Comparative Analysis of Molecular Interaction Networks: The Interplay Between Spatial and Functional Organizing Principles.

Vergleichende Analyse molekularer Interaktionsnetzwerke: Der Zusammenhang von räumlichen und funktionellen Organisationsprinzipien.

Dissertation

zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat)

eingereicht am
Institut für Biochemie und Biologie an der
Mathematisch - Naturwissenschaftlichen Fakultät
Universität Potsdam

vorgelegt von

DIPL. ING. PAWEL DUREK

Die vorliegende Arbeit wurde angefertigt am
Max-Planck-Institut für Molekulare Pflanzenphysiologie
Potsdam, Dezember 2008

Published online at the
Institutional Repository of the University of Potsdam:
http://opus.kobv.de/ubp/volltexte/2009/3143/
urn:nbn:de:kobv:517-opus-31439
[http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-31439]

Contents

Publications

Parts of this thesis have been published in peer-reviewed journals. Chapter 3 contains an integrative analysis of metabolic and protein interaction networks, published in *BMC Systems Biology*, Chapter 6, which deals witch the spatial characterization and prediction of phosphorylation sites, has been submitted to *BMC Bioinformatics*, while Chapter 7 contains the publication of the *PhosPhat* database in *Nucleic Acids Research*. Parts of the original version of the latter publication, in which I have not been involved, have been removed from the chapter, but retaining consistency. Furthermore, results of Chapter 8 dealing with concordance of experimental results has been accepted for publication in a review of plant phosphoproteomics in *Proteomics* next year.

Abstract

The study of biological interaction networks is a central theme in systems biology. Here, we investigate common as well as differentiating principles of molecular interaction networks associated with different levels of molecular organization. They include metabolic pathway maps, protein-protein interaction networks as well as kinase interaction networks.

First, we present an integrated analysis of metabolic pathway maps and protein-protein interaction networks (PIN). It has long been established that successive enzymatic steps are often catalyzed by physically interacting proteins forming permanent or transient multi-enzyme complexes. Inspecting high-throughput PIN data, it has been shown recently that, indeed, enzymes involved in successive reactions are generally more likely to interact than other protein pairs. In this study, we expanded this line of research to include comparisons of the respective underlying network topologies as well as to investigate whether the spatial organization of enzyme interactions correlates with metabolic efficiency. Analyzing yeast data, we detected long-range correlations between shortest paths between proteins in both network types suggesting a mutual correspondence of both network architectures. We discovered that the organizing principles of physical interactions between metabolic enzymes differ from the general PIN of all proteins. While physical interactions between proteins are generally dissortative, enzyme interactions were observed to be assortative. Thus, enzymes frequently interact with other enzymes of similar rather than different degree. Enzymes carrying high flux loads are more likely to physically interact than enzymes with lower metabolic throughput. In particular, enzymes associated with catabolic pathways as well as enzymes involved in the biosynthesis of complex molecules were found to exhibit high degrees of physical clustering. Single proteins were identified that connect major components of the cellular metabolism and hence might be essential for the structural integrity of several biosynthetic systems.

Besides metabolic aspects of PINs, we investigated the characteristic topological properties of protein interactions involved in signaling and regulatory functions mediated by kinase interactions. Characteristic topological differences between PINs associated with metabolism, and those describing phosphorylation networks were revealed and shown to reflect the different modes of biological operation of both network types. From a closer inspection of phosphorylation networks, we concluded that phosphorylation of kinases by other kinases primarily serves to transduce signals to the ultimate target protein with a particular effector function and in a directed fashion by way of forming kinase cascades rather than to offer regulatory capacities, for example via feedback loops. Instead, regulation was found predominantly at the level of the target protein itself, whose activity appears to be frequently modulated by several incoming kinases.

The construction of phosphorylation networks is based on the identification of specific kinase-target relations including the determination of the actual phosphorylation sites (P-sites). The computational prediction of P-sites as well as the identification of involved kinases still suffers from insufficient accuracies and specificities of the underlying prediction algorithms, and the experimental identification in a genome-scale manner is not (yet) doable. Computational prediction methods have focused primarily on extracting predictive features from the local, one-dimensional sequence information surrounding P-sites. However the recognition of such motifs by the respective kinases is a spatial event. Therefore, we characterized the spatial distributions of amino acid residue types around P-sites and extracted signature 3D-profiles. We then tested the added value of spatial information on the prediction performance. When compared to sequence-only based predictors, a consistent performance gain was obtained. The availability of reliable training data of experimentally determined P-sites is critical for the development of computational prediction methods. As part of this thesis, we provide an assessment of false-positive rates of phosphoproteomic data.

In addition, a sequence-based predictor of P-sites that implicitly captures 3D-structural aspects by including secondary and tertiary structure preferences, polarity, volume and solvent accessibility, as well as structural disorder indices was developed as part of a new database for plant-specific phosphorylation sites (*PhosPhAt*). The *PhosPhAt* database aims to present a genome-wide and consolidated view of all detected phosphorylation sites in all proteins encoded in the *Arabidopsis* genome. Augmented by a computational predictor, designed specifically to predict plant phosphorylation sites, it provides a valuable resource to the plant science community. The developed P-site predictor was applied to investigate the genome-scale distribution of functional annotations associated with proteins predicted to be phosphorylated. The predicted sites in conjunction with experimental phosphorylation sites stored in the database will provide a powerful basis for further in-depth analysis of phosphorylation motifs in orthologous and paralogous proteins from different organisms

Zusammenfassung

Ein zentrales Thema der Systembiologie ist die Untersuchung biologischer Interaktionsnetzwerke. In der vorliegenden Arbeit wurden gemeinsame sowie differenzierende Prinzipien molekularer Interaktionsnetzwerke untersucht, die sich durch unterschiedliche Ebenen der molekulareren Organisation auszeichnen. Zu den untersuchten Interaktionsnetzwerken gehörten Netzwerke, die auf metabolischen Wechselwirkungen, physikalischen Wechselwirkungen zwischen Proteinen und Kinase-Interaktionen aufbauen.

Zunächst wird eine integrativen Analyse der metabolischen Pfade und Protein Interaktionsnetzwerke vorgestellt. Es wird seit schon seit langem angenommen, dass aufeinander folgende enzymatische Schritte oft durch permanente oder transiente Multienzymkomplexe, die auf physikalischen Wechselwirkungen der involvierten Proteine basieren, katalysiert werden. Diese Annahme konnte durch die Auswertung von Ergebnissen aus Hochdurchsatz-Experimenten bestätigt werden. Demnach treten aufeinander folgende Enzyme häufiger in physikalische Wechselwirkung als zufällige Enzympaare. Die vorliegende Arbeit geht in ihrer Analyse weiter, in dem die Topologien der zugrundeliegenden Netzwerke, die auf metabolischen und physikalischen Wechselwirkungen basieren verglichen werden und der Zusammenhang zwischen der räumlichen Organisation der Enzyme und der metabolischen Effizienz gesucht wird. Ausgehend von Interaktionsdaten aus Hefe hat die Analyse der auf metabolischen und physikalischen Wechselwirkungen aufbauenden Interaktionswege eine weitgehende Korrelation der Distanzen aufgezeigt und somit eine wechselseitige Übereinstimmung der Architekturen nahegelegt. Allerdings folgen physikalische Wechselwirkungen zwischen metabolischen Enzymen anderen organisatorischen Regeln als Proteininteraktionen im allgemeinem PIN, das alle Proteininteraktionen enthält. Während Proteininteraktionen im allgemeinen PIN sich dissortativ verhalten, sind physikalische Enzyminteraktionen assortativ, d.h. dass die Anzahl der Interaktionen benachbarter Proteine im allgemeinem Netzwerk negativ und im metabolischen Netzwerk positiv korreliert. Ferner scheinen Enzyme von höherem metabolischen Durchsatz häufiger in Wechselwirkungen involviert zu sein. Enzyme der zentralen katabolischen Prozesse sowie der Biosynthese komplexer Membranlipide zeigen dabei einen besonders hohen Verknüpfungsgrad und eine dichte Clusterbildung. Einzelne Proteine wurden identifiziert, die die Hauptkomponenten des zellulären Metabolismus verbinden und so die Integrität verschiedener biosynthetischer Systeme essenziell beeinflussen könnten.

Neben dem metabolischen Aspekt der PIN wurde auch der Aspekt der Regulation sowie der Signaltransduktion, der Kinase-Interaktionen, näher analysiert. Dabei wurden charakteristische topologische Unterschiede der mit dem Metabolismus und der Phosphorylierung assoziierten PIN gefunden, die die unterschiedlichen Aufgaben beider

Netzwerke widerspiegeln. Eine nähere Untersuchung der Phosphorylierungs-Netzwerke ergab, dass die gegenseitige Phosphorylierung von Kinasen primär eher der Aufgabe der Signalleitung folgt (Kinase-Kaskaden) als ihrer gegenseitigen Regulation (Rückkopplungsprinzip). Die regulatorische Aufgabe des Netzwerks scheint überwiegend auf der Ebene der Zielproteine vorzuliegen, wo die Kreuzaktivität von unterschiedlichen Kinasen am selben Zielprotein dessen Funktionalität moduliert.

Die Rekonstruktion von Phosphorylierungs-Netzwerken basiert im Wesentlichen auf der Vorhersage von Kinase-Zielprotein Relationen und kann deshalb immer noch an der nicht genügenden Vorhersagegüte der angewandten Vorhersage-Algorithmen während der Bestimmung von Phosphorylierungsstellen (P-Stellen) und der dazugehörigen Kinasen leiden. Auch die experimentelle, genomweite Bestimmung der P-Stellen ist (noch) nicht durchführbar. Bisherige computergestützte Vorhersagemethoden beruhten für gewöhnlich auf der Auswertung charakteristischer Merkmale der lokalen, die P-Stelle umgebenden Proteinsequenz. Dieser Ansatz wird durch die Verwendung räumlicher 3D-Information in der vorliegenden Arbeit erweitert. Hierbei wird die Verteilung der Aminosäuren um die P-Stelle berechnet und spezifische 3D-Signaturen zur Vorhersage extrahiert. Beim Vergleich mit sequenz-basierten Vorhersagemethoden konnte eine konsistente Verbesserung der Vorhersage durch die Einbeziehung räumlicher Information gezeigt werden. Weiterhin wird in der vorliegenden Arbeit auch der Frage nach der Fehlerrate der experimentellen Phosphoprotein-Daten nachgegangen und ihre Verlässlichkeit bewertet. Die Verfügbarkeit eines verlässlichen Datensatzes ist bei der Entwicklung einer Vorhersagemethode ein entscheidendes Kriterium.

Neben einer auf räumlicher Information aufbauender Vorhersagemethode wurde auch eine sequenz-basierte Methode entwickelt, die implizit räumliche Aspekte wie die bevorzugte Sekundär- und Tertiärstruktur, Polarität, Volumen, Lösungsmittel-zugänglichkeit sowie strukturelle Unordnung erfasst und die als Teil einer neuen Datenbank pflanzenspezifischer P-Stellen (*PhosPhAt*) konzipiert worden ist. Mit einer genomweiten Darstellung aller im *Arabidopsis*-Genom experimentell gefundenen P-Stellen und einer rechnergestützten, speziell auf pflanzliche P-stellen zugeschnittenen Vorhersage stellt *PhosPhAt* eine wertvolle Informationsquelle für die Pflanzenforschung dar. Die entwickelte Vorhersagemethode wurde verwendet, um die genomweite Verteilung von mit vorhergesagten Phosphoproteinen assoziierten funktionellen Annotierungen zu untersuchen. Die mit hoher Wahrscheinlichkeit vorhergesagten P-Stellen bilden zusammen mit experimentellen Ergebnissen eine wertvolle Basis für eine tiefer gehende Untersuchung von Phosphorylierungs-Motiven in orthologen und paralogen Proteinen, und auch in unterschiedlichen Organismen.

Chapter 1
Introduction

The entirety of cellular processes are a tightly interlinked complex system, exhibiting interactions on different levels of organization, and synergizing to shape biological diversity and adaptation capacities to respond to environmental perturbations (**Figure 1.1**). Biological research is the effort to discover and increase our understanding of how this complex system works.

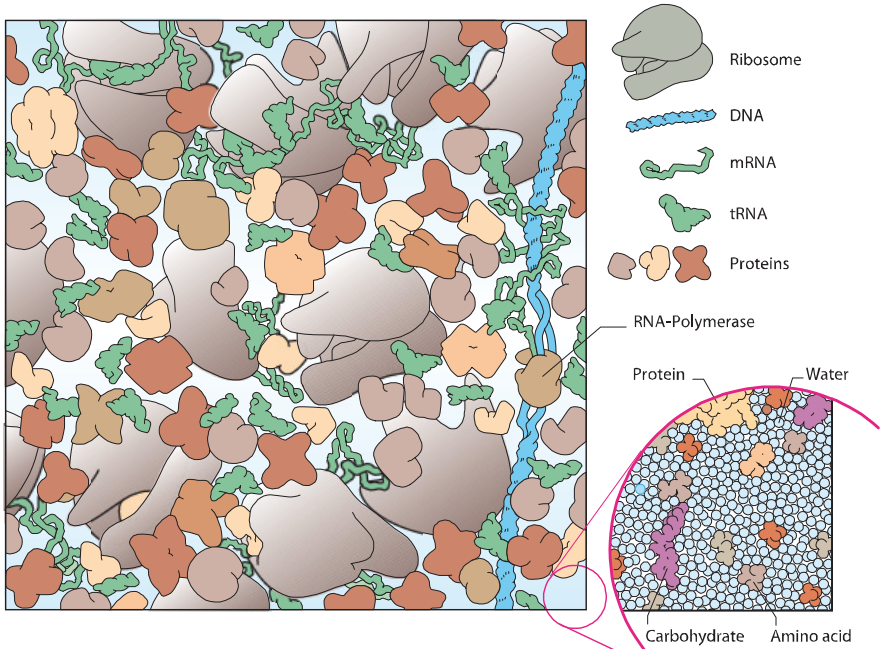


Figure 1.1 View into a bacterial cell
Cellular processes form a tightly interlinked complex system, exhibiting interactions on different levels of organization. To understand the underlying principles, well known graph-theoretical approaches may be applied. Figure reproduced from Color Atlas of Biochemistry[1].

The study of biological systems as a whole seemed for a long time inconceivable, as the available technologies showed evident limitations. Thus, classical genetics, biochemical and protein chemistry was focused on the study of pointed interactions, by tackling selected biomolecules in a time consuming, laborious manner. Eventually this strategy yielded collections of thousand of observations and, thus, allowed a partial reconstruction of biological interaction networks. Improved strategies as well as technical advances in equipment have contributed to a system-wide record of biological interactions, thus, providing the possibility to elucidate entire networks. Furthermore, the genomics and postgenomic era gave rise to a new biological field of science called systems biology, which aims opposite to the classical scientific reductionism, to integrate diverse types of

biological interactions and to discover emergent properties resulting from the many processes occurring in biological systems (**Figure 1.1**).

Typical biological interaction networks at the molecular level are metabolic networks, regulatory networks, signaling networks, and protein interaction networks. All these networks describe different types of interactions and typically are viewed as different entities. However, all these networks also contain aspects of the other network types, albeit theses aspects are not the explicit purpose of the particular network. Thus, regulatory and signaling interactions are often based on protein-protein interactions reflecting, for instance, kinase-target recognition events. Metabolic interaction networks contain aspects of the protein interaction networks as well. Successive enzymes may form enzymatic complexes influenced by protein interactions, which may in turn affect metabolic pathways. Furthermore, there is also an essential component of regulation in metabolic networks, since enzymes are regulated through interactions with substrates and products such that the appropriate conditions in the cell are maintained. As outlined here, protein-protein interactions may form the base of metabolic, signaling and regulatory network. Hence, protein interaction networks represent a complex mixture of different functional aspects of cellular processes. Therefore, it is very interesting to know, in how far these aspects affect the structure of protein interaction network itself, as well as whether an integrative analysis of different network types reveals corresponding topological properties of the architectures of the underlying networks. In the following, the principles of protein-protein, metabolic and phosphorylation interaction networks are briefly described.

1.1 Protein Interaction Networks

Protein-protein interaction is perhaps the most prototypical biological interaction, which is the most essential event for cellular functions. The structural design plan of proteins is stored in the desoxyribonucleic acid (DNA), where the information is coded by a specific sequence of four nucleotide bases Adenine, Threonine, Cystein and Guanine. The process of information transmission from DNA to proteins is called gene expression[a], and can be in principle divided into two major parts, which are the transcription of DNA to ribonucleic acid (RNA) and the following translation of RNA into protein sequence (**Figure 1.2**). Both the transcription and translation already involve complex protein–protein interactions, which correspond to regulatory and catalytic functions. The initiation of transcription starts with protein-protein interactions of transcription factors and the RNA polymerase, which itself is already a highly interconnected catalytic protein complex. In the subsequent reactions, the processing of nascent RNA by splicing and translation of

[a] Gene expression is a general term, which describes the transformation of DNA-information into functional molecules.

RNA, protein-complexes and regulatory protein interaction are involved as well. Thus, gene regulatory networks are in principle based on protein-protein interactions. The following rather unspecific protein interactions are the folding of amino-acid sequences by chaperons. Amino-acid sequences are not (yet) mature proteins, but may falsely be interpreted as such by high-throughput experiments, which involve identification of interaction partners by successive protein digestion and the determination of the amino-acid sequence. Even the next step of a protein's life, the transport of proteins to the designated compartments, involves rather unspecific protein-protein interactions. At this stage, the protein is ready to perform its designated function. However, the function may be regulated by so-called posttranslational modifications, e.g. proteolytic reactions, phosphorylations, acetylations etc., that can alter the protein structure and thus its functionality. The posttranslational modifications are accomplished by specific, enzyme-target-protein recognition events. During their life, proteins may form permanent protein-complexes or interact with other structural proteins of the cell. Of course, the designated function of a protein might be protein-protein interactions as outlined before. Finally, the ubiquitination of proteins initiates the degradation of protein, by proteases, and in both processes protein-protein interactions are involved as well.

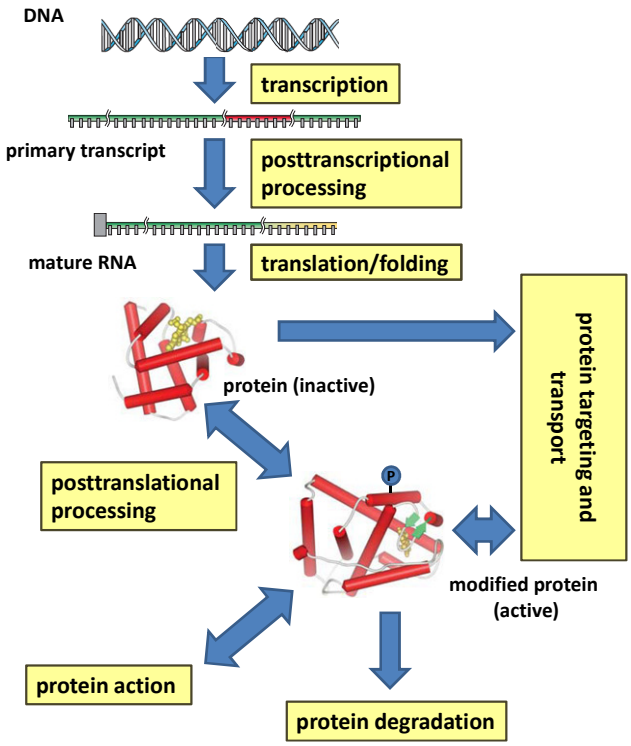


Figure 1.2 Simplified scheme of proteins' life.
The information contained in the DNA undergoes several processes to result in an active protein. In its native form, the protein may perform the designated function, which may then be regulated by further posttranslational modifications. At the end of its life, proteins are degraded by proteases. Although not all processes explicitly describe protein-protein interactions, in all shown processes protein-protein interactions are involved.

As outlined here, the raw protein interaction networks comprise a great variety of different aspects of protein-protein interactions. Thus, it is of great importance to consider the context of protein interaction network, by integrating other sources of information. By performing such strategies, the investigation of protein interaction networks may reveal novel organizing principles which are otherwise not apparent.

1.2 Metabolic Interaction Networks

Another type of biological interactions forming networks are metabolic interactions. Metabolic Interaction Networks are complete sets of metabolic and physical processes that determine the physiological and biochemical properties of a cell. As such, these networks comprise the chemical reactions of metabolism as well as the regulatory interactions that guide these reactions. In this work, only metabolic interaction networks, which comprise only metabolic reactions are discussed (**Figure 1.3**). In this sense, metabolic interactions are connections between metabolites that are involved in particular reactions (metabolite interaction networks), or interactions of successive enzymes connected via common substrates or products (enzyme interaction networks).

However, even these metabolic interaction networks are affected by protein-protein interactions. Protein-protein interactions may influence the spatial proximity of successive enzymes, since enzymes may be linked to common structural proteins, allowing a particular reaction, or even increase the metabolic performance by so-called metabolic channeling, while a spatial separation of reactions may disfavor theoretically possible but detrimental reactions from the network. We will come back to metabolic channeling in the later parts of this section. These theoretically possible reactions are not observed in experimental studies and thus also not contained in curated representations of metabolic pathways such as KEGG-maps. However, the automatic reconstruction of the metabolic interaction networks, which is based solely on reaction lists, involves the risk to losing this biological curation. The inclusion of currency metabolite such as ATP, H_2O, co-substrates, etc., for instance, may lead to irrelevant interpretations of the entire metabolic network topology (**Figure 1.3**). These metabolites may artificially connect enzymes or metabolites, which naturally are not observed to interact, as the metabolite is ubiquitously distributed in high amounts throughout the cell. The role of a particular metabolite in a reaction is often not obvious. Thus, strategies have been developed to trace the functional groups of metabolites to reveal main metabolic connections. This work takes advantage of such strategies. However, even these construction schemes suffer from omitting the compartimentation of reactions within the cell. Hence, one of the Metabolic Interaction Networks used in this study is explicitly derived from the visualization schemes of metabolic pathway from KEGG-maps.

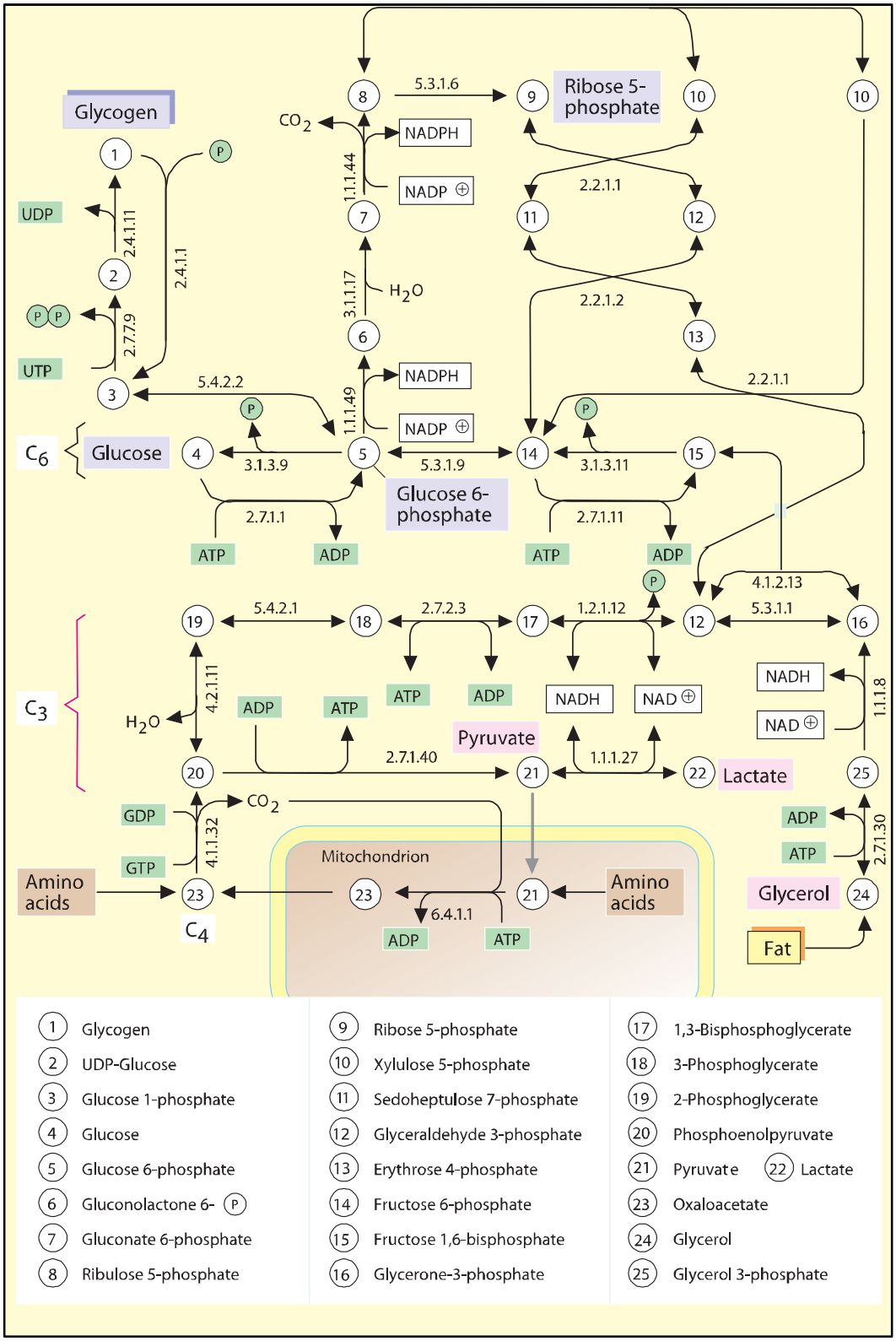


Figure 1.3 Scheme of the carbohydrate metabolism
Metabolites are converted by enzymatic reactions to shape a metabolic interaction network. Many reactions involve so-called currency metabolites such as ATP and NADH, which are ubiquitously distributed in high amounts throughout the cell. The connection of these metabolites would remarkably increase the interconnectivity of the network and probably obliterate the metabolic-pathway view of this network. Figure reproduced from Color Atlas of Biochemistry[1].

Metabolic channeling

Metabolic channeling is defined as the direct tunneling of intermediates from an active site of an enzyme to another without prior dissociation into the bulk solvent [2; 3]. The advantages of substrate channeling comprise increased catalytic efficiency by shorter transient times of movement between the active sites [4; 5] and local enrichment of substrates by bridging cooperating active sites in spatial proximity, protection from toxic intermediates, prevention of decomposition of unstable chemical compounds [6], overcome of unfavorable equilibria [7; 8; 9] and circumvention of competitive pathways [10; 11; 12; 13; 14; 15]. Although the concept of metabolic channeling was controversial at times since first proposed in the 1930ies, it is supported by metabolic control as well as experimental evidence from analysis of biochemical pathways theory [16; 17; 18; 19; 20; 21; 22].

The evidence for metabolic channeling has been found in urine and pyrimidine biosynthesis, amino acid metabolism, glycolysis and the TCA cycle, DNA replication, RNA-synthesis, and protein biosynthesis. Evidence for tunneling of intermediates was found by a number of approaches including measurement of the transient time, isotope dilution of endogenous reaction intermediates. The experiments range from the isolation of aggregates of TCA cycle by David Green in the 1940ies [23] and the absence of intermediates of the tryptophan synthetase described by Charles Ynofsky in 1958 to structural analysis of diverse complexes, and the investigation of the mechanism of channeling. Among the best characterized metabolon[b] are tryptophan synthetase, pyruvate dehydrogenase, the glycine decarboxylase system, TCA and the Calvin cycle as well as enzymes of glycolysis and fatty acid oxidation**.** All of these systems involve high-affinity, stable enzyme associations that are most amenable to experimental analysis. However, there is evidence that even parts of pathways, which were once believed to consist of soluble proteins are able to build short-living dynamic complexes by weak or transient interactions, which can dissociate or reform in response of the demand of the cell [24]. Furthermore, enzyme complexes are often associated with structural parts of the cell [16; 24], suggesting an integral analysis of protein interactions in the context of metabolic pathways by treating both levels of molecular organization as graphs and investigation of global as well as local network properties.

1.3 Phosphorylation Networks

The recognition of specific protein sites by protein kinases is central to the phosphorylation networks (**Figure 1.4**), and is followed by the covalent attachment of phosphate groups to amino acids serine, threonine, or tyrosine. Phosphorylation can

[b] The term metabolon describes a non-covalent association of several sequential enzymes involved in a metabolic pathway.

modify enzymatic activities, binding affinities, and protein conformations. It is reversible and its dynamics allows fast and precise changes in protein properties resulting in broad consequences for the protein-protein interaction network and intra-cellular signaling or even cytoskeleton remodeling as well as establishing cell-cell communication in multicellular eukaryotes[25]. Thus, the phosphorylation networks comprise both parts of the signaling as well as regulatory interaction network. Since the phosphorylation event is a directed action, it is possible to represent phosphorylation networks as directed graphs, thus, allowing the study of the effectiveness of the system, the influence of a particular kinase on the entire network as well as the stability of the information flow.

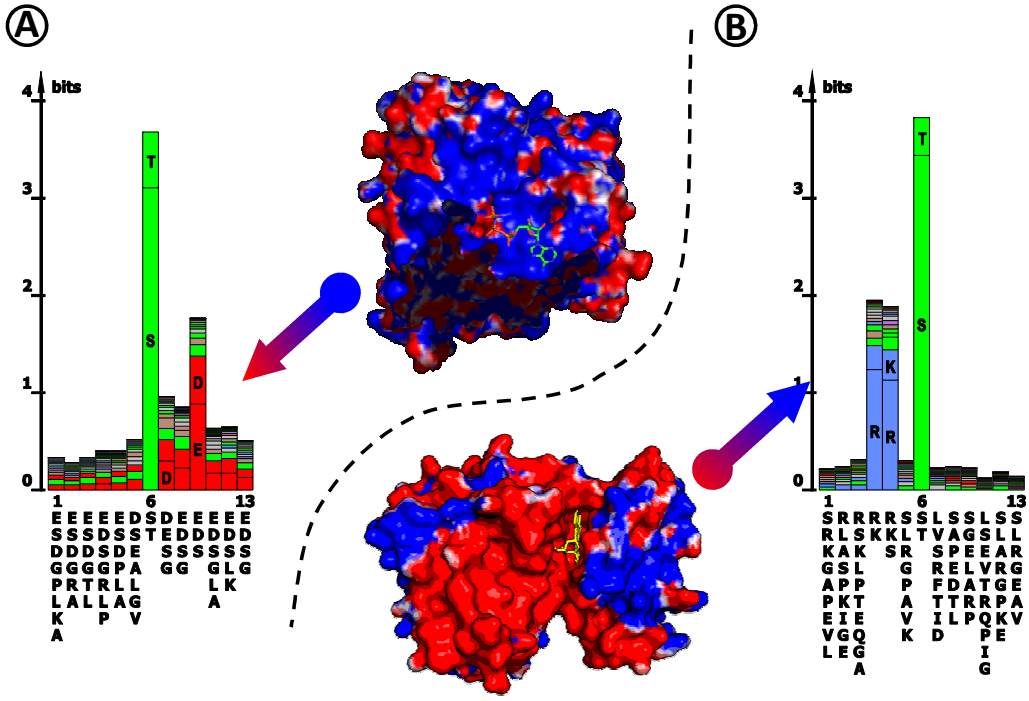


Figure 1.4 Principle of kinase-target recognition
A) The predominantly positively charged activation pocket of the CKII kinases (blue-colored surface regions) recognizes multiply presented aspartates and glutamates surrounding the central amino acid of target sequence motifs. Crystal structure of CKII with interacting ATP /1daw/. B) The negative charged activation pocket of the PKC kinases recognizes arginines and lysines in -3 and -2 position of the central amino acid of the target sequence motifs. Electrostatic surface potential calculation was generated using the Adaptive Poisson-Boltzmann Solver – APBS[26] and rendered by PyMol molecular graphics system. PKC-theta with its Inhibitor Staurosporine /1xjd/.

Classical genetics, biochemical and protein chemistry have tackled selected proteins to identify and characterize one, in special cases few, phosphorylation sites at a particular amino acid position and to describe the functionality in a switch-on–off manner. However, investigations of the same protein yielded other phosphorylation sites, revealing a more complex crosstalk between several phospho-site positions in the same protein, or even more complex dependencies between different phosphorylation sites in different proteins (**Figure 1.5** and **Figure 1.6**). Nowadays, improved experimental

strategies as well as technical advances in equipment allow large-scale identification of hundreds of phosphorylation events under different biological conditions, and provide access to a more complete record of temporal changes of the phosphorylation state in response to cellular perturbations[27; 28; 29; 30; 31; 32]. However, most data sources do not provide any information on the respective kinase, thus the reconstruction of phosphorylation interaction network is still based on low-throughput results directly or indirectly using predictor-based strategies[33]. It remains to be seen when these strategies will be replaced by reliable high-throughput results based on improved kinase-protein-chip or alternative technologies.

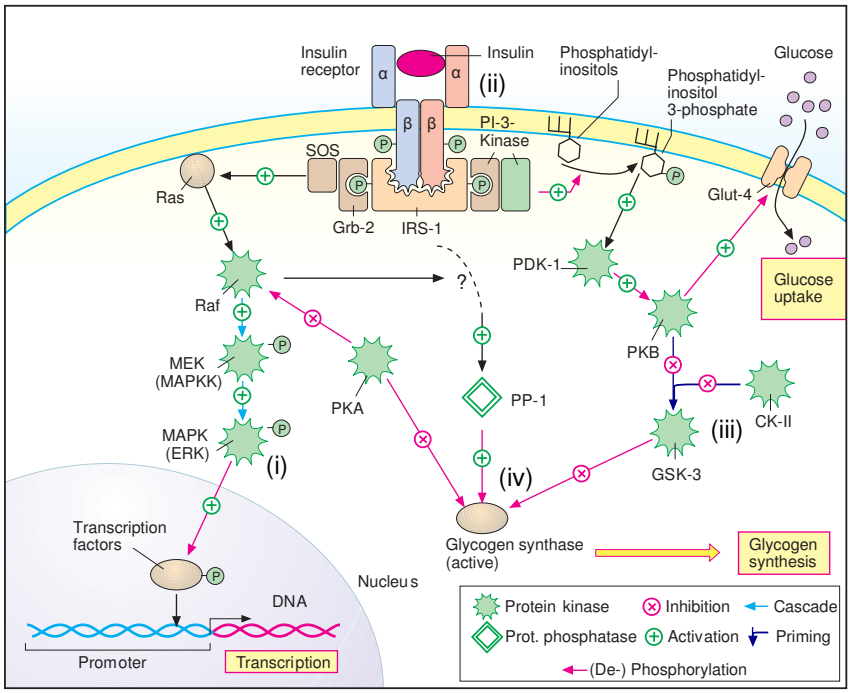


Figure 1.5 Simplified view on Insulin signal transduction.
The involved signal transduction pathways have not yet been fully explained, thus the figure presents a simplified form. The diverse effects of insulin are mediated by protein kinases that mutually activate each other in the form of enzyme cascades or directly influence the activity of glycogen synthase. (i) The kinases at the bottom of a cascade chain may influence gene transcription. (ii) The dimerisation of the insulin receptor (top) upon binding of the hormone increases the tyrosine kinase activity of the receptor as well as the release of the second messanger IP3. IP3 may itself activate kinase cascades. (iii) Priming is the introduction of novel phosphorylation recognition sites by phosphorylation. Phosphorylation of GSK3 kinase by CKII creates a phosphorylation site for the PKB-kinase. (iv) Cross-reactivities of different kinases regulate the signaling transduction as well as the activity of glycogen synthase. Within the regulation, protein phosphatases may play an equally important role, than protein kinases. Figure reproduced and altered from Color Atlas of Biochemistry[1].

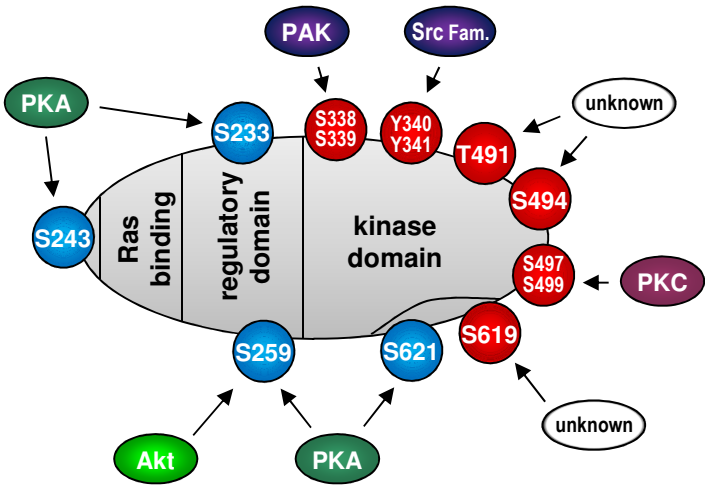


Figure 1.6 Multiple regulatory phosphorylations of Raf-kinase and its respective kinases
Independent low-throughput experiments identified multiple phosphorylation sites susceptible to phosphorylation by one of the cellular protein kinases. Phosphorylation either positively (red) or negatively (blue) affects the activity of the kinase. Hence, regulation of this enzyme is rather a matter of fine-tuned modulation than a binary switch-on-off activation-deactivation mechanism. The high number of phosphorylation sites suggests high cross-reactivities of kinases.

1.4 Thesis Overview

This thesis is concerned with the investigation of organizing principles of different network types to find correlations between the topology of networks and their function. Chapter 3, presents an integrative analysis of protein and metabolic interaction networks. In particular, various construction methods are applied to capture metabolic aspects of protein interaction networks and the topologies of underlying networks are compared. Thus, results from this section shed further light on the unifying principles of functional (metabolic) as well as physical interaction networks. In Chapter 4, the investigated network-types are compared to the dynamic phosphorylation network, revealing emergent properties of the different network types. The phosphorylation network consists of kinase-target interactions, where the reconstruction of the networks was based either on experimental evidence of protein-protein interactions or kinase-specific prediction of phosphorylation events.

We noted that the present phosphorylation network suffers from inaccuracies and low specificities of the underlying construction procedures. Thus, besides the investigation of the present networks, we also investigated new methods to improve prediction of phosphorylation sites. In Chapter 6 we investigated the spatial context of phosphorylation site to identify specific 3D-signature profiles and to include the spatial information to improve the performance of predictions. Furthermore, we noted a poor presence of information sources for plant phosphoproteomics. Consequently, a new database was established in Chapter 7, to provide a valuable resource for the plant

community by presentation of a comprehensive view of phosphorylation sites and in-depth information of the corresponding experiments. The database is further supplemented by a plant specific phosphorylation site predictor, which was specifically designed to predict plant specific phosphorylation events. Thus, the database provides a toolbox to collect experimentally identified phosphorylation sites as well as to predict and evaluate phosphorylation events.

Chapter 2
Graph theoretical concepts in the context of Biological Networks

The representation of complex biological networks as graphs and the study of their properties have contributed to an emerging system-wide approach towards studying the organizing principles of cellular and molecular processes. Well-understood graph-theoretical concepts have been applied to describe structural and dynamical properties of cellular systems and to study the relevance of network topologies on static as well as dynamic properties of the systems under investigation.

A graph consists of nodes, which are connected by directed or undirected links. In biological networks, the nodes may for instance represent proteins, enzyme, metabolites or genes, which are connected by intuitive interactions such as physical protein-protein interaction and metabolic reactions, or relations which result not directly from experimental data, such as in correlation networks. For some interactions the direction of the interaction may be of great importance. The directed interaction may for instance consider the direction of metabolic reactions or the direction of information flow in signaling networks. In graphs, the number of links of nodes is referred as the connectivity or the degree (k) of nodes.

In general, biological networks are widely accepted to be scale-free, as the probability distribution of degrees $P(k)$ for many networks was found to adhere to a heavy tailed distribution and approximately to fit the power-law $P(k) \sim k^{-\gamma}$. Furthermore, they exhibit properties of scale-free graphs such as a high clustering coefficient (neighbors interconnectivity) and short characteristic length (average distance between entities of the graph) [34; 35; 36; 37; 38].

The investigation of biological networks is not limited to a particular network type. On the contrary, integrated analysis of different network types for different levels or domains of molecular organization supports the notion that concepts are transferable from one network type to another. Ge et al. showed that gene expression and protein interaction data are correlated [39]. Kemmeren and co-workers as well as Deane et al. used gene expression data to assess confidence levels for protein interaction networks [40; 41]. Goldberg and Roth predicted genetic interactions by utilizing protein interaction data [42], and Kelly and Ideker predicted the physical context of genes [43]. Rhodes et al. used GO-annotations, integrated interlogs and expression data as well as data of protein domains, known to interact to predict protein interactions [44]. The use of gene co-expression data to identify protein interactions has also been demonstrated recently [45]. Finally, Lee et al. integrated expression, gene-fusion, phylogenetic profile, literature co-citation as well as protein interaction data to predict functional associations [46]. Recently, Huthmacher et al.

investigated metabolic interaction networks in the context of direct protein-protein interactions, revealing novel enzyme pairs potentially involved in metabolic channeling[47].

2.1 Topological properties of graphs

The small-world effect

Perhaps the most noteworthy experiments on real networks were performed by Travers and Milgram [48; 49] in the 1960s, in which letters passed on from person to person reached their designated target individual in only a small number of steps. 256 arbitrarily selected participants from Nebraska and Boston were asked to generate acquaintance chains to a target person in Massachusetts. The average number of intermediaries between starters and targets (the characteristic length; CL) for the social network was observed to be 5.2. The observation of short average distances between nodes in a network is often called the "small world effect". The small-world effect has obvious implication for the dynamics of processes taking place within the network such as information transduction.

