TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Computational Methods and Tools for Functional Analysis and Mining of High-throughput Proteomics Data

Stefka Tyanova

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Wiehenstephan für Ernährung, Landnutzung un Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. H. Luksch
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. D. Frischmann
	2. Hon.-Prof. Dr. M. Mann
	(Ludwig-Maximilians-Universität München)

Die Dissertation wurde am 15.05.2013 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Wiehenstephan für Ernährung, Landnutzung un Umwelt am 06.08.2013 angenommen.

Contents

List of Figures

List of Tables

Summary

We are drowning in information and starving for knowledge.
—Rutherford D. Roger

Rapid developments in the field of mass spectrometry-based proteomics are now enabling the identification and quantification of thousands of proteins under different biological conditions. This facilitates the generation of sufficient data to study numerous biological problems, such as the complex mechanisms underlying the intricate dynamics of signal transduction or the biological conditions resulting in the onset of severe diseases such as cancer. However, obtaining the data to the needed depth is just the first step in the process of knowledge generation, and in a second step suitable analytical tools and methods are required to complete it. This thesis provides contributions to this second part – the development and application of sophisticated tools for knowledge mining of large-scale proteomics data. In particular, computational methods for two major sub-fields in proteomics are described: post-translational modifications and clinical studies.

Nowadays, thousands of phosphorylation sites are routinely being identified in mass spectrometry-based proteomics experiments upon suitable cell stimulation and enrichment. As in many other areas characterized by fast developing technology, the functional annotation of the measured data is lagging far behind. Mapping phosphorylation sites to specific biological context, such as structural environment and interaction with other post-translational modifications, has the potential to reveal important insights into their functions. Chapter 2 of this thesis focuses on the analysis of various properties of phospho-sites and elaborates on how these may be intertwined with their functional roles. Special emphasis is placed on phosphotyrosines and the mechanisms underlying the accurate execution of their regulatory actions. The ability of phospho-sites to form functional clusters and to cross-talk

with other post-translational modifications is investigated. The observed tendencies provide strong evidence of the ability of the cell to efficiently integrate signals from distinct pathways in order to produce rapid and robust responses to various stimuli. Furthermore, they exemplify possible mechanisms for enhancing the functional and interaction spaces of phospho-proteins.

In chapter 3, the variation of phosphorylation during the cell cycle is studied with respect to the structural environment of the phospho-sites. Two groups are distinguished: sites with dynamically varying levels that are associated with intrinsically disordered regions and sites with more constant phosphorylation levels predominantly found in regular secondary structures. These results suggest that protein structure may encode distinct functions of the phospho-acceptors.

Motivated by the need for better diagnostic markers and drug targets, the second part of the thesis focuses on the development of a framework for analysis of oncoproteomics data. Accounting for processes such as degradation, secretion and localization, proteome profiling provides direct insights into the functional phenotype of cells during cancer progression. However, the large feature space (thousands of proteins) combined with a low sample size (tens of patients) and the large genetic variability characterizing the patients pose significant challenges for efficient data analysis. The framework developed here addresses these difficulties by coupling sophisticated supervised learning methods with efficient feature selection techniques.

The generic implementation of the framework supports diverse classification methods, however, due to their suitability for data sets with high dimensional feature space Support Vector Machines are employed in the analyses in this thesis. Moreover, the high prediction performance is enhanced through feature reduction methods such as ANOVA-based or SVM weights-based ranking. The rigorous design and implementation of the analysis workflow ensures maximum generalizability of the trained models and relevance of the identified discriminative features.

Furthermore, the developed framework is successfully applied to two oncoproteomics data sets. In both sets, biologically relevant features are identified and high performance predictors are built. The comparison of the selection methods reveals that, based on their underlying principles, they can be used to address different biological

and clinical questions and may be suitable to identify single biomarkers or protein sets, characteristic of the underlying mechanisms of the disease. The results clearly demonstrate that proteomics data combined with supervised learning techniques holds tremendous promise for progress in the field of personalized medicine.

Zusammenfassung

Rasante Fortschritte auf dem Gebiet der massenspektrometriebasierten Proteomik ermöglichen heute die Identifikation und Quantifizierung von tausenden Proteinen in unterschiedlichen biologischen Zuständen und somit die Generation einer umfassenden Datenbasis zur Untersuchung biologischer Fragestellungen. Dazu gehören beispielsweise die komplexen Mechanismen, die der Dynamik von Signaltransduktion unterliegen, oder die biologischen Zustände, die zur Entstehung schwerwiegender Erkrankungen wie Krebs führen. Die Erfassung der Daten in der benötigten Tiefe ist jedoch nur der erste Schritt im Prozess des wissenschaftlichen Erkenntnisgewinnes. In einem zweiten Schritt werden zur Analyse und Bewertung der gewonnenen Daten geeignete analytische Werkzeuge und Methoden benötigt. Diese Dissertation leistet einen Beitrag zu der erforderlichen Entwicklung und Anwendung fortgeschrittener Werkzeuge zum Erkenntnisgewinn aus umfangreichen Proteomikdatensätzen. Im Speziellen werden computergestützte Methoden für zwei wichtige Teilbereiche in der Proteomik beschrieben: posttranslationale Modifizierungen und klinische Studien.

Tausende von Phosphorylierungsstellen werden derzeit routinemäßig nach geeigneter Zellstimulation und Anreicherung mittels massenspektrometrie-basierter Proteomik-Analyse identifiziert. Wie in vielen Feldern, die von schnellen technologischen Entwicklungen geprägt sind, bleibt die funktionelle Einordnung und Erklärung der gemessenen Daten oft weit zurück. Die Zuordnung von Phosphorylierungsstellen zu ihrem biologischen Kontext, wie dem strukturellen Umfeld und der Interaktion mit weiteren posttranslationalen Modifikationen kann potenziell wichtige Erkenntnisse über deren Funktion liefern. Kapitel 2 dieser Dissertation konzentriert sich auf die Analyse verschiedener Eigenschaften von Phosphorylierungsstellen und führt aus, wie diese mit deren Funktion zusammenhängen können. Ein besonderer Fokus liegt auf Phosphotyrosinen und den Mechanismen, die die präzise Umsetzung ihrer

regulatorischen Wirkung ermöglichen. Untersucht wird die Fähigkeit von Phosphorylierungsstellen, funktionelle Cluster zu bilden und mit anderen posttranslationalen Modifikationen Wechselwirkungen einzugehen. Die beobachteten Tendenzen weisen auf die Fähigkeit der Zelle hin, Signale von unterschiedlichen Signalwegen effektiv zu integrieren, um schnelle und robuste Antworten auf verschiedene Stimuli zu erzeugen. Zudem werden dabei beispielhaft mögliche Mechanismen zur Erweiterung der Funktions- und Interaktionsräume von Phosphoproteinen gezeigt.

Kapitel 3 beinhaltet die Untersuchung der Variabilität von Phosphorylierungen im Verlauf des Zellzyklus mit besonderem Fokus auf die strukturelle Umgebung der Phosphorylierungsstellen. Dabei konnten zwei Gruppen von Phosphorylierungsstellen unterschieden werden: Stellen mit dynamisch veränderlichen Phosphorylierungsleveln, die mit eigentlich ungeordneten Regionen assoziiert sind, und Stellen mit eher konstanten Phosphorylierungsleveln, die vorwiegend in regelmäßigen Sekundärstrukturen gefunden werden. Dieses Ergebnis weist auf die Kodierung der Funktion von phospho-Akzeptoren durch die Proteinstruktur hin.

Angeregt durch den Bedarf an besseren diagnostischen Markern und Drug Targets ist der zweite Teil dieser Dissertations auf die Entwicklung eines Klassifizierungsrahmens zur Analyse von Onkoproteomikdaten ausgerichtet. Unter Einbeziehung von Prozessen wie Abbau, Sekretion und Lokalisierung kann proteomisches Profiling direkte Einblicke in den funktionellen Phänotyp von Zellen während der Krebsprogression liefern. Allerdings stellt die Größe des Merkmalsraumes (Tausende von Proteinen) in Kombination mit einer geringen Anzahl von Proben und genetischer Variabilität der Patienten eine groe Herausforderung für effiziente Datenanalyse dar. Der hier entwickelte Klassifizierungsrahmen widmet sich diesem Problem durch die Verbindung von anspruchsvollen überwachten Methoden des maschinellen Lernens mit effizienter Auswahl von Merkmalen.

Die generische Implementierung des Rahmens unterstützt diverse Klassifikationsmethoden, jedoch wurden im Rahmen dieser Dissertation Support Vector Machines verwendet aufgrund ihrer Eignung für Datensätze mit hochdimensionalen Merkmalsräumen. Zudem wurde die hohe Vorhersageleistung durch Merkmalsreduktionsverfahren wie ANOVA- oder SVM-wichtungsbasierten Rankings weiter gesteigert. Stringentes Design und Implementierung von Analyseabläufen stellt eine maximale

Verallgemeinbarkeit der erzeugten Modelle und Relevanz der identifizierten Merkmalen sicher.

Darüberhinaus wird im Rahmen der vorliegenden Dissertation die erfolgreiche Anwendung des Klassifizierungsrahmens auf zwei Onkoproteomikdatensätze gezeigt. In beiden Datensätzen wurden biologische relevante Eigenschaften identifiziert und leistungsstarke Prädiktoren erstellt. Ein Vergleich der Auswahlmethoden und ihrer zugrundeliegenden Prinzipien zeigt, dass diese bei der Untersuchung weiterer biologischer und klinischer Fragestellungen zur Anwendung kommen können. Zudem können sie eingesetzt werden, um einzelne Biomarker oder Proteingruppen zu identifizieren, die charakteristisch für den zugrundeliegenden Mechanismus einer Krankheit sind. Die vorliegenden Ergebnisse belegen deutlich, dass die Kombination von Proteomikdaten mit anspruchsvollen überwachten Methoden des maschinellen Lernens ein großes Potential für Fortschritte im Feld der personalisierten Medizin aufweist.

Chapter 1

Introduction

Proteins are actively involved in almost all biological activities and together with the other building blocks in the cell regulate the execution of various functions and shape the response to external and internal stimuli. The scientific area concerned with obtaining a comprehensive view of the entire complement of proteins expressed by an organism or a cell population in a specific state is proteomics [1, 2, 3]. In particular, mass spectrometry-based proteomics enables the extensive characterization and understanding of the intricate dynamics of the cellular processes. It holds an arsenal of technologies and methods to study both the absolute amount and the relative changes in the amount of proteins, including information on compartmentalization, localization and protein-protein interactions [4]. Post-translational modifications add an extra level of complexity to the proteome and present another subject of study of proteomics. Thus, proteomics has emerged as an indispensable part of systems biology allowing for detailed and accurate portrayals of the differences of functionally-relevant sub-proteomes expressed under different conditions [5].

Bioinformatics plays a vital role in the development of the field of mass spectrometry-based proteomics. On one hand, advanced algorithms enable and enhance the peptide identification and protein quantification from raw mass spectrometry files [6, 7, 8]. On the other hand, the interpretation of the large-scale quantitative data requires the development and use of sophisticated methods for analysis, as well the combination of various information resources. For example, putting proteomics data in the context of other annotations such as cellular localization, biological processes, domains and structural features is necessary to assign functional relevance. Often

the comparison of proteome profiles of complex samples is desired in order to extract valuable information on differentially expressed features. This task can be complicated by the existence of additional signals related to other sources of variability, noise and high feature dimensionality. The nature of mass spectrometry-based acquisition of proteomics data and the challenges related to the downstream analysis of such data are discussed in detail in the remaining part of the Introduction section.

1.1 Mass spectrometry-based proteomics

Recent advances in the field of mass spectrometry-based proteomics have enabled the identification and quantification of tens of thousands of proteins in highly complex biological samples under different conditions [9, 3, 10, 11, 12, 13]. The proteins in a sample are identified based on their accurate and often unique molecular weight most commonly using a shotgun proteomics approach, in which high performance liquid chromatography is directly coupled to the mass spectrometer.

1.1.1 Cell lysis and enzymatic digestion

The two currently-employed approaches of protein characterization are top-down and bottom-up proteomics. In top-down proteomics the intact proteins or protein complexes are directly subjected to analysis in the mass spectrometer, whereas in bottom-up approaches the molecules are first digested into peptides. The first step in a typical bottom-up workflow is cell lysis during which the proteins are extracted from their cellular environment. Due to the complexity of the protein mixtures, mass spectrometry is usually combined with various separation techniques at different stages of the analysis. An optional next step is the separation of the protein mixture, for example, by means of liquid chromatography or gel electrophoresis. Alternatively, proteins can be separated according to their size by SDS-PAGE, upon which the gel is cut into slices. The extracted proteins are then digested to peptides. A commonly employed enzyme is trypsin, which cuts the amino acid sequence after an arginine and lysine residue generating peptides with optimal length and charge for mass spectrometry. In order to decrease the analyte complexity different fractionation techniques are applied prior to loading the samples in a mass spectrometer. In High Performance Liquid Chromatography (HPLC) the analytes of interest are

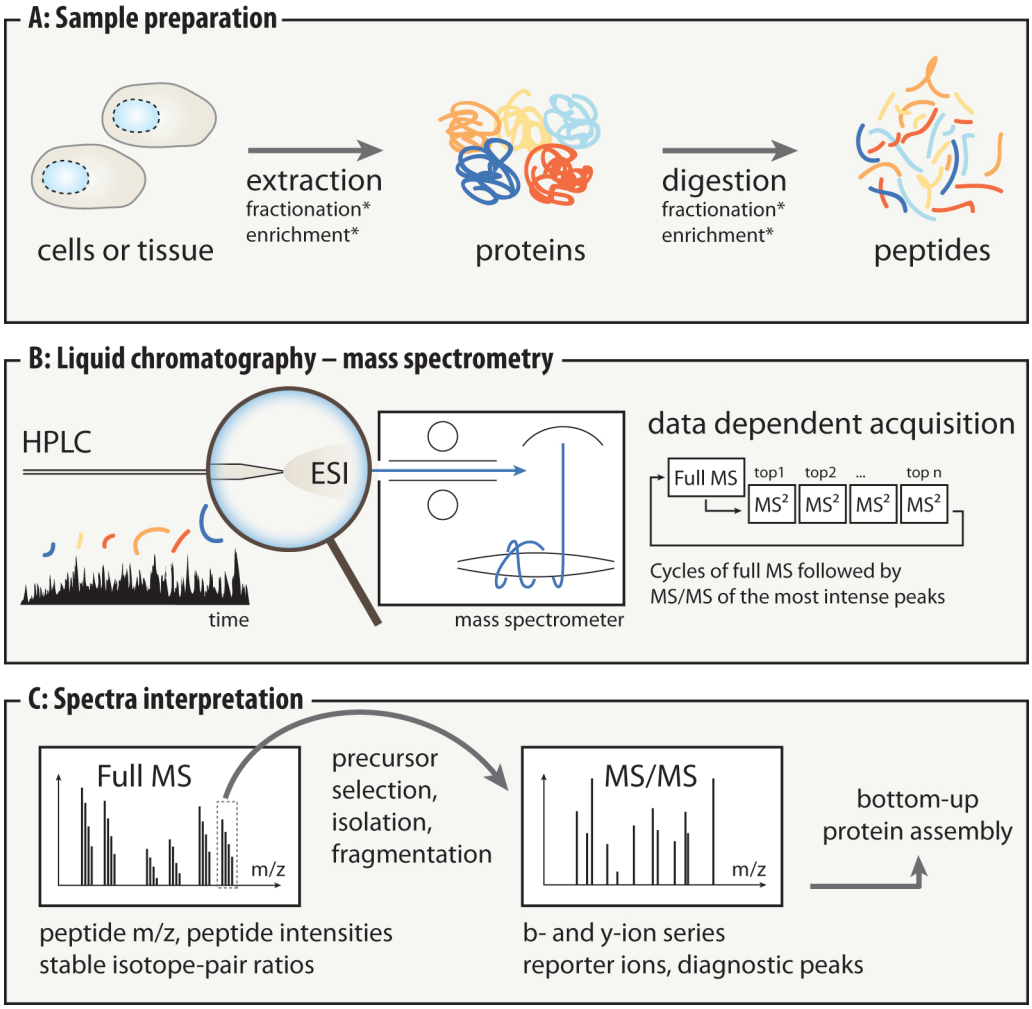


Figure 1.1: ***Shotgun proteomics workflow.*** *An outline of the shotgun proteomics workflow including:* ***A)*** *sample preparation,* ***B)*** *liquid chromatography separation and ionization and* ***C)*** *mass spectrometry analysis. Adapted with permission. from: Ḧandbook of systems biology: Concepts and Insights.; 2013, [14]. Copyright 2013 by the Academic Press.*

loaded on a column packed with specific chromatographic material (forming the stationary phase). In a standard reverse-phase HPLC format, the peptides are separated during their migration based on selective hydrophobic interaction affinities. This is accomplished by a gradient of organic solvent that elutes the peptides in order of hydrophobicity.

The main principle of peptide mass spectrometry is the dependence of the trajec-

tory of charged particles in electromagnetic fields on their mass-to-charge ratio. Thus prior to entering a mass spectrometer the analyte is transferred to the gaseous state and ionized. A wide-spread technique for converting peptides to gas phase ions is the Nobel prize winning development of the electrospray ionization technique [15]. It allows direct coupling to the effluent of the column. As the solvent evaporates from an electrosprayed droplet that has a net electric charge, the density of the charges on the droplet surface increases. After a critical mass is reached, the droplet splits into numerous offspring droplets, from which peptide ions are generated. These are injected into the mass spectrometer, where every few seconds the entire mass range is scanned. Several precursors are then isolated for fragmentation and mass spectra of their fragments are acquired. The resulting MS/MS spectra are then used in a database search to retrieve the corresponding amino acid sequences.

1.1.2 Quantification

Gaining in depth understanding of the molecular function and regulation of proteins requires quantitative information [16, 17]. Quantitative proteomics comes in two main forms: absolute and relative [18]. In absolute quantification the amount of the substance of interest is determined; for example, the exact quantity of a given biomarker in a sample in ng/mL or the copy number of a protein per cell. The absolute amount of a protein can be quantified using spike-in standards or isotopically-labeled proteins of known quantity. In relative quantification the amount of a substance is determined with respect to a measurement of another instance of the same substance, for instance, fold change in protein abundance as a result of a system perturbation.

The methods for quantification can be grouped into stable isotope labeling and label-free methods. Label-free methods employ information such as the number of peptide-identifying spectra or the summed intensities of peptide ions for a given protein. The understanding and application of the dependence of the protein abundance on the number of acquired spectra [19] has gradually evolved to utilizing a measure that fairly accurately characterizes low abundant proteins - the absolute protein expression index [20]. In other major label-free approaches the averaged normalized ion intensities of each or of the top three identified ion peptide are used for quantification [21, 22, 23].

Based on the way in which the proteins or peptides are isotopically labeled, several groups of labeling methods can be distinguished: (i) incorporation during the enzyme digestion [24, 25, 26], (ii) addition of an isotopically-labeled tag [27, 28, 29], (iii) introduction of a labeled spike-in standard of known quantity [30, 31] and (iv) metabolic labeling [32, 33, 34].

Due to its high reproducibility and high accuracy differential stable isotope labeling has become a widely-used method for quantification. It employs stable non-radioactive isotopes of amino acids and can be used both for absolute (used as a spike-in standard) and relative quantification. The method relies on the fact that the physicochemical properties of the isotopes and consequently their ionization efficiencies remain unchanged. The isotopic forms of an MS/MS identified peptide are characterized by identical elution profiles and a specific mass shift corresponding exactly to the mass differences in the labels. Upon detection of an isotopic pair the ratio between the intensity peaks of the heavy and the light version of the peptide is calculated. A major advantage of stable isotope labeling is the ability to measure multiple samples simultaneously, introducing as little experimental noise as possible by treating the samples always together upon labeling. For example, stable isotope labeling by amino acids in cell culture (SILAC) [32, 35, 36] is a labeling technique that has been successfully used in a wide range of proteomics studies [5]. Amino acids with heavy labels, such as arginine bearing six or lysine bearing eight 13C atoms (resulting in 6Da or 8Da mass shifts respectively), are introduced in the growth media of the cells. The labeled precursors become fully incorporated into all cellular proteins during cell growth and protein turnover. Cell populations grown in different media can be used in comparative studies to unravel changes in the biologically-relevant total or sub-proteomes.

1.1.3 Temporal dimension

SILAC can be also used to gain deeper insights into the biological systems by providing a non-static view of various cellular processes. A time course experiment can be set up by differentially labeling several populations of cells and keeping one of them as a control while perturbing the others. Andersen et al. described the changes in the nucleolar proteome over time and in response to three different metabolic inhibitors [37]. In their setup three instances of the same cell line were metabolically

labeled with three versions of arginine and lysine (light, medium and heavy), treated with an inhibitor of transcription and harvested at three time points. In order to gain a more accurate temporal profile the experiment was repeated several times with a common zero time point resulting in up to 9 different time points in total. Information about the temporal changes can be extremely useful in studying the cellular responses to various stimuli during signal transduction, examples of which are discussed later in this thesis.

1.1.4 Large-scale proteomics applications

Proteomics has been successfully employed in a wide range of areas, such as determining the protein composition of complex mixtures, studying the changes in phosphorylation upon a stimulus, determining protein-protein interactions and clinical studies [4]. Importantly, the advances in shotgun proteomics are now beginning to enable the characterization of complete proteomes [38].

Mass spectrometry-based proteomics is ideally suited for studying the protein components of subcellular structures [39, 40]. The main challenges related to the fractionation of the cell and the isolation of the organelle of interest are overcome through enrichment strategies [41] or with the help of protein correlation profiling [42]. The latter approach discriminates between different compartments on the basis of characteristic abundance patterns by proteins from the same compartment exhibited over density centrifugation gradients [43, 44].

Mass spectrometry-based proteomics combined with affinity purification techniques has emerged as a powerful technique for the investigation of molecular interactions [45, 46, 47]. SILAC labeling has been used to quantitatively characterize protein-protein [48, 49], protein-nucleic acids [50] and protein-peptide complexes [51, 52].

Application of mass spectrometry-based proteomics in detection and quantification of post-translational modifications, as well as in clinical studies are discusses in detail in the following sections of this thesis.

1.2 Phosphoproteomics

One of the major advantages of large-scale proteomics studies is the ability to investigate the dynamics of the cellular process at the level of post translational modifications (PTMs). Almost any PTM can be detected by mass spectrometry-based proteomics, including phosphorylation, glycosylation, ubiquitination and acetylation [53]. To date phosphorylation remains the most studied one with thousands of in vivo phosphorylation sites identified, quantified and stored in public databases. It plays an important role in the regulation of a myriad of cellular process, such as cell-cell communication, cell division, apoptosis, and signal transduction [54]. Tasks related to the proteomic analysis of post translational modifications include identification of the modified peptide/protein, localization of the modification site, quantification, assignment of a functional role , as well as possible cooperative interactions among multiple modification sites.

1.2.1 Identification of phopho-sites

Large-scale identification of phospho-sites in mass spectrometry-based experiments is a challenging task as they are often of low abundance, transient and reversible. The problem of phospho-peptides escaping detection and identification by MS analysis can be alleviated by employing enrichment strategies prior to mass spectrometry processing [55]. A plethora of such methods have been developed, among which affinity-and antibody-based ones are most widely used. Immobilized Metal Affinity Chromatography (IMAC) makes use of metal beads packed in a column, which preferentially bind phosphopeptides [56, 57]. The method has been shown to be very efficient with various types of metal ions. Nonetheless, care has to be taken to avoid binding of strongly negatively-charged peptides. A method that is generally characterized by higher affinity is titanium dioxide (TiO2) enrichment [58]. In acidified conditions a column packed with TiO2 can retain organic phosphates, whereas alkaline conditions cause elution. Strong Cation Exchange Chromatography is another powerful technique for phosphopeptide enrichment [59]. As the phosphate group reduces the solution charge of a phosphorylated peptide, its modified and unmodified forms are characterized by distinct solution charge states. These differences allow easy separation by SCX chromatography using a linear salt gradient. Together with hydrophilic interaction liquid chromatography, SCX is used rather as

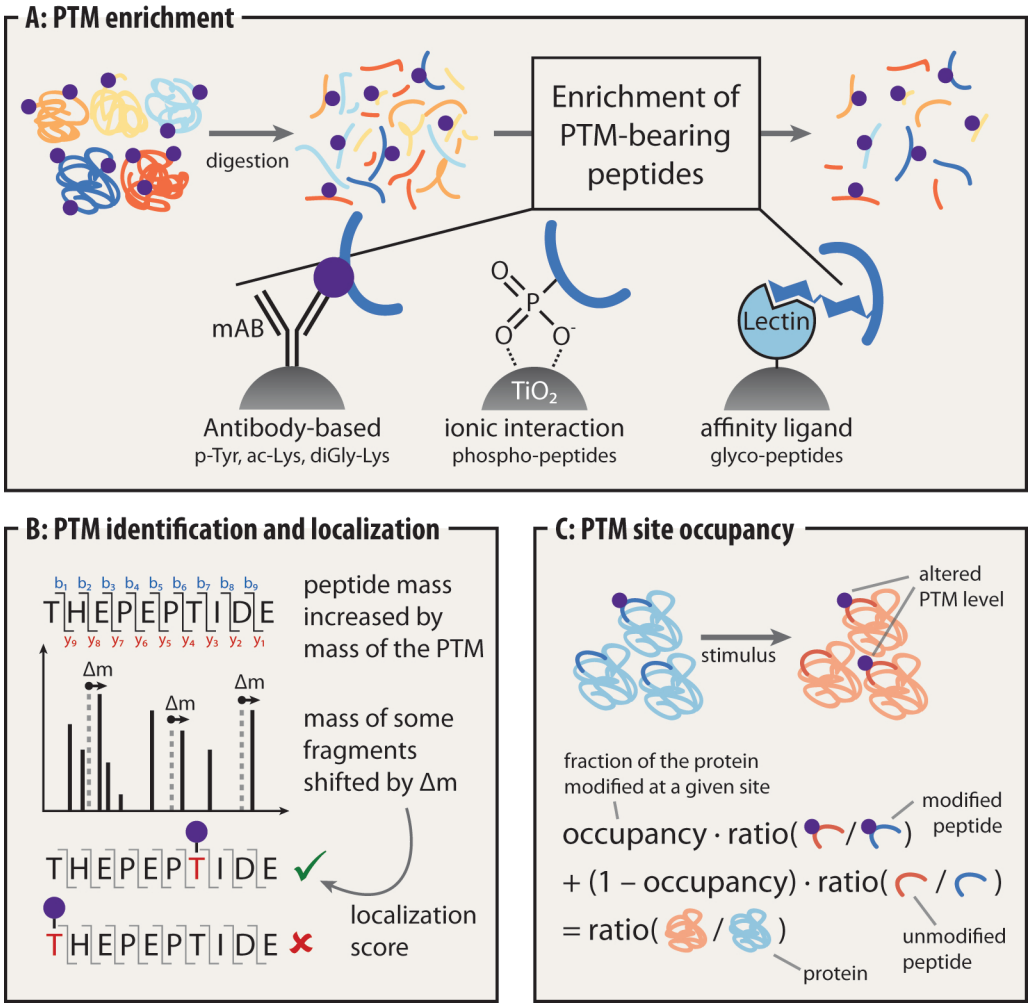


Figure 1.2: ***Mass spectrometry-based analysis of PTMs*** *An outline of the main sample preparation and computational tasks including:* ***A)*** *enrichment techniques,* ***B)*** *PTM identification and quantification and* ***C)*** *PTM quantification in the form of absolute stoichiometry. Adapted with permission from: Ḧandbook of systems biology: Concepts and Insights.; 2013, [14]. Copyright 2013 by the Academic Press.*

a prefractionation technique and is followed by additional enrichment steps. Furthermore, immunoprecipitation with specific antibodies and phosphatase inhibition are successfully applied to the enrichment of phospho-tyrosines [60].

Due to the complexity associated with the characterization of phosphopeptides, specific MS acquisition strategies are required in their detection and identification. The main approaches rely on the detection of either reporter ions or neutral losses. Dur-

ing MS analysis characteristic reporter ions (HPO3- anion with specific mass of 79 m/z) are produced by the phospho-peptides. To detect these ions the mass spectrometer must be set to negative ion mode. However, as MS/MS spectra in negative ion mode are of low quality, a switch to positive ion mode is required for sequencing. This is not the case for the precursor ion produced by phosphotyrosine (immonium ion, 216.043 m/z) as it can be detected directly in the positive mode. Neutral loss scans can be carried out in positive ion tandem MS and are directly compatible with online HPLC. In Collision Induced/Activated Dissociation (CID or CAD) phosphoserine and phosphothreonine residues generate neutral losses of 98 Da (H3PO4) or 80 Da (HPO3). Unfortunately, phosphotyrosines appear to be more stable, making the neutral loss method less applicable for their identification.

1.2.2 Localization of phospho-sites

Assignment of the phosphorylation to a particular position in the peptide is another challenge. In the case of precursor ion detection, difficulties arise when such a specific fragment ion cannot be efficiently generated or detected by the mass spectrometer. The localization can further be impaired by the presence of multiple potential phosphorylation sites. To address the problem of lack of fragmentation spectra information several computational methods have developed a probability-based score such as the post translational modification (PTM) score [61] and Ascore [62]. The PTM score is derived from information about the phospho-peptides fragments and about the presence and intensity order of diagnostic fragment ions in the MS/MS spectra. The score is computed exhaustively for each possible phosphorylation position or combinations of positions in the peptide and has been further optimized for use with the Andromeda search engine (see below) [63].

1.2.3 Quantification of phosphorylation

Most cellular processes are controlled by gradual changes at the molecular level, therefore to study in depth the functional effects of phosphorylation, quantitative information is needed. Similarly to quantification of unmodified peptides, label-based (chemical - iTRAQ and metabolic labeling - SILAC) and label-free methods can quantify abundance changes of phosphorylated peptides/proteins in response to various stimuli. To accurately assess the changes in phosphorylation the changes

in protein abundance should be taken into account. Furthermore, for many applications absolute stoichiometry is desired (the proportion of a given protein species phosphorylated at a given site at a given time point). One way to obtain this information is to combine data on the ionization efficiencies of a modified peptide and its unmodified version [64]. Olsen et al. developed a method for quantification of the stoichiometry phosphorylation that relies on three measures: the ratios of the two states (e.g. heavy and light labeled) of the modified and of the unmodified versions of a peptide and the protein ratio [65]. In their study occupancy was computed for more than 20,000 sites, showing two peaks of phosphorylation: at mitosis (possibly related to inhibition of various cellular processes) and at S-phase (regulating stress and DNA damage response). An alternative method was described by Wu et al. [66], in which the relative intensities of two differentially-labeled samples – one subjected to phosphatase inhibition and the second without inhibition – were used to obtain stoichiometries.

1.2.4 Temporal dimension of phosphorylation

Another important feature of mass spectrometry-based phospho-proteomics is the ability to add a temporal dimension to the data. Phosphorylation is the major mechanism of signal transduction and facilitates the rapid and robust response of the cell to a stimulus. Therefore having the means to study changes of phosphorylation at different time points can drastically improve our understanding of the cellular processes. Blagoev et al. and Olsen et al. have successfully studied changes in the phosphoproteome upon stimulation with the epidermal growth factor (EGF) in a time-resolved manner [67, 61, 65]. In both studies SILAC differentially-labeled cell populations were employed, using one of the states as a reference. In study of Olsen et al. three different versions of arginine and lysine residues were used, resulting in quantitative data for six time points over the cell division cycle.

1.2.5 Phosphorylation and protein structure

More than 500 kinases, encompassing about 2% of the genome of multiple organisms, are known to catalyze phosphorylation reactions [68]. Kinases are characterized by high sequence and structure conservation and can be divided into serine/threonine, tyrosine and dual-specificity (i.e. capable of modifying all 3 residues) kinases [69, 70].

The strong conservation of the catalytic region of different kinases together with the large number of serine, threonine and tyrosine residues in a typical eukaryotic cell necessitate specific mechanisms to ensure accurate substrate recognition [71]. Various in vivo experiments have demonstrated the presence of short consensus motifs on specific kinase substrates [72]. These are usually 3 to 11 residues long, found in protein regions that lack defined tertiary structure and evolve much faster than highly conserved domains [73]. Linear motifs are associated with various regulatory functions such as serving as binding regions for multiple domains (e.g. WW, SH2, SH3, PTB, 14-3-3), signal transduction and protein trafficking. Three major classes of kinases can be distinguished based on the amino acid content of their consensus motifs: basophilic, acidophilic and proline-directed. Cyclic AMP-dependent protein kinase A (PKA) is a representative of the basophilic class as positively-charged residues at positions preceding the serine/threonine residue determine the substrate specificity [74]. Proline-directed kinases require a proline residue at position +1 relative to the phosphorylation site. In the case of CDK2, for example, the substrate binding site is characterized by a specific conformation that explicitly favors a proline residue in order to satisfy a hydrogen bond from the nitrogen atom in the main chain of the substrate [75].

Although consensus motifs have been widely accepted as one of the major mechanisms underlying kinase specificity, the rules for recognition are not completely understood. For example, the presence of a certain recognition motif does not always lead to phosphorylation by the expected kinase. Furthermore, some kinases recognize linear motifs different from their consensus. Additionally, the recognition motifs of some kinase families show substantial overlap, suggesting the existence of other mechanisms facilitating substrate specificity [71]. In general, the linear consensus motifs describe the primary structure around phosphorylation sites and thus do not adequately account for all factors that may contribute to the kinase specificity. Secondary and tertiary protein structures may reveal additional recognition mechanisms, such as distant residues that lie close in space. Additional linear motifs that can be situated at a larger distance from the phosphorylation site may further increase the substrate concentration in the proximity of the kinase. Similarly, priming phosphorylation may be required to provide efficient docking interactions. Furthermore, targeting subunits and scaffolds (molecules and proteins that bind kinases and

contain domains that interact with kinase substrates) also contribute to the kinase specificity and diversify the set of sites that a kinase can modify. For instance, cyclins are targeting subunits that not only activate CDK kinases, but also contain docking domains, enabling recruiting of substrates to the kinase [75].

Another important factor influencing kinase specificity is subcellular localization. The presence of various kinases in different subcellular compartments allows better control over the concentration of the correct substrates for a given kinase and reduction of the overall number of substrates to which a kinase has access. Ubersax and Ferrell demonstrated the influence of the depth and the hydrophobiciry of the catalytic cleft, as well as the sequence complementarity between kinases and substrates on the specificity and the binding energy [71]. It may, however, turn out that it is the combination of several such mechanisms that regulates the proper substrate targeting. Alexander et al. showed how localization and sequence motifs act together to achieve kinase specificity in mitosis [76]. In particular, kinases with overlapping localization exhibited different motif preferences, whereas kinases with similar motif preferences were characterized by different localizations.

Studying the structural properties of phosphorylation sites helps to increase our knowledge of the process of phosphorylation and to gain further understanding of kinase specificity mechanisms. Multiple resources for storing structural information about phosphorylation sites are now available. Phospho3D and Phospho3D 2.0 are online applications that store and display structural information such as the local structure of ten residues-long phospho-regions and spatial regions encompassing all residues within 12Ådistance from the phospho-site [77, 78]. Furthermore, given a protein structure as input, phosphorylation sites in that protein are predicted by comparison to the structural motifs present in the database. Another effort to systematize structural data on phosphorylation is the mtcPTM database [79]. A separate web page for each protein combines information on different sites from multiple experiments, allowing for direct comparison between the phosphorylation profiles under different conditions.

Analysis of the structural properties of phospho-sites in available PDB structures and homology models revealed that they appear with higher frequencies in regions between or at the termini of structured domains, are in most cases solvent acces-

sible and less conserved than expected. In their study Durek et al. confirmed the preference of modification sites to appear predominantly in flexible regions that lack defined structure and investigated the amino acid propensity distributions in both the sequential and spatial surrounding of phospho-sites [80]. A kinase-specific analysis revealed enrichment of residues explicitly in the 3D surroundings, illustrating possible presence of specificity determinants encoded at the structural level.

Overall all studies agree on several structural properties as characteristic for phosphorylation sites: phosphorylation occurs predominantly at regions that lack defined structure and at irregular secondary structures [81, 80, 82, 83]. Specifically, Iakoucheva et al. found a strong similarity in the sequence complexity, amino acid composition and flexibility properties between phospho-sites environments and intrinsically disordered sequences. Aromatic residues were depleted, whereas the net-charged in the surrounding regions was high. In agreement with the expectation that phospho-acceptor sites need to be accessible for the modifying enzyme, the majority of them were characterized by high solvent exposure. Modification sites that appeared to be buried in the protein may be explained by inaccurate structure models or they could be associated with conformational changes that lead to residue exposure. Furthermore, these sites were found in regions with high flexibility, characterized by large B-factors (a measure of the flexibility of an amino acid) and missing electron density.

Using protein structures that contained the phosphorylated residues with the phosphate group attached, a distinction of the distribution of charged residues surrounding phospho-sites and non-modified residues has been reported [84]. The proximal charged amino acids stabilizes the phosphate group by favorable electrostatic interactions. These interactions have implications in phospho-peptide binding in signal transduction [85], in complexes formation [86] and in the regulation of activation loops [87].

Linear motifs and intrinsically disordered regions characterize the majority of phosphorylation sites, suggesting only a limited relevance of structurally-defined kinase binding regions. However, Plewczynski et al. reported that about 60% of all phosphorylation sites modified by protein kinases A and C (PKA, PKC) reside within alpha helices [88]. Furthermore, buried sites and sites present in well characterized

structures are not uncommon, implying that structural surface features and motifs may play a role in kinase recognition and specificity [89].

Structural properties have successfully been used in the development of computational methods for prediction of phosphorylation sites and such prediction tools have shown an increase in performance over those that rely only on sequential information [90, 91, 92, 93]. NetPhos utilizes contact maps computed from the sequence surrounding phosphorylated and non-phosphorylated residues in the training of artificial neural networks [83]. DISPHOS is a predictor incorporating disorder information into a combination of logistic regression models in order to discriminate between phospho-and reference sites [81]. Phos3D, PHOSIDA and the method developed by Plewczynski et al. make use of structural properties of phosphorylation sites such as secondary structure and disorder features, conservation, hydrophobicity and charge distribution and local structure segments in the training of Support Vector Machines (SVMs).

1.2.6 Functional annotation of phosphorylation sites

The advances in mass spectrometry-based proteomics have enabled the identification of thousands of phosphorylation sites, which are now stored and made available through various web-based databases [94, 95, 96, 97], for a comprehensive review see Hjerrild et al. [98]. Phosphorylation is involved in the regulation of a large number of cellular processes including cell growth, cell division, cell-cell communication, signal transduction, localization and apoptosis [99, 100, 101, 102]. However, unraveling the function of most of the phosphorylation events remains a challenging task.

Furthermore, this modification plays an important role in the regulation of protein-protein interactions [103]. Nishi et al. demonstrated a tendency of phosphorylation towards transient complexes and estimated relatively small changes in the corresponding binding energies. The additional phosphate group can lead both to activation or inhibition of an enzyme [104, 105]. At complex interfaces phosphorylation sites can promote or disrupt interactions through electrostatic or steric effects without major structural rearrangements.

Phosphorylation is involved in modification of the protein function by inducing conformational changes or by creating or eliminating binding sites [106]. The influence

of phosphorylation and other post-translational modifications on protein-protein interactions has been systematized in a domain-centric view. Seet et al. categorized modification-dependent mechanisms for domain recognition and binding and provided examples of how these regulate various cellular processes [106]. The different interaction categories include but are not limited to: (i) cooperative interactions requiring the modification of several sites in order to achieve the desired function, (ii) sequential interactions, in which initial phosphorylation of a particular site is required for a subsequent modification or conformational change and interaction, (iii) antagonistic PTMs, encompassing modification sites that prevent the interactions between adjacent sites and their binding partners.

The large number of identified phosphorylation sites with no functional annotations and the limited understanding of kinase specificity have provoked a dispute over the existence of non-functional sites [107]. Such idea finds support from evolutionary perspective, as phosphorylation sites without functions would in most of the cases cause no harm and thus would not have to be eliminated. Furthermore, functionally-neutral phosphorylation events may influence the ability of signaling networks to evolve [108]. It is estimated that only around 1% of the cellular ATP would be used up for non-functional phosphorylation, thus making the hypothesis of silent phosphorylation even more plausible. As phosphorylation often occurs on multiple sites in a protein, functional redundancy among such sites is also possible. In their study Wang et al. described a computational model of the role of multiple nonessential sites in enhancing a switch-like regulation of proteins in response to a stimulus [109]. The authors argued that a response that has both ultrasensitivity and high threshold can be achieved only when the proportion of nonessential phosphorylated sites to the total number of modification sites is optimal.

Important indicators for the functionality of a phosphorylation site are its stoichiometry and evolutionary conservation [110]. Levy et al hypothesized that highly abundant proteins are more likely to encounter random kinase interactions, which may thus lead to an on average larger number of phosphorylation sites on highly abundant proteins. Such sites were characterized by evolutionary pressure that was not higher than that of non-phosphorylated reference sites and were of low stoichiometry [111]. Nonetheless, the argument that low abundance of phosphorylation is indicative of an off-target phosphorylation and is therefore more likely to result

in non-functional sites should be considered with caution. For example, in signaling cascades only few copies of an activated protein, possibly in a specific cellular localization, are sufficient to achieve the desired response.

