Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Bioinformatics of Phosphoproteomics

Florian Gnad

aus

München

2008

Erklärung

Diese Dissertation wurde im Sinne von § 13 Abs 3 der
Promotionsordnung vom 29. Januar 1998 von Herrn
Prof. Matthias Mann betreut.

Ehrenwörtliche Versicherung

Diese Dissertation wurde selbständig, ohne unerlaubte Hilfe erarbeitet.

München, am ____________

____
(Florian Gnad)

Dissertation eingereicht am ______________

1. Gutachter Prof. Matthias Mann

2. Gutachter Prof. John Parsch

Mündliche Prüfung am 14.11.2008

Contents

1 Introduction 1-7

2 Background: Mass Spectrometry, Database Systems and ASP.NET 8-19

2.1 Mass Spectrometry based Proteomics 8-11
2.2 Database Systems 12-16
2.2.1 Components and Functions of Database Systems 12-13
2.2.2 Architecture of Database Systems 14
2.2.3 Relational Model 15
2.2.4 Query Language SQL 15-16
2.3 Web Development in ASP.NET 16-19
2.3.1 C# Language 17
2.3.2 Web Development in .NET 17-18
2.3.3 Markup Languages and HTML 18-19

3 Identification of Peptides and Phosphorylation Sites 20-24

3.1 Introduction 20
3.2 Site-specific Posttranslational Modification Scoring 20-23
3.3 Results 23
3.4 Conclusion 24

4 PHOSIDA – Phosphorylation Site Database 25-94

4.1 General Process of Knowledge Discovery in Databases 26-27
4.2 Basic Concept of PHOSIDA 27-44
4.2.1 Core Database Management of Phosphorylation Sites 28-34
4.2.1.1 Database Schema adapted for Mining 28-31
4.2.1.2 Online Database Schema 31-32
4.2.1.3 Integration of additional Biological Data 32-34
4.2.2 Kinase Motif Matching 35
4.2.3 Structural Investigation of Phosphoproteomes 35-36
4.2.4 Evolutionary Conservation of Phosphoproteomes 36-38
4.2.5 General Web Application of PHOSIDA 38-43
4.2.6 Administration Tool 43-44

4.3 Data Integration and Data Selection of various Phosphoproteomes 45
4.4 Data Transformation of Preprocessed Data 45-48
4.5 Data Mining in the Compiled Database 49-54
  4.5.1 Statistical Tests 50-51
  4.5.2 Clustering 52-53
  4.5.3 Classification 53-54
4.6 Phosphoproteome Analysis 55-88
  4.6.1 Basic Phosphoproteome Analysis 55-79
    4.6.1.1 *Homo sapiens* 55-64
      4.6.1.1.1 Phosphorylation Dynamics induced by EGF stimulation 55-60
      4.6.1.1.2 Quantitation of the Kinome across the Cell Cycle 60-64
    4.6.1.2 *Mus musculus* 64-69
      4.6.1.2.1 Mouse Liver Phosphoproteome upon Phosphatase Inhibition 64-67
      4.6.1.2.2 Solid Tumor Phosphoproteome 67-69
    4.6.1.3 *Drosophila melanogaster* 69-71
    4.6.1.4 *Saccharomyces cerevisiae* 72-73
    4.6.1.5 Prokaryotic Phosphoproteomes 74-79
  4.6.2 Gene Ontology Analysis 80-82
  4.6.3 Sequence Motif Analysis 82-84
  4.6.4 Structural Constraints on Phosphorylation Sites 85-88
4.7 Discussion 88-94

5 MAPU 2.0: Max-Planck Unified Proteome Database 95-100
5.1 Introduction 95-96
5.2 Implementation of MAPU 2.0 96-100
5.3 Discussion and Future Directions 100

6 SEBIDA – Sex Bias Database 101-107
6.1 Introduction 102-105
6.2 Implementation of SEBIDA 105-106
6.3 Discussion and Future Directions 107

7 Phosphorylation Site Prediction 108-115
7.1 Rationale 108-109
7.2 Implementation on the basis of a Support Vector Machine 109-110
7.3 Results 111-115
7.3.1 *Homo sapiens* specific Phosphosite Predictor 111-112
7.3.2 *Mus musculus* specific Phosphosite Predictor 112
7.3.3 *Drosophila melanogaster* specific Phosphosite Predictor 113
7.3.4 *Saccharomyces cerevisiae* specific Phosphosite Predictor 113
7.3.5 Prokaryotes specific Phosphosite Predictor 113
7.4 Integration of organism-specific Phosphosite Predictors in PHOSIDA 114-115
7.5 Discussion 115

8 Genome Annotation 116-124
8.1 Rationale 116-117
8.2 Mapping Proteomic Data to the Genome 117-121
8.2.1 Assignment of Mass Spectrometry obtained Proteomic Data to Genes annotated in EnsEMBL 117-118
8.2.2 PHOSIDA and MAPU as DAS Sources to annotate the Genome in EnsEMBL 118-120
8.2.3 Representation of Genome Annotations in PHOSIDA and MAPU 2.0 120-121
8.3 Results 121-123
8.4 Discussion 124

9 The Evolution of Phosphorylation 125-144
9.1 Rationale 125-127
9.2 Derivation of phylogenetic relationships and global alignments 127-129
9.3 Results 130-141
9.4 Discussion 142-144

10 Summary and Future Directions 145-147

References 148-155
Acknowledgment 156
Curriculum Vitae 157-159

Chapter 1

Introduction

Cell signalling has arguably become one of the most important aspects of modern biochemistry and cell biology (Gomperts, 2004; Hancock, 2005). The ability of organisms to perceive and correctly respond to their microenvironment is crucial to their survival. The perception of signals such as osmotic strength, pH, oxygen, light, the availability of food, and the presence of predators or competitors for food is fundamental to life. These signals provoke appropriate responses, such as motion away from toxic substances or toward food. In multicellular organisms, cells with various functions process an extensive variety of signals ranging from variations in sunlight to the presence of growth hormones. For animal cells, the interdependent metabolic activities in various tissues or the concentrations of glucose in extracellular fluids, for example, present vital signals that have to be handled. These signals convey information that is detected by receptors and converted to a cellular response. In this context, signal transduction can be defined as the conversion of information into chemical change - a universal property of living cells (Nelson and Cox, 2008).

A relatively small stimulus commonly provokes an avalanche of responses: in typical signal transduction processes the number of participating proteins increases tremendously as the process emanates from the initial stimulus, resulting in a ‘signal cascade’ (Hunter, 2000; Pawson and Nash, 2003). In many cases, the result of a signalling pathway is the posttranslational modification of target-cell proteins that change their activities. Almost all of the more than 200 kinds of posttranslational modifications that occur by covalent addition of groups to side chains are carried out by enzymes, proteins with catalytic activity. Protein phosphorylation may be the most common posttranslational modification, with tens of thousands of phosphorylation sites in the human proteome (Amanchy et al., 2005; Beausoleil et al., 2004; Olsen et al., 2006; Thelemann et al., 2005). At each phosphorylated protein a polar neutral OH side chain is converted to a tetrahedral phosphate (Figure 1.1 left panel). The introduction of negative charges has a notable effect on redistributing conformers in the microenvironment of the protein. These include conversion of unstructured regions of loops into helical regions that can drive and propagate conformational changes to other regions of the modified protein. Such conformational changes can be intramolecular or intermolecular across subunit interfaces and create docking sites for partner proteins with motifs that can specifically recognize the tetrahedral phosphate side chains.

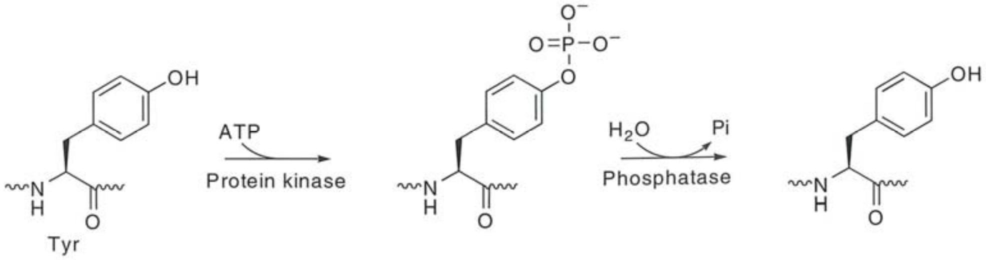


Figure 1.1: Phosphorylation and dephosphorylation processes

The phosphorylation of protein residues (serine, threonine or tyrosin) is catalyzed by protein kinases (left panel). The reaction of dephosphorylation is catalyzed by protein phosphatases (right panel) (Gomperts, 2004).

Thus, intracellular phosphorylation by protein kinases, triggered in response to extracellular signals, provides a mechanism for the cell to switch diverse processes on or off. These processes include metabolic pathways, kinase cascade activation, membrane transport and gene transcription (Schlessinger, 2000).

Two decades ago Hunter estimated that 1000 protein kinases for covalent phosphorylations of proteins are encoded in the human genome (Hunter, 1987). Manning et al. identified 518 putative protein kinase genes, which is about half of what was predicted before, but is still a very large number, constituting about 1.7% of all human genes (Manning et al., 2002b). The substrates of protein kinases in general are the side chains of specific serine, threonine, or tyrosine residues, and specificity depends on structural constraints and on the sequence context surrounding a residue. In eukaryotes, each kinase typically has a number of substrates and is usually either a serine/threonine or tyrosine kinase. However, multiple serine and threonine in a protein substrate may be phosphorylated by a given protein serine/threonine kinase. Analogously, several tyrosines may be phosphorylated by a tyrosine kinase, for instance, on the activation loop of the insulin receptor. A classification of kinases into a hierarchy of groups, families, and subfamilies on the basis of sequence comparisons aided by known biological functions yields a kinome tree (Figure 1.2).

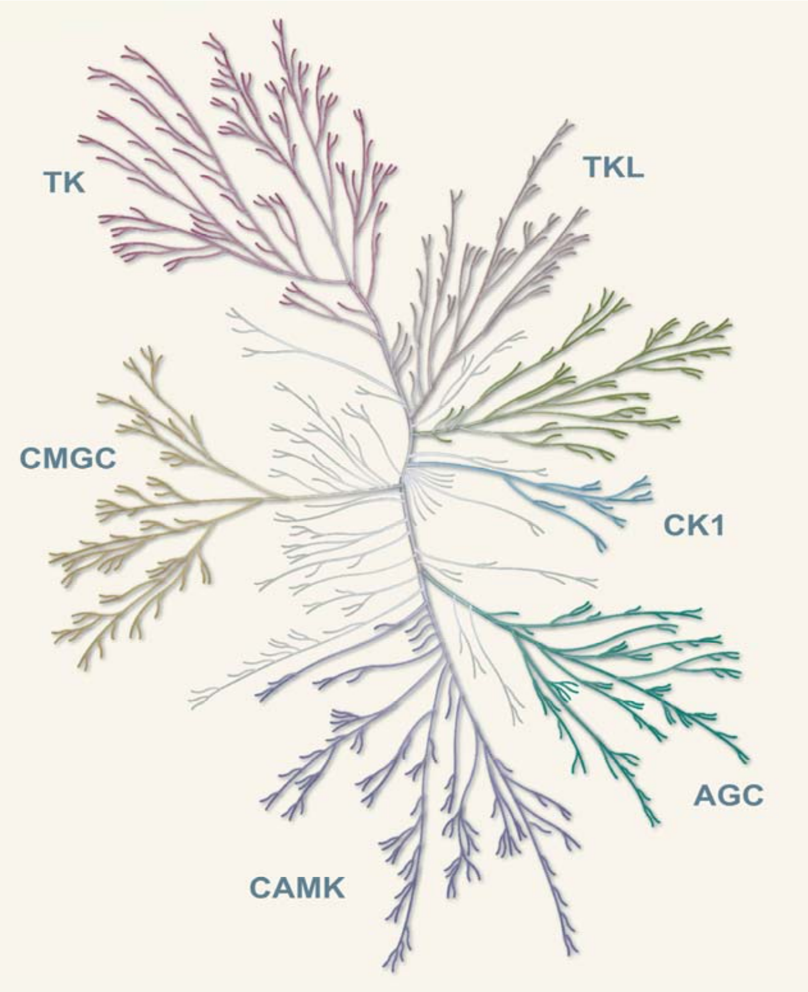


Figure 1.2: Human kinome tree

Manning et al. (Manning et al., 2002b) classified more than 500 identified kinases according to their sequence similarities and common biological functions.

In contrast, the reverse reaction of dephosphorylation (Figure 1.1 right panel) is catalyzed by protein phosphatases that are controlled in response to different stimuli so that phosphorylation and dephosphorylation are separately regulated events. Thus, protein kinase action is balanced by protein phosphatase action.

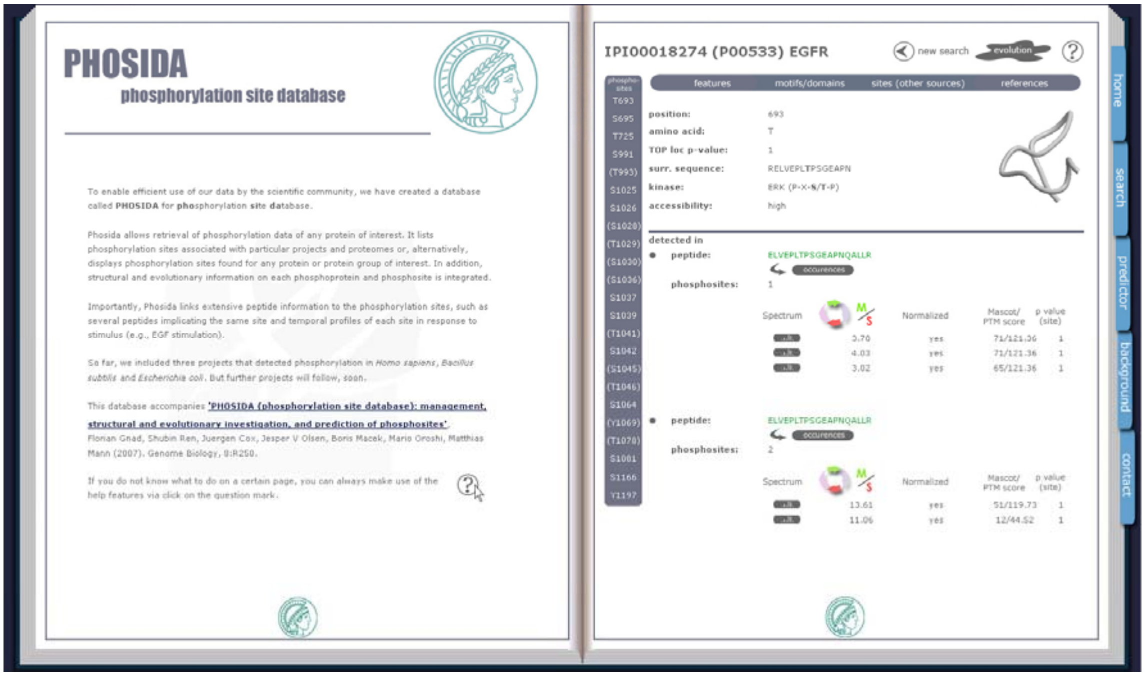
To reveal the role of phosphorylation in the cell at the proteome level, the application of mass spectrometry (MS) based technologies has proven powerful (Aebersold and Mann, 2003; Chen and White, 2004; Ficarro et al., 2002; Mumby and Brekken, 2005; Rush et al., 2005; Salomon et al., 2003). MS-based proteomics has established itself as an indispensable technology to measure proteomes of various organisms along with their phosphorylation changes. By definition, ‘the basic principle of MS is to generate ions from either inorganic or organic compounds by any suitable method, to separate these ions by their mass to charge ratio (m/z) and to detect them qualitatively and quantitatively by their respective m/z and abundance' (Kienitz, 1968). Matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) are the two techniques most commonly used to volatize and ionize the peptides (Aebersold and Mann, 2003). The basic principles and instrumentations are discussed in Chapter 2 in more detail.

The resulting mass spectra are two-dimensional representations of signal intensity versus mass to charge ratio (m/z) (Gross, 2004). The peptide or protein 'precursor' peak results from the detection of the intact ionized molecule, the molecular ion. In a separate reaction inside the mass spectrometer, termed tandem mass spectrometry or MS/MS, the precursor ion is fragmented. In the case of peptides a mass spectrum of these fragment ion peaks can be assigned to corresponding peptide sequences by scanning them against protein sequence databases (see for example (Perkins et al., 1999) for description of the popular Mascot algorithm). There is a large diversity of algorithms to solve this problem, which presents the focus of Chapter 3. However, most of these approaches are restricted to the MS based identification of peptide sequences. To determine posttranslational modifications at the site level, we constructed a probability based algorithm as described in Chapter 3.

In MS based proteomics, the typical outcome is the identification of peptides assigned to proteins. The extensive detection of sub-proteomes and sub-phosphoproteomes of living cells demands description, storage, management and recovery of the obtained data. For this purpose we created PHOSIDA (http://www.phosida.com), the Phosphorylation Site Database (Chapter 4) (Gnad et al., 2007). The aim of PHOSIDA is to comprise high quality phosphoproteomic data including quantitative information, where applicable (for example capturing cell regulation after treatment with a stimulus) (Figure 1.3). To integrate biological context and to mine features of phosphorylation on a proteome-wide scale, PHOSIDA additionally takes into account structures and evolutionary data across a variety of species as well as other protein annotations. Thus, PHOSIDA provides a rich environment to the biologist wishing to analyze phosphorylation events of proteins of interest.

The integrated large-scale datasets contain knowledge, but manual analysis exceeds human capacity. The automated computer based extraction of knowledge from comprehensive datasets is the objective of 'knowledge discovery in databases' (KDD) (Ester, 2000; Witten, 2005). To derive general constraints of phosphorylation relating to structure and conservation, we applied the KDD process to determined large-scale phosphorylation sets. The comprehensive evolutionary study of phosphorylation is explicitly described in Chapter 9.

Furthermore we developed a support vector machine (SVM) based predictor for phosphorylation (Chapter 7) (Gnad et al., 2007). SVMs are machine learning methods used for classification. Two given sets of items such as phosphorylated and non-phosphorylated residues are separated in a multidimensional space, which reflects the features of the given objects. Depending on the relative orientation in the divided feature space, an unclassified item can then be assigned to one of the given two sets. The basic principles of SVMs are discussed in Chapter 7 in more detail. The phosphorylation site predictor is integrated into PHOSIDA and makes it possible to find putative novel phosphorylation sites that have not yet been experimentally identified. Predicting novel phosphosites and matching kinase motifs on proteins of interest should be valuable for the design of biological experiments or for predicting a protein's role in a pathway.



