, May 2002.

[144] T. A. Richards, J. B. Dacks, S. A. Campbell, J. L. Blanchard, P. G. Foster, R. McLeod, and C. W. Roberts. Evolutionary origins of the eukaryotic shikimate pathway: gene fusions, horizontal gene transfer, and endosymbiotic replacements. *Eukaryotic Cell*, 5:1517–1531, Sep 2006.

[145] J. S. Richardson and D. C. Richardson. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci. U.S.A.*, 99:2754–2759, Mar 2002.

[146] I. Rigoutsos and A. Floratos. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, 14:55–67, 1998.

[147] 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*, 131:1190–1203, Dec 2007.

[148] A. Ritz, G. Shakhnarovich, A. R. Salomon, and B. J. Raphael. Discovery of phosphorylation motif mixtures in phosphoproteomics data. *Bioinformatics*, 25:14–21, Jan 2009.

[149] 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*, 3:1154–1169, Dec 2004.

[150] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, Oct 2005.

[151] J. Rush, A. Moritz, K. A. Lee, A. Guo, V. L. Goss, E. J. Spek, H. Zhang, X. M. Zha, R. D. Polakiewicz, and M. J. Comb. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.*, 23:94–101, Jan 2005.

[152] L. Rychlewski, M. Kschischo, L. Dong, M. Schutkowski, and U. Reimer. Target specificity analysis of the Abl kinase using peptide microarray data. *J. Mol. Biol.*, 336:307–311, Feb 2004.

[153] A. R. Salomon, S. B. Ficarro, L. M. Brill, A. Brinker, Q. T. Phung, C. Ericson, K. Sauer, A. Brock, D. M. Horn, P. G. Schultz, and E. C. Peters. Profiling of

tyrosine phosphorylation pathways in human cells using mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.*, 100:443–448, Jan 2003.

[154] K. Schmelzle, S. Kane, S. Gridley, G. E. Lienhard, and F. M. White. Temporal dynamics of tyrosine phosphorylation in insulin signaling. *Diabetes*, 55:2171–2179, Aug 2006.

[155] K. Schmelzle and F. M. White. Phosphoproteomic approaches to elucidate cellular signaling networks. *Curr. Opin. Biotechnol.*, 17:406–414, Aug 2006.

[156] S. R. Schmidt, F. Schweikart, and M. E. Andersson. Current methods for phosphoprotein isolation and enrichment. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, 849:154–162, Apr 2007.

[157] A. S. Schwartz and L. Pachter. Multiple alignment by sequence annealing. *Bioinformatics*, 23:e24–29, Jan 2007.

[158] D. Schwartz and S. P. Gygi. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.*, 23:1391–1398, Nov 2005.

[159] R. Schweiger and M. Linial. Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. *Biol. Direct*, 5:6, 2010.

[160] B. T. Seet, I. Dikic, M. M. Zhou, and T. Pawson. Reading protein modifications with interaction domains. *Nat. Rev. Mol. Cell Biol.*, 7:473–483, Jul 2006.

[161] Z. Serber and J. E. Ferrell. Tuning bulk electrostatics to regulate protein function. *Cell*, 128:441–444, Feb 2007.

[162] K. Shah, Y. Liu, C. Deirmengian, and K. M. Shokat. Engineering unnatural nucleotide specificity for Rous sarcoma virus tyrosine kinase to uniquely label its direct substrates. *Proc. Natl. Acad. Sci. U.S.A.*, 94:3565–3570, Apr 1997.

[163] K. Shah and K. M. Shokat. A chemical genetic screen for direct v-Src substrates reveals ordered assembly of a retrograde signaling pathway. *Chem. Biol.*, 9:35–47, Jan 2002.

[164] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13:2498–2504, Nov 2003.

[165] J. C. Smith, M. A. Duchesne, P. Tozzi, M. Ethier, and D. Figeys. A differential phosphoproteomic analysis of retinoic acid-treated P19 cells. *J. Proteome Res.*, 6:3174–3186, Aug 2007.