There has been a long-standing debate on the conservation of phosphorylation sites, accompanied by conclusions ranging from no or weak evolutionary pressure [112] to statistically significant conservation [44, 113, 114]. Currently the most widely-accepted opinion is that phosphorylation sites are characterized by an overall low level of conservation. Although for some of the sites their low conservation may reflect the functional difference between orthologous organisms, in others these non-conserved sites may indeed lack a function. In any case, modification sites, which have a characterized function, appeared to be more conserved on average than sites without functional annotation [112, 115]. Recent studies suggested that secondary structure, stoichiometry and protein abundance need to be taken into consideration to accurately estimate the level of conservation [112, 116]. The structural context of phosphorylation sites has to be considered as a control for background effects such as quickly evolving intrinsically disordered regions.

1.3 Clinical proteomics

An important focus area in proteomics is the development of sample preparation techniques and analysis methods applicable to clinical research. Investigating the proteome profiles of clinical samples poses major challenges both with respect to the experimental set-up and the analytical tools that are needed. Despite the difficulties that oncoproteomics is still facing, numerous studies have already made first steps in addressing vital clinical questions [117, 118, 119, 120, 121]. The range of cancer types studied so far includes but is not limited to: breast [122], prostate [123], ovarian [124], lung [125], colon [126], lymphoma [127], head and neck [66], and liver [128] cancer.

1.3.1 Sample preparation techniques

Clinical proteomics can be applied to a wide range of samples including body fluids – most importantly the plasma part of the blood – and tissues and each is associated with specific challenges. For example, the large dynamic range of the body fluid pro-

teomes has often necessitated targeted techniques such as MRM ([129]) or antibody enrichment ([130]). Tissue samples from patients offer a more feasible system for unbiased, in-depth proteomic analysis [131, 132, 133] and furthermore greatly benefit from advances in sample preparation such as the Filter Aided Sample Preparation (FASP) protocol [134].

1.3.2 Quantification with super-SILAC

Quantification of the tissue proteome can be achieved by using SILAC-labeled cell lines as an internal standard, thus avoiding the need for labeling the tissue itself. As patient samples are very heterogeneous, the proteome of a single cell line often does not account for the complete variability and may lead to inaccurate quantification. A new technique – super-SILAC – that circumvents this limitation was successfully used in clinical proteomics studies [135]. It uses a mix of isotopically-labeled cell lines that more adequately represents different tumor stages or subtypes, thereby allowing accurate quantification. In their paper Geiger et al. demonstrated the application of the technique to achieve deep proteome coverage of tumor tissues and accurate quantification ratios between replicates.

1.3.3 Analysis of clinical proteomics data

The application of advanced analysis techniques to oncoproteomics data is becoming increasingly popular [136, 137]. Many clinical proteomics studies directly utilize MS features in their multivariate analysis [138]. For example, Han et al. developed a feature selection method based on multi-resolution independent component analysis and combined this with powerful machine learning techniques such as linear discriminant analysis and support vector machines [139]. The authors argued that taking feature frequencies into consideration increases the reliability of the results. Adam et al. used decision trees in the classification of normal, benign hyperplasia prostate and prostate cancer samples [140]. They reported 96% accuracy with 9 selected protein-mass patterns, suggesting that multiple biomarkers are required to overcome the problem of tumor heterogeneity. In another study, artificial neural networks were applied to distinguish between two grades of astroglial tumor [141]. The authors identified two ions whose relative intensity patterns strongly discriminated between the grades. Feature selection based on principal component analysis

(PCA) combined with support vector machines were employed in a protein profiling study of the differences between Parkinson's and multiple system atrophy [142]. Marchiori et al. demonstrated the importance of feature selection techniques for handling noisy data, improving the classification accuracy and building biologically relevant models [143]. Furthermore, their article provided a discussion on a list of clinical proteomics studies making use of different machine learning techniques in combination with various feature reduction methods.

As it has only recently become possible to quantitatively characterize the protein content of patient tissues to a substantial depth, the number of clinical proteome profiling studies is still limited. Generally, due to their smaller variability cell lines provide an easier system to work with than patient samples and enable accurate segregation of different cancer subtypes. In a study of the progression of estrogen receptor-negative breast cancer tumors Geiger et al. correctly clustered the different cell lines according to the tumor stage from which they were derived based on their proteome profiles [122]. Similarly, the accurate quantification of more than 7,500 proteins allowed perfect segregation of the activated B-cell-like from the germinal-center B-cell-like lymphoma subtypes using PCA [127]. In a label-free approach Wisniewski et al. identified and quantified more than 7,500 proteins between healthy, primary carcinoma and nodal metastasis tissues [144]. Hierarchical clustering of the samples based on the protein expression patterns revealed notable differences in the proteome profiles between healthy and tumor tissues, but was unable to draw a clear distinction between primary tumors and metastatic samples.

Overall, the current results of multivariate analyses of clinical proteomics data suggest that to efficiently account for the heterogeneity of patient samples and the complexity of various diseases, such as cancer, a set of biomarkers (a signature) rather than single markers must be investigated. Surprisingly, independent studies of the same biological system often show small overlap of the putative biomarkers. Indeed, most of the identified biomarkers are never validated and are of little or no clinical use [145, 146, 147, 148]. The reasons explaining these phenomena are multiplex. On one hand, the differences in the sample preparation, instrument handling and performance strongly influence the data. On the other hand, often the lack of understanding of the nature of the data or improper use of statistical tests results in biologically-irrelevant features being selected and in an overestimation of

the prediction accuracy. Chapter 4 of this thesis focuses on the problem of accurate and efficient ways of feature selection and correct estimation of the generalizability of a model.

1.4 Large scale proteomics data analysis

1.4.1 Processing of raw mass spectrometry data

With the increasing ability to analyse complex biological samples and to generate large-scale datasets with the help of mass spectrometry-based proteomics, the need for adequate computational tools to process these data has grown dramatically. Various tools and frameworks have been developed to address tasks of peptide and consequently protein identification and quantification from the raw MS data [149, 150, 6, 151, 152, 153, 154, 155]. The computational proteomics tasks include but are not limited to: spectrum identification (extraction of peaks from raw data), matching MS/MS spectra against a sequence database and accurate protein quantification . The MaxQuant suite of algorithms is a state-of-the art tool, designed to analyse high-resolution quantitative mass spectrometry data [6]. Our general computational workflow is depicted in Figure 1.3. In a feature detection step MaxQuant first assembles the two dimensional signals formed in the mass-intensity plane over the retention time into three dimensional peak hills. The list of these features is then reduced by identification of isotope patterns using graph theory. High mass accuracy is achieved by weighted averaging and through mass recalibration by subtracting the determined systematic mass error (a function of the m/z value and the retention time [8]) from the measured mass of each MS isotope pattern. In the peptide identification step peptide fragmentation spectra are matched against a sequence database with the Andromeda search engine [63]. The scoring function used in the matching employs a binomial distribution probability formula. The experimentally determined tandem spectra are tested against theoretical fragment ions from database peptides, including reverse sequences and contaminants [156, 157].
MaxQuant also provides quantification of experiments using both label-based and label-free techniques. In SILAC-based experiments protein ratios are computed as the median of all SILAC pairs ratios of peptides associated with this protein. Detection of SILAC pairs is based on high correlation of the intensity of the two isotope

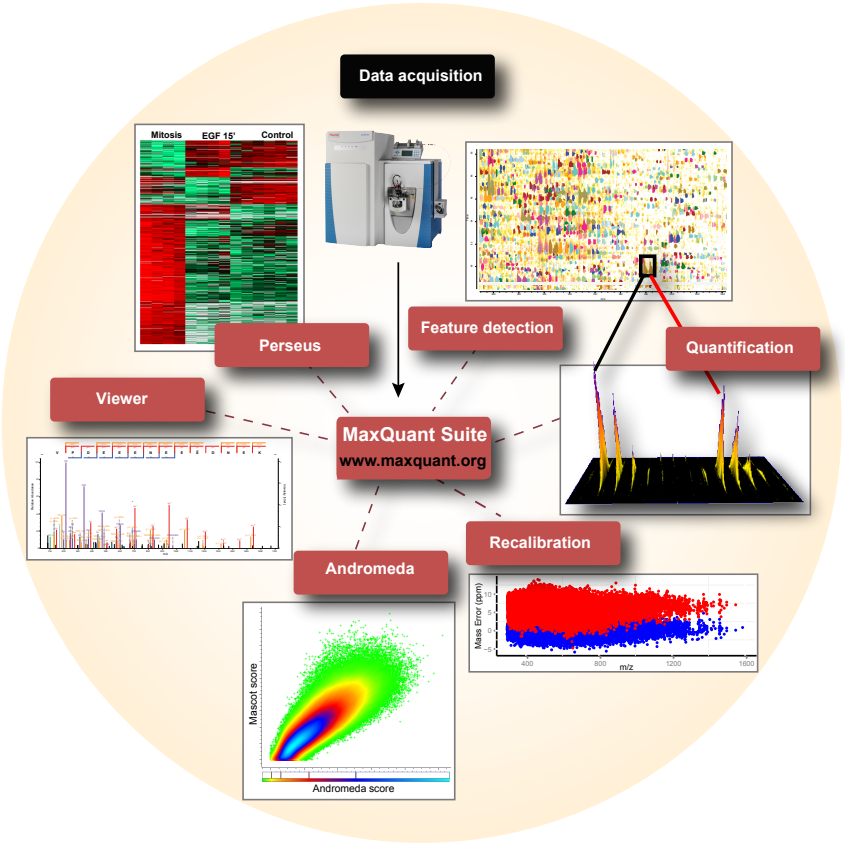


Figure 1.3: *Overview of the independent modules of the MaxQuant suite.*

patterns and agreement with the expected mass shift resulting from the mass difference between the heavy and light residue isotopes. Label-free quantification combines signal information of the same peptide in different experiments. The isotope patterns are matched across runs using peptide identifications, high mass accuracy and nonlinearly remapped retention times. The preprocessing of the raw files results in qualitative and quantitative information about the proteins in the given samples that require further downstream functional analysis. Large scale discovery studies are intended to identify as many proteins as possible, whereas comparative studies are often employed to target a specific biological question.

1.4.2 Downstream analysis of proteomics data

The ability to study biological systems under different conditions has been considered as one of the major advances in the field of mass spectrometry-based proteomics

[5]. The main goal of proteome profiling studies is finding proteomic patterns that discriminate between different biological states. The comparison of large-scale proteomics datasets harbors many analytical challenges, such as experimental noise, systematic variation between the different measurements, and large feature space. Different statistical procedures have been developed and are used routinely in the pre-processing and functional analysis of quantitative proteomics data.

The commonly used methods can be combined in several large groups: (i) differential analysis, such as t-test-based statistics, analysis of variance, (ii) unsupervised learning methods, such as hierarchical clustering, k-means clustering, and (iii) supervised learning methods: principal component analysis, linear discriminant analysis, k-nearest neighbors, support vector machines. Briefly, differential analysis employs statistical tests comparing the means of the different groups and aims at the identification of features that show significant difference between these groups. Unsupervised learning is used to find patterns in the data, enabling the distinction of different groups. In contrast, supervised learning makes use of the information about the known group identity of the samples and discovers rules to be used in the classification of new unlabeled instances.

1.4.3 Unsupervised learning techniques

Means-based methods are widely applied in differential analyses. Numerous omics studies have successfully employed t-test and ANOVA methods to identify proteins that significantly differ between groups. The t-statistics overcome the drawback of threshold-based methods by taking into account the possibly different variances between the features. One of the limitations of such tests is related to the requirement for normality of the data (i.e. the compared groups have to follow normal distribution in order for the test to be applicable). Additionally, t-test-based statistics evaluate the importance of each feature separately and thus often fail to detect groups of features that have a high relevance to the biological question of interest.

In cluster analysis, samples (and features) are grouped together in classes based on their similarity (e.g. protein expression levels). Clustering techniques are used to unravel general patterns in the data, which may serve for hypothesis generation. In hierarchical clustering, clusters can be combined or split (i.e. can be agglomer-

ative or divisive) using a metric of distance, such as Euclidean distance, between the feature vectors of the samples. Alternatively, the cluster definition can be based on the correlation between the expression profiles of the samples. For instance, a member of the centroid-based clustering methods, the k-means approach redefines the clusters through an iterative recalculation of the 'k' centroids.

Commonly used in microarray analysis [158], due to their relative algorithmic simplicity and suitability for visualization these techniques have also become a standard tool in the profiling of proteomics data [159]. Unfortunately their success decreases with increasing data complexity. For examples, studying patient samples appears to be a much more challenging task than profiling of cell lines, due to the high genetic variability of the former. As a consequence of the genetic variability the biological signal of interest is often weakened by other unrelated signals and the identification of the correct classes can be severely impaired.

1.4.4 Supervised learning techniques

The limitations of the above-described statistical analysis methods are to a large extent surmounted by the more advanced supervised learning techniques. Unlike unsupervised methods, supervised techniques incorporate the information about the class identity of each sample. The superiority of such methods to both statistical and unsupervised approaches has been shown in numerous studies [160]. However, a possible problem known as "overfitting" can arise in many situations. Overfitting results in a classifier that has a high performance in the given training set but is characterized by low generalizability (i.e. has a poor performance in independent test sets). Such situations are common in biological studies due to the combination of a large number of features with a comparatively limited number of samples. The remedy for such problems is dimensionality reduction, which can be achieved through efficient feature selection. It has been demonstrated that the performance of machine learning methods improves when they are combined with an appropriate feature selection method as irrelevant features are discarded and the noise is reduced [161]. However, bias in the feature selection can have a large influence on the results and their functional interpretation. As a consequence of improper feature scaling or sampling variation, certain features may be selected that are not relevant for the subject of interest.

Feature selection methods can be grouped in rank-based and rank-free categories. In the former features are ranked and sorted according to a certain score. The score can be computed based on a correlation measure (e.g. t test-based statistics measuring the signal to noise ratio) or on its significance (i.e. t-test p-value). Information gain has also been employed in feature ranking procedures [161]. Alternatively, the classifier's weights derived during its training can be used for the score computation [162]. The performance of a classifier also depends on the manner with which the feature selection method and the classifier are combined. The three main approaches that can be distinguished are (i) wrapper, (ii) filter and (iii) embedded and they are discussed later in the thesis.

Support vector machines are a particular class of supervised methods that are well-suited for analysis of data in high dimensional feature space, they are computationally-efficient and capable of detecting biologically-relevant signals [163, 164]. The training of an SVM classifier results in the determination of an optimal hyperplane that separates two classes. Each new sample is then classified according to its position relative to this hyperplane. The method is robust and efficient due to the use of dot products, which allows efficient computations in high dimensional space, handling of linearly-nonseparable cases and generalizability to multi-class problems. The properties and advantages of support vector machines and their application to large scale proteomics data are described in greater detail in Chapters 4 and 5 of this thesis.

1.5 Thesis motivation and organization

The above-described advances and developments in the field of mass spectrometry-based proteomics have led to a vast increase in the amount of acquired data. This thesis introduces the challenges that arise in the interpretation of these data (see Section 1: Introduction) and presents different analytical techniques and tools to address them. The two main parts are organized as follows: In part 1, analysis of the structural features of phosphorylation sites and their tendency to cluster and cross-talk with other post-translational modifications (Chapter 2), which serves as a basis for an in-depth study of the interplay between phosphorylation regulation and protein disorder (Chapter 3). Part 2 describes development of a framework for comparative analysis of proteome profiles of cancer patients (Chapter 4) in order

to achieve better subtype classification and to discover novel potential biomarkers (Chapter 5).

Part 1: Chapter 2 deals with the analysis of various properties of phosphorylation sites, such as preference for a particular structural environment, solvent accessibility and relation to domain regions. These features are then used to investigate a possible distinction between phosphorylation sites of regulatory and unknown functions. Additionally, the tendency of phospho-sites to cluster and their participation in cross-talk with other post-translational modifications are discussed in regard to the newly emerging view of a PTM code.

The large number of identified sites has opened many questions concerning their functional relevance and the possibility that the majority of them may result from unspecific kinase actions. Mapping the phosphorylation sites to particular biological contexts provides insights into their possible functional relevance. On one hand, phosphorylation sites are characterized by global features enhancing their proper functioning and regulation. On the other hand, modification sites with defined regulatory roles appear to exhibit distinct properties. Lower solvent accessibility and preferences for particular type of disordered regions, together with high evolutionary conservation deepen our understanding of the undelying machinery of signal transduction and suggest the existence of specialized mechanisms of action of regulatory phosphorylation sites. Furthermore, these may lay the foundations for classification of the functional relevance of the numerous phosphorylation events in large-scale data sets.

The regulation required to produce a robust and rapid response to various stimuli is achieved through a complex interplay between (i) multiple phosphorylation sites and (ii) phosphorylation sites and other post-translational modifications. The majority of phospho-proteins are found to contain more than one phosphorylated residue. Moreover, these sites are not randomly distributed over the entire length of the protein, but show preferences for significantly smaller distances. Similarly, modified lysine residues are found to be enriched in the proximity of phosphorylation sites. The interplay between the two modifications appear to be most prominent in the case of phospho-tyrosine sites, resulting also in the statistically significant enrichment of various GO-terms among the proteins hosting this pair of residues.

Chapter 3 presents a study of the structural properties in relation to the phosphorylation variation over the cell division cycle. It demonstrates how the structural context provides an additional regulatory mechanism. Regions with defined structure limit the number of possible phosphorylation sites, as these contribute towards a more disordered environment. Furthermore, due to the conservational properties of ordered regions, sites in such an environment are also characterized by a higher evolutionary pressure. The charged residues in the flanking regions of sites with low phosphorylation variability provide favorable electrostatic interactions with the phosphate group and add to the overall charge distribution at protein-protein interfaces. In contrast, disordered regions are often enriched in multiple phosphorylation sites. The higher phosphorylation variability of these sites is related to their regulatory function, which is further supported by the association of this group of sites to proline-related kinases (i.e. kinases with consensus motifs that contain a proline residue at a position within the close sequential proximity of the modification site). The functional and mechanistic role of proline as an important building element of regulatory motifs and its connection to the highly regulated sites in the cell cycle data are discussed in detail in that chapter. Next, the kinase preferences for specific phosphorylation variation and disorderedness level are combined to yield a reconstruction of functional classes of kinases, enhancing the discovery of additional relations and properties of kinases that were previously not known. The last part of the section focuses on the possible existence of a different time scale of phosphorylation, which may act in a conjunction with the level of phosphorylation variation and the structural context of the sites. This phenomenon may be further related to the diverse functional roles of the modification events, which contribute to a different extent to the structural stability, catalytic activation or inactivation or the interaction potential of the modified protein.

Part 2: Despite the developments and efforts in the area of cancer diagnosis and treatment there is still large room and need for improvement. Biomarkers, molecules that are produced by the cancer or by the organism as a cause or consequence of the cancer, have been extensively used in the assignment of a particular treatment and the assessment of the response to that treatment. Furthermore, their detection in screening tests in abnormal quantities is interpreted as an indicator of an early-stage cancer. However, there are serious limitations associated with the application

of the currently available biomarkers for diagnostics, especially in the latter case. For example, elevated level of the prostate-specific antigen (PSA) in the blood can either signify prostate cancer or be caused by benign prostatic hyperplasia. Due to the relatively low specificity of the this marker a large number of men who exhibit elevated levels of PSA are falsely diagnosed with prostate cancer, causing physiological distress, invasive treatment such as radiation therapy or surgery, which is often unnecessary but frequently results in bodily disfunction [165]. Similarly, alarming studies report that about 30% of women who undergo mammography screening and test positive represent overdiagnosis, often leading to overtreatment [166]. Some research estimates the accuracy rate of diagnosis among pathologist to range between 58 to 74% [167].

Therefore the second part of this thesis (Chapters 4 & 5) is motivated by the need for improvement in this area and it aims at offering analytical tools for alleviating some of these problems. The proteomics view of the cell can reflect the underlying processes very accurately by looking at and measuring the actual amount of expressed molecules and their post-translational modifications. Thus proteomics could enable scientists to unravel important mechanisms and gain better understanding of a large number of complex diseases. This can be achieved by comparative studies of the proteome profiles of patients suffering from different subtypes of a given diseases, being at different stages of the disease, or studying different responses to a certain medicine. The ultimate goal is the identification of differentially expressed proteins, whose expression level is directly related to or caused by the disease onset.

As in every other discipline important lessons can be learned from history. In the course of analysis of complex microarray data numerous possible biomarkers have been identified and reported. The majority of them, however, did not become recognized diagnostic markers or drug targets, but instead now appear to be artifacts of poor sample preparation and technical noise, combined with low sample size and to a large extent incorrect data handling [147, 148]. This thesis demonstrates how the last issue can be overcome by combining powerful machine learning techniques with properly-conducted feature selection procedures (i.e. always embedded in an external cross validation procedure see Chapter 4), especially in the case of small data sets.

Furthermore, it is now increasingly accepted that a single biomarker is not very likely to be a strong discriminator between different populations of patients. Instead, the complexity of the majority of diseases, such as cancer, necessitates the identification of a set of proteins to be used as an informative signature. Therefore, analytical techniques that are suited for the assessment of the predictive power of a group of proteins should yield biologically more relevant and thus more useful results.

Chapter 4 of this thesis focuses predominantly on the computational methods and tools, whereas Chapter 5 demonstrates the practical application of the developed framework in the analysis on two real data sets. This work shows that despite the strong genetic variability and the overall low sample size, the detection of disease-related proteins from proteomics profiles of patients is feasible. However, adequate techniques, such as supervised learning, embedded in cross validation procedures and combined with efficient feature selection methods are required.

Chapter 2

Properties of Phosphorylation sites and Functional Inference

2.1 Introduction

With tens of thousands of sites identified, phosphorylation is the most studied post translational modification to date. Similarly to other examples of fields characterized by rapid technological developments in the past, the functional interpretation and assignment of biological context to the measured data is lagging behind. Studies attempting to address functional relevance of the high-throughput sites often employ conservational information. This approach has had some intermediate success, exemplified by the controversial view on conservation due to the preference of these sites for rapidly evolving unstructured regions. Nonetheless, the intricate regulation of numerous cellular processes, which is achieved through phosphorylation, suggests the existence of well-defined mechanisms governing the action of kinases.

Such mechanisms can be encoded at the structural level and can be related to the degree of disorderedness of the local and global environment of a phospho-site, to its solvent accessibility properties or to its relation to structural domains. In this chapter general characteristics of phospho-sites are described and special focus is given to phospho-tyrosine residues, which have been less well-studied. A graphical summary of the main findings is shown in Figure 2.1.

In addition, the question of the functional relevance of phospho-acceptor residues

identified in large-scale studies is addressed through the statistical comparison of various properties of sites with annotated regulatory functions and sites of unknown function.

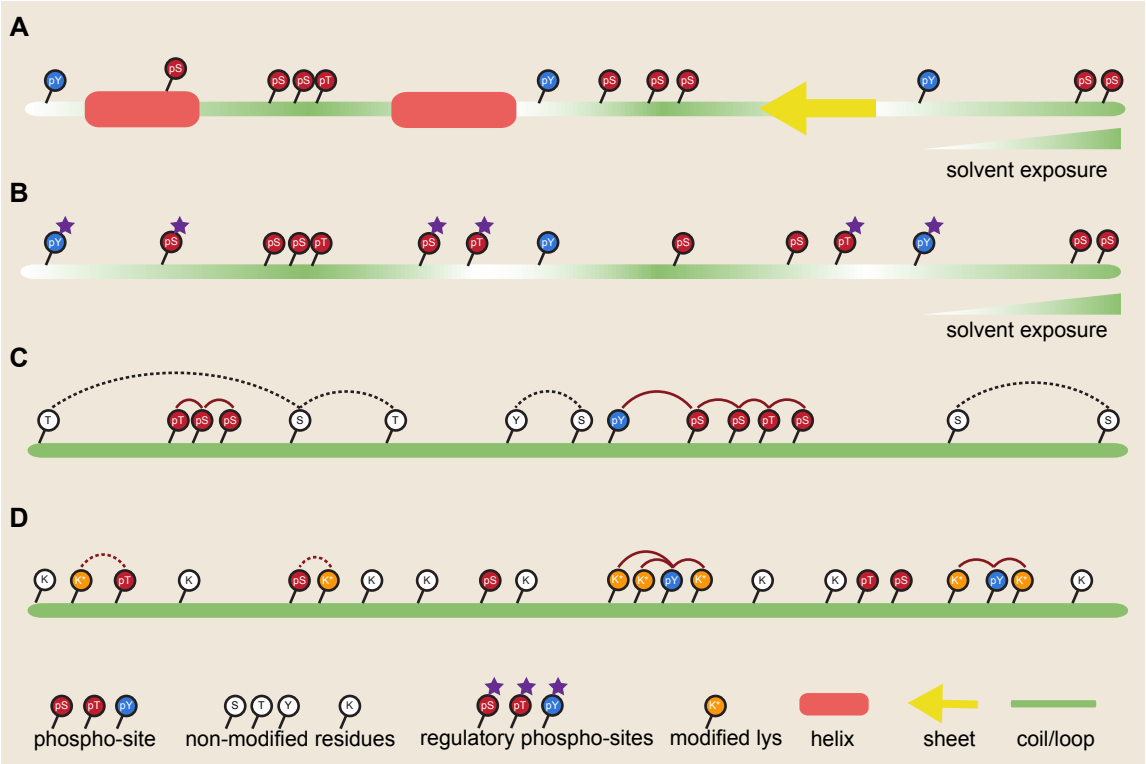


Figure 2.1: ***Schematic overview of Chapter 2.*** ***A)*** *Phosphorylated residues are predominantly found in regions that lack regulat structure. Unlike modified serines and threonines, phospho-tyrosine residues prefer less exposed areas.* ***B)*** *Phospho-sites with regulatory functions prefer less solvent accessible regions in the protein.* ***C)*** *Distances between phosphorylated residues (solid red curves) are on average smaller than distances between non-modified serine, threonine and tyrosine residues (dashed black curves).* ***D)*** *Modified lysine residues show a tendency to lie in close proximity to phospho-sites (red curves). The phenomenon is most prominent in the case of phospho-tyrosines (solid red curves).*

Post translational modifications are generally viewed as a mechanism to increase the proteome complexity and diversity. Yet another level of signal integration is added through the elaborate interactions between multiple phospho-sites and the

cross-talk between different types of modifications. The second half of this chapter focuses on the tendency of phospho-sites to form clusters and on the relation of this tendency to protein abundance levels. Furthermore, the inter-communication between phospho-sites and modified lysine residues is elucidated and a functional interpretation of this phenomenon is proposed.

In this chapter, a large-scale dataset of human phosphorylation sites comprising more than 50,000 sites and generated in our laboratory is analyzed. In agreement with previous experiments, the group of modified serines was the largest (41,009), followed by threonines (9,919), with tyrosines being the smallest group (1803). This is one of the largest sets available on phosphorylated tyrosine residues, allowing for in-depth analysis of the properties of this class of modifications.

Table 2.1: *Phosphorylation and disorder.*

		Phospho	Reference	Odds ratio	P-value
S	disordered	226087	36154	5.618783	<2.2e-16
	ordered	161210	4855		
T	disordered	117149	8572	6.856634	<2.2e-16
	ordered	126228	1347		
Y	disordered	31531	1091	4.006186	<2.2e-16
	ordered	82439	712		

The counts of both phosphorylated and non-modified (reference) serine, threonine and tyrosine residues predicted to lie ordered and disordered regions were used to build contingency tables. Preference of phospho sites to appear in disordered regions were computed using Fisher exact test and are shown by the corresponding odds ratios and their significance.

2.2 Phosphorylation and preferences for disorder and coil secondary structures

A total of 52,731 phospho-sites on 6,682 proteins measured in-house were used in this analysis of structural properties (see Section 2.9). A reference set of sites was built from all serine, threonine and tyrosine residues in the phospho-portein sequences that were not found to be modified in our data set. Next, the disorder state (ordered or disordered) was predicted for each residue using the DISOPRED program [168].

The number of modified residues predicted to lie within disordered regions was compared to the number of reference sites with similar structural context (Fig. 2.2). Fisher exact test was employed to compute odds ratios and the corresponding p-values. Clearly, phospho-acceptor sites were significantly more enriched in disordered regions than their non-modified counterparts (Table 2.1). The enrichment factors of phospho-serine and phospho-threonine residues in regions lacking defined structure were quite similar, whereas phospho-tyrosines were enriched to a lesser extent.

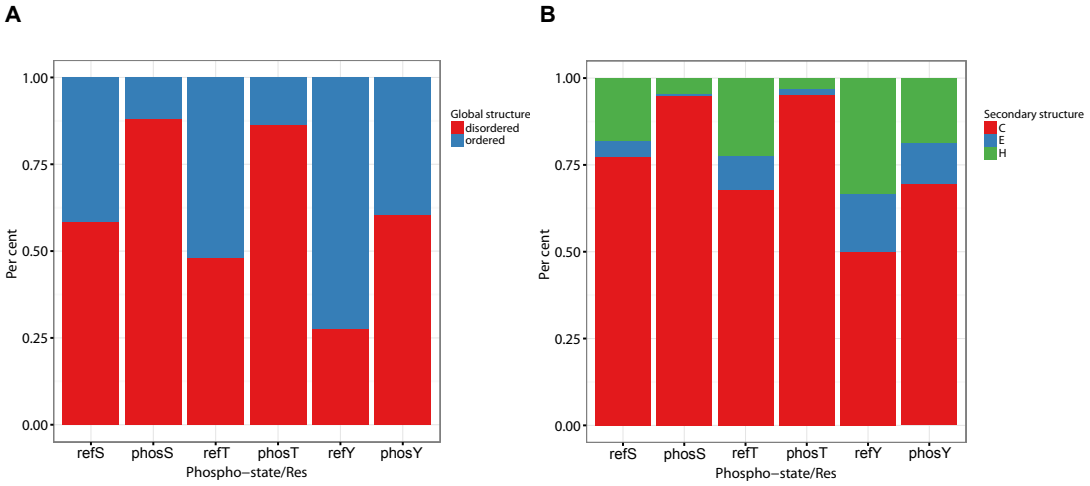


Figure 2.2: ***Structural preferences of phospho-sites.*** ***A)*** *The per cent of both phosphorylated and non-modified (reference) serine, threonine and tyrosine residues in disordered and ordered regions is shown (disordered in red and ordered in blue).* ***B)*** *Analogously, the distribution of modified and reference residues in different secondary structure regions is shown for each phospho-acceptor (coil in red, beta-strands in blue and alpha-helices in green). Phospho-sites are characterized by higher preferences for disordered and coil regions than reference residues.*

The secondary structures of all phospho-proteins were predicted with the PsiPred program [169] and the preferences of phospho-sites for specific structural environment were investigated using the corresponding non-modified sites as background. The proportion of modified residues in regions with irregular secondary structure (coils) was significantly larger than that of the reference sites (Table 2.2, p-values were computed using the proportions test implemented in R [170]). This tendency was less pronounced for phospho-tyrosine resdiues, however, around 70% of these sites where predicted to lie in coil regions. The second most preferred structural

category was α-helices, whereas the smallest proportion of phospho-sites were found in β-sheets.

Table 2.2: *Phosphorylation and secondary structure.*

		Phospho	Reference	Prop. phospho	Prop. reference	P-value
S	coil	38930	299583	0.9493038	0.7735226	<2.2e-16
	helix	1774	69276	0.0432588	0.1788705	<2.2e-16
	sheet	305	18438	0.007437392	0.047606875	<2.2e-16
T	coil	9460	165182	0.9537252	0.6787083	<2.2e-16
	helix	308	54550	0.03105152	0.22413786	<2.2e-16
	sheet	151	23645	0.01522331	0.09715380	<2.2e-16
Y	coil	1251	56817	0.6938436	0.4985259	<2.2e-16
	helix	334	37813	0.1852468	0.3317803	<2.2e-16
	sheet	218	19340	0.1209096	0.1696938	4.938e-08

Counts and proportions of phospho-sites and non-phosphorylated serine, threonine and tyrosine residues predicted to lie within α-helix, β-strand and coil regions. Preference of phospho sites for each local structural context were computed with the poportions test implemented in R. The corresponding p-values are given for each phospho-acceptor and each structural background.

Phospho-sites showed significant preferences for disordered regions and irregular secondary structures. Undermining the structure-function paradigm, intrinsically disordered regions have been shown to convey a large number of functions in the cell [171, 172, 173]. The implications of disordered regions in protein-protein interactions, in particular through increasing the number of interaction partners, ensuring flexibility of the bound complex and allowing for regulation through conformational changes, make them suitable mediators of signal transduction. Therefore, not surprisingly, many studies have reported the association of phosphorylation with disorder [81, 80], described the mechanisms of interplay between the two and presented disorder-based prediction methods for discovery of new modification sites [81]. More details on the role of intrinsic disorder for protein phosphorylation are discussed in Chapter 3 of this thesis.

2.3 Phosphorylation and solvent exposure

Using the SABLE prediction package [174] an accessibility score representing the relative solvent accessible area was computed for each residue in all phospho-proteins. High score indicates high solvent accessibility. The solvent exposure of phosphorylated residues was compared to that of the corresponding non-modified residues (Fig. 2.3). Phospho-serine and -threonine residues were significantly more exposed on average (mean accessibility score 4) than reference serine and threonine residues (mean accessibility score 3, Wilcoxon test p-value <2.2e-16). Although the same tendency was observed for phospho-tyrosine residues, they appeared more buried on average than the other two phospho-acceptor residues (accessibility score means 1 and 2 for modified and reference sites, respectively; p-value <2.2e-16).

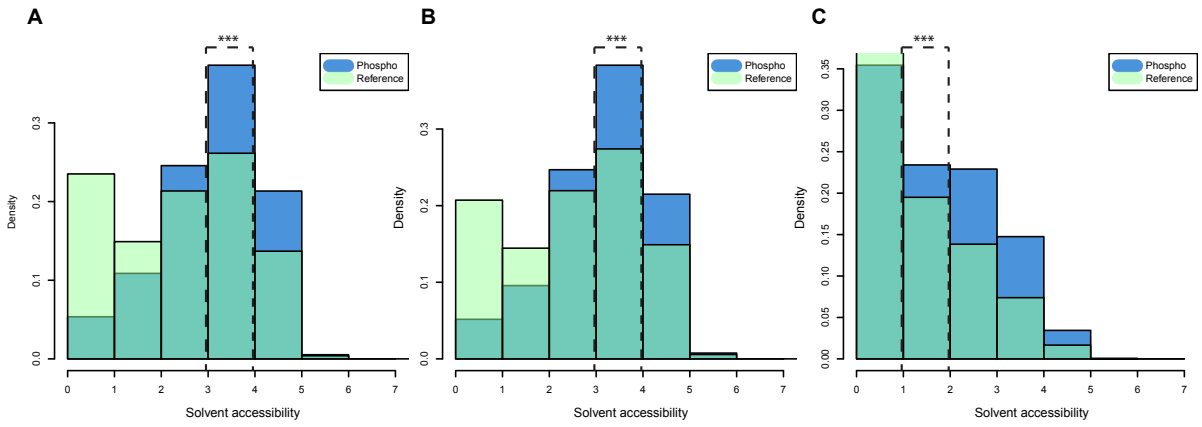


Figure 2.3: ***Solvent accessibility of phospho-sites.*** *The solvent exposure of phosphorylation sites (light green) is compared to that of their corresponding reference residues (blue):* ***A)*** *serine,* ***B)*** *threonine and* ***C)*** *tyrosine. The overlap between the two distributions is shown dark green and the corresponding means (4 and 3 respectively) of the groups are depicted by the dashed lines. In all cases the modified residues appeared to be significantly more exposed than the reference sites (Wilcoxon test).*

As expected the modification sites were more exposed on average than the modified reference sites. The polar nature of these residues underlies their preference for more solvent accessible areas in the protein. This exposure then facilitates access of the kinase to the phospho-acceptor site.

2.4 Phosphorylation and domains

We tested if phosphorylation sites tend to appear within structural domains, or rather prefer the interconnecting loops. Domain definitions were obtained from the Interpro online resource [175] and were mapped to the phospho-proteins in our data set as described in the Materials and methods section 2.9. Preferences of phospho-sites for regions between Interporo domains were computed using the Fisher exact test and the corresponding odds ratios and significance values are shown below. Considering only sites predicted to lie within disordered regions, this showed that about 42% of all modified sites resided within an Interpo domain, whereas this was the case for 60% of the reference sites. The preference for regions outside Interpro domains was even more pronounced when sites with ordered structural background were analysed – 71% of those phospho-sites were found to lie within a domain. Nonetheless, all phospho-acceptor residues in this group were significantly enriched in regions connecting domains in comparison to their corresponding non-modified residues (Table 2.3).

Table 2.3: *Tendency of phosphorylation sites predicted in ordered regions to occur outside Interpro domains.*

		Phospho	Reference	Odds ratio	P-value
S	outsideDomains	1282	30759	1.516393	<2.2e-16
	within domains	2836	103182		
T	outside domains	302	22091	1.39018	2.279e-06
	within domains	818	83183		
Y	outside domains	138	13160	1.224803	0.03935
	within domains	476	55597		

Preference of phospho-sites to appear in regions connecting Interpro domains. Contingency tables containing the counts of phospho-sites and non-modified reference sites found within and outside Interpro domain regions. Only sites predicted to lie within ordered regions were used. Fisher exact tests were computed for each phospho-acceptor residue separately. Modified sites were significantly overrepresented in regions outside Interpro domains.

There were small but highly significant differences with respect to the three phospho-acceptor residues in disordered environments (Supplementary table 6.1). Phospho-serine and -threonine sites were significantly enriched in regions outside of the interpo

domains, although the enrichment factor was relatively low. In contrast, phospho-tyrosines were found to be equally likely to appear both within and outside domains. Overall, despite the large number of phospho-sites appearing within structured domains a clear preference of phospho-serine and phospho-threonine for the interconnecting regions was evident, while no strong differences were found between phospho-tyrosines and non-modified tyrosine residues.

Phospho-serine and -threonine residues showed significant preferences for inter-domain regions even when they were predicted to lie within ordered regions. This observation is in agreement with the above-described tendency of phospho-sites to lie within intrinsically-disordered regions. In contrast, phospho-tyrosines showed only a weak preference for such regions when sites in ordered structures were considered and were found to be enriched within domains when sites in disordered structures were considered. This tendency together with the smaller preference of this phospho-acceptor residue for disordered regions and its lower solvent accessibility may facilitate the specific functional roles of this site in the cell. Tyrosine phosphorylation is involved in the regulation and execution of a wide range of processes, such as differentiation, proliferation, cell death, motility and transcriptional activation [100] and aberrant regulations are often associated with disease development [176]. The precise implementation of these functions requires a well-controlled mechanism to ensure the accurate phosphorylation of tyrosine residues, which is achieved through a complex interplay between tyrosine kinases, phosphatases and proteins with phospho-tyrosine recognition domains [177]. One general scenario of signal transduction involves the binding of some signaling molecule such as a growth factor or cytokine to the extracellular domain of a receptor tyrosine kinase, upon which an effector molecule is phosphorylated. Downstream of the signaling cascade, specificity is governed through the residues surrounding the phospho-acceptor sites, which recruit diverse but specific phospho-tyrosine binding domains. For example, the SH2 phospho-tyrosine-binding domain is characterized by strongly conserved residues at the binding pocket that stabilize the phosphate moiety [178]. Specificity and high fidelity are achieved through a second binding pocket formed by some of the conserved loop regions and beta-strands, which enhances the recognition of residues surrounding the phospho-tyrosine predominantly C-terminal to the modification site.

Furthermore, the lower solvent accessibility and less disordered environment of

phospho-tyrosines that we find in our data are in general characteristic of more hydrophobic regions. These, consequently, present suitable environments for the formation of stable protein-protein interactions, which is agreement with what is known about the role of phospho-tyrosines, which are to a large extent involved in the direct regulation of such interactions during signal transduction.

The current data set on tyrosine phosphorylation sites indicates that this modification is much less numerous than serine and threonine phosphorylation and occurs at less exposed and more structurally ordered regions. Moreover, these properties are characteristic for the binding regions of stable complexes and the sequential and structural contexts play an important role for the specificity of its functions. Hot spots were shown to be enriched in tyrosine residues [179], therefore it may be promising to investigate the overlap between pospho-tyrosines and hot spots, as they may have similar characteristics and regulate protein-protein itneractions in an analogous manner. Based on these observations it can be hypothesized that tyrosine phosphorylation results from purposive kinase interactions, is regulated in a strict manner and results in a highly specific cellular response.

2.5 Properties of regulatory phosphorylation sites

The large number of phosphorylation sites reported to date have led to the suggestion that a significant proportion of those sites may be non-functional and may result from unspecific kinase actions. Therefore, it is of high interest to be able to distinguish functionally-important from silent phosphorylation sites. To do that, a curated set of functionally-annotated phosphorylation sites was obtained from the PhosphoSitePlus resource [94] and mapped to the original set of phosphorylation sites (see Material and methods section 2.9 for details). In total 768 sites from our data set were annotated with a regulatory function. This set was then used to study various properties of the regulatory phospho-sites using as a background all other modified sites that were measured.