PHOSIDA

phosphorylation site database

To enable efficient use of our data by the scientific community, we have created a database called **PHOSIDA** for **pho**sphorylation **si**te **da**tabase.

Phosida allows retrieval of phosphorylation data of any protein of interest. It lists phosphorylation sites associated with particular projects and proteomes or, alternatively, displays phosphorylation sites found for any protein or protein group of interest. In addition, structural and evolutionary information on each phosphoprotein and phosphosite is integrated.

Importantly, Phosida links extensive peptide information to the phosphorylation sites, such as several peptides implicating the same site and temporal profiles of each site in response to stimulus (e.g., *EGF* stimulation).

So far, we included three projects that detected phosphorylation in *Homo sapiens*, *Bacillus subtilis* and *Escherichia coli*. But further projects will follow, soon.

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

If you do not know what to do on a certain page, you can always make use of the help features via click on the question mark.

IPI00018274 (P00533) EGFR

position: 693
amino acid: T
TOP loc p-value: 1
surr. sequence: RELVEPLTPSGEAPN
kinase: ERK (P-X-S/T-P)
accessibility: high

detected in

- peptide: ELVEPLTPSGEAPNQALLR
  phosphosites: 1

Spectrum	M/S	Normalized	Mascot/ PTM score	p value (site)
	3.70	yes	71/121.36	1
	4.03	yes	71/121.36	1
	3.02	yes	65/121.36	1

- peptide: ELVEPLTPSGEAPNQALLR
  phosphosites: 2

Spectrum	M/S	Normalized	Mascot/ PTM score	p value (site)
	13.61	yes	51/119.73	1
	11.06	yes	12/44.52	1

Figure 1.3: PHOSIDA (Phosphorylation Site Database)

In addition to PHOSIDA, which focuses on the database management of phosphorylation sites, we created MAPU 2.0 (http://mapuproteome.com), the Max-Planck Unified Proteome Database (Chapter 5). The main purpose of MAPU 2.0 is the storage of high throughput datasets of proteomes measured in various tissues, cell types or organellar components on the basis of our high resolution and high accuracy MS technologies. MAPU 2.0 contains several body fluid proteomes including plasma, urine, and cerebrospinal fluid. In addition, cell lines have been mapped to a depth of several thousand proteins and the red blood cell proteome has also been analyzed in depth. By employing high resolution mass spectrometry and stringent validation criteria, false positive identification rates in MAPU 2.0 are always lower than 1:100 and usually lower than 1:1000. Thus, MAPU 2.0 datasets can serve as high quality reference proteomes, for example in biomarker discovery.

Another objective of this work was the annotation of genomes on the basis of MS derived proteomic data. As mentioned above, MS is commonly applied to the identification of proteins by matching the measured spectra to sequences of known proteins that are annotated in public databases. Hence, this approach is limited to the detection of already predicted or established polypeptides. However, the original resource is the genome (Lander et al., 2001; Venter et al., 2001). It encodes all possible proteins and therefore represents the original source of the proteome. But the derivation of coding regions on the nucleotide sequence is not trivial. Current methods for gene prediction provide useful information but are still limited (Brent, 2007). It is hardly possible to predict all features of the genome from its sequence alone. Thus, the integration and validation of MS derived experimental data in a genomic context may contribute to the annotation of the genome and the identification of genes that have not been experimentally confirmed yet (Chapter 8) (Desiere et al., 2005; Fermin et al., 2006). The main idea is to assign the measured spectra to translated predicted genes or even to all potential open reading frames instead of already known proteins. In this work we assigned our proteomic data directly to genes and then we linked our proteome databases with the genome database EnsEMBL via the DAS/Proserver technology (Birney et al., 2004; Finn et al., 2007; Flicek et al., 2008).

Although not directly associated with the main topic of my PhD study, a further goal was the further development and curation of SEBIDA (www.sebida.com) – the Sex Bias Database (Chapter 6) (Gnad and Parsch, 2006). The database integrates results from multiple, independent microarray studies comparing male and female gene expression in *Drosophila melanogaster*, *Drosophila simulans* and *Anopheles gambiae*. In addition to ratios of male/female expression for each gene, SEBIDA also contains information useful for evolutionary studies, such as degree of codon bias, local recombination rates and interspecific divergence at synonymous and non-synonymous sites. Our laboratory is currently working on the quantitative evaluation of sex biased proteins on the basis of MS. This proteomic study has not been finished yet. However, we intend to analyse sex specific protein expression levels using the established SEBIDA environment in the future.

Thus, a variety of topics have been subjects of my PhD study in addition to the main focus on the bioinformatics of phosphorylation. They are tightly linked, since my study ranges from the identification of phosphorylation sites (Chapter 3) to their database storage (Chapter 4) along with other proteomic data (Chapter 5). On the basis of the created databases, which are accessible to the public community, we derived various general patterns (knowledge) (Chapter 4) with a main focus on the evolution of phosphorylation. Thus, the analysis of evolutionary constraints of phosphorylation is described in more detail in Chapter 9. The above mentioned phosphorylation site predictor that is trained on our high throughput datasets to recognize potential phosphosites mainly on the basis of features such as the surrounding sequence is described in Chapter 7. This overal workflow presents a 'Knowledge Discovery in Databases' (KDD) process as described in Chapter 4.

For the mapping of proteomic data to the genome database EnsEMBL (Chapter 8), I received a Marie Curie Fellowship and worked at the European Bioinformatics Institute (EBI) in Cambridge. My adviser at EBI was Ewan Birney, founder of the EnsEMBL database.

Chapter 2

Background: Mass Spectrometry, Database Systems and ASP.NET

2.1 Mass Spectrometry based Proteomics

Proteomics is a relatively new ‘post-genomic’ science that focuses on the large scale determination of the functional network in the cell at the protein level. It is a multifaceted field of research including a collection of various technical disciplines ranging from the experimental identification of amino acid sequences to their database storage.

Historically, protein purification was based on crude chromatographic and then on gel electrophoresis methods. In one-dimensional gel electrophoresis proteins are separated so that all proteins lie along a lane but are separated by molecular weight. In two dimensional electrophoreses (Gorg et al., 2004; Gygi et al., 2000; Rabilloud, 2002), the proteins are first separated by isoelectric point and then by molecular weight. Although this technology proved to work sufficiently well for the analysis of low complexity protein mixtures, it could not satisfy the requirements for large scale in depth proteome analysis at current requisite quality standards (Mann and Kelleher, 2008). Of all contributing disciplines, MS has established itself as the main technology of proteomics studies.

The development of two techniques – electrospray ionization (ESI) and matrix assisted laser desorption/ionization (MALDI) – in the late 1980s made essential contributions to the establishment of the rapidly evolving field of MS-based proteomics (Fenn et al., 1989; Karas and Hillenkamp, 1988). The development of these two ionization techniques encouraged the development of other decisive technologies including new mass analyses and complex multistage instruments designed to tackle the challenges of proteome analysis. In fact, it is amazing how rapidly MS has developed over the past decade. Ten years ago, the sequencing of a single protein was a remarkable achievement. Today, the determination of thousands of proteins in a single experiment is common practice. The lag between genomics, which already demonstrated the power of high-throughput analysis of biological processes, and proteomics is rapidly diminishing (Cox and Mann, 2007). MS can sequence tens of thousands of peptides from complex mixtures. Moreover, the application of quantitative proteomics using technologies such as SILAC (Stable Isotope Labeling by Amino Acids in Cell Culture), even allows to compare the relative protein abundance between different proteomes (Ong et al., 2002; Ong and Mann, 2005). By applying quantitative proteomics, functional information and temporal changes in the proteome including posttranslational modification dynamics can be captured by MS.

Although various disciplines comprising different technologies contribute to proteomics, the design of MS-based proteomics experiments is quite generic (Figure 2.1) (Aebersold and Mann, 2003):

First, the proteins to be investigated are obtained from cell lysates by affinity selection or biochemical fractionation. Sample fractionation oftentimes includes the separation into several subproteomes using gel electrophoresis.

Then proteins are degraded enzymatically to peptides. The degradation step is required, as mass spectrometry of peptides is more sensitive than mass spectrometry of proteins, where the mere entire mass is not sufficient for identification. Trypsin digestion has proven to be an especially appropriate degradation method because it yields peptides with C-terminally protonated amino acids (Arg or Lys), which fragment well in tandem MS.

Next, MS measurements are carried out in the gas phase on ionized peptides. Peptides to be analyzed are passed on to the three main components of the mass spectrometer: the ion source, the mass analyser that measures mass-to-charge (m/z) ratios of the ionized peptides and the detector that counts the number of ions at each m/z value.

Consequently, the initial step for the identification of the peptide using a mass spectrometer is the ionization in an ion source: as mentioned above, the MALDI and ESI are the two most widespread ionization technologies and have had a huge impact on the rapid development of mass spectrometry (Fenn et al., 1989; Finn et al., 2007; Karas and Hillenkamp, 1988). MALDI sublimates and ionizes the peptides out of a crystalline matrix via laser pulses, whereas ESI ionizes the peptides out of a solution. Peptides are usually separated by liquid based separation techniques such as high-pressure liquid chromatography in very fine capillaries. After electrospray ionization the multiply protonated peptides enter the mass spectrometer, where the mass analyzer presents the essential component. There are four basic types of mass analysers, namely the ion trap, time-of-flight (TOF), quadrupole and ion cyclotron resonance (ICR) instruments (Hager and Le Blanc, 2003; Marshall et al., 1998; Martin et al., 2000; Schwartz et al., 2002; Valaskovic et al., 1996). They differ in mass accuracy, resolution and sensitivity. In each case, a mass spectrum of the peptides is taken (MS^1 spectrum). A mass spectrum is the two-dimensional representation of signal intensity versus m/z (Chapter 1). Then the computer generates a list of peptides for further fragmentation. Specified ionized peptides are isolated and fragmented by collision with an inert gas at low pressure, so that a tandem (MS^2) spectrum is obtained.

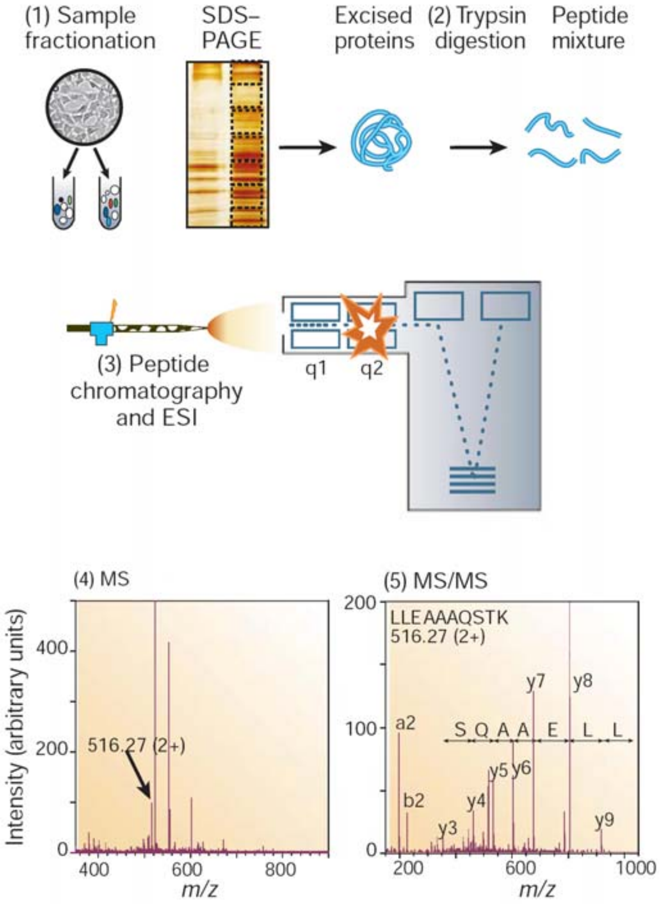


Figure 2.1: Generic mass spectrometry based proteomics approach (Aebersold and Mann, 2003)

This generic MS-based proteomics method is then followed by computational analyses as described in Chapter 4.1: The MS^1 and MS^2 spectra are matched against protein sequence databases. We use the Mascot search algorithm to match given spectra with peptide sequences (Perkins et al., 1999). The final outcome is the identity of peptides assigned to proteins. As highlighted in Chapter 3, we extended the algorithm by another probability based method that determines posttranslational modifications within specified peptide sequences at the site level. The validated results are then uploaded to a database. After transforming the integrated data for the application of computational analyses, data mining methods are then applied to derive patterns (knowledge) from the data in a KDD process (Han, 2000; Witten, 1999, 2005) (Ester, 2000).

The goal of quantitative proteomics is to determine the relative changes in expression of proteins (Ong and Mann, 2005). Translational controls and regulated degradation contribute to the biological function of proteins in addition to the regulation of the transcriptional machinery. To understand the functional impact of proteins, it is therefore indispensable to measure changes of protein expression levels in a whole biological system. Even though MS is not inherently quantitative, many techniques have been developed that supply the quantitative dimension to MS. For instance, Mann and colleagues have established a stable isotope-based technique termed stable isotope labeling by amino acids in cell culture (SILAC) (Ong et al., 2002). Cell populations grow in different metabolically labelled media (Figure 2.2): one in a medium that contains a normal ('light') amino acid and the other in a medium that contains a heavy amino acid. The heavy amino acid can contain ^{13}C instead of ^{12}C, for example. Consequently, the two proteomes can be distinguished, as each peptide appears in two forms separated by the difference between light and heavy label. The intensity difference of the two forms reflects the difference in protein amount between the two cell populations. This method makes it possible to measure protein expression changes including phosphorylation dynamics after various treatments over time. Another application is the system-wide measurement of proteome expression differences between a normal cell and a cancer cell. As illustrated in Figure 2.2, SILAC experiments can even be extended by a third label ('medium').

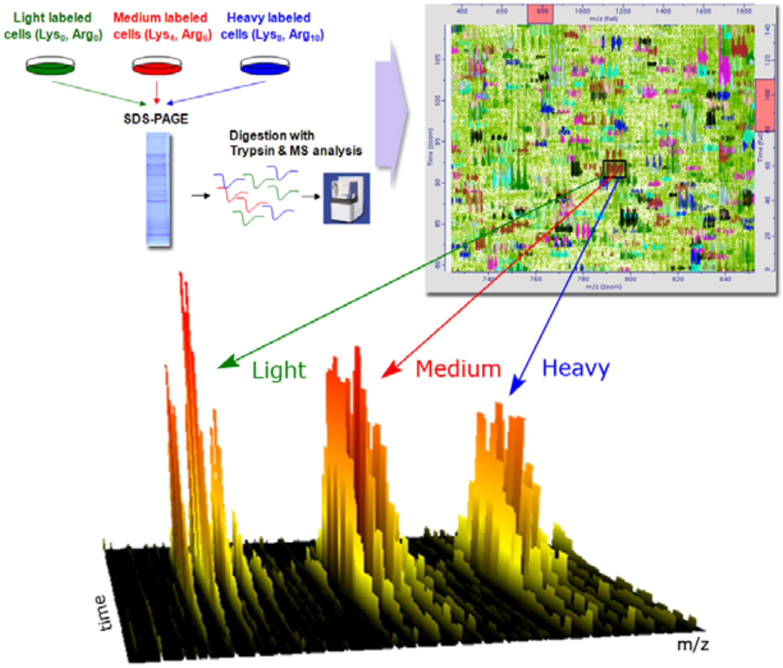

Light labeled cells (Lys$_0$, Arg$_0$)
Medium labeled cells (Lys$_4$, Arg$_6$)
Heavy labeled cells (Lys$_8$, Arg$_{10}$)
SDS-PAGE
Digestion with Trypsin & MS analysis
Light
Medium
Heavy
time
m/z

Figure 2.2: SILAC based proteome measurements (Cox and Mann, 2007)

2.2 Database Systems

The vast increase in new technologies in biology ranging from genome sequencing to mass spectrometry has led to an explosion of the amount of data. This data demands efficient description, storage, management and recovery and efficient mining to facilitate extraction of biological knowledge.

2.2.1 Components and Functions of Database Systems

The term 'database' (DB) is defined as a collection of logically linked data. 'Database Management Systems' (DBMS) are software modules designed to manage the entire database (Date, 2003; Ramakrishnan, 2003). Therefore the main function of DBMSs is to describe, store, and regain very large amounts of data. Its hierarchical layer architecture fulfils these basic functions. Another task is the separate management of transactions and metadata. In addition, an important purpose of DBMSs is to interact with external applications in two directions. On the one hand, queries have to be worked on by the conversion of descriptive statements into procedural operations (user → DB). On the other hand, data have to be presented query-dependently (DB → user).

The DB and its DBMS constitutes a 'database system' (DBS) (Figure 2.3).

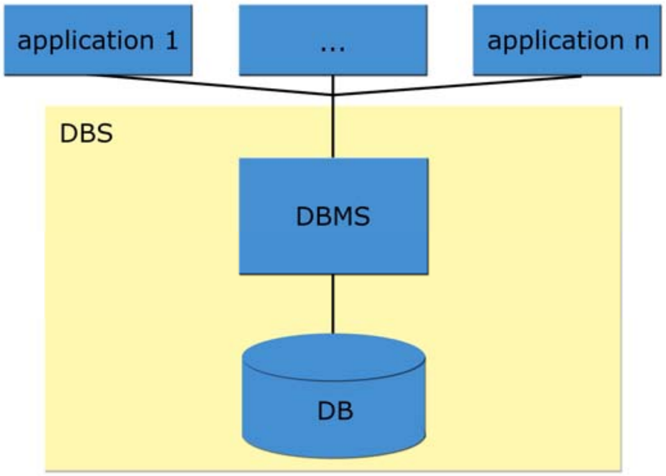


Figure 2.3: Relationship between database management system, database, and application layer

The online database GenBank exemplifies the importance of database systems (Benson et al., 2008). GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. It contains nearly 40 billion bases in about 40 million sequences (Figure 2.4a). Questions about the data must be answered efficiently, changes made to the data by different users must be applied consistently, and access to certain parts of the data must be restricted.