[166] M. B. Smolka, C. P. Albuquerque, S. H. Chen, and H. Zhou. Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases. *Proc. Natl. Acad. Sci. U.S.A.*, 104:10364–10369, Jun 2007.

[167] 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.*, 4:973–982, Nov 1994.

[168] B. Soufi, C. D. Kelstrup, G. Stoehr, F. Frohlich, T. C. Walther, and J. V. Olsen. Global analysis of the yeast osmotic stress response by quantitative proteomics. *Mol Biosyst*, 5:1337–1346, Nov 2009.

[169] C. Stark, T. C. Su, A. Breitkreutz, P. Lourenco, M. Dahabieh, B. J. Breitkreutz, M. Tyers, and I. Sadowski. PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast Saccharomyces cerevisiae. *Database (Oxford)*, 2010:bap026, 2010.

[170] 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.*, 277:1031–1039, Jan 2002.

[171] H. Steen and M. Mann. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.*, 5:699–711, Sep 2004.

[172] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122:957–968, Sep 2005.

[173] M. P. Stokes, J. Rush, J. Macneill, J. M. Ren, K. Sprott, J. Nardone, V. Yang, S. A. Beausoleil, S. P. Gygi, M. Livingstone, H. Zhang, R. D. Polakiewicz, and M. J. Comb. Profiling of UV-induced ATM/ATR signaling pathways. *Proc. Natl. Acad. Sci. U.S.A.*, 104:19855–19860, Dec 2007.

[174] S. C. Strickfaden, M. J. Winters, G. Ben-Ari, R. E. Lamson, M. Tyers, and P. M. Pryciak. A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. *Cell*, 128:519–531, Feb 2007.

[175] N. Sugiyama, H. Nakagami, K. Mochida, A. Daudi, M. Tomita, K. Shirasu, and Y. Ishihama. Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis. *Mol. Syst. Biol.*, 4:193, 2008.

[176] G. Superti-Furga, S. Fumagalli, M. Koegl, S. A. Courtneidge, and G. Draetta. Csk inhibition of c-Src activity requires both the SH2 and SH3 domains of Src. *EMBO J.*, 12:2625–2634, Jul 1993.

[177] G. Superti-Furga, K. Jonsson, and S. A. Courtneidge. A functional screen in yeast for regulators and antagonizers of heterologous protein tyrosine kinases. *Nat. Biotechnol.*, 14:600–605, May 1996.

[178] E. Szathmary, F. Jordan, and C. Pal. Molecular biology and evolution. Can genes explain biological complexity? *Science*, 292:1315–1316, May 2001.

[179] Bonner J. T. The origins of multicellularity. *Integr Biol.*, 1:28–36, May 1998.

[180] C. S. Tan, B. Bodenmiller, A. Pasculescu, M. Jovanovic, M. O. Hengartner, C. Jørgensen, G. D. Bader, R. Aebersold, T. Pawson, and R. Linding. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal*, 2:ra39, 2009.

[181] S. H. Tan, W. Hugo, W. K. Sung, and S. K. Ng. A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics*, 7:502, 2006.

[182] Y. R. Tang, Y. Z. Chen, C. A. Canchaya, and Z. Zhang. GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. *Protein Eng. Des. Sel.*, 20:405–412, Aug 2007.

[183] W. A. Tao, B. Wollscheid, R. O'Brien, J. K. Eng, X. J. Li, B. Bodenmiller, J. D. Watts, L. Hood, and R. Aebersold. Quantitative phosphoproteome analysis using a dendrimer conjugation chemistry and tandem mass spectrometry. *Nat. Methods*, 2:591–598, Aug 2005.

[184] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.*, 27:199–204, Feb 2009.

[185] J. A. Ubersax, E. L. Woodbury, P. N. Quang, M. Paraz, J. D. Blethrow, K. Shah, K. M. Shokat, and D. O. Morgan. Targets of the cyclin-dependent kinase Cdk1. *Nature*, 425:859–864, Oct 2003.