The phospho-serine sites with regulatory functions showed significantly smaller preference for disordered regions than the corresponding background modification sites of no specified function (Fig.2.4 A). This tendency was not present for threonine

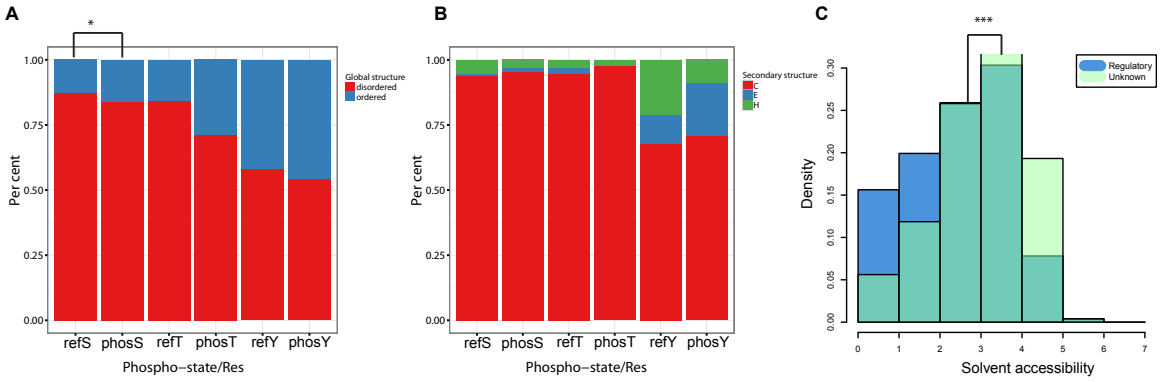


Figure 2.4: ***Structural properties of regulatory phospho-sites.*** ***A)*** *Proportions of phospho-sites of regulatory functions and of unknown funcitons in disordered (red) and ordered (blue) structural regions are compared. Regulatory phospho-serines prefer disordered regions to a significantly lesser extent than modification sites of unknown function, whereas no difference is present for the other two phospho-acceptor residues.* ***B)*** *Distributions of sites of regulatory and of unknown functions in different secondary structure regions is shown for each phospho-acceptor (coil in red, beta-strands in blue and alpha-helices in green). There are no statistically significant differences between the preferences for secondary structure elements of regulatory and functionally unannotated phospho-sites.* ***C)*** *Solvent accessibility scores of regulatory phosphorylation sites (blue) and phospho-sites of uknown funcitons (light green) are compared. The overlap between the two distributions is shown in dark green. Regulatory phospho-sites are significantly less exposed than phospho-sites of unknown function (Wilcoxon test p-valus <2.2e-16).*

and tyrosine residues. Interestingly, larger proportions of regulatory serine, threonine and tyrosine residues were found in coil regions, however, the differences were not statistically significant (Fig.2.4 B). In contrast, there was a clear difference between the preferences of the two groups of sites for solvent accessibility, revealing that the annotated sites are more buried on average (Fig. 2.4 C).

To test if the regulatory sites have different preference for the protein abundance as compared to sites with no functional annotation we compared the distributions of protein intensities (measured in our original data set) in the two groups (Fig. 2.5). Considering only sites within ordered regions, the proteins containing sites with regulatory annotation were characterized by lower intensities on average as compared to the proteins hosting sites with unknown functions. In contrast, no significant differences were found between the two groups when only disordered regions were

considered (Supplementary fig. 6.1).

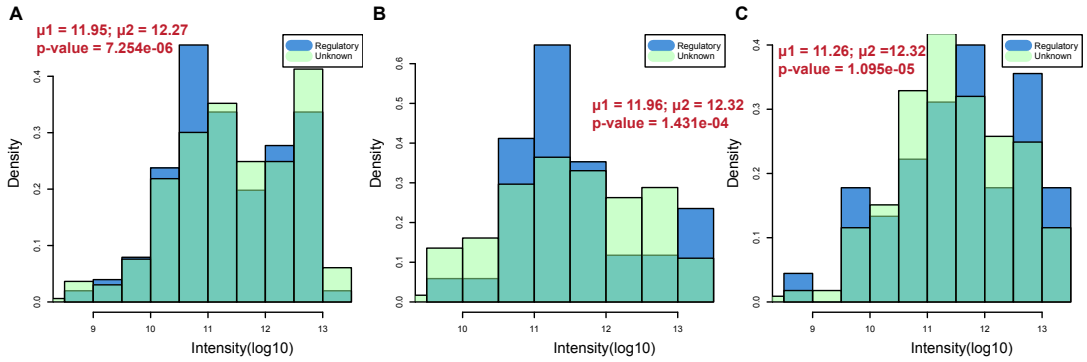


Figure 2.5: ***Regulatory phospho-sites and protein intensity.*** *Protein intensities of phospho-sites in* ***ordered regions*** *with regulatory (blue) and unknown (light green) functions are compared:* ***A)*** *serine,* ***B)*** *threonine and* ***C)*** *tyrosine. The overlap between the two distributions is shown in dark green. The corresponding group means of intensity (regulatory $\mu 1$ and unknown $\mu 2$ respectively) and the p-values computed with the Wilcoxon test are shown for each residue.*

An interesting question arising from the above-made observations is if there are major differences in the evolutionary pressure acting on the two groups of sites. To compute evolutionary rates clusters of orthologs from the EggNOG resource [180] for six eukaryote organisms were employed. The rates were then calculated using a local version of the rate4site algorithm (see [181] and Materials and methods section 2.9). The annotated sites were clearly more conserved than the background set in both ordered (Supplementary fig. 6.2) and disordered (Fig. 2.6) structural environments. The smallest difference in the conservation rate between annotated sites and sites of unknown function was observed for tyrosine residues.

Generally, the phospho-sites that were annotated as regulatory in the PhosphoSitePlus resource exhibited different properties than the rest of the sites in our set. Interestingly, the sites annotated with regulatory functions were characterized by an overall lower solvent accessibility. In addition to the lower accessibility, a very large fraction of them was found within irregular secondary structures (coils), but at the same time phospho-serine residues showed larger preference for ordered regions. As our understanding of the implications of disorder in dynamical processes in the cell increases, more evidence emerges that disorder may come in different forms. A clear distinction was found between long unstructured and regular well-structured loops,

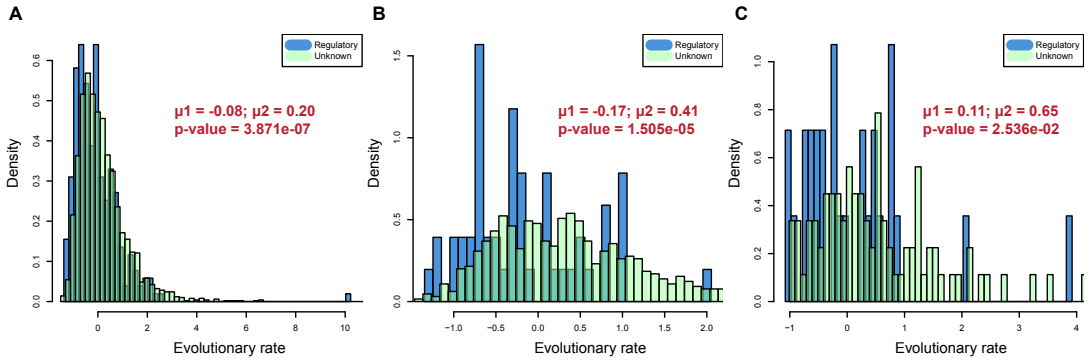


Figure 2.6: ***Evolutionary conservation of regulatory phospho-sites.*** *Evolutionary rates of phospho-sites (predicted in disordered regions) with regulatory (blue) and unknown (light green) functions are compared:* ***A)*** *serine,* ***B)*** *threonine and* ***C)*** *tyrosine. The overlap between the two distributions is shown in dark green color. Evolutionary rate means of the compared distributions (regulatory $\mu 1$ and unknown $\mu 2$ respectively) and the corresponding p-values computed with the Wilcoxon test are presented for each residue. Note that lower values correspond to higher conservation.*

mainly with respect to their length, lack of regular secondary elements and solvent accessibility [182]. Moreover, these different types of disorder may be implicated in distinct functional roles, a hypothesis which finds confirmation in the conservational analysis of Bellay et al [183]. These authors distinguished 3 levels of conservation in disordered regions and drew a link between one of them - flexible disorder (the flexibility is conserved regardless of the exact amino acid composition) and signaling pathways. Chapter 3 contains a more comprehensive discussion on the importance of disorder in the regulation of phosphorylation.

One of the mechanisms through which regulatory functions can be carried out by modification sites is conformational change. These include both changes in the immediate environment of the phospho-acceptor sites and allosteric changes influencing protein regions that lie far apart. The clustering of energetically-important residues characterized by low solvent accessibility at the interfaces of protein-protein complexes contributes to the binding free energy [179, 184]. This low solvent accessibility enhances the exclusion of bulky solvent molecules and thus prevents their interference with the affinity of the interactions, ensuring accurate execution of the regulatory functions induced by the modification sites. Demerdash et al. demonstrated that residues with allosteric function tend to be more buried than residues

that do not have such roles [185], suggesting that a more densely-populated network of residues may provide suitable environment for inducing structural changes at a distant location. Therefore the lower accessibility of the regulatory sites may be indicative of the mechanism through which they achieve their function, namely through allosteric interactions.

Not surprisingly in view of the above discussion, we found that the sites that were functionally-annotated were subject to higher conservational pressure than sites of unknown functions. This observation supports the hypothesis that these sites may be related to functionally-important regions on the protein surface and thus would impose low evolutionary rate. Interestingly, the regulatory phospho-tyrosines were only slightly more conserved than the modified tyrosines with no annotated funciton. This could well be explained by the general distribution of these residues in the protein. Tyrosine is a much less frequent residue, characterized by a bulky aromatic ring. Due to its physicochemical properties it plays an important role in structure maintenance and protein-protein interactions, necessitating its higher average conservation.

2.6 Multiple phosphorylation sites

The distribution of the number of phospho-sites per protein is shown in Fig.2.7 A). Of the total of more than 50,000 phospho-sites that were measured in 6,768 unique uniprot entries 5,121 (22%) had only a single phosphorylation site and were discarded from the rest of the analysis in this section. About half of the proteins contained between 2 and 10 sites and around 20% contained more than 10 sites.

We were interested in a possible tendency of phosphorylation sites to cluster. The distances between any two phospho-sites in a protein were computed and the smallest distance for each site was retained. Next, 1,000 randomisations of the modification positions over all serine, threonine and tyrosine residues were computed for each protein and the corresponding distance distributions were estimated (see Materials and methods section 2.9). Figure2.7 B clearly demonstrates that the measured phosphorylation sites were characterized by smaller distances than the randomized data sets, suggesting a strong and highly significant tendency of these modification sites to cluster together. This effect was larger in intrinsically disordered regions, a

tendency that underlines the capability of such regions to accumulate a large number of phosphorylation sites.

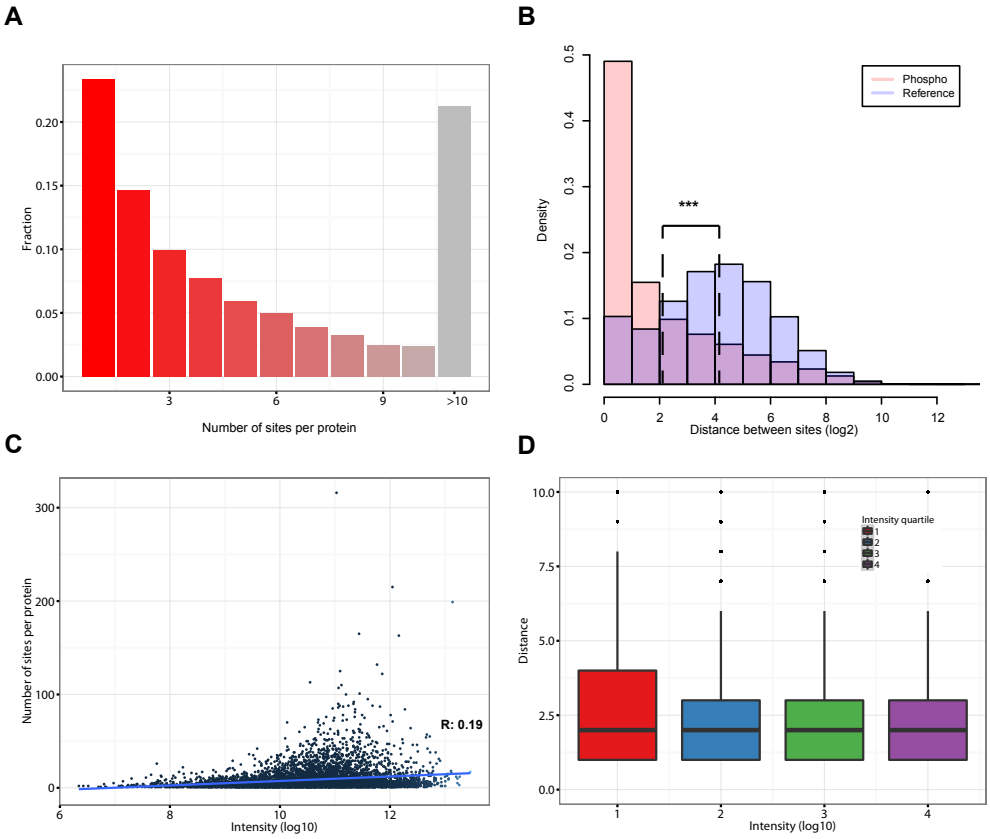


Figure 2.7: ***Characteristics of multiple phosphorylation sites.*** ***(A)*** *The distribution of proteins with specific number of phospho-sites (as a proportion of the total number of phospho-proteins) is plotted with color gradient ranging from dark red - proteins with single sites to grey - proteins with more than 10 sites. Overall, gradual decrease of the number of proteins with increasing number of sites is visible.* ***(B)*** *The distribution of distances between the measured phospho-sites (in light red) and between randomized phospho-sites (in light blue) are compared. The overlap between the two distributions is shown in dark blue. The distances are measured in amino acids, shown in log2 scale and the density of the distribution is plotted.* ***(C)*** *A scatterplot of the number of phospho-sites in a protein against the protein intensity (log10) is shown. The correlation coefficient between the two variables indicates a small positive correlation.* ***(D)*** *Distributions of distances between phospho-sites are compared between protein groups with different intensities. The group with smallest intensities (red) corresponds to the first quartile of the intensity distribution and the group with the largest intensities (purple) correponds to the last quartile.*

The importance of multiple phosphorylation sites has been discussed in a large number of studies (see [54] for a comprehensive review). For example, the mitogen-and stress-activated protein kinase (MSK1) undergoes phosphorylation on Thr581 and Ser360 and is further capable of autophosphorylation of six additional sites [186]. Two of these sites, which lie in specific motifs and loops, were shown to be determining for the catalytic activity of the N-terminal kinase domain. Precise regulation of multiple phosphorylation sites has been shown to be important in a myriad of cellular processes such as DNA damage checkpoint regulation [187], chromosome condensation initiation [188], localization [189], protein degradation [190] and regulation of transcription [191].

The mechanisms underlying the cross-talk between multiple phosphorylation sites are still not well understood. Functional implications include but are not limited to: priming phosphorylation, in which the phosphorylation of one residue is a determining factor for the modification of another [192, 193]; compensatory interactions, in which modification of one or few of all potential phosphorylation sites is enough to achieve the required function [194]; synergetic phosphorylation, in which the cumulative effect of the phosphorylation of all sites in a region determines the function of the protein [187] and exclusive phosphorylation, in which the modification of one site prevents the modification of another [195].

Using the connexin proteins family as a model, Chen et al. concluded that the implication of multiple phosphorylation sites was more complex than simply creating a binary switch controlling a single function [196]. Instead they proposed that the manifold levels of phosphorylation of multiple sites (depending on localization, order of modification and absolute stoichiometry) may be needed for the efficient integration of diverse signals and may contribute to the regulation of various processes in the cell [197, 198]. Overall, the presence of multiple phosphorylation sites increases the regulatory potential of the modified protein. For example, the proapoptotic protein BAD can be phosphorylated on three sites as a result of the activation of three distinct pathways. Phosphorylation of any of the three sites counteracted the apoptotic function of the protein, indicating the efficiency of this biological system in the integration of distinct signals [199].

Given the high number of clustered phosphorylation sites emerging from this large-

scale analysis, it would be interesting to test them in relation to some of the above functions and mechanisms.

2.7 Correlation between protein intensity and number of phosphorylation sites

Levy et al. argued that high protein abundance may increase the number of random interactions between kinases and phospho-proteins, thus resulting in a large number of non-specific modification sites [110]. To investigate this possibility, two measures from our phospho data set were used: protein intensity and the corresponding number of phospho-acceptors and the correlation between the two variables was estimated (Fig.2.7 C). In agreement with Levy et al. a trend for more abundant proteins to contain a larger number of phosphorylation sites was evident (R: 0.19, p-value <2.2e-16). However, as the confidence with which phospho-sites are identified and localized is also correlated with the protein abundance, the possibility that the above trend is related to the methodology of assigning sites cannot be excluded.

Next, we wanted to know if the observed tendency for higher abundant proteins to have more phosphorylation sites enhances their ability to cluster or rather if more abundant proteins were characterized by more clustered phospho-sites in general. The phospho-proteins were split into four classes corresponding to the intensity quartiles. The distances between phospho-sites in each class were computed in an analogous manner to that used to compute the general distribution of distances (see Materials and methods section 2.9). What we found was that the lowest abundance protein group had the least tendency for phospho-sites to cluster on average as compared to all other groups (Fig.2.7 D).

Levy et al. reasoned that a large number of the phosphorylation sites on highly abundant proteins were not functional and resulted from random kinase-substrate interactions. They supported this hypothesis by showing that these sites had lower rates of evolutionary conservation. However, another possibility is that more abundant proteins require more precise control and regulation, and one way to achieve this could be through a larger number of sites.

Signaling through phosphorylation is among the major mechanisms enabling the cell

to form a rapid and adequate response to various stimuli and stress conditions. For example, the function of the multisubunit eukaryotic translation initiation factor (elF) is regulated through its interaction with the family of repressor polypeptides (4E-BPs). Interestingly, this interaction depends on the level of phosphorylation of the 4e-BP molecule; the hypophosphorylated forms show high affinity for elF, whereas hyperphosphorylation prevents the interaction completely [200, 201].

Moreover, theoretical models showed how the interplay between multiple phosphorylation sites may enhance temporally-regulated responses, integrating various signals and gradual changes. Varedi et al. proposed a model for effective and flexible regulation of protein degradation [202]. They suggested that degradation may be achieved in a time-resolved manner through incorporation of the gradual changes in the concentration of the responsible kinases. After reaching a phosphorylation threshold (i.e. a certain number of modified sites) rapid degradation would take place. Ultrasensitivity, enhanced by the presence of multiple phosphorylation sites, appeared to be a common mechanism, implicated in the regulation of various proteins [203, 204, 205]. This phenomenon can be exemplified by a threshold behavior, controlled through variation of the requirement of the ratio of concentrations of kinases to phosphatease with respect to the number of phospho-sites. In a detailed theoretical analysis Gunawardena et al. argued that although phosphorylation at multiple sites may provide an efficient threshold, it does not necessarily cause a rapid switch [206].

Multiple phosphorylation sites are further implicated in the regulation of protein-protein or protein-nucleic acids interactions. Nishi et al. found a significant enrichment of phospho-sites at interfaces of complexes and a large overlap with hot spots [103]. Surprisingly, their results suggested that phosphorylation is more likely to cause an increase in the diversity of possible interaction partners than to lead to altered binding affinity. Furthermore, the accumulation of negatively charged phosphate groups may have an influence on the electrostatic forces between the interacting partners. Both scenarios of favorable interactions between surfaces with residue clusters of opposite charges [207] and repulsive interactions, for example, with the negatively charged DNA molecules [208], are possible. In another example, the hydrogen bonding network formed by arginine and glutamate residues during the binding of cytoplasmic linker-associated protein 2 to end-binding protein 1 was

disrupted through unfavorable electrostatic forces resulting from the addition of multiple phosphate groups [209]. Clusters of phosphate groups may also influence the structural compatibility between the interaction partners, for instance, by introducing steric hindrance [210] or by increasing the binding affinity [211]. Structural changes caused by multiple phosphorylation sites can regulate interactions regardless of if they occur at the binding interface or at more distant positions through allosteric conformational changes.

These examples strongly suggest that the individual role of phosphorylation sites may not be as important as their cumulative effect. In that case low conservation level and relatively low stoichiometry of such sites would be more readily explainable and would not indicate that these sites are non-functional. The presence of functional clusters of sites may be an important mechanism of increasing the proteome's plasticity. Compensatory phosphorylation provides alternative routes for achieving a specific outcome. This phenomenon allows for the integration of the signal of various source, while limiting the chances of failure in the cellular response. Furthermore, the requirement for modification of several sites may be an efficient way of integrating gradual changes in a time-resolved manner.

2.8 Cross-talk between phosphorylation and modified lysine sites

The complexity underlying the regulation of numerous cellular processes is further increased through various post-translational modifications working in a coordinated manner. A possible inter-dependence between phosphorylation sites and modified lysine sites was investigated through analysis of the preference of the two types of modification to occur at residues of close proximity at the sequence level. We matched experimental data on ubiquitination, acetylation and sumoylation sites that were obtained from the PhosphoSitePlus repository [94] to the phospho-protein sequences in our data set (see Materials and methods section 2.9). In total, 24,004 lysine modifications were used in the analysis. The fraction of modified to non-modified lysine residues in the vicinity of phospho-acceptor sites was computed for both the measured modified lysine positions and the their randomized distribution (see Materials and methods section 2.9). The fractions were compared at variable

distance intervals around the phospho-sites (Fig.2.8). Clearly, the measured modified lysine residues showed different behavior from the randomized data. The surrounding of all phospho-acceptor residues: serine, threonine and tyrosine - showed higher proportions of modified lysines. The phenomenon indicates a possible functional cross-talk that may facilitate the dynamic control of signal transduction under various conditions. The tendency became more pronounced at shorter distances, suggesting the existence of a specific mechanism of intercommunication between the two modification types. Interestingly, the most prominent effect was observed for phospho-tyrosine residues, which may result from specific functional pressure acting on the distribution of this combination of modifications.

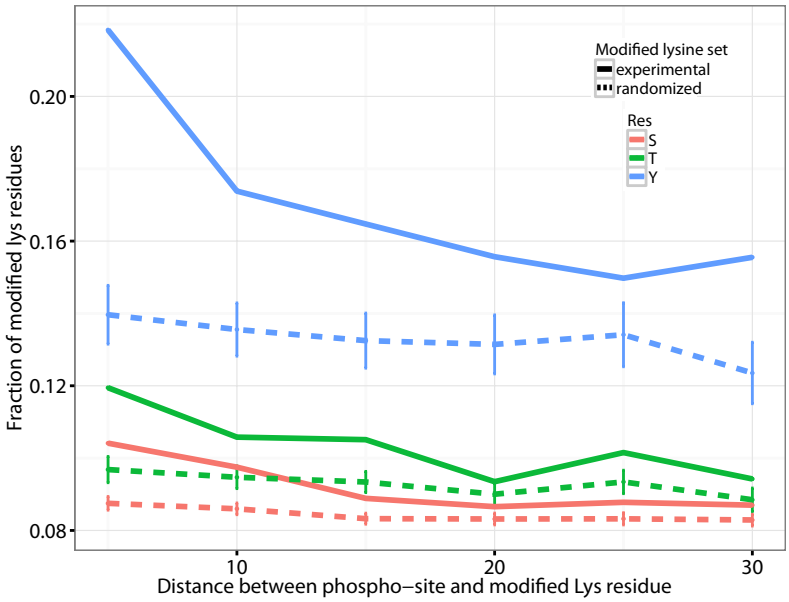


Figure 2.8: ***Cross-talk between PTMs.*** *The fraction of modified lysine residues (corresponding to ubiquitination, sumoylation and acetylation as reported in PhosphoSitePlus) and non-modified lysine residues for a particular interval of amino acids surrounding a phospho-site are represented by solid lines. The different colors correspond to the three phospho-acceptor residues: serine (red), threonine (green) and tyrosine (blue). Background fractions (represented by the dashed lines) were computed analogously, but the positions of the modified lysines were randomly re-distributed over all other lysine residues preserving the overall number of this modification.*

Various interaction patterns exist between multiple modification sites, which also

lead to distinct outcomes. The interactions between lysine modifications and phosphorylation sites can have synergistic character or be mutually exclusive with opposite functional effects [212]. Phosphorylation at multiple sites was shown to lead to enhanced transcription activity of the ATF7 and to abolish the inhibitory effect of sumoylation, most likely through conformational changes [213]. Similarly, phosphorylation of the thymine DNA glycosylase prevented acetylation of a nearby lysine site and ensured preservation of the efficient DNA repairing role of the protein [214]. In contrast, the phosphorylated form of the N-terminal tail of histone H3 was reported to act as a signal for subsequent acetylation giving an example of the synergistic modifications [215]. Furthermore, phosphorylation-dependent ubiquitination [216] and sumoylation [217] have been observed.

Recent improvements in MS-based proteomics have enabled the identification of thousands of post-translational modifications and the subsequent analysis of the cross-regulation between them. The vast amount of possible and functionally-distinct combinations of multiple sites of various modification types has led to the development of the term "post-translational modifications code" [212, 218, 219, 220], which resembles the concept of the genetic code. In their review, Nussinov et al. argued that modifications with allosteric effect (i.e. modulating distant regions in the protein) increase even further the combinatorial and thus functional space, modulated by the diverse nature of PTMs [218]. Furthermore, analysis of the co-evolution of different types of PTMs over 8 eukaryotic organisms revealed a vastly interconnected network of functionally associated modification types [220]. Phosphorylation and lysine modifications appeared central to the network due to their spatial and temporal regulatory roles. The substantial intertwining between phosphorylation and acetylation networks was further demonstrated on an organismal level [221].

The functional interplay between tyrosine phosphorylation and acetylation has harldy been addressed in large-scale experiments. Studies on individual proteins identified varying modes of cross-talk. In cortactin the two modifications did not occur simultaneously, but had similar functional effects [222], indicating possible antagonistic character. In contrast, a synergistic effect between the two was elucidated in the formation of the complex between the major herpes simplex virus type 1 DNA-binding protein and single strand DNA [223]. Tyrosine phosphorylation as well as lysine modifications are involved in the regulation of various signaling events, interactions,

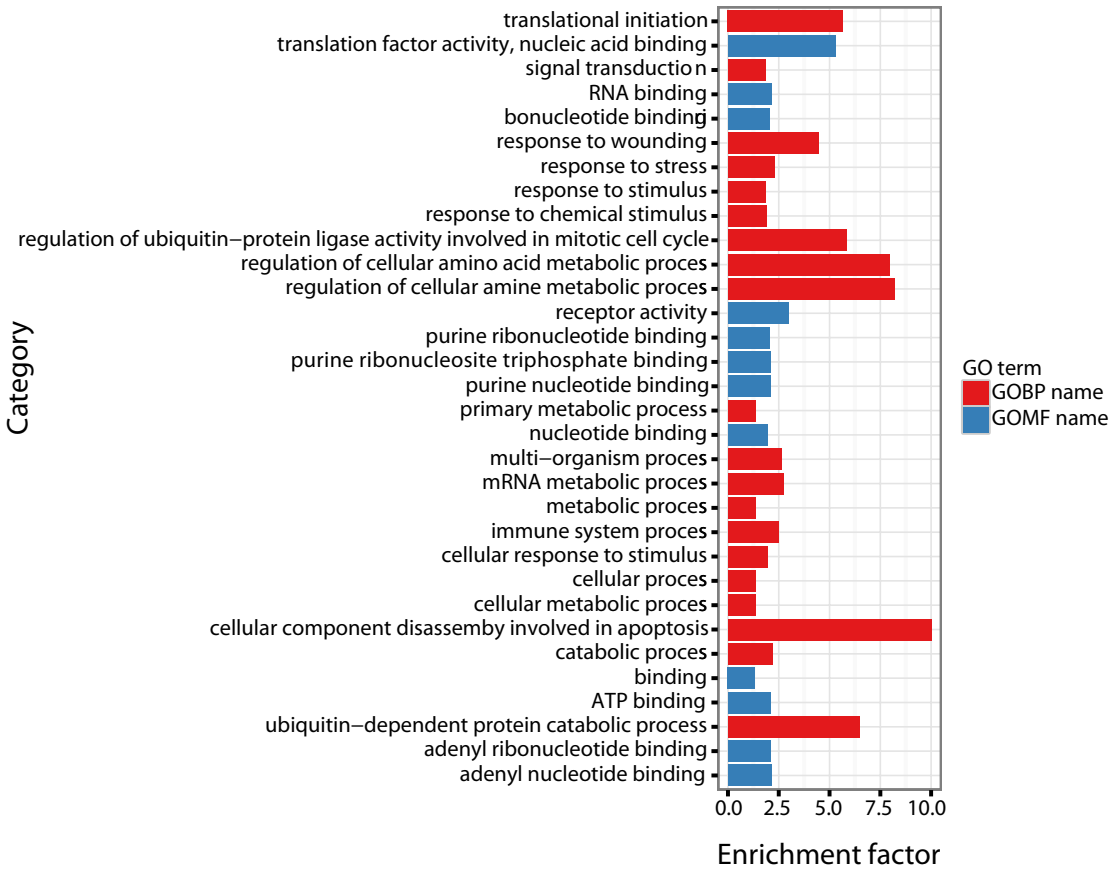


Figure 2.9: ***Cross-talk between PTMs.*** *Enrichment analysis of GO categories in the proteins set, that contains modified tyrosine and lysine residues separated by maximum 5 amino acids. Categories related to the GO term of biological processes are shown in red (GOBP), whereas GO terms representing molecular functions (GOMF) are shown in blue.*

localization and numerous cellular processes. To identify specific functions and processes, in which cross-talk between the two may have a significant implication, we performed an enrichment analysis of GO categories (Fig. 2.9). The set of proteins that contained a modified pair of lysine and tyrosine residues, lying within a distance of 5 amino acids or less, was used in the analysis. The interplay between the two appeared to be important in the regulation of binding between diverse partners, whereas the major processes that were significantly enriched were predominantly related to signal transduction and stimuli response. Interestingly, the proteins spanned a large range of cellular components, including nucleus, cytoplasm, organelles and

extracellular space. Overall, the interaction between the two modifications appears to act as a general mechanism facilitating the cellular response to various signals and in various cellular contexts.

2.9 Materials and methods

Phosphorylation, reference and regulatory sites sets definitions

An in-house phosphorylation data set was used in the analysis (courtesy of Dr. K. Sharma). High resolution mass spectrometry at the MS and MS/MS level for in-depth characterization of the phosphoproteome of a human cancer cell line (HeLa) was emplyed. Phosphopeptides were enriched using strong cation exchange based fractionation and metal complexation. The data were measured on a benchtop quadrupole Orbitrap instrument with HCD fragmentation and very high sequencing speed [224]. MS raw files were processed with the MaxQuant suite [6]. Phosphorylation sites were measured across a number of cellular conditions including mitosis and EGF stimulation. In particular, high coverage of phospho-tyrosine residues was obtained using pervanadate treatment of cells to inhibit tyrosine phosphatases. The set comprised 52,731 sites on 6, 682 proteins. The distribution of the different phospho-acceptor residues was as follows: 41,009 serines, 9,919 threonines and 1803 tyrosines.

A reference set was defined as all serine, threonine and tyrosine residues from the phospho-proteins in the phosphorylation set, which were not found to be modified in our data set.

A set of phospho-sites with regulatory functions was obtained from the PhosphoSitePlus online resource [94]. It comprised curated phosphorylation sites, i.e. sites, which were shown in literature to regulate molecular functions, biological processes, and molecular interactions. The regulatory sites were then mapped to our set of phospho-sites, resulting in 768 sites being annotated with regulatory functions.

Prediction of structural features of phosphorylation sites

The disordered state of each phospho-protein in the data set was predicted using a local installation of the DISOPRED [169] software with default parameters. To each

residues was assigned one of two states: disordered or ordered. Analogously, secondary structure predictions were computed with a local installation of the PsiPred prediction tool [169]. Each residues was associated with one of three possible assignments: 'H' for α-helix, 'E' for β-strand and 'C' for coiled regions. Solvent accessibility was predicted with the SABLE package [174]. SABLE uses sequential information to predict the relative solvent accessibility area.

Domain assignment

Domain information was downloaded from the Interpro online resource [175]. The Interpro database integrates domain predictions from various resources and thus provides a comprehensive representation of domain assignments for a protein of interest. The domain regions were mapped to the phospho-proteins in our set and depending on the relative position of each phospho-site to the end and start positions of the domains, a state 'withinDomain' or 'outsideDomain' was assigned.

Conservation of phospho-sites

Clusters of orthologs were obtained from the EggNOG resource [180]. Phospho-proteins were matched with the corresponding EggNOG cluster, if available. Next, the clusters were reduced, so that sequences from only 6 eukaryotic organisms were contained: human (9606), zebrafish (7955), Mus musculus (10090), Saccharomyces cerevisiae (4932), Arabidopsis thaliana (3702) and Caenorhabditis elegans (6239). The rate4site algorithm was then used locally to compute evolutionary rates of all phospho-sites [181]. Rate4site computes the relative evolutionary rate for each site in a multiple sequence alignment using a probabilistic evolutionary model. Based on the alignment a phylogenetic tree is computed and the resulting topolgy and branch lengths are used in the estimation of the evolutionary rates. Note that lower values correspond to higher conservation.

General statistical methods

The R statistical framework [170] and the Perseus software were used to perform the statistical tests reported in the analysis. The Wilcoxon non-parametric test was used when two distributions were compared (e.g. solvent exposure of regulatory sites with sites of unknown function). Enrichment of phospho-sites in disordered

regions was computed with the Fisher Exact test implementation in R. Enrichment in secondary structure regions was computed with the proportions test, implemented in R. The category assignment and enrichment of GO terms was done in the Perseus environment. Most of the graphics were produced using the ggplot2 package [225].

Computing distribution of distances between phosphorylation sites with respect to protein intensity

The protein intensity information measured in the phospho-data set was used to build four groups of protens based on the underlying intensity quartiles. Next, the distances between all phospho-sites in a protein were computed and the smallest distance for each site was used to build distance distributions. The distributions of distances in the four intensity groups were plotted in the R framework, using the ggplot2 package [225].

Cross-talk between various PTMs

Ubiquitination, acetylation and sumoylation data sets were obtained from the public repository PhosphoSitePlus [94]. PhosphoSitePlus is an online resource integrating information about post-translational modifications from both small-scale and large-scale studies. The modified lysine residues were mapped to the phosphorylation data set. The fraction of modified to non-modified lysine residues in the surroundingn of each phospho-site was computed. Different distance intervals were considered (e.g. +/- 5, +/-10, +/-20 amino acids surrounding the phospho-site). Next, the positions of the modified lysines were randomized over all lysine residues in the corresponding proteins and the above-described fractions were re-computed. The randomization was repeated 1,000 times creating a background distribution of random distances. The measured and randomized fractions were plotted for each phospho- serine, -threonine and tyrosine residue. The set of protein that contained modified tyrosine and lysine residues lying within a distance of maximum 5 amino acids was used in the GO term enrichment analysis.

Chapter 3

Phosphorylation dynamics over the cell cycle and intrinsic disorder

Phosphorylation at specific residues can activate a protein, lead to its localization to particular compartments, be a trigger for protein degradation and fulfill many other biological functions. Protein phosphorylation is increasingly being studied at a large scale and in a quantitative manner that includes a temporal dimension. By contrast, structural properties of identified phosphorylation sites have so far been investigated in a static, non-quantitative way. Here we combine for the first time dynamic properties of the phosphoproteome with protein structural features. At six time points of the cell division cycle we investigate how the variation of the amount of phosphorylation correlates with the protein structure in the vicinity of the modified site. We find two distinct phosphorylation site groups: intrinsically disordered regions tend to contain sites with dynamically varying levels, whereas regions with predominantly regular secondary structures retain more constant phosphorylation levels. The two groups show preferences for different amino acids in their kinase recognition motifs - proline and other disorder-associated residues are enriched in the former group and charged residues in the latter. Furthermore, these preferences scale with the degree of disorderedness, from regular to irregular and to disordered structures. Our results suggest that the structural organization of the region in which a phosphorylation site resides may serve as an additional control mechanism. They also imply that phosphorylation sites are associated with different time scales that serve different functional needs.

3.1 Introduction

Phosphorylation is a ubiquitous post-translational modification that is known to be important for the regulation of a myriad of cellular processes, among which are cell growth, apoptosis, differentiation, signal transduction and transport [54]. Rapidly evolving mass spectrometry (MS)-based technologies, innovative labeling techniques and advances in computational proteomics provide powerful means for overcoming the low abundance problem of this modification and are making it possible to obtain large-scale, high-resolution quantitative data. With these advances, not only can single protein phosphorylation experiments be done with high accuracy, but also whole-phosphoproteome studies are becoming increasingly feasible [61, 6].

Given the availability of these data, much research has been devoted to analyzing and understanding the structural features of phospho-sites. This includes creation of online resources containing structural information [78], combining data on linear motifs and structural properties [226], and development of software tools that use three-dimensional data for the prediction of phosphorylation sites (DISPHOS [81], Phos3D [80]). Large-scale studies of the structural characteristics of phosphorylation sites have focused on solvent exposure, local and global structure, amino acid context of the spatial surrounding, and structural motifs [80, 79, 84]. The mechanism of modification suggests that serine, threonine and tyrosine residues should be located on the protein surface where they are accessible for the modifying kinase [80].

The main challenge in studying structural properties of phospho-sites from experimental data is their preference for unstructured regions [81] for which electron density is often missing in X-ray structures. Disorder is strongly associated with protein-protein interactions [227]. Modified residues found within disordered regions can act as on/off switches, either promoting or inhibiting an interaction. Due to the specific structural organization of some protein kinases, in which the catalytic loop resides within a small cleft between two lobes, flexible regions within the substrates interaction surface are well suited for binding to the kinase. However, a recent systematic study suggested that kinase preference for disordered regions is only marginal [79]. Furthermore, a computational study of kinase specificity reported that approximately 60% of the sites modified by protein kinase A lie within

-helical regions [228]. These considerations raise an interesting question: can a distinction between kinases be made with respect to the level of structural organization of their substrates.

Since phosphorylation events both depend on the structural environment and influence its properties, protein structure and phosphorylation should be considered interrelated and mutually dependent. On one hand, disorder facilitates the access of a kinase to the residue to be modified. On the other hand, the addition of a phosphate moiety may lead to structural changes. Both order-to-disorder and disorder-to-order transitions upon phosphorylation have been observed in nature or studied via molecular dynamics simulations [229, 230, 231, 232]. The major driving forces of conformational changes observed upon phosphorylation are the electrostatic interactions between the negatively charged phosphate group and the surrounding charged residues. The functional roles of charged residues range from stabilization to correct substrate identification and facilitation of conformational changes.

Although numerous previous studies have focused on structural properties of phosphorylation sites [81, 80, 79] no systematic analysis has been performed combining large-scale quantitative data with structural features. To bridge this gap we here build on data from a recent study by Olsen et al., which elucidated phosphorylation site occupancy during mitosis [65]. Quantitative data were measured at six time points, corresponding to major phases of the cell division cycle. The additional temporal dimension of these data makes it possible to examine how various phosphorylation sites are dynamically regulated. Olsen et al. clustered sites according to their distinct phosphorylation patterns and similarities in regulation with the aim to infer each sites functional importance. Here, in contrast, we focus on structural properties of the phosphorylation sites and, for the first time, distinguish between two groups of sites with respect to the overall variation of phosphorylation over time.

We find that sites that lie within regular secondary structures exhibit less variable phosphorylation fold changes during the cell cycle than sites that are found in disordered regions. Analysis of the amino acid composition of the flanking regions of these two groups of sites revealed enrichment of positively charged residues and depletion of disorder-related residues such as proline, serine and threonine in the former group.