One could try to manage the data by storing it in operating file systems, but this approach has many drawbacks. A database such as GenBank would have to write extra programs to answer each question a user may want to ask about the data. These programs would be complex because of the large volume of data to be searched. Furthermore, databases have to protect data from inconsistent changes made by different users. There are many disadvantages of using file systems which even make databases indispensable. DBMSs manage the data in a robust and efficient manner. As the volume of data and the number of users grow, support by DBMSs becomes indispensable. Concrete advantages of DBS are reflected by the demands on a database system, the so called 'Codd's Rules' (Figure 2.4b) (Begg, 2004). Cood's rules demand requisite features of the databases including consistency, recovery, controlled accession, transactions and operations on the data. In addition, the main benefits of databases are physical and logical independencies (Chapter 2.2.2). These can be derived from the architecture of database systems.

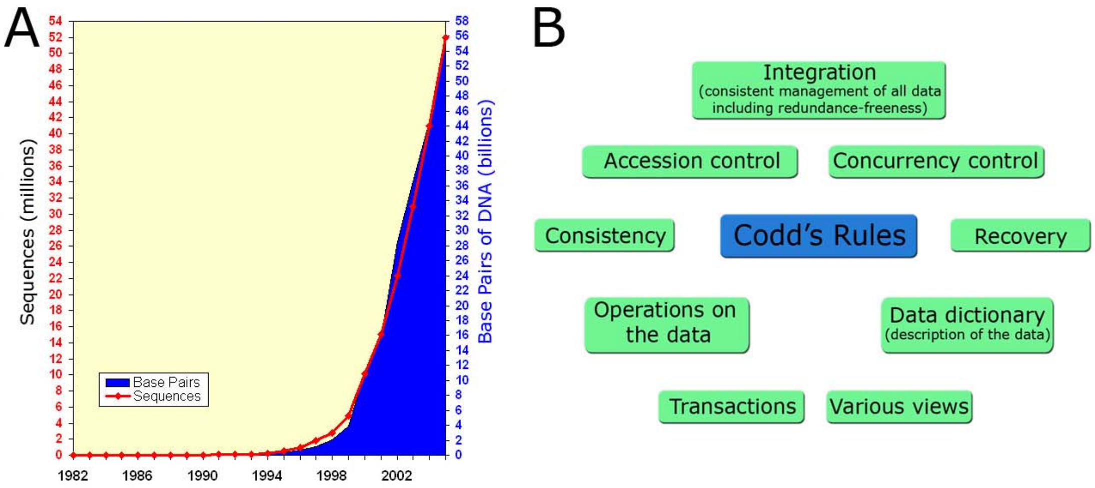


Figure 2.4: Data growth of GenBank (a) and the Codd's rules (b)

The exponential rise in GenBank data is indicative of the boost of biological data in general. Like all databases, databases managing biological data have to satisfy the Codd's Rules.

2.2.2 Architecture of Database Systems

A database system is divided into three separate tiers (Figure 2.5) (Date, 2003; Ramakrishnan, 2003). The internal view determines the physical storage of data. Its specification is dependent on the available system and it is manipulated by the Data Storage Description Language (DSDL). The conceptual tier is defined as the logical entirety of all data, whereas the Data Definition Language (DDL) devises the entire schema. Finally the external view is the collection of all application specific views. Its tools are the Data Manipulation Language (DML) and the Data Query Language (DQL). They give instructions to read and manipulate the data.

The main advantages of databases in comparison to file systems are independence between the conceptual and the internal tier (physical independence) on the one hand, and independence between the conceptual and the external tier (logical independence) on the other hand. For example, if the user asks for a sequence of a certain gene, whose descriptions and further information are stored in a database, the application does not have to care about the conceptual schema of the database. Thus, the logical independence is also called application independence. Moreover, the conceptual schema of the database is independent of its physical storage. Consequently, it is also known as implementation independence.

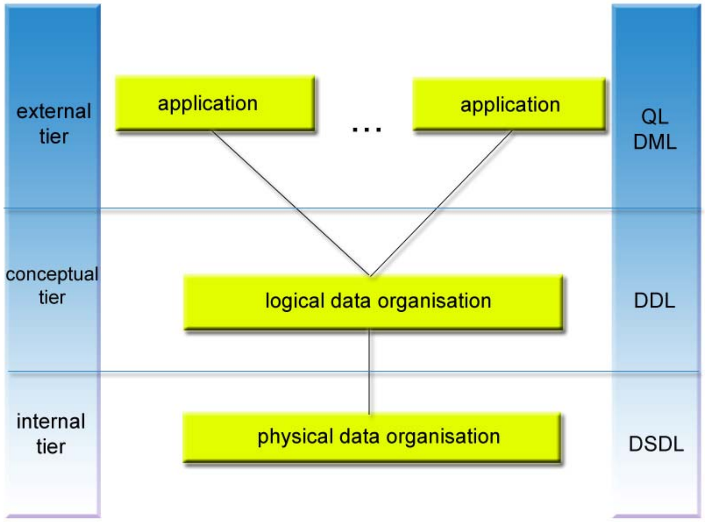


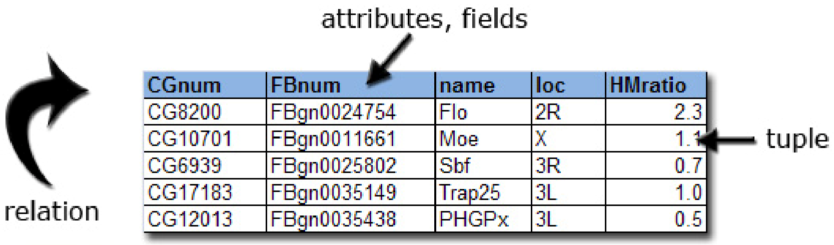
Figure 2.5: Architecture of a database system

2.2.3 Relational Model

There are many data models such as the hierarchical, object-oriented or the network model. The foremost one is the relational model, which is commonly used and widely spread (Heuer, 2000). The main construct for representing data in the relational model is the 'relation'. Its schema specifies the name of the relation, the name of each column (attribute) and the set of associated values for each attribute.

An instance of a relation is a set of tuples. They are also called 'records'. Each tuple has the same number of fields as the relation schema. A relational database can thus be defined as the collection of relations with distinct relation names.

One essential element for a relation is the primary key. It is defined as the minimal set of attributes identifying each tuple uniquely. Besides the primary key, a foreign key is the minimal set of attributes which refer to a primary key of a 'foreign' relation. Thus, various relations within a database have precisely defined relationships. Entity Relationship Models (ER models) are often used in order to describe the conceptual database scheme including various (one-to-one, one-to-many, or many-to-many) relationships between different relations. The definition of keys is associated with functional dependencies. They play important roles in the conceptual construction of a database. An example of a relation containing expression information of genes is illustrated in Figure 2.6.


attributes, fields
relation
tuple

CGnum	FBnum	name	loc	HMratio
CG8200	FBgn0024754	Flo	2R	2.3
CG10701	FBgn0011661	Moe	X	1.1
CG6939	FBgn0025802	Sbf	3R	0.7
CG17183	FBgn0035149	Trap25	3L	1.0
CG12013	FBgn0035438	PHGPx	3L	0.5

Figure 2.6: Instance of a database relation

The illustrated example of a relation contains data such as unique gene identifiers (CGnumber, FlyBase number or gene name), chromosome locations, and abundance ratios of different SILAC labels.

2.2.3 Query Language SQL

The Structured Query Language (SQL) is the most widely used relational database language (Gennick, 2006). It enables programmers to pose complex queries on datasets. It is based on relational algebra. Hence SQL is able to capture all possible relational expressions; it is relationally complete. Without the application of database systems along with a query language the analysis of large data sets such as those derived through MS-based proteomics would be inefficient. A programmer would have to write ad hoc programs for each query on data that are stored in file systems. Databases instead allow formulating formulating short statements on the data.

We used the open source database query language MySQL (http://www.mysql.com) (Reese, 2002), in order to extract information and knowledge from proteomic data (Dzeroski and Lavrac, 2007). The only disadvantage of MySQL that we experienced is the absence of a direct implementation of the frequently used 'outer join' operation. To design an 'outer join' operation on two or more tables, it is necessary to formulate a workaround by combining a left outer join operation and a right outer join operation via the 'union' operation. Except for this disadvantage, MySQL proved to be the proper tool for data queries on a relational model for very large and complex proteomics data.

2.3 Web Development in ASP.NET

One of the most recently established object-oriented programming languages is C# (Chapter 2.3.1) (Liberty, 2005a). It was designed to program the Microsoft .NET Framework (Liberty, 2005b), which is briefly described in Chapter 2.3.2. We decided to use C# because of its applicability to the Windows based Xcalibur^{TM} software that provides instrument control and data access for the entire family of mass spectrometers of the Thermo Fisher Scientific company, which are used exclusively in our group. As C# is a relatively new programming language, it has not generally been used in bioinformatics yet. Consequently, there are virtually no open access class libraries that can be shared by the public community.

However, C# provides an optimum blend of performance, simplicity and expressiveness on the basis of observations drawn from other languages such as Java and C++. It comprises all advantages of object-oriented programming and makes it possible to share self-defined classes and methods via class libraries.

Regarding web programming, C# presents the underlying language of ASP.NET, which enables programmers to encapsulate code into web controls ranging from simple HTML buttons to complex list boxes. Since the implemented dynamic web sites rely on a database to provide content, we used the ADO.NET technology (Chapter 2.3.1) to embed data retrieved from a mySQL database into dynamically created web content (Hamilton, 2003). Finally, retrieved data are dynamically represented in a structured document (web page). Its representation and design is subject to the discipline of Markup languages, namely HTML for web representations (Chapter 2.3.3) (Goodman, 2006; Musciano, 2006).

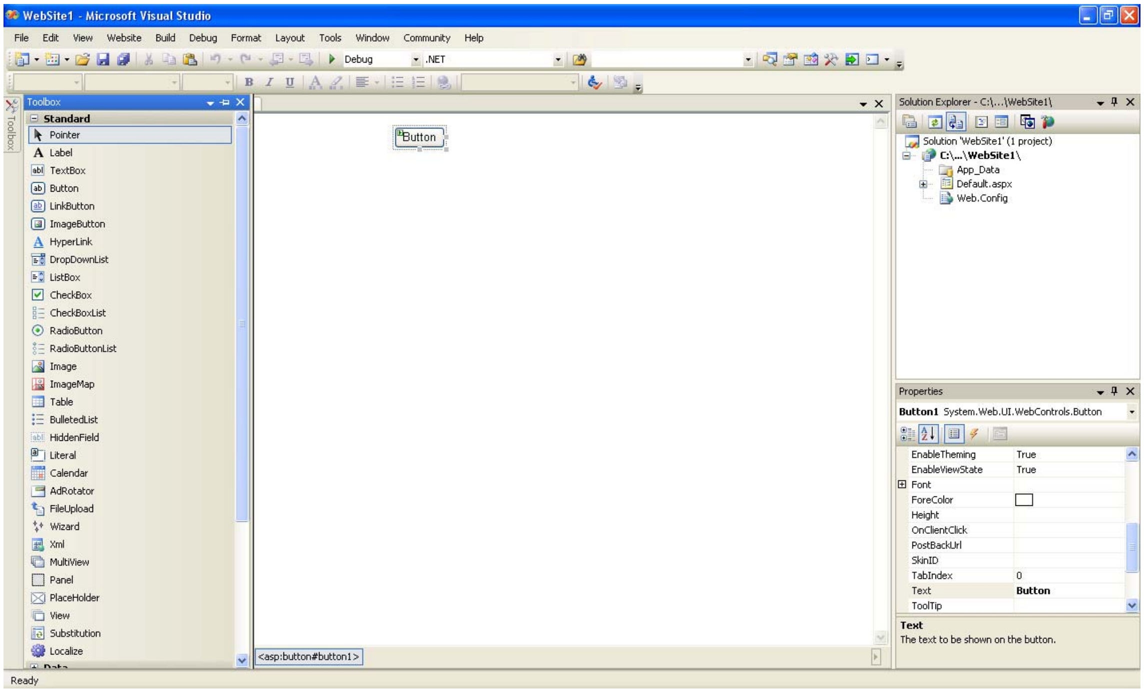
2.3.1 C# Language

The goal of the programming language C# is to provide a simple, safe, object-oriented, high-performance language for .NET development (Liberty, 2005a). C# is a very modern language, and it draws on the lessons learned over the past decades. Experienced programmers can immediately see the influence of already established languages, primarily C++, Java and Visual Basic. C# can be ideally used as a tool for programming on the .NET platform (Liberty, 2005b), especially with Visual Studio (Griffiths, 2003). As a component-based, structured, object-oriented programming language, it includes all the support for defining and working with classes. It contains keywords for defining new classes along with their properties and methods. Furthermore, it allows the implementation of the three essential requirements of object-oriented programming: encapsulation, inheritance and polymorphism. The final compilation of programming code yields a collection of files that appear to be a single executable or a single dynamic link library (DLL). These compiled files are named ‘assemblys’ and present the basic units of deployment and reuse in .NET.

In summary, C# is a very powerful programming language comprising all the strengths of object-oriented programming. It is designed for developing applications on Microsoft’s .NET platform and provides a unique solution to write dynamic web applications.

2.3.2 Web Development in .NET

‘ASP.NET is an event-driven, control-based, object-oriented architecture that generates content and dynamic client-side code from server-side code using functionality described in the *System.Web* classes of the .NET Framework’ (Cazzulino, 2004; Liberty, 2005b). This means that ASP.NET is the technology that performs server-side processing to generate the page response when receiving a web page request. After the execution of server-side code ASP.NET sends back the created web page to the browser. The event-driven feature handles the reaction to events such as when a user clicks a button. This requires the usage of elements of visual functionality known as ‘server controls’. Server controls comprise web elements such as buttons or listboxes. In principle, one can configure server controls through a Properties browser (Figure 2.7). At runtime, ASP.NET transforms the configured server controls into plain HTML code that is sent to the requesting browser. However, the design of more complex web pages such as PHOSIDA (Chapter 4) still requires the implementation of HTML code. Nevertheless, the integration of server controls presents a very strong foundation, as elements of visual functionality conform to the .NET programming model.



EnableTheming	True
EnableViewState	True
Font	
ForeColor	
Height	
OnClientClick	
PostBackUrl	
SkinID	
TabIndex	0
Text	**Button**
ToolTip	

Text
The text to be shown on the button.

Figure 2.7: Microsoft Visual Studio environment.

Server controls can be easily placed via the Toolbox (on the left) and configured via the Properties Browser (on the right).

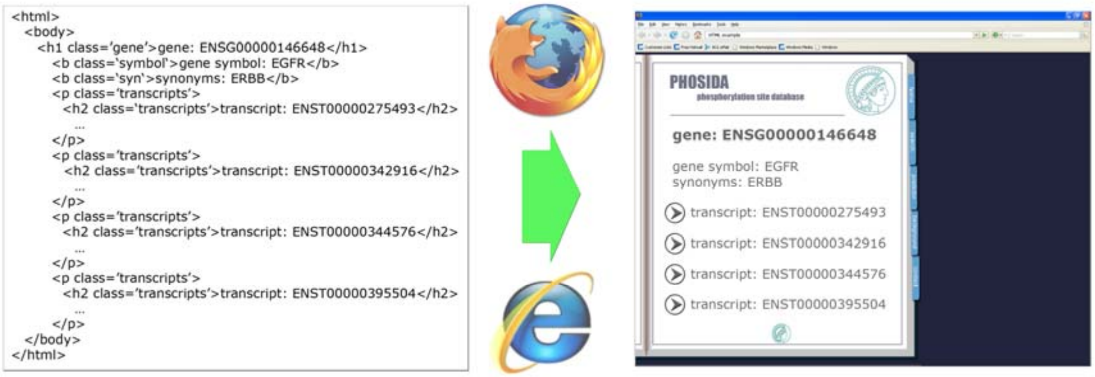
The functionality of web elements such as server controls and web forms is contained within the *System.Web* namespace. It includes a comprehensive set of ASP.NET Framework classes that enables web programmers to design multiply functional web pages in a sophisticated way. In addition, ASP.NET brings all the advantages of object-oriented programming, as all classes and methods are extensible and reusable through inheritance and polymorphism.

2.3.3 Markup Languages and HTML

Each document presents an organized set of data. This PhD thesis is also an ordered set of headings, paragraphs, and illustrations. The data in documents are arranged visually in such a way that the organization of the data is clear. This makes it easier to read the document. Analogously, we often need our computerized applications to be able to read a document and derive the structure of the data contained in it. To do this, we use ‘markup’.

Markup consists of tags that occur in the document along with the data. They specify the various elements of data within the document. All the data corresponding to an element are arranged between the opening <*element*> tag and the closing </*element*> tag. Moreover, one element is likely to embrace other elements along with their data.

Hypertext Markup Language (HTML) is generated by web applications and sent to the browser for display (Musciano, 2006). In fact, HTML is a ‘markup language’. An HTML document is a set of tags and data that allows the description of the structure of web page documents. The main purpose of the data of an HTML document is to display information in a browser window. Thus, the markup in an HTML document is intended to describe the way the browser should display the data. Figure 2.8 exemplifies an HTML document that describes general features of the EGF receptor gene.



```
<html>
  <body>
    <h1 class=’gene’>gene: ENSG00000146648</h1>
      <b class=‘symbol’>gene symbol: EGFR</b>
      <b class=‘syn’>synonyms: ERBB</b>
      <p class=’transcripts’>
        <h2 class=’transcripts’>transcript: ENST00000275493</h2>
          ...
      </p>
      <p class=’transcripts’>
        <h2 class=’transcripts’>transcript: ENST00000342916</h2>
          ...
      </p>
      <p class=’transcripts’>
        <h2 class=’transcripts’>transcript: ENST00000344576</h2>
          ...
      </p>
      <p class=’transcripts’>
        <h2 class=’transcripts’>transcript: ENST00000395504</h2>
          ...
      </p>
  </body>
</html>
```

PHOSIDA
phosphorylation site database
gene: ENSG00000146648
gene symbol: EGFR
synonyms: ERBB
transcript: ENST00000275493
transcript: ENST00000342916
transcript: ENST00000344576
transcript: ENST00000395504

Figure 2.8: Example of the web presentation of a given HTML document describing features of the EGF receptor gene (as interpreted by common web browsers such as the Internet Explorer or Mozilla Firefox)

The illustrated HTML document describes general features of the gene that encodes the epidermal growth factor precursor protein such as gene symbol and synonyms. In addition, it describes four different gene transcripts. This HTML document can be sent to a browser. As the browser is programmed to interpret tags, it is able to parse the logical structure of the document. Tags such as ‘<html>’ and ‘<b>’ are common elements that are uniformly handled by a variety of web browsers. In order to specify the display of data, Cascading Style Sheets (CSS) are used to describe a particular presentation of a document (Meyer, 2006). The application of CSS limits the scope of the web browser’s interpretation and enables the web programmer to force the browser to display the data in a defined way. In the example, the additions of tag classes refer to certain styles relating to colors, layouts, and fonts. The purpose of this chapter is not to dwell on the detailed concepts of HTML documents and Cascading Style Sheets. However, it should become obvious that the creation of dynamic user-friendly web pages is a result of the combination of the embedding of ASP server controls into HTML documents whose layouts are specified by CSS.