For biological networks, the reported characteristic lengths differ to a high degree. Almaas et al. and Bhan et al. reported a CL of around 3 for raw protein-protein interaction networks (protein interaction as derived directly from the database, without prior filtering) [50; 51]. A similar CL was reported for metabolic interaction networks as well [37; 52]. Fell and Wagner observed that the metabolic path from glutamine and pyruvate, perhaps the most central metabolites, to all other metabolites was around 2.5 in average [38; 53]. However, curation of metabolic interactions by tracing particular atoms and functional groups of molecules during reactions yield different results [54; 55]. Considering a substrate-product interaction only in case of direct evidence of chemical transfer or interchange generated a graph with a CL of around 9, i.e. substantially larger than previously suggested.

Degree distribution

To gain a more detailed insight into the topology of a network the probability distribution of the connectivity or degree of nodes can be inspected. As mentioned before, biological networks are generally believed to be scale-free, as the probability distribution of degrees: $P(k)$ for many networks was found to adhere to a heavy tailed distribution and approximately to fit the power law $P(k) \sim k^{-\gamma}$ [34; 35; 36; 37; 38], albeit it was often suggested that alternative models may fit the distribution better [56; 57; 58] or even, that they are not scale-free after all upon curration according to the biological

knowledge[59]. The approximately scale-free character of biological networks implies robustness of the networks against errors or random attacks, as the high majority of the nodes plays a less important role in the assembly of the graph [60]. One of the most popular models capturing the power-law related degree distribution was proposed by Barabási and Albert[34]. The model is often called the "rich-get-richer" effect and describes the preferential attachment of a new node to already existing highly connected nodes. The probability of the connection is assumed to be linearly proportional to the node degree and results in a scaling factor of $\gamma = 3$. However, it appears unlikely that an evolutionary process follows the rich-get-richer rule directly, measuring the node's network neighborhood. Indeed, alternative processes exist comprising gene duplication and diversification that give rise to a scale-free connectivity distribution as well [61; 62; 63]. Chung et al., for instance, proposed a duplication model, which supports the observation of the general tendency of biological networks to have scale-free exponents below two, although larger exponents have also been reported for metabolic interaction networks [37].

Clustering coefficient

It was often suggested that biological networks are functionally modular[64], which gives rise to a high average clustering coefficient ($< c >$) . Since the clustering coefficient measures the cliquishness of a particular network, that is the ratio of triangles between items in the existing compared to a fully connected network, the average clustering coefficient $< c >$ provides information on the global distribution of links (Eqs 2.1).

Eqs. 2.1. $$c_{i \in N_{k>1}} = \frac{\sum_{r,p \in N_{k>1}} A_{i,r} A_{i,p} A_{p,r}}{k_i (k_i - 1)} ; < c > = \frac{\sum_{i \in N_{k>1}} c_i}{\left| N_{k>1} \right|} ;$$

where A denotes the adjacency matrix with elements set to 1 in case of an established link between nodes and zero otherwise; k_i is the degree of node i for which c is computed, i, p, and r are indexes of all nodes in the network with $k > 1$.

A value close to unity indicates a high modular network, whereas a $< c >$ close to zero its absence. For biological networks, in particular metabolic interaction networks, a $< c >$ of around 0.6 was reported, supporting the modular character of metabolic networks.

Assortative mixing and Neighbors' Connectivity

Assortative mixing or correlation between properties of neighboring nodes in a graph, in particular the correlation of degrees, was widely studied in ecology and

epidemiology, as it has an enormous impact on understanding the spread of diseases. In principle, the assortativity (r_d) (Eq. 2.2) measures the affinity of nodes to nodes of either equal or higher degrees.

Eq. 2.2.
$$r_d = \frac{|E|^{-1}\sum_{i\in E} j_i k_i - \left[|E|^{-1}\sum_{i\in E}\frac{1}{2}\left(j_i^2 + k_i^2\right)\right]^2}{|E|^{-1}\sum_{i\in E}\frac{1}{2}\left(j_i^2 + k_i^2\right) - \left[|E|^{-1}\sum_{i\in E}\frac{1}{2}\left(j_i + k_i\right)\right]^2};$$

where r_d, is the assortativity, j and k are the degrees of nodes at the ends of the ith edge within the set of considered node pairs E and $|E|$ is their total number.

Biological networks were reported to be dissortative, e.g. the degrees of neighboring nodes in the networks are observed to be negatively correlated, such that highly connected nodes are linked to nodes with smaller connectivity.

As an alternative to the assortativity measure, the Neighbors' Connectivity distribution may be plotted $< NC(k) >$. The function $< NC(k) >$ is an increasing function for assortative networks, which means high-degree nodes tend to attach to other high-degree nodes, and a decreasing function for dissortative networks, where low-degree nodes prefer to interact with high-degree nodes $< NC(k) >$.

2.2 Centrality of nodes

Recall the acquaintance-chains experiment by Travers and Milgram (see the small-world effect), where the average number of intermediates between starters and targets was observed to be 5.2, they also observed that 48% of the chains were passing through three persons. Thus, the individuals may be designated as central[48] or betweenness-central as defined by number of shorted paths passing though a node[65; 66].

Evidence for the importance of central nodes was also observed for biological networks. Betweenness-central nodes, the so called bottlenecks, were observed to be essential[67; 68]. Furthermore, since the structure of scale-free networks implies a much faster decomposition upon focused attack on the hub nodes, as defined by the degree-centrality (DC), in comparison to random networks[60] and the reported correlation between betweenness-centrality (BN) and degree-centrality[69], it might be suggested that hub nodes of the investigated networks tend to be essential as well. Indeed, a tendency of hub nodes to be lethal when knocked out was reported by Jeong et al. [70]. However, He and Zhang suggested that increased lethality of highly connected nodes might result from a higher probability of hub nodes to participate in an essential link, rather than

stabilizing the network structure[71]. Furthermore, Yu et al. compared both measures in the context of essentiality and concluded that betweenness-centrality is a much more significant indicator for essentiality, in particular in regulatory networks[72].

While the betweenness-centrality accounts for the flow of information within a network, it does not account for the stability or robustness of the network. Here the robustness-centrality (RC) may be applied, which is defined by the effective change of the characteristic length of the graph upon removal of a particular node (**Figure 2.1**). The idea of the robustness-centrality was previously applied to identify crucial residues for enzyme activity[73]. In principle, a robustness-central node will also be central in terms of betweenness-centrality (**Figure 2.1**). However, shortest paths through betweenness-central nodes may be bypassed by equally central nodes, resulting in a lower robustness-centrality.

Although all three centrality measures will correlate to a certain degree, they all reveal different aspects of centrality in networks. We applied the betweenness-centrality, to measure the centrality in the context of metabolic efficiency and the robustness-centrality to measure the contribution of particular nodes to the structural integrity of the protein-interaction networks. While the BN accounts for essential metabolic proteins that are frequently observed to exist in several iso-forms (iso-enzymes), the RC neglects such proteins arguing that structural integrity of the networks retains preserved when such proteins are removed from the network (see Materials and Methods of Chapter 3 for the mathematical definitions).

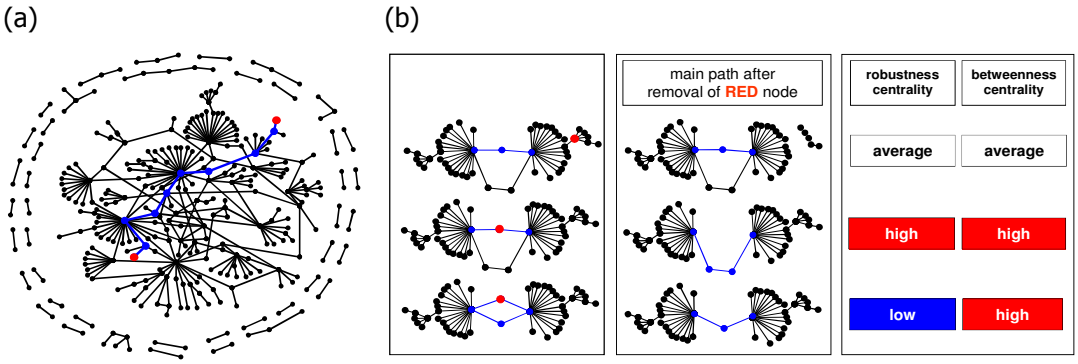


Figure 2.1 Definition of the shortest path (a); differences between the robustness-centrality and the betweenness-centrality (b). Figure (a) shows the shortest path (blue) between the red nodes. Figure (b) shows the mean path before and after the removal of the red node, as well as relative robustness and betweenness values.

Chapter 3
The integrated analysis of metabolic and protein interaction networks reveals novel molecular organizing principles

Abstract The study of biological interaction networks is a central theme of system biology. Here, we investigate the relationships between two distinct types of interaction networks: the metabolic pathway map and the protein-protein interaction network (PIN). It has long been established that successive enzymatic steps are often catalyzed by physically interacting proteins forming permanent or transient multi-enzyme complexes. Inspecting high-throughput PIN data, it was shown recently that, indeed, enzymes involved in successive reactions are generally more likely to interact than other protein pairs. In our study, we expanded this line of research to include comparisons of the underlying respective network topologies as well as to investigate whether the spatial organization of enzyme interactions correlates with metabolic efficiency.

Analyzing yeast data, we detected long-range correlations between shortest paths between proteins in both network types suggesting a mutual correspondence of both network architectures. We discovered that the organizing principles of physical interactions between metabolic enzymes differ from the general PIN of all proteins. While physical interactions between proteins are generally dissortative, enzyme interactions were observed to be assortative. Thus, enzymes frequently interact with other enzymes of similar rather than different degree. Enzymes carrying high flux loads are more likely to physically interact than enzymes with lower metabolic throughput. In particular, enzymes associated with catabolic pathways as well as enzymes involved in the biosynthesis of complex molecules were found to exhibit high degrees of physical clustering. Single proteins were identified that connect major components of the cellular metabolism and may thus be essential for the structural integrity of several biosynthetic systems.

Our results reveal topological equivalences between the protein interaction network and the metabolic pathway network. Evolved protein interactions may contribute significantly towards increasing the efficiency of metabolic processes by permitting higher metabolic fluxes. Thus, our results shed further light on the unifying principles shaping the evolution of both the functional (metabolic) as well as the physical interaction network.

3.1 Background

To ensure stable and efficient metabolic processes in cells, highly coordinated molecular interactions of the involved enzymes and metabolites are necessary. The study of spatially organizing principles of metabolic pathways has long been a research focus in cellular and molecular biology. Organelle compartmentalization and the organization of enzymatic pathways in so-called metabolons have been discussed as the main cellular-scale as well as molecular-scale organizational units to orchestrate the multiple metabolic processes inside cells and to separate as well as to integrate them in space and time. First introduced by Srere, the term metabolon describes a non-covalent association of several sequential enzymes involved in a metabolic pathway [10]. Similar to industrial assembly lines, intermediates are passed on from one enzyme to the next, referred to as metabolic channeling, leading to an optimized metabolic flux. The stability and structural integrity of metabolons varies greatly ranging from temporary associations and their dynamic formation in response to environmental changes to stable, permanent enzyme complexes [74; 75]. Furthermore, it was found that enzyme complexes are often associated with intra-cellular membrane systems[16; 24; 76], demonstrating that the spatial organization of the metabolic network is not only limited to direct physical interaction of participating enzymes, but that it also involves passive – in the context of enzymatic pathways - mediating structural cellular components.

Metabolic channeling provides several advantages such as an increase of catalytic efficiency by shorter transition times between the consecutive active sites [4; 5], local enrichment of substrates, protection from toxic intermediates by shielding them from the cellular environment, prevention of decomposition of unstable chemical compounds [6], overcoming of thermodynamically unfavorable equilibria [7; 8; 9], as well as avoidance of competitive pathways [10; 11; 12; 13; 14; 15]. Although the concept of metabolic channeling has been discussed controversially at times [16], it is now supported by metabolic control analysis as well as experimental evidence [16; 17; 18; 19; 20; 21; 22].

Recently, Huthmacher and co-workers analyzed the metabolic networks of yeast and *Escherichia coli* in the context of direct protein interactions as observed in newly available, large-scale protein-protein interaction surveys allowing a systematic scan for direct protein interactions of consecutive metabolic pathway enzymes [47; 52]. They found higher frequencies of physical interactions of enzymes sharing at least one common metabolite in the network. The chance for enzymes to physically interact was observed to be negatively correlated to the distance between enzymes in metabolic network in *E.coli* and, to a lesser degree, in yeast as well. In addition, they reported a higher probability of regulating enzymes to interact with other proteins, where regulating enzymes were defined either by a threshold of Gibbs' free energy change of the associated reaction or by their position within the network as being located at highly connecting branching

points. Furthermore, the analysis of high-throughput protein-protein interaction data yielded a number of novel candidates for metabolic channeling. Thus, the functional significance of protein-protein interactions for the metabolic pathway organization has been established and is supported by many experimental observations.

Here, we aim to expand the view on protein interactions in the context of metabolic pathways by treating both levels of molecular organization as network graphs and to investigate global as well as local network properties. The representation of complex biological networks as graphs and the study of their properties have contributed to an emerging system-wide approach towards studying the organizing principles of cellular and molecular processes. Global topological graph properties such as the degree distribution have received particular attention and have been discussed in the context of network stability and information exchange within networks [34; 35; 36; 37; 38; 50].

The integrated analysis of different network types for different levels or domains of molecular organization has been applied to transfer evidence to support particular interactions from one network type to another. Ge et al. showed that gene expression and protein interaction data are correlated [39]. Kemmeren and co-workers as well as Deane et al. used gene expression data to assess confidence levels for protein interaction networks [40; 41]. Goldberg and Roth predicted genetic interactions by utilizing protein interaction data [42], and Kelly and Ideker predicted the physical context of genes [43]. Rhodes et al. used GO-annotations, integrated interlogs and expression data as well as data of protein domains, known to interact to predict protein interactions [44]. The use of gene co-expression data to identify protein interactions has also been demonstrated recently [45]. Finally, Lee et al. integrated expression, gene-fusion, phylogenetic profile, literature co-citation as well as protein interaction data to predict functional associations[46].

In this study, we expand on the study of Huthmacher and co-workers by investigating the entire protein interaction network and its significance for metabolic networks and metabolic pathways. We extended enzymatic physical interactions to also include non-enzymatic proteins as metabolic relationships between enzymes may also be mediated by metabolically inactive interface proteins. Specifically, we investigate whether large-scale topological equivalences of both the metabolic and protein interaction network can be detected. Furthermore, as the physical organization of metabolic pathways is likely to have been under evolutionary optimization to increase metabolic throughput, we are studying here whether available flux data can be correlated to the protein interaction data supporting this hypothesis. So far, protein interaction data have been analyzed primarily across all functional categories. Here, we compare the general organization principles with those observed for the enzymatic protein subset, and report that indeed, specific differences do exist. The significance of topological parameter distributions has largely been analyzed within the context of the examined network type

itself, but not across different network types. For example, Macdonald and co-workers discovered defined relationship between fluxes going through metabolic network edges and the degree product of the connected nodes [77]. Here, we explore whether such relationships can be established across network types, in particular protein interaction and metabolic networks.

Thus, our investigations aim to establish whether unifying principles shaping the evolution of both protein interactions as well as metabolic pathways can be detected.

3.2 Results

3.2.1 Topological Properties of Interaction Networks

We start our investigations by first characterizing the global network properties of the various types of molecular networks examined in this study. Besides the two main network types, the protein interaction network and the metabolic network, further filtering and different construction methodologies were applied to reveal organizational differences between raw networks including all interactions, and networks designed specifically to capture aspects of metabolism and to also safe-guard against possible artifacts resulting from a particular reconstruction scheme.

Protein Interaction Networks (PIN)

The raw PIN (rPIN, see Methods) derived from the merged databases of DIP and BIOgrid comprises 5,438 proteins involved in 39,766 physical interactions. The network does not differentiate whether the interaction between two proteins is transient or permanent, or under which conditions the proteins were found to interact, or the functional relevance of the association. As the PINs used in this study are represented as undirected graphs, the functionality of an interaction cannot be resolved. A kinase interacting with a protease may activate the protease or be degraded by it.

The connectivity distribution $P(k)$ of the rPIN can be approximated by a power-law function with $P(k) \approx k^{-\gamma}$, where γ - the scale-free exponent - is the slope of the linear regression line in the double-logarithmic diagrams (**Figure 3.1**A). The value of γ was observed as 1.6 for the rPIN. The deviation from a straight line in the double-logarithmic suggests that a better fit may be obtained by introducing a mixture of power law and exponential degree distribution as was observed similarly for the Drosophila protein interactome [78] and other molecular networks [56; 57; 58; 59]. As it is typical for biological networks, the great majority of proteins show a small number of links whereas few proteins have up to 330 interactions. The rPIN network graph is characterized by a relatively short characteristic length (CL) of 3.49 ± 0.01, i.e. the average path between any two proteins was observed to lead through 3.49 nodes. Of the 39,766 physical

interactions, 15,232 occurred in the cytosol, 58 between two membrane associated proteins, and 298 were interactions between a membrane associated protein and a soluble protein. The sub-cellular localization information of 16,739 interactions was incomplete. GO-cellular component annotations for 7,439 interactions were inconsistent; i.e. participating proteins were reported in different compartments, and have thus been discarded from the analysis.

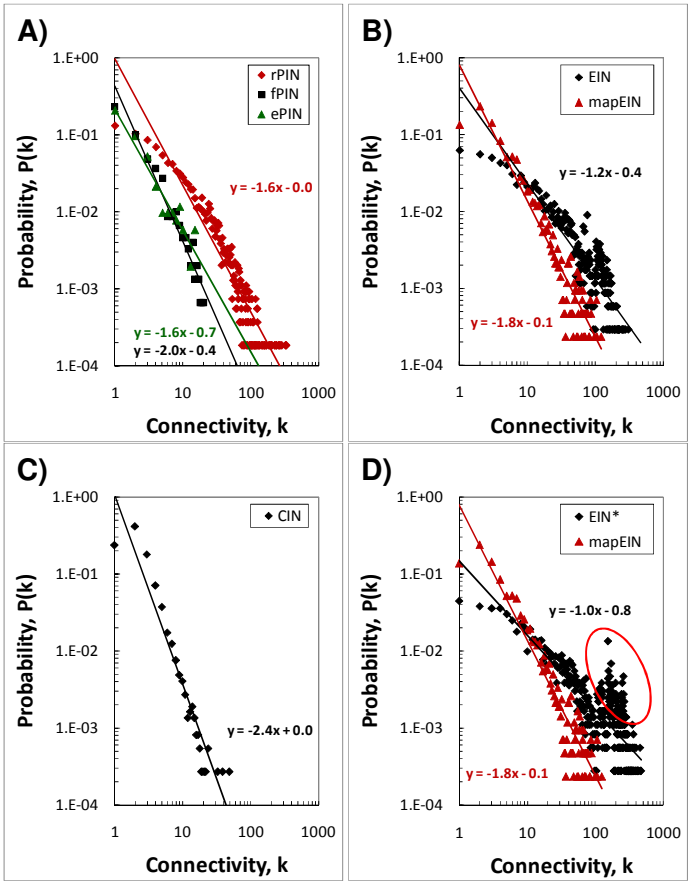


Figure 3.1: Connectivity distribution *P(k)* for the raw (rPIN), filtered (fPIN) and enzymes-only ePIN (A) as well as EIN, EIN from KEGG-Maps, mapEIN (B) and CIN (C). All distributions show approximately power law behavior $P(k) \approx k^{-\gamma}$, where k is the degree of a node, and the slope γ representing the scale-free exponent given in the graphs with some noteworthy deviations for the rPIN and the EIN. The EIN*, a variant of EIN that includes interactions derived from relations (D) between small ubiquitously occurring molecules such as H^+, NH_3, H_2O, CO_2 and metals as well as co-enzymes and co-substrates, such as CoA, NADH+, FAD, SAM (see Appendix A for complete list), is characterized by a distribution deviating from the power law distribution (red oval) for high connectivity values, k. The solid lines correspond to linear regression lines applied to the raw data in log-log scale with the associated linear equation indicated in the graph.

To analyze aspects of the protein interaction network that are specifically associated with metabolic functions, we identified proteins of rather non-metabolic functions and processes and their associated interactions. The rPIN comprises 1,186

proteins related to DNA processing functions with 21,952 associated interactions, 297 protein-degradation related proteins involved in 4,999 interactions, and 267 kinase-phosphatase associated proteins with 8,251 associations as well as 2,300 other-non-metabolic rather unspecific proteins involved in 34,230 interactions. All these interactions were partially overlapping as proteins from different groups were also reported to interact. After removing these interactions, the remaining nodes span a graph of 1,517 proteins, which can be considered to be the key molecular components responsible for maintaining the metabolic machinery. We will refer to this graph as the *filtered* PIN (fPIN). Of the 1,517 proteins, 522 represent enzymes annotated with an EC-number. The fPIN comprises 1,086 links, with 289 interactions between enzyme pairs. One third of all nodes are included in the graph's giant component, the largest connected sub-graph. We left unconnected nodes in the graph as the absence of interactions of such proteins may also be significant. In comparison to the raw network, the number of enzymes (869 in the rPIN) is lower, because in the fPIN, non-metabolic enzymes such as protein kinases and protein phosphatases have been excluded.

Further removal of proteins not assigned to at least one EC-number led to the enzyme-only-PIN (ePIN), a graph comprising only enzymes and the interactions between them. Its giant component contains 19% of all nodes. Thus, with applied filtering, the PIN became progressively disintegrated.

As observed for the rPIN, and even more convincingly, the connectivity distribution of both networks, the fPIN and ePIN, follows a power law behavior with respective scale-free exponents of 2.0 for the fPIN, and 1.6 for the ePIN (**Figure 3.1**A). Compared to the rPIN and explained by the removal of many non-specific interactions, the fraction of highly-connected nodes is reduced in the fPIN and ePIN with a simultaneous increase of unconnected nodes. The characteristic length (*CL*) of the fPIN is 8.16 and for the ePIN 6.22, which is approximately twice as long as the *CL* associated with rPIN (3.49) suggesting that, in particular, highly connected nodes providing shortcuts have been removed in the fPIN and ePIN compared to the rPIN even though the networks as such are smaller as nodes have been deleted. Note that impossible paths (no connection between nodes) have not been included in the calculation of *CL*.

The average cluster coefficient of the rPIN was determined as 0.16, 0.39 for the fPIN, and 0.41 for the ePIN indicating increased modularity of the two filtered PINs compared to the raw protein interaction network. While the rPIN shows a negative correlation of degrees associated with neighboring nodes, i.e. it is dissortative, the fPIN and the ePIN revealed a positive correlation. The assortativity (r_d) was calculated as -0.11 for the rPIN, 0.15 for the fPIN, and 0.16 for the ePIN. All correlations are significant with associated p-values of 1.0E-101, 1.0E-6, and 0.005, respectively. The negative correlations in the rPIN can be explained by the high dissortativity of protein sub-networks that have been discarded in the filtered PINs. The graph comprising relations

between kinase-phosphatase associated proteins shows a dissortativity of -0.36, DNA-related proteins of -0.12, protein-degradation -0.26 and other-non-metabolic proteins of -0.22. Consistent with these findings, the distribution of the Neighbors' Connectivity increases with increasing connectivity of nodes for the fPIN and ePIN, albeit moderately - yet significantly, and decreases for the rPIN with increasing degree of nodes (**Figure 3.2**A).

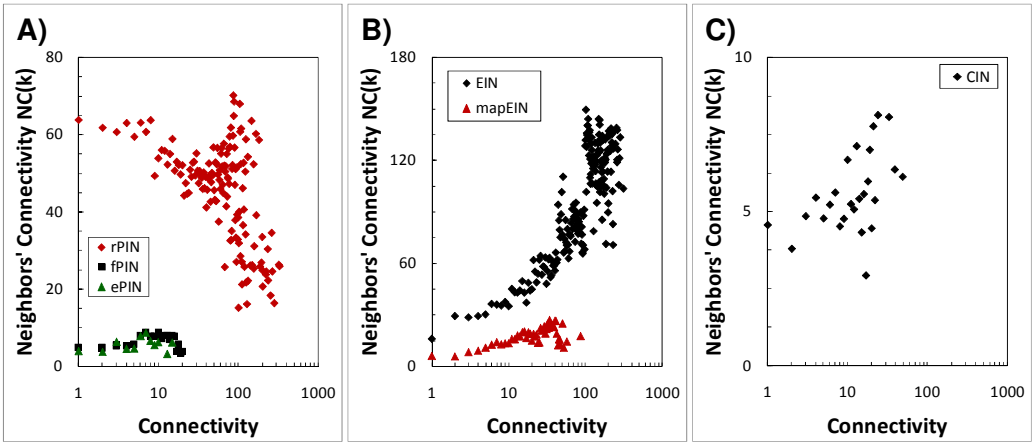


Figure 3.2: Neighbors' Connectivity (NC) as a function of connectivity of the respective reference node for the (A) rPIN, fPIN and only ePIN as well as EIN, (B) mapEIN, and (C) CIN. Plotted points correspond to mean values at a particular connectivity value. Linear regression analysis applied to the raw NC-connectivity value pairs yielded for the rPIN (Pearson correlation coefficient, *r*, and associated p-value, *p*): *r*=-0.10, *p*=9.7E-15; fPIN: *r*=0.16, *p*=7.1E-6; ePIN: *r*=0.21, *p*=1.3E-3; EIN: *r*=0.70, *p*=0; mapEIN: *r*=0.47, *p*=2.6E-109; and CIN: *r*=0.08, *p*=1.9E-7. Note: the number of raw values per plotted mean value decreases rapidly with increasing connectivity. Thus, the apparent cluster of points in the rPIN near connectivity values near 100 is not statistically significant.

Thus, the organizing principles governing protein interactions between proteins involved in metabolic functions appear to be different than for other functional categories. While PINs generally are dissortative, protein interactions associated with metabolic functions appear to be assortative; i.e. enzymes preferentially interact with other enzymes of similar degree (connectivity).

The Metabolic Interaction Networks (MIN)

We analyzed three different realizations of metabolic interaction networks (MIN) each representing metabolic pathways from a different perspective. The first two representations of metabolic interaction networks are the Enzyme Interaction Networks (EIN) and the EIN derived from KEGG pathway maps (mapEIN), where the nodes of the graph are enzymes with assigned EC-numbers. In the EIN, two enzymes are linked if

they are associated by at least one product-substrate relationship. For constructing the mapEIN, we extracted relations from KEGG pathway maps directly rather than scanning reaction lists for product-substrate relationships as done for the EIN. While the EIN comprises a large number of enzymes and their relations, the mapEIN may capture better the established biochemical knowledge of metabolic pathways. The Compound (Metabolite) Interaction Network (CIN) represents a third representation of MINs. In this graph, nodes are metabolites, and links are drawn between them if they are connected by at least one reaction.

The EIN comprises 3,435 nodes representing unique EC-numbers. The connectivity distribution, *P(k),* of the graph follows approximately a scale-free distribution with an estimated scale-free exponent γ of 1.8 (**Figure 3.1**B). As observed for the rPIN, a deviation from a simple power law behavior is evident (see above). However, the distribution follows a power law only if small ubiquitously occurring, so-called currency metabolites, such as H^+, NH_3, H_2O, CO_2, and metal ions as well as co-enzymes and co-substrates, like CoA, NADH+, FAD, SAM are excluded (**Figure 3.1**D). Including these compounds significantly increases the degree of the enzymes interacting with them resulting in a distribution *P(k)* deviating from the power law distribution for high connectivity values (**Figure 3.1**D). Upon including currency metabolites, the total number of edges increases from 60,622 to 140,260 and characteristic length, *CL*, decreases from 3.64 to 3.00.

The mapEIN comprises 1,957 nodes connected by 6,395 relations. The scale-free exponent of the connectivity distribution, γ, was computed as 1.2 with an increased probability of nodes to be less connected as compared to the EIN, where many more relationships between enzymes are possible simply via their possible substrate-product relationships. The *CL* of the mapEIN network was determined as 6.62.

The third representation, the CIN, comprises 3,702 metabolites connected by 4,868 links. As done for the EIN, the currency metabolites, co-enzymes and co-substrates have been removed prior to analysis. The connectivity distribution of the CIN exhibits a scale-free exponent, γ, of 2.4 and CL of 12.3 (**Figure 3.1**C).

All three MIN graphs are assortative with assortativity values, r_d, of 0.43, 0.26, and 0.09 in the EIN, mapEIN, and CIN, respectively. Correspondingly, an increasing Neighbors Connectivity, *NC(k)*,was observed for increasing connectivity, *k* (**Figure 3.2**B,C).The high assortativity value for the EIN probably results from the construction procedure. The EIN was constructed by scanning for product-substrate relationships. As reactions are generally treated as reversible, so that the lists of substrate and products are interchangeable, all enzymes sharing a metabolite may be linked through substrate-product relations and form a complete sub-graph. While the high assortativity of the EIN may originate from the reconstruction method possibly resulting in too many connections, this may not be the case for the mapEIN as the interactions have been

curated manually. However, many reactions in the KEGG-maps are known to be performed by isoenzymes carrying different EC numbers. Since reactions are treated as reversible, isoenzymes will be considered connected as the product of one isoenzyme can be the substrate of another, even though it is the same reaction they are catalyzing. Thus, a set of isoenzymes will form a fully-connected sub-graph, also including the enzymes of the preceding or subsequent reaction step as each isoenzyme is connected to them. The reconstruction of the CIN avoids this problem. This third representation of MINs is closest to the biological and intuitive understanding of metabolic pathways. A pathway in this sense is the path from a first substrate to a final product. The difference in the respective construction methods is also reflected by the average clustering coefficient (CC), where the CC for EIN was 0.67, EIN from KEGG-Maps 0.47, and 0.06 for the CIN, respectively.

A summary of global network properties for the PINs and MINs investigated in this study is provided in **Table 3.1**.

Table 3.1: Summary of global network properties associated with the different types of PINs and MINs investigated in this study

	PIN			MIN		
	rPIN	fPIN	ePIN	EIN	mapEIN	CIN
Number of nodes	5438	1517	522	3435	1957	3702
Number of edges	39766	1086	298	60622	6395	4868
Number of enzymes	869	522	522	3435	1957	n.a.
Giant component	5415	510	99	3276	1674	3374
Characteristic length, CL	3.49	8.16	6.22	3.64	6.62	12.30
<Cluster coefficient>	0.16	0.39	0.41	0.67	0.47	0.06
<Neighbors' connectivity>	57.17	2.65	1.97	55.60	10.34	4.41
Assortativity	-0.11	0.15	0.16	0.43	0.26	0.09

Brackets indicate mean values. Properties are explained in detail in the Materials and methods section.

3.2.2 Correlation of Protein Interaction Networks (PINs) and associated Metabolic Interaction Networks (MINs)

Nodes in the EIN and mapEIN represent enzymes. It is therefore possible to link enzymes found in PINs to the EIN and mapEIN via their annotated EC numbers. Enzymes from PINs can be linked to metabolites from the CIN network via the enzyme (EC number) -substrates and -product relationships. Thus, it is possible to directly relate network distances of proteins (enzymes) across both network types (PINs and MINs) allowing us to study how metabolic network or pathway distances are reflected in Protein Interaction Networks.

We evaluated the distribution of the shortest paths between distance pairs in the PINs and MINs, comparing the actually observed distribution to distributions generated

by 1,000 randomly constructed networks (see Materials and methods). The over- or under-representation of the distances were judged by the z-score of observed frequencies (**Figure 3.3**). We applied the analysis to all PINs and related them to the EIN, the mapEIN, and the CIN.

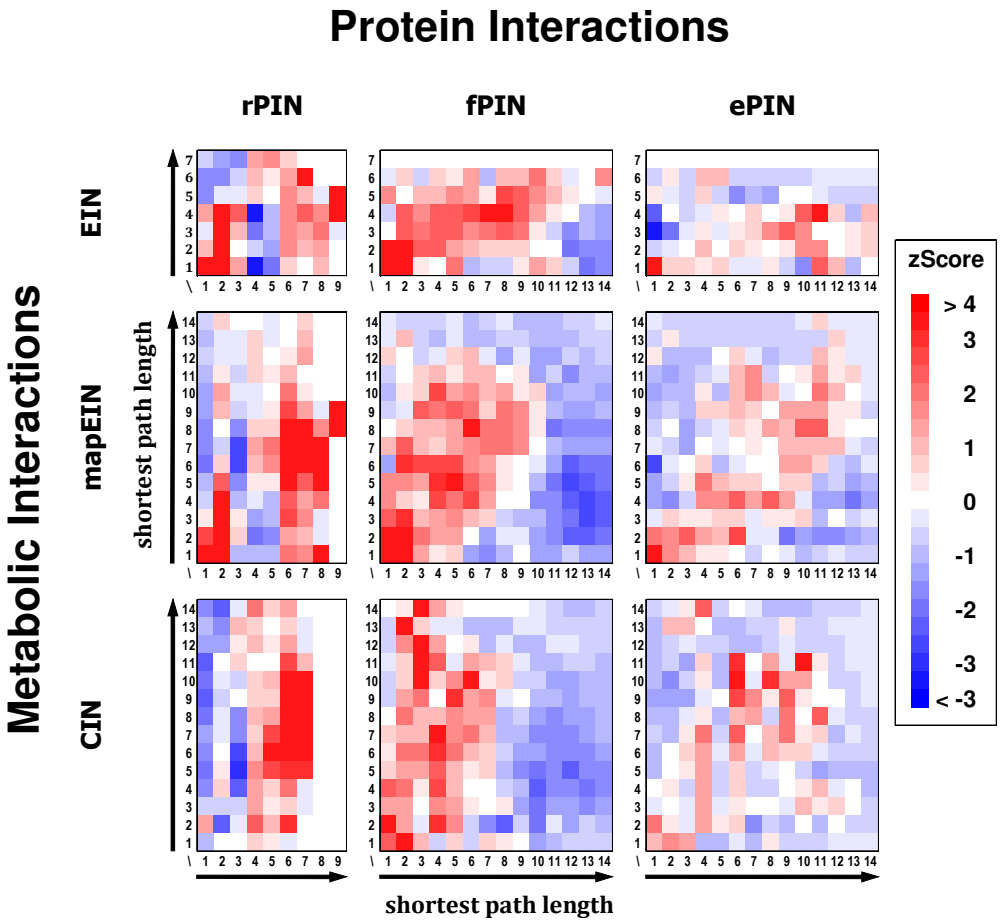


Figure 3.3: Enrichment and depletion of the abundance of shortest path pairs between nodes represented in both the PIN and MIN. The enrichments and depletions were judged by the *z*-score of the frequency of observations in comparison to randomized distribution with red-color indicating enrichment and blue-color depletion relative to randomized networks.

A direct correspondence between the protein interaction networks and metabolic networks; i.e. the physical organization of enzyme interactions follows directly its reaction pathway network, would be reflected by red-colored squares - indicating increased occurrence compared to random expectation - along the diagonal in the pair-distance matrices shown in **Figure 3.3**. Indeed, the distributions of the enrichments and depletions of the distance-pairs reveal an overall correlation of the shortest paths in PINs and MINs. All PINs show a strong enrichment of direct interactions; i.e. distance 1, in relation to the EIN and mapEIN. Furthermore, an overall correlation of distance pairs with

increased numbers of observations relative to the random background (red squares along the diagonal, blue squares primarily off-diagonally) for all PIN-MIN comparisons is evident, especially for the fPIN (**Figure 3.3**, central column). Interestingly, enzymes appear more closely related (shorter distances between them) in the fPIN in comparison to their distances derived from their metabolic network association (mapEIN and CIN), as the off-diagonal pattern of red-colored squares indicates a skewed distribution towards larger shortest paths between linked proteins in the MIN compared to their distances in PINs. Thus, it appears that enzymes catalyzing enzymatic steps of some medium distance, i.e. not directly subsequent to each other, are physically brought into contact via protein-protein interactions involving proteins that are not necessarily directly participating in the actual metabolic pathway. For the EIN, above-diagonal enrichments were not observed when compared to the fPIN and ePIN, possibly a consequence of the network reconstruction procedure that allows very many interactions leading to a highly connected metabolic network. As the fPIN contains both enzymes and structural proteins, some proteins included in this network may function as connector or bridging proteins holding distant parts of metabolic pathways together. In the ePIN, such proteins have been filtered out leaving only enzymes in the PIN. Here, the enrichment pattern follows the main diagonal, but at weaker significance as the absolute numbers are smaller. A direct comparison of shortest paths between enzyme pairs connected via valid paths in both the fPIN and the ePIN yielded a mean distances of 5.3 for the fPIN, and 6.2 for the ePIN (p=0; paired, two-tailed t-test, N=5,531). Thus, metabolic enzymes are brought into spatial proximity - by way of protein-protein interactions – via interactions mediated by non-metabolically active proteins.

		PIN		
		rPIN	**fPIN**	**ePIN**
MIN	**EIN**	0.07	0.18	0.35
	mapEIN	0.09	0.21	0.44
	CIN	0.09	0.14	0.22

Table 3.2 Correlation between pairwise network distances of proteins in MINs compared to their respective distance in PINs. Listed are the Pearson correlation coefficients, *r*. All correlations were highly significant ($p <$ 1.0E-40).

The overall Pearson correlation values, *r*, for distance-pairs (PIN, MIN) are listed in **Table 3.2**. All correlations are highly significant (p < 1.0E-40). Thus, in all comparisons, a positive correlation of the organization of protein-protein relations was observed between their enzymatic pathway organization and their corresponding physical organization. The correlation is strongest when enzyme-only protein networks are compared to MINs, in particular to KEGG-map derived metabolic pathways (mapEIN). The PIN-MIN correlations were observed to become more pronounced, when more

relevant (with regard to metabolism) PINs were considered and increase steadily from rPIN to fPIN, with greatest correlations observed for the ePIN. Thus, it is no contradiction that the correlations for the rPIN are low, but a result, because many unspecific interactions included in the rPIN were eliminated in the other PINs. The reported correlation coefficients (**Table 3.2**) were computed over the entire range of network distances including distant pairs for which correlations can be expected to be low. Correspondingly, correlation values increased significantly when remote pairs were discarded (Appendix B).

3.2.3 Correlation of metabolic fluxes carried by enzymes and their Protein Interaction Network properties

On the basis of measured relative metabolite flux rates of yeast growing in a glucose medium, we evaluated the correlation of network cluster coefficients of the involved enzymes in PINs to the flux rate carried by the enzymes. The flux rates were estimated by Blank and colleagues based on a global metabolic network model of *S.cerevisae* [79] and a flux balance analysis based on large-scale ^{13}C-isotope tracer experiments [80]. Our analysis revealed high PIN clustering coefficients for high flux enzymes decreasing with the decrease of the relative flux rates (**Figure 3.4**). Further analysis revealed that the connectivity as well as the betweenness-centrality are also positively correlated with the flux rates carried by the associated enzymes (**Table 3.3**). Thus, highly connected and central enzymes (in PINs) are enzymes carrying high fluxes. Furthermore, enzymes preferentially interact with enzymes of similar flux rates. A strong positive correlation of flux rates of physically interacting proteins was observed (correlation coefficient of 0.52) in the fPIN and ePIN.

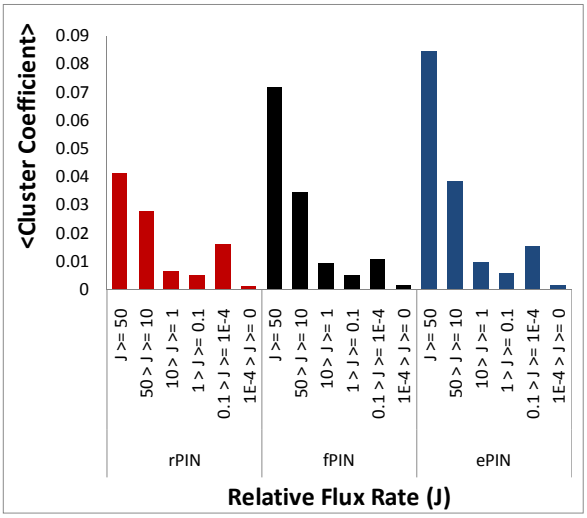


Figure 3.4. Average PIN cluster coefficient of enzymes binned by relative flux rates, *J*.

Table 3.3: Pearson correlation coefficient, *r*, and associated *p*-values between connectivity (degree), centrality as judged by betweenness of enzymes (nodes in PINs) and the respective relative flux rates, as well as the correlation coefficient of the relative flux rates between neighboring (physically interacting) enzymes in the three PINs examined in this study.

Correlation between	rPIN		fPIN		ePIN	
	R	p-Value	r	p-Value	r	p-Value
Connectivity~Flux rate of nodes	0.15	4.76E-06	0.23	1.08E-04	0.25	8.99E-05
Centrality~Flux rate of nodes	0.14	2.16E-05	0.11	1.24E-02	0.32	9.34E-13
Flux Rate~Flux rate of neighbors	0.24*	5.25E-25	0.52	1.20E-41	0.52	1.45E-40

*Correlation coefficient, *r*, differs in rPIN from values obtained in fPIN and ePIN, because enzyme pairs are included in the rPIN that were filtered out in the other two PINs because of inconsistent subcellular localization annotation. They are identical for the fPIN and ePIN as the two networks contain the same set of enzymes.