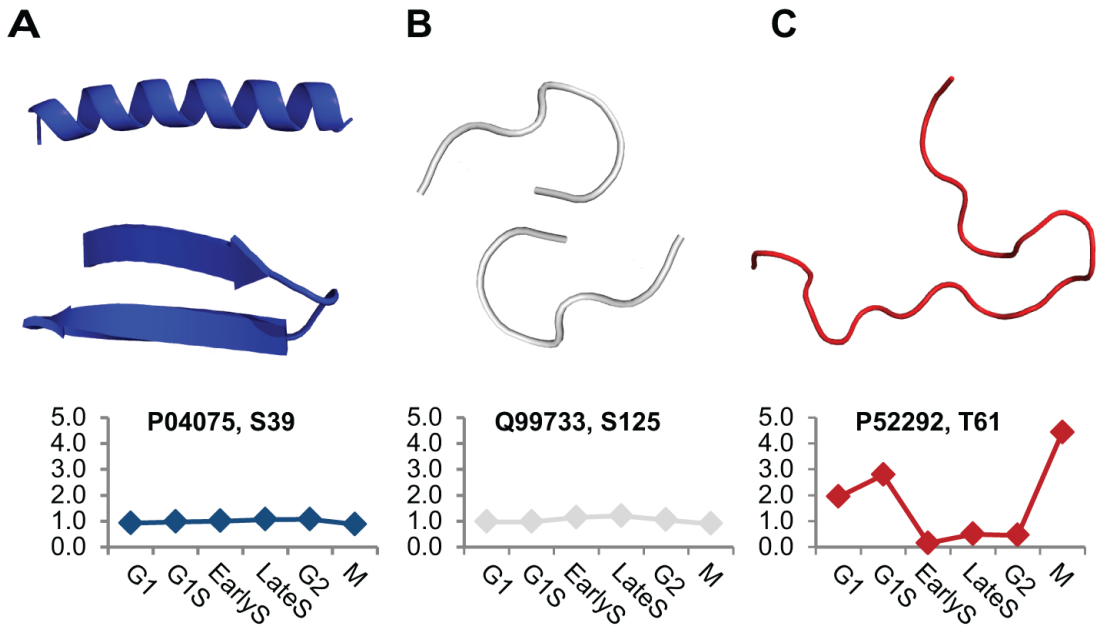


Figure 3.1: ***Temporal phosphorylation patterns of phospho-sites with distinct structural properties.*** *Phosphorylation fold changes of three sites (UniProt accession number and residue identification number are given) during the six time points is shown together with their corresponding local structure. From left to right the phosphorylation variation over the six time points increases, together with the level of disorder: from* ***(A)*** *regular secondary structure (-helix or -sheet) through* ***(B)*** *irregular coils and loops to* ***(C)*** *disordered regions. Phospho-serine residues (pS) within regular regions and loops show small fluctuations in their phosphorylation levels, while larger changes occur in disordered regions.*

3.2 Variation of phosphorylation in disordered regions

Using the data from the Olsen et al. investigation [65], we here computed the overall variation of the phosphorylation ratios during six time points of the cell cycle and investigated the differences between the sites with small variation as opposed to the sites with large variation. The original data set comprised 6,027 proteins with 20,443 unique phosphorylation sites. We retained only those sites that had quantitative information for all six time points available (1,059 proteins with 5,173 sites). The phospho-site variability is calculated as the standard deviation of the phosphorylation ratios over the six time points measured during the cell cycle.

We sought to investigate a possible relation between the structural organization of the environment in which a modified residue is found and the experimentally measured changes in phosphorylation during the cell cycle. To do so, we compared the phosphorylation variation of two groups of sites. These two groups were composed of sites that reside in ordered regions and sites that lie within disordered regions as predicted with DISOPRED [168]. In agreement with previously observed tendency we found over 90% of the modified residues to lie within disordered regions (4,675 sites versus 498 sites). Figure 1 shows three examples from our large-scale dataset, illustrating a non-variable site on a regular secondary structure (-helix), a slightly variable site on a short loop and a variable site in a disordered region.

Our results revealed notable differences in the distributions of phosphorylation variations of the two sets (Kolmogorov-Smirnov test p-value 6.6E-13). The sites associated with structurally characterized regions were found to exhibit smaller changes in phosphorylation during the cell cycle (median 1.77) as compared to sites located in disordered regions (median 2.22, Figure 2A).

3.3 Phosphorylation variability scales with the level of structural order

Having investigated the difference between ordered and disordered regions on a global scale, we next predicted protein secondary structure in more detail using PsiPred [169]. First we classified sites into regular structures (92 in -helices or 53 in -sheets) and sites with irregular structures (5,028 in loops, turns and coils). Phosphorylation in regular secondary structures showed smaller variation over the six time points of the cell cycle. This effect was small but statistically significant (ANOVA p-value 1.8E-04).

Although there is a large intersection between ordered structures and regular secondary structures, and the terms are often used interchangeably, the two sets are not identical. We observed that a large number of regions predicted as coil by PsiPred are predicted as ordered by DISOPRED. This reflects a distinction between ordered and disordered coils. A major difference between these two groups of coils is the length distribution of their elements (p-value 4.12E-114): ordered coils are much

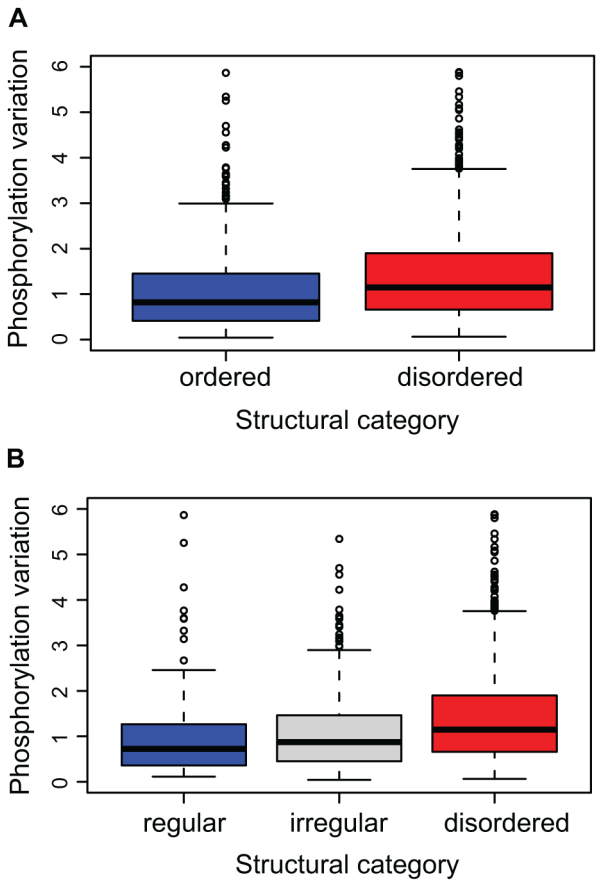


Figure 3.2: ***Comparison of phosphorylation variation of sites within different structural categories. A)*** *Sites within ordered regions (blue) show smaller variation of the phosphorylation fold change over the cell cycle than those within disordered regions (red). The significance of the observation has been tested with Kolmogorov-Smirnov test (p-value 6.6E-13).* ***(B)*** *The variation of phosphorylation changes over the cell cycle scales with the structural propensities of the phosphorylated residues: from lowest in regular structures (blue) to highest in disordered regions (red). The observed differences were found to be significant by ANOVA test (p-value 3.02E-09).*

shorter on average as they mainly correspond to turns and short loops connecting regular secondary structures. By contrast, disordered regions are longer and represent large protein regions lacking defined structure (see Text S1 for details).

In order to take this distinction into account, we redefined the structural environments into three categories: regular structures (predicted as helix/sheet and ordered), irregular structures (predicted as coil and ordered), and disordered regions

(predicted as coil and disordered) (Figure 2B). We found significant differences in the variation of the phosphorylation ratios between these distinct structural groups (ANOVA p-value 3.02E-09). Sites within ordered structural environments appeared to be subjected to the lowest level of regulation during the cell cycle (median 1.65). Interestingly, a distinction emerged between coils (median 1.83) and disordered regions (median 2.22), signifying that the latter exhibited the largest variation in phosphorylation changes. We speculate that the increased variation of phosphorylation in longer, disordered coils correlates with their higher solvent exposure, which makes them more easily accessible for both kinases and phosphatases. Overall, our data shows that the phosphorylation variation of a site clearly scales with the level of order of its structural context (i.e. the tendency of a site to be found within a regular, irregular or disordered region).

3.4 Amino acid content of flanking regions of phosphorylation sites

We wanted to investigate if sites with distinct phosphorylation patterns over the cell cycle differ not only according to structural context, but also with respect to the amino acid content in their local sequence environment. A two-sample logo [233] was computed to contrast the two data sets, using the highly variable sites as a negative set (Figure 3). For each position and each possible amino acid, a two sample t-test was used to evaluate the null-hypothesis that the vectors of residues at a given position in both the positive and negative data sets (i.e. low and high variation) come from the same distribution. We found statistically significant enrichment of charged amino acids and depletion of proline, serine and threonine in the surrounding of sites with small phosphorylation variability (p-value ¡ 0.05). Additional comparisons of the amino acid distributions of the two sets against a background distribution accounting for structural differences using the composition profiling technique [234] revealed similar trends (see Text S1 for the detailed analysis and results).

The enrichment of serine and threonine residues in the vicinity of the detected phosphorylation sites could correlate with additional modification events. To check this hypothesis, we determined if multiple phosphorylation sites are found with higher preference in disordered regions. Phosphorylation sites that had at least

one neighboring phosphorylation site in both ordered and disordered regions were compared. A neighbor was defined as any phosphorylated residue that lies within +/-1,2,3,4, or 5 residue-long flanking region of a given modification site. Regardless of which of these five cut-offs was chosen, multiple phosphorylation sites were always highly significantly enriched in disordered regions (Table 3.1).

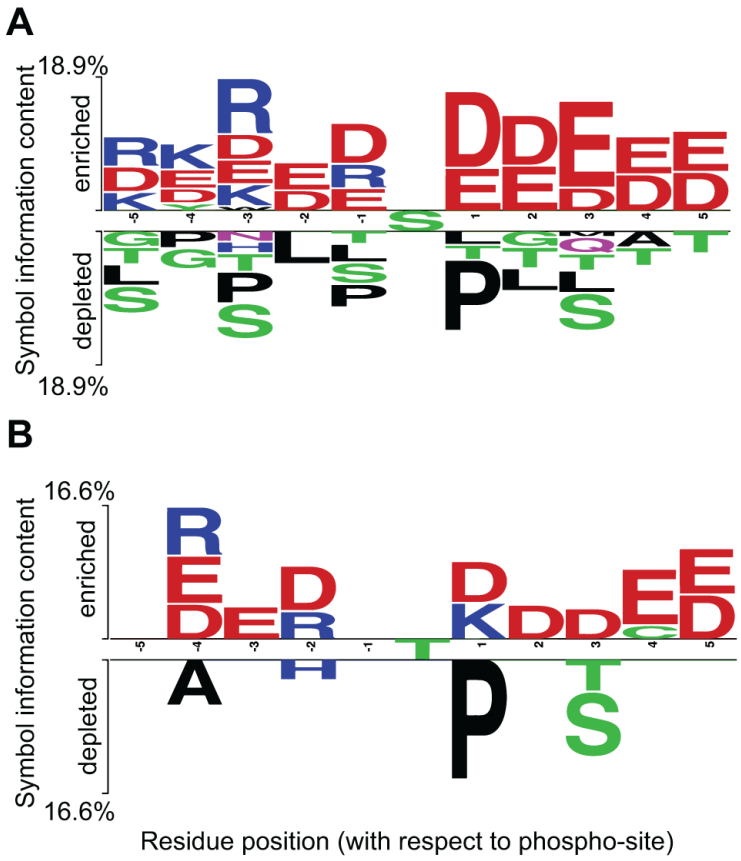


Figure 3.3: ***Two sample logo of flanking regions of phosphorylation sites of low versus high phosphorylation variation.*** *Amino acids in the top and bottom parts* ***(A)*** *central residue serine and* ***(B)*** *central residue threonine) represent residues, which are enriched or depleted correspondingly in the flanking regions of sites with small phosphorylation variation. Strong preferences are found for charged residues such as arginine, aspartate, and glutamate. In contrast, the majority of the amino acids that are more frequent in the negative set (i.e. variable phosphorylation set) are disorder-related e.g. proline, serine and glycine.*

Table 3.1: *Enrichment of multiple phosphorylation sites in disordered regions.*

Distance+-	Odds ratio	P-value
1	1.45	8.15E-04
2	1.64	1.80E-07
3	1.69	6.86E-08
4	1.90	3.72E-10
5	1.96	2.36E-10

The enrichment of additional phosphorylation sites at different distances from the central modified residues was computed. The odds ratios were calculated with the Fishers Exact test implemented in R. Multiple phosphorylation sites were found significantly more often in disordered regions for any of the considered distances.

3.5 Conservation of phospho-sites with different structural context

Next, we were interested in potential differences in evolutionary constraints on the phospho-sites in structured and disordered regions. When analyzing conservation it is important to take into account the different evolutionary rates of disordered and ordered regions. We therefore compared conservation scores between phosphorylated serines, threonines and tyrosines with control serine, threonine and tyrosine residues with a similar structural background. We define the set of control residues, as all potential phosphorylation sites that were not found to be phosphorylated in the study of Olsen et al [61].

As expected, phospho-sites that were predicted to lie in regular regions appeared significantly more conserved than phospho-sites in disordered regions (p-value 3.23E-120), due to the more conserved structural background of the former (Figure 4). In agreement with a previous study [82] modified residues in regions that lack defined structure were more conserved than the control serine, threonine and tyrosine residues with the same surrounding environment (Mann-Whitney Wilcoxon test p-value 3.4E-03). The same holds true for phospho-serine, phospho-threonine and phospho-tyrosine in ordered regions as compared to their equivalent control sets (p-value 2.24E-16). Despite the small size of the effect (groups means -0.38, -0.28, 0.14

and 0.22 for pS/pT/pY ordered, S/T/Y ordered, pS/pT/pY disordered and S/T/Y disordered respectively) the higher evolutionary pressure on phosphorylated residues suggests functional importance of these sites in a broad range of species.

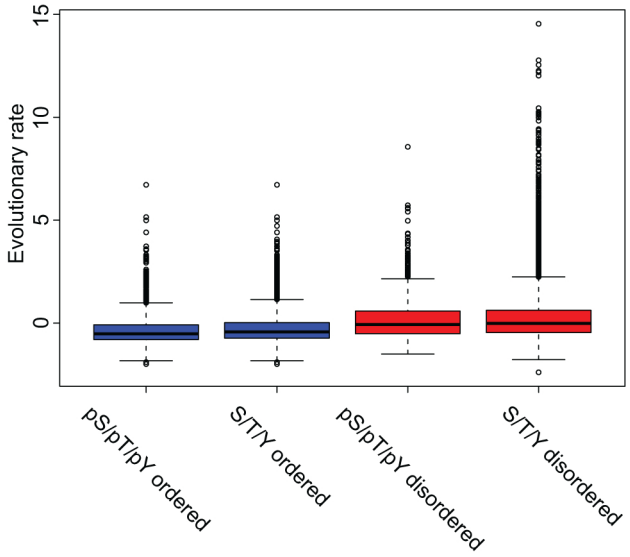


Figure 3.4: ***Conservation of phosphorylated sites versus conservation of control sites taking into account local structure.*** *Lower values correspond to slower evolutionary rate and higher conservation. Phosphorylation sites predicted to lie within regular structures (in blue, pS/pT/pY regular) appeared to be more conserved than their equivalent non-phosphorylated residues from the same proteins (p-value 2.24e-16). The same tendency was present for modified sites in disordered regions (in red, pS/pT/pY disordered), which were also subjected to a statistically significant slower evolutionary rate than their control set (p-value 3.4E-03). Phosphorylation sites in regular structures showed higher conservation than that of phosphorylation sites in irregular structures (p-value 3.23E-120).*

3.6 Motif decomposition with the 2D annotation enrichment technique

We next asked if different groups of kinases would exhibit preferences for less variable or highly variable phosphorylation sites. To identify kinase recognition motifs that show similar behavior with respect to two variables protein disorder and phosphorylation variation, we used the recently described 2D Annotation Enrichment technique (see Materials and methods and [235]). It employs a two dimensional gen-

eralization of the nonparametric two-sample test to detect preferences of a certain group of elements for two numerical attributes simultaneously relative to all other elements. The motifs separation is plotted in Figure 5 (the complete data are available in Table 2). The general trend between disorder and phosphorylation variation is reflected in the plot as sites with more disordered background show also higher variability. For individual kinases a very clear separation reflecting their preference for specific amino acids in their consensus motifs becomes apparent. Overall, four classes can be distinguished: (i) tyrosine kinases (black squares), (ii) proline-directed kinases (red circles), (iii) non-proline directed kinases with charged residues in their substrate recognition motif (green and blue triangles) and (iv) proline-oriented kinases, which contain a proline residue in their motif (red triangles and pentagons).

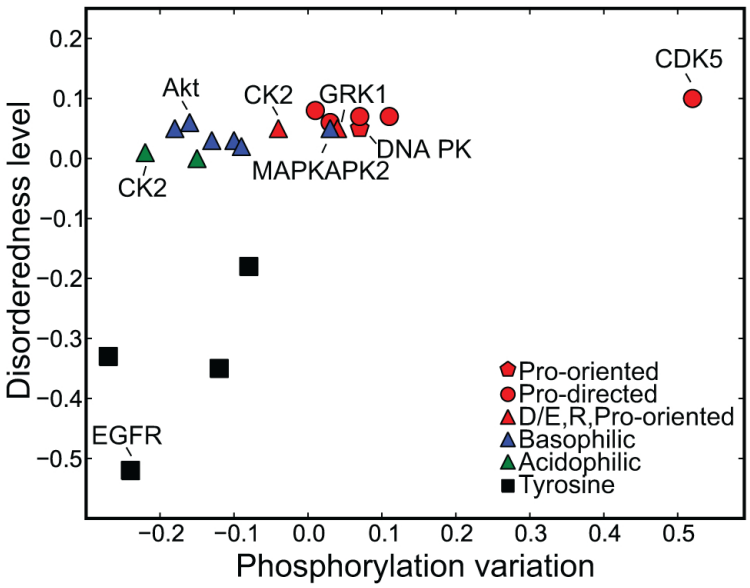


Figure 3.5: ***Kinase motif decomposition based on phosphorylation variability and structural preferences.*** *The preferences of various kinases for sites with specific structural background and phosphorylation variation were calculated by the 2D Annotation Enrichment technique (see Materials and methods). In general four classes can be distinguished: (i) tyrosine kinases (black squares), (ii) proline-directed kinases (red circles), (iii) non-proline directed kinases with charged residues in their substrate recognition motif (green and blue triangles corresponding to acidophilic and basophilic kinases respectively) and (iv) proline-oriented kinases, which contain a proline residue in their motif at position different from +1 relative to the modification site (red triangles and pentagons).*

The class of tyrosine kinases shows a strong preference for low phosphorylation

variability and structured regions, whereas the other three classes favor more disordered regions, but span a wider range of phosphorylation variation values. Also among the latter three groups higher quantitative variability is clearly associated with higher disorder. As seen in Figure 5, basophilic and acidophilic kinases occupy the regions on the graph corresponding to low phosphorylation variation, whereas proline-directed kinases are located on the right part of the graph, demonstrating their preference for more variable sites. The motif corresponding to the highest variability is the consensus motif for the proline-directed CDK5 kinase, which is in agreement with the important regulatory role of this enzyme during the cell cycle.

The proline-related class represented by two acidophilic, one basophilic and one atypical protein kinases shows preferences for intermediate level of disorderedness and phosphorylation variation properties. The Casein kinase II is characterized by various substrate recognition motifs [236], but the main differences are related to presence or absence of a proline residue preceding the phosphorylated site. These two motifs show distinct structural and variation preferences the former type being more similar to proline-directed kinases and the latter to non-proline ones. The occurrence of the G protein-coupled receptor kinase 1 (GRK1) near the proline-directed kinases (red triangle) can be explained analogously by the presence of a proline residue in the consensus sequence for that kinase. Interestingly, the reported consensus motifs of the MAPKAPK2 kinase (blue triangle) do not contain proline residues, however, it is still grouped together with the more variable kinase motifs. After careful examination of the amino acid composition of substrates of the MAPKAPK2 kinase in our data set we found multiple examples that contained proline within +/-6 residue window around the phospho-site. Together with our structural analysis, this suggests that this residue may play an important role in the substrate recognition. Overall, the group of proline-oriented kinases has similar preferences for disorder and phosphorylation variability as the proline-directed group. This observation also extends to the functional relevance of the member kinases to the regulation of the cell cycle. For example, the DNA-dependent protein kinase (DNA-PK) is involved in stress response and DNA repair and is known to play a role in the progression of the cell cycle [237]. Furthermore, the MAPKAPK2 kinase is involved in DNA repair processes and thus can provide an alternative to checkpoints activation [238].

Proline-directed kinases such as PLK1 are known to actively regulate the progression of the cell division cycle, thus implying that disordered regions (which are enriched for prolines) are subjected to regulation and therefore to variable phosphorylation patterns. We therefore checked if the tendency of phosphorylation variability to scale with the level of disorder persists if we control for proline-directed kinases and excluded all sites modified by such from the data set. Indeed, the effect of lower phosphorylation variability being associated with ordered regions and higher - with disordered regions remained the same.

Table 3.2: *Enrichment of kinase recognition motifs with specific preferences for disorder and phosphorylation variation.*

Variation	Disorder	Benj. Hoch. FDR	Names
-0.16	0.06	3.20E-03	Akt kinase
-0.27	-0.33	7.67E-05	ALK kinase
-0.15	0.00	6.19E-05	b-Adrenergic Receptor kinase
-0.13	0.03	7.49E-05	Calmodulin-dependent protein kinase II
-0.22	0.01	2.71E-10	Casein kinase II
-0.04	0.05	1.88E-05	Casein Kinase II
0.52	0.10	6.32E-03	CDK5 kinase
0.07	0.05	2.88E-03	DNA dependent Protein kinase
-0.24	-0.52	5.72E-05	EGFR kinase
0.07	0.07	6.97E-07	ERK1,2 Kinase
0.04	0.05	2.15E-05	G protein-coupled receptor kinase 1
0.01	0.08	5.73E-04	Growth associated histone HI kinase
0.03	0.06	1.76E-03	GSK3 kinase
0.11	0.07	1.04E-09	GSK-3, ERK1, ERK2, CDK5
-0.12	-0.35	1.76E-03	JAK2 kinase
-0.18	0.05	1.84E-03	MAPKAPK1 kinase
0.03	0.05	4.63E-03	MAPKAPK2 kinase
-0.10	0.03	1.26E-05	PKA kinase
-0.09	0.02	8.09E-05	PKC kinase
-0.08	-0.18	1.83E-04	Src kinase

The enrichment of kinase substrate motifs with specific preferences for disorder and phosphorylation variation was calculated with the 2D annotation enrichment technique [235].

3.7 Discussion

Previous studies had already found a prefeence of phosphorylation to occur in loops or disordered regions [81, 80]. However, those studies generally did not have access to the dynamics of phosphorylation and they therefore based their analysis on the absence or presence of phosphorylation sites alone. Here, we instead made use of a large-scale quantitative phosphorylation data set to investigate a possible relation between the structural features of phosphorylation sites with their degree of regulation. This allowed us to contrast the behavior of less variable sites to those that were dynamically regulated. Our data clearly demonstrate that the propensity of phosphorylation sites to be regulated during the cell division cycle is related to the level of structural organization of the environment in which these sites reside. Furthermore, we discovered that this effect occurs in a graded manner: regions with regular structure are least likely to harbor regulated phosphorylation sites, followed by irregular regions (short loops or random coils). Note that over 90% of the sites were found within disordered structures and their high phosphorylation variability relates them to regulated phosphorylation events.

Interestingly, the sets of sites within ordered loops and disordered structures showed significant differences. It has been shown before that different flavors of disordered regions exist with regards to their lengths, amino acid composition, and the conformational transitions that they undergo upon binding [239, 182]. Liu et al. defined regions with no regular secondary structures (NORs) as one specific category of disordered regions. They demonstrated that NORs differ significantly from regular structured loops and argued that these might have different functional implications, a hypothesis which finds support in our study.

Functional analysis of the highly variable set of sites revealed enrichment of cell cycle-related, biosynthesis and cellular organization and localization processes (Table S1). Some examples are RNA, DNA and mRNA processing, localization and transport, regulation of gene expression and biosynthesis. Cell cycle-associated processes such as regulation of the different phases of the cycle, DNA replication and repair, telomere organization and maintenance and chromatin assembly were also strongly over-represented in the variable set of sites.

Phosphorylation is an important mechanism for regulation of a myriad of intricate processes during cell division. A detailed study of the cell cycle regulation through phosphorylation focused on functional analysis of protein groups that are up or down regulated at specific time points [61]. These were the proteins that contained sites that reached phosphorylation peaks at S or M phases. As expected, proteins involved in mitotic and cell cycle processes were shown to be maximally phosphorylated at mitosis. Interestingly, Olsen et al. found proteins that regulate metabolic processes to be weakly phosphorylated during S phase and highly phosphorylated at mitosis. An explanation to this discovery is the possibly inhibitory character of phosphorylation on proteins that regulate metabolic processes, as protein synthesis and related functions tend to shut down during mitosis. Furthermore, DNA replication takes place during S phase, which rationalizes the up-regulation through phosphorylation of various proteins involved in DNA replication repair. High phosphorylation of cytokinesis-related proteins in S phase appears to play an important role in the control of the correct segregation of the two daughter cells.

The tendency of modification sites in regular structures to be less variable may be facilitated by proximal charged residues acting as stabilizers of the phosphate group. Charged flanking regions offer a suitable environment for hosting a phosphate group and allow for favorable interactions that potentially result in phosphorylation acting on a longer time scale. For instance, these favorable interactions could reduce the efficiency of phosphatases in removing a phosphate group, thereby contributing to the tendency for smaller variation in the phosphorylation level that we observe in our data. In contrast, negatively charged residues could lead to repulsion-driven conformational changes and polarization of the entire protein surface by creating clusters of negatively charged residues.

Several mechanisms that are known from literature furthermore contribute to the observed tendency for structural rather than regulatory phosphorylation sites to be present in ordered regions. Specific structural changes due to phosphorylation include stabilization of the N-termini of -helices via favorable interactions of the added phosphate group with the helix backbone [240]. This is effected by the interaction of the phospho group with the helix dipole moment. Yet, the same modification introduced at the C-terminus would have the opposite effect [241]. The optimal stabilizing position for the phosphate group was estimated as -2 relative to the N-cap

of a helix. Additionally, favorable electrostatic interactions between proximal positively charged residues (e.g. at a helix cap) and the phosphate group can enhance helix formation. The stabilizing effect of salt bridges formed between a phosphate group and a lysine side-chain has been recognized as one of the strongest possible -helix inducers [242]. In contrast, the phosphate-guanidinium interaction leads to disruption of the local regular structure [243]. Phosphorylation has also been reported to cause conformational changes in -sheets and disruption of -hairpins. In those cases repulsive interactions with an aromatic tryptophan residue in the spatial vicinity of the phospho-site are observed [244]. A related question that arises from our investigation is to what extent the phosphorylation variability of a site is connected to a role in the overall structural re-arrangements of a protein. A phosphorylation event can alter the energy that is required for a conformational change [232], and thus hinder or facilitate it. Further experiments including 3D structural information or computational models are needed to increase our understanding of the interplay between structure and phosphorylation.

Multiple experimental studies show the regulatory role of modification sites that show variation in their phosphorylation patterns and lie within intrinsically disordered regions For example, the cyclin-dependent kinase inhibitor 1B (p27) is an intrinsically unstructured protein, which is multiply phosphorylated and regulates the cell cycle by inhibition of cyclin-dependent kinases (CDKs) [245]. The disorderedness of p27 plays an important role in keeping the complex formed between CDK and p27 flexible. Due to this flexibility the segment, which blocks the ATP binding site becomes exposed. This allows a tyrosine residue to become accessible for phosphorylation, upon which the space previously occupied by the inhibitor becomes available for ATP binding. Then the partially reactivated CDK phosphorylates p27 at another residue, which leads to its degradation and allows CDK to regain full activity and guide the progression through the cell cycle [246].

In another example, multiple phosphorylation sites on the transcription regulator Retinoblastoma protein (Rb) influence its ability to interact with transcription factors and other regulatory proteins. A detailed structural study reports that the different phospho-sites found within disordered regions induce distinct conformational changes and also serve different functional roles [247]. For instance, one of the modified residues decreases the affinity of Rb for binding the transcription factor

E2F by reordering the pocket domain. At the same time another modified site at a loop in the pocket domain induces complete blocking of E2F binding.

We found that the set of sites with varying phosphorylation patterns was enriched in amino acids associated with disorder, specifically Pro, Gly and Ser. Interestingly the same sites were more likely to have additional modified residues in their vicinity. Phosphorylation of a protein often occurs at several distinct residues and it has been reported that modification sites tend to cluster and function in a cooperative manner [248]. Mathematical models suggest that this phenomenon leads to an increase in the sensitivity and robustness of the cellular response [206] and may promote a switch-like behavior [203]. In such a case, the exact position of a modification site in a cluster would not be a determining factor on its own, but would rather contribute to a cumulative effect. It would be worth studying how different levels of phosphorylation variability in regions with different structural organization may be implicated in the cellular regulation of the cell cycle. Multiple phosphorylation sites with highly dynamic phosphorylation patterns may be suitable for both rapid and robust response. In contrast, the robustness of the response of sites within regular regions might be achieved on a longer time scale and be related to longer lasting effects of phosphorylation.

We showed that phosphorylated residues tend to be more conserved than their equivalent non-modified residues. Conservation of phosphorylated residues has been a broadly debated issue [82, 112], but the general consensus appears to be that the overall conservation of phospho-sites is low. Even though statistically it is significantly stronger than that of the equivalent non-modified sites, the effect size is relatively small. Possible explanations include (i) loss and gain of phospho-sites at different positions in disordered regions, likely due to clusters of sites acting as functional units regardless of the exact sequence position [249] and (ii) potential silent phosphorylation events [112].

The idea that it is the cluster of phosphorylation sites that plays a functional role is becoming increasingly accepted [248, 206]. The functional roles of multiple phosphorylated residues span a wide range: (i) targeting for sub-cellular localization, (ii) targeting for degradation, (iii) control of protein-protein and protein-nucleic acid interactions (often through electrostatic effects) and (iv) enhancement of a robust

and rapid response to a stimulus [54]. Furthermore, mechanisms of priming phosphorylation are also well-known [250].

Here we showed that disordered regions harbor variable sites, which tend to be surrounded by additional phosphorylated sites. This raises the possibility that the variability of these sites is related to some of the above-described phenomena. It is known that disordered regions can facilitate a large number of interaction partners, and that multiple sites can control their association and dissociation. Given the wide range of functions of multiple phosphorylated sites in disordered regions, a larger variability in their phosphorylation patterns may provide an adequate functional mechanism to effect the desired regulation. In contrast, structural regularity imposes certain constrains on the less variable sites. The necessity of evolutionary conservation of the structure tends to prevent the accumulation of disorder-associated serine and threonine residues and a consequent change of their positions. Furthermore, the more rigid structure implies a more limited number of interactions partners. Therefore, we reason that the requirement for regulation for these sites in structured regions can be smaller.

Our data allowed us to investigate the kinase preferences of phosphorylation sites with high vs. low levels of regulation. Tyrosine kinases and kinases that require charged residues in their substrate recognition motives clearly preferred sites with smaller phosphorylation variation, whereas proline-directed kinases were clearly associated with sites that were dynamically regulated. Proline is known to be a helix and sheet breaker, due to the planarity of its side-chain. Proline lacks an NH backbone donor to form a hydrogen bond and thus disrupts the formation of regular hydrogen bond patterns, which are the basis of regular structure formation. Due to its unique stereochemistry the proline residue can adopt two different conformational states cis and trans and a large number of folded proteins contain both states of the residue. The intrinsic conformational changes resulting from the proline isomerization play an important role in determining the function, ligand recognition and interactions of the protein [251]. For instance, certain kinases, such as MAPKs and CDK2 preferentially modify substrates with the trans isomer [75]. Proline isomerization in a S/TP motif, where S/T is phosphorylated, can also control the opposite step dephosphorylation, as some phosphatases appear to be conformation-specific and prefer the trans state [252]. Therefore, the preference of proline-directed kinases

for sites with higher variation illustrates a connection between dynamic regulation and disordered regions.

3.8 Materials and methods

Computation of phosphorylation sites with high and low variability over the cell cycle

In the data set underlying our analysis [65], human HeLa S3 cells were labeled with SILAC [5, 32] to produce three different isotopic forms of lysine and arginine (light, medium and heavy). The light and heavy isotopes were synchronized in six different stages of the cell cycle, while the medium one was kept non-synchronized as a reference. Relative quantification of protein abundances (protein ratios) and/or phosphorylation (phospho-peptide ratios) were computed by taking the ratio between two cell states at each time point (i.e. synchronized heavy-labeled cells in S phase and non-synchronized medium-labeled cells). In order to account for the possible influence of protein abundance, changes in the phosphorylation ratios between the reference and the stimulated cells were normalized by the protein change. We mainly focused on the phosphorylation ratios as they were available for a larger number of sites compared to the absolute occupancy values.

The data set contained information about the UniProt id of the phosphorylated protein, sequence positions of phosphorylated residues, and quantitative measures of phosphorylation (normalized phosphorylation ratio) at 6 time points (i.e. cell cycle phases: G1, G1S, Early S, Late S, G2 and M). In total 1,059 proteins and 5,173 phosphorylation sites with measured phospho-ratios for each of the six time points of the cell cycle were used in the analysis.

In order to assure that the observed phenomena are not due to the properties of the chosen subset of sites, we repeated the analysis of phosphorylation variation between different structural groups with data sets containing five (5,254 sites), four (8,537 sites), and three (8,731 sites) time points only (see Text S1 for details). Although slight fluctuations were observed, the main tendencies remained stable and the conclusions did not change. Therefore, no bias in the reduced data set (i.e. the one

containing information about all six time points) was found.

Phosphorylation variation value for each modified site was computed as the standard deviation of the phosphorylation ratios over the six time points. High variation corresponds to sites with temporal variation of phosphorylation ratios (e.g. a peak is observed in S phase), while low variation describes those sites that retain constant or slightly variable phosphorylation fold change during the cell cycle.

Structure prediction and structural categories

The secondary structure of each site was predicted with PsiPred [169]. Each site was assigned one of three possible states: H for -helix (92 sites) E for -sheet (53 sites), and C for random coil, turn or loop region (5,028 sites).

An intrinsic disordered state was also predicted for each site using DISOPRED [168] with standard settings. We found 498 sites to be in the order state while the remaining 4,675 were predicted to be in the disorder state.

Based on a combination of secondary structure and disorder predictions, we defined three distinct structural categories for each phosphorylated site: (i) regular regions (helices and sheets in ordered regions, 145 sites), (ii) irregular regions (coils in ordered regions, 353 sites), and (iii) disordered regions (coils in disordered regions, 4,675 sites).

Statistics

Statistical analyses were performed within the R environment [170] and using the in-house statistics work frame Perseus. The lattice package was utilized for comparing distributions of phosphorylation variation in different structural categories. Differences between distributions were assessed with the standard non-parametric Kolmogorov-Smirnov test. In the case of three structural categories, analysis of variance of the phosphorylation fold change was performed using the structural category as an independent variable. Data on phosphorylation site variation and structure predictions are available in the Supporting material (Table S2). Enrichment of functional Gene Ontology (GO) categories was performed with the GOrilla tool [253].

Evolutionary analysis of phosphorylation sites

We performed conservation analysis on phosphorylated residues in ordered and dis-

ordered regions. The proteins from our data set were mapped to pre-computed EggNOG groups of orthologs [180]. We used the maximum likelihood-based rate4site algorithm to build phylogenetic trees from the EggNOG clusters and to compute residue-based evolutionary rates [181]. Lower evolutionary scores correspond to stronger conservation.

The control sets of sites were defined as all serine, threonine and tyrosine residues from the phospho-proteins that were not measured to be phosphorylated in our data set with equivalent structural background (i.e. disordered and ordered as predicted by PsiPred [169]).

Enrichment of proximal phosphorylation sites in disordered regions
We tested if disordered regions are enriched in multi-phosphorylation sites, as compared to ordered regions. We considered phosphorylation sites with at least one modified neighbor as multi phospho-sites. A neighbor residue is defined as a phosphorylated serine, threonine or tyrosine located within +/- 1,2,3,4 or 5 residue-long flanking regions of a central phospho-site. For each cut-off length, we built a contingency table. Each contingency table contained the number of sites with and without neighboring phospho-sites for both ordered and disordered regions. The significance of the enrichment was estimated with the Fishers Exact Test.

2D Annotation Enrichment Technique
The 2D Annotation Enrichment technique [235] enables analysis of the preference of a certain group of elements (i.e. phosphorylation sites, characterized by the same consensus motif) for two numerical attributes simultaneously relative to all other elements (in our case all other phosphorylation sites). It employs a two dimensional generalization of the nonparametric two-sample test and uses the Benjamini-Hochberg method to correct for multiple hypotheses testing. We used the default settings to distinguish the statistically significant groups, corresponding to false discovery rate ¡ 0.01. We used the Human Protein Reference Database motif definitions in this analysis [236].

Two sample logos
The difference between the amino acid content of the flanking regions of the sites with low and the sites with high phosphorylation variation was computed, assessed

and visualized with the help of the Two Sample Logo method [233]. The highly variable set was used as the negative set. Residues significantly enriched in a certain position are shown above the horizontal line in the logo.

Chapter 4

Framework for efficient feature selection and classification of cancer proteome profiles

4.1 Introduction

Clustering techniques have been shown to perform well on proteome profiles derived from cell lines due to the relatively low sample complexity and the lack of high biological variability. For example, Geiger et al. studied the quantitative differences in the proteomes of cell lines from different stages of breast cancer [254]. First they used T-test-based statistics to determine the proteins that showed significant variations. From the total of 7,800 proteins approximately 50% were found to significantly change their expression levels and these were retained for further analysis. Using unsupervised hierarchical clustering the authors were able to distinguish two main groups that exactly corresponded to tne basal and luminal subtypes. Subsequent analysis using T-test statistics resulted in the identification of potential biomarkers. In another study the proteomes of two diffuse large B-cell lymphoma subtypes were compared [127]. Clear subtype segregation was achieved with hierarchical clustering of the expression profiles of the samples and the groups were shown to be separable by the first component of a principal component analysis. The actual complexity of comparative analysis of proteomics experiments was illustrated in the study by Wisniewski et al. [132]. The proteomes of tissue samples from normal mucosa, pri-

mary and metastatic colon cancers were compared. More than 7,500 proteins were quantified using the MaxQuant label free quantification algorithm [6]. Between the normal mucosa and the two cancer tissue types more than 1,800 proteins were significantly differentially expressed. These strong differences were easily detected in an unsupervised comparison, in which the normal tissues clustered together instead of clustering according to patients of origin. The unsupervised approach, however, failed to detect the subtle differences between the proteomes of the primary and the metastatic cancer sets. Instead, in this case the samples were grouped together in corresponding pairs (primary and its nodal metastasis coming from the same patient). This result exemplifies the limitations of unsupervised techniques in cases when the signal due to biological variability is much stronger than that coming from the disease stage and clearly demonstrates the need for more sophisticated analysis methods (probably in addition to larger data sets).

Supervised learning methods are now becoming the state-of-the-art techniques in the analysis of large scale mass spectrometry-based data. Decision trees [255, 256], Bayesian neural networks [257] and support vector machines (SVMs) [136] are among the methods that are used. However, in order to overcome the limitations associated with the nature of the data, such methods need to be combined with efficient feature selection and noise reduction techniques [258, 259]. Furthermore, special care has to be taken to assure that all methods are used correctly and overfitting during the evaluation of the classifier's accuracy is minimized.

In this chapter the performance of SVMs in combination with efficient feature selection techniques is explored and discussed in detail. A feature ranking technique based on the weights of the instances computed during the training of a classifier is implemented in a recursive manner to improve the quality and the relevance of the selected features. All methods are integrated as a plug-in into the Perseus framework, which enables downstream analysis of large-scale proteomics data. This 'Learning' plug-in supports classification and prediction, feature selection and parameter optimization and allows any desired combination between the implemented methods. The user is prompted to use cross validation during both processes of classification and feature selection in order to guarantee good generalizability of the results. The dataframe used in the software is a matrix containing instances in the columns and features in the rows by default, but a swapped orientation is also supported. The

output is also in the form of matrices, which can be conveniently used for further functional analysis in the Perseus framework.

4.1.1 Support Vector Machines (SVMs)

The Support Vector Machines technique allows building accurate classifiers in both linearly-separable and linearly-inseparable problems. The binary nature of SVMs, however, requires their extension to multi-class problems. In the simplest case of a two-class problem with linearly-separable classes, a classifier is trained to find a linear decision boundary (a hyperplane) in a high-dimensional space that separates the groups in the data. The decision hyperplane is defined by a weights vector w (a normal perpendicular to the hyperplane) and an intercept term b (Eq.4.1), where x denotes the set of feature vectors. As numerous such hyperplanes exist, the main task becomes the maximization of the distance between the hyperplane and the nearest training examples, known as margin maximization, where the margin equals $2/|w|$. The instances lying on and within the margin constitute the support vectors, which are used in the actual prediction of the class of a new unlabeled instance. The margin size can be further controlled to allow for misclassifications in the training set with appropriate penalties (Eq. 4.2). Using the so-called soft margin increases the classifier's generalizabilty. The size of the soft margin is controlled by the penalty parameter C (large values correspond to large penalties for misclassification and resemble a hard margin classifier) and a slack variable ξ which measures the degree of misclassification.

$$D(x) = w \cdot x + b \tag{4.1}$$

where w is the weights vector and b is a bias value.

$$min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{4.2}$$

$$\text{subject to} \quad y_i(w^T x_i + b) < 1 - \xi_i \quad and \quad \xi \geq 0 \tag{4.3}$$

In the case of linearly separable classes, the decision function is a linear discriminant function based on the weighted sums of the training instances plus some bias (Eq 4.1) and all instances can be separated without errors. Often the underlying patterns in the data do not allow linear separation and instead require the definition of complex

functions to build a good classifier. Support vector machines use kernels to deal with such situations. With the help of kernels the original finite space is mapped to a high dimensional feature space. The hyperplanes in the higher dimensional space are represented by all points defining a set, whose inner product with a vector in that space is constant. As the SVMs depend on the data only through dot products, it is possible to compute the dot products even at a high dimension at low cost by applying the so-called kernel trick, i.e. the dot product is replaced by a kernel function. The kernel functions should be such that the distance between any two points in space x_i and x_j is defined in the transformed space and has a relation to the distance in the original space. Some of the most commonly used kernels are:

$$linear: \quad K(x_i, x_j) = x_i^T x_j \tag{4.4}$$

$$sigmoid: \quad K(x_i, x_j) = tanh(\gamma x_i^T x_j + r) \tag{4.5}$$

$$radial \quad basis: \quad K(x_i, x_j) = exp(-\gamma |x_i - x_j|^2), \gamma > 0 \tag{4.6}$$

$$polynomial: \quad K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0, \tag{4.7}$$

where γ is the slope, d is the degree of the polynomial and r is a constant.