Chapter 3

Identification of Peptides and Phosphorylation Sites

3.1 Introduction

Many approaches and algorithms have been described in the literature for peptide and protein identification by searching a sequence database using MS data (Sadygov et al., 2004). Although reported methods differ in their detailed implementation, the general concept is similar: The experimental data are compared with peptide and peptide fragment mass values calculated on the basis of cleavage rules applied to the protein sequences in the specified database.

To assign measured spectra to peptide sequence, we use the search engine Mascot (Perkins et al., 1999), which is based on probability scoring. Mascot is a well established software used in many MS laboratories for protein identification by searching sequence databases. It is primarily optimized for the identification of sequence stretches (peptides) based on the presence of calculated fragment ions in the tandem spectra. Identified peptide sequences are assigned to protein entries afterwards. However, proper site specific location of posttranslational modifications is not a strength of Mascot but it is critical, as many biological processes are regulated through the modification of specific residues.

Hence, we established a probability based algorithm that measures the probability of correct phosphorylation site localization. We applied our method in a fully automated fashion via the PHOSIDA upload system (Chapter 4.2) enabling us to investigate identified phosphoproteomes on the site level.

3.2 Site-specific Posttranslational Modification Scoring

The post-translational modification (PTM) score used for localization of the phosphorylation sites is an extension of the MS^3 score described by Olsen and Mann (Olsen and Mann, 2004), and was described in Olsen et al. (Olsen et al., 2006). The binomial distribution score is used to compute the probability for all individual serine, threonine and tyrosine residues to be phosphorylated in a phosphopeptide identified by MASCOT.

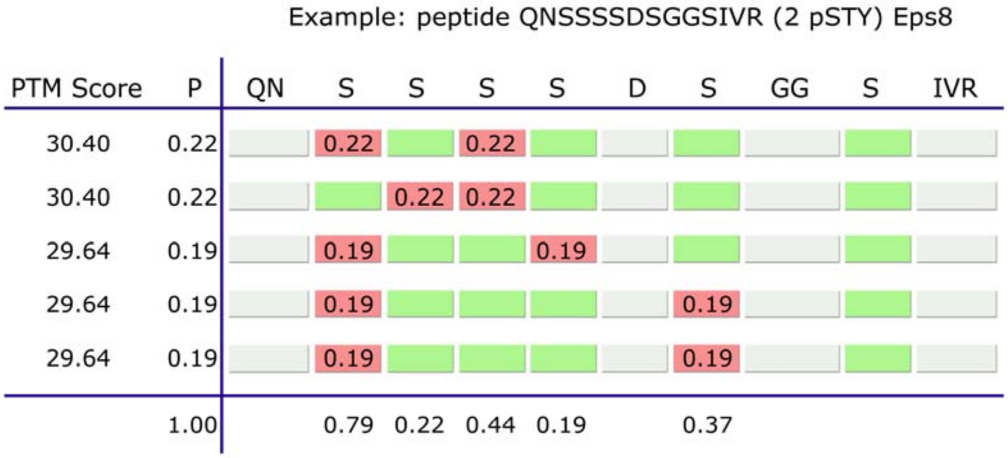
In an ion trap MS/MS spectrum (e.g. from LTQ Orbitrap or LTQ-FT instruments) fragments are matched with a mass tolerance of +/- 0.5 Da. As a result, one fragment ion can be matched per m/z unit throughout the mass range and there are 100 'bins' for the fragments per hundred m/z interval. To compute the binomial distribution score, the top most intense fragment ions per 100 m/z bins in a spectrum are considered. The algorithm automatically discards most of the ions and keeps only the top four most intense one per 100 m/z units, which therefore have 4% chance (0.04) of matching randomly (Andersen et al., 2003). For a true match, the most intense fragment ions are expected to match the peptide sequence-specific b- and y-type ions.
The binomial distribution score probability (P) is calculated as:

$$
\begin{aligned}
P(k) &= \binom{n}{k} \cdot p^{k} \cdot (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!} \cdot p^{k} \cdot (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!} \cdot 0.04^{k} \cdot (0.96)^{n-k}
\end{aligned}
$$

where n over k is the number of permutations of a subset of k elements (matches) in a set of n elements (total number of possible b and y ions in the mass range). The probability of a putative b- or y-ion to match one of the experimental fragment masses by chance is simply 4/100 or 0.04, independently of the mass range considered, because we allow four measured masses per 100 Da. For some applications, six instead of four peaks per 100 m/z interval are retained.
To make the PTM score comparable to the probability-based MASCOT score, we compute the Post-Translational Modifiation (PTM) score in the same way:

$$\text{PTM Score} = -10 \cdot \log_{10}(P(k))$$

The algorithm calculates the PTM scores for all possible phosphorylation site combinations within a given phosphopeptide sequence by successively placing the number of phospho groups (known from the measured peptide molecular weight) on each serine, threonine or tyrosine in turn. To calculate the probability of phosphorylation for all candidate sites, all phosphorylation site combinations showing a PTM score higher than the maximum score minus five are taken into account. The value of five was chosen on an empirical basis as it turned out to retain most of the possible phosphorylation sites in the peptide. For each candidate combination i with a PTM score PTM_i, the corresponding probability p_i is given to all assigned phosphorylation sites. Subsequently, the p value for the phosphorylation probability of each candidate site is calculated as the sum of probabilities p_i of all candidate phosphopeptides and normalized, so that the sum of all resulting site-specific localization probabilities is equal to the number of phosphorylation events in the specified peptide. Table 3.1 gives an example of phosphorylation site combinations along with the resulting localization probabilities of each candidate site.



Example: peptide QNSSSSDSGGSIVR (2 pSTY) Eps8

PTM Score	P	QN	S	S	S	S	D	S	GG	S	IVR
30.40	0.22		0.22	🟩	0.22	🟩		🟩		🟩	
30.40	0.22		🟩	0.22	0.22	🟩		🟩		🟩	
29.64	0.19		0.19	🟩	🟩	0.19		🟩		🟩	
29.64	0.19		0.19	🟩	🟩	🟩		0.19		🟩	
29.64	0.19		0.19	🟩	🟩	🟩		0.19		🟩	
	1.00		0.79	0.22	0.44	0.19		0.37			

Table 3.1: Derivation of localization probabilities of candidate phosphorylation sites in a given phosphorylated peptide

The table shows a specific example of the calculation of site-specific localization probability values (doubly phosphorylated peptide of Eps8). The five top scoring possibilities for phosphorylation have PTM scores from 30.4 to 29.64. Corresponding probabilities (P) reflect the proportional probability for each phosphorylation site combination and add up to one. They are assigned to the two phosphorylation sites in each case. Next, probabilities are summed up for each candidate site.

To deduce the exact localization of phosphorylation events within a given phosphopeptide along with the corresponding probabilities from the given spectrum, the algorithm was embedded into the PHOSIDA upload system. It was first applied to a large-scale study in which we investigated the phosphoproteome in human cells exposed to EGF stimulation (Chapter 4.6.1.1.1) (Olsen et al., 2006). For the first time, we were able to identify the phosphoproteome in a site-specific way without manual derivation of the exact position of phosphorylation sites. To test the algorithm for a defined set of phosphopeptides with known phosphorylation sites, we analyzed synthetic phosphopeptides available in our laboratory and phosphopeptides derived from tryptic digests of bovine caseins by LC-MS on the LTQ-Orbitrap. The phosphorylation sites on caseins are highly validated in the literature and were taken from Thingholm et al. (Thingholm et al., 2008). All phosphopeptides harbored at least one non-phosphorylated residue. We first calculated the PTM scores for all phosphopeptide spectra and subsequently determined phosphorylation site probabilities. In addition, we tested the PTM scoring on the dataset from a previous large-scale study on the phosphoproteome identified in pheromone treated yeast cells. Gruhler et al. identified 700 phosphopeptides, for which phosphorylated residues were manually assigned (Gruhler et al., 2005).

3.3 Results

In the case of synthetic peptides, 27 out of 37 phosphorylated residues were correctly assigned with a p value of 1, with no false positive assignment. For all phosphorylation sites with a probability value of 0.75 or higher, which we define as class I sites, precision was still 94%. Figure 3.1a presents the corresponding precision-recall curve. ’Recall’ is the proportion of true positives to the sum of true positives and false negatives, whereas ’precision’ describes the number of true positives out of all predicted positives. Briefly, an ideal precision-recall curve would stay at a precision value of 1 (only true positives) until all true cases have been ‘recalled’ (recall value of 1, see also Chapter 7). Furthermore we plotted the recall, also termed as ‘sensitivity’, against given PTM-localization probability cutoffs (Figure 3.1b). The test on the basis of manually evaluated phosphopeptides determined in yeast cells yielded 92% precision relating to the correct assignment of class I phosphorylation sites, which satisfy a localization probability of 0.75. The corresponding precision-recall curve and the correlation diagram reflecting the sensitivity of the algorithm at different probability cutoffs are illustrated in Figure 3.1c and Figure 3.1d respectively.

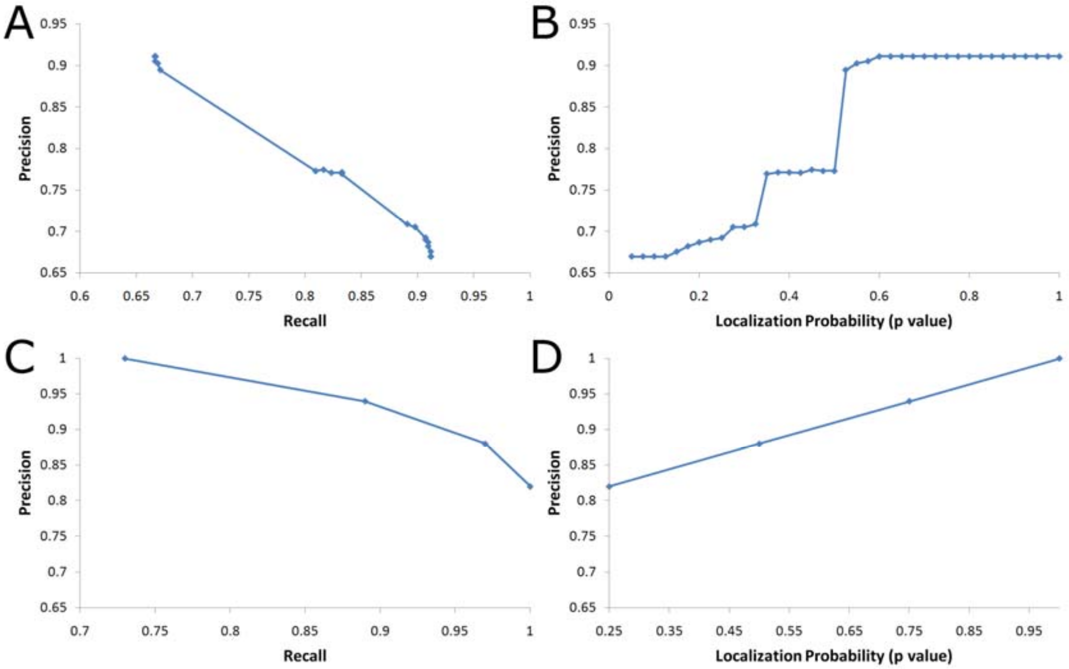


Figure 3.1: Validation of the PTM algorithm on the basis of synthetic phosphopeptides, phosphopeptides derived from tryptic digests of bovine caseins (a, b) and phosphopeptides identified in pheromone treated yeast cells (c, d)

3.4 Conclusion

Several algorithms have been described in the literature for protein identification by searching a sequence database using MS data. The probability based Mascot scoring algorithm assigns peptide sequences to MS/MS spectra and enables the user to judge whether the result is statistically significant. It has all the advantages of probability based approaches but is primarily optimized for the identification of peptide sequences. However, proper posttranslational modification location is also critical because many biological processes are regulated through the modification of specific residues. The entire concept of the phosphorylation site database (Chapter 4) also demands proper phosphorylation site placement. Therefore, we have developed a probability-based approach to calculate the likelihood of matching given ions to specific phosphorylation site locations. The algorithm is embedded in the PHOSIDA upload system (Chapter 4.2) and allows the calculation of localization-specific probability for each phosphorylation site within the given data set. The algorithm was originally described in the study of the human phosphoproteome upon EGF stimulation (Chapter 4.6.1.1.1) and enabled the automated site-specific investigation of high throughout phosphodata for the first time. We routinely apply the algorithm to all large scale studies of phosphoproteomes in our laboratory. It was originally implemented to derive site specific localizations of phosphorylation events on the basis of results from our open source MS computational platform MSQuant (www.msquant.org). Meanwhile, we have integrated this probability methodfully into the MaxQuant software, the current computational proteomics platform of our laboratory. To evaluate the method we analyzed the accuracy on the basis of known and manually validated phosphorylated peptides. The main finding of this test was that more than 90% of the phosphorylation sites were predicted correctly using a 0.75 cutoff relating to the resulting localization probability. The evaluation – including precision-recall curves - established that our approach is very accurate and efficiently extends the fragments-to-sequence-assignment from the peptide level to the residue level. We define phosphorylation sites, which satisfy a localization probability of 0.75, as ‘class I sites’. The integration of the described probability-based algorithm in the automated Phosida upload process allows the site specific investigation of identified phosphoproteomes. It is indispensable for the large scale analysis of various constraints of phosphorylation events including evolution (Chapter 9). In addition, the Phosida web application shows the exact localization probability of each determined phosphorylated site enabling web users to validate the residue specific assignment of posttranslational modification events within specified peptide sequences (Chapter 4.2.5).

Chapter 4

PHOSIDA – Phosphorylation Site Database

PHOSIDA, the phosphorylation site database, integrates thousands of high-confidence *in-vivo* phosphosites identified by MS-based proteomics in various species (Gnad et al., 2007). It comprises phosphoproteomes of various organisms ranging from bacteria including *Escherichia coli* and *Bacillus subtilis* to eukaryotes including yeast and human. It contains around 7000 phosphorylation sites that have been determined in human cancer cells upon EGF stimulation (Olsen et al., 2006). Since the objective of many of our phosphostudies was to quantify a given *in-vivo* phosphoproteome using SILAC (Chapter 2.1), PHOSIDA makes it possible to check phosphorylation changes after certain treatments such as growth factor stimulation and kinase/phosphatase inhibition by small molecules. On the protein level, PHOSIDA includes general information such as sequences, isoelectric points (pIs), motifs, active sites, binding sites, domains, gene ontology classifications and associated literature. These annotations are mainly derived from the SwissProt database, which is cross-linked to our database containing peptide identifications. On the phosphosite level, PHOSIDA provides information about matching kinase motifs, MS specific identification scores including localization probabilities, predicted secondary structures, and the residue conservation within a multitude of different species. Importantly, the underlying environment allows the automated integration of determined phosphoproteomes along with corresponding annotations from various sources. In addition, further information relating to evolution and structure are derived via a self-constructed pipeline. To establish a consistent database management, integrated projects have to be preprocessed and transformed in a uniform manner. This ensures that various projects can be compared in a very simple and fast way. Moreover, it allows the mining of phosphoproteomes of various organisms in a standardized way. The whole process constitutes a KDD process (Chapter 1).

Chapter 4.1 gives an overview of the general process of knowledge discovery in databases. Chapter 4.2 provides insights into the basic concepts of PHOSIDA. The application of the KDD process on MS specific datasets and its implementation into PHOSIDA are described in chapters 4.3 – 4.6. The description of the practical implementation of the KDD process is rounded off by a discussion (Chapter 4.7).

4.1 General Process of Knowledge Discovery in Databases (KDD)

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of very large amounts of data and the pressing need for turning such data into useful information and knowledge (Han, 2000). However, the abundance and diversity of data, coupled with the need for powerful data analysis tools, has been described as a 'data rich but information poor' situation. The fast-growing, tremendous amount of data that are collected and stored in databases has far exceeded human ability for comprehension. Thus data analysis including data mining can uncover important data patterns not accessible to direct inspection.

The widening gap between data and information demands a systematic development of data mining tools turning data tombs into 'golden nuggets' of knowledge. For example, classification of biological data such as protein folds, association rules detecting metabolic pathways and clustering of protein structure are essential data mining applications to gain information (Mitchell, 1997).

The KDD process is dedicated to derive knowledge from large scale datasets generated by the application of high throughput mass spectrometry technologies. The actual KDD process is applied to data that have already been processed via the typical MS data workflow and thus consist of assigned and quantified SILAC peptides (Chapter 2.1). These data are validated by manual inspection or by the specification of specific cutoffs governing the false positive rate. This rate of false positives can be estimated by the inclusion of reversed protein sequences in the database used for the identification yielding a ‘decoy database’ (Elias et al., 2005). The overall process can be defined as an integration process. It presents the preliminary data flow before the integration into the database, the first step of the actual KDD process (Figure 4.1).

The KDD process is frequently equated with the term ‘data mining’ and this is definitely the most important and essential procedure. However, it is only one step in the whole process of knowledge discovery: Data integration and data selection (Chapter 4.3) are the first steps of KDD, followed by data transformation (Chapter 4.4). Only then can data mining methods (Chapter 4.5) be applied (Figure 4.1). The final step is the evaluation of the obtained results linked to validation and presentation of the gained knowledge (Chapter 4.6). Frequently one has to go some steps backward in the KDD process, if the evaluation of the results is not satisfactory.