3.2.4 Physical interactions in high-throughput catabolic pathways and synthesis pathways of complex metabolites

To gain further insight, we studied the physical organization of enzymes carrying high fluxes in greater detail. The large-scale flux analysis in yeast by Blank and co-workers [80] comprised 1,038 reactions (745 distinct reactions) encoded by 672 genes of which 610 can be found in the rPIN. Of the distinct reactions, 28% (208 reactions) have reaction rates greater than 1 relative to a glucose uptake of 100 (arbitrary flux units) and can be summarized in a global glucose utilization scheme (**Figure 3.5**A). In this scheme, 16 reactions, 2.1% of the 1,038 reactions considered by Blank, have flux rates greater than 50, corresponding to 61 proteins found in the PIN and were contained in all three PIN variants studied here. The reactions include two transport reactions of the products of the fermentative glycolysis with no annotated gene assigned to these steps, the glucose uptake, and the reaction performed by the ATPase complex. The remaining reactions are involved in the fermentative glycolysis as shown in **Figure 3.5**B. Glycolysis describes the utilization of glucose as an energy source upon its degradation to pyruvate. Depending on the culture conditions, pyruvate may either be fully degraded to CO_2 by the enzymes of the TCA-cycle within the aerobic glycolysis, where the pyruvate dehydrogenase (PDH) connects glycolysis with the TCA-cycle enzymes, or to ethanol by pyruvate decarboxylase (PDC) and alcohol dehydrogenase (ADH) within the fermentative glycolysis, when O_2 is limiting. The enzymes of the fermentative glycolysis are highly interconnected with each other by many physical interactions detected between the associated enzymes (**Figure 3.5**B). An exception is the pyruvate kinase (PYK1, CDC19) which is not physically linked to any of the other enzymes of the pathway, as well as the 6-phosphofructokinase (PFK). Instead, PYK interacts with PDH.

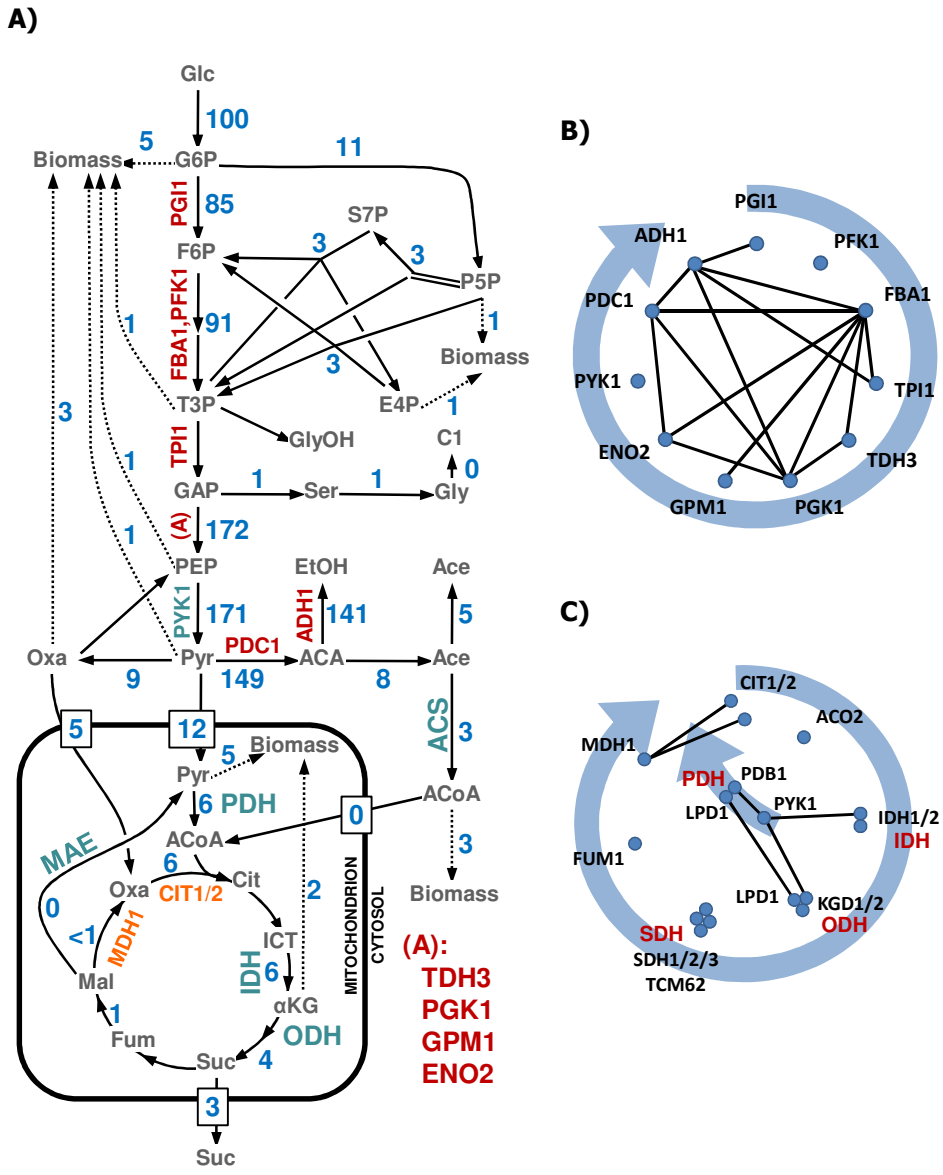


Figure 3.5: (A) Generalized view of the utilization of glucose by yeast growing in a glucose medium. (B) Interaction cluster of enzymes involved in fermentative glycolysis. (C) Interaction clusters of enzymes of the TCA-cycle as well as the prior reactions of the pyruvate kinase (PYK1) and pyruvate dehydrogenase (PDH). Figure (A): Only reactions with relative flux-rates greater than 1 relative to a glucose uptake of 100 (arbitrary flux units) are shown (blue numbers). The protein names represent enzymes responsible for the respective reaction. Three notable interaction clusters of enzymes can be found indicated by orange, cyan, and red color. Figure (B) and (C): The enzymes are ordered in clockwise direction according to the catabolic step during the process. Black edges indicate interaction in the protein-protein interaction network.

Only a minor fraction of pyruvate (flux rate of 6 relative to the glucose uptake rate flux rate set to 100) appears to be channeled to the TCA-cycle, that is 3% (as one glucose molecules may lead to the formation of two pyruvate molecules) of the initial

glucose influx is processed by the TCA-cycle enzymes. The production of pre-stage substrates of amino acids rather than energy production is the main function of the TCA-cycle. The flux rates decrease to 1 beyond the succinate dehydrogenase (SDH) reaction step. This path leads through the PYK1, PDH and the following enzymes of the TCA-cycle: citrate synthase (CIT), isocitrate deyhdrogenase (IDH), 2-oxoglutarate complex (KGD), succinyl-CoA synthetase (LSC) and SDH. The enzymes of the TCA exhibit a relatively low number of physical interactions (**Figure 3.5**C). The interactions are mainly pooled in enzyme complexes, SDH (SDH1/2/3, TCM62), the LSC (LSC1/2) and KGD (KGD1/2, LPD1), performing the individual reactions steps of the TCA-cycle. Only the reactions of the malate dehydrogenase (MDH1) and the CIT are physically connected. However, taking the prior reactions of the PYK1 and PDH into account, the TCA reactions reveal a more dense interaction cluster. The PDH interacts with KGD sharing the common subunit lipoamide dehydrogenase (LPD1). The PYK1 interacts with PDH, KGD as well as IDH (**Figure 3.5**C).

The remaining reported direct physical interactions contained in the fPIN between enzymes detected within metabolic pathways are distributed throughout anabolic pathways. Most interactions are found in the biosynthesis of ergosterol, ubiquinone, sphingolipid and glucogen synthesis. Single links between enzymes can be found in biosynthetic pathways of pyrimidine, leucine, isoleucine, and lysine. Within the fatty acid synthesis pathway, the malic enzyme (MAE1) interacts with the alpha subunit of FAS and Acetyl-CoA carboxylase (ACC1) (Appendix C).

Figure 3.5B, C and Appendix C provides a comprehensive account of all reported protein interactions mapped to canonical metabolic pathways from the SGD database; i.e. for pathways not included in this figure, no protein interaction was contained in the PIN.

3.2.5 Central proteins in the fPIN

Analyzing centrality as judged by the z-score of the change of the characteristic length of the graph after removal of a particular node identified enzymes with the most influence on the cohesion of the interactome. The ten most influencing proteins are listed in **Table 3.4**. ATP14 exhibits the most influence on the characteristic length of the fPIN. The H-chain of the ATP synthase is one of 17 polypeptides building up the complex (Figure 3.6A). While only interacting with a relatively low number of other subunits of its own complex, it interacts with the Complex IV (Cytochrome c) of the respiratory chain, via COX5B and Complex III (Cytochrome b-c1) via QCR8. Furthermore, it interacts with the FBA1 from the glycolysis pathway, a central enzyme assembling the glycolytic cluster.

Table 3.4 Ten most central proteins in the fPIN as judged by the *z*-score of the change of the characteristic length of the graph after removal of a particular protein, robustness-centrality (RB)

Protein	RB	BN	DC	Protein description
ATP14	17.17	25.44 (1)	5.14 (14)	ATP synthase H chain, mitochondrial
FBA1	11.34	6.71 (6)	4.45 (20)	Fructose-bisphosphate aldolase
TSC13	8.49	10.28 (3)	3.07 (33)	Enoyl reductase, very long fatty acid elongation
IFA38	7.73	1.26 (37)	8.59 (4)	Oxidoreductase, fatty acid elongation
COX1	7.69	4.43 (14)	5.14 (15)	Cytochrome c oxidase subunit 1
SER3	7.49	-0.32 (1449)	5.83 (8)	D-3-phosphoglycerate dehydrogenase 1
COX5B	7.19	3.53 (17)	1.00 (117)	Cytochrome c oxidase chain Vb, mitochndr.
URA2	6.74	2.72 (21)	3.07 (31)	Bifunctional glutamine-dependent carbamoyl-phosphate synthase, Aspartate carbamoyl-transferase
GPM1	6.60	2.94 (20)	2.38 (42)	Phosphoglycerate mutase 1
MAE1	6.46	9.46 (4)	0.31 (160)	NAD-dependent malic enzyme, mitochondrial

For comparison to other centrality measures, the *z*-scores and respective rank (in parentheses) of the betweenness-centrality (BN) as well as degree-centrality (DC) were evaluated. The overall correlation between the centrality was RB~BN 0.70 (*p*-Value: 4.35E-220), RB~DC 0.59 (*p*-Value: 1.44E-144) and BN~DC 0.25 (*p*-Value 3.37E-23).

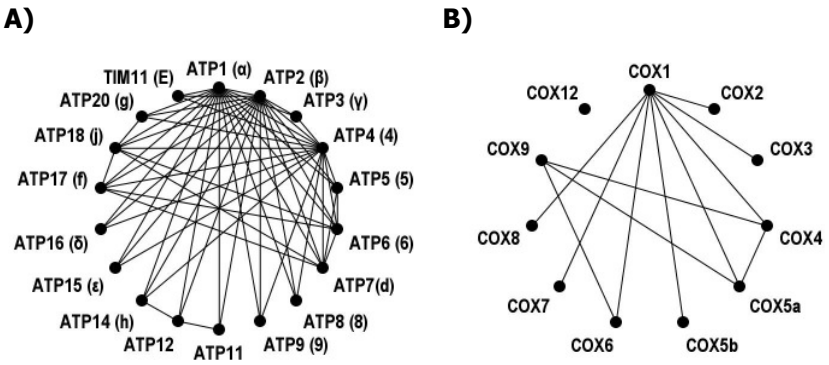


Figure 3.6: The physical interaction clusters of the A) ATP synthase, and B) Cytochrome c. The identity of enzymes is given by their gene symbols and detected interactions between them denoted by solid lines.

COX1 and COX5B are two of 11 subunits of Cytochrome b-c1 (**Figure 3.6**B). While COX1 plays an essential role in the assembly of the complex, the role of COX5b is the interaction with ATP synthase. The FBA1 and GPM1 are part of the glycolytic cluster. Taken together, the glycolysis pathway and the respiratory chain are tightly connected via physical interactions illustrated in **Figure 3.7**A.

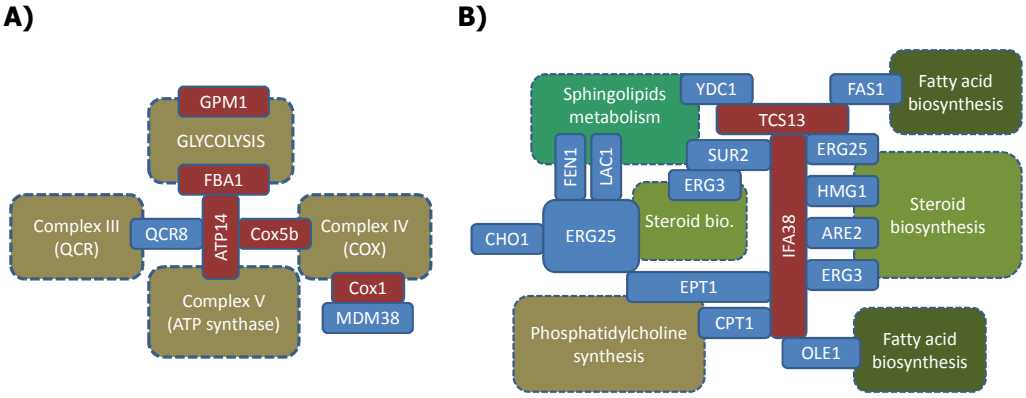


Figure 3.7 Crosstalk via protein interactions between enzymes and proteins in A) glycolysis and the respiratory chain (ubiquinol cytochrome-c reductase complex (Complex III); cytochrome c oxidase (Complex IV) and ATP synthase (Complex V)) as well as B) systems for biosynthesis of membrane lipids (biosynthetic systems for sphingolipids, fatty acid, steroids as well as for phosphatidylcholine). Individual proteins are highlighted in blue. If belonging to the group of 10 most central proteins of the fPIN (see Table 3.4), they are colored red. Boxes with dashed perimeter indicate larger metabolic systems. Touching protein boxes indicate interaction between the proteins, whereas protein boxes emerging from metabolic systems denote participation of that protein in this system.
(ARE2: Acyl-CoA:sterol acyltransferase; ATP14: H chain of ATP synthase; CHO1: phosphatidylserine synthase; Cox5b and Cox1: Subunit Vb and I of cytochrome c oxidase; CPT1: Cholinephosphotransferase; EPT1: sn-1,2-diacylglycerol ethanolamine- and cholinephosphotranferase; ERG25: C-4 methyl sterol oxidase; ERG3: C-5 sterol desaturase; FAS1: Beta subunit of the fatty acid synthase; FBA1: Aldolase; FEN1: Fatty acid elongase, involved in sphingolipid biosynthesis; GPM1: phosphoglycerate mutase; HMG1: HMG-CoA reductase; IFA38: oxidoreductase, fatty acids elongation; LAC1: Ceramide synthase; MDM38: required for k+/H+ exchange; OLE1: Delta(9) fatty acid desaturase; QCR8: Subunit 8 of ubiquinol cytochrome-c reductase complex; Sur2: Sphinganine C4-hydroxylase; TSC13: enoyl reductase, very long fatty acids elongation; YDC1: Alkaline dihydroceramidase, sphingolipids metabolism.)

TSC13 and IFA38, which are responsible for the elongation of very long fatty acids, connect enzymes involved in the biosynthesis of membrane lipids by interacting with enzymes from the biosynthesis of fatty acids, steroids and related metabolites, phosphatidyl -choline, -serine and -ethanol amine, suggesting that the pathways are brought into spatial proximity via protein-protein interactions (**Figure 3.7**B).

For comparison, other centrality measures for the top 10 most influencing proteins are listed in Table 3.4. While an overall correlation between the centrality measures is evident (correlation coefficients and associated *p*-values are provided in the legend of Table 3.4), each centrality measure identifies particular aspects of centrality and does not correspond directly to the robustness measure used here.

3.3 Discussion

Our investigations integrated protein interaction networks with metabolic networks to study the extent to which metabolic pathways; i.e. functional processes, are pre-formed in the underlying structural interaction network, i.e. the "plumbing" of cellular components. The networks examined here were derived from different sources of information and provide different views on the metabolic as well as protein interaction systems.

We discovered that sub-systems of the entire protein-protein interaction network may follow specific organizing principles. While interactions associated with signaling and other regulatory processes (e.g. transcriptional regulation via DNA-interaction associated proteins) were found to be dissortative; i.e. proteins of high degree interact with proteins of low degree, interactions between metabolic enzymes were observed to be assortative such that enzymes frequently interact with other enzymes of similar degree (**Table 3.1**). Regulatory processes may often involve hierarchical one-to-many associations such as master regulators (e.g. kinases) and their respective individual target proteins. Physical interactions between metabolic enzymes, on the other hand, appear to generally follow a more horizontal organization with enzymes participating in larger complexes or sequential one-to-one interaction chains. Nonetheless, we identified interaction hub enzymes that are located at central positions integrating several metabolic systems and whose removal would severely impact the structural integrity of larger portions of the metabolism-focused metabolic network (fPIN, **Table 3.4**, **Figure 3.7**A,B).

When dealing with characteristics of protein-protein interaction network, possible technological as well as biases introduced by targeted scientific interest always are a concern. To best avoid this problem, it would be ideal to use strictly unbiased datasets for analysis. However, such fully unbiased datasets are not available (yet) as this would require nothing less than an identification of all true and relevant protein-protein interactions occurring inside cells. Presently, we have to resort to the best available unbiased datasets generated by high-throughput screens. As the BIOGRID data contains information about the source of information, it is possible to evaluate the assortativity of the biggest subsets in the database that were obtained from high-throughput experiments, namely Krogan et. al [81] comprising 1,669 nodes involved in 2,682 interactions, and Gavin et al. [82] with 2682 nodes involved in 8,138 interactions. Reducing the filtered PIN to these subsets yields two sub-fPINs comprising 364 nodes involved in 411 interactions, and 757 nodes involved in 475 interactions, respectively. The reduced fPINs exhibit an assortativity of 0.31 and 0.15, respectively, confirming the results obtained for the whole fPIN. Correspondingly, for the rPIN, a reduced assortativity was obtained for both datasets with 0.15 for the Gavin set (N_{nodes}=1,669, $N_{interactions}$=10,992) and -0.01 (N_{nodes}= 2,682 and $N_{interactions}$=8,138) for the Krogan set. Thus, given the

available datasets, the increased positive assortativity of filtered/enzyme PINs does not appear to be resulting from a bias towards well-studied enzymes.

We generated three different versions of the metabolic interaction network (MIN). The enzyme interaction network, EIN, was introduced to capture all possible metabolic interactions between enzymes, whereas the mapEIN transformed the pathway knowledge available in KEGG into a metabolic network. The compound interaction network, CIN, was created as an alternative and focuses on main metabolites as network nodes rather than enzymes. With regard to our main research focus, the topological equivalence of protein interactions and metabolic pathways, all three versions yielded significant positive correlations between the respective shortest paths across both network types (**Table 3.2**). Thus, the reported results proved robust against details of the network reconstruction approach. All three MIN-versions were reported here with positive assortativity (**Table 3.1**) while a negative assortativity of metabolic networks has been reported elsewhere (degree correlation coefficient of -0.24 [83]). We note that this difference is caused by the elimination of ubiquitous (currency) metabolites and the inclusion of only main metabolic substrates and products in this work. Including all metabolites in the CIN yielded a degree correlation coefficient $r_{d,}$ of -0.3 and an increased mean cluster coefficient of 0.7. As currency metabolites such as ATP follow a one-to-many network motif, thereby also introducing many more edges in the network, the dissortativity obtained when including them as well as the increased mean cluster coefficient can be rationalized.

The decision on the exact procedure to generate metabolic networks must remain operational and may dependent on the objective of the study at hand. Defining metabolic networks based on carbon atomic traces in metabolic reactions resulted in different topological characteristics of metabolic networks than for the commonly used approaches [59].

The interaction networks investigated in this study vary regarding their graph-parameters, such as characteristic length (*CL*) values and scale-free exponents and also differ from some networks reconstructed. Generally, biological networks tend to be scale-free with associated scale-free exponents reported below two, which was suggested to result from evolutionary mechanism driven by gene duplication [61]. However, larger exponents have been reported for the CIN [37]. Joang et al. analyzed compound interaction networks of 43 organisms. The average *CL* was observed as 3.29 ± 0.11 and the average scale-free exponent as 2.18 ± 0.09. While the scale-free exponent reported here (2.4) is in line with the reported average value, the CL reported here is much larger (12.3). The reasons for the difference can be attributed mainly to the different approaches taken to reconstruct the CIN, and the filtering mechanisms applied to remove compounds that are not directly relevant for main biochemical pathway routes such as co-factors. Here, we followed the concept of main metabolite relations introduced by Kotera and co-workers

and annotated accordingly in the KEGG database [54; 55]. By contrast, the networks of Joang et al. comprised all relations between all substrates and products, including currency metabolites such as H_2O or ATP rendering the *CL* much smaller.

In their analysis of the E.coli metabolic pathway network, Wagner and Fell reported a mean *CL* of 3.8. This value compares favorably with our value (3.64) for the yeast EIN, which corresponds to the network analyzed by Wagner and Fell [38]. A similar value has also been reported by Huthmacher et al. (*CL*=3) [47]. Similarly, the scale-free exponents agree well (-1.3 Wagner and Fell; -1.2 reported here for yeast). Kotera et al. [55] reported a *CL* of 9 for the equivalent of our CIN. The larger value we obtained (*CL*=12.3) is explained by the exclusion of currency metabolites in our analysis. The newly introduced mapEIN (CL=6.62) is not directly comparable to previous studies. It was constructed to capture our accumulated knowledge of biochemical pathways represented in KEGG and with nodes represented by enzymes, not compounds.

For the rPIN, our reported values for graph properties such as *CL* and scale-free exponent agree well with previously reported data [50; 51]. For the other PIN types studied by us, no comparative data are available.

In their analysis of protein interaction data in the context of metabolic pathways, Huthmacher and co-workers [47] focused on direct interactions between enzymes catalyzing consecutive metabolic reaction steps. Here, we expanded the scope of an integrative analysis by also showing that such correlation between metabolic and protein interaction data is discernable even at larger distances. Of course, an increased probability for consecutive enzymes to interact naturally leads to correlations at larger distances as well, even though the significance can be expected to drop. We showed that such large-scale topological correspondence between both the PIN and MIN indeed exists adding further evidence for the significance of physical interactions for the functioning of metabolic reactions. Our analysis also revealed that shortest paths between two enzymes appear to be shorter in the PIN compared to their distance when analyzed in the metabolic network (**Figure 3.3**, elevated z-scores above the diagonal), especially when the allowed physical interactions also include proteins not actively participating in enzymatic reactions (fPIN). A direct comparison of shortest paths between enzyme pairs connected via valid paths in both the fPIN and the ePIN yielded a mean distances of 5.3 for the fPIN, where non-metabolically active proteins are still included, compared to 6.2 for the ePIN. Thus, our analyses suggest that such metabolically passive proteins may function as interface components to spatially organize enzymatic pathways.

The functional significance of topological parameters of molecular networks has largely been analyzed within the context of the examined network type itself such as the reported relationship between fluxes passing through metabolic network edges (reactions) and the degree product of the connected nodes [77], but not across different network types. Here, we showed that such relationships can also be established across

different network types such that topological parameters of enzymes within the context of protein interactions have relevance for their functional, metabolic context. In particular, we observed that metabolic flux rates are positively correlated with degree and centrality of enzymes in their PIN (**Table 3.3**). We interpret this observation as evidence for a co-evolutionary adaptation of both network types. High-flux enzymes are physically interacting with many other enzymes such that metabolic substrates and products can be passed on to subsequent enzymes quickly and efficiently.

On the technical side, it has to be borne in mind that our knowledge of protein interactions is certainly incomplete and may contain many false positive interactions [84; 85] and the employed technologies may skew the datasets towards particular interactions [86]. Furthermore, since we used sub-cellular localization information to eliminate potential false positive protein associations, this information, too, is to some degree based on predictions alone and may contain erroneous assignments. However, the fact that we did observe significant correlations between protein interactions and metabolic pathways despite the noise in the data may suggest that the true topological correspondence may actually be even stronger than reported here.

3.4 Conclusions

Our results reveal topological equivalences between the protein interaction network and the metabolic pathway network. Evolved protein interactions may contribute significantly towards rendering metabolic processes more efficient by permitting increased metabolic fluxes. Thus, our results shed further light on the unifying principles shaping the evolution of both the functional (metabolic) as well as the physical interaction network.

3.5 Materials and Methods

Because yeast represents a model organism with comprehensive experimental as well as annotation data available for both protein-protein interactions as well as metabolic reaction pathways, we focus our investigations on *Saccharomyces cerevisiae.*

Protein Interaction Networks (PINs)

To study protein interaction networks (PINs) from a global perspective as well as by focusing on enzymatic proteins alone, we generated three different network graphs describing protein-protein interactions. The raw, unfiltered PIN (rPIN) was constructed by extracting physical interactions reported in the protein interaction databases DIP, version 20060402 [87] and BIOGrid, version 2.0.21 [88], respectively. Based on available gene ontology (GO) annotation information, proteins involved in processes related to protein translation, DNA-transcription and associated regulatory processes, such as transcription or translation factors, as well as proteins involved in the assembly of chromatin structures were labeled as 'DNA-related'. Proteins involved in degradation and related regulatory proteins were labeled as 'degradation-related', protein kinases and phosphatases and related regulators labeled as 'kinase-phophatase-related'. Additionally, we defined a set of proteins as 'other non-metabolic proteins'. This set comprised proteins assigned to unspecific functions and processes as judged by their available Gene Ontology (GO) annotation such as binding to unfolded proteins, protein targeting, protein transport, protein tagging as well as other post transcriptional modifications other than phosphorylation, which were labeled as kinase-phophatase-related. Proteins assigned to any of the above sets and associated physical interactions were removed from the rPIN to generate a second PIN, the filtered protein-protein interaction graph (fPIN). Thus, in the fPIN, all protein interactions of proteins involved functions other than metabolism – as judged by their GO-annotation - were removed from the rPIN. It also included proteins with currently unknown function. A third PIN including only enzymes as judged by an assigned EC number was also generated and designated as the ePIN.

Interactions with inconsistent localization annotation according to the available gene ontology, GO:Cellular-Component annotation information; i.e. interactions between proteins located in different sub-cellular compartments, were discarded from the fPIN and ePIN as well. In case of membrane-embedded proteins, interactions between proteins localized in different, but neighboring compartments were retained.

GO-Annotations for yeast genes were obtained from the SGD database [89]. The evidence codes for the gene ontology were not considered. The GO-annotation information used to sub-set the protein data is available in Appendix A.

Metabolic Interaction Networks (MINs)

Metabolic Interaction Networks (MINs) are represented in this study by Enzyme Interaction Networks (EINs) focusing on metabolic reactions as well as a Compound Interaction Networks (CIN) establishing links between metabolites directly.

The metabolic reaction lists form KEGG [90], YeastCyc [91], and the set of metabolic reactions obtained from a whole-genome metabolic reconstruction approach, in the following referred to as the Förster-Set [79], were merged and used to reconstruct the Enzyme Interaction Network (EIN). The corresponding network graph is a representation of EC numbers and associated reactions and their metabolic interactions. Two nodes are considered connected if they share at least one common substrate or product. Ubiquitously occurring molecules, so-called currency metabolites, such as H^+, NH_3, H_2O, CO_2, and metal ions as well as coenzymes and co-substrates such as CoA, NADH+, FAD, SAM have been excluded from the analysis. In total, 51 metabolites were excluded (see Appendix A for the complete list).

The applied connectivity conditions to generate the EIN may produce links that are theoretically possible, but that have not been experimentally verified yet. To reflect the available biological knowledge, we also generated a metabolic network from curated pathway maps, the mapEIN. For the reconstruction of the mapEIN, the relations between enzymes were extracted directly from the xml-description files of the pathway maps from the KEGG database. Two nodes in this graph are connected if both are associated with at least one common metabolite node in a map.

The EINs reflect relationships between enzymes or reactions, respectively. Alternatively, a metabolic network can be reconstructed considering metabolites themselves as nodes. Such a network, the compound (metabolite) interaction network (CIN), was constructed utilizing the reaction lists from KEGG. Two metabolites are considered connected if both are recognized as a main substrate-product reaction as annotated in KEGG, respectively. As for EINs, currency metabolites and co-enzymes and co-substrates have been discarded from consideration. The YeastCyc and Förster-Set was not considered for the construction of the CIN as both databases do not differentiate between main and side substrates or products, respectively.

Topological properties of networks

To characterize global as well as local properties of the molecular interaction networks analyzed in this study, we computed several well established graph-theoretic network parameters.

The characteristic length (*CL*) describes the average shortest path of a graph, i.e. the expected shortest distance between any two different nodes. The *CL* was calculated applying Equation 3.1.

Eq.3.1

$$CL = \frac{1}{|E|}\sum_{(i,j)\in E} d(i,j);$$

$$E = \{(i,j) \in \{1,2,...,N\} x \{1,2,...,N\} : \infty > d(i,j) \geq 0 \wedge j > i\},$$

where $d(i,j)$is the distance (shortest path) between nodes i and j, n is the total number of nodes, E defines the set of considered node pairs and $|E|$ is their total number. Distances between unconnected node pairs were not considered.

The connectivity distribution, $P(k)$, was calculated according to Equation 3.2:

Eq.3.2 $$P(k) = \frac{N_k}{N},$$

where k is the degree of nodes, i.e. the number of links associated with a node, N is the total number of nodes, and N_k is the number of nodes of degree k.

The directionality of links was not considered. Biological networks were shown to follow scale-freeness according to a power law degree distribution with $P(k) \sim k^{\gamma}$, where γ is the scale-free exponent [34; 35; 36; 37; 38], which was estimated by the slope of the linear regression line of degree distributions in log-log diagrams.

The cluster coefficient (c) is a measure of modularity of a graph. It measures to which degree the neighborhood of a node resembles a complete; i.e. fully connected graph. The cluster coefficient and its mean value were calculated according to Eq. 3.3 [92].

Eq. 3.3. $$c_{i\in N_{k>1}} = \frac{\sum_{r,p\in N_{k>1}} A_{i,r}A_{i,p}A_{p,r}}{k_i(k_i-1)}; \quad < c >= \frac{\sum_{i\in N_{k>1}} c_i}{|N_{k>1}|},$$

where A denotes the adjacency matrix with elements set to 1 in case of an established link between nodes and zero otherwise; k_i is the degree of node i for which c is computed, i, p, and r are indexes of all nodes in the network with $k > 1$.

The Neighbors' Connectivity $< NC(k) >$ measures the affinity of nodes of a particular degree to interact with nodes of either higher, similar, or lower degree. The Neighbors' Connectivity, NC, of a particular node is the average degree of its neighboring nodes [93]. $< NC(k) >$ is the average NC for nodes of degree k. It is an increasing function of k when a graph is assortative, i.e. high-degree nodes preferentially tend to interact with degrees of similar, high degree. The function is decreasing when high-degree nodes preferentially interact with nodes of lower degree; then the graph is said to be dissortative. Assortativity is defined as the Pearson correlation of the degrees of

neighbors, r_d. If the distribution is uniform, r_d equals zero, otherwise r_d is positive for assortative graphs or negative for dissortative graphs. The assortativity was measured according to an algorithm proposed by Newman [94] (Eq. 3.4).

Eq. 3.4.
$$r_d = \frac{|E|^{-1}\sum_{i\in E} j_i k_i - \left[|E|^{-1}\sum_{i\in E}\frac{1}{2}\left(j_i^2 + k_i^2\right)\right]^2}{|E|^{-1}\sum_{i\in E}\frac{1}{2}\left(j_i^2 + k_i^2\right) - \left[|E|^{-1}\sum_{i\in E}\frac{1}{2}\left(j_i + k_i\right)\right]^2},$$

where r_d, is the assortativity, j and k are the degrees of nodes at the ends of the ith edge within the set of considered node pairs E and $|E|$ is their total number, as notated for Eq. 3.1.

Correlation of Metabolic and Protein Interaction Networks

The PINs were related to the EIN and mapEIN via protein - EC number relations; i.e. proteins (enzymes) were identified in both network types and, subsequently, their pairwise distance computed in both network types. EC-number annotations were taken from KEGG [90], YeastCyc [91], SGD [89] and Expasy [95]. The relation of PINs to the CIN followed from indirect protein – EC-number-metabolite mappings according to the annotation information in KEGG. Nodes were considered equivalent in both network types, if for a metabolite (node in the CIN) the corresponding protein (node in the PIN) was identified via its EC number annotation and its list of main metabolites associated with the reaction catalyzed by the enzyme.

For nodes with representation in both networks, the respective shortest distances were correlated. The distribution of distances within the PINs and MINs were evaluated by consideration of all such node couples resulting in abundance matrices. The two dimensions of the abundance matrix are the respective distances in the PINs and MINs, and the elements contain the observed counts for the respective distances pairs. Note that proteins may be assigned to more than one EC number and can be represented multiple times in the EINs. Likewise, unique EC numbers may be assigned to multiple proteins. The EC-numbers may comprise multiple metabolites as well. All such possible relations between the PIN and MINs were considered.

We evaluated enrichments and depletions of particular distance fields in the abundance matrix by comparing the actual counts to counts obtained from 1,000 randomly produced PIN-MIN correlations. For the randomization, protein-names within the PIN were shuffled among the graph's nodes. In this procedure, the nodes of the PINs were randomly assigned to a protein name leading to alteration of protein – EC number relations while preserving the topology of the graphs. Statistical enrichment and

depletion of actual counts versus random expectation were judged by the z-score (Eq. 3.5) of a particular element of the abundance matrix.

Eqs. 3.5

$$\sigma_{\text{rand},d_{\text{PIN}},d_{\text{MIN}}} = \sqrt{\frac{\sum_{\text{rand}=1}^{1000}\left(n_{\text{rand},d_{\text{PIN}},d_{\text{MIN}}} - \left\langle n_{\text{rand},d_{\text{PIN}},d_{\text{MIN}}}\right\rangle\right)^2}{1000}};$$

$$zScore_{d_{\text{PIN}},d_{\text{MIN}}} = \frac{n_{\text{observed},d_{\text{PIN}},d_{\text{MIN}}} - \left\langle n_{\text{rand},d_{\text{PIN}},d_{\text{MIN}}}\right\rangle}{\sigma_{rand,d_{\text{PIN}},d_{\text{MIN}}}},$$

where n is the number of times a particular distance pair d_{PIN} and d_{MIN} was observed ($n_{observed}$) or obtained in random networks (n_{rand}), brackets indicate mean values, and $d_{PIN} > 0$ and $d_{MIN} > 0$ (see next paragraph).

Treatment of multi-enzyme complexes

If subunits belonging to the same multi-enzyme complex carried identical EC numbers, their distance was considered zero and their network relationship was not analyzed further in the correlation analysis as the minimum distance included in the analysis is one. If they carried different EC numbers, their distances were computed as for any other enzyme pair given the available data.

The centrality of nodes

The centrality of nodes in PINs was measured either by their betweenness (BN) according to the algorithms proposed by Newman [66], or by the influence on the average shortest path between enzymes (CL_{EC}), according to the Equations 2.5. While BN corresponds to the number of shortest paths leading through a particular node, the latter centrality measure evaluates the changes on the average shortest path length of a graph after removal of a particular node. For each node, a *z*-score of CL_{EC} was calculated to judge the centrality of a node (Eqs. 3.6).

Eqs. 3.6

$$\text{zScore}_i = \frac{(\text{CL}_i - \langle \text{CL}_{\text{EC}} \rangle)}{\sigma_{\text{CL}}};$$

$$CL_{EC} = \frac{1}{|E_{EC}|}\sum_{(i,j)\in E_{EC}} d(i,j);$$

$$\sigma_{CL} = \sqrt{\frac{\sum_{i\in N_{EC}}\left(\text{CL}_{\text{EC},i} - \langle CL_{EC}\rangle\right)^2}{N_{EC}}};$$

$$E_{EC} = \{(i,j) \in \{1,2,...,N\} x \{1,2,...,N\} : \infty > d(i,j) \geq 0 \wedge j > i \wedge i,j \in enzymes\}$$

Notation as for Eq. 2.1.

Correlation of PINs and metabolic flux rates

For correlating PINs and metabolic flux rates, we used flux rate data from a large-scale ^{13}C-flux analysis from Blank and colleagues [80]. In this approach, flux rates of enzymes of the global metabolic network of yeast strain iFF708 [79] were estimated by flux balance analysis. In particular, we used data of flux rates measured in yeast growing in a glucose-containing medium resulting in flux data for 747 unique reactions catalyzed by 672 enzymes. The enzymes were divided into a group of enzymes with flux rates greater than 50, enzymes with flux rates between 10 and 50, flux rates of 0.1 to 1, 1.0E-4 and 0.1, and 0 to 1.0E-4 relative to an uptake of glucose set to 100 (arbitrary flux units). While the flux rates were divided according to a logarithmic scale, the range 1.0E-4 to 0.1 had been chosen to yield similar numbers of enzymes in all bins.

Metabolic pathways

Metabolic pathways and associated proteins used in this study were taken from the SGD database. For the fatty acid synthesis pathway, malic enzyme was assumed as a source of NADPH and the malat dehydrogenase as a source of Acetyl-CoA and added to the pathway.

Chapter 4
Topology of Phosphorylation-Networks

4.1 Background

In Chapter 3, we described the global network properties of protein interaction networks and metabolic networks. Filtering and different construction methodologies were applied to reveal organizational differences between raw networks including all interactions, and networks designed specifically to capture aspects of metabolism.

For this purpose, we identified proteins of rather non-metabolic functions and processes and their associated interactions. The rPIN comprised 1,186 proteins related to DNA processing functions with 21,952 associated interactions, 297 protein-degradation related proteins involved in 4,999 interactions, and 267 kinase-phosphatase associated proteins with 8,251 associations. All these interactions were partially overlapping as proteins from different groups were also reported to interact. After removing these interactions, the remaining nodes span a graph of 1,517 proteins, the fPIN, which was considered to be key molecular components responsible for maintaining the metabolic machinery. The analysis of the networks revealed a remarkable difference of global structural properties of the raw network upon filtering of sub-networks. The characteristic length (*CL*) of the fPIN was 8.16, which was approximately twice as long as the *CL* associated with rPIN (3.49) suggesting that, in particular, highly connected nodes providing shortcuts have been removed. The average cluster coefficient of the rPIN was determined as 0.16 and 0.39 for the fPIN indicating increased modularity. And finally, we observed a positive correlation of degrees of neighboring nodes in the fPIN, i.e. the assortativity was determined as 0.15. By contrast, for rPIN a negative correlation of -0.11 was shown. The dissortativity of the rPIN was suggested to result from the high dissortativity of protein sub-networks discarded in the filtered PINs. The graph comprising relations between DNA-related proteins showed a dissortativity of -0.12, protein-degradation -0.26 and the kinase-phosphatase associated proteins of -0.36. The graph comprising only kinase associated proteins, exhibits an even higher dissortativity (r_d = -0.40) and a characteristic length of 3.69, which is very close to that of the CL for rPIN. However, since the sub-network is based on a reported protein interactions, without available information on the function of the interaction, its limited correspondence to the phosphorylation network is evident. The protein interactions do not necessarily correspond to kinase-target recognition events, they may correspond to an alternative function: regulatory events, protein interactions involved in protein-

production, degradation and transport as well as association to structural proteins of the cell, are conceivable. Furthermore, for kinase-kinase interactions, determination of the directions of the interactions is difficult.

Recently, Linding et al. established a resource for exploring phosphorylation networks[33]. In their approach they reconstructed the human phosphorylation network by integrating phosphorylation site information and five fundamentally different types of interaction evidence: genomic context, physical protein interactions and gene co-expression, manually curated pathway databases and automatic literature mining. Their construction strategy comprised two steps and yielded predictions of kinase-target interactions. In the first step, the related kinase family for each reported phosphorylation site was predicted, applying a set of 20 kinase-specific predictors (NetphosK[96] and Scansite[97]). The best candidate kinases within the appropriate kinase families were then identified from a protein network of functional associations by calculating the proximity of the most probable path connecting them. The context was thus used as a filter that eliminated many false-positive predictions obtained from sequence motifs.

In this chapter, the global properties of the Linding's phosphorylation networks are determined. Two strategies were pursued, the interpretation of the phosphorylation network as an undirected network as well as a directed network. The directed network considers the flow of information from the kinase to the related target and therefore allows to study signaling properties of the network.

4.2 Results

Linding's network (networKIN) comprises 1,777 human proteins, including 68 kinases and 4,481 interactions. We pursued two strategies to get more insight into the phosphorylation network. We represented the network both as a directed and an undirected graph. It is very important to understand the differences of both networks. Nodes in the undirected network interact in a bidirectional manner, while nodes in the directed network are only linked downstream the phosphorylation signaling cascade. This means in particular that kinases, acting on the same protein, are indirectly connected by their common target in the undirected network, while no path exists between both kinases in the directed network. Thus, since the network contains only 68 nodes with out-going links, 1,709 nodes are assigned with a degree of zero.

Both networks exhibit closely related, short, characteristic lengths (CL). The CL for the undirected graph was determined as 3.10, while a CL of 3.18 for the directed graph was determined. Since the directed network reflects the downstream signaling of kinases, its observed CL represents the average length of signaling paths. Both networks are observed to be highly dissortative (r_d = -0.56), i.e. highly connected nodes prefer to interact with nodes of lower degrees. Furthermore, the undirected network exhibits a

mean cluster coefficient $< C >$ of 0.19, which follows the intuitive expectation of relatively high modularity of the phosphorylation network. However, only a $< C >$ of 0.002 was computed for the directed network. The difference between both $< C >$ implies that modularity in the phosphorylation network is mainly established by cross-reactivity of kinases, which, in the directed network, are not apparent. Indeed, an undirected phosphorylation network, which only consists of interactions between kinases, exhibits an even lower $< C >$ of 0.06. Here, it should be noted that the modularity of the phosphorylation networks corresponds to the regulation mechanism of cellular functions. A particular phosphorylation state reflects a (fine-tuned) adjustment to current environmental conditions. This (fine-tuned) adjustment is constituted by balanced, counteracting activations and deactivations, and is established both on the kinase levels, where further phosphorylation is regulated by cross interacting kinases, as well as by cross-reaction of kinases with concurrent activation and deactivation of a common target. Considering the observed $< C >$, the prominent role of cross-reactivates might be suggested. However, since the interactions in the phosphorylation networks are not weighted according to the effect of phosphorylation, the estimation of the influence of both regulation modes is not possible.