A more detailed explanation of the SVMs can be found in the following SVM methods papers [163, 164, 260, 261].

4.1.2 Feature selection

The importance of feature selection for classification has long been recognized [258, 259]. Using a subset of the total number of features often leads to an improvement in the prediction performance of the classifier as noisy and irrelevant features are discarded. Furthermore, focusing only on highly-relevant and specific features may help to gain a deeper understanding of the underlying patterns in the data. Training a prediction with a smaller number of features further improves its speed and efficiency. Dimensionality reduction is particularly useful in classification problems emerging from proteomics data, as they are often characterized by a relatively small sample size and a large feature space. Essentially, two main applications of feature selection are distinguishable: (i) dimensionality reduction to create a more accurate and generalizable predictor and (ii) discovery of a small subset of features that discriminate well between the given classes and in the case of clinical proteomics can

be utilized as a disease-related signature.

Based on how they interact with the classification model, feature selection algorithms fall in three large categories: wrapper, filter and embedded (for more details see [262]). The filter approach is applied prior to the classification, is independent from the classification method and makes use only of the general characteristics of the data. The ability of each feature to discriminate between the different classes is estimated separately regardless of its interactions with the other features often using statistical tests such as ANOVA. All features meeting certain criteria are retained during the training and testing of the classifier. In contrast, wrapper techniques require a predefined classification method. The classification method is used as a black box in the evaluation of the predictive power of a feature or a set of features. The assessment of the goodness of the selected set is usually done in a cross-validation procedure. Embedded techniques incorporate the feature selection process directly into the training procedure of the classifier.

A common problem of numerous microarray and proteomics classification studies is the inaccurate use of feature selection procedures. Often feature selection is performed prior to the classification and the evaluation of the predictive power of the selected subset of feature is done on the complete training set. This practice results in overfitting, imprecise estimation of the classifier's accuracy and the identification of biologically-irrelevant or simply unspecific proteins. One possible way to overcome this problem and to increase the number of the selected disease-associated proteins is embedding the feature selection in a cross validation procedure. In this way, estimation of the classifier's accuracy is always performed on a subset of the entire set of instances (i.e. within a cross-validation procedure) regardless of the choice of a feature ranking method. This procedure is followed here.

4.2 Recursive feature elimination embedded in cross validation

An outline of the feature selection process using the SVM weights-based ranking method (see Materials and methods section 4.4) and the generation of a ranked features list is shown in Fig. 4.1. In the first step the data are split into q fractions

(in the case of Leave-one-out cross validation q equals the number of instances). The training set is always formed by $q-1$ fractions and the remaining fraction is used as a test set. Only the examples from the training set are used in the feature ranking process.

We have generalized the Recursive feature elimination (RFE) procedure [162] for multi-class problems as outlined in Figure 4.1. There are two methods for handling multi-class data: One-versus-Rest and Each-versus-Each. In the first approach, the number of trained classifiers equals the number of classes, as each classifier is designed to separate one particular class from all the rest. In the second approach, $c(c-1)$ binary classifiers (corresponding to the number of possible combinations of two groups) are trained, where c equals the number of classes. For example, if the data contain 3 classes, in the One-versus-Rest approach 3 binary classifiers are trained. Each of the classifiers is used to rank all features using the SVM-RFE procedure described in the Material and methods section of this chapter, resulting in 3 separate lists of ranked features.The three lists are then united into a single final ranked list as outlined in Figure 4.1. First the best feature from the first list is added to the final ranked list, then the best feature from the second list and so on until the best features from all lists are included. Next, the second best feature from the first list is added and this process continues until all features have been added to the final list (double entries are not allowed).The united ranked list is then used to determine the optimal number of features to be used in training a classifier. To do so, subsets of ranked features of different sizes are used for training and testing and the classifier's accuracy is recorded. The optimal number of features is determined in a cross-validation procedure as outlined next.

Sample set: X with size n, feature set: F
Procedure outline:

1 : **for** i in $(0:n)$
2 : $X_{train} = X(0:i; i+1:n); \quad X_{test} = X_i$
3 : Rank all features F in X_{train}
4 : **for** f in $0:F$.
5 : Restrict X_{train}, X_{test} to f features
6 : Train(X_{train})
7 : Test(X_{test})

The dataset is split into training and test sets according to one of the standard cross validation procedures: Leave-one-Out, n-fold cross validation or random sampling. Then a feature selection procedure is applied on the training set. In the case of the SVM-RFE, a binary classifier is trained and the computed SVMs weights are used to calculate the features ranks. To determine the optimal number of features, the size of the feature subset is gradually varied with a defined step (starting from 1 and ending with the total number of features) and the accuracy of the trained classifier using these subsets of features is estimated on each test set. The procedure results in an error rate curve as a function of the number of features, allowing for accurate determination of the optimal size or a range of optimal feature set sizes (Fig. 4.1).

4.3 Implementation in the Perseus environment

To enable a larger group of scientists who are not necessary experts in the field of supervised learning to apply such analytical techniques, the above-described methods were implemented and integrated into the above mentioned statistical analysis environment Perseus. Perseus offers a user-friendly graphical interface, which integrates state-of-the-art statistical tools for functional analysis of large-scale proteomics data. It is one of the main cores of the MaxQuant suite for identification and quantification of peptides and proteins from mass spectrometric measurements. Upon protein identification and quantification with MaxQuant the user can perform a complete downstream analysis within the Perseus module. The software is implemented in C# in a practical format consisting of main core, to which plug-ins are added. This format allows easy integration of an unlimited number of independent tools that can then be used together. Already the software includes a wide range of functionalities, such as data transformation and normalization, statistical tests, functional annotations, enrichment tests, clustering methods and many others.

The supervised learning-related methods are implemented as part of the "Learning" plug-in and are designed to work with the rest of the tools. The standard data frame of Perseus is a matrix, in which instances (samples) are stored as columns and features (in the case of proteome studies - proteins) are stored as rows. Thus the input of the Learning plug-in is also a matrix, however, both orientations - samples as columns and rows are accepted. In a standard proteomics analysis workflow the

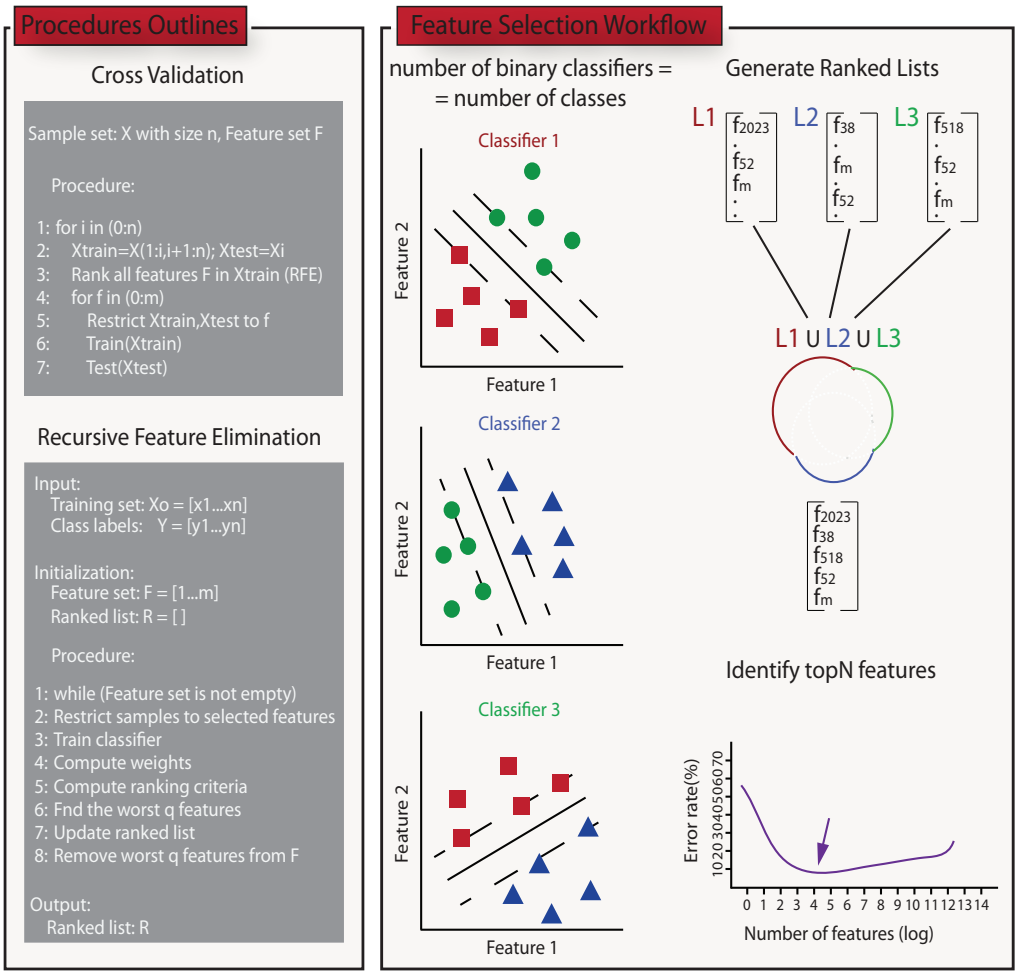


Figure 4.1: ***Outline of the feature selection framework with SVM-RFE implementation.*** *The feature selection is embedded in a corss-validation procedure to avoid overfitting. Feature ranking is performed on a subset of all samples and the prediction accuracy of subsets of ranked features of different sizes is then tested on the remaining test data set. The recursive feature elimination procedure allows optimization of the feature ranking at every next cycle as weak features are excluded and better decision functions are computed. Feature selection in multi-class problems is achieved by building c binary classifiers resulting in c ranked lists of features, which are then combined into a single final ranked list. The search for the optimal number of features is performed on this final list.*

MaxQuant pre-processed raw files can be uploaded into Perseus and subjected to downstream analysis (Fig. 4.2). The first step is always transformation and nor-

malization of the data in order to meet the data assumptions of various statistical methods and to diminish the influence of outliers. After normalization the data are ready to be used in the "Learning" module.

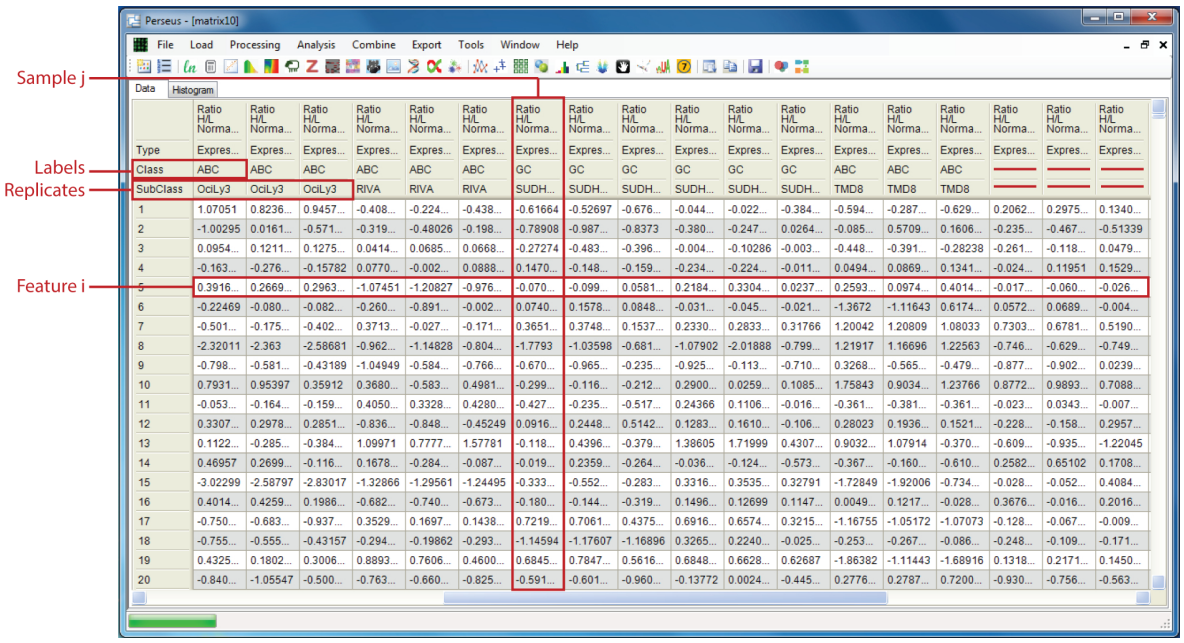


Figure 4.2: ***Loading of clinical proteomics data into Perseus.*** *Proteomics profiles of patients from two clinically-distinct lymphoma subtypes are loaded into Perseus. The patient samples are displayed in columns and each row contains protein expression over the different samples. The known classes are specified in the 'Class' annotation row: ABC and GC lymphoma subtype correspondingly. The second annotation row 'SubClass' contains information about replicate measurements (triplicates in the lymphoma example). The last three samples lack labels and are used as a test set.*

4.3.1 Classification form

Input form: The Classification option of the Learning module allows assessment of the classifier's accuracy and prediction of labels of new unlabeled instances (samples). Figure 4.3 shows the options available in the Classification form. The classifier's accuracy is estimated with the Cross-validation (CV) option. The standard CV methods are all implemented: n-fold with the number of fractions being modifiable by the user; Leave-one-out and Random sampling, where both the size of

the sample to be used for training and the number of samplings can be specified by the user. As the samples can be organized both in columns and rows the correct orientation of the matrix has to be specified together with the correct labeling of the samples (Fig. 4.3). Samples that lack labels are automatically used as a test

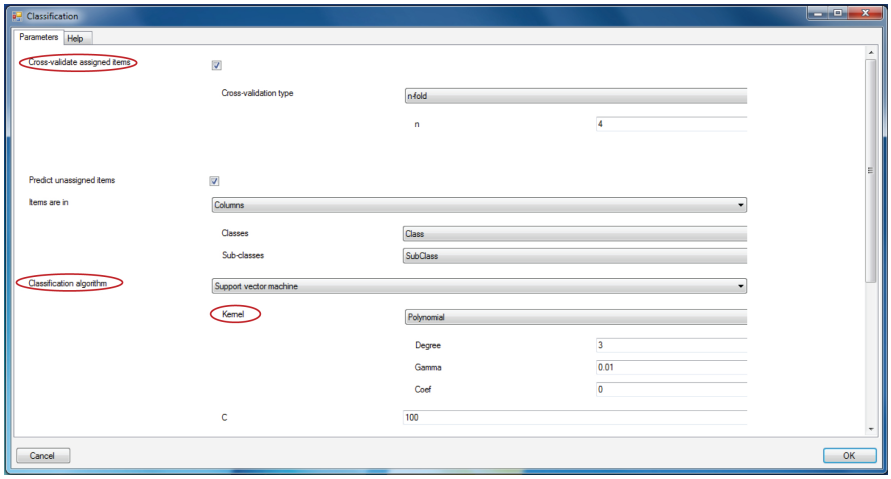


Figure 4.3: ***Perseus Classification input form.*** *In the example Classification with n-fold cross validation procedure is selected. The 'Predict unassigned items' option is selected to enable prediction of samples with unknown labels. The data organization is specified in the 'Items are in' fields as follows: 'Class' corresponds to the sample labels and 'SubClass' - to the replicates grouping of the samples. Support vector machines is chosen as the 'Classification algorithm' with Polynomial kernel and default parameter settings.*

dataset and their labels are predicted using the classifier trained on the labeled data. In case the data have been measured in replicates, second grouping indicating the replicate samples, has to be specified in the Sub-classes option. This information is then taken into account during the cross-validation procedure by always using all replicates together for training or testing, which is important to avoid overfitting. Several classification methods are currently supported: K Nearest Neighbors and Support Vector Machines. The LibSVM library ([263]) translated into C# is used as a basis for the SVM-based classification. The user can select from four kernels (discussed in the Materials and methods section of this chapter): linear, polynomial, radial basis function and sigmoidal. Upon selection of a kernel, the associated parameters become modifiable. An optional step is to include feature selection directly in the classification or cross validation processes. The user is required to specify the

desired number of features to be used in the learning process. The complete feature set is then ranked using the user-defined ranking parameters and the top n features are retained. The feature selection procedure is discussed in detail in Section 4.3.3.

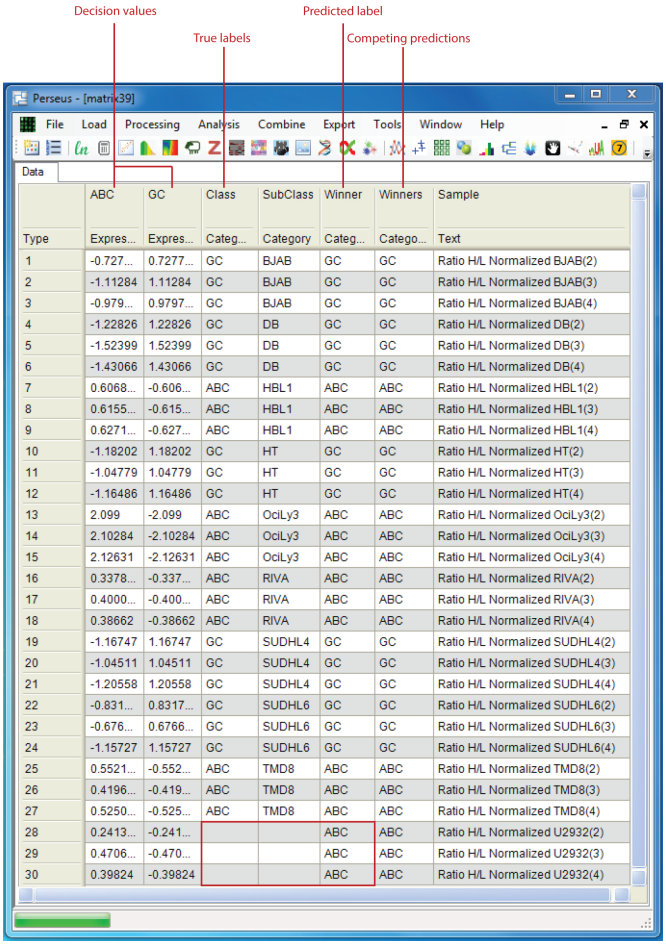
Output description: The output matrix is shown in Figure 4.4. The first columns contain the decision values generated during the training of the classifier and used in the class prediction. The original labels of the data (both main and replicates labeling), as well as the sample names are indicated. The actual prediction is stored in the 'Winner' column, whereas competing labels that have been assigned to the instance during the cross-validation procedure are displayed in the 'Winners' column. This misclassification column can be a good indicator of wrongly assigned labels in the training data, especially when a strong tendency towards a particular label is present.

4.3.2 Parameter optimization form

Performance of an SVM-based classifier depends on the choice of kernel function that best describes the patterns in the data. However, for optimal performance optimization of kernel-specific parameters is often necessary. The parameter optimization option in the Learning module of Perseus allows for a parameters space search (Fig. 4.5).

Input form: The input form is similar to that of the Classification option and specification of the cross validation procedure, the kernel type and the data organization are required. All parameters related to the selected kernel function can be optimized either on a one-by-one basis or in a two dimensional scan. The number of values to be tested, the step size with which the values are altered and the operation which is used to alter the values (addition or multiplication) can be defined by the user. For each value or combination of values of the kernel parameters a classifier is trained and tested on the specified data.

Output form: The prediction error rate associated with each parameter or parameter set is computed and stored in the output table. The parameters corresponding to the lowest error rate can directly be used in a classification procedure or alternatively can be subjected to further parameter refinement with a smaller step size.



	ABC	GC	Class	SubClass	Winner	Winners	Sample
Type	Expres...	Expres...	Categ...	Category	Categ...	Catego...	Text
1	-0.727...	0.7277...	GC	BJAB	GC	GC	Ratio H/L Normalized BJAB(2)
2	-1.11284	1.11284	GC	BJAB	GC	GC	Ratio H/L Normalized BJAB(3)
3	-0.979...	0.9797...	GC	BJAB	GC	GC	Ratio H/L Normalized BJAB(4)
4	-1.22826	1.22826	GC	DB	GC	GC	Ratio H/L Normalized DB(2)
5	-1.52399	1.52399	GC	DB	GC	GC	Ratio H/L Normalized DB(3)
6	-1.43066	1.43066	GC	DB	GC	GC	Ratio H/L Normalized DB(4)
7	0.6068...	-0.606...	ABC	HBL1	ABC	ABC	Ratio H/L Normalized HBL1(2)
8	0.6155...	-0.615...	ABC	HBL1	ABC	ABC	Ratio H/L Normalized HBL1(3)
9	0.6271...	-0.627...	ABC	HBL1	ABC	ABC	Ratio H/L Normalized HBL1(4)
10	-1.18202	1.18202	GC	HT	GC	GC	Ratio H/L Normalized HT(2)
11	-1.04779	1.04779	GC	HT	GC	GC	Ratio H/L Normalized HT(3)
12	-1.16486	1.16486	GC	HT	GC	GC	Ratio H/L Normalized HT(4)
13	2.099	-2.099	ABC	OciLy3	ABC	ABC	Ratio H/L Normalized OciLy3(2)
14	2.10284	-2.10284	ABC	OciLy3	ABC	ABC	Ratio H/L Normalized OciLy3(3)
15	2.12631	-2.12631	ABC	OciLy3	ABC	ABC	Ratio H/L Normalized OciLy3(4)
16	0.3378...	-0.337...	ABC	RIVA	ABC	ABC	Ratio H/L Normalized RIVA(2)
17	0.4000...	-0.400...	ABC	RIVA	ABC	ABC	Ratio H/L Normalized RIVA(3)
18	0.38662	-0.38662	ABC	RIVA	ABC	ABC	Ratio H/L Normalized RIVA(4)
19	-1.16747	1.16747	GC	SUDHL4	GC	GC	Ratio H/L Normalized SUDHL4(2)
20	-1.04511	1.04511	GC	SUDHL4	GC	GC	Ratio H/L Normalized SUDHL4(3)
21	-1.20558	1.20558	GC	SUDHL4	GC	GC	Ratio H/L Normalized SUDHL4(4)
22	-0.831...	0.8317...	GC	SUDHL6	GC	GC	Ratio H/L Normalized SUDHL6(2)
23	-0.676...	0.6766...	GC	SUDHL6	GC	GC	Ratio H/L Normalized SUDHL6(3)
24	-1.15727	1.15727	GC	SUDHL6	GC	GC	Ratio H/L Normalized SUDHL6(4)
25	0.5521...	-0.552...	ABC	TMD8	ABC	ABC	Ratio H/L Normalized TMD8(2)
26	0.4196...	-0.419...	ABC	TMD8	ABC	ABC	Ratio H/L Normalized TMD8(3)
27	0.5250...	-0.525...	ABC	TMD8	ABC	ABC	Ratio H/L Normalized TMD8(4)
28	0.2413...	-0.241...			ABC	ABC	Ratio H/L Normalized U2932(2)
29	0.4706...	-0.470...			ABC	ABC	Ratio H/L Normalized U2932(3)
30	0.39824	-0.39824			ABC	ABC	Ratio H/L Normalized U2932(4)

Figure 4.4: ***Perseus Classification Result form.*** *The first two columns of the Classification result matrix contain the actual decision values on which the classifier bases its prediction. The 'Class', 'SubClass' and 'Sample' columns contain the original data set assignments when available. Predictions are made for the samples with unknown class. The predicted class is displayed in the 'Winner' column. The 'Winners' column contains competing predictions, i.e. labels different from the final prediction that have been assigned to a given sample during some of the steps of the cross validation.*

4.3.3 Feature selection form

The Feature optimization option facilitates the extraction of features that discriminate best between the subgroups in the data. Currently several rank-based methods are implemented in the software: (i) ANOVA, (ii) Golub score-based and (iii) SVM weights-based. A detailed description of the underlying theory behind each of them

A B

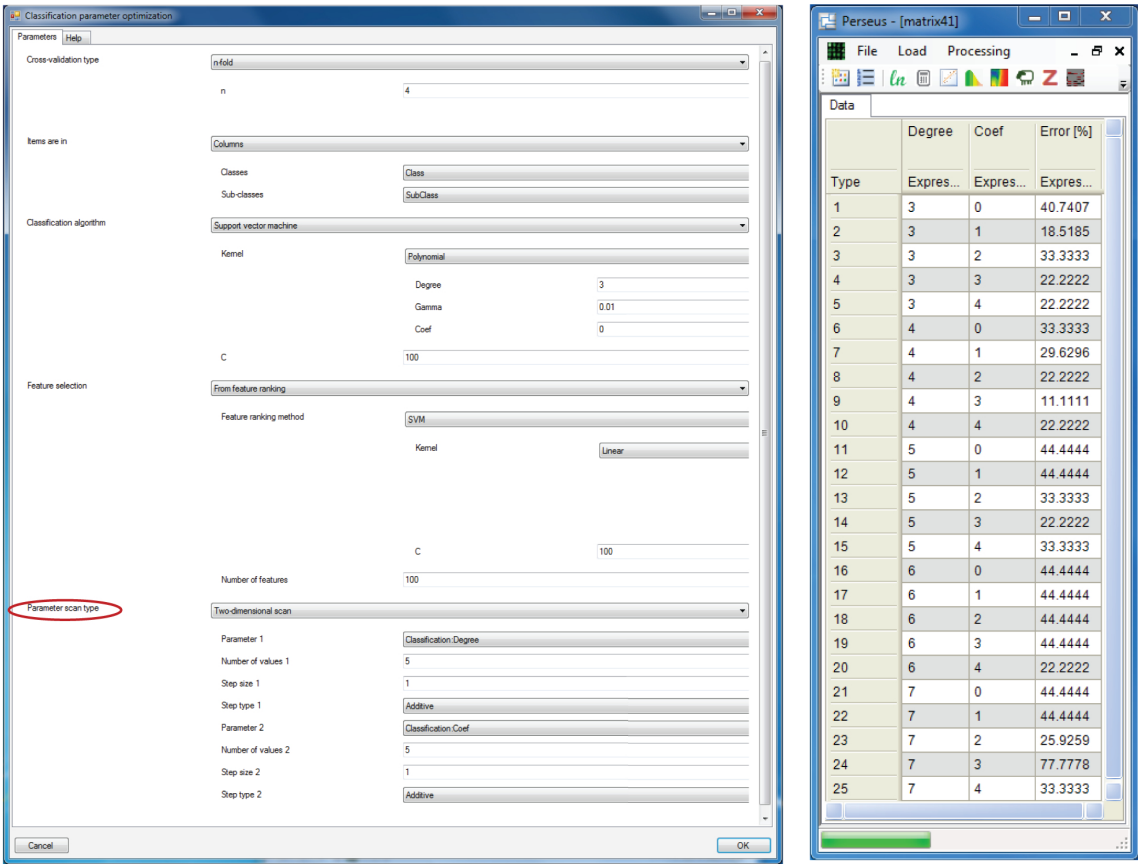


Figure 4.5: ***Perseus Parameter optimization input and result forms.*** ***A)*** *Various parameters of both the classification algorithm and the feature selection method can be optimized. In the example polynomial kernel is selected and the classification degree and coefficient are set for optimization in the dual parameter scan option. The number of values to be tested for each parameter, the starting value and the size of the step with which the value is altered can be set by the user.* ***B)*** *The parameter values tested as well as the associated error rate of the classifier are shown in the output matrix.*

is available in Section 4.4. Briefly, the labeled set of instances is split into training and test subsets and the selected ranking method is used to rank all features in the training set. The top F features are then used for training and prediction, where top F is gradually increased using the step size defined by the user.

Input form: The parameters of the input form are shown in Figure 4.6. Feature

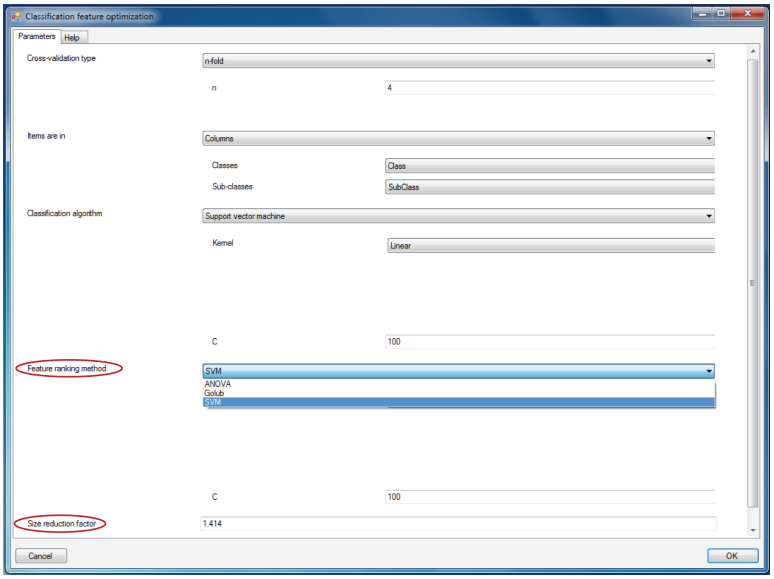


Figure 4.6: ***Perseus Classification feature optimization input form.*** *First, the classification algorithm and the cross validation type have to be specified. Cross validation is always enforced in order to ensure the generalizability of the results. Various feature ranking methods can be used, including SVM weights-based, ANOVA and Golub score-based methods. The specific parameters for each ranking method can be set by the user. The size of the step, with which the number of features in a set to be tested is varied, can be specified in the 'size reduction factor' option.*

selection is always performed within a cross-validation procedure in order to avoid overfitting. Moreover, a reliable estimation of predictor's accuracy and selection of biologically-relevant features are ensured, see Section 4.2).

Output form: Two output tables are generated (Fig. 4.7). The first one is a copy of the original input table with an additional column: 'Ranks', which contains the rank computed for each feature with the selected ranking method. Sorting the column in ascending order displays the best predictive features on top. The information about the optimal number of features, i.e. the smallest subset of ranked features that generates the highest classification accuracy, is stored in the second output table. It summarizes the classification accuracy and the corresponding number of features on which the classifier has been trained. Depending on the initial biological question there are two main usages of the optimal number of features. On one hand, a new classifier can be trained using only the top F selected features (using the same

classification and feature ranking parameters and setting the number of features to F in the Classification form). On the other hand, the selected subset of features can be subjected to further functional analysis in order to gain a deeper insight into underlying biological processes.

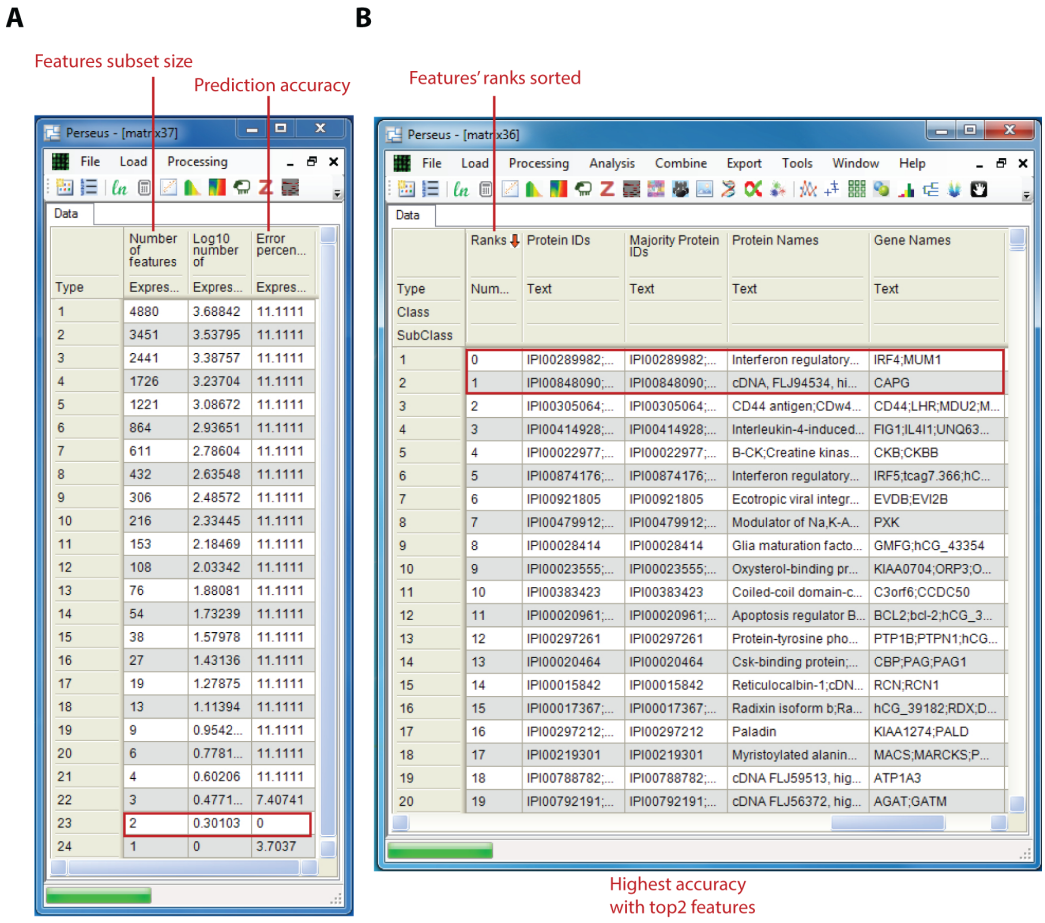


Figure 4.7: ***Perseus Classification feature optimization result forms. A)*** *The first output matrix of the feature optimization form contains information about the error rates corresponding to the different features sets. In the example a classifier trained on the top2 ranked features has the highest prediction accuracy.* ***B)*** *The second output matrix shows a copy of the original input matrix plus the additional column 'Ranks', containing the ranks computed during the feature selection procedure. Sorting the Ranks column displays the features with the highest predictive power on top of the list.*

4.4 Materials and methods

Classification problems are defined by instances in the form of vectors of features. In the case of clinical proteomics experiments the instances correspond to the patient samples and the features are the proteins expression values. The goal of supervised learning is to infer a rule from a set of instances with known labels, which can be used to assign labels to new instances. Support Vector Machines is a cutting edge supervised learning technique, which is well-suited for mining of high dimensional data.

A large range of feature ranking methods exist including correlation-based, multivariate selection-based, causal relevance-based methods and others [259]. In the simplest class: correlation-based methods the evaluation of the predictive power of the features depends on the feature correlation to a particular class.

Golub's score-based method

The correlation coefficient defined by Golub [264] is defined by the mean and the standard deviation of the distribution of values of a given feature in a certain class (Eq. 4.8):

$$w_i = \frac{(\mu_i(+) - \mu_i(-))}{\sigma_i(+) + \sigma_i(-)}, \tag{4.8}$$

where μ_i and σ_i are the average and standard deviation of the values of feature i in the positive and the negative class respectively.

Features weights can be viewed as signal-to-noise ratio estimations, resulting in the assignment of high scores to features that have larger between-class difference than within-class difference. In the original report of Golub et al. an equal number of the features with the largest positive and negative scores formed the final set of discriminative features. Alternatively the absolute value of the score can be used as a measure of its predictive power [265].

ANOVA-based method

More advanced ranking methods are the T-test statistics-based methods. Analysis of variance (ANOVA) measures the significance of the variation of the response variable attributable to the differences between the groups as opposed to the difference

within the groups. The generalizability of the method to a multi-class problem turns it into a powerful feature selection tool. Unlike the Golub method it uses the features variance as an estimation of the noise level. The calculated ANOVA p-values are ranked and converted into feature ranks.

The advantage of correlation-based methods is that they are relatively simple and usually fast to execute. Such methods are efficient in excluding features that are weak predictors, but do not take into account interactions between the features, thus eliminating potentially good predictors.

SVM weights-based feature selection

Unlike other feature ranking methods that estimate the predictive potential of single features, the SVM weights-based method assesses the capability of a set of features to distinguish between the different classes. In this method the weights computed for each instance during the training of the classifier are used to compute the feature ranks. The weights of an SVM predictor are based only on a limited subset of the training examples - the support vectors. Consequently the feature scores are influenced only by the most discriminative instances, as opposed to the scores generated by the Golub ranking method, which averages over the complete training set. The feature score vector (Eq. 4.9) is a linear combination of the support vectors and their non-zero weights.

$$w = \sum_k \alpha_k x_k y_k, \tag{4.9}$$

where $alpha_k, x_k$ and y_k are the weight, value and class of the kth instance.

SVM-Recursive Feature Elimination (RFE)

As it would be computationally- and time-expensive to test the predictive power of all possible subsets of features, more efficient strategies are needed. Guyon et al. [162] proposed a Recursive Feature Elimination (RFE) technique, which enables optimal feature subset discovery in a reasonable time. RFE is an instance of the backward elimination algorithm in which a classifier is trained, all features are ranked and the worst feature or worst several features are removed (see: Recursive Feature Elimination Procedure Outline). The procedure is repeated until no features remain

in the training set. The recursive manner of the procedure ensures that every next run results in a more precise decision function and further optimizes the scores of the features.

Recursive Feature Elimination Procedure Outline:

1 : Restrict instances to the selected features set S: $X = X_0[0 : S]$

2 : Train classifier: SVMTrain(X,Y)

3 : Compute weights: $\overrightarrow{w} = \sum_k \alpha_k x_k y_k$

4 : Compute ranking criteria: $c = w^2$.

5 : Find the worst m features: $f = argmin(c, m)$

6 : Update ranked list R: $R = [S(f), R]$

7 : Remove the worst m features from the feature set S: $S = [S : f]$

Chapter 5

Cancer classification: applications

" . . . there is moderate certainty that the benefits of PSA-based screening for prostate cancer do not outweigh the harms . . . "
(USPSTF, Annals of Internal Medicine, 22 May 2012)

5.1 Introduction

The recent advances in the field of mass spectrometry-based proteomics provide a promising platform for clinical studies. The newly developed super-SILAC technique enhances the identification and accurate quantification of a large number of proteins from patient samples [135, 122]. A heterogeneous cell lines mixture - super-SILAC - has been shown to be a good representative of the complexity and the large biological variability of tissue samples. A standard clinical proteomics work-flow utilizes formalin-fixed paraffin-embedded (FFPE) tissues in combination with the FASP protocol [266]. The MaxQuant computational framework enables the processing of the raw files obtained from mass spectrometric measurements and results in the identification and quantification of thousands of proteins, ready for downstream analysis [6]. Thus mass spectrometry-based clinical proteomics may provide the means to improve the current state of disease detection and staging, as well as to develop accurate tools for prognosis and prediction of the therapeutic outcome and of the probability of recurrence of a disease.

Among the main goals of clinical proteomics are classification of tumor-specific proteomics profiles and identification of potential biomarkers and drug targets. The

large biological variability that characterizes tissue samples poses a big challenge for the applicability of statistical and analytical methods for classification of disease subtypes that have previously only been shown to perform well on cell lines [127]. The task is further complicated by a usually small sample size combined with high dimensional feature space.

Prostate cancer is the 6th leading cause of death for men in the world [267] with an incidence of more than 200,000 new cases per year in the United States alone. A large number of markers that have the potential to distinguish between healthy and malignant tissues are available or are under clinical trials [268]. The prostate-specific antigen PSA is a major biomarker used in the early detection of the disease and in the evaluation of the chances of biochemical recurrence. Elevated levels of PSA in the blood are used as indicators of possible development of a malignant prostate formation or of biochemical recurrence of an already treated tumor [269]. By far most of the deaths associated with this form of cancer are caused by metastases. It is therefore of high clinical importance to deepen our understanding of the molecular functions of the oncogenic activators and drivers of the transformation of the primary tumor cells into metastatic ones.

Breast cancer is among the most common cancers affecting women and a major cause of death. From clinical perspective 3 major breast cancer subtypes can be distinguished: estrogen receptor positive (ER+), Epidermal Growth Factor Receptor 2 positive (Her2+/ErbB2) and triple negative (TNBC), in which none of the 3 pertinent receptors: ER, progesterone receptor PR or Her2 are expressed. This main classification is based on the recognition of three biomarkers that are extensively used in diagnostics and treatment - Her2, ER and PR. As the disease is characterized by large heterogeneity even finer subtypes can be defined. For example, based on microarray data 5 classes have been characterized: basal-like, luminal A and B, ErbB2-overexpressing and normal breast-like tumors [270, 271].

Despite the efforts in diagnosis, prognosis, and treatment the ability to distinguish between different subtypes is still limited and offers large space for improvement. The presence or absence of overexpression of any of the above-mentioned breast cancer biomarkers has been shown to be of clinical importance and in some cases also sufficient to guarantee a positive outcome of a treatment. The use of a drug

that targets the Her2 receptor - Herceptin - is highly effective in the class of Her2+ patients [272, 273]. The situation is very different for patients falling in the TNBC class. Despite the known mutation in the BRCA1 gene that increases the chance of women to develop this type of cancer [274], the optimal treatment options are still very limited. The main treatment is based explicitly on chemotherapy with overall poor survival prognosis, which clearly points to the need for discovery of new drug targets [275]. Furthermore, elevated PSA level can also be caused by benign prostatic hyperplasia or even by conditions unrelated to the prostate. Characterized by high sensitivity but low specificity, the usefulness of screening tests relying on the PSA biomarker has been seriously questioned [165, 268, 276].