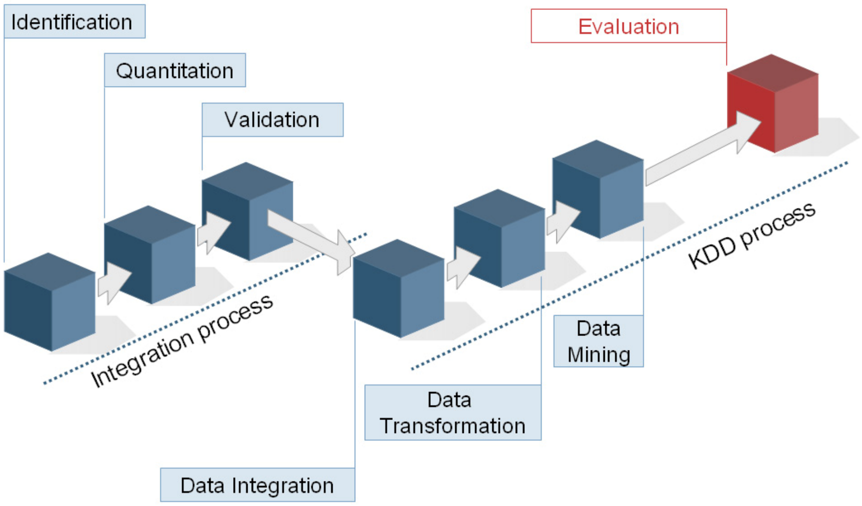


Figure 4.1: Process of KDD applied to mass spectrometry determined (phospho-) proteomes

4.2 Basic Concepts of PHOSIDA

PHOSIDA was developed to retrieve and analyze phosphosites from large-scale high-confidence phosphoproteomics experiments including quantitative data that describe the response of biological systems to various treatments. Thus it is the first phosphorylation site database to explicitely store quantitative data of site-specific phosphorylation changes. PHOSIDA also matches kinase motifs to phosphosites and illustrates the structural environments and conservation of phosphorylated residues.

As mentioned above, the final result of an MS based proteomics approach is the identity of peptides. The mapping of determined peptides to protein entries is challenging, as peptides can match to several protein sequences. This problem is addressed in PHOSIDA by a many-to-many mapping between phosphopeptide sequences and protein entries in the sequence database.

One of the fundamental strengths of PHOSIDA lies in the high quality of the *in vivo* data contained in the database and in the very large size of its *in vivo* data sets. PHOSIDA presents the most comprehensive database storing not only phosphosites identified in eukaryotic cells, but also phosphosites detected in prokaryotic cells.

4.2.1 Core Database Management of Phosphorylation Sites

As the primary goal of PHOSIDA is not only to make identified phosphorylation sites available to the public community, but also to derive biological context relating to phosphorylation events in the cell, there are two different PHOSIDA versions: One database scheme is designed to allow automated mining of phosphosets (Chapter 4.2.1.1), whereas the other database scheme is constructed for web usage (Chapter 4.2.1.2).

4.2.1.1 Database Schema adapted for Mining

The integration of phosphorylated peptides into PHOSIDA (version 1.1) is based on validated data processed via MASCOT (Perkins et al., 1999) and MSQuant (Andersen et al., 2003). The MASCOT software assigns measured spectra to peptide sequences (identification process), whereas the MSQuant software quantifies identified peptides. The final result is a list of detected peptides along with a variety of features such as charge status, MASCOT identification scores and quantitative data. Furthermore, all theoretical combinations of modifications of each peptide are listed along with posttranslational modification (PTM) scores as calculated by a probability based algorithm (Chapter 3). This combinatorial listing provides the basis for the derivation of the probability for each residue to be phosphorylated within the given peptide.

For each peptide, its sequence, number of phosphorylated residues, Mascot score, PTM score, and quantitative data are uploaded to the PHOSIDA database. In some cases, the experimental design requires the inclusion of additional attributes such as cellular localizations. The PHOSIDA 1.1 upload also comprises a procedure that assigns each peptide to a specific protein entry of the corresponding database. The assignment of peptides that occur uniquely in one protein of the given database is unambiguous, however, peptides that occur in several proteins are assigned to the protein that shows the highest total number of identified peptides (this is the most likely protein form to be present in the measured proteome). The many-to-one assignment between peptides and corresponding proteins is essential to derive general patterns from non-redundant data. Many-to-many relationships between non-unique peptides and proteins as used for the online application (Chapter 4.2.1.2) would artificially increase the number of identified proteins yielding misleading results. The database relation ‘peptides’ contains all identified peptides distinguishable by their sequence and number of phosphorylations (Figure 4.2). Each peptide entry is uniquely indexed by the ‘pep_id’ identifier. Thus, the ‘pep_id’ presents the primary key of this relation (Chapter 2.2). Usually, many measured instances correspond to a single peptide entry due to varying charge states, duplicate experiments, etc. The database relation ‘peptides_sub’ contains each measured entity. Its primary key is termed ‘subpep_id’. Since there are several instances associated with one peptide, the relationship between ‘peptides’ and ‘peptides_sub’ is one-to-many. The attribute ‘pep_id’ serves as foreign key linking the table ‘peptides_sub’ to ‘peptides’.

The SILAC technology allows the quantitation of peptides in three different conditions using light, medium and heavy amino acid labelling (Chapter 2.1). If one is interested in the intensity distribution in more than three different conditions, one has to combine multiple SILAC based experiments. Two SILAC experiments can compare five conditions because one common point is needed for normalization. To combine quantitative data from parallel SILAC experiments, we assign abundance levels of the top scoring peptide instances observed in one specified experimental condition to the associated peptide entry. Combined quantitative data are integrated into the relation ‘peptides’, whereas quantitative data for each instance are integrated into the relation ‘peptides_sub’.

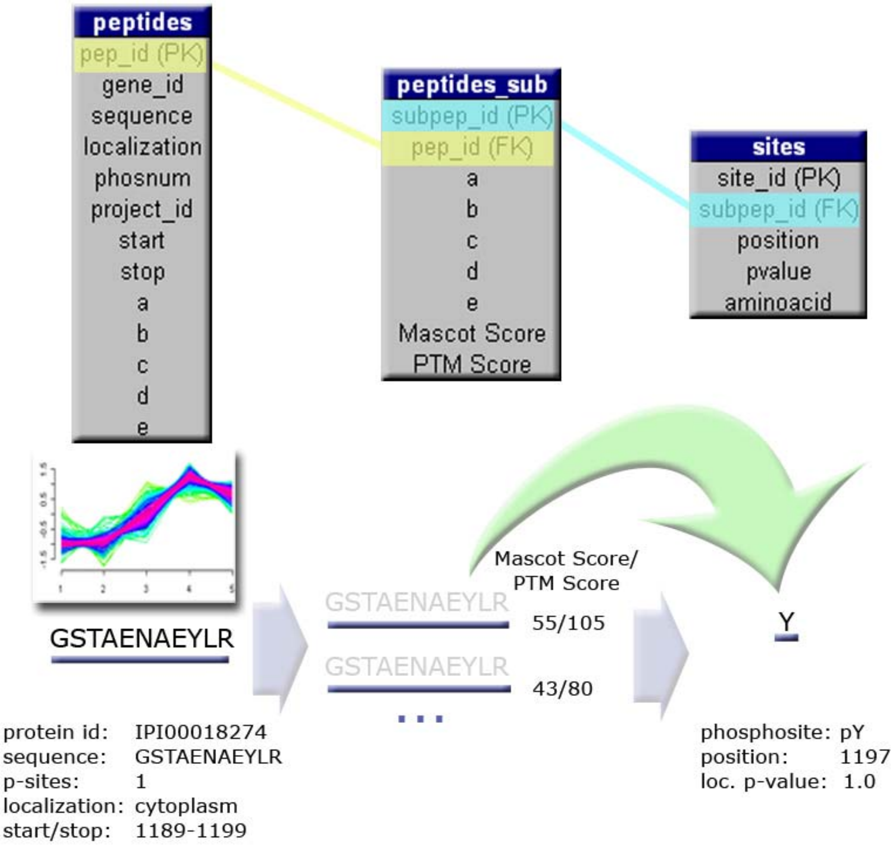


Figure 4.2: Basic database schema of PHOSIDA 1.1

In addition to the integration of phosphorylated peptides, associated phosphorylation sites are uploaded, too (Figure 4.2). For each peptide instance, the corresponding phosphorylated residues are stored in relation ‘sites’. Each entry contains the position of the phosphosite in the protein sequence, the localization probability, and the type of amino acid. Thus, there are many instances for each peptide instance in the case of multiple phosphorylation and ambiguous site phosphorylation. This results in a one-to-many relationship between the database relations ‘peptides_sub’ and ‘sites’. As apparent from the database schema (Figure 4.2), PHOSIDA database version 1.1 is peptide based. Consequently, quantitative data of peptides are directly assigned to all residues that are phosphorylated within each peptide instance.

In contrast to PHOSIDA version 1.1, the second database version (1.2) is predominantly phosphorylation site based. The upload process is also different: The upload process of database version 1.1 is based on a single result file generated by MSQuant. In contrast, the upload process of database version 1.2 is based on several result files generated by the new computational proteomics environment, MaxQuant. The result files list identifies peptides and phosphorylated residues separately. Each file is cross-linked via unique identifiers. Therefore, the concept of the MaxQuant result files already reflects the logical schema of the database (Figure 4.3). Furthermore, calculated localization p-values of phosphosites and the correct protein assignments are already provided by MaxQuant. The idea of a site-specific database schema is primarily reflected by the fact that quantitative data are directly assigned to phosphorylation sites in a sophisticated manner: The quantitation of posttranslationally modified residues is based on taking the median of the quantitative data of all peptides containing the given modified residue. Hence, the database relation ‘sites’ is the most comprehensive table including the maximum localization probabilities observed in all corresponding peptides, assigned protein identifiers, amino acid types, quantitative data, and further features. For each phosphosite, the top scoring peptide instance is stored in the relation ‘peptides_sub’. The database relation ‘sites’ is linked to ‘peptides_sub’ via ‘subpep_id’ identifiers. The relationship between the tables ‘peptides’ and ‘peptides_sub’ is the same as the one of PHOSIDA version 1.1.

The initial upload of identified phosphorylated peptides is followed by a number of further processes that contribute to the KDD process.

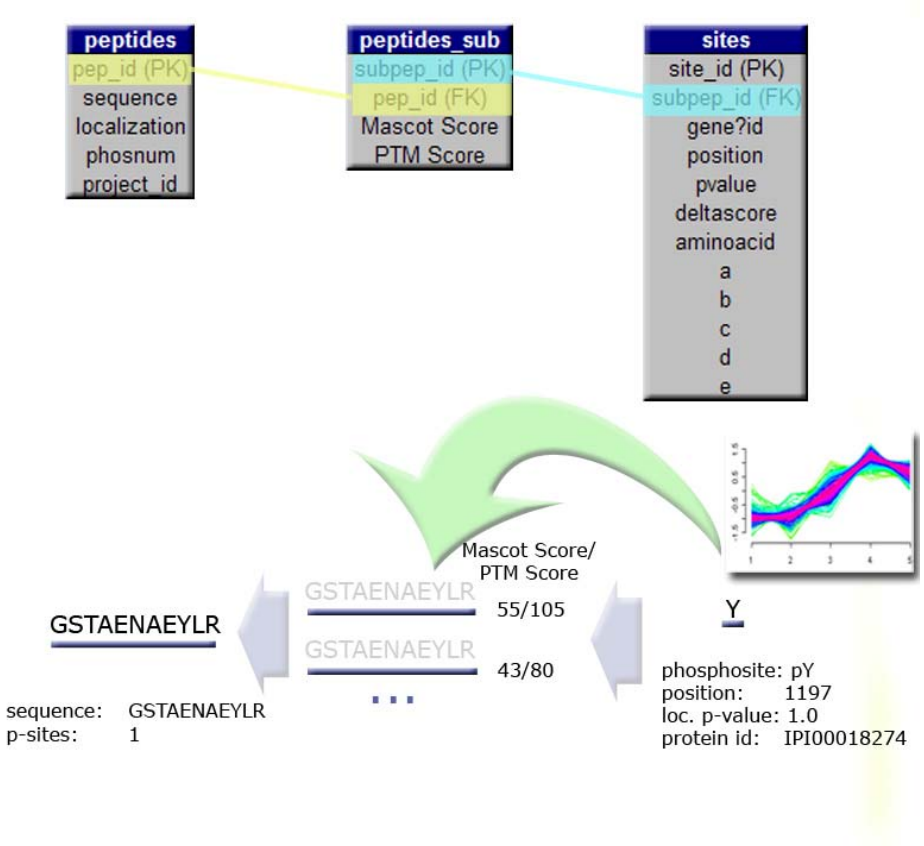


Figure 4.3: Basic database schema of PHOSIDA 1.2

4.2.1.2 Online Database Schema

The concept of the database schema providing the basis for the web applications is different from the one of the PHOSIDA versions described in Chapter 4.2.1. The transformation between the two database schemes is carried out automatically. Depending on the underlying quantitation software one can distinguish between a peptide-based online database schema (Figure 4.4) and a site-based online database schema. The only difference between the two online database schemas is that quantitative data are assigned to peptides in the one scheme, whereas quantitative data are attributed to phosphorylated residues in the other scheme.

In contrast to the database schemes designed for mining, the online database schemas are based on the principle of many-to-many protein-peptide assignments. Hence, each peptide is assigned to all proteins that contain the given peptide sequence. This relationship is reflected in the database relation ‘idmatch__[project_id]’, as it assigns each identified peptide to all corresponding proteins. Therefore, each protein potentially shows a multitude of peptides stored in relation ‘idpower__[project_id]’. The correct peptide-protein assignment is predicated on the assumption that the higher the total number of assigned peptides, the higher the probability that the given protein was identified.

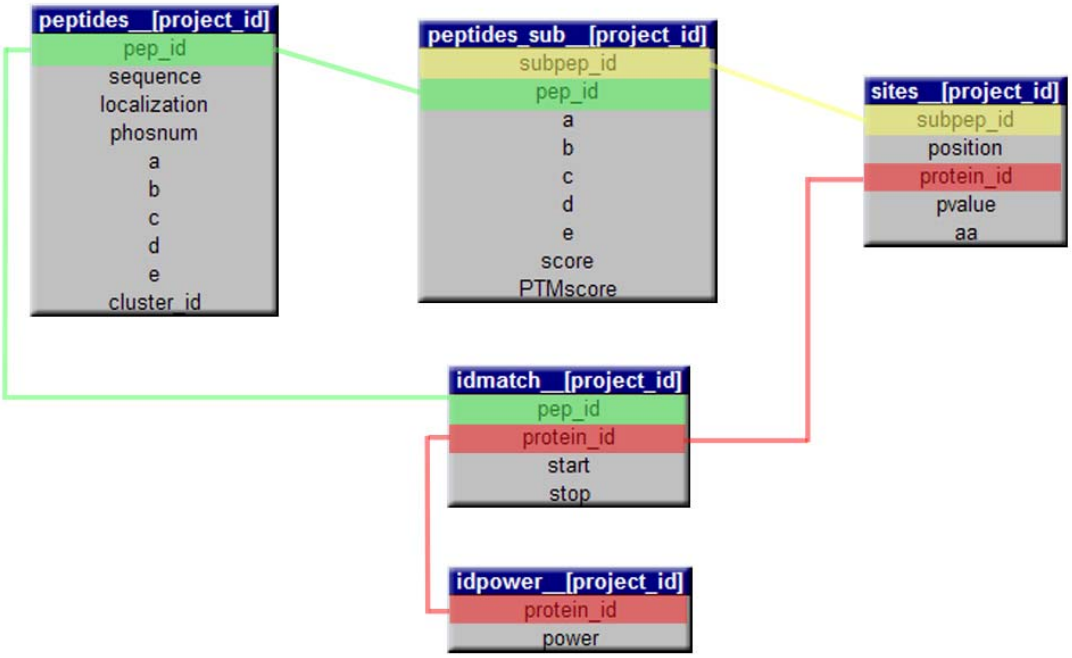


Figure 4.4: Online database schema

4.2.1.3 Integration of additional Biological Data

In addition to the upload of identified phosphosites along with their corresponding peptides, further biological data sources have to be integrated, in order to fulfil the requirements of mining and to enable web users to derive a biological context for any protein of interest.

It is obvious to include general protein features that are outlined in the database that was used for the peptide identification. For example, in the case of the International Protein Index (IPI) database (Kersey et al., 2004), the downloadable files contain general descriptions, features such as pI and molecular weight, and gene symbols besides the sequence of each protein. These attributes are integrated into PHOSIDA (Figure 4.5). They are not only used for mining, but also for a more comprehensive illustration of each phosphorylated protein on the web. Therefore, the inclusion of additional protein characteristics is added to the database that is required for mining and to the online database as well.

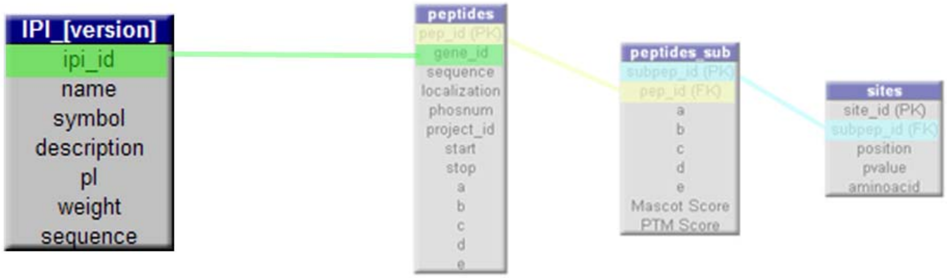


Figure 4.5: Integration of additional protein features into the PHOSIDA database

The Gene Ontology (GO) annotation is another valuable data resource (Ashburner et al., 2000). The GO project is a collaborative effort to address the need for consistent functional descriptions of gene products in different databases. The three organizing principles of GO are molecular function, biological process and cellular component. Many gene products are associated with a multitude of functions, processes, or cellular localizations. The Gene Ontology Annotation (GOA) database (Camon et al., 2004) provides GO annotations to protein entries of the IPI database, for example. Its inclusion requires only one additional database relation (Figure 4.6).

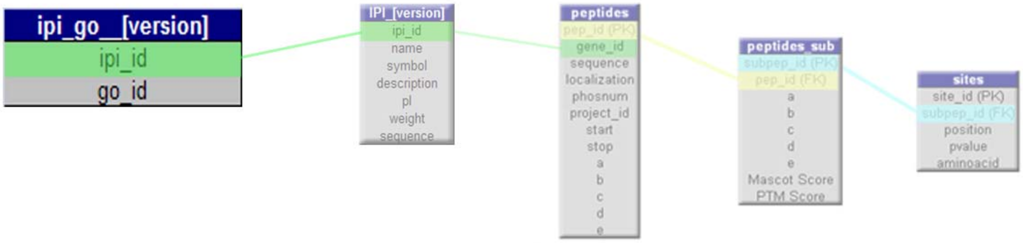


Figure 4.6: Integration of GO annotations into the database

In addition, the protein database SwissProt provides a high level of annotation ranging from the domain structure of a protein to post-translational modifications and corresponding literature (Bairoch and Apweiler, 1996). Therefore, it constitutes an excellent resource to gain deeper insight into biological context. In particular, the integration of annotated post-translational modifications makes it possible to determine if an identified phosphorylation site is novel. However, the inclusion of protein annotations from various databases presents a challenge, since the annotations might be based on different products of the same gene. For example, the epidermal growth factor receptor precursor protein has only one entry in SwissProt, which can be uniquely identified by its accession number ‘P00533’. In contrast, the IPI database contains four different entries for the same protein due to various splice forms (IPI00018274, IPI00221346, IPI00221347, IPI00221348). Thus, comprehensive sequence mappings between various databases are required to combine protein annotations of various sources.