The influence of particular kinases on the entire phosphorylation network may further be investigated by computing the components of the directed network. Starting from a particular kinase, the respective corresponding component comprises the nodes, which can be reached from the kinase. It is constituted by the shortest paths only. Thus, the components represent signaling trees, where the initial signal, the phosphorylation by the root kinase is passed on to the leaf targets. For phosphorylation networks, the sizes of components are meaningful quantities. In particular, they reflect the possible influence of a particular kinase on the entire network. In addition, the depth of the respective component is the longest signaling path and the betweenness-centrality weights the nodes according to their central role, as a hub node, in the signaling network. While the giant component of the undirected network comprises the entire network (all 1,777 nodes), the size of the giant component of the directed network was discerned as 1,692, and was initiated by the cAMP-dependent protein kinase beta. Overall, 37% (25 out of 68) of kinases included in networKIN were observed to be capable to exert influence on the activity of more than 92% of the entire network (> 1650 nodes), revealing a high complexity of the phosphorylation network (**Table 4.2**). The longest signaling path passes 7 kinases and is initiated by the Casein Kinase I (component depth of 9). The most betweenness-central kinase was determined as Casein kinase II in the undirected network and the proto-oncogene tyrosine-kinase SRC, in the directed network (**Table 4.2**).

A summary of global network properties for the phosphorylation networks, the protein interaction networks rPIN and fPIN as well as the metabolic interaction networks mapEIN is provided in **Table 4.1**.

Table 4.1 Summary of global network properties associated with the different types of PINs and MINs investigated in this study

	PIN		phosphorylation networks			MIN
	rPIN	fPIN	kPIN	Unidir.	direct	mapEIN
Number of nodes	5438	1517	4999	1777	1777	1957
Number of edges	39766	1086	7471	4713	4713	6395
Number of enzymes	869	522	139*	68*	68*	1957
Giant component	5415	510	4999	1777	1692	1674
Characteristic length, CL	3.49	8.16	3.69	3.10	3.18	6.62
<Cluster coefficient>	0.16	0.39	0.15	0.19	0.002	0.47
<Neighbors' Connectivity>	57.17	2.65	47.41	229.84	0.12	10.34
Assortativity	-0.11	0.15	-0.40	-0.56	-0.56	0.26

Brackets indicate mean values. Properties are explained in greater detail in the Materials and methods section of 0.; *kinases

Table 4.2 Topological properties of kinases in the phosphorylation network: networKIN. The size and depth of the instituted component, signaling tree, were determined. Furthermore, the centrality of kinases in the undirected and directed network were indicated by the rank of the kinases in the betweenness-centrality scale.

Gene symbol	ENSEMBL-ID (ENSP00000*)	Signaling-tree size	Signaling-tree depth	Centrality undir.	Centrality direct.	applied predictor	Kinase name
PRKACB	***319504**	**1692**	7	38	11	PKA	**cAMP-dependent prot. kinase ß-catalytic unit**
CSNK1D	*324464	1663	7	38	27	CKI	Casein kinase I, delta isoform
CSNK1A1	***261798**	1658	**9**	38	31	CKI	**Casein kinase I, alpha isoform**
CDC42BPA	*295191	1652	7	38	51	DMPK	CDC42-binding protein kinase alpha isoform A
MAPK14	*229795	1651	5	36	6	p38MAPK	Mitogen-activated protein kinase 14
RPS6KA1	*263975	1651	8	32	24	RSK	Ribosomal protein S6 kinase alpha 1
CDK2	*266970	1651	4	36	5	cdk5	Cell division protein kinase 2
IGF1R	*268035	1651	7	33	12	INSR	Insulin-like growth factor I receptor precursor
LYN	*276497	1651	6	30	26	SRC	Tyrosine-protein kinase LYN
ATM	*278616	1651	7	36	28	ATM	Serine-protein kinase ATM
PRKCA	*284384	1651	7	31	7	PKC	Protein kinase C, alpha type
INSR	*303830	1651	6	8	10	INSR	Insulin receptor precursor
PRKCB1	*305355	1651	5	35	8	PKC	Protein kinase C, beta type
CDC2	*306043	1651	6	10	4	cdk5	Cell division control protein 2 homolog
	*309591	1651	7	36	17	PKA	cAMP-dependent protein kinase, alpha-catalytic unit
PRKDC	*313420	1651	6	12	13	DNAPK	DNA-dependent protein kinase catalytic subunit
RPS6KA3	*316010	1651	7	13	3	RSK	Ribosomal protein S6 kinase alpha 3
MAPK9	*321410	1651	7	36	21	p38MAPK	Mitogen-activated protein kinase 9
FYN	*321899	1651	5	28	20	SRC	Proto-oncogene tyrosine-protein kinase FYN
GSK3B	*324806	1651	6	36	2	GSK3	Glycogen synthase kinase-3 beta
LCK	*328213	1651	6	14	22	SRC	Proto-oncogene tyrosine-protein kinase LCK
CSNK1E	*338432	1651	8	29	15	CKI	Casein kinase I, epsilon isoform
SRC	***341571**	1651	6	**1**	16	SRC	**Proto-oncogene tyrosine-protein kinase Src**
-	***341595**	1651	5	15	**1**	CKII	**Casein kinase II, alpha chain**
MAPK8	*345944	1651	7	34	14	p38MAPK	Mitogen-activated protein kinase 8

Table 4.2 (continued)

Gene symbol	ENSEMBL-ID (ENSP00000*)	Signaling-tree size	Signaling-tree depth	Centrality undir.	Centrality direct.	applied predictor	Kinase name
CSNK2A2	*262506	213	3	38	9	CKII	Casein kinase II, alpha' chain (CK II)
MAPK10	*226594	103	2	26	23	p38MAPK	Mitogen-activated protein kinase 10
MAPK1	*215832	94	2	27	37	Erk1	Mitogen-activated protein kinase 1
PRKG2	*264399	83	4	38	66	PKG	cGMP-dependent protein kinase 2
PRKG1	*332522	82	3	16	19	PKG	cGMP-dependent protein kinase 1, beta isozyme
AKT1	*270202	74	1	37	30	PKB	RAC-alpha serine/threonine kinase
MAPK3	*263025	57	3	37	44	Erk1	Mitogen-activated protein kinase 3
SGK	*237305	50	1	11	25	PKB	Serine/threonine-protein kinase Sgk1
ERBB4	*342235	41	3	38	47	EGFR	Receptor protein-tyrosine kinase erbB-4 precursor
EGFR	*275493	37	2	2	18	EGFR	Epidermal growth factor receptor precursor
HCK	*262651	31	2	19	29	SRC	Tyrosine-protein kinase HCK
ERBB2	*269571	19	2	18	35	EGFR	Receptor protein-tyrosine kinase erbB-2 precursor
CDK3	*293215	13	1	38	45	cdk5	Cell division protein kinase 3
MAPK13	*211287	11	1	38	38	p38MAPK	Mitogen-activated protein kinase 13
RPS6KB1	*225577	8	1	9	32	RSK	Ribosomal protein S6 kinase
CSF1R	*286301	8	1	27	52	PDGFR	Macrophage colony stimul. fact. I receptor precurs.
KIT	*288135	8	2	3	33	PDGFR	Mast/stem cell growth factor receptor precursor
-	*297518	8	1	17	46	cdk5	-
ABL1	*298467	8	1	5	34	Abl	Proto-oncogene tyrosine-protein kinase ABL1
ERBB3	*267101	7	1	7	40	EGFR	Receptor protein-tyrosine kinase erbB-3 precursor
MAPK11	*310750	7	1	38	53	p38MAPK	Mitogen-activated protein kinase 11
PDGFRA	*257290	6	2	38	65	PDGFR	Alpha platelet-derived growth factor receptor
PRKCQ	*342894	6	1	20	36	PKC	Protein kinase C, theta type
BTK	*337233	5	1	22	54	Abl	Tyrosine-protein kinase BTK
ITK	*231189	4	1	16	60	Abl	Tyrosine-protein kinase ITK/TSK
PRKCL1	*242783	4	1	27	43	PKC	Protein kinase C-like 1
ROCK1	*261535	4	1	38	62	DMPK	Rho-associated, coiled-coil containing prot. kinase 1
PDGFRB	*261799	4	1	4	50	PDGFR	Beta platelet-derived growth factor receptor precrs.
PRKCG	*263431	4	1	38	48	PKC	Protein kinase C, gamma type
PRKCZ	*288816	4	1	23	41	PKC	Protein kinase C, zeta type
AKT2	*309428	4	1	37	56	PKB	RAC-beta serine/threonine protein kinase
CLK1	*326830	4	1	38	58	Clk2	Dual specificity protein kinase CLK1
FGFR1	*337247	4	1	6	42	PDGFR	Basic fibroblast growth factor receptor 1 precursor
RPS6KA5	*261991	3	1	38	63	RSK	Ribosomal protein S6 kinase alpha 5
CAMK2A	*305090	3	1	25	55	CaM-II	Calcium/calmodulin-dependent protein kinase II
MAPK7	*316736	3	1	38	64	Erk1	Mitogen-activated protein kinase 7
YES1	*324740	3	1	21	39	SRC	Proto-oncogene tyrosine-protein kinase YES
FGR	*001380	2	1	24	49	SRC	Proto-oncogene tyrosine-protein kinase FGR
MAPK12	*215659	2	1	27	59	p38MAPK	Mitogen-activated protein kinase 12
PTK6	*217185	2	1	27	61	SRC	Tyrosine-protein kinase 6
CSNK1G2	*255641	2	1	38	66	CKI	Casein kinase I, gamma 2 isoform
CLK4	*316948	2	1	38	66	Clk2	Dual specificity protein kinase CLK4
FGFR3	*339824	2	1	27	57	PDGFR	Fibroblast growth factor receptor 3 precursor

4.3 Discussion

We presented an analysis of the phosphorylation network proposed by Linding et al. [33]. The network considers kinase-target interaction events and therefore allows the construction of a directed network. The topological properties of the undirected networKIN and kPIN were determined to be broadly concordant, albeit the kPIN does not explicitly consider kinase-target interactions, and the phosphorylation networks are based

on evolutionarily distant organisms. The CL for the undirected networKIN was determined as 3.10, the $< C >$ as 0.19 and the assortativity r_d as -0.56, while a CL of 3.69, $< C >$ of 0.15 and r_d of -0.40 were determined for kPIN. We observed similar topological tendencies for the rPIN, described in 0, suggesting once again an evident influence of the signaling sub-networks on the global topologies of the rPIN.

The average clustering coefficient $< C >$ of the directed phosphorylation network was determined as 0.002. Thus, the directed networKIN exhibited higher modularity when compared to the directed networKIN, suggesting intense cross-reactivity of kinases in phosphorylation networks. The cross-reactivity is one of the general regulatory modes of phosphorylation networks. Thereby, the regulation mechanism of cellular functions reflects a balance of contradicting activations and deactivations, which may generally be established either on the kinase level (by regulation of phosphorylation) or on the target level (by concurrent cross-reactivities of kinases on common targets). Since the undirected phosphorylation network, consisting solely of interactions between kinases, exhibits an even lower $< C >$ of 0.06, a prominent role of regulation on target level might be suggested. However, since the interactions in the phosphorylation networks are not weighted according to the effect of phosphorylation, a reliable estimation of the influence of both regulation modes is not feasible.

For each kinase, the respective corresponding component, or signaling tree, was computed. Surprisingly, the kinase which corresponds to the giant component was determined as PKA. One would rather expect a receptor kinase to shape the largest network. However, one should keep in mind, that the connection between the receptor kinases and many cellular kinases, as the PKA, is established via second messengers, such as cAMP , which are not considered within the networKIN. Furthermore, it should be noted that 37% of kinases in the network are capable to influence more than 92% of the entire phosphorylation network, such that the differences in the size of the respective components are only modest. Thus, the prominent role of PKA must be quantified. The depth of the respective signaling tree may be more representative. The longest path was determined for CKI (9), although even this kinase is not a classical initiator of signaling cascades. Finally, we classified the hub properties of kinases. The kinases with the highest betweenness-centrality were determined as the SRC-proto-oncogene tyrosine kinase in the directed and the CKII kinase in the undirected networKIN.

The global properties of different network types in the rPIN highlight again the necessity of careful curation of protein interactions, according to the investigated aspects. While protein interaction networks comprising metabolic aspects are characterized as assortative, and to exhibit a relative long characteristic length and high modularity, phosphorylation networks are dissortative, with shorter characteristic length and are less modular.

However, considering the gap between the number of kinases in the human genome, which was estimated approximately as 500[98], and the number of kinases in the phosphorylation network proposed by Linding et al, the weakness of the networKIN becomes evident. It might be suggested that the classification of phosphorylation sites according to their respective kinases may still suffer from limited numbers of kinase-specific predictors, which generally results from a limited number of experimentally determined kinase-phospho-site annotations. It remains to be seen, if the advances in protein-chip-technologies, e.g. epitope phosphorylation by monoclonal kinases will bring more clarity to this issue.

4.4 Methods

We obtained the list of kinase-target relations for the phosphorylation network as proposed by Linding et al. from the networKIN resource (http://networkin.info/download.php). Within this list, respective kinases and their associated targets are annotated by a unique ENSMBL-ID. Descriptions of the involved proteins are provided. We perused two strategies for the representation of the phosphorylation network. The networKIN was both represented as an undirected and directed graph, in which only interactions downstream the kinase-target interactions were established.

For the construction of the kPIN, the sub-network comprising the kinase-phosphatases related proteins and their interactions were used (see Chapter 3, Materials and Methods). The kinase-phosphates network was refined by including of proteins with GO:Annotations reflecting kinase activities only and their interactions.

For all three phosphorylation networks, we computed the global properties according to algorithms described in Chapter 3, Materials and Methods.

Chapter 5
Classification using Support Vector Machines

This chapter briefly introduces the key concepts of classification using Support Vector Machines. In particular, the principles underlying the training of support vector machines, the reduction of dimensionality of input data by principle component analysis prior to the training as well as the final assessment and comparison of classification algorithms are described. We will use support vector machines to predict novel phosphorylation sites in proteins, utilizing the sequence-information of phosphosite motifs as well as 3D-spatial profiles of phosphosites in Chapter 6 and Chapter 7.

5.1 Support Vector Machines

Support Vector Machines (SVMs) provide an efficient toolbox for classification problems. The SVMs were founded in statistical learning theory by Vapnik 1998[99; 100; 101]. The main advantages of SVMs are bounds on the generalization error (risk), the possibility of formulation as a quadratic programming problem with available efficient optimization algorithms for finding the global optimum, the robustness in terms of overfitting, and the possibility for application to high dimensional problems.

SVMs are computer algorithms that learn by example to assign labels to objects. In principle, a SVM tries to optimize a particular mathematical function, the decision function, to determine the differences between two groups and separate them in an optimal way. To understand SVMs it is necessary to understand few basic concepts behind the SVM: (i) the separating hyperplane, (ii) the maximum-margin hyperplane, (iii) the soft margin and (iv) the kernel function.

The separating hyperplane and maximum-margin hyperplane

The human eye is very good at pattern recognition. A glance on **Figure 5.1**a allows drawing a separation line between the different groups. Subsequently, a label might be assigned to a yet not annotated item, by a simple evaluation of the position of the item relative to the drawn separating line (**Figure 5.1**). While the association to a particular group depends on the position of an item with respect to the separation line (above or below), the certainty or reliability of the decision corresponds to the distance to the separation line. The SVM tries to find a separation line or, in case of higher dimensional space, a separation hyperplane, which separates the groups in an optimal

way. The hyperplane is defined by $\langle w * x_i \rangle + b = 0$, where w is the normal vector and b the threshold. Thus, the decision function is $f(x) = \langle w * x \rangle + b$. The sign of the decision value, $sgn(\langle w * x_i \rangle + b)$, implies the relevance to a particular group, its absolute value, $abs(\langle w * x_i \rangle + b)$, the significance of the prediction. The larger the absolute decision value, the more reliable the decision. However, as an infinite number of hyperplanes are capable of separating the training groups, the task of training a SVM is to find the optimal hyperplane, whereas the optimal hyperplane is defined by the maximal margin to training items. This assures a better generalization of the classifier (**Figure 5.1**b). The margin is defined by the distance of the closest data points to the hyperplane, which is $2/\|w\|^2$. Thus, the task is to maximize $2/\|w\|^2$, or equivalently to minimize $\|w\|^2/2$. Furthermore, as the normal, w, may be expressed as a linear combination of the training vectors, $w = \sum \alpha_i y_i x_i$, the task is to minimize the magnitude of the weighting vector α. Consequently, the decision function for linear separable groups, $f(x) = sgn(\langle w * x_i \rangle + b) = sgn(\sum \alpha_i y_i \langle x_i, x \rangle + b)$, is defined by support vectors (training samples for which $\alpha_i > 0$) and a threshold b.

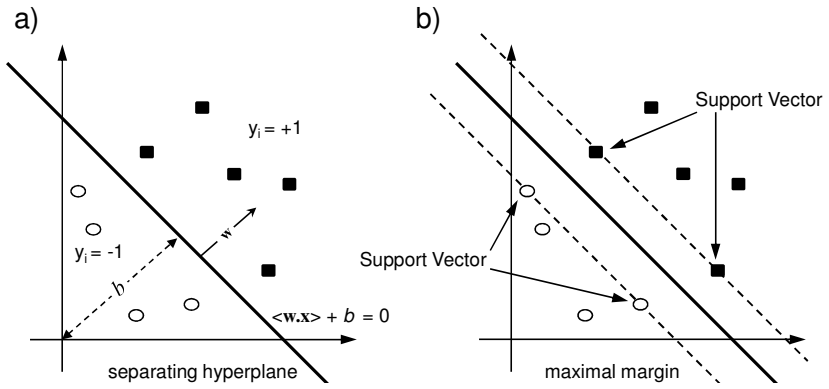


Figure 5.1: Classification strategy of a Support Vector Machine.
(a) During training, the SVMs try to find the optimal separating hyperplane. The separating hyperplane is defined by $f(x) = \langle w * x \rangle + b = 0$. Items are related to a particular group, given by the relative position of the item to the hyperplane. While the sign $f(x)$ reveals the respective class, the distance to the hyperplane defines the reliability of the prediction. (b) The optimal separating hyperplane is defined by the maximal margin, while a margin is defined by the closest training items to the separating hyperplane, the support vectors.

The soft margin

The special case of ideally and linearly separable groups is very rare. In most cases no hyperplane may be defined, and the goal is to minimize the classification error. To handle such cases, the SVM algorithm was extended by an error term, the slack variable ξ_i, allowing some data to be classified wrongly. The slack variable ξ_i is $\xi_i = \|w\|$ for wrong and $\xi_i = 0$ for correct classification. Consequently, the extended task of the

training process now is to minimize $\|w\|^2/2 + C\sum\xi$, where the cost factor C is introduced to balance the task of maximizing the margin and correct classification.

The kernel function

To understand the notion of a "kernel function", a classification problem is introduced in **Figure 5.2**a. Each item is described by a single value. The two groups are not linearly separable and even the introduced slack variable does not provide a possibility for a prediction. Separation is finally made possible by adding a new dimension (**Figure 5.2**b). Here, the one-dimensional input space is mapped onto a two-dimensional feature space, by a simple addition of a new dimension containing the square of the original value, i.e. applying the polynomial projection function $\Phi(x_i)$: $x_i \rightarrow (x_i, x_i^2)$. Furthermore, the decision function contains a scalar product of the feature vectors, $f(x) = sgn(\sum \alpha_i\, y_i \langle \Phi(x), \Phi(x_i)\rangle + b)$. Thus, the extended decision function is defined as $f(x) = sgn(\sum \alpha_i\, y_i k\langle x, x_i\rangle + b)$, where $k\langle x, x_i\rangle$ is the kernel function and i denotes support vectors. However, mapping into very high-dimensional spaces can be problematic, due to the so-called curse of dimensionality. With increasing number of variables, the number of possible solutions also increases exponentially. Consequently, it becomes harder for any algorithm to select a correct solution. One would like to use a kernel function that is likely to allow the data to be separated, but without introducing too many irrelevant dimensions. The choice of optimal kernel function, however, in most cases relies on trial and error. Thus, to eliminate redundancies in the input space a dimension reduction prior to the training step is advisable. Equations 5.1 summarizes the commonly used kernel functions:

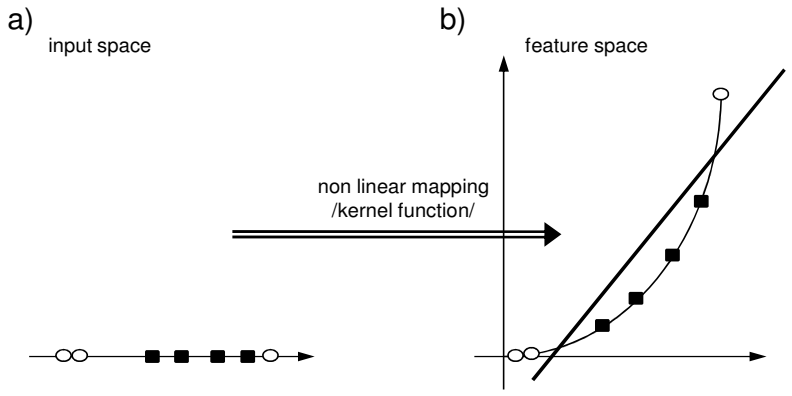


Figure 5.2: Non-linear mapping of the input space by the kernel function.
(a) Each training item is described by only one variable. The two groups are not linearly separable. (b) To allow the separation, the input space is extended by new dimensions, here by the polynomial projection function $\Phi(x_i)$: $x_i \rightarrow (x_i, x_i^2)$. As the decision function contains a scalar product of the projections of the feature vectors, $f(x) = sgn(\sum \alpha_i\, y_i \langle \Phi(x), \Phi(x_i)\rangle + b)$, the projection vectors may be computed once and stored within the kernel, such that the scalar product may be accessed by the kernel function $k\langle x, x_i\rangle = \langle \Phi(x), \Phi(x_i)\rangle$.

Eq. 5.1

linear: $k\langle x, x_i\rangle = \langle x, x'\rangle$

polynomial: $k\langle x, x_i\rangle = (scale\langle x, x'\rangle + offset)^{degree}$

gaussian (radial basis): $k\langle x, x'\rangle = e^{-\sigma|x-x'|^2}$,

where $scale$, $offset$ and $degree$ as well as σ are optimization parameters.

5.2 Dimension reduction via Principal Component Analysis (PCA)

Real biological data often consist of huge data sets comprising thousands of variables describing dozens to hundreds of samples. It is not unusual that the number of variables exceeds the number of samples, making the assessment of the importance of a particular variable to capture the differences among the sample set difficult. Furthermore, with increased number of variables, data becomes hard to handle. One would like to select the important variables prior to further analysis, that is to perform a dimensionality reduction without decreasing the information content of the data set. Principal Component Analysis (PCA) is a well-known method of dimensionality reduction. In essence, PCA tries to find an uncorrelated, linear transformation of the input set of variables, while retaining as much as possible the variation of the origin variables. Whenever two variables are correlated, the information gained by the addition of the second variable is reduced. Therefore, dimensionality reduction by projection of the input space onto an alternative uncorrelated output space is justified. The goal of PCA is to find an uncorrelated linear transformation of the original p variables designed by the input space $X = [x_1, x_2, \dots, x_p]$ to a set of k predictive variables defined by the computed space $T = [t_1, t_2, \dots, t_k]$ with high variance. Geometrically, the PCA transforms the original variable to a new coordinate system obtained by rotating the original system. The new dimensions, the principle components, represent axes with the maximal variance and are ordered by the amount of variance in the original data they account for.

For the purpose of prediction, PCA has at least two major advantages, when applied prior to classification by SVMs: an procedural improvement and a possible improvement of the performance of the predictor. Procedural improvement comprises decreased data size as well as the reduction of computation time and memory requirements for the subsequent algorithm. Improvement of performance comprises possible improvements by direct elimination of correlations and indirect potential improvement via fine-tuning the predictor, while removing misleading variables. While the application of between-group PCA might be more obvious in the context of discrimination of the respective group, as it considers the differences of variance between

groups, we observed that the application of PCA prior to a SVM prediction for the prediction of phosphorylation sites outperforms the predictions based on the between-group PCA, even if they are subsequently linked to SVMs.

5.3 Assessing and comparing classification algorithms

Perhaps the simplest and most intuitive measure of the performance of a predictor is the accuracy, which is the proportion of the correct predictions: $Accuracy = \frac{(T_N+T_P)}{(T_N+T_P+F_P+F_N)}$, where T_N is the number of true negative, T_P true positive, F_P false positive and F_N false negative predictions. The accuracy is often used to assess and compare predictions of phosphorylation sites, although this measure suffers from several drawbacks. Firstly, the comparison of different predictors is critical, if the sizes of the test sets differ and, secondly, the accuracy is prune to biased results. A predictor reporting only negative predictions will show high accuracies if the negative set exceeds the size of the positive set, although the prediction of the positive set is completely wrong. To circumvent this problem, specificity and sensibility of the prediction often accompanies the reported accuracy. An alternative measure is Matthew's Correlation Coefficient (CC) (Eq. 5.2), which is balancing the negative and positive sets. While a complete recall at the expense of the precision is reflected as zero, a CC of '-1' indicates exactly wrong and '+1' a perfect prediction. However, perhaps the most commonly used method to assess the performance is to draw a receiver operating characteristics (ROC). In the ROC chart, the true positive rate is plotted versus the false positive rate in respect to decreasing decision thresholds. The performance of a predictor is defined by the area under the ROC curve (AUC). While an AUC of 0.5 denotes a random prediction, values below 0.5 indicate wrong predictions and an AUC of 1 a perfect prediction. Both *CC* and AUC may be applied to fine-tune the classifier and to compare different predictors.

To estimate the classification error and the generalizability, a prediction approach must be tested on an independent test set. This may be achieved by performing a k-fold cross–validation. Here, the training set is initially partitioned into k-subsets. The predictor is then trained by all but one partition, which is subsequently classified. Finally, the performance is estimated applying above mentioned measures on the latter classification. The leave-one-out cross-validation corresponds to an N-fold cross validation, where a single observation is left out each time.

Eq 5.2. $$CC_{\text{cor.coef.}} = \frac{T_P T_N - F_P F_N}{\sqrt{(T_N + F_N)(T_N + F_P)(T_P + F_N)(T_P + F_P)}}$$

Definition of Matthew's correlation coefficient (CC) where T_P stands for true positives, F_N: false positives, T_N: true negatives, and F_N: false negatives.

Chapter 6
Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction

Abstract

Phosphorylation of proteins plays a crucial role in the regulation and activation of metabolic and signaling pathways and constitutes an important target for pharmaceutical intervention. Central to the phosphorylation process is the recognition of specific target sites by protein kinases followed by the covalent attachment of phosphate groups to the amino acids serine, threonine, or tyrosine. The experimental identification as well as computational prediction of phosphorylation sites has proved to be a challenging problem. Computational methods have focused primarily on extracting predictive features from the local, one-dimensional sequence information surrounding phosphorylation sites.

In this work, our goal was to characterize the spatial context of phosphorylation sites and to assess its usability for improved phosphorylation site prediction. We identified 750 non-redundant, experimentally verified sites with 3D-structural information available in the protein data bank (PDB) and grouped them according to their respective kinase family. We studied the spatial distribution of amino acids around phosphoserines, phosphothreonines and phosphotyrosines to extract signature 3D-profiles. To test the added value of using spatial information for the computational prediction of phosphorylation sites, Support Vector Machines were applied using both sequence as well as structural information. Characteristic spatial distributions of amino acid residue types around phosphorylation sites were indeed discernable, especially when kinase-family-specific target sites were analyzed. When compared the sequence-only based predictors developed as part of this study and other, publicly available predictors, a small but consistent performance improvement was obtained when the prediction was informed by 3D-context information.

While sequence-only based methods were observed to harbor most of the discriminatory information with regard to phosphorylation, spatial context information was identified as relevant for the recognition of kinases and their cognate target sites and can be used for an improved prediction of phosphorylation sites.

6.1 Background

Protein phosphorylation is a ubiquitously occurring posttranslational modification influencing many molecular processes in all complex cells. The recognition of phosphorylation sites by specific kinases and the subsequent phosphorylation generally leads to an alteration of structure, function, or protein binding properties of the target protein which has evolved as a mechanism to respond to environmental changes via phosphorylation-triggered complex signaling networks and cascades as well as playing a crucial role in the regulation of enzymes or transporters in metabolic processes [32; 102; 103; 104].

The study of phosphorylation events has been a central research topic in molecular biology for many years. Given the high number of candidate phosphorylation sites, efforts to experimentally identify and verify them all remain challenging. These difficulties motivated the development of computational methods to predict potential phosphorylation sites. Most established computational prediction methods rely solely on the local sequence surrounding the target amino acid residue. The developed prediction methods range from simple amino acid sequence pattern recognition methods to Markov Models, Neuronal Networks, and advanced machine learning methods such as Support Vector Machines [97; 105; 106; 107; 108; 109]. Many of them have been made publicly available and yield results with reasonable sensitivity and specificity, but they generally suffer from either over- or undercalling candidate sites as optimal parameters found for one particular protein target class cannot be generalized to all phosphorylation motifs [108; 110]. Recognizing that the information content increases significantly when the respective kinase families associated with their targets are considered separately, approaches to predict phosphorylation sites in a kinase-family specific manner based on family-specific local sequence motifs have also been presented [97; 105; 106; 107].

The acceptable performance of local-sequence-only methods, together with reports that phosphorylation sites appear to be preferentially located in unstructured regions of proteins suggesting a limited relevance of any structurally well-defined binding epitopes for the specific recognition of kinases and their substrate proteins [109] , appear to justify focusing exclusively on local sequence patterns rather than 3D-structural context information. However, the significantly increased number of experimentally determined phosphorylation sites by proteomics technologies with simultaneously available 3D structures of the associated proteins in recent years and published analyses suggesting that target sites may very well assume defined structural conformations and, furthermore that phosphorylation sites may be surrounded by specific 3D-structural environments [111; 112] motivated us to revisit the issue of the role of 3D-structural information for the specific recognition of kinases and their substrate proteins.

In a recently published systematic comparative and structural analysis of protein phosphorylation, Jiménez and co-workers [111] reported that serine and threonine phosphorylation sites exhibit only a marginal tendency to occur preferentially in structurally more flexible loops with approximately 35% actually being located in α-helices or β-strands, which can be assumed as relatively rigid structural secondary structural elements. For tyrosine sites, no tendency to occur more frequently in loops was detectable. Furthermore, they reported that a substantial number of phosphorylation sites (15%) are actually buried inside the protein and not exposed to the solvent. An increased significance of 3D-structural context for these locations is evident. Plewczynski and co-workers reported that as many as 60% of phosphorylation sites for the kinase families protein kinase A and C (PKA, PKC) are located in α-helical regions [107]. Thus, a significant number of phosphorylation sites are actually located in structurally defined regions in which defined structural surface features and motifs may turn out to be relevant.

When studying sequence motifs associated with the protein kinase A and G (PKA, PKG kinase families), the consensus target sequence was determined as xRRxSx [113; 114]. However, of 273 target motifs for PKA in the Phospho.ELM database [110], 5.5 % do not contain any arginine, and 1.5 % neither arginine nor lysine in the sequential neighborhood of six residues in both directions relative to the central serine. Of 32 targets for PKG kinases, 9.3 % of target sites do not contain any arginine, and in 6 % of the targets, both arginine and lysine is absent. This observation implies that some recognition features may perhaps be localized outside of the local sequence, such that the positive-charge bearing amino acids defining the required electrostatic potential surface for binding may be contributed from sequentially distant, but spatially close rather than sequence-local sites.

In the light of these observations, it appears possible that, although the local amino acid sequence may contain a significant portion of the information contents with regard to phosphorylation, the actual local 3D environment may contribute appreciably to the specificity of the kinase - target protein molecular recognition event.

Although there have been several approaches to use structural information for improved prediction of phosphorylation, they generally resulted in only modest success rates [96; 112]. These unsatisfactory results can possibly be explained by an insufficient number of annotated, experimentally determined structures as well as by focusing on general structural properties such as secondary structure, rather than trying to define 3D-motifs based on spatial amino acid distributions.

Fan and Zhang characterized phosphorylation sites in their spatial, protein-structural context using a simplified "Altman" shell model with a radius of 16Å and found only minor differences of the amino acid composition around phosphorylation sites compared to average protein composition [112; 115]. However, by analyzing phosphorylation

sites across all kinase families, any motif that may be specific for particular kinase classes may have been masked. The identification of kinase-family-specific sequence motifs supports this view. These amino acid preferences may also be detectable using a protein structural approach which considers spatial proximity rather than sequence proximity alone.

Plewczynski and co-workers applied molecular modeling to characterize the local structural context of phosphorylation sites [107]. In their approach, protein sequences are compared to a library of short sequence and structure motifs via a sequence matching algorithm, adapted for local 3D-structure prediction. They achieved significantly improved prediction accuracy of phosphorylation events by means of similarity scores to a library of PKA and PKC targets and conclude that "sequence information ought to be supplemented with additional structural context information […] for more successful predictions of phosphorylation sites in proteins."

The use of structural information for improved phosphorylation site predictions has also been explored by Blom et al., the authors of the popular sequence-only-based NetPhos predication server [96]. In this approach, probabilities of contacts between Cα atoms of residues within spatial neighborhoods of phosphorylation sites and non-phosphorylation sites are calculated, so called contact or distance maps. In a second step, the probabilities of contacts of residues from sequences are then calculated according to those maps and used for prediction purposes. This led to markedly improved sensitivity of the prediction of phosphorylated tyrosine sites which the authors interpreted as an indication of the relevance of tertiary structural information not reflected in the sequence alone. However, this approach also led to an increase of false negative sites and, as a consequence, to overall worse prediction results.

The goal of this work was to characterize the phosphorylation sites by spatial amino acid propensity distributions to generate spatial signature motifs and the subsequent assessment of this information to improve the sequence-only-based prediction of phosphorylation sites.

As previous studies have shown that "one-fits-all" approaches; i.e. parameterization of the prediction method irrespective of kinase-family, have led to only modest success rates, we investigate here whether considering kinase-family specific 3D-motifs may reveal greater information contents and, thereby, yield improved prediction results. Our method is based on Support Vector Machines (SVM) [99; 100]. SVMs have been used in a wide range of problems in the area of molecular biology including analyses of microarray data[116; 117; 118], string matching [119; 120], drug design [121], protein fold recognition [122] and prediction of phosphorylation sites using sequence information [96; 106; 107].

We observed that 3D-motifs are indeed detectable, especially when studying kinase families individually and obtained improved prediction results by including 3D information in the prediction. We also implemented a sequence-only approach that

implicitly captures 3D structural preferences associated with each of the different amino acid types by using 530 amino acid features which include also the generally accepted phosphorylation site features such as hydrophobicity, solvent accessibility as well as secondary and tertiary structure preferences, polarity, volume and solvent accessibility, structural disorder indices and others. The predictor has recently been developed by our group as part of a database of plant-specific phosphorylation sites. The predictor was shown to accurately identify plant phosphorylation sites and to outperform commonly available predictors [123].

6.2 Results

To characterize the general structural properties of phosphorylation sites (phos-sites) and to compare them to unphosphorylated sites (non-phos sites), we first analyzed secondary structural assignments, relative side chain solvent accessibility, and the crystallographic B-factor as a measure for local structural rigidity. A statistically significant tendency for serine as well as tyrosine phosphorylation sites to be more exposed to the solvent was detected. Threonine sites were also more exposed, albeit statistical significance could not be established (**Figure 6.1**, **Table 6.1**). While these observations follow the intuitive expectation that phosphate-group attachment sites should be more exposed, the magnitude of this difference appears surprisingly low (**Figure 6.1**). However, one has to bear in mind that serine, threonine and tyrosine, themselves polar amino acids, have an innate tendency to be exposed to water. The distribution of crystallographic B-factors of phos-sites in comparison to non-phos sites were also observed to differ (**Figure 6.1**). Phos-sites are more often found associated with the largest B-Factor (Bin 9), i.e. regions of greater structural flexibility, albeit significant *p*-values of differences were only observed for serine sites (**Table 6.1**). Phosphorylated serines and threonines are more frequently found in random coil regions and less in α-helical or β-strand regions than their unphosphorylated counter-parts (**Figure 6.1**). For tyrosine, no such preferences for secondary structural type were detectable, except for a marginally increased frequency of phosphorylated sites to occur more often in turns.

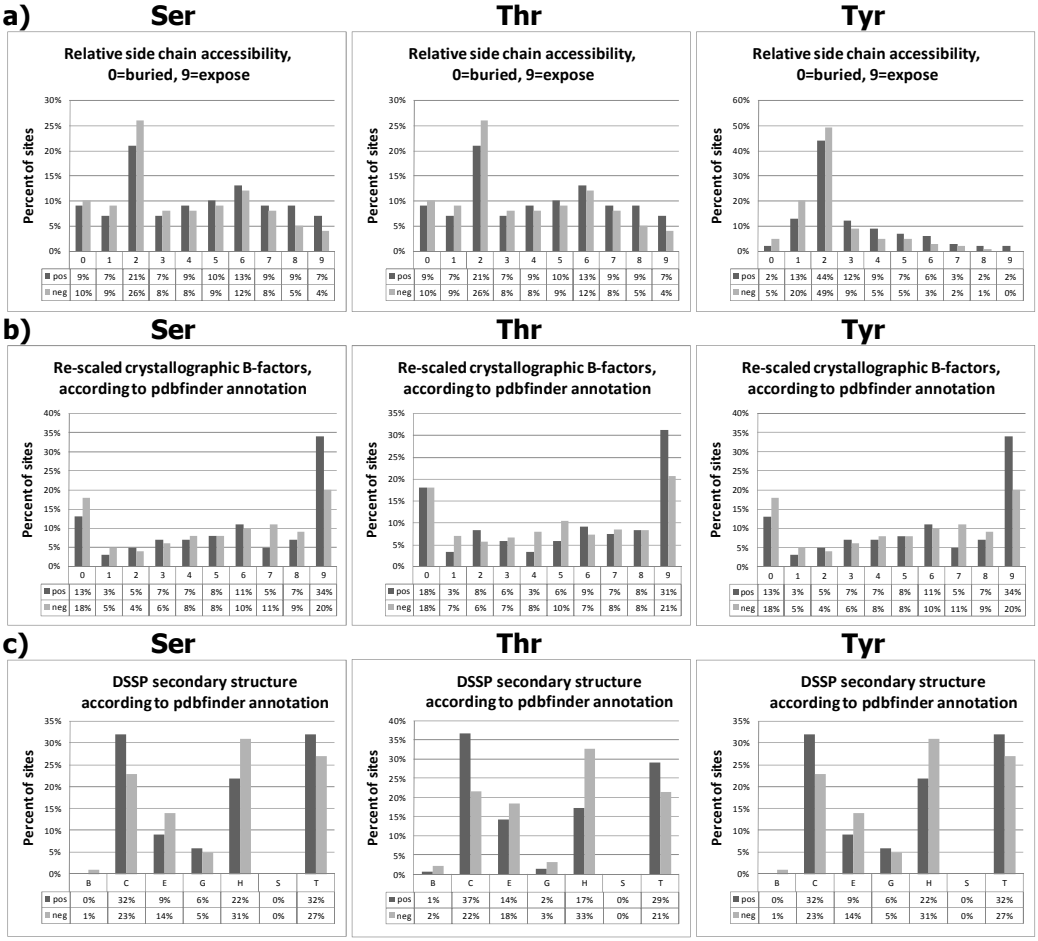


Figure 6.1 Comparison of general structural properties associated with phosphorylated (pos.) vs. non-phosphorylated (neg.) residues, serine left column, threonine middle column, tyrosine right column. Annotations were taken from PDBFinder [124]. (a) Side chain accessibility to solvent relative to the large possible accessibility for serine. (b) Re-scaled crystallographic B-factors describe the attenuation of x-ray scattering caused by thermal motion or quenched disorder and is applicable measure for local structural rigidity. B-Factors from pdb-structures in the range of [10..40] are mapped to the range [0..9] by PDBFinder; 0 signifying rigid structures, 9 – indicating unresolved, rather flexible structural regions. (c) DSSP secondary structure association. B = residue in isolated beta-bridge; C=Loop, irregular stretches; E = extended strand, participates in beta ladder; G = 3-helix (3/10 helix); H = alpha helix; S = bend ; T = hydrogen bonded turn.

Table 6.1 Statistics for significance of the observed differences of solvent accessibility and crystallographic B-factor of phosphorylated (pos) vs. non-phosphorylated (neg) for serine, threonine and tyrosine sites

Property	mean-Values		p-Values	
	Positive set	Negative set	t-Test	Mann-Whitney
		Serine-sites		
Accessibility	4.25	3.70	1.32 E-03	1.93 E-03
B-Factor	5.65	4.93	7.40 E-04	2.79 E-04
		Threonine-sites		
Accessibility	3.92	3.57	1.49 E-1	2.07 E-01
B-Factor	5.30	4.76	1.11 E-1	8.26 E-02
		Tyrosine-sites		
Accessibility	2.98	2.36	7.33 E-07	2.52E-06
B-Factor	5.56	5.15	6.96 E-02	2.68E-02

6.2.1 Characterization of the spatial environment of phosphorylation sites

We determined the propensities for the different amino acid residue types to occur in the spatial vicinity of the phosphorylated serines, threonines, and tyrosines, both for the sequential neighborhood as well as for the spatial environment. By separating the two, our goal was to identify possible 3D-signature motifs. In a third analysis, both contributions were combined to assess the relative contribution of the sequence and structural environment. As explained in the Method section, across all phosphorylation sites, we calculated the propensity values as log-odds ratios of the relative occurrences of amino acid types within distances from 2 to 10 Å from central phosphorylated amino acid residue and display the results in radial-radial cumulative propensity plots (RCP-plots) in which red-colored segments signify statistically significant enrichment relative to a reference set, and blue-colorings depletion.