Overall, the currently-known single biomarkers fail to account for the large heterogeneity associated with breast cancer subtypes and they do not reliably estimate the risks of prostate cancer development and recurrence. This can result in wrong diagnosis and overdiagnosis and therefore in suboptimal treatment of the patient. Furthermore, better understanding of the molecular processes orchestrating the development and progression of cancer subtypes is needed in order to address the need for more efficient personalized medicine. Proteome analysis may offer a more direct insight into the functional phenotype of the cells than mRNA studies do as it takes into account processes such as protein degradation, secretion and localization. Consequently, it is reasonable to hypothesize that proteome profiling of cancer patients could enable the discovery of more suitable biomarkers or sets of biomarkers with high impact on both accurate diagnosis and drug development.

5.2 Data preprocessing

5.2.1 Data transformation

Various preprocessing steps are usually applied to any type of data prior to the main analysis. The first step in working with proteomics data is removing ('filtering') unreliable and noisy data, such as contaminants and hits from the reverse database (see Chapter 1, Quantification). In the proteome profiling of cancer tissues using the super-SILAC technology, the protein quantity is measured and reported as a ratio

between the heavy (the cell lines mix) and the light (the tissue samples) sets. As one is interested in comparing the tissue samples in different conditions, the data are inverted for convenience. The log2 transformation of the data diminishes the dependence on the absolute magnitude and helps to avoid the strong influence of outliers and skewed distributions. If the data are approximately normally distributed, upon log2 transformation there is an even distribution of negative and positive values centered around the 0 value, which makes the interpretation of the up and down regulation differences between the groups more intuitive.

5.2.2 Missing values imputation

Due to the nature of the proteomics data and the current limitations of the mass spectrometry technology missing values are a common problem. It is however reasonable to assume that the missing values result from low-intensity entries that thus could not be quantified. One rational strategy to impute such missing values is to fit a normal distribution to the values in each sample, from which values will be randomly drawn to fill in the missing data points. Following the above-made assumption the new distribution is shifted towards lower values and its width is narrowed down. It is important to avoid creating a second peak in the original distribution by carefully controlling the two parameters (down-shift and width). The effects of imputation and of the amount of valid values that is required per feature (protein group) is discussed in the subsequent sections of this chapter. Overall, the results suggested that filtering for 70-80% valid values is a good compromise between discarding valuable information and introducing noise in the data (i.e. caused by the imputation of a larger number of missing values).

5.2.3 Data normalization

In case the protein ratio distributions of the instances appear to be shifted, normalization is performed in order to center them around zero. This can be achieved by subtraction (if the data are log-transformed) of the most frequent value or of the median of that distribution from all elements of that instance. Some particular analysis techniques strongly benefit from normalization of the features. For example, the underlying data patterns become much better visible in clustering when the features are z-scored (i.e. follow a distribution with mean 0 and standard deviation

1). Furthermore, normalization strategies such as z-scoring and placing the data into some interval (e.g. [-1;+1]) minimize the influence of outliers in the data on the performance of the analytical tools.

5.3 Classification of breast cancer subtypes

Proteome profiles of 40 breast cancer patients were measured on a high performance Q Exactive instrument. Protein extraction and digestion were performed using the FFPE-FASP protocol [266] and a super-SILAC mix [135] of five cell lines was employed to obtain accurate quantification of the proteins in the tissue samples. The data set includes patients from three main breast cancer subtypes: 14 Her2+, 13 ERPR+ and 13 triple negative samples, respectively. The aim of this study is to identify a discriminative set of features that can increase our understanding of the distinct mechanisms governing the development of the subtypes and possibly discover new potential biomarkers and to estimate their predictive power. The influence of various feature selection methods and data normalization techniques on the performance of the predictor is also discussed.

The original data contained 12,500 identified protein groups with more than 8,000 proteins quantified per sample. Data acquisition is described in detail in the Matrials and methods secion 5.5. The data were filtered for reverse hits and contaminants, inverted, log2 transformed and filtered for different cutoffs of valid values. The missing values were imputed as described in section 5.2.2 above, using a width parameter of 0.2 and a downshift by 1.0 standard deviations. Furthermore, several data normalization techniques were used to reduce the influence of outliers and of the possible bias introduced during the imputation on the final result of the classifier.

5.3.1 Effect of valid values filtering on the prediction accuracy

Support vector machine classifiers were trained and tested on several subsets derived from the original breast cancer data, generated by varying the minimal number of required valid values (available protein ratios). Four conditions were tested: (i) 50% valid values, corresponding to protein measurements available in at least 20 samples,

(ii) 70%, (iii) 80% and (iv) 100%, i.e. only those features were kept, for which quantitative information was available for all 40 samples. The remaining missing values were then imputed as described in Section 5.2.2 using values randomly drawn from a fitted Gaussian distribution with width 0.2 and shift of 1 standard deviations to the left of the center of the original sample distribution.

The effect of data normalization was further tested for each derived data set by comparing performances of classifies on non-normalized data, z-scored data and data scaled to the [-1;1] interval. The classifiers were combined with different feature selection methods (see Section 4.1.2) and the summary of the results is shown in Figure 5.1.

Overall, the accuracy curves plotted as a function of the number of features were similar when the same feature selection method was used, regardless of the number of missing values. However, choosing an optimal filter for valid values did influence the quality of the classifier and the biological relevance of the selected features. Imposing the most stringent requirement of 100% valid values strongly reduced the features size, but also led to losing important information as visible from Fig. 5.1. The optimal accuracy that was reached by applying feature selection on this data set based on ANOVA or Golub score ranking was significantly lower than that when only 50% valid values were required. This was because FOXA1 and AAC1, two of the top ranked features in the larger data set, were eliminated at the valid values filtering step.

The results of the classification of the data set with lower requirements for valid values (50, 70 and 80%) were highly similar and the differences can be readily explained by the precision of the accuracy estimation and the small variability introduced during the imputation by random values. In all sets, in which means-based feature ranking methods were used, the classifier reached its optimal accuracy with either the top 3 (50 and 70% valid values) or the top 4 (80% valid values) features, where three of these four were the same as top3 in the other two sets of selected features. The variation in the optimal number of features selected with the SVM weights-based ranking method was greater: 214 (50%), 521 (70%) and 333 (80%). The similar trends of the classifiers performances demonstrate the overall small effect that the number of missing values has on the prediction power of the classifiers.

The variation in the selected sets of proteins was more tightly related to the type of ranking method, which is discussed in detail in Section 5.3.3.

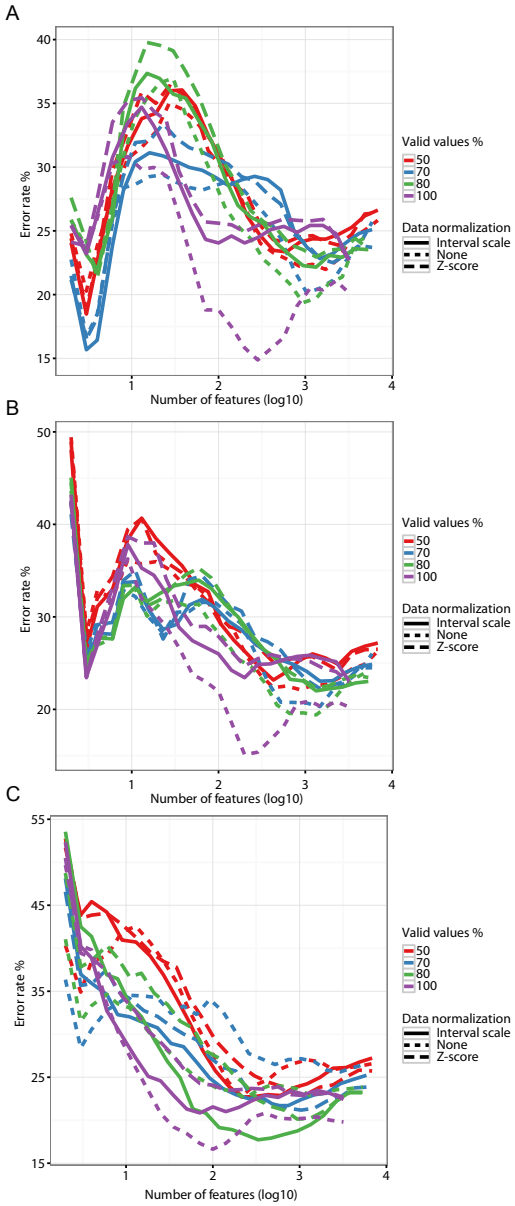


Figure 5.1: ***Influence of valid values filtering on classification accuracy.*** *Classification error rate as a function of the number of features is shown for three distinct feature selection methods* ***A)*** *ANOVA-based ranking,* ***B)*** *Golub score-based ranking, and* ***C)*** *SVM weights-based ranking. Each of the feature selection methods is applied on several data sets varying according to the percentage of valid values filtering and to the technique used to normalize the features prior to the analysis.*

In summary, datasets with missing values between 50 and 80% are suitable for classification and feature selection tasks. However, one can argue that the more tolerant approach of keeping all features for which data were available in 50% of the samples may result in the introduction of more noise upon imputation and thus in overall lower reliability of the classifiers. Therefore, a threshold of 70-80% valid values is recommended as it would allow preservation of the majority of the important features and at the same time limit the noise introduced by imputation.

5.3.2 Influence of feature normalization on the prediction accuracy

Analysing directly the original data without any preprocessing may lead to unexpected behavior of the classifier or to the selection of features that are irrelevant to the biological question of interest. Some of the problems that may necessitate data transformation prior to the analysis are: (i) influence of outliers in the data that interfere with the biological signal, (ii) features that appear to be on different scales with respect to the magnitude of their variation, causing them to become incomparable and (iii) imputed missing values that increase the noise level in the data.

Z-scoring is a linear transformation that normalizes the distribution of each feature to have a mean of 0 and a standard deviation of 1. As the different features may vary with different magnitude, it is often advisable to normalize the data to improve the features comparability. Essentially, z-scoring brings small and large differences to the same scale. However, this normalization technique is not insensitive to outliers in the data. Furthermore, bringing the variability of the features to similar scales makes the task of means-based feature ranking methods to identify clear winners among the ample of features more difficult.

Scaling the feature values to some interval is similar to the standard score normalization in the sense that the differences between the features are minimized, but the extent to which the small differences are amplified is smaller.

Generally, normalization techniques do not strongly influence the performance of the predictors (Fig. 5.1). It can, however, be seen that in the dataset containing only the original measurements (no missing values) the highest accuracy was

reached when no normalization was performed. The other data sets, in which a certain percentage of missing values was retained and imputed, the predictors benefitted from the rescaling of the data. Although the differences were not large the normalization functioned as a filter for any noise that could have been introduced by the imputation and slightly improved the performance.

5.3.3 Effect of feature ranking methods on accuracy

Interestingly, notable differences were present in the performance of the classifiers with respect to the feature selection method with which they were combined (Fig. 5.2). Means-based methods (ANOVA p-value and Golub score) reached highest accuracy with the top3 or top4 ranked features. This result is not surprising as the strength of such methods lies in identifying single features that show high discriminative power. However, the accuracy dropped significantly as new features were added. Although each additional feature may be a good candidate on its own, ANOVA and Golub score-based ranking methods ignore the interplay between features and thus fail to recognize sets of features with good predictive properties. In contrast, the prediction accuracy of the classifier, combined with SVM weights-based feature ranking, improved upon increasing the number of ranked features in the test set until a maximum was reached. This is because this ranking method estimates the goodness of complete sets of features as opposed to single features.

The performance curves obtained with the different ranking methods provided two important pieces of information. The means-based methods detected three features that carried most of the predictive power in the features set, strongly suggesting that these three proteins are potential biomarkers. The much larger optimal feature sets identified by the SVM weights-based ranking method, may instead provide information about the biological mechanisms underlying the disease onset, rather than about single potential markers. The larger signatures were also characterized by a smaller overlap betewen the different runs of the cross validation procedure (shown by the average ranks estimated during the cross validation procedures). This suggests that a signature would consist of features that complement each other with less regard of their individual contributions. Suppose such a set included features from different pathways or of different biological modules, the SVM weights-based ranking method would not favor a particular feature from that pathway or module.

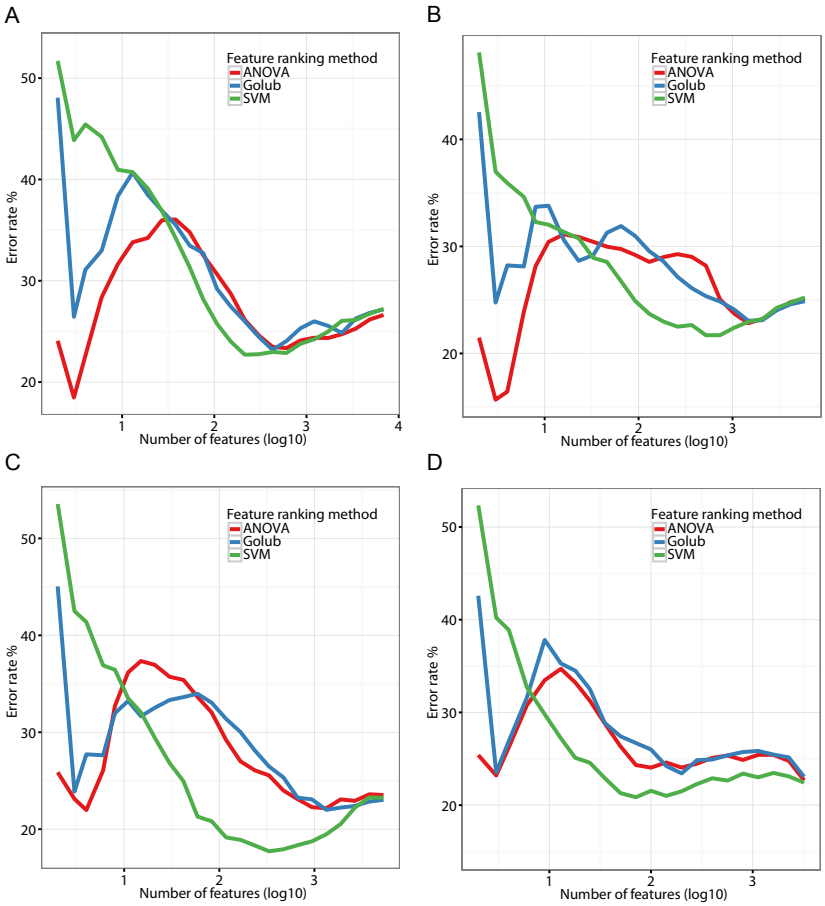


Figure 5.2: ***Influence of feature selection methods on classification accuracy.*** *The prediction accuracy is plotted as a function of the number of features. The four panels summarize the results for data sets generated by different valid values requirements:* ***A)*** *50%,* ***B)*** *70%,* ***C)*** *80% and* ***D)*** *100%. In each panel, classifiers using three distinct feature ranking methods are depicted: SVM weights-based ranking in green, ANOVA-based ranking in red and Golub score-based ranking in blue.*

In the different runs of the cross validation it would pick any of the features, as long as the overall set has a high predictive value.

In summary, means-based methods unambiguously identified the features that showed the largest between-class variation, whereas SVM weights-based ranking selected any of several possible features that added predictive information to the complete set. Therefore, the two techniques can be used to answer distinct biological questions

and may have complementary clinical applications.

5.3.4 Biological relevance of the selected features (GO and GSEA enrichment analysis)

An obvious question in feature selection tasks concerns the relevance of the selected features to the problem of interest, here the subtyping of breast cancer. Literature search showed that the three features that always received high ranking with the means-based ranking approach are important in the onset of breast cancer. The v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 (ErbB2), also called Her2, which was the top ranked feature regardless of the ranking method of choice, is an already established biomarker. ErbB2 is the defining marker in the diagnostics and treatment of Her2+ breast cancer patients [277, 278, 279]. The fact that this protein was picked from the proteomics data set of 8000 quantified proteins, provides strong evidence for the quality of the data and the analysis. The second top feature was the protein encoded by the human anterior gradient-2 (AGR2). Significant co-expression of the AGR2-encoded protein and the estrogen receptor has been reported, see for instance [280]. Additionally, there is strong evidence that AGR2 may act as a metastasis inducer [281] and it was further demonstrated to have separate prognostic value for survival [282]. The last feature that appeared to be necessary to achieve the highest accuracy with the means-based ranking methods was a transcription factor, the Forkhead box protein A1 (FOXA1). Not much is known about FOXA1 in breast cancer, but a recent study has shown that breast cancer-associated SNPs alter the binding of the FOXA1 protein to DNA, which in turn regulates the estrogen receptor α (ER) function [283]. This observation provides strong support for the relevance of the FOXA1 protein as a potential biomarker in relation to the ER+ subtype.

A 4th feature appeared to be highly important in the 80% valid values filtered data set - the Growth factor receptor-bound protein-7 (GRB7), an SH2 domain containing protein involved in growth factor signaling. GRB7 has already been identified as both a prognostic and recurrence marker in gene expression studies [284, 285]. It has been related to the Her2 protein but has also been shown to carry predictive and diagnostic information on its own, thus adding yet another promising feature to the list of known and potential biomarkers [286]. Thus the extensive support found in

literature provides strong evidence that feature selection with means-based ranking procedures is a powerful technique for discovery of potential biomarkers.

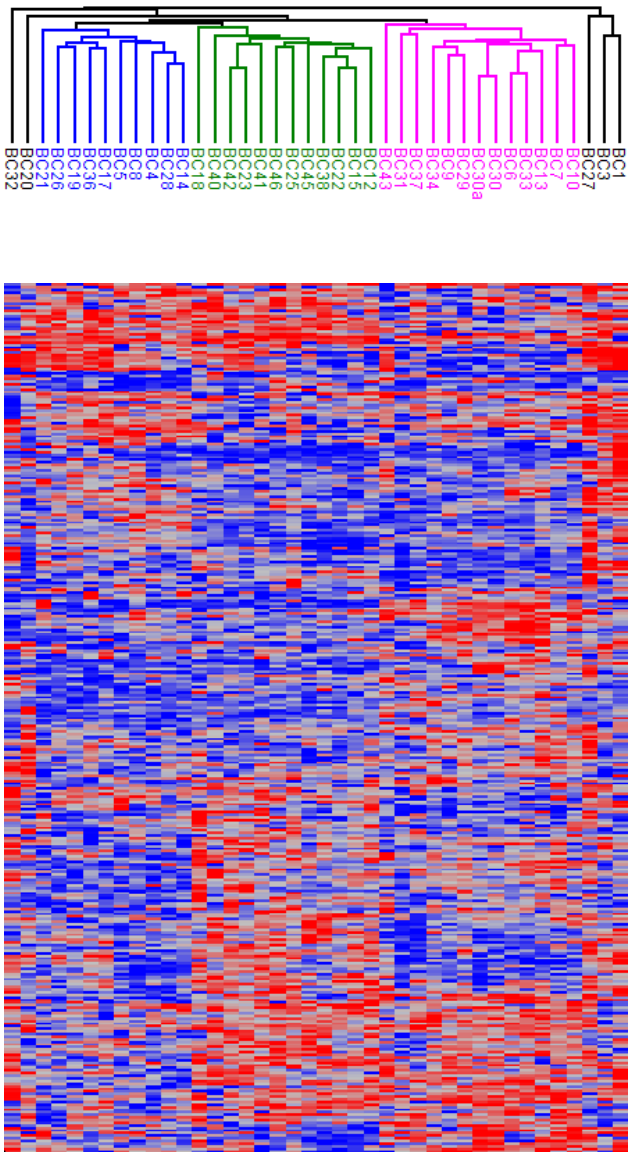


Figure 5.3: ***Hierarchical clustering of patient samples.*** *The 40 patient samples are clustered based on the proteome profiles of the 333 features selected with the SVM weights-based ranking method. The 3 main breast cancer subtypes are shown in blue (ERPR+), green (TNBC) and purple (HER2+), respectively.*

Although the accuracy with the top 3 features was around 80%, this still means that

about 20% of all patients would be misclassified, illustrating the need for further improvement. The SVM weights-based ranking method reached highest performance (82% accuracy) with the top 333 features. Although this number may be too high to provide a clinically-applicable signature, it already offers a good platform for analysis of the underlying biological processes. Furthermore, the selected by the SVM weights-based method set contained the above-described top4 proteins at highly-ranked positions, which is a strong indication of its biological relevance.

Hierarchical clustering of the patient samples based on the selected 333 features resulted in 3 clear clusters corresponding to the ERPR+ breast cancer subtype in blue, TNBC in green and Her2+ in purple (Figure 5.3). Furthermore, patterns of group-specific differentially-regulated clusters of features emerged. Some clusters were up-regulated in a specific class, while others were up-regulated in two of the classes but not in the third one. In total, 35 of all samples were correctly grouped, whereas five formed separate clusters. Inspection of the quality of the wrongly-clustered samples showed that they contained noticeably fewer identified and quantified proteins. It is therefore highly probable that exactly the larger number of missing values and the consequent imputation increased the noise level and made these instances more challenging for segregation. This also provides a useful yardstick for the quality of proteomics datasets needed in the classification of breast cancer subtypes.

The selected subset of 333 features was tested for enrichment of GO-terms and GSEA categories. Interestingly, several breast cancer gene sets contained in the GSEA database were significantly enriched: (i) SMID BREAST CANCER LUMINAL B UP, (ii) SMID BREAST CANCER BASAL DN, (iii) DOANE BREAST CANCER ESR1 UP and others (Fig. 5.4). The most enriched gene ontology term was the major histocompatibility complex I (MHC I) class peptide loading complex. The MHC class I molecules are involved in the cellular immune response in cancer: decreased levels lead to poor defense mechanisms ([287]). The strong enrichment of this category may point to a connection of the selected proteins to general processes involved in cancer development.

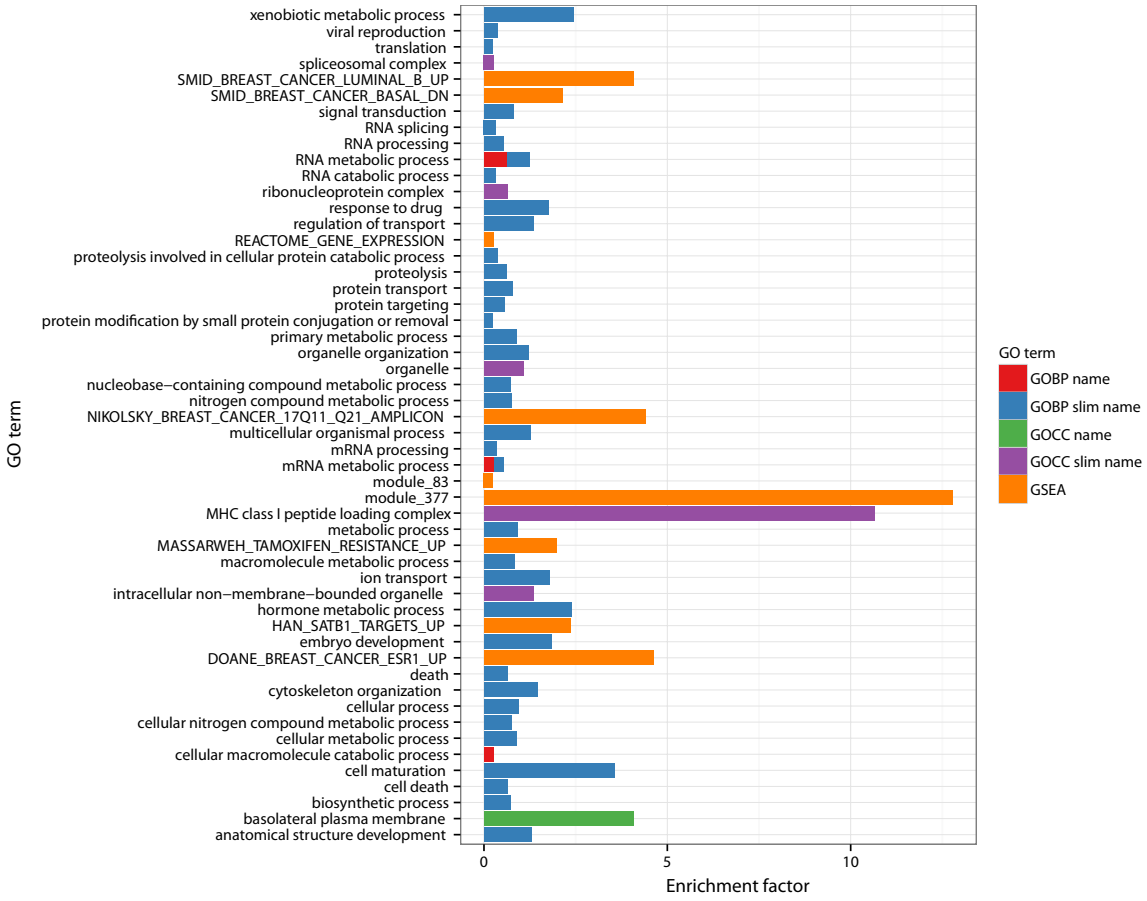


Figure 5.4: ***Enrichment of GO-terms and GSEA categories.*** *Cellular Compartments (GOCC) in green and Biological Processes (GOBP) in red gene ontology terms or their broader versions (GOCC slim in purple and GOBP slim in blue), GSEA categories in orange that were significantly enriched in the selected set of 333 ranked features and their corresponding enrichment factors are shown.*

5.4 Classification of relapse and non-relapse prostate cancer subtypes

Upon radical prostatectomy some patients suffer a recurrence of the prostate cancer, currently detected by elevated levels of the prostate specific antigene (PSA). The timely diagnosis of the patient with respect to one of the two possible classes - recurring or non-recurring - would allow for a more rational and efficient treatment increasing the patient's survival chances or sparing patients from unnecessary treat-

ment. The goal of this project is to investigate if features that strongly discriminate between the relapse and non-relapse patient samples can be extracted based on their proteomic profiles and to evaluate their predictive power.

In total more than 9,800 proteins were identified and quantified in 20 samples from prostate cancer patients that were characterized as either 'relapse' or 'non-relapse', depending on the level of PSA. Standard pre-processing was applied to the initial set, including log2 transformation of the ratios, filtering for valid values (similarly to the breast cancer set four thresholds were tested) and normalization of the features. Support vector machines based classification was used in combination with two feature ranking methods: ANOVA-based and AVM weights-based (Fig. 5.5).

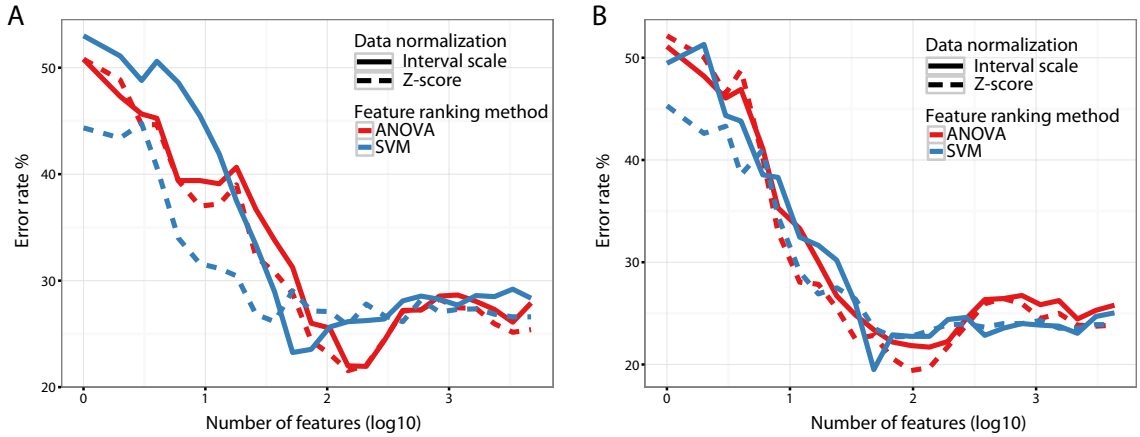


Figure 5.5: ***Influence of feature selection methods on classification accuracy.*** *The prediction accuracy is plotted as a function of the number of features. The two panels summarize the results for data sets with different levels of missing values:* ***A)*** *70%,****B)*** *80% were required respectively. Classifiers using SVM weights-based ranking are shown in blue and ANOVA-based ranking – in red. The solid lines represent feature values scaling to interval, whereas the dashed line – z-scoring.*

Classifiers trained on sets derived by different missing values filtering showed similar trends. In contrast to the results in the breast cancer subtype classification, none of the ranking methods was able to detect strong class discriminators exemplified by the size of the features set that had the best prediction accuracies (Fig. 5.5).

The ANOVA-based ranking reached 78% accuracy with the top 70 (80% valid values filtering) and top 150 (70% valid values filtering) features. Using the SVM weights in the feature ranking resulted in sets of about 50 features and an overall accuracy ranging between 77 and 80%. Examination of the top ranked features revealed important proteins with known implications in the development and progression of prostate cancer. The set included various proteins such as chaperones, kinases and other signaling proteins, some of which are described in the following paragraphs.

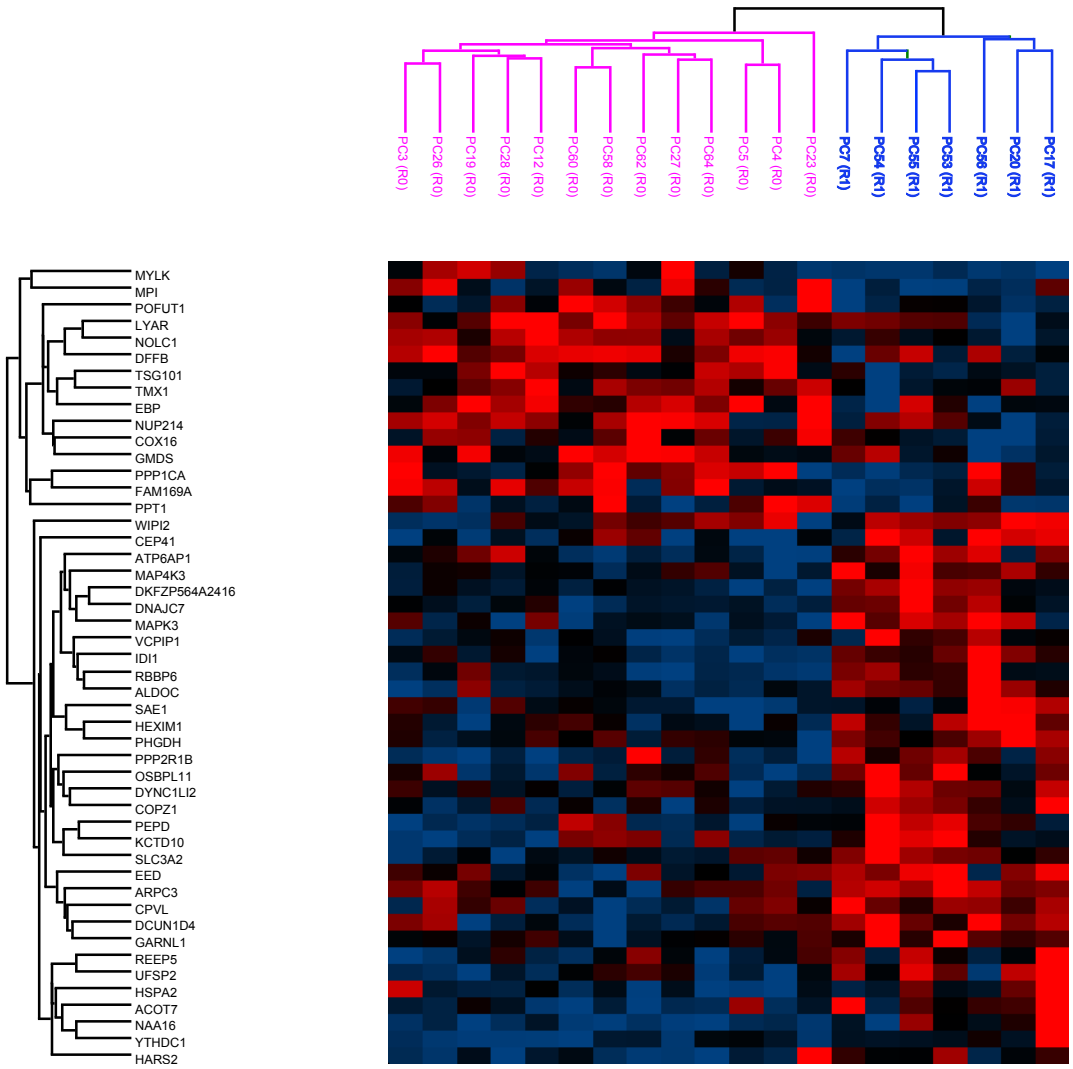


Figure 5.6: ***Hierarchical clustering of patient samples.*** *The 20 prostate cancer patient samples were clustered based on the proteome profiles of the 48 features (80% accuracy) selected with an SVM weights-based ranking method. The relapse (R1) and non-relapse (R0) classes are shown in red and blue respectively.*

One of the top-ranked features was the DNAJ homolog subfamily C member 7 (DNAJC7), which is a member of the heat shock protein family 40 and is known to regulate the molecular functions of the chaperones HSP70 and HSP90 [288]. In particular, DNAJC7 has been shown to be involved in the regulation of the folding of the progesterone receptor [289]. In the study of Bonkhoff et al. the progesterone receptor (PR) has been detected in a significant number of metastatic and recurrent prostate cancers and was further shown to correlate with prostate tumor progression [290]. The authors argued that as progesterone is considered a marker of estrogen activity, the high expression of PR at late stage progression demonstrates the ability of prostate cancer cells to escape androgen deprivation and instead utilize estrogens for growth and maintenance, which clearly demonstrates how DNAJC7 may be involved in cancer recurrence.

Another protein in the top ranked features was the mitogen-activated protein kinase 3 (MAPK3). Together with other molecules important in signaling cascaded, de-regulation of the MAPK3 (also known as ERK1) is associated with the onset of various cancer types including prostate cancer [291]. Extracellular Signal-Regulated Protein Kinases (ERK1) is implicated in proliferation, differentiation, angiogenesis, motility and invasiveness [292], which suggests the importance of this protein for prostate cancer recurrence. Moreover, constituent activation of the MAPK3 kinase has been shown to be implicated in hypersensitivity of PSA to reduced levels of androgen [293].

A feature with high predictive power was also the nucleolar and coiled-body phosphoprotein 1 (NOLC1) – a phosphoprotein that is localized in the nucleolus and due to its ability to bind nuclear localization signals it has been hypothesized to play a role as a chaperone for transport between the nucleolus and the cytoplasm. In addition, it is involved in the synthesis of rRNA and ribosomes. Several lines of evidence point at the possible importance of this feature as an oncogene. NOLC1 has been shown to bind to the promoter of the tumor protein p53 while enhancing the progression of nasopharyngael carcinoma [294]. Interestingly, one of the transcription regulators of NOLC1 was found to be the transcription factor NF-kB [295]. NF-kB influences the expression level of many genes, which are known to be involved in major cellular processes. Missregulation in its function often has detrimental consequences, including tumorgenesis [296]. Moreover, a clear connection between the

function of NF-kB and both prostate cancer development [297, 298] and recurrence after prostatectomy [299, 300] has been established.

The described literature provides substantial evidence, that despite the lack of clear biomarkers or strong signature discriminating between the relapse and non-relapse prostate cancer subtypes, disease-relevant features were present in the top selected set.

Hierarchical clustering of the samples using only the features selected with the SVM weight-based ranking method (upon filtering for 80% valid values, 48 proteins with accuracy 80%) resulted in the correct grouping of the samples in their corresponding categories – relapsed or non-relapsed (Supplementary figure 5.6). However, no GO terms or GSEA categories were significantly enriched in the selected set of proteins. Moreover, the average ranks of the selected features showed a large deviation from the final ranks (Pearson correlation coefficient of 0.61), which indicates that a large number of sets of features with similar predictive power can be derived from the current data set. This observation together with the small sample size and the questionable reliability of the prostate specific antigene, which was used in the initial categorization of the samples explain why the task of identification of a set of features that discriminates well between relapse and non-relapse patients appeared challenging. Strategies for improving the prediction accuracy and the relevance and generalizability of the selected features are discussed in detail in the Outlook section of this thesis (Chapter 6).

In summary, the machine learning methods and feature ranking techniques described in Chapter 4 were employed in the analysis of two clinical proteomics data sets. The first study addressed the challenges in accurate breast cancer sub-class diagnosis. A 3-class predictor was trained on the major breast cancer subtypes: ER+, Her2+ and TNBC and its accuracy to distinguish between the proteomic profiles of the 3 groups was estimated in a cross-validation procedure. The highest accuracy reached was 80% after feature selection with RFE-SVM procedure (see Section 4.2).

The second study aimed at revealing important regulators of prostate cancer that can be used to distinguish between patients that experience reccurring cancer upon radical prostatectomy and patients that remain free of symptoms. An SVM-based

classifier in combination with feature selection reached prediction accuracy of 80% in a random sampling cross validation procedure.

Special emphasis was placed on the data preprocessing required prior to any analysis and the effects it had on the classification accuracy. Furthermore, the outputs of different feature selection techniques were compared, providing important insights for the optimal analysis framework. The classification and feature selection results demonstrate the potential that proteomics data hold for clinical applications. Clearly they may bring the desired transformation of personalized, adequate and time-efficient treatment from a research field into a rountine practice one step closer.

5.5 Materials and methods

5.5.1 Data acquisition

The specially developed clinical proteome workflow developed in our group was used for the sample preparation of the tumor tissues. Protein extraction and digestion were performed from formalin-fixed paraffin-embedded (FFPE) tumor tissues following the FASP (Filter Aided Sample Preparation) protocol [134]. To enrich the content of cancer cells in the samples the tissues were dissected.

The super-SILAC mix strategy [135] was employed to account for the high tissue heterogeneity and to allow for accurate quantification of the proteome content. Stable-isotope labeling by amino acids in cell culture (SILAC) has become a standard method for cell line protein quantification [32]. Cell lines are labeled through the incorporation of stable heavy versions of essential amino acids, typically lysine and arginine. The labeling produces a mass difference of 8 and 10 mass units, respectively, for each tryptic peptide, thus allowing for accurate quantification of the proteins in the sample. As complete metabolic incorporation is required, this procedure is not directly applicable to analysis of human tissues. Instead, metabolically-labeled cell lines are used as internal standards and mixed with the tissue sample lysates [301]. The standard mass spectrometric analysis steps are then performed on the mixed samples together. However, to achieve in depth and accurate quantification of the majority of the relevant proteins, the SILAC-labeled proteome should be

sufficiently similar to the proteome of interest. The identification and quantification of the deep proteome of the complex tissue sample was thus enhanced by using a mixture of five SILAC-labeled cell lines as opposed to using a single cell line. The use of the mixture of variable cell lines (e.g. from different cancer grades and stages) allowed to account more fully for the tumor cells heterogeneity [302].

The proteomes were measured on the high-precision Q Exactive instrument [224] in the relatively short time - 6 fractions of 4 hour gradients per sample.

5.5.2 Protein identification and quantification

The raw mass spectrometric data files were processed with the MaxQuant computational framework. All standard setting were kept, including advanced options such as 'Match between runs' and 'Advanced ratio estimation'. The breast cancer data were analysed with two separate parameter groups: one for the tissue samples and an additional one for the super SILAC mix with only heavy labels. The use of the mix in the analysis improves the overall identification rates.

5.5.3 Prostate cancer dataset

Tumor samples were obtained from 20 patients. Based on the PSA levels upon radical prostatectomy the patients were classified into two groups: 9 samples showing chemical relapse (i.e. high PSA level) and 11 - lacking chemical relapse (retaining low PSA level). In total 9,828 proteins were identified and quantified.

5.5.4 Breast cancer dataset

Tumor samples were obtained from 40 breast cancer patients. The samples were categorized into the 3 main subgroups: 14 Her2+, 13 ERPR+ and 13 triple negative samples. In total 12,555 proteins were identified and quantified with more than 8,000 proteins in each sample.

5.5.5 Data analysis

The methods used in the data analysis (support vector machines with recursive feature elimination embedded in cross validation) and the software framework are

described in detail in Chapter 4.

Chapter 6

Conclusions and outlook

Overall, this thesis makes contributions to the analysis of two main types of proteomics experiments: phospho-proteomics and onco-proteomics data. The first part of the thesis focuses on the analysis of various properties of phosphorylation sites and their applicability to address the difficult task of distinguishing functional from silent modification sites. The second part describes the development of a computational framework for feature extraction and classification of patient samples based on their proteomics profiles and its successful application to two cancer data sets.

Phospho-tyrosines exhibit structural properties that are significantly different from those of modified serine and threonine residues

All phosphorylation residues showed a statistically significant preference for disordered and irregular regions. However, this tendency was much less pronounced for phospho-tyrosines, a large proportion of which appeared in ordered regions. In addition, they were found to be much less solvent accessible and appeared with similar frequencies in domain and inter-domain regions, with a slight preference for the former when sites in disordered environment were considered. Interestingly, such properties are characteristic for the interfaces of stable protein-protein complexes, suggesting a possible overlap between the two environments. This hypothesis is in agrement with our current knowledge of the role of modified tyrosine residues in the regulation of interactions. Furthermore, as tyrosine residues are generally known to make an important contribution towards the free binding energy of many complexes, it would be interesting to investigate the intersection between phospho-tyrosines and

known hot spots.