To align protein sequences of various databases, we used the basic local alignment search tool BLAST (Altschul et al., 1990). It allows rapid sequence comparisons and creates alignments that optimize a measure of local similarity. BLAST searches for high scoring sequence alignments using a heuristic approach that approximates the Smith-Waterman algorithm (Smith and Waterman, 1981) but is much faster. It is the most popular bioinformatics tool in use today due to its speed and accuracy. To align protein sequences, we used the software BLASTP. It is optimized for the comparison of amino acid sequences. The automated comparisons between corresponding protein sequences of various databases result into the database relation ‘map_[organism]_[db1]_[db2]’, which stores the generated alignments (Figure 4.7).

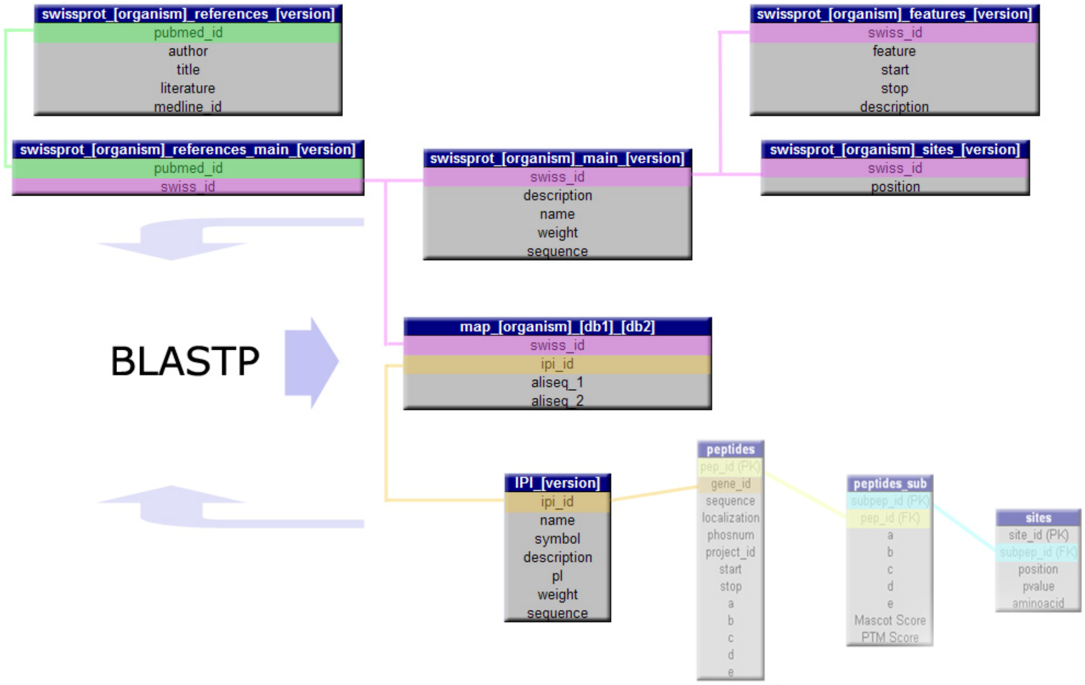


Figure 4.7: Database integration of several databases such as IPI and SwissProt requires comprehensive sequence alignments for a merged protein annotation

4.2.2 Kinase Motif Matching

Protein phosphorylation levels are essential for understanding the basic principles of signalling pathways in both normal and diseased cell states (Chapter 1) (Pawson and Scott, 2005). The derivation of consensus sequences (motifs) for protein kinase sites of phosphorylation is essential to estimate the 'kinase affiliation' of substrates. Consensus sequences are primarily deduced from *in-vitro* incubations of kinases with a combinatorial peptide library and ATP. In addition, there are many algorithms that extract motifs *in-silico*. Among these, an iterative statistical approach proved to be the best performing method to identify protein phosphorylation motifs from large-scale data sets (Schwartz and Gygi, 2005). With verified kinase motifs in hand, one is in principle able to determine the kinase responsible for a given protein substrate phosphorylation of interest. However, previous experience has shown that one has to check the matching of consensus sequences on the site level, as many proteins are substrates of different kinases and participate in different pathways. Therefore, for each phosphorylated site, the matching consensus sequences are illustrated in PHOSIDA (Chapter 4.2.5). PHOSIDA checks 34 different consensus sequences of various human kinases such as casein kinase and glycogen synthase kinase against each site.

Besides the estimated assignments of kinases for each phosphorylated residue of interest, the inclusion of kinase motif matches makes it possible to check the over- and underrepresentation of matching consensus sequences in a given large-scale dataset of phosphorylation sites (Chapter 4.5.1). The significance of motif matches provides insight into the overall kinase distribution that initiated the phosphorylation of specified substrates.

4.2.3 Structural Investigation of Phosphoproteomes

Previous studies have already shown that phosphorylation sites are mainly located in parts of proteins without regular structure (Dunker et al., 2002; Iakoucheva et al., 2004). To verify this observation on the basis of our large-scale studies and to enable users to investigate the structural context of each phosphorylation site of interest (Chapter 4.2.6), we performed large-scale solvent accessibility calculation as well as secondary structure prediction employing the SABLE 2.0 program (Wagner et al., 2005). The predicted structural constraints of each residue of a given phosphorylated protein are stored in the database relation 'structures_[project_id]' (Figure 4.8). Besides the predicted secondary structures and solvent accessibilities scaling from 0 (low accessibility) to 9 (high accessibility), the corresponding residue specific validation scores are stored. An inner join with the database relations that contain the identified phosphorylation yields the virtual relation ‘sites_structure_[project_id]’. It comprises the structural context of each identified phosphorylation site.

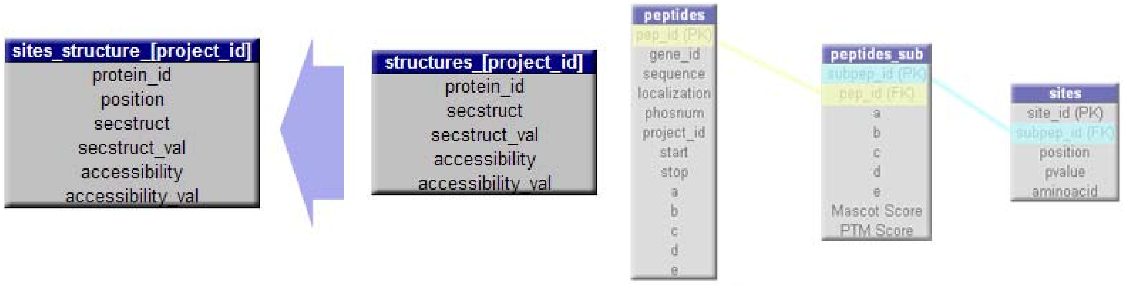


Figure 4.8: Adding structural context to the PHOSIDA database

4.2.4 Evolutionary Conservation of Phosphoproteomes

The generation of high-throughput data of posttranslationally modified proteomes of various species enables us to answer the following questions relating to the conservation of phosphorylation events: Did an identified phosphorylated protein of a given species such as human already occur in distantly related species such as bacteria? Is there a highly conserved protein that is orthologous to a specified phosphorylated protein, and, if so, is the homologous protein also phosphorylated in the other organism? Can one observe any differences relating to the conservation on various levels ranging from the evolutionary preservation of the protein to the conservation of the specified phosphorylated residue? These questions and further basic issues relating to conservation can be answered by the application of appropriate algorithms that try to find the highest similarity between protein sequences by aligning them in a fast and accurate way: To find homologous proteins, we used BLASTP (Chapter 4.2.1.3) (Altschul et al., 1990). We defined proteins to be homologous, if the resulting E-values reflecting the significance of sequence similarities were lower than 10^{-5}, which is a frequently used cutoff to determine homology. To distinguish proteins that are homologous within one species (paralogs) and proteins that are homologous between species (orthologs), we used a two-directional BLASTP approach (O'Brien et al., 2005).

Since BLAST is a heuristic approach that approximates the Smith-Waterman algorithm (Smith and Waterman, 1981), it creates sequence alignments that show a very high local similarity. If two given sequences do not also show high overall sequence similarity, the resulting sequence alignment will not cover the entire lengths of both sequences. Hence, we used the software Needle (Rice et al., 2000), which is based on the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). It generates global alignments that cover the total lengths of protein sequences. The only disadvantage is the processing time, as the method involves dynamic programming. However, it is the most accurate method to align sequences globally and is guranteed to find the best global alignment.

Therefore, the combination of BLASTP, which approximates homology relationships between proteins, and Needle, which generates global alignments of homologs, is an appropriate method to measure the degree of conservation of phosphorylation on various levels. The classified phylogenetic relationships between phosphorylated proteins and proteins of other selected species covering the phylogenetic tree representatively are stored in database relations 'orthologs_[project_id]' and 'alignments_[project_id]' (Figure 4.9). The relation 'orthologs_[project_id]' indicates the homology of each phosphoprotein (no homology, paralog, or ortholog), whereas relation 'alignments_[project_id]' stores the global aligned sequences along with the overall sequence similarity and the accession number of the homologous protein.

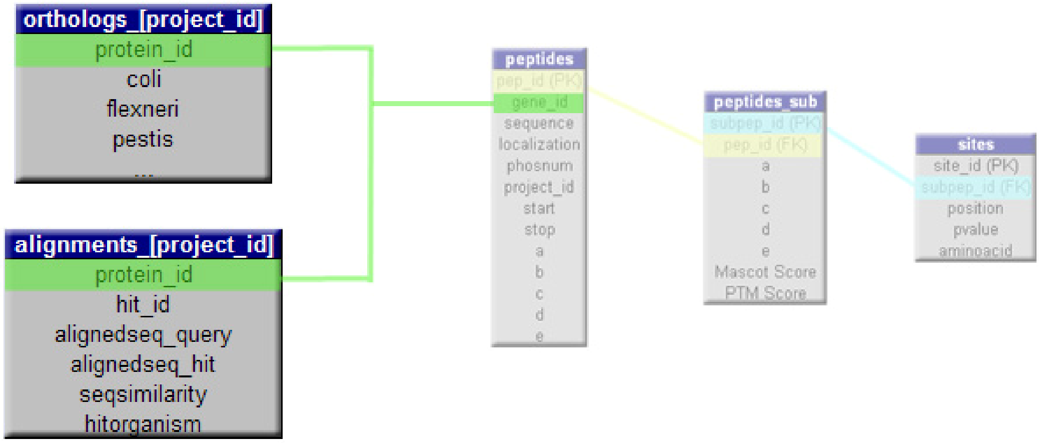


Figure 4.9: Database integration of phylogenetic relationships and global alignments of homologous proteins

The inclusion of derived phylogenetic relationships along with the resulting global protein sequence alignments allows the investigation of the conservation of phosphorylation events on various levels ranging from the overall protein level to the specific phosphorylated residue.

Given high-throughput data of phosphorylated proteomes, one is then able to draw global conclusions about the conservation of phosphorylation events.

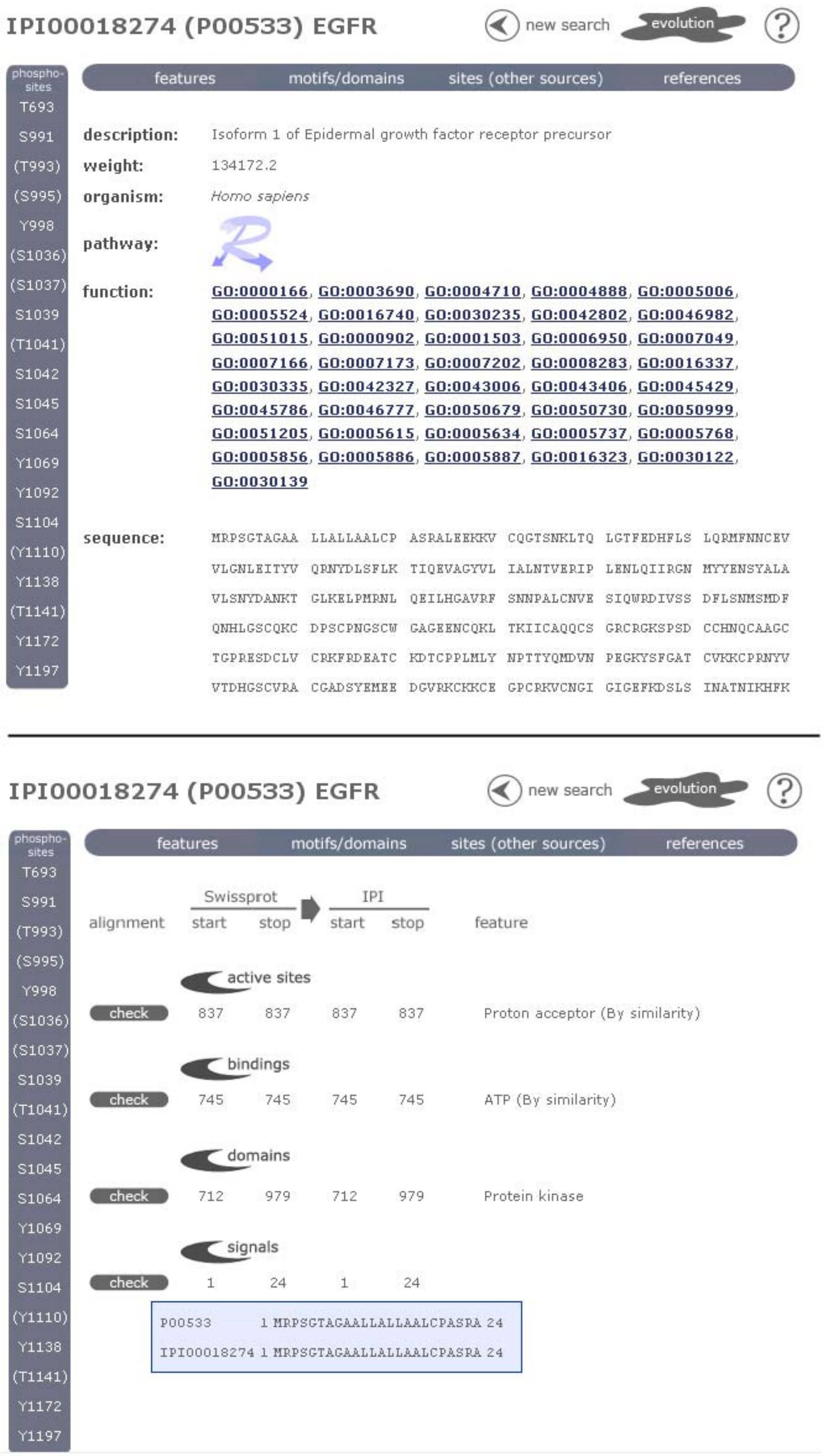
Furthermore, the internet implementation of PHOSIDA enables web users to check the conservation of any phosphorylated protein or phosphorylated site of interest (Chapter 4.2.5). A detailed report of the results regarding the evolution of phosphorylation is described in Chapter 9.

4.2.5 General Web Application of PHOSIDA

The web user is able to search for any protein of interest within a specified organism for phosphorylation sites. In the cases of mouse and human, it is possible to search via IPI accession number, SwissProt identifier, gene symbol, protein name, peptide sequence, protein sequence or any substring that matches with the description of a given protein.

For each protein, the user is presented with general features such as isoelectric point (pI), molecular weight, sequence, and description at the protein level in addition to corresponding GO accession numbers, which are directly hyperlinked to the detailed description of the annotation at the Gene Ontology website (www.godatabase.org) (Figure 4.10 upper panel). In addition, a hyperlink to the Reactome database (Joshi-Tope et al., 2005) is provided, if the specified protein is annotated in Reactome (www.reactome.org). As Reactome is a curated knowledgebase of biological pathways, the phosphorylation events can then be associated to candidate pathways. Furthermore, annotated protein features such as active sites, binding sites, domains, and signal sequences are derived from the SwissProt database (Chapter 4.2.1.3) and displayed, when clicking the ‘motifs/domains’ button.