When all serine, threonine, and tyrosine phosphorylation sites irrespective of their association with a particular kinase family are analyzed, both the sequence logos and the spatial profile of phosphorylated serines show only very little information contents (**Figure 6.2**). Only small differences relative to the reference set of un-phosphorylated sites were detectable as reflected by the only few colored segments in the RCP-plots indicating enrichment or depletion. For all three target amino acid types, most information appears to be contained in the local sequence and not in the spatial environment. By considering amino acids irrespective of their sequential proximity ("combined" graph), essentially no significant differences to the reference set of un-phosphorylated sites were found. This agrees well with results reported by Fan and Zhang who characterized structural microenvironments of phosphorylation sites within 16 Å from the central residue only and observed no evidence for significant amino acid propensities to fall within radial distance of 16 Å [112]. Interestingly, in the local sequence neighborhood, tyrosine residues – an amino acid that itself is target of phosphorylation events - appear to be depleted relative to the reference dataset in serine- and threonine-targeted phosphorylation sites. However, this depletion appears to be compensated by tyrosine residues found in the spatial environment such that overall ("combined" graph), no significant depletion of tyrosine residues in the environment of serine- and threonine phosphorylation sites was detectable.

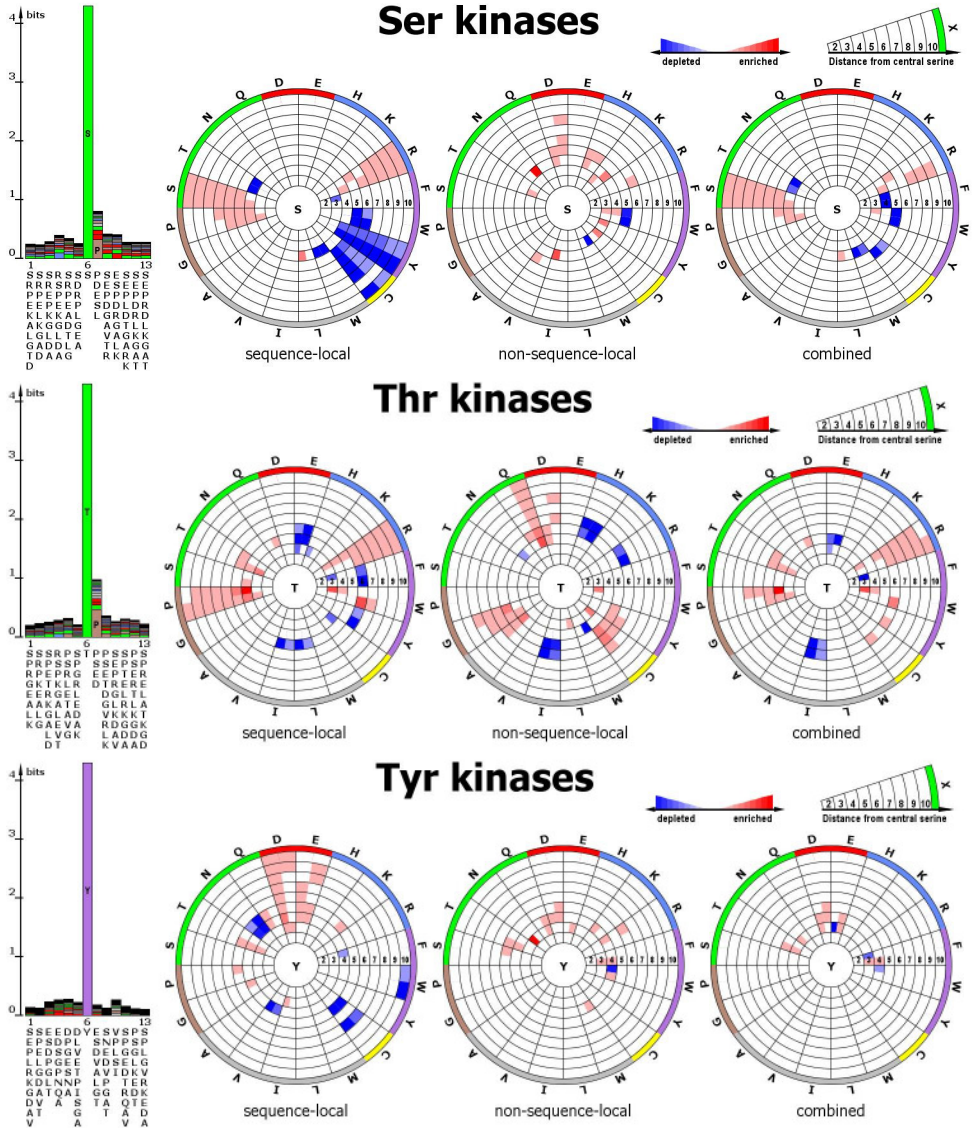


Figure 6.2 Sequence logos and radial cumulative propensity plots (RCP-plots) illustrating enrichment as well as depletion of particular amino acid types in the local sequence (sequence logo), sequence-local spatial environment including the 6 flanking amino acid residues on either side of the central serine/threonine/tyrosine, (left RCP-plot), spatially-local, but non-sequence local; i.e. excluding residues in the flanking sequence (middle plot), and combined information (right RCP-plot). For every amino acid type, the two different sub-sectors correspond to the statistics obtained by using the closest detected atom and the interaction center, respectively, and in clockwise order.

Kinase-family specific phosphorylation motifs

For the set of serine protein kinase sequences whose target proteins were found by screening the protein database; i.e. the structure of the target proteins is known, we constructed a phylogenetic tree and computed the corresponding sequence logos of the targets associated with kinase group (**Figure 6.3**) [125] to obtain an overview of the evolutionary relationships of the kinase sequences and their respective targets. Note: for tyrosine and threonine kinases, such analysis was not possible (with the exception of the

PTK group of tyrosine specific kinases) because of a lack of annotated kinase-target pairs with known structure. In agreement with results from previous studies, the sequence logos of serine kinase targets associated with the main serine-kinase families can be clustered into several groups [96; 108; 126; 127]. Evolutionarily close kinase-groups tend to also share common features in their respective targets. The major groups of targets are characterized by proline residues next to the central serine (CMGC kinase group except CK II) or a glutamate (ATM), a second group with negatively charged sequences (CMGC IV: CK II). The AGC kinase-group as well as the CaMK kinase-group comprise kinases with positively charged targets.

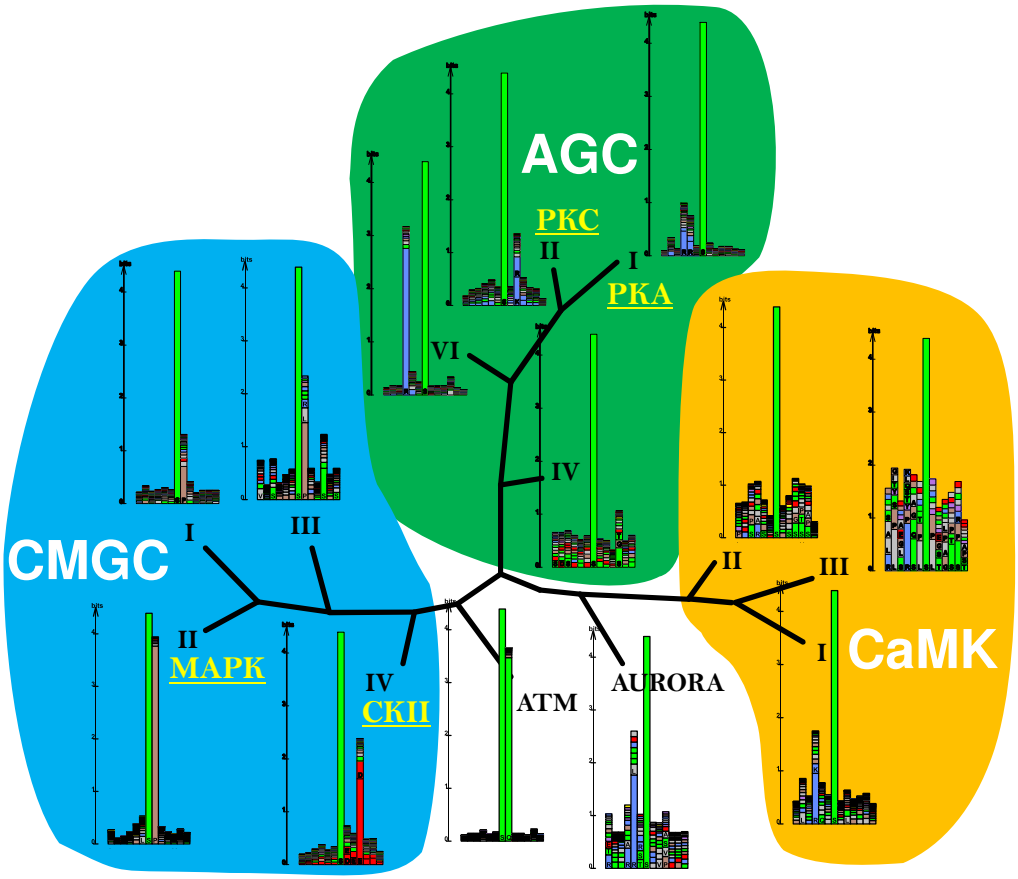


Figure 6.3 Phylogenetic tree of serine-kinase groups whose targets can be found in the protein structure database (PDB) according to the original Hanks and Hunter classification scheme [128] and associated sequence logos [125]. Kinases with high similarity tend to share similar targets. The major classes of kinase targets are characterized by a proline and glutamate next to the central serine, CMGC group I, II, II and respectively ATM, a group with preferentially negatively charged amino acid residues, CMGC IV and AGC IV, and a large group of targets with an arginine and lysine at the second or third position relative to the central serine, CaMK-Group and AGC-Group except the AGC IV sub family. For kinase families PKA, PKC, as well as CKII and MAPK most targets with resolved structure were available and were used for kinase family-specific predictors in this study.

These enrichments are well captured by the sequence logos and are also reflected in the RCP-plots for the spatial environment considering sequence-local residues. In addition to the detected enrichments (red-colored segments), the RCP-plots

also highlight significant depletions of amino acid types (blue segments, **Figure 6.2**) that are not immediately apparent from the sequence logo plots alone.

In the following, we investigated the targets associated with the main kinase families in more detail. In particular, we were interested to uncover potential 3D-signature motifs beyond the established sequence motifs that can be revealed when investigating individual kinase families rather than across all sites. Such motifs would become evident as colored segments found in the "non-sequence-local" graph, but not found in the "sequence-local" graph. We will refer to those motifs as 3D-signature motifs. We showed the RCP-plots for kinase families with 12 or more targets with known 3D-structures limiting our analyses to only selected kinase families.

Serine Sites

The AGC group

The AGC kinase group consists of kinases recognizing serine targets with an arginine or lysine residue at a distance of 2-3 residues relative to the central serine within the local protein sequence and includes the PKA and PKC as well as GRK, BARK, MARK, PKB, PKG and RSK kinase families which are not included in the study of spatial motifs presented here for paucity of corresponding data. Furthermore, the local sequence-based spatial profile is characterized by lower than expected occurrences of tryptophan and glutamate. Interestingly, the elevated occurrences of the positively charged amino acids arginine and lysine – the hallmark for the AGC kinase group – appear confined to the sequence-local neighborhood. An enrichment of arginine or lysine in the spatial context of PKA was not detectable. In the structural neighborhood ("non-sequence-local" graphs), the counts for both amino acids are not increased relative to the reference distribution. The PKC motifs exhibit an additional enrichment of serine in the sequence-local neighborhood, accompanied by a pronounced depletion of the amino acid residues histidine, glutamate, and tryptophan. The PKA motifs were observed to be depleted of the amino acid cysteine. For both families, PKA and PKC, a depletion of the hydrophobic amino acids alanine and leucine in the non-sequence-local neighborhood and an additional depletion of isoleucine in PKA motifs was detected (**Figure 6.4**).

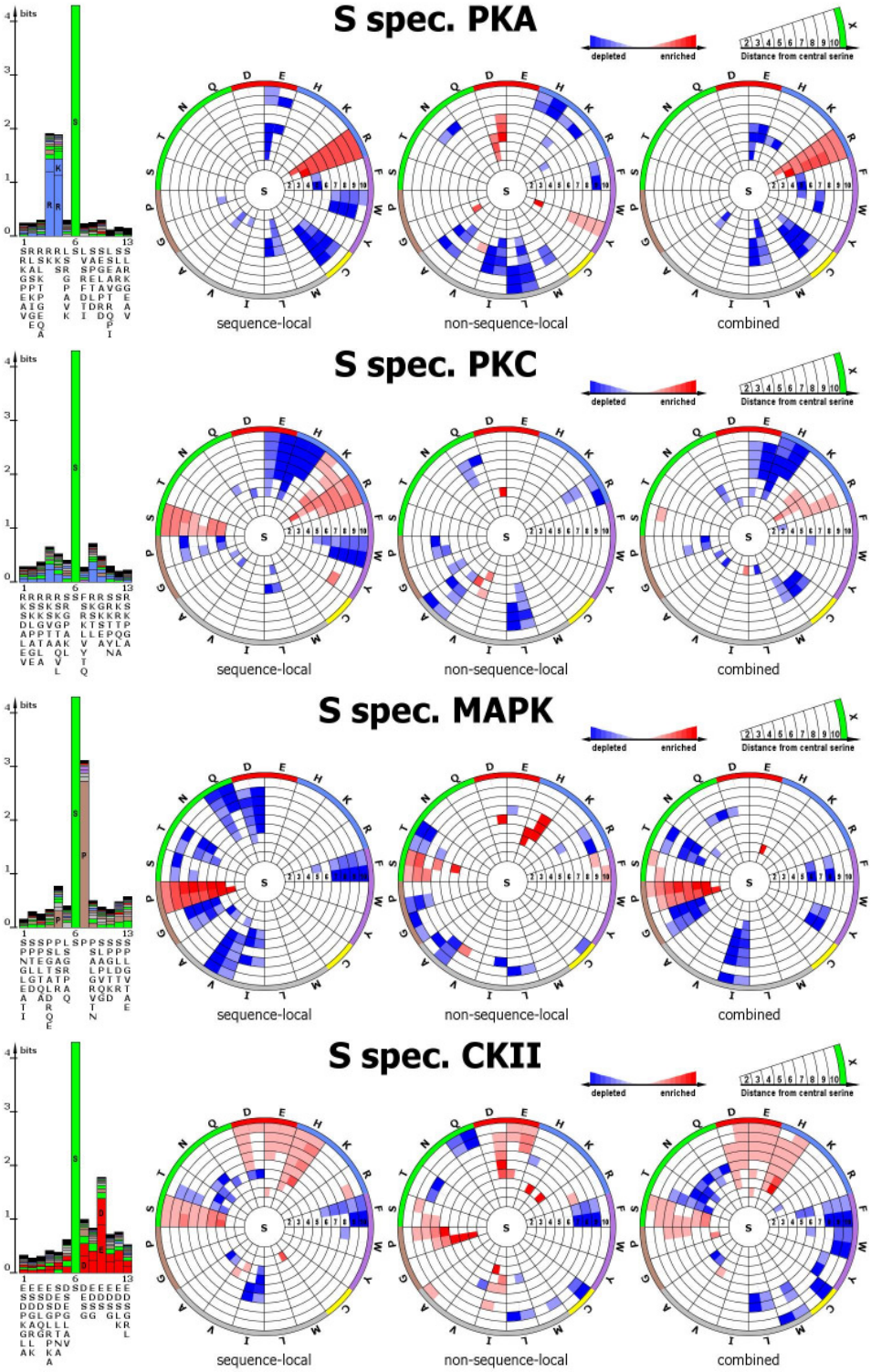


Figure 6.4 Sequence logos and radial cumulative propensity plots (RCP-plots) of serine kinase specific sequence motifs, illustrating enrichment as well as depletion of particular amino acid types in the local sequence (sequence logo), sequence-local spatial environment including the 6 flanking amino acid residues on either side of the central serine/threonine/tyrosine, (left RCP-plot), spatially-local, but non-sequence local; i.e. excluding residues in the flanking sequence (middle plot), and combined information (right RCP-plot). For every amino acid type, the two different sub-sectors correspond to the statistics obtained by using the closest detected atom and the interaction center, respectively, and in clockwise order.

The CMGC group

Proline residues flanking the phosphorylated serines are the hallmark sequence feature of targets associated with CMGC kinase group which includes the CDK, CKII kinase families (**Figure 6.3**) as well as MAPK and CDC. The CKII and MAPK were included in the spatial study as the number of structurally annotated targets was sufficient. The CKII family from the CMGC IV group, even though grouped into the CMGC group, does not follow the Pro-next-to-Ser rule. Its location in the serine-kinase phylogenetic tree is near the branching point between the CMGC branch and ATM family.

In the sequence-local environment of the MAPK, no enrichments of amino acids besides proline were detectable. Instead, depletions of eight amino acid types, glutamine, asparagine, phenylalanine, isoleucine, valine as well as glycine, serine and threonine were detected. In the non-sequence-local environment of target serines, serine and histidine residues were observed to be overrepresented.

The active sites of CKII kinases are characterized by positively charged surfaces [129]. This positive charge density is mirrored by negatively charged aspartate and glutamate in the sequence-local and non-sequence-local spatial neighborhood. Furthermore, the RCP-plots reveal enrichments of serine and histidine the sequence-local and proline in the non-sequence-local RCP pattern. A depletion of phenylalanine is observed at distances of 7 Å and greater for both patterns, while a depletion of threonine, asparagine and isoleucine is only detectable in the sequence-local spatial context.

Tyrosine sites

The PTK group

The PTK group comprises tyrosine phosphorylating kinases and as such is not included in the introduced phylogenetic tree of serine targeting kinases. The sequence-local spatial context of SRC-kinase family (PTK I) – for which sufficient data for analysis was availabe - is enriched in aspartate, proline, leucine, alanine, and tryptophan in the non-sequence-local spatial context. Depletions of several amino acids were also detectable, most consistently cysteine (**Figure 6.5**).

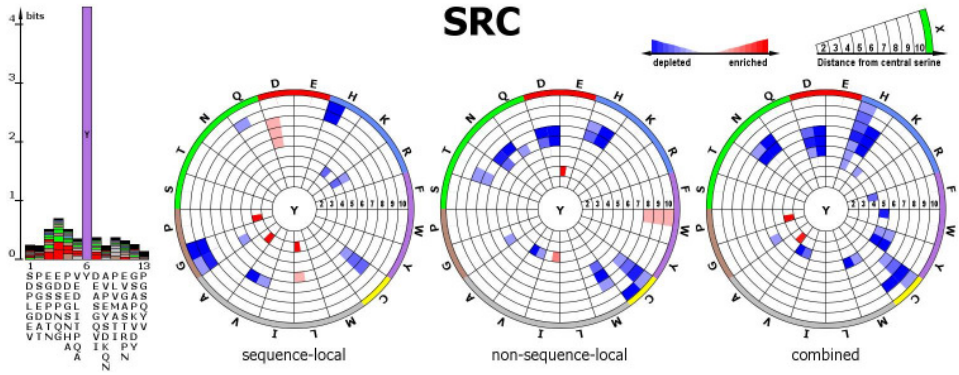


Figure 6.5 Sequence logos and radial cumulative propensity plots (RCP-plots) of SRC kinase specific sequence motifs. Notation as in **Figure 6.4**.

In summary, all kinase-family specific RCP-plots reveal specific spatial profiles and more information contents than were detectable when sites were investigated across all kinase families (**Figure 6.4** and **Figure 6.5**). The profiles comprise signatures of sequential motifs and discern spatial preferences which cannot be identified by inspecting the local sequence alone. All profiles show significant patterns of enrichments as well as depletions of particular amino acids within the spatial neighborhood of the phosphorylated target amino acid

6.2.2 Computational prediction of phosphorylation events using 3D-information

We now turn to investigating whether incorporating 3D structural information can be used to improve the sensitivity and specificity of phosphorylation site predictions.

Comparative analysis of prediction performance, Kinase-family-specific predictions

For the general, kinase-family unspecific prediction of phosphorylated serine, threonine, and tyrosine sites, the SVM-predictors based on local sequence information alone that have been developed as part of this study were observed to perform at comparable or even slightly better performance levels compared to NetPhos and DisPhos, and consistently better compared to KinasePhos as judged by the area under the receiver operating characteristic (AUC) from 10-fold cross-validation test (**Table 6.2**).

Similarly, for the kinase-family specific predictions, the AUC-based performance of NetPhos and our SVM-based method was comparable or even in favor of our SVM (**Table 6.2**) giving us an appropriate best possible sequence-information-alone baseline to assess the effect of adding 3D-structural information on the prediction accuracy when added to the SVM.

While the magnitude of performance gain when including 3D-profile information was relatively small compared to the estimated standard error, for all target sites and across all kinase-families and target residue types, a consistent increase in performance was obtained suggesting that including 3D information does indeed improve the sensitivity and specificity of phosphorylation site prediction.

Similar conclusions can be drawn from comparing prediction accuracies rather than AUCs (**Table 6.3**). Unlike in the case of AUC, where it was impossible to compute AUC values for the KinasePhos 2.0 prediction program because of non-returned score values, here it was possible to obtain relevant values also for the KinasePhos 2.0 prediction program. Again, adding 3D-information to using only sequence information resulted in modest (up to 5 percentage points), yet consistently improved predictions for all three target amino acid types as well as kinase-family specific targets such that best prediction results were always obtained by using our 3D-information enriched SVM-based

prediction method with the exception of the kinase families PKC and MAPK for which the performance was virtually identical compared to our sequence-only SVM, but still better than the other prediction programs included in this study. The most significant gain was obtained for serines sites followed by tyrosine and threonine sites.

Table 6.2 Results from the cross-validation of the various prediction approaches

Kinase family	Kinase group	N	sequence-only	Spatial-information enriched	NetPhos 3.1b	Disphos 1.3	KinasePhos 2.0*
Ser kinases	/	363	0.74 ± 0.02	**0.79 ± 0.02**	0.69 ± 0.02	0.73 ± 0.05	0.63 ± 0.05
PKA	AGC I	34	0.91 ± 0.04	**0.94 ± 0.04**	0.91 ± 0.03		
PKC	AGC II	31	0.83 ± 0.05	**0.87 ± 0.04**	0.78 ± 0.05		
MAPK	CMGC II	12	0.89 ± 0.07	**0.91 ± 0.06**	0.78 ± 0.09		
CKII	CMGC IV	19	0.73 ± 0.07	**0.78 ± 0.07**	0.76 ± 0.07		
Thr kinases	/	134	0.72 ± 0.03	**0.74 ± 0.03**	0.66 ± 0.03	0.72 ± 0.06	0.66 ± 0.05
Tyr kinases	/	253	0.69 ± 0.02	**0.71 ± 0.02**	0.65 ± 0.02	0.56 ± 0.06	0.54 ± 0.05
SRC	PTK I	24	0.72 ± 0.07	**0.79 ± 0.06**	0.62 ± 0.07		
unspecific predictor		750	0.71 ± 0.01	**0.75 ± 0.01**	0.67 ± 0.01	0.68 ± 0.03	0.63 ± 0.03

The sequence-only and Spatial-information enriched methods were developed as part of this study and compared to NetPhos 3.1b that includes the kinase-specific predictor NetPhos/K, DisPhos1.3 and KinasePhos2.0. As KinasePhos reports only decision values of positively predicted sites, the evaluation of kinase specific prediction was not possible due to missing score values for sites not predicted to be phosphorylated. However, the kinase-specific predictions were feasible as KinasePhos essentially reports all submitted sites as being phosphorylated by at least one kinase. For the evaluation of the predictor, the highest reported decision value was used for each site. Best performing methods are printed in bold-face.

Table 6.3 Prediction Performance as measured by accuracy

Kinase family	Kinase group	N	sequence-only	Spatial-information enriched	NetPhos 3.1b	Disphos 1.3	KinasePhos 2.0
Ser kinases	/	363	0.69 ± 0.01	**0.73 ± 0.01**	0.64 ± 0.01	0.68 ± 0.01	0.5
PKA	AGC I	34	0.83 ± 0.03	**0.88 ± 0.02**	0.82 ± 0.02		0.71 ± 0.02
PKC	AGC II	31	**0.82 ± 0.02**	**0.82 ± 0.02**	0.72 ± 0.02		0.65 ± 0.03
MAPK	CMGC II	12	**0.89 ± 0.04**	**0.89 ± 0.04**	0.69 ± 0.02		0.61 ± 0.05
CKII	CMGC IV	19	0.70 ± 0.03	**0.74 ± 0.04**	**0.74 ± 0.02**		0.62 ± 0.03
Thr kinases	/	134	0.68 ± 0.01	**0.69 ± 0.01**	0.63 ± 0.01	0.66 ± 0.03	0.5
Tyr kinases	/	253	0.65 ± 0.01	**0.67 ± 0.01**	0.62 ± 0.01	0.53 ± 0.02	0.5
SRC	PTK I	24	0.70 ± 0.03	**0.75 ± 0.03**	0.57 ± 0.01		0.70 ± 0.04
unspecific predictor		750	0.66 ± 0.01	**0.69 ± 0.01**	0.63 ± 0.01	0.62 ± 0.01	0.5

Results from the cross-validation of the various prediction approaches. The sequence-only and Spatial-information enriched methods were developed as part of this study and compared to NetPhos 3.1b that includes the kinase-specific predictor NetPhos/K, Disphos1.3 and KinasePhos2.0. The size of the negative set was adjusted to the size of the positive sites, ensuring equal sizes of the sets and a comparison to original reports of accuracies of alternative prediction approaches. In the case of the kinase unspecific prediction of KinasePhos2.0, all sites were predicted to be phosphorylated by at least one kinase. Best performing methods are printed in bold-face.

6.3 Discussion

In this work, we focused on the characterization and prediction of phosphorylation sites. Serine is the most frequent target amino acid residue type for phosphorylation followed by threonine and tyrosine. Here, we pursued two major themes: the analysis phosphorylation in a kinase family specific fashion, and the investigation whether phosphorylation sites are characterized by specific 3D structural motifs or epitopes constituted by amino acid residues that are not necessarily close in sequence, thereby providing additional information that can help in predicting phosphorylation sites for proteins with known structure or with available structural models. We used the simple radial distance to define structural motifs. Ideally, one would also include angular information as well. However, much larger datasets with determined structures would be necessary to derive reliable statistical data for more refined approaches. Even by applying only this simple model, we observed that 3D-structural context information is indeed discernable, even though most of the information contents appears to reside primarily in the local sequence, as judged by the sequence-local kinase unspecific RCP-plots and the modest increased performance when adding spatial information to sequence-only based predictors. The most pronounced improvement of prediction of phosphorylation sites by augmenting sequence-only prediction by spatial information was obtained for targets of serine kinases. However, also for the prediction of threonine and tyrosine targets a performance gain was obtained when adding 3D information.

As several experimental techniques have been established to detect proteins that specifically bind to phosphorylated sites based on immobilized peptides (pull-down assays and peptide chips [130; 131; 132]); i.e. the binding epitope is reconstituted from the sequence-local amino acid residues alone, the results obtained in this study lend further support to such approaches. Based on the findings obtained for our dataset, spatial information is discernable, but may not be absolutely critical to define the binding epitope, although the proving will require experimental comparisons of binding efficiencies for known interacting partners based on the complete as well as local peptide sequence.

It has been reported that phosphorylation is preferentially occurring in unstructured; i.e. flexible regions of proteins [109]. These conclusions resulted from sequence-based predictions of flexibility of phosphorylated and non-phosphorylated sites and is also supported by the reasonable prediction performance by DisPhos1.3 for serine and threonine. The prediction of phosphorylation sites by DisPhos is based on a prior predication of local flexibility. However, many phosphorylation sites were found in regions of clearly defined secondary structures (**Figure 6.1**). We further investigated this by comparing the crystallographic B-factor as well as secondary structural class for phosphorylated and unphosphorylated serine sites (**Figure 6.1**). In the latter, loop

regions may represent rather unstructured segments, even though it does not mean that this regions are structurally flexible. Flexibility may be better captured by the reported B-factor. There were statistically significant differences of B-factors for phosphorylated compared to non-phosphorylated serine sites detectable, albeit the differences are not that large. Of course, we only included those proteins in our investigation with an available crystallographic structure; including atomic coordinate information for the targeted peptide segment itself. It may be possible that, by only using fully resolved structures that we needed in order to detect possible 3D-motifs, we excluded phosphorylation events in unstructured regions right from the start. Follow-up studies need to be performed to address this question more systematically by mapping sites that were found in peptide-based methods (mass spectroscopy) and to map them to available protein structures and to gather statistics how often phosphorylation sites map to regions that cannot be resolved crystallographically.

A major problem in any effort to develop a computational predictor arises from the difficulty to define a reliable true-negative set; *i.e.* sites that are truly unphosphorylated. As phosphorylation is condition-dependent, experimental screens may well be incomplete as it is impossible to explore all environmental conditions under which phosphorylation events may occur. Even sites that are buried and inaccessible for phosphorylating kinases in one protein state may become exposed upon conformational changes and become phosphorylated [111; 133]. Thus, even burial state cannot be used to rule out phosphorylation. Even more so, as the numerical value for solvent accessibility may oftentimes suggest that a serine is buried, even though it is actually a surface residue, but occluded by neighboring side chains and not buried deep in the protein's core. Indeed we observed serine and tyrosine sites to be significantly, in term of significant *p*-values, more exposed to the solvent. However the inspection of the distributions of the accessibility yields only little differences. The assumption that buried amino acids cannot become phosphorylated and using it as criterion for the construction of a negative set may, in fact, be misleading. The resulting predictors will tend to predict accessibility of target sites rather than the possibility of phosphorylation. An alternative way for defining negative sets is including all candidate sites (serine, tyrosine, or threonine residues) except experimentally verified phosphorylation sites with the reasoning that such a true-negative set will at least be depleted in true-positive sites. In this study, we followed this approach, realizing that this may represent a source of error.

An estimated two to five percent of eukaryotic genomes codes for kinase genes grouped into different kinase families [98; 134] and 30% of all proteins are estimated to be phosphorylated as judged by proteomics screens [114; 135]. Mirroring the many different kinases catalyzing the addition of phosphate group to proteins, the high diversity of their cognate phosphorylation target sites is a major obstacle for a reliable prediction of phosphorylation. In addition, experimental evidence suggests that the kinases are to

some degree unspecific and are capable of phosphorylating a wide spectrum of substrates [114]. On the other hand, evidence for sequence-encoded specificity on the side of phosphorylation target has also been presented. For example, the prediction accuracy of phosphorylation sites in plant proteins was shown to increase substantially when the computational methods were trained on plant proteins versus methods trained primarily on animal proteins suggesting kingdom specific differences of phosphorylation target sites [123; 136].

The high diversity of targets of particular kinases and the number of possible phosphorylated proteins accompanied with the pleiotropicity of kinases appear to contradict a specific regulatory role of phosphorylation. However, the specificity for the actual target site may not be the only source of kinase specificity and sensitivity of the regulatory system. In fact, it was shown that subcellular compartmentation accompanied with recognition of secondary target sites relatively distant to the catalytic domains are crucial for further selectivity and specificity. While 3D motifs near the actual target site for phosphorylation have been at the center of our investigations, for the kinase family CDK, in particular the kinase CDK2 [137], it has been reported that secondary sites, protein surface site distant from the actual phosphorylation site may determine binding specificity of kinases with their target protein. The systematic identification and characterization of such secondary recognition sites appear therefore worthwhile [138]**.** Kinase activation in kinase cascades by posttranslational modification, formation of protein complexes as well as priming of phosphorylation further enhance the sensitivity of the phosphorylation system [114].

In summary, inclusion of 3D-structural information led to a small, yet consistent improvement of prediction accuracy. The reliable prediction of phosphorylation sites and the identification of associated kinase enzymes are important steps that will ultimately lead to a deeper understanding of complex signaling events in cellular systems.

6.4 Methods

Creation of phosphorylation site datasets (phos-Sets)

The dataset of phosphorylation sites was obtained from the Phospho.ELM database [110]. The amino acid residue annotated as phosphorylated (Ser/Thr/Tyr) was placed in the middle position of the 13-mer peptide with six amino acid residues on either side flanking the central position extracted from the native sequence of the respective protein harboring the site. Incomplete (i.e. truncated) motifs were discarded. The data set comprised 14,629 putative motifs (10,769 serine, 2,095 threonine, and 1,765 tyrosine motifs). To identify associated protein structures and the actual conformations and locations of the peptide motifs within their three-dimensional context, we screened the Protein Data Base (PDB) for protein structures containing the 13-mer peptide sequence motifs associated with phosphorylation sites based on exact sequence matches. The search yielded a set of 14,337 exact matches (Ser: 10,769, Thr: 2,095, and Tyr: 1,765 matches) in 6,596 different protein chains (Ser: 4,337, Thr: 2,086, Tyr: 2,765 chains) associated with 1,231 unique sequence motifs (Ser: 632, Thr: 240, Tyr: 359 unique motifs); i.e. many motifs were found multiple times in different protein structures. Considering only structures with complete atomic coordinates for the phosphorylation motif and choosing the structure with the best crystallographic resolution in case of identical sequence motif hits, we obtained a set of 750 non-redundant structural phosphorylation motifs (Ser: 363, Thr: 134, Tyr: 253 structural motifs). For a subset comprising 307 motifs (Ser: 164, Thr: 59, Tyr: 84 motifs), information of their respective phosphorylating kinases was available as well, and the associated motifs were classified into respective kinase families.

Creation of non-phosphorylation site datasets (non-phos-Sets)

We removed the phos-Set motifs from the sequences of the respective protein chains with known protein structure. From the remaining sequence fragments, we extracted all non-overlapping Ser/Thr/Tyr site motifs. The resulting sets of sites served as the true-negative set. While our approach cannot guaranty that these extracted sites are truly unphosphorylated, we expect this dataset to be at least depleted in true phosphorylation sites.

When kinase-family-specific phosphorylation events are analyzed, the true-positive counts are heavily outnumbered by true-negative sites posing the risk of dominating influences of the negative set rather than the positive set. In particular, the false-negatives; i.e. sites that we grouped as unphosphorylated that may, however, become phosphorylated under different conditions may then obscure any discernible signal. To alleviate this problem, while at the same time keeping a sufficient number of

examples for training purposes, we reduced the negative set for kinase specific predictions by randomly eliminating sites from the non-phos-Set until the negative sets were no more than twice as large as the positive sets.

Construction of the phylogenetic tree of serine-kinases

Sequence motifs associated with putatively phosphorylated serines are partly annotated with their respective phosphorylating kinase and can be grouped into kinase families and groups according to the classification scheme proposed by Hanks and Hunter augment by the AURORA and ATM kinase group. Considering only kinase groups with known targets, a phylogenetic tree (dendrogram) was built from representative sequences using the CLUSTALW package [139; 140]. For each group, we calculated sequence logos from all respective targets, i.e. also targets which are not represented in the protein database PDB [125].

General structural properties of phosphorylated and unphosphorylated sites

Secondary structural assignments, relative side chain accessibilities and crystallographic B-Factors were obtained from the PDBFINDER II database (ftp://ftp.cmbi.ru.nl/pub/molbio/data/pdbfinder2/) [124; 141; 142].

Calculation of spatial amino acid propensity profiles, Radial Cumulative Propensity (RCP) plots

Propensity ratios (odds-ratios) defined by the normalized counts of a particular amino acid type around sites in the phos-Set relative to their counts observed around sites in the non-phos-Set representative set within radial distances ranging from 2 to 10 Å from the central Ser/Thr/Tyr were calculated according to Equations 6.1. We used two different distance measures. Amino acid residues were considered to lie within a given radial cutoff distance if i) the distance between the putatively activated oxygen (ß-hydrogen) in case of a central serine and threonine, or γ-carbon in case of tyrosine and any atom of that residue was shorter than the given cutoff distance, or, if ii) the distance between the interaction centers of residues as proposed by Park et al. [143] fell within a given radial distance cutoff. Radial distance-dependent propensity ratios for all 20 amino acid types are illustrated graphically in radial cumulative propensity plots (RCP-plot). These plots reflect the cumulative spatial amino acid residue propensity profile around phosphorylation sites. We differentiate between radial profiles associated with i) sequence-local amino acids, i.e. amino acid residues located within 6 residues from the central serine in the protein sequence, and ii) non-local amino acid residues; i.e. residues that are outside the local sequence environment (>6 residue positions), and, iii), the general spatial profile irrespective of the amino acid position in the protein sequence. The

20 radial sectors associated with the different amino acid types are divided into two sub-sectors according to the two different distance measures used. The significance of the obtained propensities for increased or decreased occurrences relative to random expectation around phosphorylated sites was assessed by estimating the standard error of the odds-ratios, S_E , as proposed by Levitt (Equations 6.1) [144]. Odds-ratios signifying over- or underrepresentation were considered statistically significant if odds-ratio > 2 and (odds-ratio – S_E) > 1 with odds-ratios inverted in cases where the propensity ratio was below 1; i.e. observed less than expected by chance.

Eqs. 6.1.

$$f_{k,r} = \frac{\#AA_{k,r,\text{in pos. set}}}{\#AA_{s,r,\text{in pos. set}}} \qquad g_{k,r} = \frac{\#AA_{k,r,\text{in ref. set}}}{\#AA_{s,r,\text{in ref. set}}}$$

$$odds\text{-}ratio_{k,r,i} = \begin{cases} g_{k,r} / f_{k,r}, & f_{k,r} \le g_{k,r} \\ f_{k,r} / g_{k,r}, & f_{k,r} > g_{k,r} \end{cases}$$

$$S_{E\ k,r,i} = \begin{cases} f_{k,r}^{-1} \dfrac{\sqrt{g_{k,r}(1-g_{k,r})}}{\#AA_{s,r,\text{in ref. set}}}, & f_{k,r} \le g_{k,r} \\ g_{k,r}^{-1} \dfrac{\sqrt{f_{k,r}(1-f_{k,r})}}{\#AA_{s,r,\text{in pos. set}}}, & f_{k,r} > g_{k,r} \end{cases}$$

Calculation of amino acid propensity ratios for the estimation of average depletion or enrichment given a particular motif set. #AAk/s is the count for amino acids, where k is the amino acid type and s is the count for all amino acids; r is the considered radius of the distance to the central serine/threonine/tyrosine, f is the relative frequency of amino acid in a particular set, and g the relative frequency of the amino acid k in the reference, non-phos set.

Prediction approach, evaluation of prediction performance

To predict phosphorylation sites from sequence and to evaluate the effect of using structural information on prediction performance, we applied Support Vector Machines (SVMs), first using sequence information alone and, subsequently, enriched by the spatial information. We used the "kernlab" R-package developed by Alexandros and co-workers[145] applying the default "Radial Basis kernel" with automated "sigma estimation". We evaluated the area under the Receiver Operator Characteristic (ROC)-curve (AUC) from a 10-fold cross-validation to quantify the performance of predictors and to compare the obtained results to prediction results obtained by using NetPhos and DisPhos [96; 108; 109].

The 10-fold cross-validation was based on training of the predictor on 9 out of 10 parts of the randomly ordered data set and subsequent classification of the remaining part. The test is repeated for all 10 possible partitions of the dataset. The classification results are then used for measuring the performance of the predictor. The developed

classifiers based on Support Vector Machines included general, kinase-family unspecific serine, threonine and tyrosine predictors; i.e. the parameters were trained across all proteins irrespective of annotated kinase family, as well as predictors specific for the serine-centric PKA, PKB, MAPK, and CKII kinase family as well as tyrosine-centric SRC kinase family, for which at least 12 annotated targets or more were contained in the dataset. For threonine target sites, the respective kinase-family annotation information yielded only data sets of insufficient size for statistical analyses. The area under the ROC-curve (AUC) from 10-fold cross-validation was compared among different prediction approaches and programs to judge, whether the addition of spatial information can improve the prediction performance. Perfect prediction results would yield an AUC of 1, while guessing the outcome would, on average, yield AUCs of 0.5.

Feature-vectors (FV) for the implemented Support Vector Machines

The feature-vector (FV) used for the Support Vector Machines consisted of chemical-physical amino acid properties for the sequence-information-only approach and an additional spatial information component for the spatial prediction approach. For the amino acid property components of the FV, we utilized values from the collection of 530 commonly used indices provided by the AAindex database [146] including hydrophobicity, solvent accessibility preferences, secondary and tertiary structure preferences, polarity, volume and solvent accessibility, structural disorder indices and others. The vector consisted of 530 x 12 dimensions for every index and position around the central serine, threonine, or tyrosine, where the components were values from the respective index and 530 dimensions for the average index value of the particular sequence motif. The latter dimensions were introduced to cover the general properties of the motifs, e.g. negative charge or high flexibility. To reduce the dimensionality of the FVs as well as to eliminate correlations between components, principle component analysis (PCA) was performed on the $D_{FV} x N$ data matrix, where D_{FV} is the number of components of the FV, and N is the number of example peptide sequences in the training set, and the components of the FV were replaced by the resulting principle components with non-zero Eigenvalues explaining the entire variance in the dataset. Note: as there are fewer examples (N peptide sequences in the training set) than dimensions, the dimensionality (Eigenvectors with non-zero Eigenvalues) can be at most $N - 1$. The PCA was performed independently for the serine, threonine, and tyrosine motifs. The entire variance in all independent datasets was covered by 228 principal components.

The spatial information component consisted of the normalized distribution ratios according to Equations 6.2. The ratios of amino acid residues within the local sequence, outside the local sequence, and irrespective of the position in the protein sequence were used for distances in a range of 2 to 10 Å between the putatively activated oxygen (β-

hydrogen) in case of a central serine and threonine, or γ-carbon in case of tyrosine and the closest atom of all other amino acid residues, or between the interaction centers proposed by Park and coworkers [143].