Tyrosine phosphorylation is involved in the regulation of a myriad of cellular processes and deviations from its proper modification often have severe consequences. Therefore, based on the distinct properties of phospho-tyrosines, it can be hypothesized that the majority of these sites have a strictly regulated role and are well-suited for accurate control of protein-protein interactions both through direct effect on the binding affinity and the structural fit and through indirect allosteric changes.

Phosphorylation sites with regulatory functions are more evolutionary conserved, more buried and prefer coil regions

The question of the existence of non-functional phosphorylation sites is becoming more and more urgent due to the rapid increase in the number of identified sites that lack a defined function. This possibility is supported by the unexpectedly low conservation characterizing the majority of these sites and the hypothesis that more abundant proteins are prone to the random actions of kinases, thus resulting in unspecific phosphorylation events. Indeed, the results in this thesis show that the phospho-acceptor sites with an assigned regulatory function had different properties compared to those with unknown function. Their preference for more ordered coil regions and lower solvent exposure is indicative of the specific environment that is required to facilitate their proper recognition and functioning. Furthermore, the higher evolutionary conservation suggests that the modification of these sites have important functional roles.

In summary, the specific properties of regulatory phosphorylation sites may become a useful instrument to distinguish functional from silent modifications. However, it would be difficult to claim the lack of functional relevance of any site, as it is possible that what is lacking is the knowledge and the understanding of its function. Therefore, a more likely hypothesis is that such differences between the various modification sites may enable the identification of distinct functional classes of phospho-sites and in general of post-translational modifications (e.g. regulatory versus fine-tuning as a part of a set of modifications) and provide deeper insights into the underlying mechanisms.

The functional analysis of phospho-sites would greatly benefit from methods that

on one hand assess the conservation of the functional role of a site, regardless of its positional conservation, and on the other hand distinguish between conserved and evolving disordered regions. Phosphorylation sites that lie within structurally-defined regions or that are associated with specific conformational changes, such as loop activation, are often facilitated by complex networks of hydrogen bonding with the surrounding residues. Therefore, it is natural that such sites exhibit strong evolutionary pressure. In contrast, phosphorylation sites present in disordered regions are characterized by much lower conservation, in line with the rapidly evolving nature of these regions. Often such sites are involved in the regulation of protein-protein interactions by influencing various electrostatic and steric properties. Thus the actual position of the phospho-acceptor site may not be as important as the overall effect of the phosphate group on the particular region in the protein. It would therefore be interesting to systematically investigate the existence of an alternative form of conservation, in which not the position, but the function is retained. Furthermore, the properties of disordered regions need to be considered with great caution while investigating conservation of modification sites, due to the existence of different types of disordered regions. The possibility that kinase consensus motifs shift along the disordered regions or that they are transformed into motifs of other kinases may also contribute to the better understanding of the evolutionary pressure acting on phospho-sites.

Multiple phospho-sites and PTM cross-talk increase the complexity of the proteome and add evolutionary flexibility

The investigation of the co-occurrence of phosphorylated residues and modified lysines revealed that the two appeared at shorter distances than would be expected if they were scattered randomly over the protein length. This tendency provides strong evidence that the cross-talk between the different types of modifications may add another layer of complexity to the intricate system of regulation in the cell. It boosts the number of functional states of a protein and further fine-tunes its interaction network. Additionally, the presence of various modifiable residues allows the incorporation of signals from distinct pathways and facilitates the precise timing and regulation of the cellular response. The proportion of modified lysine residues in the close proximity of phospho-tyrosines was the highest among the three phospho-acceptor sites, which further supports the hypothesis that the majority of

the phosphorylated tyrosines have a well-defined regulatory role.

Phosphorylation dynamics scales with structural propensities

Our data allowed us to investigate the kinase preferences of phosphorylation sites with high vs. low levels of regulation. The results highlight the central role of proline, as a disorder-promoting residue that is also part of regulatory motifs [303]. The directing role of proline together with the multiple functions associated with disorder explain the more variable character of phosphorylation of sites with these properties that we observe in our study. Furthermore, our statistical analysis of the interplay between structure and phosphorylation variation in relation to specific kinase recognition motifs presents a new approach of describing and classifying protein kinases. We showed that the combination of both properties can be used to gain conceptual and specific insights into regulation. We were able to reproduce known relations and to identify new links between kinases, which may reflect functional dependencies emerging from common regulatory behavior and structural preferences.

In conclusion, we have related the tendency of phosphorylation sites to be dynamically regulated throughout the cell cycle with the structural features of the sites. While we have found clear relations between phosphorylation dynamics and protein structure, we are only scratching the surface of what we believe could be an exciting new area at the interface of proteomics and structural biology.

Phosphorylation sites have a tendency to cluster in regions that lack defined structure. An interesting question arising from this phenomenon is if proximal phosphorylation sites can compensate each other (i.e. if only the total effect plays a role) or if they act in concert (i.e. each of them is regulated in a specific manner). To address this question quantitative information in the form of absolute protein quantification and quantification of the exact amount of phosphorylation of each site (occupancy) is required. A suitable set-up for such investigation would involve different perturbations and various time points at which changes in occupancy are measured. In a particular scenario a cluster of phospho-sites may retain an overall constant amount of phosphorylation, which would suggest additive function or presence of silent phosphorylation events. In different settings, the sites in the cluster may be regulated in a similar fashion indicating the distinct functional role of each

of them. Such a study would clearly add to our understanding of both the interplay between multiple phosphorylation sites and their functional relevance.

Classification of cancer patients based on their proteome profiles requires feature selection embedded in cross validation

The efficient treatment of cancer patients greatly depends on accurate subtype classification, which despite the available and long established biomarkers is not always straightforward and often suboptimal. Clinical mass spectrometry-based proteomics is becoming an increasingly powerful technology for addressing the needs for better diagnostic and discovery of novel biomarkers. The current advances in sample preparation and quantification of proteins in tissues enable the characterization of thousands of proteins from patient samples, thus providing ample data encoding the disease mechanisms and its impact on the organism. The size and complexity of these data, however, necessitate the development and use of sophisticated analytical tools. A classification framework was developed that allows feature selection in a rigorous manner and enables any scientist to perform supervised learning on proteomics data. It addresses the main challenges, related to the tasks of signature detection and subsequent sample classification: (i) high biological variability among patient samples and (ii) large feature space combined with low sample size.

The framework is built as a plug-in (called 'Learning') to the statistical software Perseus, becoming increasingly popular in the analysis of large-scale proteomics data. Its generic implementation allows the addition of various supervised learning methods, however, the focus in this thesis is in particular on support vector machines. The framework provides the modules: classification and prediction of clinical proteomics samples, feature selection and parameter optimization. The convenient implementation enables scientist who are not necessarily experts in learning theory to employ supervised learning methods to analyse complex data. In the classification module, the accuracy of a classifier trained on the data can be estimated using any of three cross validation procedures. Furthermore, this classifier can then be used to predict the labels of new unlabeled samples. The parameters required for optimal performance of the support vector machines can be conveniently explored in the 'Parameter optimization' module of the framework.

Perhaps the most powerful module of the three is the 'Feature optimization' option. In addition to improving the classifier's performance, feature selection allows the identification of disease-relevant features with high clinical relevance. The framework supports various feature selection methods. The key to recognizing a reliable set of features is to avoid overfitting during the feature selection procedure. Performing feature selection on the entire data set is still a common mistake, which leads to the identification of features that discriminate well between the classes of the training samples, but have poor performance in independent test sets. The developed framework overcomes this problem by enforcing feature selection to be always embedded in a cross validation procedure, ascertaining maximum generalizability of the results.

Thus, an easy to use analytical framework was developed, that provides the user with the ability to employ various state-of-the art methods for discovering patterns underlying complex proteomics data. The software combines an user-friendly graphical interface with rigorous implementation of the analytical procedures, ensuring their proper application and reliability of the results.

The detection of biomarkers is plausible and holds great promise for the future

Chapter 5 demonstrated the successful application of the earlier described analysis framework (see Chapter 4) to feature selection and classification of the subtypes of two types of cancers – prostate and breast cancer – based on protein expression data.

The optimized analysis showed that despite the limited number of samples it is possible to distinguish disease-related patterns and to extract biologically-relevant features from large-scale proteomics data sets. Furthermore, the comparison of the different ranking methods revealed that due to their underlying principles they can serve different purposes. Means-based methods are tailored for the detection of single features that strongly discriminate between the classes and ignore any interactions among the features. These methods thus result in an overall small number of features and are suitable for the discovery of potential biomarkers that could be directly used in diagnostics and treatment. In contrast, the SVM weights-based

ranking method estimates the predictive power of complete sets of features. The highest accuracy using this method is reached with a much larger number of ranked features and it can therefore be used to increase our understanding of the mechanisms and implications of the disease.

In summary, clinical proteomics combined with rigorously-applied data mining techniques holds great promise for improvements in the field of personalized medicine. The foundations have already been laid, however, several improvements would lead to even better performance of the developed framework.

In a continuation of the work in this thesis it would be highly interesting to extend the presented methods to different proteomics platforms. A medical test used on daily basis has to be fast and reliable, therefore a procedure that is able to achieve high accuracy and precision in diagnostics as well as in correct treatment assignment from small amounts of samples and in a short time is highly desirable. For instance, single-shot proteomics provides such a platform by skipping the pre-fractionation step and by making use of highly sensitive mass spectrometers [304]. A systematic research is, however, needed to assess if the depth of the data measured in this way would be enough for clinical applications.

Another currently unaccounted for aspect is the regulation of the functions of proteins through post-translational modifications. Modifications can alter the structure and the function of the proteins adding a new level of complexity to the cellular organization and the underlying mechanisms. PTMs are often responsible for the activation of oncogenes or the inactivation of tumor-suppressor genes, turning them into important factors for the process of cancer development. For instance, tyrosine kinases have become one of the largest classes of drug targets for cancer treatment [305] due to their regulatory role in a wide range of cellular processes. It is therefore of high interest to make use of the information at the post-translational level (i.e. modification profiles) in order to build accurate classifiers and unravel informative features.

Furthermore, the developed framework for analysis would greatly benefit from a larger cohort of samples. Due to the fast speed of development of mass spectrometry-based proteomics the profiling of tens and even of hundreds of samples will soon be

feasible. In addition to increasing the number of samples, coupling tumor tissue to healthy tissue from the same patient would provide the ultimate remedy to the heterogeneity problem. The presence of healthy tissue can be used to compute ratios of the protein expression levels between tumor and healthy tissues in order to cancel the genetic variability and intensify the disease-related signal. In addition, a collection of tumor samples from different areas in the tumor may help to avoid misclassifications resulting from the existence of different cell subpopulations in the same tumor [306].

Bibliography

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

[2] B. F. Cravatt, G. M. Simon, and J. R. Yates, 3rd. The biological impact of mass-spectrometry-based proteomics. *Nature*, 450(7172):991–1000, Dec 2007.

[3] J. Cox and M. Mann. Is proteomics the new genomics? *Cell*, 130(3):395–398, Aug 2007.

[4] J. Cox and M. Mann. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu Rev Biochem*, 80:273–299, Jun 2011.

[5] M. Mann. Functional and quantitative proteomics using silac. *Nat Rev Mol Cell Biol*, 7(12):952–958, Dec 2006.

[6] J. Cox and M. Mann. Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, 26(12):1367–1372, Dec 2008.

[7] J. Cox and M. Mann. Computational principles of determining and improving mass precision and accuracy for proteome measurements in an orbitrap. *J Am Soc Mass Spectrom*, 20(8):1477–1485, Aug 2009.

[8] J. Cox, A. Michalski, and M. Mann. Software lock mass by two-dimensional minimization of peptide mass errors. *J Am Soc Mass Spectrom*, 22(8):1373–1380, Aug 2011.

[9] A. I. Lamond, M. Uhlen, S. Horning, A. Makarov, C.l V. Robinson, L. Serrano, F. U. Hartl, W. Baumeister, A. K. Werenskiold, J. S. Andersen, M. Vorm, O.and Linial, R. Aebersold, and M. Mann. Advancing cell biology through proteomics in space and time (prospects). *Mol Cell Proteomics*, 11(3):O112.017731, Mar 2012.

[10] M. Mann, R. C. Hendrickson, and A. Pandey. Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem*, 70:437–473, 2001.

[11] M. Mann, S. E. Ong, M. Grnborg, H. Steen, O. N. Jensen, and A. Pandey. Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol*, 20(6):261–268, Jun 2002.

[12] T. Nilsson, M. Mann, R. Aebersold, J. R. Yates, 3rd, A. Bairoch, and J. J. M. Bergeron. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods*, 7(9):681–685, Sep 2010.

[13] P. Mallick and B. Kuster. Proteomics: a pragmatic perspective. *Nat Biotechnol*, 28(7):695–709, Jul 2010.

[14] M. Walhout, M. Vidal, and J. Dekker, editors. *Handbook of systems biology: Concepts and Insights.* Elsevier, 2013.

[15] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, Oct 20101989.

[16] L. N. Mueller, M. Brusniak, D. R. Mani, and R. Aebersold. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J Proteome Res*, 7(1):51–61, Jan 2008.

[17] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem*, 404(4):939–965, Sep 2012.

[18] S. Ong and M. Mann. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol*, 1(5):252–262, Oct 2005.

[19] H. Liu, R. G. Sadygov, and J. R Yates, 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem*, 76(14):4193–4201, Jul 2004.

[20] P. Lu, C. Vogel, R. Wang, X Yao, and E. M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25(1):117–124, Jan 2007.

[21] P. V. Bondarenko, D. Chelius, and T. A. Shaler. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem*, 74(18):4741–4749, Sep 2002.

[22] M. Ono, M. Shitashige, K. Honda, T. Isobe, H. Kuwabara, H. Matsuzuki, S. Hirohashi, and T. Yamada. Label-free quantitative proteomics using large peptide data sets generated by nanoflow liquid chromatography and mass spectrometry. *Mol Cell Proteomics*, 5(7):1338–1347, Jul 2006.

[23] J. M. Asara, H. R. Cristofk, L. M. Freimark, and L. C. Cantley. A label-free quantification method by ms/ms tic compared to silac and spectral counting in a proteomics screen. *Proteomics*, 8(5):994–999, Mar 2008.

[24] O. A. Mirgorodskaya, Y. P. Kozmin, M. I. Titov, R. Krner, C. P. Snksen, and P. Roepstorff. Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)o-labeled internal standards. *Rapid Commun Mass Spectrom*, 14(14):1226–1232, 2000.

[25] K. C. S. Rao, V. Palamalai, J. R. Dunlevy, and M. Miyagi. Peptidyl-lys metalloendopeptidase-catalyzed 18o labeling for comparative proteomics: application to cytokine/lipolysaccharide-treated human retinal pigment epithelium cell line. *Mol Cell Proteomics*, 4(10):1550–1557, Oct 2005.

[26] M. Bantscheff, B. Dmpelfeld, and B. Kuster. Femtomol sensitivity post-digest (18)o labeling for relative quantification of differential protein complex composition. *Rapid Commun Mass Spectrom*, 18(8):869–876, 2004.

[27] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, 17(10):994–999, Oct 1999.

[28] Y. Lu, P. Bottari, F. Turecek, R..i Aebersold, and M. H. Gelb. Absolute quantification of specific proteins in complex mixtures using visible isotope-coded affinity tags. *Anal Chem*, 76(14):4104–4111, Jul 2004.

[29] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*, 3(12):1154–1169, Dec 2004.

[30] J. J. Kusmierz, R. Sumrada, and D. M. Desiderio. Fast atom bombardment mass spectrometric quantitative analysis of methionine-enkephalin in human pituitary tissues. *Anal Chem*, 62(21):2395–2400, Nov 1990.

[31] O. Stemmann, H. Zou, S. A. Gerber, S. P. Gygi, and M. W. Kirschner. Dual inhibition of sister chromatid separation at metaphase. *Cell*, 107(6):715–726, Dec 2001.

[32] S. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1(5):376–386, May 2002.

[33] H. Zhu, S. Pan, S. Gu, E. M. Bradbury, and X. Chen. Amino acid residue specific stable isotope labeling for quantitative proteomics. *Rapid Commun Mass Spectrom*, 16(22):2115–2123, 2002.

[34] H. Jiang and A. M. English. Quantitative analysis of the yeast proteome by incorporation of isotopically labeled leucine. *J Proteome Res*, 1(4):345–350, 2002.

[35] S. Ong and M. A practical recipe for stable isotope labeling by amino acids in cell culture (silac). *Nat Protoc*, 1(6):2650–2660, 2006.

[36] S. Hanke, H. Besir, D. Oesterhelt, and M. Mann. Absolute silac for accurate quantitation of proteins in complex mixtures down to the attomole level. *J Proteome Res*, 7(3):1118–1130, Mar 2008.

[37] J. S. Andersen, Y. W. Lam, A. K. L. Leung, S. Ong, C. E. Lyon, A. I. Lamond, and M. Mann. Nucleolar proteome dynamics. *Nature*, 433(7021):77–83, Jan 2005.

[38] L. M. F. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frhlich, T. C. Walther, and M. Mann. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217):1251–1254, Oct 2008.

[39] J. R Yates, 3rd, A. Gilchrist, K. E. Howell, and J. J. M. Bergeron. Proteomics of organelles and large cellular structures. *Nat Rev Mol Cell Biol*, 6(9):702–714, Sep 2005.

[40] T. C. Walther and M. Mann. Mass spectrometry-based proteomics in cell biology. *J Cell Biol*, 190(4):491–500, Aug 2010.

[41] V. K. Mootha, J. Bunkenborg, J. V. Olsen, M. Hjerrild, J. R. Wisniewski, E. Stahl, M. S. Bolouri, H. N. Ray, S. Sihag, M. Kamal, N. Patterson, E. S. Lander, and M. Mann. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell*, 115(5):629–640, Nov 2003.

[42] J. S. Andersen, C. J. Wilkinson, T. Mayor, P. Mortensen, E. A. Nigg, and M. Mann. Proteomic characterization of the human centrosome by protein correlation profiling. *Nature*, 426(6966):570–574, Dec 2003.

[43] T. P. J. Dunkley, R. Watson, J. L. Griffin, P. Dupree, and K. S. Lilley. Localization of organelle proteins by isotope tagging (lopit). *Mol Cell Proteomics*, 3(11):1128–1134, Nov 2004.

[44] D. J. L.. Tan, H. Dvinge, A. Christoforou, P. Bertone, A. Martinez Arias, and K. S. Lilley. Mapping organelle proteins and protein complexes in drosophila melanogaster. *J Proteome Res*, 8(6):2667–2678, Jun 2009.

[45] A. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dmpelfeld, A. Edelmann, M. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. 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(7084):631–636, Mar 2006.

[46] N. J. Krogan, G. Cagney, D. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrn-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643, Mar 2006.

[47] A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik, and JR Yates, 3rd. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol*, 17(7):676–682, Jul 1999.

[48] M. Selbach and M. Mann. Protein interaction screening by quantitative immunoprecipitation combined with knockdown (quick). *Nat Methods*, 3(12):981–983, Dec 2006.

[49] N. C. Hubner, A. W. Bird, J. Cox, B. Splettstoesser, P. Bandilla, I. Poser, A. Hyman, and M. Mann. Quantitative proteomics combined with bac transgeneomics reveals in vivo protein interactions. *J Cell Biol*, 189(4):739–754, May 2010.

[50] F. Butter, M. Scheibe, M. Mrl, and M. Mann. Unbiased rna-protein interaction screen by quantitative proteomics. *Proc Natl Acad Sci U S A*, 106(26):10626–10631, Jun 2009.

[51] W. X. Schulze and M. Mann. A novel proteomic screen for peptide-protein interactions. *J Biol Chem*, 279(11):10756–10764, Mar 2004.

[52] S. Hanke and M. Mann. The phosphotyrosine interactome of the insulin receptor family and its substrates irs-1 and irs-2. *Mol Cell Proteomics*, 8(3):519–534, Mar 2009.

[53] O. N. Jensen. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol*, 7(6):391–403, Jun 2006.

[54] P. Cohen. The regulation of protein function by multisite phosphorylation–a 25 year update. *Trends in biochemical sciences*, 25(12):596–601, 2000.

[55] J. Fla and D. Honys. Enrichment techniques employed in phosphoproteomics. *Amino Acids*, 43(3):1025–1047, Sep 2012.

[56] D. C. Neville, C. R. Rozanas, E. M. Price, D. B. Gruis, A. S. Verkman, and R. R. Townsend. Evidence for phosphorylation of serine 753 in cftr using a novel metal-ion affinity resin and matrix-assisted laser desorption mass spectrometry. *Protein Sci*, 6(11):2436–2445, Nov 1997.

[57] J. Ye, X. Zhang, C. Young, X. Zhao, Q. Hao, L. Cheng, and O. N. Jensen. Optimized imac-imac protocol for phosphopeptide recovery from complex biological samples. *J Proteome Res*, 9(7):3561–3573, Jul 2010.

[58] Y. Ikeguchi and H. Nakamura. Determination of organic phosphates by column-switching high performance anion-exchange chromatography using on-line preconcentration on titania. *Anal Sci*, 1997.

[59] S. A. Beausoleil, M. Jedrychowski, D. Schwartz, J. E. Elias, J. Villn, 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(33):12130–12135, Aug 2004.

[60] A. Amoresano, A. Carpentieri, C. Giangrande, A. Palmese, G. Chiappetta, G. Marino, and P. Pucci. Technical advances in proteomics mass spectrometry: identification of post-translational modifications. *Clin Chem Lab Med*, 47(6):647–665, 2009.

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

[62] S. A. Beausoleil, J. Villn, S. A. Gerber, J. Rush, and S. P. Gygi. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol*, 24(10):1285–1292, Oct 2006.

[63] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J.r V. Olsen, and M. Mann. Andromeda: a peptide search engine integrated into the maxquant environment. *J Proteome Res*, 10(4):1794–1805, Apr 2011.

[64] J. A. J. Steen, H. Steen, A. Georgi, K. Parker, M. Springer, M. Kirchner, F. Hamprecht, and M. W. Kirschner. Different phosphorylation states of the anaphase promoting complex in response to antimitotic drugs: a quantitative proteomic analysis. *Proc Natl Acad Sci U S A*, 105(16):6069–6074, Apr 2008.

[65] J. V. Olsen, M.l Vermeulen, A. Santamaria, C. Kumar, M. L. Miller, L. J. Jensen, F. Gnad, J. Cox, T. S. Jensen, E. A. Nigg, S. Brunak, and M. Mann. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal*, 3(104):ra3, 2010.

[66] Z. Wu, J. B. Doondeea, A. M. Gholami, M. C. Janning, S. Lemeer, K. Kramer, S. A. Eccles, S. M. Gollin, R. Grenman, A. Walch, S. M. Feller, and B. Kuster. Quantitative chemical proteomics reveals new potential drug targets in head and neck cancer. *Mol Cell Proteomics*, 10(12):M111.011635, Dec 2011.

[67] B. Blagoev, S. Ong, I. Kratchmarova, and M. Mann. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat Biotechnol*, 22(9):1139–1145, Sep 2004.

[68] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, Dec 2002.

[69] S. K. Hanks and T. Hunter. Protein kinases 6. the eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J*, 9(8):576–596, May 1995.

[70] X. Zeng, K. Tamai, B. Doble, S. Li, H. Huang, R. Habas, H. Okamura, J. Woodgett, and X. He. A dual-kinase mechanism for wnt co-receptor phosphorylation and activation. *Nature*, 438(7069):873–877, Dec 2005.

[71] J. A. Ubersax and J. E. Ferrell, Jr. Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol*, 8(7):530–541, Jul 2007.

[72] R. B. Pearson and B. E. Kemp. Protein kinase phosphorylation site sequences and consensus specificity motifs: tabulations. *Methods Enzymol*, 200:62–81, 1991.

[73] N. E. Davey, K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, and T. J. Gibson. Attributes of short linear motifs. *Mol Biosyst*, 8(1):268–281, Jan 2012.

[74] B. E. Kemp, D. J. Graves, E. Benjamini, and E. G. Krebs. Role of multiple basic residues in determining the substrate specificity of cyclic amp-dependent protein kinase. *J Biol Chem*, 252(14):4888–4894, Jul 1977.

[75] N. R. Brown, M. E. Noble, J. A. Endicott, and L. N. Johnson. The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. *Nature cell biology*, 1(7):438–43, 1999.

[76] J. Alexander, D. Lim, B. A. Joughin, B. Hegemann, James R .A. H., T. Ehrenberger, F. Ivins, O. Sessa, F.and Hudecz, E. A. Nigg, A. M. Fry, A. Musacchio, P. T. Stukenberg, K. Mechtler, J. Peters, S. J. Smerdon, and M. B. Yaffe. Spatial exclusivity combined with positive and negative selection of phosphorylation motifs is the basis for context-dependent mitotic signaling. *Sci Signal*, 4(179):ra42, 2011.

[77] A. Zanzoni, G. Ausiello, A. Via, P. F. Gherardini, and M. Helmer-Citterich. Phospho3d: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res*, 35(Database issue):D229–D231, Jan 2007.

[78] A. Zanzoni, D. Carbajo, F. Diella, P. F. Gherardini, A. Tramontano, M. Helmer-Citterich, and A. Via. Phospho3d 2.0: an enhanced database of three-dimensional structures of phosphorylation sites. *Nucleic acids research*, 39(Database issue):D268–71, 2011.

[79] J. L. Jimenez, 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 biology*, 8(5):R90, 2007.

[80] P. Durek, C. Schudoma, W. Weckwerth, J. Selbig, and D. Walther. Detection and characterization of 3d-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC bioinformatics*, 10:117, 2009.

[81] L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O'Connor, J. G. Sikes, Z. Obradovic, and A. K. Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic acids research*, 32(3):1037–49, 2004.

[82] 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 biology*, 8(11):R250, 2007.

[83] N. Blom, S. Gammeltoft, and S. Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*, 294(5):1351–1362, Dec 1999.

[84] J. Kitchen, R. E. Saunders, and J. Warwicker. Charge environments around phosphorylation sites in proteins. *BMC structural biology*, 8:19, 2008.

[85] B. A. Joughin, B. Tidor, and M. B. Yaffe. A computational method for the analysis and prediction of protein:phosphopeptide-binding sites. *Protein Sci*, 14(1):131–139, Jan 2005.

[86] A. C. A. Roque and C. R. Lowe. Lessons from nature: On the molecular recognition elements of the phosphoprotein binding-domains. *Biotechnol Bioeng*, 91(5):546–555, Sep 2005.

[87] A. Krupa, G. Preethi, and N. Srinivasan. Structural modes of stabilization of permissive phosphorylation sites in protein kinases: distinct strategies in ser/thr and tyr kinases. *J Mol Biol*, 339(5):1025–1039, Jun 2004.

[88] D. Plewczyski, A. Tkacz, A. Godzik, and L. Rychlewski. A support vector machine approach to the identification of phosphorylation sites. *Cell Mol Biol Lett*, 10(1):73–89, 2005.

[89] A. Via, F. Diella, T. J. Gibson, and M. Helmer-Citterich. From sequence to structural analysis in protein phosphorylation motifs. *Front Biosci*, 16:1261–1275, 2011.

[90] P. Puntervoll, R. Linding, C. Gemnd, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. A. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferr, 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. Kster, 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(13):3625–3630, Jul 2003.

[91] J. C. Obenauer, L. C. Cantley, and M. B. Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 31(13):3635–3641, Jul 2003.

[92] J. Kim, J. Lee, B. Oh, K. Kimm, and I. Koh. Prediction of phosphorylation sites using svms. *Bioinformatics*, 20(17):3179–3184, Nov 2004.

[93] D. Schwartz and S. P. Gygi. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol*, 23(11):1391–1398, Nov 2005.

[94] P. V. Hornbeck, J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, and M. Sullivan. Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res*, 40(Database issue):D261–D270, Jan 2012.

[95] H. Dinkel, C. Chica, A. Via, C. M. Gould, . J. Jensen, T. J. Gibson, and F. Diella. Phospho.elm: a database of phosphorylation sites–update 2011. *Nucleic Acids Res*, 39(Database issue):D261–D267, Jan 2011.

[96] F. Gnad, J. Gunawardena, and M. Mann. Phosida 2011: the posttranslational modification database. *Nucleic Acids Res*, 39(Database issue):D253–D260, Jan 2011.

[97] C. Schaab, T. Geiger, G. Stoehr, J. Cox, and M. Mann. Analysis of high accuracy, quantitative proteomics data in the maxqb database. *Mol Cell Proteomics*, 11(3):M111.014068, Mar 2012.

[98] M. Hjerrild and S. Gammeltoft. Phosphoproteomics toolbox: computational biology, protein chemistry and mass spectrometry. *FEBS Lett*, 580(20):4764–4770, Sep 2006.

[99] L.N. Johnson and D. Barford. The effects of phosphorylation on the structures and function of proteins. *Annu. Rev. Biophys. Biomol. Struct.*, 22:199–232, 1993.

[100] T. Hunter. Tyrosine phosphorylation: thirty years and counting. *Curr Opin Cell Biol*, 21(2):140–146, Apr 2009.

[101] L. N. Johnson. The regulation of protein phosphorylation. *Biochem Soc Trans*, 37(Pt 4):627–641, Aug 2009.

[102] T. Pawson and J. D. Scott. Protein phosphorylation in signaling–50 years and counting. *Trends Biochem Sci*, 30(6):286–290, Jun 2005.

[103] H. Nishi, K. Hashimoto, and A. R. Panchenko. Phosphorylation in protein-protein binding: effect on stability and function. *Structure*, 19(12):1807–1815, Dec 2011.

[104] A. A. Russo, P. D. Jeffrey, and N. P. Pavletich. Structural basis of cyclin-dependent kinase activation by phosphorylation. *Nat Struct Biol*, 3(8):696–700, Aug 1996.

[105] J. P I. Welburn, J. A. Tucker, T. Johnson, L. Lindert, M. Morgan, A. Willis, M. E. M. Noble, and J. A. Endicott. How tyrosine 15 phosphorylation inhibits the activity of cyclin-dependent kinase 2-cyclin a. *J Biol Chem*, 282(5):3173–3181, Feb 2007.

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

[107] G. E. Lienhard. Non-functional phosphorylations? *Trends Biochem Sci*, 33(8):351–352, Aug 2008.

[108] D. S. Tawfik. Messy biology and the origins of evolutionary innovations. *Nat Chem Biol*, 6(10):692–696, Oct 2010.

[109] L. Wang, Q. Nie, and G. Enciso. Nonessential sites improve phosphorylation switch. *Biophys J*, 99(6):L41–L43, Sep 2010.

[110] E. D. Levy, S. W. Michnick, and C. R. Landry. Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information. *Philos Trans R Soc Lond B Biol Sci*, 367(1602):2594–2606, Sep 2012.

[111] C. Heng Tan and G. D. Bader. Phosphorylation sites of higher stoichiometry are more conserved. *Nat Methods*, 9(4):317; author reply 318, Apr 2012.

[112] C. R. Landry, E. D. Levy, and S. W. Michnick. Weak functional constraints on phosphoproteomes. *Trends Genet*, 25(5):193–197, May 2009.

[113] V. E. Gray and S. Kumar. Rampant purifying selection conserves positions with posttranslational modifications in human proteins. *Mol Biol Evol*, 28(5):1565–1568, May 2011.

[114] J. Boekhorst, B. van Breukelen, A. Heck, Jr, and B. Snel. Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol*, 9(10):R144, 2008.

[115] R. Malik, E. A. Nigg, and R. Krner. Comparative conservation analysis of the human mitotic phosphoproteome. *Bioinformatics*, 24(12):1426–1432, Jun 2008.

[116] 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(12):2642–2653, Dec 2010.

[117] W. C. S. Cho. Contribution of oncoproteomics to cancer biomarker discovery. *Mol Cancer*, 6:25, 2007.

[118] R. J. Simpson, O. K. Bernhard, D. W. Greening, and R. L. Moritz. Proteomics-driven cancer biomarker discovery: looking to the future. *Curr Opin Chem Biol*, 12(1):72–77, Feb 2008.

[119] S. M. Hanash, C. S. Baik, and O. Kallioniemi. Emerging molecular biomarkers–blood-based strategies to detect and monitor cancer. *Nat Rev Clin Oncol*, 8(3):142–150, Mar 2011.

[120] M. Mann. Proteomics for biomedicine: a half-completed journey. *EMBO Mol Med*, 4(2):75–77, Feb 2012.

[121] E. S. Baker, T. Liu, V. A. Petyuk, K. E. Burnum-Johnson, Y. M. Ibrahim, G. A. Anderson, and R. D. Smith. Mass spectrometry for translational proteomics: progress and clinical implications. *Genome Med*, 4(8):63, Aug 2012.

[122] T. Geiger, S. F. Madden, W. M. Gallagher, J. Cox, and M. Mann. Proteomic portrait of human breast cancer progression identifies novel prognostic markers. *Cancer Res*, 72(9):2428–2439, May 2012.

[123] B. Matharoo-Ball, G. Ball, and R. Rees. Clinical proteomics: discovery of cancer biomarkers using mass spectrometry and bioinformatics approaches–a prostate cancer perspective. *Vaccine*, 25 Suppl 2:B110–B121, Sep 2007.

[124] J. L. Hays, G. Kim, I. Giuroiu, and E. C. Kohn. Proteomics and ovarian cancer: integrating proteomics information into clinical care. *J Proteomics*, 73(10):1864–1872, Sep 2010.

[125] A. Taguchi, K. Politi, S. J. Pitteri, W. W. Lockwood, V. M. Faa, K. Kelly-Spratt, C. Wong, Q. Zhang, A. Chin, K. Park, G. Goodman, A. F. Gazdar, J. Sage, D. M. Dinulescu, R. Kucherlapati, R. A. Depinho, C. J. Kemp, H. E. Varmus, and S. M. Hanash. Lung cancer signatures in plasma based on proteome profiling of mouse tumor models. *Cancer Cell*, 20(3):289–299, Sep 2011.

[126] C. R. Jimenez, J. C. Knol, G. A. Meijer, and R. J. A. Fijneman. Proteomics of colorectal cancer: overview of discovery studies and identification of commonly identified cancer-associated proteins and candidate crc serum markers. *J Proteomics*, 73(10):1873–1895, Sep 2010.

[127] S. J. Deeb, R. C J. D'Souza, J. Cox, M. Schmidt-Supprian, and M. Mann. Supersilac allows classification of diffuse large b-cell lymphoma subtypes by their protein expression profiles. *Mol Cell Proteomics*, 11(5):77–89, May 2012.

[128] M. S. Ritorto and J. Borlak. Combined serum and tissue proteomic study applied to a c-myc transgenic mouse model of hepatocellular carcinoma identified novel disease regulated proteins suitable for diagnosis and therapeutic intervention strategies. *J Proteome Res*, 10(7):3012–3030, Jul 2011.

[129] T. A. Addona, S. E. Abbatiello, B. Schilling, S. J. Skates, D. R. Mani, D. M. Bunk, C. H. Spiegelman, L. J. Zimmerman, A. L. Ham, H. Keshishian, S. C. Hall, S. Allen, R. K. Blackman, C. H. Borchers, C. Buck, H. L. Cardasis, M. P. Cusack, N. G. Dodder, B. W. Gibson, J. M. Held, T. Hiltke, A. Jackson, E. B. Johansen, C. R. Kinsinger, J. Li, M. Mesri, T. A. Neubert, R. K. Niles, T. C. Pulsipher, D. Ransohoff, H. Rodriguez, P. A. Rudnick, D. Smith, D. L. Tabb, T. J. Tegeler, A. M. Variyath, L. J. V., A. Wahlander, S. Waldemarson, M. Wang, J. R. Whiteaker, L. Zhao, N. L. Anderson, S. J. Fisher, D. C. Liebler, A. G. Paulovich, F. E. Regnier, P. Tempst, and S. A. Carr. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat Biotechnol*, 27(7):633–641, Jul 2009.

[130] N. L. Anderson, N. G. Anderson, L. R. Haines, D.l B. Hardie, R. W. Olafson, and T. W. Pearson. Mass spectrometric quantitation of peptides and proteins using stable isotope standards and capture by anti-peptide antibodies (siscapa). *J Proteome Res*, 3(2):235–244, 2004.

[131] N. Nagaraj, J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Pbo, and M. Mann. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol*, 7:548, 2011.

[132] J. R. Winiewski, K. Du, and M. Mann. Proteomic workflow for analysis of archival formalin fixed and paraffin embedded clinical samples to a depth of 10,000 proteins. *Proteomics Clin Appl*, Oct 2012.

[133] P. Ostasiewicz, D. F. Zielinska, M. Mann, and J. R. Winiewski. Proteome, phosphoproteome, and n-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry. *J Proteome Res*, 9(7):3688–3700, Jul 2010.

[134] J. R. Winiewski, A. Zougman, N. Nagaraj, and M. Mann. Universal sample preparation method for proteome analysis. *Nat Methods*, 6(5):359–362, May 2009.

[135] T. Geiger, J. Cox, P. Ostasiewicz, J. R. Wisniewski, and M. Mann. Super-silac mix for quantitative proteomics of human tumor tissue. *Nat Methods*, 7(5):383–385, May 2010.

[136] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, and R. A. Clark. Data mining techniques for cancer detection using serum proteomic profiling. *Artif Intell Med*, 32(2):71–83, Oct 2004.

[137] A. A. Alaiya, B. Franzn, G. Auer, and S. Linder. Cancer proteomics: from identification of novel markers to creation of artifical learning models for tumor classification. *Electrophoresis*, 21(6):1210–1217, Apr 2000.

[138] F. Schleif, T. Villmann, B. Hammer, and M. van der Werff. *Computational Intelligence in Biomedicine and Bioinformatics Studies in Computational Intelligence : Analysis of Spectral Data in Clinical Proteomics by Use of Learning Vector Quantizers.* Springer Link, 2008.

[139] H. Han. A high performance profile-biomarker diagnosis for mass spectral profiles. *BMC Syst Biol*, 5 Suppl 2:S5, Dec 2011.

[140] B. Adam, Y. Qu, J. W. Davis, M. D. Ward, M. A. Clements, L. H. Cazares, O J. Semmes, P. F. Schellhammer, Y. Yasui, Z. Feng, and Jr. G. L. Wright. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res*, 62(13):3609–3614, Jul 2002.

[141] G. R. Ball, S. Mian, F. Holding, R. O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I. O. Ellis, C. Creaser, and R. C. Rees. An integrated approach utilizing artificial neural networks and seldi mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*, 18(3):395–404, 2002.

[142] N. Ishigami, T. Tokuda, M. Ikegawa, M. Komori, T. Kasai, T. Kondo, Y. Matsuyama, T. Nirasawa, H. Thiele, K. Tashiro, and M. Nakagawa. Cerebrospinal fluid proteomic patterns discriminate parkinson's disease and multiple system atrophy. *Mov Disord*, 27(7):851–857, Jun 2012.

[143] E. Marchiori, C. R. Jimenez, M. West-Nielsen, and N. H. H. Heegaard. Robust svm-based biomarker selection with noisy mass spectrometric proteomic data. In *Proceedings of the 2006 international conference on Applications of Evolutionary Computing*, EuroGP'06, pages 79–90, Berlin, Heidelberg, 2006. Springer-Verlag.

[144] J. R. Winiewski, P. Ostasiewicz, K. Du, D. F. Zieliska, F. Gnad, and M. Mann. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol Syst Biol*, 8:611, 2012.

[145] K. A. Phillips, S. Van Bebber, and A. M. Issa. Diagnostics and biomarker development: priming the pipeline. *Nat Rev Drug Discov*, 5(6):463–469, Jun 2006.

[146] E. P. Diamandis. Cancer biomarkers: can we turn recent failures into success? *J Natl Cancer Inst*, 102(19):1462–1467, Oct 2010.

[147] S. E. Kern. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. *Cancer Res*, 72(23):6097–6101, Dec 2012.

[148] E. P. Diamandis. The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? *BMC Med*, 10:87, 2012.

[149] A. Rauch, M. Bellew, J. Eng, M. Fitzgibbon, T. Holzman, P. Hussey, M. Igra, B. Maclean, C. W. Lin, A. Detter, R. Fang, V. Faca, P. Gafken, H. Zhang, J. Whiteaker, J. Whitaker, D. States, S. Hanash, A. Paulovich, and M. W. McIntosh. Computational proteomics analysis system (cpas): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J Proteome Res*, 5(1):112–121, Jan 2006.

[150] O. Rinner, L. N. Mueller, M. Hublek, M. Mller, M. Gstaiger, and R. Aebersold. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nat Biotechnol*, 25(3):345–352, Mar 2007.

[151] S. Park, J. D. Venable, T. Xu, and J. R. Yates, 3rd. A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat Methods*, 5(4):319–322, Apr 2008.

[152] M. Brusniak, B. Bodenmiller, D. Campbell, K. Cooke, J. Eddes, A. Garbutt, H. Lau, S. Letarte, L. N. Mueller, V. Sharma, O. Vitek, N. Zhang, R. Aebersold, and J. D. Watts. Corra: Computational framework and tools for lc-ms discovery and targeted mass spectrometry-based proteomics. *BMC Bioinformatics*, 9:542, 2008.