In the case of the human proteome, protein assignments were based on the IPI database, which is cross-referenced with the SwissProt database by PHOSIDA. Entries of both databases that correspond to the same proteins were aligned to derive the exact positions of these protein features. Already annotated phosphosites derived from SwissProt are mapped to the IPI sequences in the same way and listed when clicking the ‘sites (other sources)’ button. The aligned regions can be visualized via ‘check alignment’ buttons (Figure 4.10 lower panel). Furthermore, corresponding literature references are provided.



description: Isoform 1 of Epidermal growth factor receptor precursor

weight: 134172.2

organism: *Homo sapiens*

pathway:

function: GO:0000166, GO:0003690, GO:0004710, GO:0004888, GO:0005006, GO:0005524, GO:0016740, GO:0030235, GO:0042802, GO:0046982, GO:0051015, GO:0000902, GO:0001503, GO:0006950, GO:0007049, GO:0007166, GO:0007173, GO:0007202, GO:0008283, GO:0016337, GO:0030335, GO:0042327, GO:0043006, GO:0043406, GO:0045429, GO:0045786, GO:0046777, GO:0050679, GO:0050730, GO:0050999, GO:0051205, GO:0005615, GO:0005634, GO:0005737, GO:0005768, GO:0005856, GO:0005886, GO:0005887, GO:0016323, GO:0030122, GO:0030139

sequence:

```
MRPSGTAGAA  LLALLAALCP  ASRALEEKKV  CQGTSNKLTQ  LGTFEDHFLS  LQRMFNNCEV
VLGNLEITYV  QRNYDLSFLK  TIQEVAGYVL  IALNTVERIP  LENLQIIRGN  MYYENSYALA
VLSNYDANKT  GLKELPMRNL  QEILHGAVRF  SNNPALCNVE  SIQWRDIVSS  DFLSNMSMDF
QNHLGSCQKC  DPSCPNGSCW  GAGEENCQKL  TKIICAQQCS  GRCRGKSPSD  CCHNQCAAGC
TGPRESDCLV  CRKFRDEATC  KDTCPPLMLY  NPTTYQMDVN  PEGKYSFGAT  CVKKCPRNYV
VTDHGSCVRA  CGADSYEMEE  DGVRKCKKCE  GPCRKVCNGI  GIGEFKDSLS  INATNIKHFK
```

alignment	Swissprot start	Swissprot stop	IPI start	IPI stop	feature
active sites					
check	837	837	837	837	Proton acceptor (By similarity)
bindings					
check	745	745	745	745	ATP (By similarity)
domains					
check	712	979	712	979	Protein kinase
signals					
check	1	24	1	24	

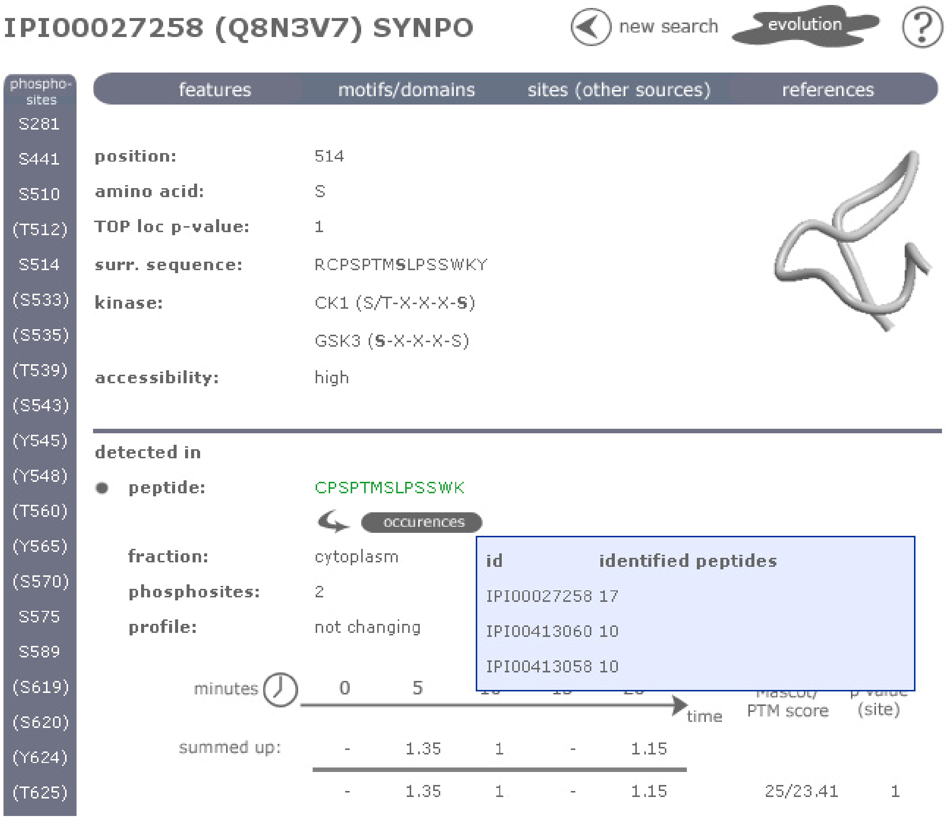
```
P00533      1 MRPSGTAGAALLALLAALCPASRA 24
IPI00018274 1 MRPSGTAGAALLALLAALCPASRA 24
```

Figure 4.10: Illustration of general protein features in PHOSIDA

PHOSIDA shows the description, sequence, weight, gene ontology annotation of each phosphoprotein (upper panel). In addition, PHOSIDA displays annotated domains, binding sites, active sites, and signaling regions along with the aligned sequences between the IPI database and the SwissProt database (lower panel).

Moreover, all phosphorylated sites that have been identified in the project are displayed on the left panel. This presentation allows to check immediately whether phosphorylation sites occur within known domains or other sequence regions potentially associated with signaling such as binding sites are phosphorylated. In such cases, one can link the phosphorylation event to its potential functional consequences. If the localization probability is lower than 0.75, it is enclosed in round brackets. When users click on any of the displayed phosphosites, the surrounding sequence and matching kinase motifs are shown (Figure 4.11).
Often, several phosphopeptides covering the same phosphosite are measured by mass spectrometry. These peptides are also listed along with their localization probabilities, Mascot scores, and PTM scores for each instance. Depending on the experimental design, PHOSIDA contains quantitative data including time-resolved data, where applicable, of each phosphopeptide. Figure 4.11 shows corresponding ratios and clustered time courses as illustrated in PHOSIDA. These data are listed separately for peptides as a function of their sequences, degrees of phosphorylation, and further categories, such as experimental design or fraction (for example nuclear or cytosolic). When moving the mouse over the 'occurences' button, protein entries sharing the same phosphopeptide of interest are listed along with the number of unique peptides that have been measured in one experimental project. Each peptide is color coded according to the protein assignment: if the peptide sequence is marked in green, the selected protein has the maximum number of peptides in comparison to all other proteins that contain the same peptide. If the protein assignment is ambiguous because of another protein with the same number of identified peptides, the peptide is highlighted in blue. Red indicates that other proteins exceed the number of detected peptides in comparison to the selected phosphoprotein. Each feature of PHOSIDA is explained in the help menu, which is accessible via the 'background' menu or via clicking on the 'question mark' button at the page of interest.

Furthermore, as depicted in Figure 4.11, the predicted structural attributes of each phosphorylation site are visualized in PHOSIDA. The solvent accessibility is classified into 'low', 'medium', and 'high'. Secondary structures are classified into 'loop/turn', 'α helix', and 'β sheet'.


IPI00027258 (Q8N3V7) SYNPO

position: 514
amino acid: S
TOP loc p-value: 1
surr. sequence: RCPSPTMSLPSSWKY
kinase: CK1 (S/T-X-X-X-S)
GSK3 (S-X-X-X-S)
accessibility: high

detected in
● **peptide:** CPSPTMSLPSSWK

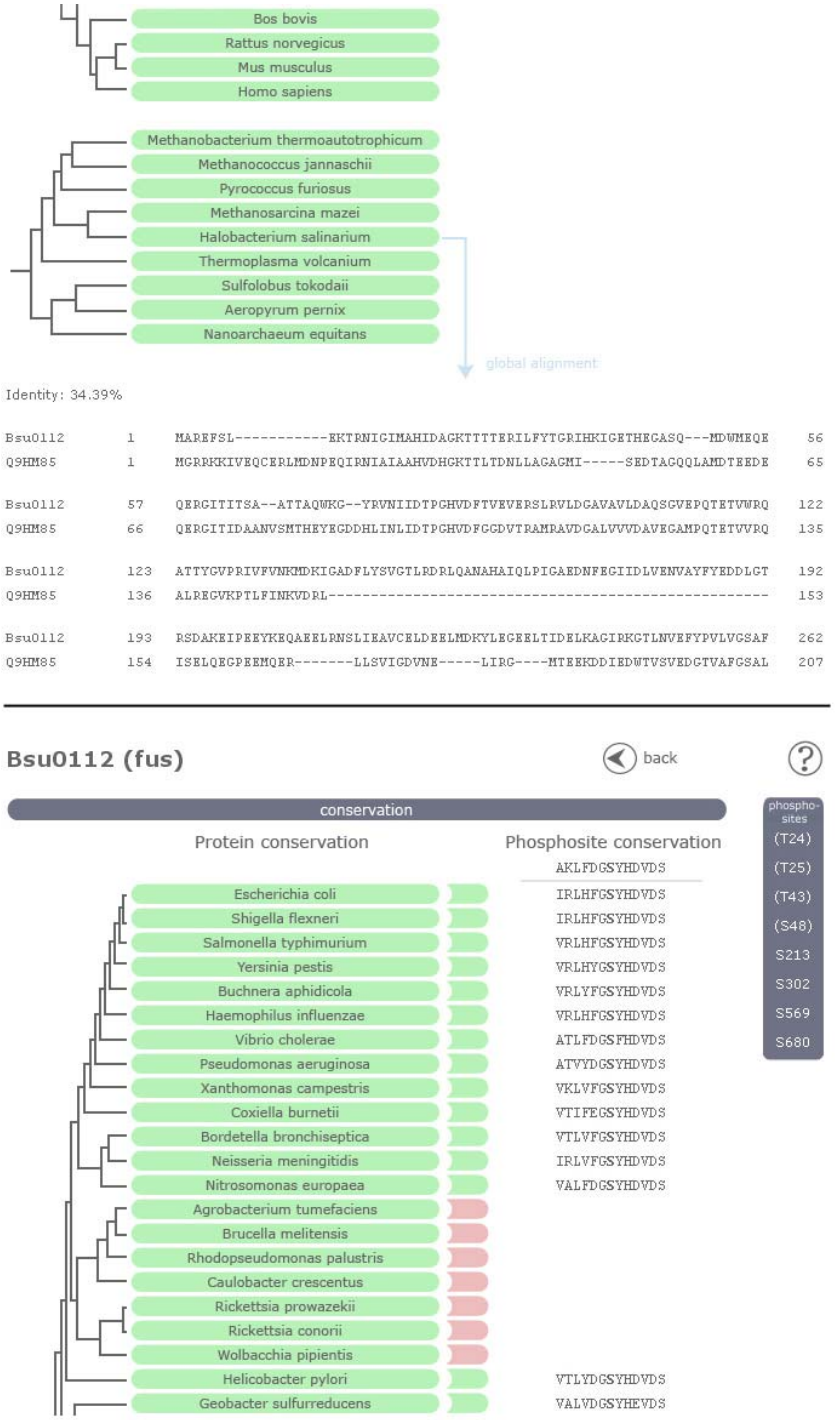
fraction: cytoplasm
phosphosites: 2
profile: not changing

id	identified peptides
IPI00027258	17
IPI00413060	10
IPI00413058	10

minutes	0	5				Mascot/ PTM score	p value (site)
summed up:	-	1.35	1	-	1.15		
	-	1.35	1	-	1.15	25/23.41	1

Figure 4.11: Phosphorylation site specific information in PHOSIDA

The evolutionary section of PHOSIDA displays the results of the homology searches (Chapter 4.2.5) using an approximate phylogeny of all investigated species (Figure 4.12 upper panel). Taxonomic divisions are displayed on-screen when the cursor is pointed at the phylogenetic tree. If the selected phosphoprotein is not homologous to any protein of a particular organism, that organism is highlighted in red. If the similarity between the sequence of the phosphoprotein and its homologous protein was the most significant one in both directions, the given organism is highlighted in green. A higher similarity between the sequence of the homologous protein and another protein of the organism of the selected phosphoprotein suggests paralogy, indicated in blue. The full global alignment between the given phosphoprotein and the orthologous protein of a specific organism is shown when the web user clicks on the organism button (Figure 4.12 upper panel). In addition, all phosphorylation sites that have been measured in our laboratory are listed on the right side. If users click on a phosphorylation site of interest, the conservation status of the selected phosphorylation site is indicated in red or green, whith green indicating conservation (Figure 4.12 lower panel). For conserved phosphosites, the alignment of the surrounding sequence is displayed. With alignments between the phosphorylation site of interest and protein sequences from 70 organisms, PHOSIDA enables users to check the conservation of each site of each protein of interest. Furthermore, the conservation of matching motifs can immediately be checked. This enables the user to distinguish conserved motifs around the phosphosite from other motifs that also match the phosphosite but are not conserved and may thus be less likely to be functionally important or have appeared only recently in evolution.



```
Identity: 34.39%

Bsu0112    1     MAREFSL-----------EKTRNIGIMAHIDAGKTTTTERILFYTGRIHKIGETHEGASQ---MDWMEQE    56
Q9HM85     1     MGRRKKIVEQCERLMDNPEQIRNIAIAAHVDHGKTTLTDNLLAGAGMI-----SEDTAGQQLAMDTEEDE    65

Bsu0112    57    QERGITITSA--ATTAQWKG--YRVNIIDTPGHVDFTVEVERSLRVLDGAVAVLDAQSGVEPQTETVWRQ   122
Q9HM85     66    QERGITIDAANVSMTHEYEGDDHLINLIDTPGHVDFGGDVTRAMRAVDGALVVVDAVEGAMPQTETVVRQ   135

Bsu0112    123   ATTYGVPRIVFVNKMDKIGADFLYSVGTLRDRLQANAHAIQLPIGAEDNFEGIIDLVENVAYFYEDDLGT   192
Q9HM85     136   ALREGVKPTLFINKVDRL----------------------------------------------------   153

Bsu0112    193   RSDAKEIPEEYKEQAEELRNSLIEAVCELDEELMDKYLEGEELTIDELKAGIRKGTLNVEFYPVLVGSAF   262
Q9HM85     154   ISELQEGPEEMQER-------LLSVIGDVNE-----LIRG----MTEEKDDIEDWTVSVEDGTVAFGSAL   207
```

Bsu0112 (fus)

Protein conservation	Phosphosite conservation
	AKLFDGSYHDVDS
Escherichia coli	IRLHFGSYHDVDS
Shigella flexneri	IRLHFGSYHDVDS
Salmonella typhimurium	VRLHFGSYHDVDS
Yersinia pestis	VRLHYGSYHDVDS
Buchnera aphidicola	VRLYFGSYHDVDS
Haemophilus influenzae	VRLHFGSYHDVDS
Vibrio cholerae	ATLFDGSFHDVDS
Pseudomonas aeruginosa	ATVYDGSYHDVDS
Xanthomonas campestris	VKLVFGSYHDVDS
Coxiella burnetii	VTIFEGSYHDVDS
Bordetella bronchiseptica	VTLVFGSYHDVDS
Neisseria meningitidis	IRLVFGSYHDVDS
Nitrosomonas europaea	VALFDGSYHDVDS
Agrobacterium tumefaciens	
Brucella melitensis	
Rhodopseudomonas palustris	
Caulobacter crescentus	
Rickettsia prowazekii	
Rickettsia conorii	
Wolbacchia pipientis	
Helicobacter pylori	VTLYDGSYHDVDS
Geobacter sulfurreducens	VALVDGSYHEVDS

Figure 4.12: Illustration of phylogenetic relationships and global alignments between phosphorylated proteins and homologous proteins (upper panel) and phosphosite conservation in PHOSIDA (lower panel)

Besides the online display of phosphorylation sites on different levels ranging from the protein level to the residue level under various aspects including conservation, phosphorylation changes, and structural constraints, PHOSIDA also contains other sections that are explained in detail in other chapters of this study:
With thousands of phosphorylation sites in hands, we next trained a support vector machine (SVM) that distinguishes between positive and negative instances on the basis of various features such as the surrounding amino acid sequence. Thus, the SVM is capable of predicting phosphorylation sites *in-silico*. This enables researches to detect possible phosphorylation sites for any protein of interest. This application of the PHOSIDA predictor can be used as the first step in planning an experiment. The implementation, accuracy, field of application, and web usage of the prediction method are subject of Chapter 7.

Finally, we used measured proteomic data to annotate the genome. This approach provides insight not only into the encoding of phosphorylated residues on the genome, but also enables to connect the Phosida databases with genome databases such as the EnsEMBL database. The inclusion of the online genome annotation section in PHOSIDA, the direct linkage to genome databases, and the integration of PHOSIDA annotated proteomic data in genome databases via the Distributed Annotation System (DAS) source technology are discussed in Chapter 8.

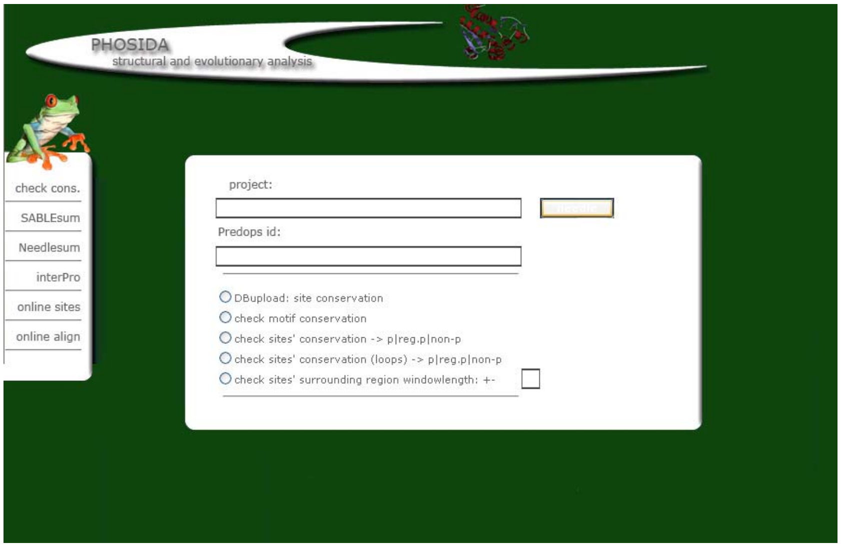
4.2.6 Administration Tool

To facilitate the administration and management of the phosphorylation site database along with its associated mining methods, we created three web based administration tools:
The main maintenance application allows the automated upload of large-scale phosphorylation datasets to the database version of PHOSIDA that is appropriate for the application of mining tools (Chapter 4.2.1.1) as well as to the online database (Chapter 4.2.1.2). To upload the data of a specific project, the only required inputs are the file paths to the resulting MSQuant or MaxQuant files, the corresponding protein database, which already has to have been uploaded, and optional filtering criteria relating to probability scores. Furthermore, it is possible to upload sequences, gene symbols, accession numbers, descriptions, and molecular weights of proteins from various public databases such as IPI (Kersey et al., 2004), SwissProt (Bairoch and Apweiler, 1996), FlyBase (Grumbling and Strelets, 2006), TIGR (Kirkness and Kerlavage, 1997), NCBI (Benson et al., 2008), and SGD (Cherry et al., 1998) on the basis of given FASTA files. Then, the entries of different databases can be automatically cross-referenced via BLAST alignments (Chapter 4.2.1.3)

resulting in additional automatically generated database relations. Biological data such as Gene Ontology annotations and SwissProt annotations can also be uploaded in an automated way. Another important feature of any sequence based database is the update to the most current database release. To compare the results of various experiments within one species, the data have to be ideally predicated on the same database release. Therefore, the database management tool includes methods to reassign all phosphorylated peptides and phosphorylated sites to a newer database release. Besides the general database management, it is also possible to use various mining methods on the database directly. For example, this allows the derivation of significant patterns relating to kinase assignments and the creation of comprehensive tables that provide an overall overview of the large-scale data.