Eqs. 6.2

$$f_{k,r,i} = \frac{\#AA_{k,r,i}}{\#AA_{s,r,i}} \qquad g_{k,r} = \frac{\#AA_{k,r}}{\#AA_{s,r}}$$

$$odds\text{-}ratio_{k,r,i} = \begin{cases} 1 - g_{k,r} / 2f_{k,r,i}, & g_{k,r} <= f_{k,r,i} \\ f_{k,r,i} / 2g_{k,r}, & g_{k,r} > f_{k,r,i} \end{cases}$$

Calculation of spatial propensity ratios of amino acid distributions for usage in Support Vector Machines. #AAk/s number of amino acid, where k is the amino acid type and s are all amino acids, r is the considered radius of the distance to the central serine/threonine/tyrosine, motif i from the sample set, f is the frequency of amino acid in a motif i, g is the frequency of the amino acid k in the representative non-phos set.

Comparison to NetPhos, Disphos-1.3 and KinasePhos2.0

We compared the AUC from the 10-fold cross-validation results obtained by using NetPhos, NetPhosK, DisPhos, KinasePhos2.0. NetPhos and NetPhosK are both part of the NetPhos-3-1b package. While NetPhos was designed to generally predict serine, threonine, and tyrosine phosphorylation events, NetPhosK includes kinase-specific predictors. DisPhos is based on SVM, utilizing the binary representation of the motif sequence, the relative frequencies of amino acids in that sequence as well as outputs from predictors for structural disorder and secondary structure[109]. Furthermore, the Feature Vectors are supplemented by amino acid properties covering the sequence complexity, net-charge and aromatic content, hydrophobic moment, and hydrophobicity as well as values according to a flexibility and surface exposure scale. Thus, the features used by DisPhos are comparable to features applied in our predictors. While NetPhos and DisPhos are predictors for phosphorylation events, KinasePhos2.0 was developed to identify the respective kinase[147], comprising over 50 kinase-specific predictions. The server is reported to yield highly accurate results also for the general prediction of phosphorylation events and, therefore, a good benchmark for the kinase specific predictors developed here.

For comparison with DisPhos, we submitted 60 randomly selected protein sequences covering at least 50 positive and 100 negative motifs for serine, threonine, and tyrosine sites to the DisPhos 1.3 server "http//core.ist.temple.edu/pred/pred/predict". Although 60 protein sequences are only a small subset of the total of 869 protein structures, the sequences covered 14% serine, 37% threonine and 20% tyrosine sites. Sites being reported as predicted by similarity to the training sequences, as assigned by DisPhos

were removed to avoid self-recognition. A similar procedure was applied to KinasePhos 2.0, however the KinasePhos2.0 server as well as NetPhos do not provide information of possible self-recognition events. We submitted the above mentioned protein sequences to the KinasePhos 2.0 server "http://kinasephos2.mbc.nctu.edu.tw/" setting the specificity value to "default". The comparison proved difficult as only positively predicted (phosphorylated) sites, i.e. sites which were predicted by a decision value above 0.5 were returned. This made the computation of the AUC for specific predictors impossible, as not for all submitted sites, a decision value (score) was available. However, as essentially all sites from the training and test set, irrespective of whether they were positive or negative were predicted to be phosphorylated by the server by at least one kinase, assessment of the performance of the prediction by evaluation of the AUC for kinase unspecific prediction was still possible. Out of 1,335 submitted serines, 1,288 (97%) were predicted as being phosphorylated. The corresponding ratios for threonines were 1,098 out of 1,124 (98%), and tyrosines 713 out of 723 (98%). Before the evaluation of the ROC curve, for each site, the highest reported decision values were determined. For a meaningful comparison, the size of the results from DisPhos and NetPhos were adjusted to reflect a ratio of 1:2 between the positive and negative set. This was performed by random removal of results from the positive or negative set, respectively. Subsequently, the AUC was computed and compared.

Comparison to NetPhos, Disphos 1.3 and KinasePhos 2.0 judged by accuracy

As an alternative measure of performance, we also computed the accuracy defined as the proportion of correct predictions (true positive or true negative predictions) among the predictions made (Eq. 6.3):

Eq.6.3
$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N},$$

where T_P are true positive, T_N-true negative, F_P-false positive, and F_N-false negative predictions.

This measure also allows our prediction approach to be compared directly to other available prediction programs, especially KinasePhos 2.0. For computing accuracies, a decision threshold for the assignment of a site to a particular group must be set. The positive assignment threshold for our predictors was set to zero. Negative decision values were judged as predicted to be non-phosphorylated and positive decision values to be phosphorylated. For the other predictors, this value was set to 0.5 as they reflect probabilities. For kinase-specific predictions, sequences from the training set were

submitted to the KinasePho2.0. server. For assessing the performance associated with a particular kinase family, only the results of the corresponding family were evaluated as relevant predictions. As the prediction reports usually estimate the accuracies based on equal sizes of the positive and negative set, the negative sets were adjusted by random removal of the respective prediction results to reflect this ratio. This adjustment was performed 1,000 times with different random removals and the mean accuracy as well as the standard deviation was determined.

Chapter 7
PhosPhAt: A database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor

Abstract

The *Arabidopsis* Protein Phosphorylation Site Database (*PhosPhAt*) provides a valuable resource to the plant science community and can be accessed through the following link: http://phosphat.mpimp-golm.mpg.de. *PhosPhAt* aims to present a genome-wide and consolidated view of all detected phosphorylation sites in all proteins encoded in the *Arabidopsis* genome. Furthermore, it offers a computational predictor of phosphorylation sites trained on the experimentally verified sites contained in the database.

At time of publication of *PhosPhAt* the database contained 1,187 unique tryptic peptide sequences encompassing 1,053 *Arabidopsis* proteins. Among the characterized phosphorylation sites, there were over 1000 with unambiguous site assignments, and nearly 500 for which the precise phosphorylation site could not be determined. The database is searchable by protein accession number, physical peptide characteristics, as well as by experimental conditions (tissue sampled, phosphopeptide enrichment method).

Currently *PhosPhAt* comprises a total of 7,257 phosphosites contained in 6,282 phosphopeptides collected from several published datasets and is still growing. The phosphosites are mapped to 2,107 proteins. For each protein, a phosphorylation site overview is presented in tabular form with detailed information on each identified phosphopeptide.

We have utilized a set of 802 experimentally validated serine phosphorylation sites to develop a method for prediction of serine phosphorylation (pSer) in *Arabidopsis*. An analysis of the current annotated *Arabidopsis* proteome yielded in 27,782 predicted phosphoserine sites distributed across 17,035 proteins. These prediction results are summarized graphically in the database together with the experimental phosphorylation sites in a whole sequence context.

The increasing availability of proteome-wide phosphorylation data across different species and kingdoms will enable researchers to conduct cross-species comparative analyses to identify conserved phosphorylation signatures, as well as sites of evolutionary adaption mediated via species-specific phosphorylation.

7.1 Background

Phosphorylation is the most studied posttranslational modification (PTM) involved in signaling. The principle of activation and inactivation of proteins by phosphorylation as well as the function of phosphorylated residues as docking sites for protein scaffolds and complex assemblies has been well characterized in the field of mammalian signal transduction [148; 149; 150; 151]. In the field of plant biology, the focus so far has been on the analysis of phosphorylation of specific proteins and protein families [152; 153] and on the study of very specific signaling pathways [154; 155], mainly using genetic tools.

In recent years, several large-scale proteomics initiatives have been started to detect phosphorylation sites in various different species, including plants, using MS-based proteomics and to investigate the role of site-specific phosphorylation events in response to changing biological conditions [156; 157; 158; 159]. A number of global studies of plant protein phosphorylation sites have been carried out on various tissues and under a variety of biological conditions ranging from biotic and abiotic stresses to changing nutrient environments [160; 161; 162; 163]. These datasets were made available in large supplementary or printed tables with different specific information for each peptide, making these large tables difficult to handle in comparative analyses. There is currently no resource in the plant field that collects such information and makes it available to the community in a readily searchable format, thereby providing the possibility for added value through combined and comparative data interpretation.

While a number of phosphorylation databases are available, they are generally concentrated on studies undertaken in mammalian and prokaryotic systems. Phosida[164] (http://www.phosida.de/) contains large scale data from in house studies of *Homo sapiens* and *Bacillus subtilis*; The Phosphorylation Site Database (http://vigen.biochem.vt.edu/xpd/xpd.htm) contains phosphorylation information from prokaryotic organisms; Phospho.ELM[110] (http://phospho.elm.eu.org/) contains validated phosphorylation sites from eukaryotic systems but is heavily biased towards mammalian systems, while PhosphoSite[165] (http://www.phosphosite.org/) is a curated site that focuses on vertebrate systems. The model plant *Arabidopsis thaliana* is a significant focus of international plant research (http://www.masc-proteomics.org/ for *Arabidopsis thaliana* proteomics) and is currently only poorly represented by existing phosphorylation databases. Furthermore, although several computational algorithms to predict phosphorylation sites given the amino acid sequence of target proteins alone have been developed[97; 108], they all suffer from low accuracies applied on plant species[166]. Because experimentally determined sites (and non-sites) needed as training sets for the development of prediction algorithm have been largely limited to non-plant species, no dedicated plant-specific computational prediction algorithm was available. Correspondingly, the performance of available prediction programs was notoriously poor

when applied to plant proteins. For example, while generating many false positive predictions, the available prediction programs, trained primarily on mammalian phosphorylation sites, missed 40% of experimentally determined sites in the plant *Arabidopsis* [166]. Evidently, there seems to exist kingdom-specific differences of phosphorylation target site motifs as well as their cognate kinases[136]. Furthermore, many animal-specific kinases are not present in plants. Therefore, we believe that the *PhosPhAt* service combining experimental results with pSer prediction will be a valuable addition to current phosphorylation databases and to the plant research community in general.

Several other authors were involved in the establishment of the database. Their contributions were mainly focused on an intensive curation and selection of reliable datasets. Our contribution to *PhosPhAt* was the establishment of the database, from the initial step of formatting basic data tables and detection of inconsistencies in the data, to the setup of the database server and integration of a plant specific phosphorylation site prediction. Furthermore, a subsequent genome-scale prediction of the phosphorylation sites in *Arabidobsis* revealed first insights into the distribution of phosphorylation sites among the functionalities of the *Arabidobsis* genome.

7.2 Results

7.2.1 Database overview

The *PhosPhAt* database uses a MySQL relational database operating on a Linux based operating system. The web-based graphical user interface allows the construction of SQL (structured query language) queries through standard HTML forms. Complex database queries are created with pull-down menus that retrieve data through purpose-built PHP scripts that interact with the MySQL tables in *PhosPhAt*.

The database is comprised of two distinct tables (**Figure 7.1**): The first table (phosphat) contains the experimental phosphopeptide information and comprises data from several published large- and medium scale phosphoproteomic analyses [158; 160; 161; 162; 163] as well as unpublished sites identified in authors' labs. Each entry is a unique experimentally measured precursor ion (m/z)[c] and not a composite entry. This is an important feature of the *PhosPhAt* database as it tracks each piece of experimental data, and provides links also to the actual experimental mass spectra deposited in PROMEX (http://promex.mpimp-golm.mpg.de; [167]). With a link to this spectral library on the 'Result Table' users can download the precursor mass-to-charge ratio and the corresponding CID-spectrum[159; 168].

[c] A mass spectrum is an intensity vs. m/z (mass-to-charge ratio) plot, representing a chemical analysis. Hence, the mass spectrum of a sample is a pattern representing the distribution of components by the mass-to-charge ratio in a sample.

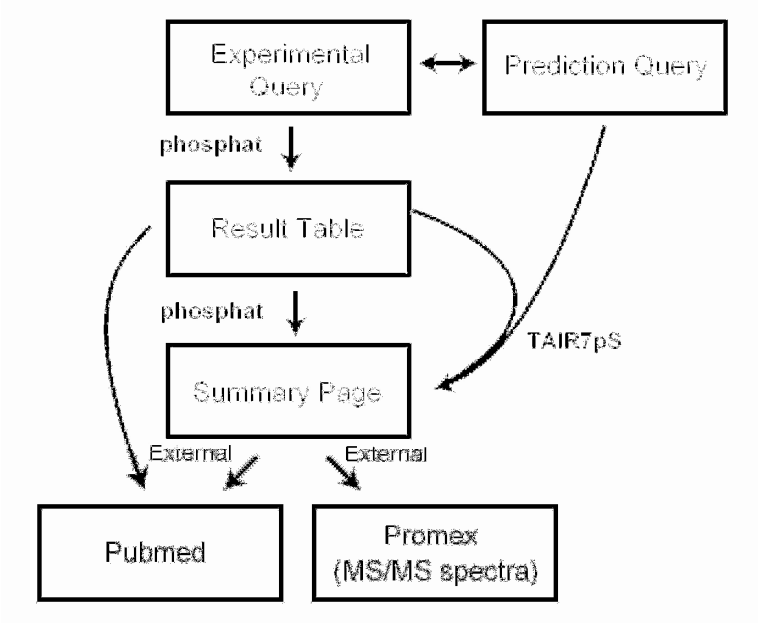


Figure 7.1 Schematic diagram outlining the structure of the *PhosPhAt* service illustrating the two main query entry points to query experimental data and pSer prediction information. Both services merge into a common output at the 'Summary Page' on which the prediction results are displayed on top of the page and all experimental phosphopeptides for the given AGI code are listed below. In instances where no experimental phosphopeptides are available, only the prediction result will be displayed. External links to published references at PubMed and MS/MS data at the ProMex mass spectral library [167] are also shown.

The second table, the prediction table (TAIR7pS), contains pSer predictions for the entire *Arabidopsis* annotated proteome comprising 31,921 proteins (release 7 from 2007-04-25) available from The *Arabidopsis* Information Resource (www.arabidopsis.org; [169]). The prediction table contains precompiled pSer prediction scores for a total of 928,449 serine residues. Phosphorylation sites are marked as 'defined' if the precise location of the phosphorylated amino acid has been unambiguously determined by mass spectrometric analysis. This usually implies manual interpretation of mass spectra and additional scoring algorithms [164]. These 'defined' sites are marked with brackets and a lowercase p, e.g. (pS), (pT), (pY). Phosphorylation sites marked as 'undefined' were not clearly resolved by the mass spectrometric experiments. These sites are marked as lowercase letters in brackets, e.g. (s), (t), (y). Often, the 'undefined' sites are two putatively phosphorylated amino acids in close proximity in the peptide and the difference between these options could not be interpreted based on the mass spectrum. The 'undefined' sites are often only a subset of the serines, threonines, or tyrosines in the tryptic peptide. If no statement can be made on the location of the phosphorylation site, the modified tryptic peptide sequence is displayed with the remark 'site not determined'.

At time of publication of *PhosPhAt*, the experimental data table contained 1,187 defined tryptic peptides matching 1,053 distinct proteins from the model plant *Arabidopsis thaliana*. Currently *PhosPhAt* comprises a total of 7,257 phosphosites contained in 6,282 phosphopeptides collected from several published datasets and is still growing. The phosphosites are mapped to 2,107 proteins. The distribution of these sites is 84% on serine, 13% on threonine, and 3% on tyrosine.

The entry page of the *PhosPhAt* database provides two general search strategies: (i) browsing multiple instances of experimental phosphorylation sites via the tab 'Query Experimental Data', and (ii) displaying a summary of phosphorylation site prediction of one locus with a concurrent display of experimental sites via the tab 'Query Prediction Data'.

The query via 'Experimental Data' provides access to the experimentally verified phosphorylation sites by physical parameters of the peptide (charge state, number of modifications, mass accuracy), methodological aspects (enrichment method, digesting enzyme, mass analyzer), biological context (tissue, cellular compartment, experimental condition), or research group (published datasets, research groups). A list of proteins of interest can also be submitted using the AGI gene code format. The user will then be directed to the 'Result Table' on which, depending on the query, all experimentally identified phosphorylated peptides are displayed for every protein in a tabular form. Each AGI code in the 'Result Table' provides a link to the 'Summary Page' outlining all experimental information for that locus as well as pSer prediction.

The 'Summary Page' details experimentally validated/identified peptides for a given AGI code with each phosphopeptide displayed in its own table. The database has been specifically designed to capture as much information as possible for each experimentally identified phosphopeptide and thus a 'composite' entry for each site has not been used. In many cases, site level redundancy in the form of multiple experimental phosphopeptide entries for one phosphorylation site can be observed on this page. Each phosphopeptide entry provides a link to MS/MS spectra housed in the ProMEX [167] database (if available; http://promex.mpimp-golm.mpg.de) as well as a link to the PubMed reference (if data published).

The 'Query Prediction Data' tab also serves as entry point to the database and allows queries using single AGI codes. This tab provides a direct link to the 'Summary Page' where experimental and pSer predictions for the AGI code entry are outlined for the amino acid sequence of the retrieved entry. As outlined above, this page also provides a detailed breakdown of all phosphorylation modification data (if available) for this locus.

7.2.2 The *Arabidopsis* pSer predictor

Protein phosphorylation is of paramount importance for understanding biochemical regulation. Because of restricted experimental approaches for *in vivo*-site determination, the computational prediction of phosphorylation sites is a complementary and helpful tool. Using the gathered experimentally verified data from our database as a training set, we used a Support Vector Machine (SVM) approach to classify candidate serine sites. Computed SVM decision values greater than zero indicate a positive prediction of a phosphorylation event, while negative values predict serine residues not to be phosphorylated. Greater absolute decision values indicate greater confidence in the prediction. In the 'Summary Page', candidate serines are tagged with mouse-over information pop-up-boxes of experimental evidence as well as prediction results (SVM decision value). In the displayed sequence, serines are colored red if experimentally verified and they are underlined when positively predicted with a *decision value* > 0 by the computational classifier.

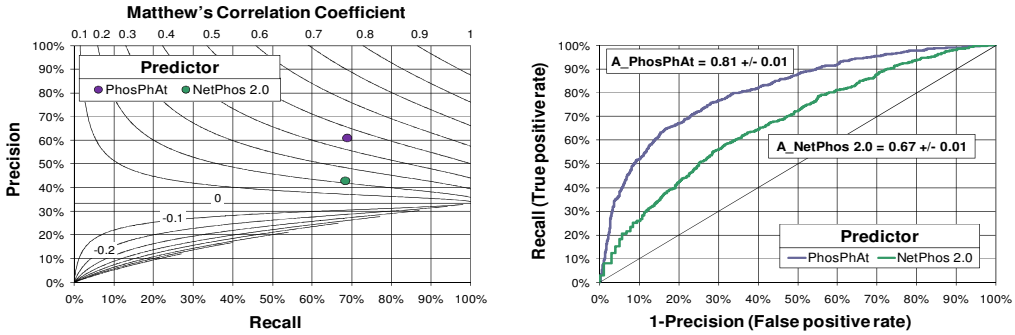


Figure 7.2 Prediction performance of the pSer predictor in comparison to NetPhos 2.0 [170]. The recall rate vs. the associated precision is plotted. The curved lines indicate lines of equal correlation coefficient. In the diagram, improved classification performance is indicated for predictors falling into the upper right corner. Performance results for our classifier correspond to results obtained in the 10-fold cross-validation. The classifier NetPhos 2.0 was applied to our dataset without training. b) Receiver operating characteristics curves of the prediction by pSer predictor in comparison to NetPhos 2.0 [170]. In the diagram, improved classification performance is indicated for predictors with increased area under the ROC. The area under the ROC curve was A1=0.81 ± 0.01 for the pSer predictor and A2=0.67 ± 0.01 for NetPhos and was significantly better with a z-score=(A1-A2)/SE(A1-A2) of 24.1 corresponding to a p-value of 3.3E-128 in the limiting case of a normal distribution according to the algorithm proposed in [171].

A comparison of the prediction performance of the plant-specific pSer predictor and the generic NetPhos 2.0 [170] reveals a significant improvement of recall, precision, as well as Matthew's Correlation Coefficient (CC) for *Arabidopsis* proteins (**Figure 7.2**). The CC reached with our plant-specific pSer predictor was 0.46 and, thus, significantly better than the CC for NetPhos 2.0 (CC = 0.22). In a 10-fold cross-validation test, 69% of phosphorylated serine sites from the training set were correctly recognized compared to

68% recall for the NetPhos 2.0 server. Of the predicted sites, 61% were experimentally verified phosphoserine sites while the precision achieved with NetPhos 2.0 was 43%. The comparison of the receiver operating characteristic (ROC) curves revealed a highly significant improvement of the prediction performance with a z-score of 24.1 according to the algorithm proposed by Hanley et al[171] corresponding to a p-value of 3.3E-128 in the limiting case of a normal distribution. The area under the ROC curve for the *PhosPhAt* plant-specific pSer predictor was 0.81 ± 0.01 and 0.67 ± 0.01 for NetPhos, respectively (**Figure 7.2**).

7.2.3 Genome-scale prediction of phosphorylation sites

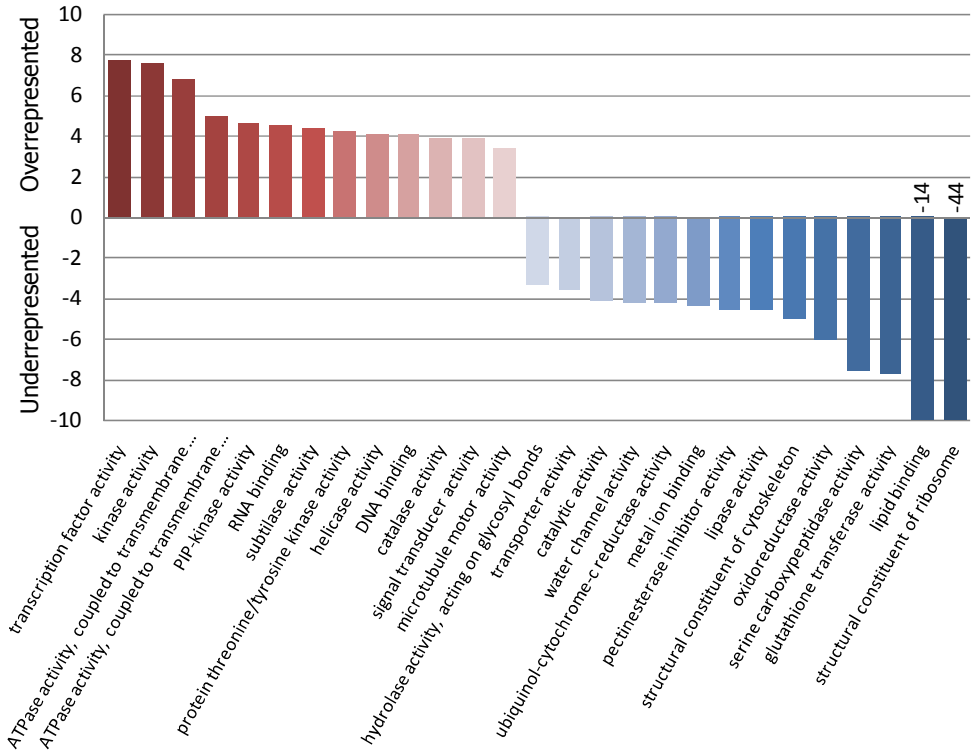


Figure 7.3 Negative log(pValues) from Fisher exact test on the occurrences of GO:function terms associated with predicted phosphoproteins. *P-values* were corrected for multiple testing by using the False Discovery Rate (FDR) formalism[172]. Overrepresented GO:terms are colored red, underrepresented blue. GO:terms were included if $p_{FDR}<0.001$. GO:annotations were taken from TAIR [169]. To avoid training bias, phosphorylation sites used during the training of the classifier have been removed in the Fisher exact test. Only GO assignments with evidence categories: direct assay, mutant phenotype, physical and genetic interaction as well as sequence of structural similarity have been considered.

The TAIR7pS table comprises a total of 928,449 serine site motifs in 31,921 protein sequences. Of those, 27,782 serines, distributed in 17,035 proteins (14,339 unique genes), about 2 phosphorylation sites per phosphorylated protein were predicted to be phosphorylated with high confidence (*decision value* > 1), which makes up approximately half of the annotated *Arabidopsis* proteome. For 176,442 serines, medium

confidence (0 < *decision value* < 1) was predicted and for 435,231 serines, the computed *decision value* was below -1 indicative of high-confidence negative predictions; i.e. no phosphorylation. The absolute *decision values* above 1 correspond to an accuracy of 95%.

In order to test for over- and under representation of predicted phosphorylation sites in different functional categories based on GO:annotations [173], we applied the Fisher exact test to the GO:term classified prediction result. Proteins involved in regulatory and signaling processes, including kinases, are significantly overrepresented in the set of highly confident phosphorylated proteins while housekeeping, structural components; e.g. cytoskeleton and other enzymatic functions are underrepresented (**Figure 7.3**).

7.3 Discussion

The *PhosPhAt* database has been initiated to provide a resource that consolidates our current knowledge of mass spectrometry-based identified phosphorylation sites in the model plant *Arabidopsis*. It is combined with a phosphoserine site prediction tool specifically trained on *Arabidopsis* serine phosphorylation site motifs. Thus, our database not only serves as a searchable knowledge base for experimentally identified phosphorylation sites, but in addition also provides a powerful resource for the characterization and annotation of yet unidentified phosphoserine sites in *Arabidopsis.* The value of the *PhosPhAt* resource thus lies in the possibility for comparative analysis of experimental sets, confirmation of experimental phosphorylation sites, by providing evidence from different published and unpublished sources, and in the implementation of prediction, where experimental evidence is not (yet) available.

Using the available experimentally verified *Arabidopsis* sites as a training set, a predictor based on a Support Vector Machine was developed. The performance of this plant-species-specific predictor was shown to generate significantly more accurate predictions than prediction programs trained on non-plant species thus confirming the kingdom and species-specific difference of the phosphorylation machinery. The resulting set of predicted sites comprises 27,782 high-confidence phosphorylated serine residues residing in 17,035 proteins; i.e., about 2 phosphorylation sites per phosphorylated protein (this ratio is about 2.8 for experimentally verified NR sites in *PhosPhAt*). Compared to the 2,107 proteins with detected phosphorylation sites contained in the *PhosPhAt* database, and considering the high accuracy of the *PhosPhAt* predictor, the potentially still large gap between the observed and the actual phosphoproteome becomes apparent. In a 10-fold cross-validation protocol, the *PhosPhAt* predictor generated a FPR of 22% while 70% of the predictions (TPR) turned out to be supported by experimental data. However, as with any score-generating classifiers, these performance rates (i.e., the trade-off between recall and precision) can be adjusted by setting the prediction score cut of value

appropriately. Since all computational methods rely on training sets of actual sites, they will inevitably inherit their respective bias towards certain proteins or sites as well. This bias can only be minimized by using diverse training data sets, also with regard to experimental platform.

When analyzing the functional annotations associated with proteins predicted to be phosphorylated, proteins involved in signaling and regulatory processes, including many kinases, were overrepresented, while house-keeping functions and structural component proteins; e.g., cytoskeleton, were associated more with the non-phosphorylated set. The predicted sites with highest decision values in combination with the experimental phosphorylation sites provide a powerful basis for further in-depth analysis of phosphorylation motifs in orthologous and paralogous proteins also between different organisms[174]. Thus, our dataset provides a rich resource for computational biologists interested in the study of conservation of phosphorylation sites and discovery of such conserved sites across protein classes and plant species. Furthermore, in short-term future, *PhosPhAt* will be extended by the prediction of threonine and tyrosine phosphorylation events and the precompiled phosphorylation site prediction will be replaced by the prediction algorithm. The integration will provide a toolbox for prediction of phosphorylation sites in submitted protein sequences also from different plant species. It remains to be seen how generalizable the *PhosPhAt Arabidopsis*-centric prediction method will turn out to be, when applied to phosphorylation sites detected in other plant species and whether a generic plant-phosphorylation prediction method will replace the *PhosPhAt* predictor or whether species-specific prediction methods may have to be developed.

7.4 Methods

Prediction of *Arabidopsis* phosphorylated serines using SVMs

A non-redundant dataset of 802 experimentally verified phosphorylated serines with defined sequence position in 859 unique *Arabidopsis* proteins and associated local sequence motifs consisting of 13 amino acid residues with 6 residues on either side of the central serine was used as the training set for the development of the computational classifier. Non-redundancy was assured by removing identical sequence motifs. As shown in **Figure 7.4**, by removing identical motifs and their associated protein sequences, a non-redundant protein dataset was obtained; i.e. proteins were non-redundant even when considering their entire amino acid sequence. The vast majority of proteins had sequence identity levels to any other protein in the dataset of less than 25%. After removing the positive motifs as well as experimentally identified phosphorylation sites of low confidence from the set of 859 proteins, all other remaining, non redundant and non-

overlapping serine site motifs formed the true-negative set. To avoid abundance bias towards the negative class (unphosphorylated serine sites), the raw set of 49,314 true negative serine sites was reduced by randomly eliminating sites from the set until the negative set was no more than twice as large as the positive set. This final datasets served as the true-negative set.

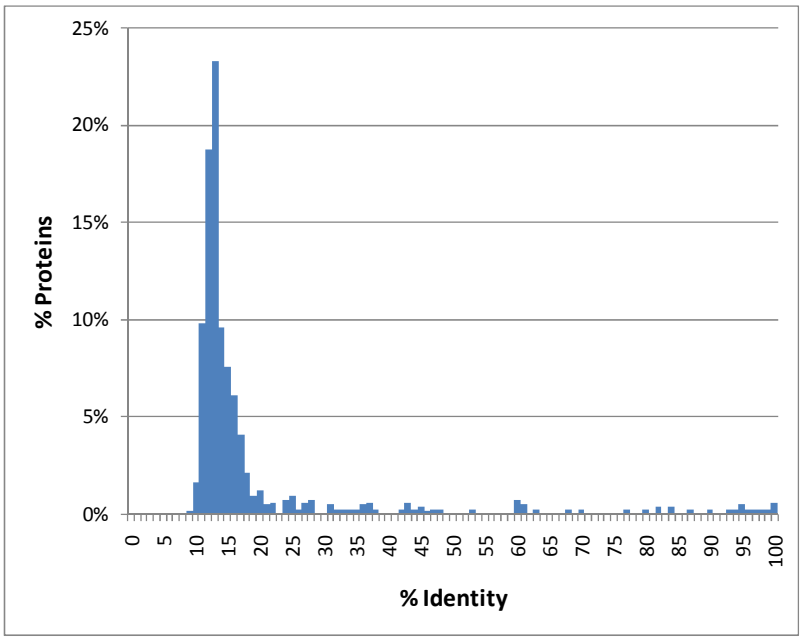


Figure 7.4 Frequency distribution of maximal pairwise sequence identity of all proteins used in the training set versus all other proteins in the training dataset; i.e. every protein sequence was aligned to all other protein sequences using ClustalW and the maximal sequence identity was recorded in the histogram.

We used the svm-light package developed by Joachims and co-workers [101]. The feature vector (FV) used for the Support Vector Machines consisted of a binary representation of the sequence of amino acids and numeric representation of their chemical-physical properties. The sequence information part was represented by a vector consisting of 240 elements (12 x 20 with 6 residues on either side of the central serine and 20 amino acid types). Each component of the vector was set to 1 in case of an occurrence of the particular amino acid type in the respective position. For the amino acid property part of the FV, we utilized data from the collection of 530 commonly used indices provided by the AAindex database [175] including hydrophobicity, solvent accessibility preferences, secondary and tertiary structure preferences, polarity, volume, solvent accessibility, as well as structural disorder indices. The resulting vector consisted of 530 x 12 elements representing every index and position around the central serine. Furthermore, the average values of the 530 different properties over the considered sequence window were appended to the FV. Optimal parameters for the kernel decision function, as judged by the highest obtained Matthew's Correlation Coefficient (CC) value (Eq. 5.2 p. 63), have been determined by using the built-in Leave-One-Out (LOO) test for

all possible parameter combinations for degree of the polynomial function with degrees ranging from 2 to 4 and error weighting values (cost factor) ranging from 1 to 2.5 in 0.25 increments (21 possible parameter combinations).

Eq. 7.1 ;

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}; \mathrm{Re}\,call = \frac{T_P}{T_P + F_N}; \mathrm{Pr}\,ecision = \frac{T_P}{T_P + F_P},$$

where T_P are true positive, T_N-true negative, F_P-false positive, and F_N-false negative predictions.

Prediction performance and estimated generalization error was evaluated by implementing a 10-fold cross-validation test. The training set was randomly divided into ten sets of equal size by also assuring that in all sets, the number of negative examples was twice as large as the positive set. The SVM-based classifier was trained on all but one set and tested on the set which was taken out. This process was repeated for all ten sets to evaluate precision, recall, and CC as well as receiver operating curve (ROC) and the area under the ROC [171; 176] for comparison with NetPhos [170]. The classifier NetPhos 2.0 was applied to our dataset without training as it was not technically possible to perform a 10-fold cross-validation for NetPhos 2.0. Statistical tests on the performance difference (difference of the area under the ROC and estimation of the standard error of the area) were performed according to the algorithm proposed by Hanley et al. [171]. In addition recall, precision and accuracy were evaluated according to equation 7.1.

Recall, precision, and Matthew's Correlation Coefficient (CC) are commonly used performance indicators of computational classifiers. While recall and precision describe the quantity and quality respectively of the prediction of the positive targets, Matthew's Correlation Coefficient describes the overall quality of the prediction, balancing the true and false results. CC of -1 means exactly wrong predictions. A complete recall at the expense of precision would be reflected by a CC of zero. A perfect prediction would yield a CC of 1.

Chapter 8
Assessment of false positive rates of phospho-proteomic data

8.1 Background

The utilization of high-throughput technologies necessitates an assessment of false-positive identifications and general estimation of error. The proteomics community established curation rules to generate truly large-scale data on phosphoproteins and phosphorylation sites utilizing existing techniques. Although the extent of false positive rate for phosphopeptides and phosphorylation sites is not known and their experimental confirmation is not easily possible at present, a confidence measure may be generated by statistical analysis of the overlap of phosphosites between two independent datasets. Furthermore, predictor-based strategies supporting the identification of false positives are conceivable. Since phosphorylation site predictors assign confidence values to predicted sites, a newly identified phosphorylation site might be further supported by a prediction result.

In this study, we evaluated the overlap of experimental reports of phosphorylation sites in general as well as the overlap of low-throughput and high-throughput observations. Since strategies, based on prediction for an identification of false positives are not yet possible, as a set of false positives is not available, we confined this issue to a feasibility study, evaluating the correlation between the confidence values derived from prediction and corresponding experimental evidence. Furthermore, we compared the prediction performance based on confidence values for phosphorylation sites confirmed by multiple experiments.

The results will support the development strategies concerning the identification of false positives in experimental results.

8.2 Results

8.2.1 Concordance of experimental reports

For estimation of experimental confidence of identified phosphorylation sites in the *PhosPho.Elm* database as well as *PhosPhAt,* the number of identifying experiments was assigned to each phosphosite and the overlap between experiments was evaluated. The overlap is surprisingly high between the phosphoproteins and phosphosite. As shown in **Figure 8.1**, phosphoproteins and phosphosites were found concordantly in up to four independent high-throughput (HTP) experiments. From a total number of 2,602 human

phosphoproteins in the Phospho.ELM database, 27.6% could be confirmed by at least two experiments. A lower overlap was found for phosphoproteins from rodents (mouse and rat). Here, 17.1% out of 1,532 HTP-proteins were identified in at least two experiments. On the phosphosite level, 15.7% of 7,891 human and 11.5% of rodent phosphosites were found concordantly by at least 1122 two experiments. The percentage of confirmed *Arabidopsis* phosphoproteins or phosphosites contained in the *PhosPhAt* database was 21.7% or 10.6%, respectively, when filtering for phosphoproteins and phosphosites confirmed by at least two HTP experiments.

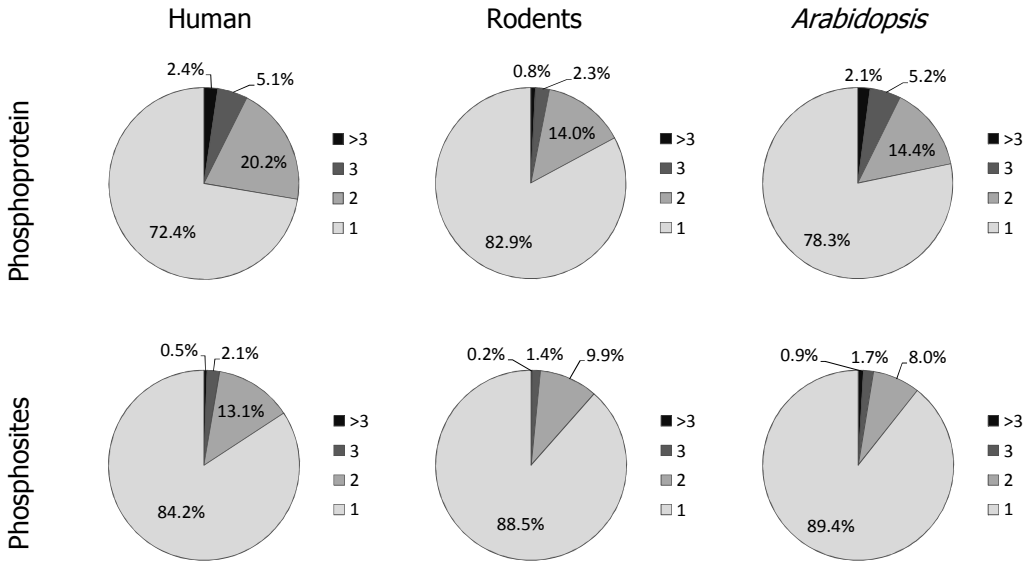


Figure 8.1 Identified phosphoproteins and phosphosites in Human, rodents, and *Arabidopsis*. The Phospho.ELM (for Human and Rodents) and *PhosPhAt* (for *Arabidopsis*) were used for bioinformatics analyses. Numbers of independent experiments are based on number of different Medline entries.

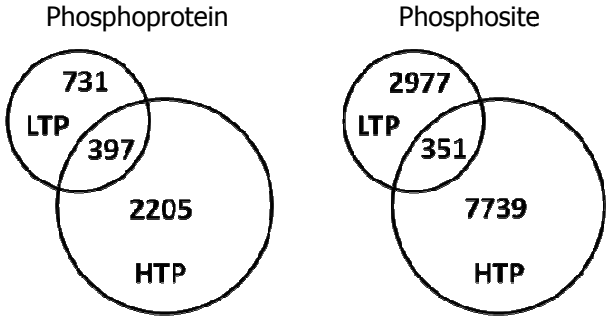


Figure 8.2 Venn diagram of phosphoproteins and phosphosites detected in the human samples using high-throughput (HTP) and low-throughput (LTP) approaches as per the Phospho.ELM database.

As a raw estimate of the number of false positives (FP) and false negatives (FN), we evaluated the overlap between the results received from low-throughput (LTP) and high-throughput experiments, using the Phospho.ELM database, that, unlike *PhosPhAt,*

comprises results from both LTP and HTP experiments (**Figure 8.2**). We found that 35.2% of human phosphoproteins identified in the LTP experiments (397 out of 1,128 proteins) have also been identified in HTP experiments, while 15.3% of phosphoproteins of the HTP experiments (397 out of 2,602) were also verified by the LTP experiments. On the phosphosite level, 10.5% of the 3,328 phosphosites from the LTP experiments were found in the HTP experiments, while 4.3% of 8,090 phosphosites from the HTP experiments were only found in LTP experiments (**Figure 8.2**).

8.2.2 Correlation of the confidence values and the number of publication reports as well as the computed AUC

Besides the general task of computational predictors to identify phosphorylation sites, predictor based confirmation strategies for estimating false positives in experiments are conceivable. Such strategies require high prediction accuracies, which might presently not be sufficient. However, we observed evidence in support of such a strategy. For each evaluated predictor in this study, we observed a significant correlation between the decision value (confidence value), an estimation of the reliability of prediction, and the number of publications of the site. For serine sites, a Pearson Correlation Coefficient, r, of 0.44 was observed, for threonine sites of 0.23 and for tyrosine sites of 0.27, when decision values from the spatial-information-based predictor (see Chapter 6) are considered. All correlations were observed to be statistically significant (p-Value below 3e-21) (**Table 8.1**). Furthermore, the performance of a predictor, as judged by the area under the ROC curve, was observed to increase, when only results assigned to multiple experimental setups were considered (**Figure 8.3**). The AUC for the spatial-information-based serine specific predictor increased from 0.79 ± 0.02 to 0.85 ± 0.02, for threonine sites from 0.74 ± 0.03 to 0.85 ± 0.02 and for tyrosine sites 0.71 ± 0.02 to 0.75 ± 0.02, revealing a positive relation of the number of concordant identifications of phosphorylation sites and accuracies of their prediction.