[153] D. May, W. Law, M. Fitzgibbon, Q. Fang, and M. McIntosh. Software platform for rapidly creating computational tools for mass spectrometry-based proteomics. *J Proteome Res*, 8(6):3212–3217, Jun 2009.

[154] J. Oh, S. Pan, J. Zhang, and J. Gao. Msq: a tool for quantification of proteomics data generated by a liquid chromatography/matrix-assisted laser desorption/ionization time-of-flight tandem mass spectrometry based targeted quantitative proteomics platform. *Rapid Commun Mass Spectrom*, 24(4):403–408, Feb 2010.

[155] P. Mortensen, J. W. Gouw, J. V. Olsen, S. Ong, K. T. G. Rigbolt, J. Bunkenborg, J. Cox, L. J. Foster, A. J. R. Heck, B. Blagoev, J. S. Andersen, and M. Mann. Msquant, an open source platform for mass spectrometry-based quantitative proteomics. *J Proteome Res*, 9(1):393–403, Jan 2010.

[156] R. G. Sadygov, D. Cociorva, and J. R. Yates, 3rd. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods*, 1(3):195–202, Dec 2004.

[157] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–214, Mar 2007.

[158] H. Chipman, T. Hastie, and R. Tibshirani. *"Clustering Microarray Data" Statistical Analysis of Gene Expression Microarray Data.* CRC press, 2003.

[159] B. Meunier, E. Dumas, I. Piec, D. Bchet, M. Hbraud, and J. Hocquette. Assessment of hierarchical clustering methodologies for proteomic data mining. *J Proteome Res*, 6(1):358–366, Jan 2007.

[160] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, 2001.

[161] G. Forman, I. Guyon, and A. Elisseeff. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.

[162] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422, March 2002.

[163] B. Boser, I. Guyon, and V. Vapnik. An training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

[164] V.N. Vapnik. *Statistical learning theory.* Wiley Interscience, 1998.

[165] J. R. Prensner, M. A. Rubin, J. T. Wei, and A. M. Chinnaiyan. Beyond psa: the next generation of prostate cancer biomarkers. *Sci Transl Med*, 4(127):127rv3, Mar 2012.

[166] P.C. Gtzsche and M. Nielsen. Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews 2011*, 2011.

[167] P. L. Fitzgibbons. Atypical lobular hyperplasia of the breast: a study of pathologists' responses in the college of american pathologists performance improvement program in surgical pathology. *Arch Pathol Lab Med*, 124(3):463–464, Mar 2000.

[168] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones. The disopred server for the prediction of protein disorder. *Bioinformatics*, 20(13):2138–9, 2004.

[169] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202, 1999.

[170] Team R Development Core. R: A language and environment for statistical computing. 2010.

[171] P. Romero, Z. Obradovic, and A K. Dunker. Natively disordered proteins: functions and predictions. *Appl Bioinformatics*, 3(2-3):105–113, 2004.

[172] A. K. Dunker, C. J. Oldfield, J. Meng, P. Romero, J. Y. Yang, J. W. Chen, V. Vacic, Z. Obradovic, and V. N. Uversky. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, 9 Suppl 2:S1, 2008.

[173] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic. Flavors of protein disorder. *Proteins*, 52(4):573–584, Sep 2003.

[174] R. Adamczak, A. Porollo, and J. Meller. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, 56(4):753–767, Sep 2004.

[175] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J A. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P. D. Thomas, C. H. Wu, C. Yeats, and S. Yong. Interpro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*, 40(Database issue):D306–D312, Jan 2012.

[176] P. Lahiry, A. Torkamani, N. J. Schork, and R. A. Hegele. Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nat Rev Genet*, 11(1):60–74, Jan 2010.

[177] W. A. Lim and T. Pawson. Phosphotyrosine signaling: evolving a new cellular communication system. *Cell*, 142(5):661–667, Sep 2010.

[178] B. A. Liu, B. W. Engelmann, and P. D. Nash. The language of sh2 domain interactions defines phosphotyrosine-mediated signal transduction. *FEBS Lett*, 586(17):2597–2605, Aug 2012.

[179] A. A. Bogan and K. S. Thorn. Anatomy of hot spots in protein interfaces. *J Mol Biol*, 280(1):1–9, Jul 1998.

[180] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, and P. Bork. eggnog v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic acids research*, 40(Database issue):D284–9, 2012.

[181] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18 Suppl 1:S71–7, 2002.

[182] A. Schlessinger, J. Liu, and B. Rost. Natively unstructured loops differ from other loops. *PLoS computational biology*, 3(7):e140, 2007.

[183] J. Bellay, S. Han, M. Michaut, T. Kim, M. Costanzo, B. J. Andrews, C. Boone, G. D. Bader, C. L. Myers, and P. M. Kim. Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol*, 12(2):R14, 2011.

[184] O. Keskin, B. Ma, and R. Nussinov. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol*, 345(5):1281–1294, Feb 2005.

[185] O. N. A. Demerdash, M. D. Daily, and J. C. Mitchell. Structure-based predictive models for allosteric hot spots. *PLoS Comput Biol*, 5(10):e1000531, Oct 2009.

[186] C. E. McCoy, D. G. Campbell, M. Deak, G. B. Bloomberg, and J. S. C. Arthur. Msk1 activity is controlled by multiple phosphorylation sites. *Biochem J*, 387(Pt 2):507–517, Apr 2005.

[187] G. Wang, X. Tong, S. Weng, and H. Zhou. Multiple phosphorylation of rad9 by cdk is required for dna damage checkpoint activation. *Cell Cycle*, 11(20):3792–3800, Oct 2012.

[188] A. Van Hooser, D. W. Goodrich, C. D. Allis, B. R. Brinkley, and M. A. Mancini. Histone h3 phosphorylation is required for the initiation, but not maintenance, of mammalian chromosome condensation. *J Cell Sci*, 111 (Pt 23):3497–3506, Dec 1998.

[189] M. Jackman, C. Lindon, E. A. Nigg, and J. Pines. Active cyclin b1-cdk1 first appears on centrosomes in prophase. *Nat Cell Biol*, 5(2):143–148, Feb 2003.

[190] N. Watanabe, H. Arai, Y. Nishihara, M. Taniguchi, N. Watanabe, T. Hunter, and H. Osada. M-phase kinases induce phospho-dependent ubiquitination of somatic wee1 by scfbeta-trcp. *Proc Natl Acad Sci U S A*, 101(13):4419–4424, Mar 2004.

[191] C. I. Holmberg, S. E. F. Tran, J. E. Eriksson, and L. Sistonen. Multisite phosphorylation provides sophisticated regulation of transcription factors. *Trends Biochem Sci*, 27(12):619–627, Dec 2002.

[192] S. Sdelci, M. Schtz, R. Pinyol, M. T. Bertran, L. Regu, C. Caelles, I. Vernos, and J. Roig. Nek9 phosphorylation of nedd1/gcp-wd contributes to plk1 control of -tubulin recruitment to the mitotic centrosome. *Curr Biol*, 22(16):1516–1523, Aug 2012.

[193] A. Gingras, S. P. Gygi, B. Raught, R. D. Polakiewicz, R.t T. Abraham, M. F. Hoekstra, R. Aebersold, and N. Sonenberg. Regulation of 4e-bp1 phosphorylation: a novel two-step mechanism. *Genes Dev*, 13(11):1422–1437, Jun 1999.

[194] R. Ben-Levy, I. A. Leighton, Y. N. Doza, P. Attwood, N. Morrice, C.r J. Marshall, and P. Cohen. Identification of novel phosphorylation sites required for activation of mapkap kinase-2. *EMBO J*, 14(23):5920–5930, Dec 1995.

[195] S. Liokatis, A. Sttzer, S. J. Elssser, F. Theillet, R. Klingberg, B. van Rossum, D. Schwarzer, C. D. Allis, W. Fischle, and P. Selenko. Phosphorylation of histone h3 ser10 establishes a hierarchy for subsequent intramolecular modification events. *Nat Struct Mol Biol*, 19(8):819–823, Aug 2012.

[196] V. C. Chen, J. W. Gouw, C. C. Naus, and L. J. Foster. Connexin multi-site phosphorylation: mass spectrometry-based proteomics fills the gap. *Biochim Biophys Acta*, 1828(1):23–34, Jan 2013.

[197] E. S. Knudsen and J. Y. Wang. Differential regulation of retinoblastoma protein function by specific cdk phosphorylation sites. *J Biol Chem*, 271(14):8313–8320, Apr 1996.

[198] M. Thomson and J. Gunawardena. Unlimited multistability in multisite phosphorylation systems. *Nature*, 460(7252):274–277, Jul 2009.

[199] J. M. Lizcano, N. Morrice, and P. Cohen. Regulation of bad by camp-dependent protein kinase is mediated via phosphorylation of a novel site, ser155. *Biochem J*, 349(Pt 2):547–557, Jul 2000.

[200] B. Raught and A. C. Gingras. eif4e activity is regulated at multiple levels. *Int J Biochem Cell Biol*, 31(1):43–57, Jan 1999.

[201] A. Gingras, B. Raught, and N. Sonenberg. Regulation of translation initiation by frap/mtor. *Genes Dev*, 15(7):807–826, Apr 2001.

[202] S. M. Varedi K, A. C. Ventura, S. D. Merajver, and X. N. Lin. Multisite phosphorylation provides an effective and flexible mechanism for switch-like protein degradation. *PLoS One*, 5(12):e14029, 2010.

[203] 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(6863):514–21, 2001.

[204] C. Salazar and T. Hfer. Allosteric regulation of the transcription factor nfat1 by multiple phosphorylation sites: a mathematical analysis. *J Mol Biol*, 327(1):31–45, Mar 2003.

[205] S. Legewie, N. Blthgen, R. Schfer, and H. Herzel. Ultrasensitization: switch-like regulation of cellular signaling by transcriptional induction. *PLoS Comput Biol*, 1(5):e54, Oct 2005.

[206] J. Gunawardena. Multisite protein phosphorylation makes a good threshold but can be a poor switch. *Proceedings of the National Academy of Sciences of the United States of America*, 102(41):14617–22, 2005.

[207] J. W. Harbour, R. X. Luo, A. Dei Santi, A. A. Postigo, and D. C. Dean. Cdk phosphorylation triggers sequential intramolecular interactions that progressively block rb functions as cells move through g1. *Cell*, 98(6):859–869, Sep 1999.

[208] D. Vuzman, Y. Hoffman, and Y. Levy. Modulating protein-dna interactions by post-translational modifications at disordered regions. *Pac Symp Biocomput*, 17:188–199, 2012.

[209] P. Kumar, M. S. Chimenti, H. Pemble, A. Schnichen, O. Thompson, M. P. Jacobson, and T. Wittmann. Multisite phosphorylation disrupts arginine-glutamate salt bridge networks required for binding of cytoplasmic linker-associated protein 2 (clasp2) to end-binding protein 1 (eb1). *J Biol Chem*, 287(21):17050–17064, May 2012.

[210] D. O. Cowley and B. J. Graves. Phosphorylation represses ets-1 dna binding by reinforcing autoinhibition. *Genes Dev*, 14(3):366–376, Feb 2000.

[211] R. L. Daugherty and C. J. Gottardi. Phospho-regulation of beta-catenin adhesion and signaling functions. *Physiology (Bethesda)*, 22:303–309, Oct 2007.

[212] T. Hunter. The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Mol Cell*, 28(5):730–738, Dec 2007.

[213] B. Camuzeaux, J. Diring, P. Hamard, M. Oulad-Abdelghani, M. Donzeau, M. Vigneron, C. Kedinger, and B. Chatton. p382-mediated phosphorylation and sumoylation of atf7 are mutually exclusive. *Journal of molecular biology*, 384:980–991, 2008.

[214] R. D. Mohan, D. W. Litchfield, J. Torchia, and M. Tini. Opposing regulatory roles of phosphorylation and acetylation in dna mispair processing by thymine dna glycosylase. *Nucleic Acids Res*, 38(4):1135–1148, Mar 2010.

[215] W. S. Lo, L. Duggan, N. C. Emre, R. Belotserkovskya, W. S. Lane, R. Shiekhattar, and S. L. Berger. Snf1–a histone kinase that works in concert with the histone acetyltransferase gcn5 to regulate transcription. *Science*, 293(5532):1142–1146, Aug 2001.

[216] D.I. Lin, O. Barbash, K.G. Kumar, J.D. Weber, J.W. Harper, A.J. Klein-Szanto, A. Rustgi, Fuchs S.Y., and J.A. Diehl. Phosphorylation-dependent ubiquitination of cyclin d1 by the scf(fbx4-alphab crystallin) complex. *Mol Cell*, 24(3):355–366, Nov 2006.

[217] J. Kang, C. B. Gocke, and H. Yu. Phosphorylation-facilitated sumoylation of mef2c negatively regulates its transcriptional activity. *BMC Biochem*, 7:5, 2006.

[218] R. Nussinov, C. Tsai, F. Xin, and P. Radivojac. Allosteric post-translational modification codes. *Trends Biochem Sci*, 37(10):447–455, Oct 2012.

[219] P. Creixell and R. Linding. Cells, shared memory and breaking the ptm code. *Mol Syst Biol*, 8:598, 2012.

[220] P. Minguez, L. Parca, F. Diella, D. R. Mende, R. Kumar, M. Helmer-Citterich, A. Gavin, V. van Noort, and P. Bork. Deciphering a global network of functionally associated post-translational modifications. *Mol Syst Biol*, 8:599, 2012.

[221] V. van Noort, J. Seebacher, S. Bader, S. Mohammed, I. Vonkova, M. J. Betts, S. Khner, R. Kumar, T. Maier, M. O'Flaherty, V. Rybin, A. Schmeisky, E. Yus, J. Stlke, L. Serrano, R.t B. Russell, A. J. R. Heck, P. Bork, and A. Gavin. Crosstalk between phosphorylation and lysine acetylation in a genome-reduced bacterium. *Mol Syst Biol*, 8:571, 2012.

[222] E. Meiler, E. Nieto-Pelegrn, and N. Martinez-Quiles. Cortactin tyrosine phosphorylation promotes its deacetylation and inhibits cell spreading. *PLoS One*, 7(3):e33662, 2012.

[223] W. T. Ruyechan and J. W. Olson. Surface lysine and tyrosine residues are required for interaction of the major herpes simplex virus type 1 dna-binding protein with single-stranded dna. *J Virol*, 66(11):6273–6279, Nov 1992.

[224] A. Michalski, E. Damoc, J. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann, and S. Horning. Mass spectrometry-based proteomics using q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Mol Cell Proteomics*, 10(9):M111.011015, Sep 2011.

[225] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[226] C. M. Gould, F. Diella, A. Via, P. Puntervoll, C. Gemund, S. Chabanis-Davidson, S. Michael, A. Sayadi, J. C. Bryne, C. Chica, M. Seiler, N. E. Davey, N. Haslam, R. J. Weatheritt, A. Budd, T. Hughes, J. Pas, L. Rychlewski, G. Trave, R. Aasland, M. Helmer-Citterich, R. Linding, and T. J. Gibson. Elm: the status of the 2010 eukaryotic linear motif resource. *Nucleic acids research*, 38(Database issue):D167–80, 2010.

[227] J. H. Fong, B. A. Shoemaker, S. O. Garbuzynskiy, M. Y. Lobanov, O. V. Galzitskaya, and A. R. Panchenko. Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis. *PLoS computational biology*, 5(3):e1000316, 2009.

[228] D. Plewczynski, A. Tkacz, A. Godzik, and L. Rychlewski. A support vector machine approach to the identification of phosphorylation sites. *Cellular & molecular biology letters*, 10(1):73–89, 2005.

[229] E. E. Metcalfe, N. J. Traaseth, and G. Veglia. Serine 16 phosphorylation induces an order-to-disorder transition in monomeric phospholamban. *Biochemistry*, 44(11):4386–96, 2005.

[230] L. M. Espinoza-Fonseca, D. Kast, and D. D. Thomas. Thermodynamic and structural basis of phosphorylation-induced disorder-to-order transition in the regulatory light chain of smooth muscle myosin. *Journal of American Chemistry Society*, 130:1220812209, 2008.

[231] S. Tait, K. Dutta, D. Cowburn, J. Warwicker, A. J. Doig, and J. E. McCarthy. Local control of a disorder-order transition in 4e-bp1 underpins regulation of translation via eif4e. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41):17627–32, 2010.

[232] E. S. Groban, A. Narayanan, and M. P. Jacobson. Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PLoS computational biology*, 2(4):e32, 2006.

[233] V. Vacic, L. M. Iakoucheva, and P. Radivojac. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, 22(12):1536–7, 2006.

[234] V. Vacic, V. N. Uversky, A. K. Dunker, and S. Lonardi. Composition profiler: a tool for discovery and visualization of amino acid composition differences. *BMC bioinformatics*, 8:211, 2007.

[235] J. Cox and M. Mann. 1d and 2d annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics*, 13(Suppl 16):S12, 2012.

[236] T. S. Prasad, K. Kandasamy, and A. Pandey. Human protein reference database and human proteinpedia as discovery tools for systems biology. *Methods in molecular biology*, 577:67–79, 2009.

[237] F. Watanabe, K. Shinohara, H. Teraoka, K. Komatsu, K. Tatsumi, F. Suzuki, T. Imai, M. Sagara, H. Tsuji, and T. Ogiu. Involvement of dna-dependent protein kinase in down-regulation of cell cycle progression. *The international journal of biochemistry & cell biology*, 35(4):432–40, 2003.

[238] H. C. Reinhardt, A. S. Aslanian, J. A. Lees, and M. B. Yaffe. p53-deficient cells rely on atm- and atr-mediated checkpoint signaling through the p38mapk/mk2 pathway for survival after dna damage. *Cancer cell*, 11(2):175–89, 2007.

[239] J. Liu, H. Tan, and B. Rost. Loopy proteins appear conserved in evolution. *Journal of molecular biology*, 322(1):53–64, 2002.

[240] J. L. Smart and J. A. McCammon. Phosphorylation stabilizes the n-termini of alpha-helices. *Biopolymers*, 49(3):225–33, 1999.

[241] C. D. Andrew, J. Warwicker, G. R. Jones, and A. J. Doig. Effect of phosphorylation on alpha-helix stability as a function of position. *Biochemistry*, 41(6):1897–905, 2002.

[242] N. Errington and A. J. Doig. A phosphoserine-lysine salt bridge within an alpha-helical peptide, the strongest alpha-helix side-chain interaction measured to date. *Biochemistry*, 44(20):7553–8, 2005.

[243] L. N. Johnson and M. O'Reilly. Control by phosphorylation. *Current opinion in structural biology*, 6(6):762–9, 1996.

[244] A. J. Riemen and M. L. Waters. Controlling peptide folding with repulsive interactions between phosphorylated amino acids and tryptophan. *Journal of the American Chemical Society*, 131(39):14081–7, 2009.

[245] C. A. Galea, Y. Wang, S. G. Sivakolundu, and R. W. Kriwacki. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry*, 47(29):7598–609, 2008.

[246] L. M. Espinoza-Fonseca. Dynamic optimization of signal transduction via intrinsic disorder. *Molecular bioSystems*, 8(1):194–7, 2012.

[247] J. R. Burke, G. L. Hura, and S. M. Rubin. Structures of inactive retinoblastoma protein reveal multiple mechanisms for cell cycle control. *Genes and development*, 26(11):1156–66, 2012.

[248] R. Schweiger and M. Linial. Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. *Biology direct*, 5(1):6, 2010.

[249] 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(5948):1682–6, 2009.

[250] M. Koivomagi, E. Valk, R. Venta, A. Iofik, M. Lepiku, E. R. Balog, S. M. Rubin, D. O. Morgan, and M. Loog. Cascades of multisite phosphorylation control sic1 destruction at the onset of s phase. *Nature*, 480(7375):128–31, 2011.

[251] A. H. Andreotti. Native state proline isomerization: an intrinsic molecular switch. *Biochemistry*, 42(32):9515–24, 2003.

[252] X. Z. Zhou, O. Kops, A. Werner, P. J. Lu, M. Shen, G. Stoller, G. Kullertz, M. Stark, G. Fischer, and K. P. Lu. Pin1-dependent prolyl isomerization regulates dephosphorylation of cdc25c and tau proteins. *Molecular cell*, 6(4):873–83, 2000.

[253] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10:48, 2009.

[254] T. Geiger, A. Wehner, C. Schaab, J.n Cox, and M. Mann. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics*, 11(3):M111.014050, Mar 2012.

[255] A. Vlahou, J. O. Schorge, B. W. Gregory, and R. L. Coleman. Diagnosis of ovarian cancer using decision tree classification of mass spectral data. *J Biomed Biotechnol*, 2003(5):308–314, 2003.

[256] G. Ge and G W. Wong. Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics*, 9:275, 2008.

[257] J. Yu and X. Chen. Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data. *Bioinformatics*, 21 Suppl 1:i487–i494, Jun 2005.

[258] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.

[259] I. Guyon. Practical feature selection: from correlation to causality. In *Mining Massive Data Sets for Security*. IOS Press, 2008.

[260] V. N. Vapnik. An overview of statistical learning theory. *IEEE Trans Neural Netw*, 10(5):988–999, 1999.

[261] N. Cristianini and J. Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA, 2000.

[262] Y. Saeys, I. Inza, and P. Larraaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, Oct 2007.

[263] C. Chang and C. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[264] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 1999.

[265] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, Oct 2000.

[266] J. R. Winiewski, P. Ostasiewicz, and M. Mann. High recovery fasp applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. *J Proteome Res*, 10(7):3040–3049, Jul 2011.

[267] P. D. Baade, D. R. Youlden, and L. J. Krnjacki. International epidemiology of prostate cancer: geographical distribution and secular trends. *Mol Nutr Food Res*, 53(2):171–184, Feb 2009.

[268] F. H. Schrder. Review of diagnostic markers for prostate cancer. *Recent Results Cancer Res*, 181:173–182, 2009.

[269] M. K. Brawer, M. P. Chetner, J. Beatie, D. M. Buchner, R. L. Vessella, and P. H. Lange. Screening for prostatic carcinoma with prostate specific antigen. *J Urol*, 147(3 Pt 2):841–845, Mar 1992.

[270] C. M. Perou, T. Srlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lnning, A. L. Brresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, Aug 2000.

[271] T. Srlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lnning, and A. L. Brresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–10874, Sep 2001.

[272] P. K. Morrow, G. M. Wulf, J. Ensor, D. J. Booser, J. A. Moore, P. R. Flores, Y. Xiong, S. Zhang, I. E. Krop, E. P. Winer, D. W. Kindelberger, J. Coviello, A. A. Sahin, R. Nuez, G. N. Hortobagyi, D. Yu, and F. J. Esteva. Phase i/ii study of trastuzumab in combination with everolimus (rad001) in patients with her2-overexpressing metastatic breast cancer who progressed on trastuzumab-based therapy. *J Clin Oncol*, 29(23):3126–3132, Aug 2011.

[273] N. L. Spector and K. L. Blackwell. Understanding the mechanisms behind trastuzumab therapy for human epidermal growth factor receptor 2-positive breast cancer. *J Clin Oncol*, 27(34):5838–5847, Dec 2009.

[274] W. D. Foulkes, I. M. Stefansson, P. O. Chappuis, L. R. Bgin, J. R. Goffin, N. Wong, M. Trudel, and L. A. Akslen. Germline brca1 mutations and a basal epithelial phenotype in breast cancer. *J Natl Cancer Inst*, 95(19):1482–1485, Oct 2003.

[275] S. Bayraktar and S. Glck. Molecularly targeted therapies for metastatic triple-negative breast cancer. *Breast Cancer Res Treat*, Jan 2013.

[276] U.S. Preventive Services Task Force. Screening for prostate cancer: recommendation statement. *Am Fam Physician*, 87(4):Online, Feb 2013.

[277] J. S. Ross, E. A. Slodkowska, W. F. Symmans, L. Pusztai, P. M. Ravdin, and G. N. Hortobagyi. The her-2 receptor and breast cancer: ten years of targeted anti-her-2 therapy and personalized medicine. *Oncologist*, 14(4):320–368, Apr 2009.

[278] R. Schillaci, P. Guzmn, F. Cayrol, W. Beguelin, M. C. Daz Flaqu, C. J. Proietti, V. Pineda, J. Palazzi, I. Frahm, E. H. Charreau, E. Maronna, J. C. Roa, and P. V. Elizalde. Clinical relevance of erbb-2/her2 nuclear expression in breast cancer. *BMC Cancer*, 12:74, 2012.

[279] F. Henjes, S. Bender, C .and von der Heyde, H. A. Braun, L .and Mannsperger, C. Schmidt, S. Wiemann, M. Hasmann, S. Aulmann, T. Beissbarth, and U. Korf. Strong egfr signaling in cell line models of erbb2-amplified breast cancer attenuates response towards erbb2-targeting drugs. *Oncogenesis*, 1(7):e16, 2012.

[280] D. A. Thompson and R. J. Weigel. hag-2, the human homologue of the xenopus laevis cement gland gene xag-2, is coexpressed with estrogen receptor in breast cancer cell lines. *Biochem Biophys Res Commun*, 251(1):111–116, Oct 1998.

[281] D. Liu, P. S. Rudland, D. R. Sibson, A. Platt-Higgins, and R. Barraclough. Human homologue of cement gland protein, a novel metastasis inducer associated with breast carcinomas. *Cancer Res*, 65(9):3796–3805, May 2005.

[282] F. R. Fritzsche, E. Dahl, S. Pahl, M. Burkhardt, J. Luo, E. Mayordomo, T. Gansukh, A. Dankof, R. Knuechel, C. Denkert, K. Winzer, M. Dietel, and G. Kristiansen. Prognostic relevance of agr2 expression in breast cancer. *Clin Cancer Res*, 12(6):1728–1734, Mar 2006.

[283] R. Cowper-Sal lari, X. Zhang, J. B. Wright, S. D. Bailey, M. D. Cole, J. Eeckhoute, J. H. Moore, and M. Lupien. Breast cancer risk-associated snps modulate the affinity of chromatin for foxa1 and alter gene expression. *Nat Genet*, 44(11):1191–1198, Nov 2012.

[284] O. Giricz, V. Calvo, S. C. Pero, D. N. Krag, J. A. Sparano, and P. A. Kenny. Grb7 is required for triple-negative breast cancer cell invasion and survival. *Breast Cancer Res Treat*, 133(2):607–615, Jun 2012.

[285] Y. Nadler, A. M. Gonzlez, R. L. Camp, D. L. Rimm, H. M. Kluger, and Y. Kluger. Growth factor receptor-bound protein-7 (grb7) as a prognostic marker and therapeutic target in breast cancer. *Ann Oncol*, 21(3):466–473, Mar 2010.

[286] B. Ramsey, T. Bai, A. Hanlon Newell, M. Troxell, B. Park, S. Olson, E. Keenan, and S. Luoh. Grb7 protein over-expression and clinical outcome in breast cancer. *Breast Cancer Res Treat*, 127(3):659–669, Jun 2011.

[287] P. M. Dimberu and R. M. Leonhardt. Cancer immunotherapy takes a multi-faceted approach to kick the immune system into gear. *Yale J Biol Med*, 84(4):371–380, Dec 2011.

[288] A. Brychzy, T. Rein, K. F. Winklhofer, F. U. Hartl, J. C. Young, and W. M. J. Obermann. Cofactor tpr2 combines two tpr domains and a j domain to regulate the hsp70/hsp90 chaperone system. *EMBO J*, 22(14):3613–3623, Jul 2003.

[289] N. S. C. Moffatt, E. Bruinsma, C. Uhl, W. M. J. Obermann, and D. Toft. Role of the cochaperone tpr2 in hsp90 chaperoning. *Biochemistry*, 47(31):8203–8213, Aug 2008.

[290] H. Bonkhoff, T. Fixemer, I. Hunsicker, and K. Remberger. Progesterone receptor expression in human prostate cancer: correlation with tumor progression. *Prostate*, 48(4):285–291, Sep 2001.

[291] J. S. Sebolt-Leopold and R. Herrera. Targeting the mitogen-activated protein kinase cascade to treat cancer. *Nat Rev Cancer*, 4(12):937–947, Dec 2004.

[292] G. Rodrguez-Berriguete, B. Fraile, P. Martnez-Onsurbe, G. Olmedilla, R. Paniagua, and M. Royuela. Map kinases and prostate cancer. *J Signal Transduct*, 2012:169170, 2012.

[293] R. E. Bakin, D. Gioeli, R. A. Sikes, E. A. Bissonette, and M. J. Weber. Constitutive activation of the ras/mitogen-activated protein kinase signaling pathway promotes androgen hypersensitivity in lncap prostate cancer cells. *Cancer Res*, 63(8):1981–1989, Apr 2003.

[294] Y. Hwang, T. Lu, D. Huang, Yu. Kuo, C. Kao, N. Yeh, H. Wu, and C. Lin. Nolc1, an enhancer of nasopharyngeal carcinoma progression, is essential for tp53 to regulate mdm2 expression. *Am J Pathol*, 175(1):342–354, Jul 2009.

[295] X. Gao, Q. Wang, W. Li, B. Yang, H. Song, W. Ju, S. Liu, and J. Cheng. Identification of nucleolar and coiled-body phosphoprotein 1 (nolc1) minimal promoter regulated by nf-b and creb. *BMB Rep*, 44(1):70–75, Jan 2011.

[296] M. Karin. Nuclear factor-kappab in cancer development and progression. *Nature*, 441(7092):431–436, May 2006.

[297] S. Huang, C. A. Pettaway, H. Uehara, C. D. Bucana, and I. J. Fidler. Blockade of nf-kappab activity in human prostate cancer cells is associated with suppression of angiogenesis, invasion, and metastasis. *Oncogene*, 20(31):4188–4197, Jul 2001.

[298] C. D. Chen and C. L. Sawyers. Nf-kappa b activates prostate-specific antigen expression and is upregulated in androgen-independent prostate cancer. *Mol Cell Biol*, 22(8):2862–2870, Apr 2002.

[299] Q. Zhang, B. T. Helfand, T. L. Jang, L. J. Zhu, L. Chen, X. J. Yang, J. Kozlowski, N. Smith, S. D. Kundu, G. Yang, A. A. Raji, B. Javonovic, M. Pins, P. Lindholm, Y. Guo, W. J. Catalona, and C. Lee. Nuclear factor-kappab-mediated transforming growth factor-beta-induced expression of vimentin is an independent predictor of biochemical recurrence after radical prostatectomy. *Clin Cancer Res*, 15(10):3557–3567, May 2009.

[300] I. A. Voutsadakis, P. J. Vlachostergios, D. D. Daliani, F. Karasavvidou, G. Kakkas, G. Moutzouris, M. D. Melekos, and C. N. Papandreou. Cd10 is inversely associated with nuclear factor-kappa b and predicts biochemical recurrence after radical prostatectomy. *Urol Int*, 88(2):158–164, 2012.

[301] Y. Ishihama, T. Sato, T. Tabata, N. Miyamoto, K. Sagane, T. Nagasu, and Y. Oda. Quantitative mouse brain proteomics using culture-derived isotope tags as internal standards. *Nat Biotechnol*, 23(5):617–621, May 2005.

[302] T. Geiger, J. R. Wisniewski, J. Cox, S. Zanivan, M. Kruger, Y. Ishihama, and M. Mann. Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nat Protoc*, 6(2):147–157, Feb 2011.

[303] E. A. Nigg. Mitotic kinases as regulators of cell division and its checkpoints. *Nature reviews. Molecular cell biology*, 2(1):21–32, 2001.

[304] S. S. Thakur, T. Geiger, B. Chatterjee, P. Bandilla, F. Frhlich, J. Cox, and M. Mann. Deep and highly sensitive proteome coverage by lc-ms/ms without prefractionation. *Mol Cell Proteomics*, 10(8):M110.003699, Aug 2011.

[305] A. Arora and E. M. Scholar. Role of tyrosine kinase inhibitors in cancer therapy. *J Pharmacol Exp Ther*, 315(3):971–979, Dec 2005.

[306] A. Marusyk and K. Polyak. Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta*, 1805(1):105–117, Jan 2010.

Appendix

Table 6.1: *Tendency of phosphorylation sites, predicted in disordered regions, to occur outside Interpro domains.*

		Phospho	Reference	Odds ratio	P-value
S	outsideDomains	17947	104779	1.037986	0.004705
	withinDomains	10718	64951		
T	outsideDomains	4186	52072	1.104311	0.0001264
	withinDomains	2610	35853		
Y	outsideDomains	484	14054	0.9700555	0.6697
	withinDomains	357	10056		

Preference of phospho-sites to appear in regions connecting Interpro domains. Contingency tables containing the counts of phospho-sites and non-modified reference sites found within and outside Interpro domain regions. Only sites predicted to lie within disordered regions were used. Fisher exact tests were computed for each phospho-acceptor residue separately. Modified sites were significantly overrepresented in regions outside Interpro domains.

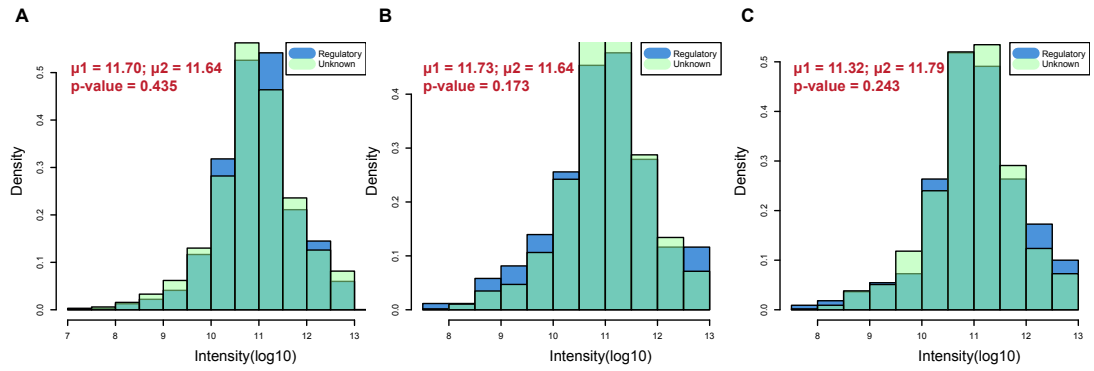


Figure 6.1: ***Regulatory phospho-sites and protein intensity.*** *Protein intensities of phospho-sites in* ***disordered regions*** *with regulatory (blue) and unknown (light green) functions are compared:* ***A)*** *serine,* ***B)*** *threonine and* ***C)*** *tyrosine. The overlap between the two distributions is shown in dark green. The corresponding group means of intensity (regulatory $\mu 1$ and unknown $\mu 2$ respectively) and the p-values computed with the Wilcoxon test are shown for each residue.*

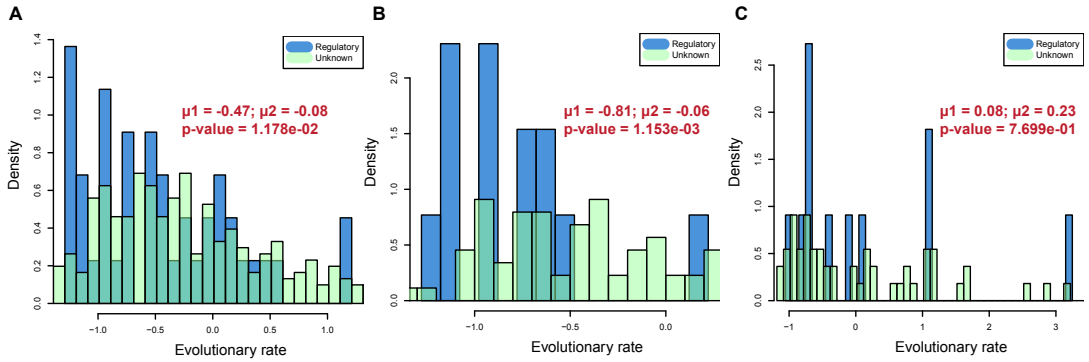


Figure 6.2: ***Evolutionary conservation of regulatory phospho-sites.*** *Evolutionary rates of phospho-sites (predicted in ordered regions) with regulatory (blue) and unknown (light green) functions are compared:* ***A)*** *serine,* ***B)*** *threonine and* ***C)*** *tyrosine. The overlap between the two distributions is shown in dark green color. Evolutionary rate means of the compared distributions (regulatory $\mu 1$ and unknown $\mu 2$ respectively) and the corresponding p-values computed with the Wilcoxon test are presented for each residue. Note that lower values correspond to higher conservation.*

Table 6.2: *Preference of regulatory phosphorylation sites for disordered regions.*

		Regulatory	Unknown	Odds ratio	P-value
S	disordered	460	4655	0.7724114	0.03884
	ordered	87	680		
T	disordered	101	1249	1.340972	0.144
	ordered	41	233		
Y	disordered	43	180	0.8563674	0.6102
	ordered	36	129		

Preference of regulatory phosphorylation sites versus phosphorylation sites of unknown funciton to appear in disordered regions are shown by odds ratios and their significance (based on Fisher's exact test).

Table 6.3: *Preference of regulatory phosphorylation sites for secondary structure elements.*

		Regulatory	Unknown	Prop. phospho	Prop. reference	P-value
S	coil	523	5005	0.9561243	0.9381443	0.1119
	helix	17	275	0.03107861	0.05154639	0.04598
	sheet	7	55	0.01279707	0.01030928	0.7469
T	coil	139	1403	0.9788732	0.9466937	0.1409
	helix	3	45	0.02112676	0.03036437	0.7177
	sheet	0	34	0.00000000	0.02294197	0.1292
Y	coil	56	209	0.7088608	0.6763754	0.6757
	helix	7	65	0.08860759	0.21035599	0.02024
	sheet	16	35	0.2025316	0.1132686	0.05627

The preference of regulatory phosphorulation sites versus phosphorylation sites of unknown funciton to appear in coil regions are shown by the corresponding proportions in each structural group and the corresponding p-values (Proportions test).

Acknowledgements

The implementation and completion of this thesis would have not been possible without the help, encouragement and support by several people that are acknowledged here:

I would like to thank Prof. Dmitrij Frishman for all the support, pieces of advice and encouragement during the completion of this thesis. I am especially thankful for the scientific freedom and trust and the possibility to gain experience in various aspects of science.

I would like to express my deep gratitude to Prof. Matthias Mann for giving me the opportunity to work on exciting projects, for the tremendous help in paper writing, his generosity and the inspirational supervision that I have received.

Additionally, I want to thank Prof. Harald Luksch, who was so kind to agree to become the head of my defense committee.

Special thanks to Jürgen Cox for being an unlimited source of knowledge, ideas and always having good solutions and clever answers to the hardest problems and reviewers' questions.

Furthermore, I would like to express my gratitude to: Tami Geiger for the inspiring collaborations and fruitful discussions, Kirti Sharma for sharing her knowledge both on the beautiful world of signaling and the laws of life; Nadin, Michal and Rochelle for all the stimulating discussions and pleasant coffee breaks and lunches; Richard Sheltema for his constructive criticism, realistic pessimism and the geeky talks; Dirk Walther for critical reading and help with the German translation of parts of this thesis and the entire Mann department for the great working atmosphere, bringing professionalism, fun and friendship together.

I am very thankful that I had the chance to work with Martin Sturm, who helped me a lot during my first steps as a PhD student; Andre Jehl who introduced me in a fun way to the world of pathogens; Dmitry Suplatov for sharing his expertise in structural biology and offering vital help during my trips in Moscow.

Special thanks to Andrey Chursov for offering expert advice on machine learning and data analysis problems and patiently answering all my questions. I cherish very much our existential talks that also thought me a lot about management and prioritization of daily life tasks.

I am especially grateful to Claudia Luksch for always being so positive, helpful and dedicated to assisting students with all issues. I am obliged to her for helping me out with the translation of parts and all the bureaucracy associated with the submission of this thesis.

Last but not least, I say " a huge thank you" to Milena Makaveeva, Simon Leis, Daniel Cernea and my family who shared the ups and downs, never lost trust in me and keep filling my life with love, happiness and meaning.

Glossary/Abbreviations

AGR2	–	Anterior gradient 2 homolog
CID	–	Collision-Induced Dissociation
DNAJC7	–	DNAJ homolog subfamily C member 7
ER	–	Estrogen Receptor
FASP	–	Filter-aided Sample Preparation
FFPE	–	Formalin-fixed Parafin-embedded
FOXA1	–	Forkhead box protein A1
GO	–	Gene Ontology
GOBP	–	Gene Ontology Biological Process
GOMF	–	Gene Ontology Molecular Function
GRB7	–	Growth factor receptor-bound protein-7
Her2	–	Human epidermal growth factor receptor 2
HCD	–	Higher energy Collisional Dissociation
HPLC	–	High Performance Liquid Chromatography
HSP	–	Heat Shock Protein
MAPK3	–	Mitogen-activated protein kinase 3
NOLC1	–	Nucleolar and coiled-body phosphoprotein 1
PCA	–	Principal Component Analysis
PR	–	Progesterone Receptor
PSA	–	Prostate-Specific Antigene
PTMs	–	Post Translational Modifications
RFE	–	Recursive Feature Elimination
SILAC	–	Stable Isotope Labeling by Amino acid in Cell culture
SVMs	–	Support Vector Machines
TNBC	–	Triple Negative Breast Cancer