[186] A. Ultsch and H. P. Siemon. Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. In *Proceedings of International Neural Networks Conference (INNC)*, pages 305–308, Paris, 1990. Kluwer Academic Press.

[187] J. Villen, S. A. Beausoleil, S. A. Gerber, and S. P. Gygi. Large-scale phosphorylation analysis of mouse liver. *Proc. Natl. Acad. Sci. U.S.A.*, 104:1488–1493, Jan 2007.

[188] C. Vogel and C. Chothia. Protein family expansions and biological complexity. *PLoS Comput. Biol.*, 2:e48, May 2006.

[189] B. Vogelstein and K. W. Kinzler. Cancer genes and the pathways they control. *Nat. Med.*, 10:789–799, Aug 2004.

[190] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork. STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, 35:D358–362, Jan 2007.

[191] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, May 2002.

[192] J. Wan, S. Kang, C. Tang, J. Yan, Y. Ren, J. Liu, X. Gao, A. Banerjee, L. B. Ellis, and T. Li. Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. *Nucleic Acids Res.*, 36:e22, Mar 2008.

[193] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, 337:635–645, Mar 2004.

[194] 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.*, 104:5860–5865, Apr 2007.

[195] 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.

[196] 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:W588–594, Jul 2007.

[197] S. Wright. Evolution in Mendelian Populations. *Genetics*, 16:97–159, Mar 1931.

[198] 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.

[199] Y. Xue, J. Ren, X. Gao, C. Jin, L. Wen, and X. Yao. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell Proteomics*, 7:1598–1608, Sep 2008.

[200] M. B. Yaffe, G. G. Leparc, J. Lai, T. Obata, S. Volinia, and L. C. Cantley. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.*, 19:348–353, Apr 2001.

[201] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322:104–110, Oct 2008.

[202] A. Zarrinpar, S. H. Park, and W. A. Lim. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature*, 426:676–680, Dec 2003.

[203] B. Zhai, J. Villen, S. A. Beausoleil, J. Mintseris, and S. P. Gygi. Phosphoproteome analysis of Drosophila melanogaster embryos. *J. Proteome Res.*, 7:1675–1682, Apr 2008.

[204] 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*, 4:1240–1250, Sep 2005.

[205] B. Zheng, J. H. Jeong, J. M. Asara, Y. Y. Yuan, S. R. Granter, L. Chin, and L. C. Cantley. Oncogenic B-RAF negatively regulates the tumor suppressor LKB1 to promote melanoma cell proliferation. *Mol. Cell*, 33:237–247, Jan 2009.

[206] H. Zheng, P. Hu, D. F. Quinn, and Y. K. Wang. Phosphotyrosine proteomic study of interferon alpha signaling pathway using a combination of immunoprecipitation and immobilized metal affinity chromatography. *Mol. Cell Proteomics*, 4:721–730, Jun 2005.

[207] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Ger-

stein, and M. Snyder. Global analysis of protein activities using proteome chips. *Science*, 293:2101–2105, Sep 2001.

[208] H. Zhu, J. F. Klemic, S. Chang, P. Bertone, A. Casamayor, K. G. Klemic, D. Smith, M. Gerstein, M. A. Reed, and M. Snyder. Analysis of yeast protein kinases using protein chips. *Nat. Genet.*, 26:283–289, Nov 2000.

[209] J. Zuberek, A. Wyslouch-Cieszynska, A. Niedzwiecka, M. Dadlez, J. Stepinski, W. Augustyniak, A. C. Gingras, Z. Zhang, S. K. Burley, N. Sonenberg, R. Stolarski, and E. Darzynkiewicz. Phosphorylation of eIF4E attenuates its interaction with mRNA 5' cap analogs by electrostatic repulsion: intein-mediated protein ligation strategy to obtain phosphorylated protein. *RNA*, 9:52–61, Jan 2003.