Table 8.1 Pearson correlation coefficient of the pdecision values and number of experimental reports of serine, threonine and tyrosine sites

	HTP	LTP	XTP	HTP	LTP	XTP	HTP	LTP	XTP
	Serine-sites			Threonine-sites			Tyrosine-sites		
sequence -only	0.24 p= 2.23e-15	0.28 p= 2.79e-21	0.36 p= 4.97e-35	0.14 p= 5.89e-03	0.21 p= 1.65e-05	0.24 p= 1.02e-06	0.16 p= 5.19e-06	0.22 p= 4.75e-10	0.26 p= 7.46e-13
spatial-inf.	0.22 p= 5.63e-13	0.37 p= 8.37e-37	0.44 p= 5.62e-52	0.15 p= 2.09e-03	0.26 p= 2.00e-07	0.29 p= 5.29e-09	0.17 p= 1.26e-06	0.23 p= 1.10e-10	0.27 p= 8.01e-14
NetPhos	0.13 p= 1.68e-05	0.23 p= 2.59e-14	0.27 p= 2.64e-19	0.08 p= 0.113	0.30 p= 6.10e-10	0.31 p= 2.72e-10	0.12 p= 1.21e-03	0.23 p= 2.95e-10	0.25 p= 7.33e-12

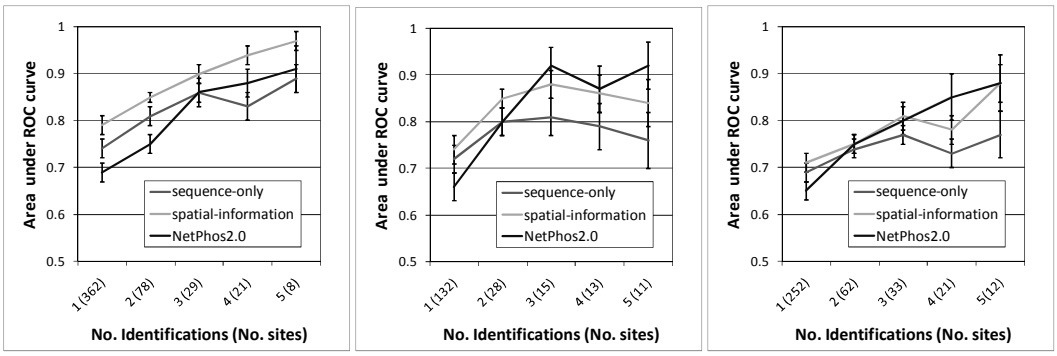


Figure 8.3 Area under the ROC curve (AUC) according to the number of experimental evidences for phosphorylation sites for serine (left chart), threonine (middle chart) and tyrosine (right chart). The AUCs were computed including results reported by at least 1, 2, 3 and 4 experimental setups. For each computation, non-phos sites were randomly chosen to form a negative set of a size twice as large as the positive set to reflect the training setup. For AUCs of sequence-only based and spatial information supported prediction results from 10-fold-cross-validations were used, whereas for NetPhos2.0, results received from the server were used.

8.3 Discussion

We evaluated the overlap between results from independent studies as well as results obtained from low-throughput (LTP) and high-throughput (HTP) experiments to obtain an impression of reproducibility of data as well as a raw estimation of the false positive rates in high throughput data. In both cases, we observed that this overlap is surprisingly high. However, although the confirmation of phosphoproteins and phosphosites from independent studies is a good approach for true positive identifications, the absence of overlap is not an immediately an indication of falsely-detected phosphorylation events. It is obvious that non-overlapping phosphoproteins and phosphosites could appear in different environmental as well as experimental conditions. Considering the possible differences, the high overlap rates indicate highly reliable data. Although the evaluation of overlap provides a confidence of experimental results, the estimation of false positive rates in numbers is not (yet) feasible.

We observed a significant correlation between the confidence values from predictions and the number of experimental identifications of phosphosites. Furthermore, we observed that the prediction performance for more confident phosphorylation sites, as found concordantly in multiple independent experiments, is much higher in comparison to the predictors performance obtained for all phosphosites. Thus, we believe that further confidence on experimental data might indeed be provided by an integration of prediction results.

8.4 Methods

Concordance of experimental reports

For each phosphorylation site as well as phosphoprotein, the number of independent experiments was assigned. Two experiments were assumed to be independent, if the phosphosite or phosphoprotein was reported by independent Medline entries. Subsequently, the overlap of experiments was evaluated for phosphosites and phosphoproteins from human and rodents (PhosPho.Elm) and *Arabidobsis* (*PhosPhAt*). Additionally, the overlap between the results obtained from low-throughput (LTP) and high-throughput (HTP) experiments was evaluated for human and rodent phosphosites and phosphoproteins using the Phospho.ELM database, as unlike *PhosPhAt,* it comprises results from both LTP and HTP experiments.

Correlation of confidence values and number of publication reports as well as the computed AUC

Decision values from the 10-fold-cross-validation of the sequence-only as well as spatial-information-based prediction algorithm introduced in Chapter 3 were used to evaluate the correlation between decision values (confidence values) and the number of publication reports. For the computation of the correlation for NetPhos2.0 results, the corresponding probability values were applied, as a 10-fold-cross-validation for NetPhos2.0 was not feasible.

Computation of the AUC, which corresponds to different evidence levels, was performed by including phosphosites assigned with a minimal number of experimental reports and a twice as large negative set, consisting of sites not found in any experiment. The 1:2 ratio of the utilized positive and negative set was chosen to reflect training conditions of the sequence-only and spatial-information-based predictor. The negative set size was adjusted to the positive size by random removal of the overhanging non-phos sites. The adjustment procedure was performed 1,000 times for each level of experimental evidence. Subsequently, the mean AUC as well as the standard deviation was computed.

General Discussion

Perhaps the most important feature of biological systems is the ability to adjust to environmental conditions that emerged from evolutionary pressure. The fundament of biological systems are interactions, which, represented as biological interaction networks (graphs), may be investigated by means of graph-theory-based algorithms. Advances in technology in recent years contributed to this approach. High-throughput experiments yielded system-wide records of biological dependencies, and provided an excellent source of information to investigate entire systems and thus created a new perspective of a comprehensive view on cellular processes.

This work was concerned with a small subset of biological interaction networks. We investigated metabolic interaction networks, phosphorylation networks and protein interaction networks. One strategy to investigate biological networks is the comparative study of organization principles of the underlying networks. However, we have to pose the question, whether a comparative study of different network types provides answers for biological questions. Certainly yes, as can be seen in Chapter 3, for instance: we successfully identified broad correspondences of topologies between metabolic interaction networks and protein interaction networks, proving that metabolic interaction networks are indeed reflected in protein interaction networks. These correspondences, however, become only apparent upon integration of different information types such as GO:Annotations. Hence, integration allows the focused analysis of a particular aspect of raw protein interaction networks. Alongside the analysis of protein interaction networks designed specifically to capture aspects of metabolism (fPIN), we also investigated the signaling and regulatory aspects of protein interactions in Chapter 4. In particular, two kinase interaction networks were investigated. While the kPIN was constituted by kinase interactions directly derived from protein interaction networks, and as such did not explicitly cover kinase-target interactions, the networKIN was constructed via an integrative determination of kinase-target relations by kinase-specific predictions[33]. The latter construction scheme allowed the interpretation of the networKIN also as a directed graph, thus to investigate its dynamic properties.

We observed broad correlations of topological tendencies between the fPIN and metabolic interaction network and also between the kPIN and networKIN, although their construction methodologies were different. The topological properties of the rPIN were observed to better reflect the phosphorylation networks than the metabolic networks. In particular, metabolic networks such as metabolic interaction networks and fPIN were observed to be assortative, i.e. they showed a positive correlation between the degrees of neighboring nodes. In contrast, the rPIN and the phosphorylation networks were observed to be dissortative. Moreover, the metabolic networks exhibit greater characteristic lengths and higher modularity than phosphorylation networks and rPIN.

These properties reflect the different modes of operation of both network types. While short information paths are established in signaling-networks, to ensure fast signal transduction, longer assembly lines are necessary for the synthesis of complex biomolecules in metabolite interaction networks. Since cooperation is necessary for an efficient production system, enzymes are organized in tightly interlinked modules, a property that is reflected in higher cluster coefficients of enzymes interaction networks. This modularity is also reflected in the fPIN. Hence, protein interactions support the cooperation of enzymes by ensuring their spatial proximity. The difference of the cluster coefficient of the directed and undirected phosphorylation networks ($< cc >_{directed}\ networKIN = 0.19; < cc >_{directed}\ networKIN = 0.002$) suggests that the inter-phosphorylation between kinases, reflected in kinase cascades, is mainly responsible for signal transduction, while the regulatory tasks of the network is rather performed by cross-acting kinases on the same target than cross-regulation of kinases, in a feedback manner (Chapter 4). Also, the differences in the assortativity of the metabolic networks and the phosphorylation networks reflect the different biological activities of these networks. While the synergized regulation of a high number of proteins, by the same kinase, cause the dissortativity of phosphorylation networks, the assortativity of enzyme-, metabolite- and fPIN corresponds to an effective infrastructure, where central tightly connected nodes supply less connected peripheral routes.

Further insights into interaction networks

The integrative analysis of the fPIN and metabolic interaction networks revealed novel organization principles of metabolic networks. Short connections between enzymes are established to support performance of the central metabolism as well as the production of secondary metabolites in metabolic networks. The connectivity, betweenness-centrality as well as the cluster coefficient were shown to correlate with the flux rates of the associated enzymes. In addition, the flux rates of physically interacting enzymes were shown to correlate as well. However, the organization of connections between enzymes does not follow the classical rules of metabolic channeling. Enzymes involved in glycolysis, for instance, rather constitute a metabolic cluster established by few highly connected enzymes than an enzymatic chain constituted by interactions between successive enzymes. We identified these linking proteins introducing a new measure, the robustness-centrality.

Also the investigation of the phosphorylation networks provided interesting insights. In particular, we found that 38% of 68 included kinases in networKIN are capable to influence almost the entire network (Chapter 4), suggesting an overlapping influence of the involved kinases and revealing once again the high complexity of the regulatory network.

How complete are the descriptions of interaction networks?

Although different information types were already integrated into the analysis, it may be suggested that the integration of further available information would lead to a more accurate description of cellular networks. For instance, the integration of metabolomic and proteomic data to capture the abundance of proteins (enzymes) and metabolites may reveal novel insights into the metabolic interaction networks as well as protein interaction networks. Also, the collections of integrated interactions used in this study are far from complete. Hart et al. suggested that considering possible high false-positive rates of protein interaction data yeast protein interaction networks are roughly only 50% complete[177]. An even greater gap between the observed and the actual number of interactions may be expected for phosphorylation networks (Chapter 7). Moreover, the investigated phosphorylation networks contained only a small number of the total counts of kinases. The networKIN, for instance, contained only 68 kinases, while the complete human kinome was estimated to comprise 500 different kinases[98]. Furthermore, all evaluated interaction networks in this study may suffer from possible technological as well as other biases introduced by targeted scientific interest[86], such that these networks may not reflect the real interaction networks properly. We noted that the possible biases do not influence our conclusions (Chapter 3). However, it should be expected that for a more in-depth analysis these biases might be more critical[84].

False-positive-rates (FPR) of the interaction networks

Along with the increased efficiency of novel technologies, the number of false positive observations increases as well. The identification of false positives or at least the estimation of the false-positive-rates is an imperative task for an accurate research. There are many approaches for the assessment of false-positive-rates, including the computation of overlaps between independent experiments (confirmation by independent experimental setups), integration of annotations (proteins with similar functions are more likely to interact) as well as comparison of network properties (true edges and false edges are suggested to have different network properties). Saito et al., for instance, proposed a method to assess confidence in protein interaction networks by introducing a so called "interaction generality" measure[178; 179]. Their idea was based on the assumption that true edges and false edges have different network properties. Thus, for instance, interactions of so called "sticky" proteins (proteins with many connections), may reflect technical artifacts or not be biologically meaningful. The rPIN comprised a number of proteins with 100 and more connections. These interactions were filtered out in the fPIN, significantly affecting the topology of the protein interaction network. However, these highly connected proteins were included in the phosphorylation network. Kinases are expected to contain many interactions. Hence, raw protein interaction networks may be suggested to contain many dynamic interactions such as kinase-target recognition

events, which are not necessarily false. Although we could not confirm that these interactions were indeed kinase-target recognition events, the broad correspondence of the topological properties of the kPIN and the networKIN may support such suggestions. The networKIN explicitly contained kinase-target relations. Consequently, it should be suggested that proteins of different biological functions may exhibit different network properties, and the consideration of even such "sticky" interactions yields biologically meaningful conclusions. Another topological measure to gain a confidence measure for protein interactions was proposed by Goldberg and Roth. The basic idea of their approach was the assumption that true interactions might be clustered at higher levels clustered than false. However, since proteins with general functions may be involved in multiple complexes, this assumption is possibly wrong[42]. We believe that the best method to study errors in interaction networks is the integration of experimental details as well as different data sources as proposed by Gilchrist and Wagner[180], and Bader et al.[181]. However, the comparison of different but related network types by integrative analysis as shown for metabolic interaction network and fPIN as well as kPIN and networKIN, provided evidence for the reliability of the underlying protein interaction networks. Thus, we believe that the false-positive rates of protein interaction may be overestimated, and the estimated numbers may result from disregard of the function of the protein interaction and eventually of misinterpretation of overwhelming numbers of links. However, we are also convinced that high-throughput experiments have non-trivial error rates. A balance between the quality and quantity of data is always a concern. Despite their large size, even the largest current dataset have far more false negatives, i.e. observations that are not yet recorded, than true positives[84]. Another problem for the estimation of the quality of data is the lack of a "gold standard". Sometimes there is little other than high-throughput data, in other cases, when a gold standard is available, the high-throughput data is often so huge that the "gold standard" cannot validate most of the true data. The MIPS dataset, for instance, which is frequently used as a gold standard to assess the reliability of protein interaction networks, covers solely tightly connected protein complexes, i.e. it does not consider other types of protein interactions, and is magnitudes smaller than datasets derived from high-throughput experiments[182]. Therefore, Leach et al. proposed a different technique to assess confidence, which does not need a gold standard. The completeness as well as the error rates of two independent experiments are estimated by comparison of overlaps between these two datasets and a further independent reference set, which not necessarily has to be a gold standard[183]. However, the confidence measure obtained from even this approach is limited, since the independence of the datasets as well as the independence of the reference set must be guaranteed, which is most often not possible. At the same time, the evaluated datasets may be biased towards the identification of a certain type of

interactions, yielding low overlaps, and consequently assumed to have higher false-positive-rates[84; 86].

Even the proposed networKIN may suffer from insufficient accuracies of the underlying construction procedure, which is based on kinase-specific predictions. Kinase-specific predictors may produce false links when the specificity and sensibility of the underlying predictors is not sufficient or the motifs of several kinases have similar motifs. It remains to be seen if the same conclusions as gained in Chapter 4 may be drawn when the phosphorylation networks become more complete and the prediction methods more accurate. Another issue is the assessment of false-positive observations of phosphorylation in experimental data and the question, whether the probabilistic confidence assessments of protein interactions are transferable to phosphoproteomics. Although the proteomics community established curation rules to generate truly large-scale data on phosphoproteins and phosphorylation sites, the false-positive-rates for phosphopeptides and phosphorylation sites, are not known and their experimental confirmation presently not easily doable. A confidence measure may be obtained by statistical analysis of the overlap of phosphosites between independent datasets, similarly to approaches performed on protein interaction networks. We observed that the overlap between independent experiments is surprisingly high, and much higher than comparable overlaps of high-throughput protein interaction results. Considering that non-overlapping phosphoproteins and phosphosites could appear in different environmental as well as experimental conditions, the results suggest that the data are highly reliable. However, this consideration also reveals limitations of this application of this approach on the dynamic phosphorylation networks. We therefore propose a different method to address this issue, which is based on prediction of already identified phosphorylation sites. In principle, the reliability of experimental identifications should correlate to the assigned confidence values from predictions. However, since a set of false positives is not yet available, the estimation of the performance of the method is yet not possible. Indeed, we observed a significant correlation of confidence values derived from prediction and the number of concordant experiments. Furthermore, the prediction performance, as judged by the area under the ROC curve, remarkably increases when only phosphorylation sites, found concordantly in multiple independent experiments, are considered. However, since the prediction results are based on similarities to phosphorylation sites included in the training, a bias on a particular type of phosphorylation sites, which is more frequently found in experiments, is possible. Thus, further strategies have to be developed to estimate similarities among phosphorylation motifs and to eliminate redundancies in the training set. However, since the predictor, is designed to learn common properties of phosphorylation sites, and to assign a prediction value according to the similarity of the predicted site to the highlighted properties, the elimination of redundancies may contradict the accuracy of the predictor, albeit also

reduce the risk of overfitting. Thus, the next essential step towards the development of a predictor-based confidence algorithm appears to be a definition of a reliable false positive set for confirmation and fine-tuning of the method.

Prediction of phosphorylation sites

Besides the general issue to reconstruct reliable phosphorylation networks, there is still a demand for functional characterization of post-translational modifications (PTM) of proteins, in particular also phosphorylation events. This demand is additionally amplified by rapid technological progress and the accompanied identification of thousands of novel proteins in recent years. Given the high number of candidate phosphorylation sites, efforts to experimentally identify and verify them all remain challenging. Most present computational methods to predict potential phosphorylation sites are primarily based on extracting predictive features from the local, one-dimensional sequence information surrounding phosphorylation sites. However, kinase-target recognition events are a three-dimensional event. Consequently, the integration of spatial context of phosphosites is justified. In Chapter 6, we characterized phosphorylation sites by specific 3D structural motifs or epitopes constituted by amino acid residues that are not necessarily close in sequence but in spatial proximity, thereby providing additional information for the prediction of phosphorylation sites for proteins with known structure or with available structural models. We observed that 3D-motifs are indeed detectable, especially when studying kinase families individually and obtained improved prediction results by including 3D information in the prediction. This conclusion was drawn by comparison of publicly available predictors and also our own, highly accurate predictor based on sequence-information only. Since only a small part (approximately 6%) of identified phosphorylation sites were used to train the predictor, as only these sites were structurally characterized, it may be suggested that adding novel structures into the study will further increase the performance of the prediction in the near future.

Also our sequence-only based predictor implicitly captures 3D structural preferences like hydrophobicity, solvent accessibility as well as secondary and tertiary structure preferences, polarity, volume and solvent accessibility, structural disorder. The sequence-only-based prediction method was developed to fairly access the contribution of spatial information to the performance of predictions, but also to contribute to the plant science community as an integral part of *PhosPhAt*: since it has been trained on plant specific phosphosites it outperforms other common available predictors, usually trained on non-plant species, when used for the prediction plant phosphosites.

PhosPhAt

The *PhosPhAt* database has been established to provide a resource that consolidates our current knowledge of mass spectrometry-based identified

phosphorylation sites in the model plant *Arabidopsis thaliana*, albeit work is ongoing to include information for other plant species as well. Combined with a phosphosite prediction tool that was specifically trained on *Arabidopsis* serine phosphorylation site motifs, it not only serves as a searchable knowledge base for experimentally identified phosphorylation sites, but in addition also provides a powerful resource for the characterization and annotation of yet unidentified phosphoserine sites. The PhosPhAt predictor was shown to accurately identify plant phosphorylation sites and to outperform commonly available predictors, which usually suffer from low accuracies applied to plant species, since the training of these predictors has been largely limited to non-plant species. Based on the *Arabidopsis*-specific predictor, a genome-scale prediction of phosphoserine sites was performed, revealing a raw estimation of the distribution of phosphoproteins among functional annotations. The predicted sites with highest decision values in combination with the experimental phosphorylation sites will provide a powerful basis for further in-depth analysis of phosphorylation motifs in orthologous and paralogous proteins also between different organisms [174]. This issue will gain further support by replacement of pre-computed predictions with an prediction algorithm for user defined input, for serine tyrosine and threonine sites (in progress), allowing the prediction of phosphorylation sites in submitted unknown protein sequences. It remains to be seen, how generalizable the *PhosPhAt Arabidopsis*-centric prediction method will turn out to be when applied to phosphorylation sites detected in other plant species and whether a generic plant-phosphorylation prediction method will replace the *PhosPhAt* predictor or whether species-specific prediction methods may have to be developed.

Outlook

Besides the on-going work on integration of the prediction algorithms into *PhosPhAt*, the development of confirmation strategies for the assessment of false-positives in phosphoproteomics data and refinement of the phosphorylation network, by application of improved prediction methods to identify kinase-target relations, there are also ideas emerging to expand the phosphorylation networks by including phosphatase-target relations as well as signaling events performed by secondary messengers such as cAMP, cGMP, inositol trisphosphat or Ca^{2+} (or calmodulin respectively). Moreover, the characterization of distinct posttranslational modifications may be of interest.

Furthermore, since the activities of metabolic pathways are subject to precise regulation in order to adjust the synthesis and degradation of metabolites to physiological requirements, the integration of regulatory aspect into the metabolic interaction networks, may reveal emergent properties of this networks. Present investigations are mainly focused on modulation of enzymatic activities by ligands, including modulations affected by the availability of substrates and coenzymes (transport between compartments and regeneration of coenzymes) as well as competition of different

metabolites and feedback inhibitions by end products. However, metabolic pathways may also be regulated by distinct mechanisms. Thus, the activity of enzymatic reactions is strongly influenced by transcriptional control, i.e. induction or repression on gene expression level and the affected abundance of enzymes, as well as interconversion of inactive enzymes into the catalytically active form by proteolysis reactions or reversible posttranslational modifications.

Conclusions

Our results reveal topological equivalences between the protein interaction network and the metabolic pathway network, upon filtering of non-metabolic aspect of raw protein interaction networks. Evolved protein interactions may contribute significantly towards rendering metabolic processes more efficient by permitting increased metabolic fluxes. Thus, our results shed further light on the unifying principles shaping the evolution of both the functional (metabolic) as well as the physical interaction network.

The different functional aspects of the raw protein interaction networks reveal the necessity of a careful curation of the protein interactions, according to the particular interest. Protein interaction networks comprising metabolic aspects exhibit different topological properties when compared to protein interaction networks focused on phosphorylation events. Consequently, the reported properties of biological networks must be refined and also updated according to the actual scientific results.

The study of phosphorylation networks is by far not complete. The networks still suffer of insufficient accuracies of underlying predictions and specificity, thus limiting their reliability. We contributed to this issue by studying the inclusion of further available information to the performance of predictions. Two additional sources of information were considered: spatial context of the phosphorylated amino acids as well as the focus on species specific information. In both cases, we effectively increased the prediction performance. Another contribution for the investigation of phosphorylation networks is established by a database for plant specific phosphorylation sites, *PhosPhAt*. The database provides a valuable resource for plant science community and as such fills the gap of poor represented sources for plant specific phosphoproteomics.

Glossary and Abbreviations

$< c >$	average cluster coefficient; a graph-theoretical property, which describes the neighbors' interconnectivity of nodes, thus the modularity of a graph
$< NC(k) >$	Neighbors' Connectivity; a graph-theoretical function, which describes the correlation of degrees (k) of nodes
Aaindex	collection of commonly used indices of amino acid properties
accuracy	in the context of prediction: proportion of correct predictions
AGI (gene code)	*Arabidopsis* Genome Initiative; gene-ID for genes annotated in the *Arabidopsis* Information Resource (TAIR)
AUC	area under the ROC curve; estimator of performance of prediction; yields a number between 0 and 1, with 1 indicating perfect prediction, 0.5 random prediction, and <0.5 a worse than random prediction accuracies.
BN	betweenness (centrality); a graph-theoretical property, which describes the number of transpassing shortest paths
CC	Matthews' Correlation Coefficient; estimator of performance of prediction; yields a number between -1 and 1, with 1 indicating perfect prediction, 0 indicating a random prediction and <0 a worse than random prediction levels.
CIN	Compound (Metabolite) Interaction Network; type of metabolic interaction network, where metabolites involved in the same reaction are connected
CL	characteristic length; a graph-theoretical property, which describes the average shortest path between nodes in a graph
confidence value	in the context of prediction: decision value (see decision value)
connectivity	in graph-theoretical context: the degree of nodes
CV	cross-validation; method to estimate the generalization of a predictor
decision value	prediction response value from a SVM approach, where the sign denotes the respective class while the absolute value denotes the reliability of the prediction, higher values mean higher reliability
degree	in graph-theoretical context: number of links part of nodes
degree-centrality	in graph-theoretical context: a centrality measure based on degrees of nodes
EIN	Enzyme Interaction Network, an interaction network, where enzymes of successive reactions are connected; derived from reaction lists
ePIN	enzyme Protein Interaction Network; protein interaction network consisting only of enzyme interactions

FN	false negative(s)
FP	false positive(s)
fPIN	rPIN after removal (filtering) of proteins related to DNA processing, protein-degradation, kinase-phosphatase and other-non-metabolic rather unspecific functions; is assumed to capture aspects of metabolism
FPR	false-positive-rate
FV	Feature Vector; in the context of SVM: a vector comprising the predictive variables of a sample
GO	Gene Ontology
HTP	high-throughput (experiments)
KEGG	Kyoto Encyclopedia of Genes and Genomes
kPIN	kinases Protein Interaction Network; phosphorylation network, comprising protein interactions of kinases
LOO	Leave-One-Out test; method for cross validation of predictors
LTP	low-throughput (experiments)
mapEIN	map Enzyme Interaction Networks; enzyme interaction network, derived directly from the xml-description files, thus reflecting KEGG-maps
MIN	Metabolic Interaction Network
networKIN	phosphorylation network; based on prediction of kinase-target relations[33]
PCA	principal component analysis
PDB	Protein Data Bank
phosphorylation Network	interaction networks, consisting only of kinase interactions
PIN	protein interaction network, network of physical interactions between proteins
PKA, PKB, PKG, PKC, RSK, CDK, CDC, CKII, ATM	kinase families
PLS	Partial Least Squares
precision	in the context of prediction: performance measure (contrasted with recall), the fraction of true positives predicted as positive,

	same as true-positive-rate
P-Site	phosphorylation site
PTK, CMGC, AGC, CAMK	kinase groups
RC	robustness centrality; centrality measure, based on the change of the characteristic length upon removal of a node
RCP-Plot	Radial Cumulative Propensity Plot
r_d	assortativity; a graph-theoretical property, which describes the correlation of degrees of neighboring nodes
recall	in the context of prediction, performance measure (contrasted with precision), the fraction of correctly predicted positives, same as true-positive-rate
ROC	Receiver Operating Characteristic; plot of the true positive rate versus the false-positive-rate in respect to increasing decision thresholds.
rPIN	raw Protein Interaction Network; protein interaction network directly derived from databases for physical protein interactions.
SGD	*Saccharomyces* Genome Database
SVM	Support Vector Machine
TCA cycle	tricarboxylic (citric) acid *cycle*
TN	true negative(s)
TP	true positive(s)
TPR	true-positive-rate; proportion of correctly identified positives
γ	scaling exponent in the power law $P(k) \sim k^{-\gamma}$

Bibliography

1. Koolman, J. & Roehm, K. H. (2005). *Color Atlas of Biochemistry* 2nd edit, Thieme, Stuttgart · New York.
2. Winkel, B. S. (2004). Metabolic channeling in plants. *Annu Rev Plant Biol* **55**, 85-107.
3. Huang, X., Holden, H. M. & Raushel, F. M. (2001). Channeling of substrates and intermediates in enzyme-catalyzed reactions. *Annu Rev Biochem* **70**, 149-80.
4. Easterby, J. S. (1981). A generalized theory of the transition time for sequential enzyme reactions. *Biochem J* **199**, 155-61.
5. Westerhoff, H. V. & Welch, G. R. (1992). Enzyme organization and the direction of metabolic flow: physicochemical considerations. *Curr Top Cell Regul* **33**, 361-90.
6. Rudolph, J. & Stubbe, J. (1995). Investigation of the mechanism of phosphoribosylamine transfer from glutamine phosphoribosylpyrophosphate amidotransferase to glycinamide ribonucleotide synthetase. *Biochemistry* **34**, 2241-50.
7. Ushiroyama, T., Fukushima, T., Styre, J. D. & Spivey, H. O. (1992). Substrate channeling of NADH in mitochondrial redox processes. *Curr Top Cell Regul* **33**, 291-307.
8. Ovadi, J., Huang, Y. & Spivey, H. O. (1994). Binding of malate dehydrogenase and NADH channelling to complex I. *J Mol Recognit* **7**, 265-72.
9. Dewar, M. J. & Storch, D. M. (1985). Alternative view of enzyme reactions. *Proc Natl Acad Sci U S A* **82**, 2225-9.
10. Srere, P. A. (1987). Complexes of sequential metabolic enzymes. *Annu Rev Biochem* **56**, 89-124.
11. Wakil, S. J., Stoops, J. K. & Joshi, V. C. (1983). Fatty acid synthesis and its regulation. *Annu Rev Biochem* **52**, 537-79.
12. Batke, J. (1989). Channeling of glycolytic intermediates by temporary, stationary bi-enzyme complexes is probable in vivo. *Trends Biochem Sci* **14**, 481-2.
13. Keleti, T. & Ovadi, J. (1988). Control of metabolism by dynamic macromolecular interactions. *Curr Top Cell Regul* **29**, 1-33.
14. Ovadi, J. & Keleti, T. (1978). Kinetic evidence for interaction between aldolase and D-glyceraldehyde-3-phosphate dehydrogenase. *Eur J Biochem* **85**, 157-61.
15. Vertessy, B. & Ovadi, J. (1987). A simple approach to detect active-site-directed enzyme-enzyme interactions. The aldolase/glycerol-phosphate-dehydrogenase enzyme system. *Eur J Biochem* **164**, 655-9.
16. Srere, P. A. (2000). Macromolecular interactions: tracing the roots. *Trends Biochem Sci* **25**, 150-3.
17. Cornish-Bowden, A. & Cardenas, M. L. (1993). Channelling can affect concentrations of metabolic intermediates at constant net flux: artefact or reality? *Eur J Biochem* **213**, 87-92.
18. Pettersson, G. (1991). No convincing evidence is available for metabolite channelling between enzymes forming dynamic complexes. *J Theor Biol* **152**, 65-9.
19. Wu, X. M., Gutfreund, H., Lakatos, S. & Chock, P. B. (1991). Substrate channeling in glycolysis: a phantom phenomenon. *Proc Natl Acad Sci U S A* **88**, 497-501.
20. Ro, D. K. & Douglas, C. J. (2004). Reconstitution of the entry point of plant phenylpropanoid metabolism in yeast (Saccharomyces cerevisiae): implications for control of metabolic flux into the phenylpropanoid pathway. *J Biol Chem* **279**, 2600-7.

21. Degenring, D., Rohl, M. & Uhrmacher, A. M. (2004). Discrete event, multi-level simulation of metabolite channeling. *Biosystems* **75**, 29-41.
22. Kholodenko, B. N., Westerhoff, H. V., Schwaber, J. & Cascante, M. (2000). Engineering a living cell to desired metabolite concentrations and fluxes: pathways with multifunctional enzymes. *Metab Eng* **2**, 1-13.
23. Green , D. E. (1958). Studies in organized enzyme systems. *Harvey Lect.* **52**, 177–227
24. Ovadi, J. & Srere, P. A. (2000). Macromolecular compartmentation and channeling. *Int Rev Cytol* **192**, 255-80.
25. Seet, B. T., Dikic, I., Zhou, M. M. & Pawson, T. (2006). Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* **7**, 473-83.
26. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **98**, 10037-41.
27. Pandey, A., Podtelejnikov, A. V., Blagoev, B., Bustelo, X. R., Mann, M. & Lodish, H. F. (2000). Analysis of receptor signaling pathways by mass spectrometry: identification of vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *Proc Natl Acad Sci U S A* **97**, 179-84.
28. Andersson, L. & Porath, J. (1986). Isolation of phosphoproteins by immobilized metal (Fe3+) affinity chromatography. *Anal Biochem* **154**, 250-4.
29. Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villen, J., Li, J., Cohn, M. A., Cantley, L. C. & Gygi, S. P. (2004). Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* **101**, 12130-5.
30. Rush, J., Moritz, A., Lee, K. A., Guo, A., Goss, V. L., Spek, E. J., Zhang, H., Zha, X. M., Polakiewicz, R. D. & Comb, M. J. (2005). Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat Biotechnol* **23**, 94-101.
31. Ibarrola, N., Molina, H., Iwahori, A. & Pandey, A. (2004). A novel proteomic approach for specific identification of tyrosine kinase substrates using [13C]tyrosine. *J Biol Chem* **279**, 15805-13.
32. Johnson, L. N. & Barford, D. (1993). The effects of phosphorylation on the structure and function of proteins. *Annu Rev Biophys Biomol Struct* **22**, 199-232.
33. Linding, R., Jensen, L. J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M. B. & Pawson, T. (2008). NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res* **36**, D695-9.
34. Barabasi, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509-12.
35. Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**, 47.
36. Albert, R. (2005). Scale-free networks in cell biology. *J Cell Sci* **118**, 4947-57.
37. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature* **407**, 651-4.
38. Wagner, A. & Fell, D. A. (2001). The small world inside large metabolic networks. *Proc Biol Sci* **268**, 1803-10.
39. Ge, H., Liu, Z., Church, G. M. & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29**, 482-6.
40. Kemmeren, P., van Berkum, N. L., Vilo, J., Bijma, T., Donders, R., Brazma, A. & Holstege, F. C. (2002). Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* **9**, 1133-43.
41. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* **1**, 349-56.

42. Goldberg, D. S. & Roth, F. P. (2003). Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A* **100**, 4372-6.
43. Kelley, R. & Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* **23**, 561-6.
44. Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A. & Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol* **23**, 951-9.
45. Ramani, A. K., Li, Z., Hart, G. T., Carlson, M. W., Boutz, D. R. & Marcotte, E. M. (2008). A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. *Mol Syst Biol* **4**, 180.
46. Lee, I., Date, S. V., Adai, A. T. & Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science* **306**, 1555-8.
47. Huthmacher, C., Gille, C. & Holzhutter, H. G. (2008). A computational analysis of protein interactions in metabolic networks reveals novel enzyme pairs potentially involved in metabolic channeling. *J Theor Biol* **252**, 456-64.
48. Travers, J. & Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry* **32**, 425-443.
49. Milgram, S. (1967). The Small World Problem. *Psychology Today* **2**, 60-67.
50. Almaas, E. (2007). Biological impacts and context of network theory. *J Exp Biol* **210**, 1548-58.
51. Bhan, A., Galas, D. J. & Dewey, T. G. (2002). A duplication growth model of gene expression networks. *Bioinformatics* **18**, 1486-93.
52. Huthmacher, C., Gille, C. & Holzhutter, H. G. (2007). Computational analysis of protein-protein interactions in metabolic networks of Escherichia coli and yeast. *Genome Inform* **18**, 162-72.
53. Fell, D. A. & Wagner, A. (2000). The small world of metabolism. *Nat Biotechnol* **18**, 1121-2.
54. Kotera, M., Hattori, M., Oh, M., Yamamoto, R., Komeno, T., Yabuzaki, J., Tonomura, K., Goto, S. & Kanehisa, M. (2004). RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions. *Genome Informatics* **15**, P062.
55. Kotera, M., Okuno, Y., Hattori, M., Goto, S. & Kanehisa, M. (2004). Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J Am Chem Soc* **126**, 16487-98.
56. Amaral, L. A., Scala, A., Barthelemy, M. & Stanley, H. E. (2000). Classes of small-world networks. *Proc Natl Acad Sci U S A* **97**, 11149-52.
57. Stumpf, M., Ingram, P., Nouvel, I. & Wiuf, C. (2005). Statistical Model Selection Methods Applied to Biological Networks. In *Transactions on Computational Systems Biology III*, pp. 65-77.
58. Stumpf, M. P., Wiuf, C. & May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci U S A* **102**, 4221-4.
59. Arita, M. (2004). The metabolic world of Escherichia coli is not small. *Proc Natl Acad Sci U S A* **101**, 1543-7.
60. Albert, R., Jeong, H. & Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature* **406**, 378-82.
61. Chung, F., Lu, L., Dewey, T. G. & Galas, D. J. (2003). Duplication models for biological networks. *J Comput Biol* **10**, 677-87.
62. Pastor-Satorras, R., Smith, E. & Sole, R. V. (2003). Evolving protein interaction networks through gene duplication. *J Theor Biol* **222**, 199-210.

63. Vázquez, A., Flammini, A., Maritan, A. & Vespignani, A. (2003). Modeling of Protein Interaction Networks. *Complexus* **1**, 38-44.
64. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. (1999). From molecular to modular cell biology. *Nature* **402**, C47-52.
65. Freeman, L. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry* **40**, 35-41.
66. Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E* **64**, 016132.
67. Joy, M. P., Brock, A., Ingber, D. E. & Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *J Biomed Biotechnol* **2005**, 96-103.
68. Hahn, M. W. & Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* **22**, 803-6.
69. Goh, K. I., Oh, E., Kahng, B. & Kim, D. (2003). Betweenness centrality correlation in social networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **67**.
70. Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411**, 41-2.
71. He, X. & Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS Genet* **2**, e88.
72. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3**, e59.
73. del Sol, A., Fujihashi, H., Amoros, D. & Nussinov, R. (2006). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol Syst Biol* **2**, 2006 0019.
74. Mathews, C. K. (1993). The Cell Bag of Enzymes or Network of Channels. *J Bacteriol* **175**, 6377-6381.
75. Spivey, H. O. & Ovadi, J. (1999). Substrate channeling. *Methods* **19**, 306-21.
76. Giege, P., Heazlewood, J. L., Roessner-Tunali, U., Millar, A. H., Fernie, A. R., Leaver, C. J. & Sweetlove, L. J. (2003). Enzymes of glycolysis are functionally associated with the mitochondrion in Arabidopsis cells. *Plant Cell* **15**, 2140-51.
77. Macdonald, P., Almaas, E. & Barabasi, A. (2005). Minimum spanning trees on weighted scale-free networks. *Europhys. Lett.* **72**, 308-314.
78. Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., Jr., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J. & Rothberg, J. M. (2003). A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727-36.
79. Forster, J., Famili, I., Fu, P., Palsson, B. O. & Nielsen, J. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res* **13**, 244-53.
80. Blank, L. M., Kuepfer, L. & Sauer, U. (2005). Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol* **6**, R49.
81. Krogan, N. J., Peng, W. T., Cagney, G., Robinson, M. D., Haw, R., Zhong, G., Guo, X., Zhang, X., Canadien, V., Richards, D. P., Beattie, B. K., Lalev, A., Zhang, W., Davierwala, A. P., Mnaimneh, S., Starostine, A., Tikuisis, A. P., Grigull, J., Datta,

N., Bray, J. E., Hughes, T. R., Emili, A. & Greenblatt, J. F. (2004). High-definition macromolecular composition of yeast RNA-processing complexes. *Mol Cell* **13**, 225-39.

82. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7.
83. Newman, M. E. (2003). The structure and function of complex networks. *SIAM REVIEW* **45**, 167–256.
84. Han, J. D., Dupuy, D., Bertin, N., Cusick, M. E. & Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* **23**, 839-44.
85. Sprinzak, E., Sattath, S. & Margalit, H. (2003). How Reliable are Experimental Protein-Protein Interaction Data? *Journal of Molecular Biology* **327**, 919-923.
86. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.
87. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. & Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res* **28**, 289-91.
88. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A. & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-9.
89. Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. & Botstein, D. (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26**, 73-9.
90. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res* **30**, 42-6.
91. Caspi, R., Foerster, H., Fulcher, C. A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S. Y., Tissier, C., Zhang, P. & Karp, P. D. (2006). MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* **34**, D511-6.
92. Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**, 440-2.
93. Pastor-Satorras, R., Vazquez, A. & Vespignani, A. (2001). Dynamical and correlation properties of the internet. *Phys Rev Lett* **87**, 258701.
94. Newman, M. E. (2002). Assortative mixing in networks. *Phys Rev Lett* **89**, 208701.
95. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**, 3784-8.
96. Blom, N., Gammeltoft, S. & Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* **294**, 1351-62.
97. Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res* **31**, 3635-41.
98. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science* **298**, 1912-34.

99. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. *Springer*.
100. Joachims, T. (1999). Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning. *MIT-Press*.
101. Joachims, T. (1999). Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Boston.
102. Denhardt, D. T. (1996). Signal-transducing protein phosphorylation cascades mediated by Ras/Rho proteins in the mammalian cell: the potential for multiplex signalling. *Biochem J* **318 (Pt 3)**, 729-47.
103. Yaffe, M. B. & Cantley, L. C. (1999). Signal transduction. Grabbing phosphoproteins. *Nature* **402**, 30-1.
104. Nishida, E. & Gotoh, Y. (1993). The MAP kinase cascade is essential for diverse signal transduction pathways. *Trends Biochem Sci* **18**, 128-31.
105. Xue, Y., Li, A., Wang, L., Feng, H. & Yao, X. (2006). PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics* **7**, 163.
106. Kim, J. H., Lee, J., Oh, B., Kimm, K. & Koh, I. (2004). Prediction of phosphorylation sites using SVMs. *Bioinformatics* **20**, 3179-84.
107. Plewczynski, D., Tkacz, A., Godzik, A. & Rychlewski, L. (2005). A support vector machine approach to the identification of phosphorylation sites. *Cell Mol Biol Lett* **10**, 73-89.
108. Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. & Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**, 1633-49.
109. Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z. & Dunker, A. K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* **32**, 1037-49.
110. Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N. & Gibson, T. J. (2004). Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* **5**, 79.
111. Jimenez, J. L., Hegemann, B., Hutchins, J. R., Peters, J. M. & Durbin, R. (2007). A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. *Genome Biol* **8**, R90.
112. Fan, S. C. & Zhang, X. G. (2005). Characterizing the microenvironment surrounding phosphorylated protein sites. *Genomics Proteomics Bioinformatics* **3**, 213-7.
113. Kemp, B. E. & Pearson, R. B. (1990). Protein kinase recognition sequence motifs. *Trends Biochem Sci* **15**, 342-6.
114. Pinna, L. A. & Ruzzene, M. (1996). How do protein kinases recognize their substrates? *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1314**, 191-225.
115. Bagley, S. C. & Altman, R. B. (1995). Characterizing the microenvironment surrounding protein sites. *Protein Sci* **4**, 622-35.
116. Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F. & Mewes, H. W. (2005). Gene selection from microarray data for cancer classification--a machine learning approach. *Comput Biol Chem* **29**, 37-46.
117. Qin, J., Lewis, D. P. & Noble, W. S. (2003). Kernel hierarchical gene clustering from microarray expression data. *Bioinformatics* **19**, 2097-104.
118. Pirooznia, M. & Deng, Y. (2006). SVM Classifier - a comprehensive java interface for support vector machine classification of microarray data. *BMC Bioinformatics* **7 Suppl 4**, S25.
119. Vert, J. P. (2002). Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. *Pac Symp Biocomput*, 649-60.

120. Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. & Muller, K. R. (2000). Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics* **16**, 799-807.
121. Burbidge, R., Trotter, M., Buxton, B. & Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput Chem* **26**, 5-14.
122. Ding, C. H. & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **17**, 349-58.
123. Heazlewood, J. L., Durek, P., Hummel, J., Selbig, J., Weckwerth, W., Walther, D. & Schulze, W. X. (2008). PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. *Nucleic Acids Res* **36**, D1015-21.
124. Hooft, R. W., Sander, C., Scharf, M. & Vriend, G. (1996). The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. *Comput Appl Biosci* **12**, 525-9.
125. Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**, 6097-100.
126. Kreegipuu, A., Blom, N. & Brunak, S. (1999). PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res* **27**, 237-9.
127. Schwartz, D. & Gygi, S. P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol* **23**, 1391-8.
128. Hanks, S. K., Quinn, A. M. & Hunter, T. (1988). The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42-52.
129. Niefind, K., Putter, M., Guerra, B., Issinger, O. G. & Schomburg, D. (1999). GTP plus water mimic ATP in the active site of protein kinase CK2. *Nat Struct Biol* **6**, 1100-3.
130. Reimer, U., Reineke, U. & Schneider-Mergener, J. (2002). Peptide arrays: from macro to micro. *Curr Opin Biotechnol* **13**, 315-20.
131. Rychlewski, L., Kschischo, M., Dong, L., Schutkowski, M. & Reimer, U. (2004). Target specificity analysis of the Abl kinase using peptide microarray data. *J Mol Biol* **336**, 307-11.
132. Mah, A. S., Elia, A. E., Devgan, G., Ptacek, J., Schutkowski, M., Snyder, M., Yaffe, M. B. & Deshaies, R. J. (2005). Substrate specificity analysis of protein kinase complex Dbf2-Mob1 by peptide library and proteome array screening. *BMC Biochem* **6**, 22.
133. Zhou, T., Sun, L., Humphreys, J. & Goldsmith, E. J. (2006). Docking interactions induce exposure of activation loop in the MAP kinase ERK2. *Structure* **14**, 1011-9.
134. Hunter, T. (1987). A thousand and one protein kinases. *Cell* **50**, 823-9.
135. Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., McCartney, R. R., Schmidt, M. C., Rachidi, N., Lee, S. J., Mah, A. S., Meng, L., Stark, M. J., Stern, D. F., De Virgilio, C., Tyers, M., Andrews, B., Gerstein, M., Schweitzer, B., Predki, P. F. & Snyder, M. (2005). Global analysis of protein phosphorylation in yeast. *Nature* **438**, 679-84.
136. Weckwerth, W. & Selbig, J. (2003). Scoring and identifying organism-specific functional patterns and putative phosphorylation sites in protein sequences using mutual information. *Biochem Biophys Res Commun* **307**, 516-21.
137. Cheng, K. Y., Noble, M. E., Skamnaki, V., Brown, N. R., Lowe, E. D., Kontogiannis, L., Shen, K., Cole, P. A., Siligardi, G. & Johnson, L. N. (2006). The role of the

phospho-CDK2/cyclin A recruitment site in substrate recognition. *J Biol Chem* **281**, 23167-79.

138. Remenyi, A., Good, M. C. & Lim, W. A. (2006). Docking interactions in protein kinase and phosphatase networks. *Curr Opin Struct Biol* **16**, 676-85.
139. Higgins, D. G. & Sharp, P. M. (1989). Fast and sensitive multiple sequence alignments on a microcomputer. *Comput Appl Biosci* **5**, 151-3.
140. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80.
141. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56-68.
142. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637.
143. Park, Y. & Helms, V. (2006). Assembly of transmembrane helices of simple polytopic membrane proteins from sequence conservation patterns. *Proteins* **64**, 895-905.
144. Levitt, M. (1978). Conformational preferences of amino acids in globular proteins. *Biochemistry* **17**, 4277-85.
145. Alexandros, K., Alexandros, S., Kurt, H. & Achim, Z. kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software* **11**.
146. Kawashima, S. & Kanehisa, M. (2000). AAindex: amino acid index database. *Nucleic Acids Res* **28**, 374.
147. Wong, Y. H., Lee, T. Y., Liang, H. K., Huang, C. M., Wang, T. Y., Yang, Y. H., Chu, C. H., Huang, H. D., Ko, M. T. & Hwang, J. K. (2007). KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res* **35**, W588-94.
148. Chung, H. J., Sehnke, P. C. & Ferl, R. J. (1999). The 14-3-3 proteins: cellular regulators of plant metabolism. *Trends in Plant Science* **4**, 367-371.
149. Yaffe, M. B. (2002). Phosphotyrosine-binding domains in signal transduction. *Nature Reviews Molecular Cell Biology* **3**, 177-186.
150. Pawson, T. (2004). Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell* **116**, 191-203.
151. Pawson, T. & Gish, G. D. (1992). SH2 and SH3 domains: from structure to function. *Cell* **71**, 359-362.
152. Camoni, L., Iori, V., Marra, M. & Aducci, P. (2000). Phosphorylation-dependent interaction between plant plasma membrane H(+)ATPase and 14-3-3 proteins. *Journal of Biological Chemistry* **275**, 99919-9923.
153. Hrabak, E. M., Chan, C. W., Gribskov, M., Harper, J. F., Choi, J. H., Halford, N., Kudla, J., Luan, S., Nimmo, H. G., Sussman, M. R., Thomas, M., Walker-Simmons, K., Zhu, J. K. & Harmon, A. C. (2003). The Arabidopsis CDPK-SnRK superfamily of protein kinases. *Plant Physiology* **132**, 666-680.
154. Wang, X., Goshe, M. B., Sonderblom, E. J., Phinney, B. S., Kuchar, J. A., Li, J., Asami, T., Yoshida, S., Huber, S. C. & Clouse, S. D. (2005). Identification and functional analysis of in vivo phosphorylation sites of the Arabidopsis Brassinosteroid-insesnitive 1 receptor kinase. *The Plant Cell* **17**, 1685-1703.
155. Yoshida, S. & Parniske, M. (2005). Regulation of plant symbiosis receptor kinase through serine and threonine phosphorylation. *Journal of Biological Chemistry* **280**, 9203-9209.

156. Sugiyama, N., Nakagami, H., Mochida, K., Daudi, A., Tomita, M., Shirasu, K. & Ishihama, Y. (2008). Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis. *Mol Syst Biol* **4**, 193.
157. Nühse, T. S., Stensballe, A., Jensen, O. N. & Peck, J. (2003). Large-scale analysis of *in vivo* phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. *Molecular and Cellular Proteomics* **2**, 1234-1243.
158. Wolschin, F. & Weckwerth, W. (2005). Combining metal oxide affinity chromatography (MOAC) and selective mass spectrometry for robust identification of in vivo protein phosphorylation sites. *Plant Methods* **1**, 1-10.
159. Wolschin, F., Lehmann, U., Glinski, M. & Weckwerth, W. (2005). An integrated strategy for identification and relative quantification of site-specific protein phosphorylation using liquid chromatography coupled to MS2/MS3. *Rapid Communications in Mass Spectrometry* **19**, 3626-3632.
160. Nühse, T. S., Stensballe, A., Jensen, O. N. & Peck, S. C. (2004). Phosphoproteomics of the Arabidopsis plasma membrane and a new phosphorylation site database. *Plant Cell* **16**, 2394-23405.
161. Benschop, J. J., Mohammed, S., O'Flaherty, M., Heck, A. J., Slijper, M. & Menke, F. L. (2007). Quantitative phospho-proteomics of early elicitor signalling in Arabidopsis. *Molecular and Cellular Proteomics*.
162. Niittylä, T., Fuglsang, A. T., Palmgren, M. G., Frommer, W. B. & Schulze, W. X. (2007). Temporal analysis of sucrose-induced phosphorylation changes in plasma membrane proteins of Arabidopsis. *Molecular and Cellular Proteomics* **in press**.
163. de la Fuente van Bentem, S., Anrather, D., Roitinger, E., Djamei, A., Hufnagl, T., Barta, A., Csaszar, E., Dohnal, I., Lecourieux, D. & Hirt, H. (2006). Phosphoproteomics reveals extensive in vivo phosphorylation of Arabidopsis proteins involved in RNA metabolism. *Nucleic Acids Research* **34**, 3267-3278.
164. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P. & Mann, M. (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635-648.
165. Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E. & Zhang, B. (2004). PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **4**, 1551-61.
166. Peck, S. C. (2006). Phosphoproteomics in Arabidopsis: moving from empirical to predictive science. *J Exp Bot* **57**, 1523-7.
167. Hummel, J., Niemann, M., Wienkoop, S., Schulze, W., Steinhauser, D., Selbig, J., Walther, D. & Weckwerth, W. (2007). ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. *BMC Bioinformatics* **8**, 216.
168. Glinski, M. & Weckwerth, W. (2005). Differential multisite phosphorylation of the trehalose-6-phosphate synthase gene family in Arabidopsis thaliana: A mass spectrometry-based process for multiparallel peptide library phosphorylation analysis. *Molecular and Cellular Proteomics* **4**, 1614-1625.
169. Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L. A., Bhattacharyya, D., Bhaya, D., Sobral, B. W., Beavis, W., Meinke, D. W., Town, C. D., Somerville, C. & Rhee, S. Y. (2001). The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Research* **29**, 102-105.

170. Blom, N., Gammeltoft, S. & Brunak, S. (1999). Sequence- and structure-based prediction of eucaryotic protein phosphorylation sites. *Journal of Molecular Biology* **294**, 1351-1362.
171. Hanley, J. A. & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **3**.
172. Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289-300.
173. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25-29.
174. Weckwerth, W. & Selbig, J. (2003). Scoring and identifying organism-specific functional patterns and putative phosphorylation sites in protein sequences using mutual information. *Biochemical and Biophysical Research Communications* **307**, 516-521.
175. Kawashima, S. & Kanehisa, M. (2000). AAindex: amino acid index database. *Nucleic Acids Research* **28**, 374.
176. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861-874.
177. Hart, G. T., Ramani, A. K. & Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol* **7**, 120.
178. Saito, R., Suzuki, H. & Hayashizaki, Y. (2002). Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Res* **30**, 1163-8.
179. Saito, R., Suzuki, H. & Hayashizaki, Y. (2003). Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics* **19**, 756-63.
180. Gilchrist, M. A., Salter, L. A. & Wagner, A. (2004). A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics* **20**, 689-700.
181. Bader, J. S., Chaudhuri, A., Rothberg, J. M. & Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* **22**, 78-85.
182. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. & Weil, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**, 31-4.
183. Leach, S., Gabow, A., Hunter, L. & Goldberg, D. S. (2007). Assessing and combining reliability of protein interaction sources. *Pac Symp Biocomput*, 433-44.

Appendix A.

GO Terms used for identifying Protein Degradation/Ubiquitin associated proteins

GO:Term ID	Description
GO:0000151	ubiquitin ligase complex
GO:0001509	legumain activity
GO:0004175	endopeptidase activity
GO:0004176	ATP-dependent peptidase activity
GO:0004177	aminopeptidase activity
GO:0004178	leucyl aminopeptidase activity
GO:0004179	membrane alanyl aminopeptidase activity
GO:0004180	carboxypeptidase activity
GO:0004182	carboxypeptidase A activity
GO:0004185	serine carboxypeptidase activity
GO:0004186	carboxypeptidase C activity
GO:0004187	carboxypeptidase D activity
GO:0004190	aspartic-type endopeptidase activity
GO:0004191	barrierpepsin activity
GO:0004194	pepsin A activity
GO:0004196	saccharopepsin activity
GO:0004197	cysteine-type endopeptidase activity
GO:0004198	calpain activity
GO:0004221	ubiquitin thiolesterase activity
GO:0004222	metalloendopeptidase activity
GO:0004226	Gly-X carboxypeptidase activity
GO:0004239	methionyl aminopeptidase activity
GO:0004240	mitochondrial processing peptidase activity
GO:0004243	mitochondrial intermediate peptidase activity
GO:0004244	mitochondrial inner membrane peptidase activity
GO:0004247	saccharolysin activity
GO:0004250	aminopeptidase I activity
GO:0004252	serine-type endopeptidase activity
GO:0004262	cerevisin activity
GO:0004274	dipeptidyl-peptidase IV activity
GO:0004287	prolyl oligopeptidase activity
GO:0004289	subtilase activity
GO:0004298	threonine endopeptidase activity
GO:0004839	ubiquitin activating enzyme activity
GO:0004839	ubiquitin activating enzyme activity
GO:0004842	ubiquitin conjugating enzyme activity
GO:0004843	ubiquitin-specific protease activity
GO:0004866	endopeptidase inhibitor activity
GO:0004867	serine-type endopeptidase inhibitor activity
GO:0005680	anaphase-promoting complex
GO:0006511	ubiquitin-dependent protein catabolic process
GO:0006512	ubiquitin cycle
GO:0008054	cyclin catabolic process
GO:0008233	peptidase activity
GO:0008234	cysteine-type peptidase activity
GO:0008235	metalloexopeptidase activity
GO:0008236	serine-type peptidase activity

GO:Term ID	Description
GO:0008237	metallopeptidase activity
GO:0008423	bleomycin hydrolase activity
GO:0008450	O-sialoglycoprotein endopeptidase activity
GO:0008451	X-Pro aminopeptidase activity
GO:0008487	prenyl-dependent CAAX protease activity
GO:0008641	small protein activating enzyme activity
GO:0008717	D-alanyl-D-alanine endopeptidase activity
GO:0008769	X-His dipeptidase activity
GO:0009003	signal peptidase activity
GO:0009049	aspartic-type signal peptidase activity
GO:0016574	histone ubiquitination
GO:0016806	dipeptidyl-peptidase and tripeptidyl-peptidase activity
GO:0016929	SUMO-specific protease activity
GO:0017039	dipeptidyl-peptidase III activity
GO:0019005	SCF ubiquitin ligase complex
GO:0019778	APG12 activating enzyme activity
GO:0019779	APG8 activating enzyme activity
GO:0019781	NEDD8 activating enzyme activity
GO:0019787	small conjugating protein ligase activity
GO:0019789	SUMO ligase activity
GO:0019948	SUMO activating enzyme activity
GO:0030414	protease inhibitor activity
GO:0030433	ER-associated protein catabolic process
GO:0030693	caspase activity
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
GO:0031146	SCF-dependent proteasomal ubiquitin-dependent protein catabolic process
GO:0031371	ubiquitin conjugating enzyme complex
GO:0031386	protein tag
GO:0031463	Cul3-RING ubiquitin ligase complex
GO:0032435	negative regulation of proteasomal ubiquitin-dependent protein catabolic process
GO:0042292	URM1 activating enzyme activity
GO:0042576	aspartyl aminopeptidase activity
GO:0042787	protein ubiquitination during ubiquitin-dependent protein catabolic process
GO:0043130	ubiquitin binding
GO:0043161	proteasomal ubiquitin-dependent protein catabolic process
GO:0043162	ubiquitin-dependent protein catabolic process via the multivesicular body pathway
GO:0043224	nuclear SCF ubiquitin ligase complex
GO:0043328	protein targeting to vacuole during ubiquitin-dependent protein catabolic process via the MVB pathway
GO:0051436	negative regulation of ubiquitin ligase activity during mitotic cell cycle
GO:0051443	positive regulation of ubiquitin ligase activity

GO Annotations used for identifying Kinase/Phosphatase proteins

GO:Term ID	Description
GO:0000155	two-component sensor activity
GO:0000156	two-component response regulator activity
GO:0000158	protein phosphatase type 2A activity
GO:0000159	protein phosphatase type 2A complex
GO:0000163	protein phosphatase type 1 activity
GO:0000164	protein phosphatase type 1 complex
GO:0003924	GTPase activity
GO:0004672	protein kinase activity
GO:0004673	protein histidine kinase activity
GO:0004674	protein serine/threonine kinase activity
GO:0004679	AMP-activated protein kinase activity
GO:0004680	casein kinase activity
GO:0004681	casein kinase I activity
GO:0004682	protein kinase CK2 activity
GO:0004683	calmodulin regulated protein kinase activity
GO:0004684	calmodulin-dependent protein kinase I activity
GO:0004685	calcium- and calmodulin-dependent protein kinase activity
GO:0004691	cAMP-dependent protein kinase activity
GO:0004693	cyclin-dependent protein kinase activity
GO:0004694	eukaryotic translation initiation factor 2alpha kinase activity
GO:0004696	glycogen synthase kinase 3 activity
GO:0004697	protein kinase C activity
GO:0004702	receptor signaling protein serine/threonine kinase activity
GO:0004707	MAP kinase activity
GO:0004708	MAP kinase kinase activity
GO:0004709	MAP kinase kinase kinase activity
GO:0004712	protein threonine/tyrosine kinase activity
GO:0004713	protein-tyrosine kinase activity
GO:0004721	phosphoprotein phosphatase activity
GO:0004722	protein serine/threonine phosphatase activity
GO:0004723	calcium-dependent protein serine/threonine phosphatase activity
GO:0004725	protein tyrosine phosphatase activity
GO:0004727	prenylated protein tyrosine phosphatase activity
GO:0004860	protein kinase inhibitor activity
GO:0004861	cyclin-dependent protein kinase inhibitor activity
GO:0004862	cAMP-dependent protein kinase inhibitor activity
GO:0004864	protein phosphatase inhibitor activity
GO:0004871	signal transducer activity
GO:0004872	receptor activity
GO:0004888	transmembrane receptor activity
GO:0004930	G-protein coupled receptor activity
GO:0004932	mating-type factor pheromone receptor activity
GO:0004933	mating-type a-factor pheromone receptor activity
GO:0004934	mating-type alpha-factor pheromone receptor activity
GO:0005034	osmosensor activity
GO:0005057	receptor signaling protein activity
GO:0005083	small GTPase regulator activity
GO:0005955	calcineurin complex
GO:0008138	protein tyrosine/serine/threonine phosphatase activity
GO:0008158	hedgehog receptor activity
GO:0008287	protein serine/threonine phosphatase complex

GO:Term ID	Description
GO:0008330	protein tyrosine/threonine phosphatase activity
GO:0008349	MAP kinase kinase kinase kinase activity
GO:0008597	calcium-depend. protein serine/threonine phosphatase regulator activity
GO:0008599	protein phosphatase type 1 regulator activity
GO:0008601	protein phosphatase type 2A regulator activity
GO:0008603	cAMP-dependent protein kinase regulator activity
GO:0008605	protein kinase CK2 regulator activity
GO:0015071	protein phosphatase type 2C activity
GO:0016299	regulator of G-protein signaling activity
GO:0016301	kinase activity
GO:0016538	cyclin-dependent protein kinase regulator activity
GO:0017017	MAP kinase phosphatase activity
GO:0019207	kinase regulator activity
GO:0019209	kinase activator activity
GO:0019211	phosphatase activator activity
GO:0019828	aspartic-type endopeptidase inhibitor activity
GO:0019887	protein kinase regulator activity
GO:0019888	protein phosphatase regulator activity
GO:0019912	cyclin-dependent protein kinase activating kinase activity
GO:0030295	protein kinase activator activity
GO:0030695	GTPase regulator activity
GO:0033550	MAP kinase tyrosine phosphatase activity
GO:0035174	histone serine kinase activity
GO:0043539	protein serine/threonine kinase activator activity

GO Annotations used for identifying DNA-related proteins

GO:Term ID	Description
GO:0000014	single-stranded DNA specific endodeoxyribonuclease activity
GO:0000049	tRNA binding
GO:0000124	SAGA complex
GO:0000149	SNARE binding
GO:0000175	3'-5'-exoribonuclease activity
GO:0000179	rRNA (adenine-N6,N6-)-dimethyltransferase activity
GO:0000182	rDNA binding
GO:0000213	tRNA-intron endonuclease activity
GO:0000215	tRNA 2'-phosphotransferase activity
GO:0000339	RNA cap binding
GO:0000400	four-way junction DNA binding
GO:0000403	Y-form DNA binding
GO:0003677	DNA binding
GO:0003678	DNA helicase activity
GO:0003680	AT DNA binding
GO:0003682	chromatin binding
GO:0003684	damaged DNA binding
GO:0003688	DNA replication origin binding
GO:0003689	DNA clamp loader activity
GO:0003690	double-stranded DNA binding
GO:0003697	single-stranded DNA binding
GO:0003700	transcription factor activity
GO:0003701	RNA polymerase I transcription factor activity
GO:0003702	RNA polymerase II transcription factor activity

GO:Term ID	Description
GO:0003704	specific RNA polymerase II transcription factor activity
GO:0003706	ligand-regulated transcription factor activity
GO:0003709	RNA polymerase III transcription factor activity
GO:0003711	transcriptional elongation regulator activity
GO:0003712	transcription cofactor activity
GO:0003713	transcription coactivator activity
GO:0003714	transcription corepressor activity
GO:0003723	RNA binding
GO:0003724	RNA helicase activity
GO:0003729	mRNA binding
GO:0003735	structural constituent of ribosome
GO:0003743	translation initiation factor activity
GO:0003746	translation elongation factor activity
GO:0003747	translation release factor activity
GO:0003887	DNA-directed DNA polymerase activity
GO:0003889	alpha DNA polymerase activity
GO:0003890	beta DNA polymerase activity
GO:0003891	delta DNA polymerase activity
GO:0003893	epsilon DNA polymerase activity
GO:0003894	zeta DNA polymerase activity
GO:0003895	gamma DNA-directed DNA polymerase activity
GO:0003896	DNA primase activity
GO:0003899	DNA-directed RNA polymerase activity
GO:0003905	alkylbase DNA N-glycosylase activity
GO:0003906	DNA-(apurinic or apyrimidinic site) lyase activity
GO:0003908	methylated-DNA-[protein]-cysteine S-methyltransferase activity
GO:0003910	DNA ligase (ATP) activity
GO:0003917	DNA topoisomerase type I activity
GO:0003918	DNA topoisomerase (ATP-hydrolyzing) activity
GO:0004003	ATP-dependent DNA helicase activity
GO:0004004	ATP-dependent RNA helicase activity
GO:0004045	aminoacyl-tRNA hydrolase activity
GO:0004402	histone acetyltransferase activity
GO:0004406	H3/H4 histone acetyltransferase activity
GO:0004407	histone deacetylase activity
GO:0004479	methionyl-tRNA formyltransferase activity
GO:0004482	mRNA (guanine-N7-)-methyltransferase activity
GO:0004484	mRNA guanylyltransferase activity
GO:0004808	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase activity
GO:0004809	tRNA (guanine-N2-)-methyltransferase activity
GO:0004810	tRNA adenylyltransferase activity
GO:0004811	tRNA isopentenyltransferase activity
GO:0004813	alanine-tRNA ligase activity
GO:0004814	arginine-tRNA ligase activity
GO:0004815	aspartate-tRNA ligase activity
GO:0004816	asparagine-tRNA ligase activity
GO:0004817	cysteine-tRNA ligase activity
GO:0004818	glutamate-tRNA ligase activity
GO:0004819	glutamine-tRNA ligase activity
GO:0004820	glycine-tRNA ligase activity
GO:0004821	histidine-tRNA ligase activity
GO:0004822	isoleucine-tRNA ligase activity
GO:0004823	leucine-tRNA ligase activity

GO:Term ID	Description
GO:0004824	lysine-tRNA ligase activity
GO:0004825	methionine-tRNA ligase activity
GO:0004826	phenylalanine-tRNA ligase activity
GO:0004827	proline-tRNA ligase activity
GO:0004828	serine-tRNA ligase activity
GO:0004829	threonine-tRNA ligase activity
GO:0004830	tryptophan-tRNA ligase activity
GO:0004831	tyrosine-tRNA ligase activity
GO:0004832	valine-tRNA ligase activity
GO:0004844	uracil DNA N-glycosylase activity
GO:0005671	Ada2/Gcn5/Ada3 transcription activator complex
GO:0006608	snRNP protein import into nucleus
GO:0006609	mRNA-binding (hnRNP) protein import into nucleus
GO:0008079	translation termination factor activity
GO:0008094	DNA-dependent ATPase activity
GO:0008134	transcription factor binding
GO:0008135	translation factor activity, nucleic acid binding
GO:0008159	positive transcription elongation factor activity
GO:0008173	RNA methyltransferase activity
GO:0008174	mRNA methyltransferase activity
GO:0008175	tRNA methyltransferase activity
GO:0008193	tRNA guanylyltransferase activity
GO:0008251	tRNA specific adenosine deaminase activity
GO:0008301	DNA bending activity
GO:0008419	RNA lariat debranching enzyme activity
GO:0008534	oxidized purine base lesion DNA N-glycosylase activity
GO:0008650	rRNA (uridine-2'-O-)-methyltransferase activity
GO:0008989	rRNA (guanine-N1-)-methyltransferase activity
GO:0010390	histone monoubiquitination
GO:0015999	eta DNA polymerase activity
GO:0016149	translation release factor activity, codon specific
GO:0016251	general RNA polymerase II transcription factor activity
GO:0016423	tRNA (guanine) methyltransferase activity
GO:0016424	tRNA (guanosine) methyltransferase activity
GO:0016428	tRNA (cytosine-5-)-methyltransferase activity
GO:0016429	tRNA (adenine-N1-)-methyltransferase activity
GO:0016431	tRNA (uridine) methyltransferase activity
GO:0016439	tRNA-pseudouridine synthase activity
GO:0016455	RNA polymerase II transcription mediator activity
GO:0016563	transcriptional activator activity
GO:0016564	transcriptional repressor activity
GO:0016565	general transcriptional repressor activity
GO:0016566	specific transcriptional repressor activity
GO:0016944	RNA polymerase II transcription elongation factor activity
GO:0017005	tyrosyl-DNA phosphodiesterase activity
GO:0017136	NAD-dependent histone deacetylase activity
GO:0017150	tRNA dihydrouridine synthase activity
GO:0019237	centromeric DNA binding
GO:0019843	rRNA binding
GO:0030188	chaperone regulator activity
GO:0030337	DNA polymerase processivity factor activity
GO:0030371	translation repressor activity
GO:0030515	snoRNA binding

GO:Term ID	Description
GO:0030528	transcription regulator activity
GO:0030620	U2 snRNA binding
GO:0031202	RNA splicing factor activity, transesterification mechanism
GO:0031490	chromatin DNA binding
GO:0032041	NAD-dependent histone deacetylase activity (H3-K14 specific)
GO:0032777	Piccolo NuA4 histone acetyltransferase complex
GO:0033100	NuA3 histone acetyltransferase complex
GO:0035267	NuA4 histone acetyltransferase complex
GO:0042054	histone methyltransferase activity
GO:0042134	rRNA primary transcript binding
GO:0042162	telomeric DNA binding
GO:0042393	histone binding
GO:0042800	histone lysine N-methyltransferase activity (H3-K4 specific)
GO:0043140	ATP-dependent 3' to 5' DNA helicase activity
GO:0043141	ATP-dependent 5' to 3' DNA helicase activity
GO:0043166	H4/H2A histone acetyltransferase activity
GO:0043189	H4/H2A histone acetyltransferase complex
GO:0045129	NAD-independent histone deacetylase activity
GO:0045182	translation regulator activity
GO:0046695	SLIK (SAGA-like) complex
GO:0046969	NAD-dependent histone deacetylase activity (H3-K9 specific)
GO:0046970	NAD-dependent histone deacetylase activity (H4-K16 specific)
GO:0050072	m7G(5')pppN diphosphatase activity
GO:0051500	D-tyrosyl-tRNA(Tyr) deacylase activity
GO:0051864	histone demethylase activity (H3-K36 specific)

GO Annotations used for identifying other, non-metabolic proteins

GO:Term ID	Description
GO:0000054	ribosome export from nucleus
GO:0000055	ribosomal large subunit export from nucleus
GO:0000056	ribosomal small subunit export from nucleus
GO:0000059	protein import into nucleus, docking
GO:0000060	protein import into nucleus, translocation
GO:0000149	SNARE binding
GO:0000208	nuclear translocation of MAPK during osmolarity sensing
GO:0000268	peroxisome targeting sequence binding
GO:0000290	deadenylation-dependent decapping
GO:0001671	ATPase stimulator activity
GO:0003923	GPI-anchor transamidase activity
GO:0003924	GTPase activity
GO:0004175	endopeptidase activity
GO:0004596	peptide alpha-N-acetyltransferase activity
GO:0004857	enzyme inhibitor activity
GO:0004860	protein kinase inhibitor activity
GO:0004871	signal transducer activity
GO:0005084	Rab escort protein activity
GO:0005085	guanyl-nucleotide exchange factor activity
GO:0005086	ARF guanyl-nucleotide exchange factor activity
GO:0005088	Ras guanyl-nucleotide exchange factor activity
GO:0005089	Rho guanyl-nucleotide exchange factor activity
GO:0005093	Rab GDP-dissociation inhibitor activity

GO:Term ID	Description
GO:0005094	Rho GDP-dissociation inhibitor activity
GO:0005095	GTPase inhibitor activity
GO:0005096	GTPase activator activity
GO:0005097	Rab GTPase activator activity
GO:0005098	Ran GTPase activator activity
GO:0005099	Ras GTPase activator activity
GO:0005100	Rho GTPase activator activity
GO:0006605	protein targeting
GO:0006606	protein import into nucleus
GO:0006607	NLS-bearing substrate import into nucleus
GO:0006610	ribosomal protein import into nucleus
GO:0006611	protein export from nucleus
GO:0006612	protein targeting to membrane
GO:0006613	cotranslational protein targeting to membrane
GO:0006614	SRP-dependent cotranslational protein targeting to membrane
GO:0006616	SRP-dependent cotransl. prot. target. to membrane, translocation
GO:0006617	SRP-dep. cotransl. prot. target. to mem., signal sequence recognition
GO:0006620	posttranslational protein targeting to membrane
GO:0006623	protein targeting to vacuole
GO:0006625	protein targeting to peroxisome
GO:0006626	protein targeting to mitochondrion
GO:0006627	mitochondrial protein processing
GO:0006886	intracellular protein transport
GO:0007092	anaphase-promoting complex activation during mitotic cell cycle
GO:0008047	enzyme activator activity
GO:0008060	ARF GTPase activator activity
GO:0008139	nuclear localization sequence binding
GO:0008308	voltage-gated ion-selective channel activity
GO:0008320	protein carrier activity
GO:0008538	proteasome activator activity
GO:0008565	protein transporter activity
GO:0015664	nicotinamide mononucleotide permease activity
GO:0016538	cyclin-dependent protein kinase regulator activity
GO:0016558	protein import into peroxisome matrix
GO:0016560	protein import into peroxisome matrix, docking
GO:0016562	protein import into peroxisome matrix, receptor recycling
GO:0016598	protein arginylation
GO:0016925	protein sumoylation
GO:0017112	Rab guanyl-nucleotide exchange factor activity
GO:0019211	phosphatase activator activity
GO:0030150	protein import into mitochondrial matrix
GO:0030189	chaperone activator activity
GO:0030190	chaperone inhibitor activity
GO:0030234	enzyme regulator activity
GO:0030970	retrograde protein transport, ER to cytosol
GO:0031204	posttranslational protein targeting to membrane, translocation
GO:0031386	protein tag
GO:0032258	CVT pathway
GO:0032527	protein exit from endoplasmic reticulum
GO:0042719	mitochondrial intermembrane space protein transporter complex
GO:0042992	negative regulation of transcription factor import into nucleus
GO:0042994	cytoplasmic sequestering of transcription factor
GO:0043001	Golgi to plasma membrane protein transport

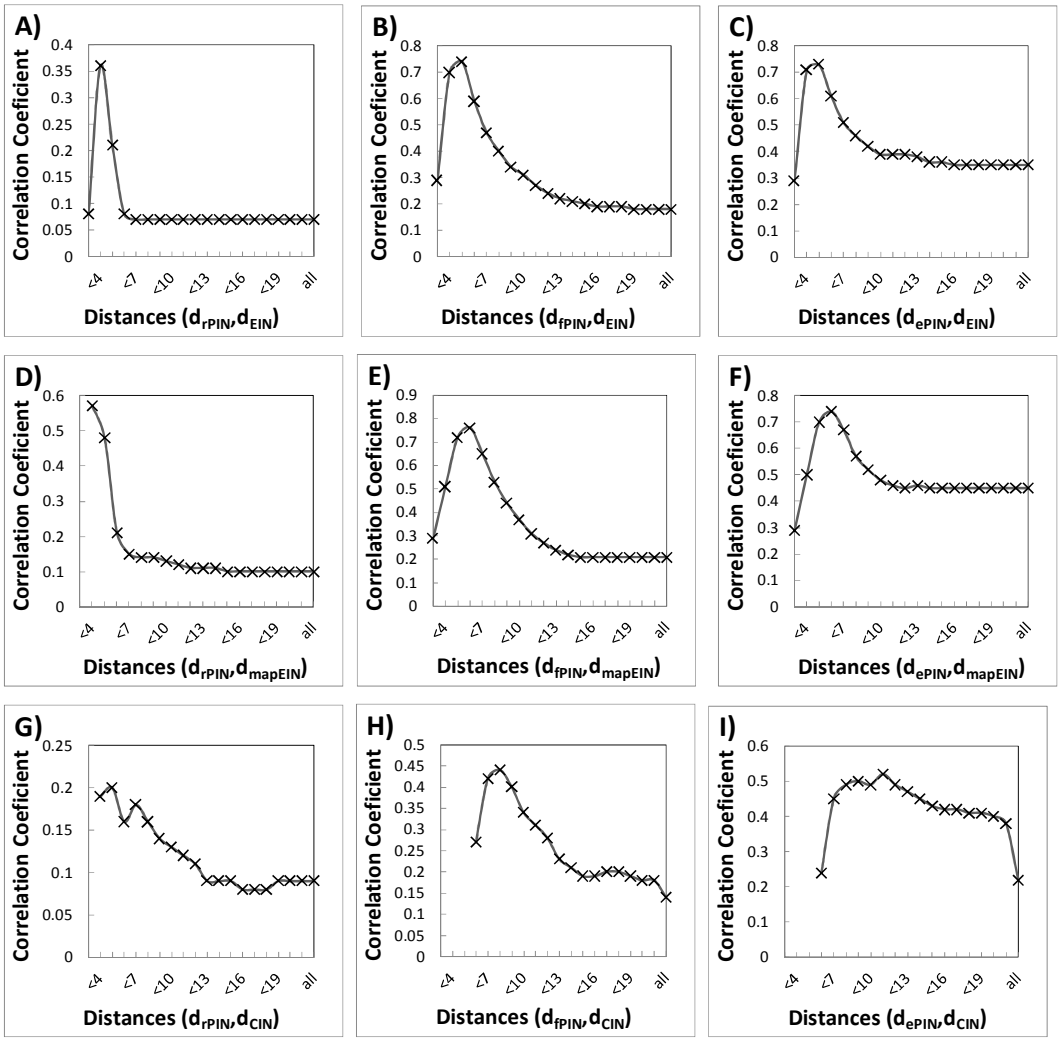
GO:Term ID	Description
GO:0043328	Prot. targeting to vacuole during ubiquitin-dependet catabolic proc.
GO:0045039	protein import into mitochondrial inner membrane
GO:0045040	protein import into mitochondrial outer membrane
GO:0045041	protein import into mitochondrial intermembrane space
GO:0045046	protein import into peroxisome membrane
GO:0045047	protein targeting to ER
GO:0048306	calcium-dependent protein binding
GO:0051082	unfolded protein binding

Currency metabolites, co-factors removed from the metabolic network

KEGG-ID	Metabolite	KEGG-ID	Metabolite
C00001	H2o	C00050	Metal
C00002	ATP	C00055	CMP
C00003	NAD+	C00061	FMN
C00004	NADH	C00063	CTP
C00005	NADPH	C00070	Copper
C00006	NADP+	C00075	UTP
C00007	Oxygen;O2	C00076	Calcium;Ca2+
C00008	ADP	C00080	H+
C00009	Orthophosphate,Pi	C00105	UMP
C00010	CoA	C00112	CDP
C00011	CO2	C00113	PQQ
C00013	Pyrophosphate;PPi	C00115	Chloride
C00014	NH3	C00120	Biotin; Coenzyme R
C00015	UDP	C00125	Ferricytochromec; Cytochromec3+
C00016	FAD		
C00018	Pyridoxalphosphate	C00126	Ferrocytochromec; Cytochrome C 2+;
C00019	S-Adenosyl-L-methionine; SAM		
C00020	AMP	C00138	Reducedferredoxin
C00021	S-Adenosylhomocysteine; SAH	C00139	Oxidizedferredoxin
C00023	Iron	C00144	GMP
C00027	H2O2	C00175	Cobalt;Co2+
C00028	Acceptor	C00194	Cobamidecoenzyme
C00030	Reduced acceptor	C00205	hn;Light
C00034	Manganese	C00238	Potassium;K+
C00035	GDP	C00291	Nickel;Ni2+
C00038	Zinc	C01352	FADH2
C00044	GTP		

Appendix B.

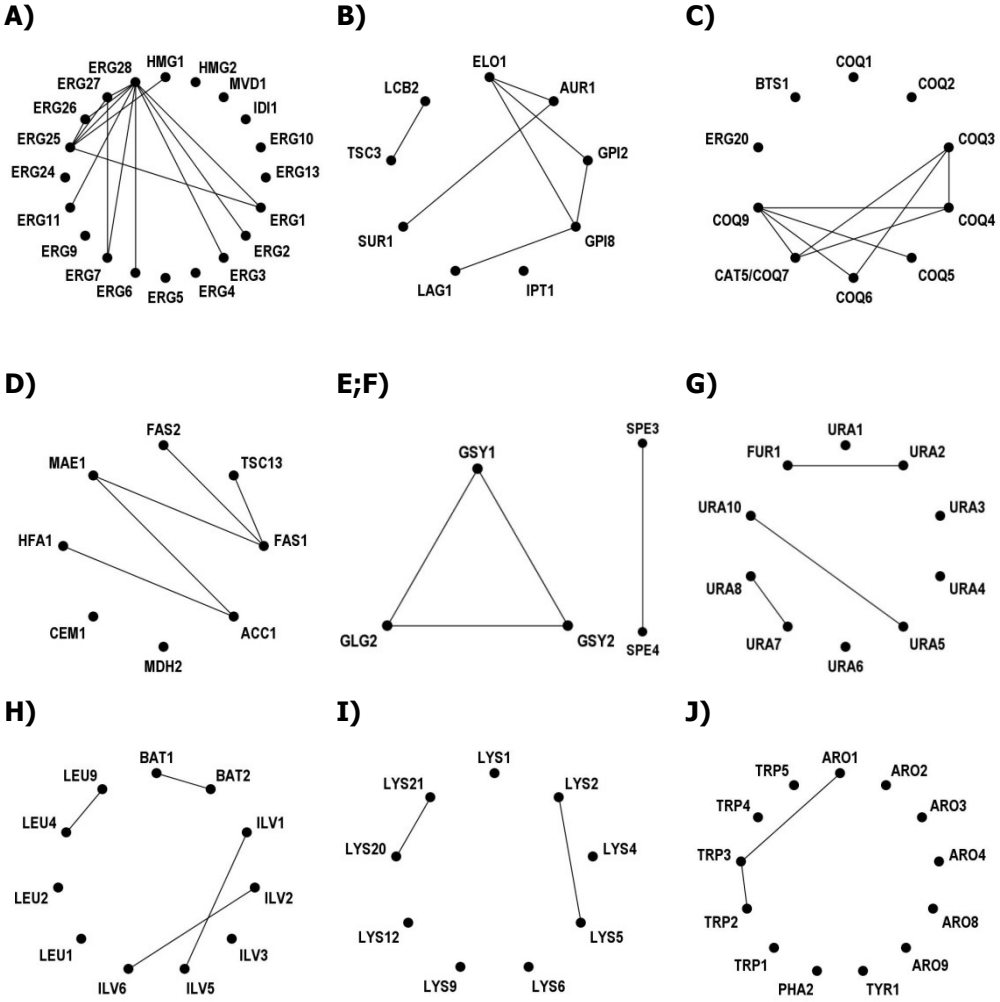
Distribution of correlations according to the included distances



Correlation Coefficient in dependence of considered distances in PINs and MINs. The correlation of distances couples were calculated including distance couples shorter than thresholds of a range form d_{PIN} and d_{MIN} <3 to all distances. Only significant correlations were shown in the figures. All correlations exhibit a p-Value < 1E-7. A) rPIN and EIN; B) fPIN and EIN; C) ePIN and EIN; D) rPIN and mapEIN; E) fPIN and mapEIN; F) ePIN and mapEIN; G) rPIN and CIN; H) fPIN and CIN and I) ePIN and EIN I). The correlations usually decrease with the considered distances, as the correlations in short distances influence distances in longer distance.

Appendix C.

Detected physical interaction of enzymes involved in selected pathways



A) ergosterol biosynthesis; B) sphingolipid biosynthesis; C) ubiquinone biosynthesis; D) fatty acid biosynthesis; E) glycogen biosynthesis; F) last step of polyamine biosynthesis; G) de novo biosynthesis of pyrimidine ribonucleotides; H) superpathway of leucine, isoleucine, and valine biosynthesis; I) lysine biosynthesis; J) superpathway of phenylalanine, tyrosine and tryptophan biosynthesis. In picture A) HMG1/2, MVD1, IDI1 and ERG10/13 are part of the mevalonate pathway. All pathways are derived from the SGD Database. Only enzymes contained in the PIN are visualized, i.e. the pathways are not complete in a biochemical sense. For the fatty acids biosynthesis, the malic enzymes (MAE) as well as the malate dehydrogenase (MDH2) were included as sources of NADPH and AcetylCoA. Enzymes are abbreviated by their gene symbols and detected interactions between them are denoted by connecting lines.

Selbständigkeitserklärung

Hiermit erkläre ich, dass die vorliegende Arbeit an keiner anderen Hochschule eingereicht sowie selbständig und nur mit den angegebenen Mitteln angefertigt wurde.

Potsdam, Dezember 2008
Pawel Durek