Another administrative web based tool (Figure 4.13) is specialized on the derivation of phylogenetic relationships and the creation of global alignments between phosphorylated proteins and homologous proteins of more than 70 other species (Chapter 4.2.4). Moreover, it integrates predicted structural features (Chapter 4.2.3) and generates and integration data that are relevant for the evolutionary and structural analysis of phosphorylated proteins. It also creates input files that are used to train the support vector machine that distinguishes between phosphorylated and non-phosphorylated residues taking their primary sequence environment, structural context, and conservation into account (Chapter 7).

Finally, the purpose of the third management tool is the application of various extensive analyses that assess the conservation on various levels ranging from the protein level to the phosphorylation site level.



project:

Predops id:

○ DBupload: site conservation
○ check motif conservation
○ check sites' conservation -> p|reg.p|non-p
○ check sites' conservation (loops) -> p|reg.p|non-p
○ check sites' surrounding region windowlength: +- ☐

Figure 4.13: PHOSIDA administration tool

4.3 Data Integration and Data Selection of various Phosphoproteomes

As already pointed out in Chapter 4.1, the initial steps of the Knowledge Discovery in Databases Process (KDD) are selection and integration of data. Besides the required integration of public protein databases and other data that are relevant to derive a biological context (Chapter 4.2.1.3), the most essential datasets for my thesis were large-scale phosphosets generated in our laboratory (Chapter 4.6). Since we are very confident that our mass-spectrometry based technology assures very high accuracy at a false discovery rate lower than 1% for peptide identification, we rely primarily on high throughout data measured in our group. To assess the novelty of our data and to check the overlap with other datasets, we also integrated phosphorylation sites that are annotated in SwissProt. As our group is working on a variety of projects on different organisms, our data presents an optimal resource to gain insight into basic biological principles ranging from the activation of certain pathways upon different treatments to the derivation of general constraints on phosphorylation events.

4.4 Data Transformation of Preprocessed Data

According to the general ‘knowledge discovery in databases‘ (KDD) process data selection and integration is followed by data transformation (Ester, 2000) into a readable format for data mining. In general this includes standardizing values, deleting irrelevant attributes, or converting numerical values into discrete values. Since we ignored irrelevant attributes in the data integration process, the task of attribute deletion can be omitted.

In order to verify the overlaps of large-scale data between different projects, for example, joins of relevant relations are required. This is implemented by a single SQL statement demonstrating the strength of database techology. In order to avoid duplicate data, it does not make sense to create a ‘real’ relation for each join. The solution to this problem is given by the idea of ‘views’. If each ‘join’ statement yielded a real relation, data redundancy would increase. Avoiding data redundancy is the main purpose of virtual relations. Thus, to derive the number of shared identified peptides between different experiments, for instance, one has to join the two corresponding tables that store the peptides identified in a certain experiment. This results in the creation of a virtual relation that contains peptides, which are common in two given experiments (Figure 4.14).

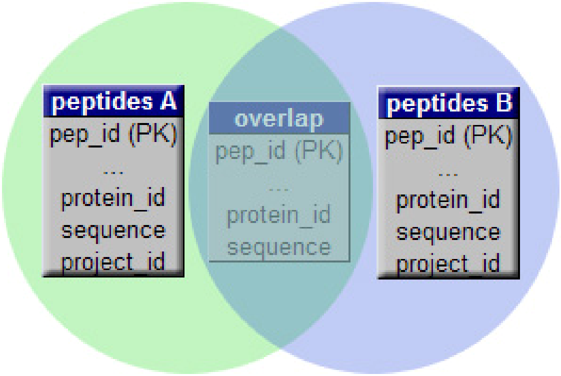


Figure 4.14: Database join of two relation instances ('peptides A', 'peptides B') containing detected peptides of a given project results into the creation of a virtual relation ('overlap')

To deduce the overlaps of phosphoproteomes between different organisms, we used the database relations that store evolutionary information such as homology between species for each integrated phosphorylated protein (Chapter 4.2.4). This simple way of dealing with data once stored in a consistent format once again underlines the benefit of databases.

As discussed in Chapter 4.5, the PHOSIDA database schema that stores non-redundant data such as 1:1 assignments between peptides and proteins (Chapter 4.2.1.1) is the one used for data mining of phosphoproteomes. On the one hand, we implemented mining tools in the language C# including statistical tests such as the χ^2-tests to check significant overrepresentations of matching kinase motifs. These self coded methods rely on a consistent database schema with categorical requirements for data storage and applications including mining tools to derive significant patterns from the managed data. Another prime example is the training of the support vector machine (Chapter 7). The implementation of organism-specific predictors requires consistent database storage to obtain positive instances, namely phosphorylation sites along with their surrounding sequence, as training sets. On the other hand, we used already established public mining tools that are freely available to the community (Chapter 4.5). The software Cytoscape (Maere et al., 2005; Shannon et al., 2003) determines whether certain GO categories describing cell components, functions, or biological processes are significantly overrepresented in a given set of proteins in comparison to the whole gene ontology annotation of a specified species. Although established mining methods are relatively easy to handle and user friendly, the required input files have to satisfy certain format specifications. Such format stringencies demand conversions of accession numbers, combinations of various annotation sets and further formatting. Hence the PHOSIDA administration tool (Chapter 4.2.6) includes various C# classes that enable the database administrator to create differently formatted files that are required as input for these mining tools (Figure 4.15). Underlying joins between relations storing protein annotations and relations containing phosphoproteome data and accession number conversions, for example, are executed automatically.

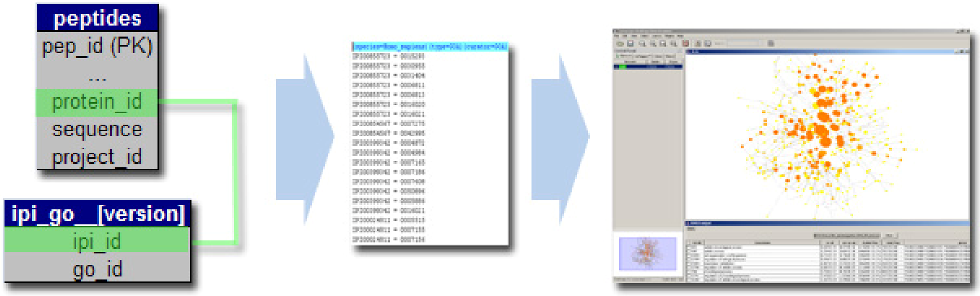


Figure 4.15: The PHOSIDA administration tool allows the conversion of accession numbers, joins on various annotation tables, or specified formatting of files required as input for certain mining methods such as Cytoscape

Besides the mining of integrated large scale data, the web application of PHOSIDA also demands an appropriate transformation of uploaded data. One prime example is the unification of different project specific subdatabases into one comprehensive organism specific database (Figure 4.16). Because of regularly updated versions of various databases, the spectrum-to-peptide assignments are often based on different database releases. For example, the identification of the human phosphoproteome identified upon epidermal growth factor stimulation (Chapter 4.6.1.1.1) was based on the human IPI database version 3.24, whereas the study of cell cycle dependent phosphorylation dynamics of kinases in human cells (Chapter 4.6.1.1.2) was based on IPI version 3.13. To unify the two subdatabases into one consistent database comprising both detected phosphoproteomes in human, it is indispensible to transfer the given data to a common database version. This is also required to determine the overlaps between large scale studies. Therefore, we reassigned the detected phosphorylated peptides to a more current database version, resulting in new peptide-to-protein assignments. Along with the amino acid sequence of a database entry the positions of identified phosphosites within the protein sequence can also change. In very few cases (less than 1%), identified peptide sequences cannot be reassigned to a more updated database release. Although the number of peptides that are not present in a more current database version is miniscule, this shows that databases do loose correct protein sequences between versions. With phosphoproteomic data assigned to a common database, it is possible to compare various phosphorylation changes observed under different treatments together using the PHOSIDA web page. The reassignment of peptides to an up-to-date database release was also essential to unite the different subdatabases annotated in the former version of the proteome database MAPU resulting in the new release of MAPU 2.0 (Chapter 5).

Finally, the reassignment of identified peptides to another database was also one of the main underlying principles of the genome annotation study using the genomic database EnsEMBL as for assigning peptides to gene transcript entries (Chapter 8).

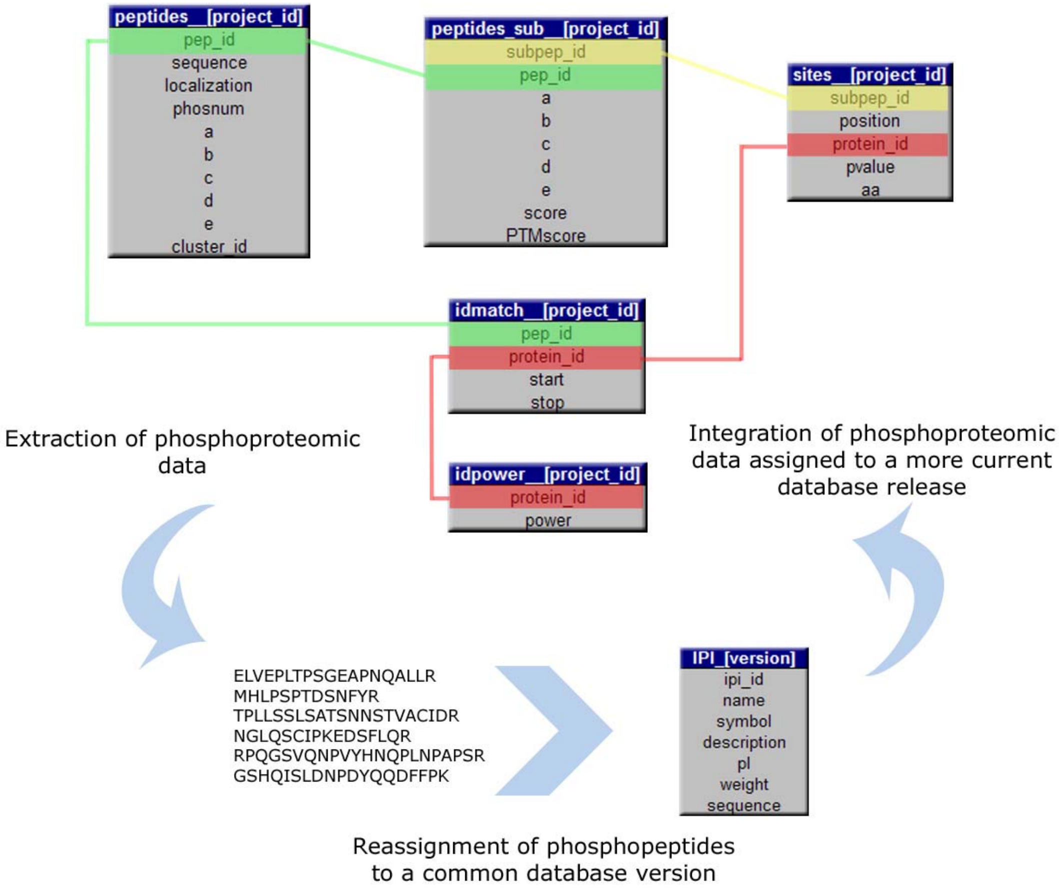


Figure 4.16: To unify various large scale data, the identified phosphopeptides have to be reassigned to a shared and more current database version

4.5 Data Mining in the Compiled Database

Data mining can be defined as the application of efficient algorithms that detect valid patterns in the data automatically (Han, 2000; Mitchell, 1997). There is nothing new about seeking patterns in data: Farmers seek patterns in crop growth, hunters seek patterns in animal migration behavior, and football managers seek weaknesses in the opponent team. However, we are overwhelmed with data. It has been estimated that the amount of data stored in the world's databases doubles every 20 months. Many decisions in our life are recorded in databases ranging from buying milk in the supermarket to ordering a 'Hed Kandi' music CD via the internet. The entrepreneur then tries to find opportunities deriving patterns from the customer's behavior and using this for business advantage. Association rules, for example, are used in 'market basket analysis'. On the basis of a priori algorithms, this data mining approach tries to find out which items are frequently bought together using the cash scanner records. The derived information then suggests certain shop design variants. The world wide web has also contributed decisively to the avalanche of information. Probably much of the entire human knowledge is stored in databases and illustrated in the internet. Another example is the field of biotechnology itself: As outlined in Chapter 1, high throughput technologies such as the microarrays measuring the expression levels of thousands of genes and mass spectrometry determining thousands of proteins quantitatively produce a vast amount of data. The same tendency can be observed in genome sequencing, as a new completely sequenced eukaryotic genome is in the news nearly every month. These trends underline the need for automated approaches that extract information and knowledge out of these raw diamonds (data). Regarding phosphorylation events in the cell, statistical tests can be used to determine significantly overrepresented proteins that contribute to a certain biological process. As an example, we used the Cytoscape Plugin BINGO (Maere et al., 2005) to find overrepresented gene ontology annotations including cell component localization in a given set of phosphorylated proteins. Another statistical method, named Motif-X (Schwartz and Gygi, 2005) and already introduced above, extracts significantly overrepresented consensus sequences from a set of sequences. Thus, this iterative statistical approach is suited to extract potential kinase motifs from a set of sequences surrounding determined phosphorylation sites. Furthermore, the Ka/Ks calculator also provides several statistical approaches ranging from the Nei and Gojobori calculation to the Goldman and Yang approach to derive the selective pressure on proteins (Zhang et al., 2006). Besides the application of such freely available mining tools, we designed various statistical mining methods implemented in C# and accessible via the PHOSIDA administration tool, as described above. These self implemented tests comprise the χ^2 tests to check the statistical significance of frequencies of identified phosphosites that match with a given kinase motif, and the Fisher test to test variances in conservation between phosphorylated residues and non-phosphorylated counterparts. Applied statistical tests are described in Chapter 4.5.1.

Clustering is another data mining method that we applied to our phosphoproteomic datasets (Chapter 4.5.2). The main idea of clustering is to divide a given set of data into several groups (clusters). In each cluster, assigned members should be as similar to each other as possible, whereas members of different clusters should be as dissimilar as possible. With quantitative data describing phosphorylation changes after treatment, the clustering approach was applied to distinguish phosphorylation sites that are immediately affected by a specified stimulus and those whose response follows in the latter parts of the flow providing negative feedback.

Support Vector Machines are part of the arsenal of ‘machine learning’ and they try to distinguish two given datasets according to their features, which are transformed in a high dimensional vector space, with each dimension reflecting a certain feature. Creation of a separating hyperplane the divides up the two given datasets and enables classification of new objects according to their position in the vector space relative to the hyperplane. This classification approach was used to predict phosphorylation sites (Chapter 4.5.3) and it is described in detail in Chapter 7.

4.5.1 Statistical Tests

The Chi Square Test is a very simple and basic method to check whether two given distributions are significantly different (independent) in a statistical sense. Thus, the χ^2-test is often used to estimate whether a given distribution correlates with the expected one. In the case of contingency tables with one degree of freedom, χ^2 is the difference between the expected frequency and the observed frequency squared and divided by the expected frequency:

$$\chi^2 = \frac{(\text{observed frequency - expected frequency})^2}{\text{expected frequency}}$$

The formula makes clear that a high χ^2 value reflects a high discrepancy from the expected frequency. Hence, this statistical approach can be applied to determine whether a given kinase motif matches significantly with the identified phosphorylation sites. To assess the number phosphosites matching an expected motif, we estimated the chance for each kinase motif to match with a given phosphosite according to the amino acid composition of the motif and the relative frequencies of each amino acid in the entire specified proteome. Another application is the proportion of homologous phosphoproteins to non-homologous phosphoproteins in comparison to their non-phosphorylated counterparts.

The χ^2-test is a simple test exemplifying mathematical methods, which are integrated in the PHOSIDA analysis pipeline among other statistical tests. It can be applied to any given phosphorylation site dataset via the PHOSIDA administration interface.

In contrast to the application of these statistical tests, the PHOSIDA analysis pipeline also comprises methods that create specified formatted files, which can be used as input for advanced statistical methods such as Motif-X (Schwartz and Gygi, 2005). This iterative statistical approach tries to derive consensus sequences that are significantly overrepresented in a given set of phosphorylation sites. A peptide data set is used for background probability calculations, and a set of detected phosphorylation sites along with their surrounding six amino acids is used as positive set. Both sets are converted into position weight matrices, where each cell presents the frequency of a certain amino acid on a specified position around the phosphosite. Based on the two resulting matrices, a binomial probability matrix is created reflecting the significance of each residue on a certain position. On the basis of a greedy recursive search, highly correlated position/residue pairs are then derived. After deleting all instances that match with the extracted motif, the method searches iteratively again until no significant consensus sequence can be found. We used this statistical method to extract potential kinase motifs from identified phosphorylation sites of various species.

Cytoscape is another open source bioinformatics software platform that we used to gain knowledge from the derived data. It is a platform for visualizing biological pathways and molecular interaction networks. We used the Java-based tool BiNGO (Biological Network Gene Ontology tool) to determine which gene ontology categories are statistically overrepresented in a set of identified phosphorylated proteins (Maere et al., 2005). BiNGO is implemented as a plugin for Cytoscape. Using various statistical tests such as the binomial test and the hypergeometric test, BiNGO tries to find significantly overrepresented functions, biological processes, and cellular component localizations comparing the given set of phosphorylated proteins with the whole proteome of the investigated species. Again, the application of BiNGO is directly connected to the PHOSIDA analysis pipeline providing all required input data in the specified formats.

4.5.2 Clustering

‘Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters’ (Witten, 2005).

The objects’ attribute values are usually transformed into a hyperdimensional feature space in order to calculate distance measures reflecting their dissimilarity. Figure 4.17 illustrates a two dimensional clustering resulting into three different groups (clusters). Each axis reflects a certain attribute value of a given object. The three different clusters are obvious by visual inspection. This visual grouping is highly intuitive because of the human brain’s highly evolved capacity for image and pattern recognition. Clustering analysis has been widely used in applications ranging from market analysis to microarray gene expression data analysis. The application of clustering to large scale datasets containing objects that can be described by multiple features has led to the design of a large number of different clustering approaches. Hierarchical methods, grid-based methods, density-based methods, or partitioning methods solve the problem of grouping given objects. Each approach has its advantages and disadvantages depending on the set of data.

We applied the Fuzzy C-Means (FCM) algorithm (Futschik and Carlisle, 2005), a partitioning method, in order to group the quantitative data reflecting phosphorylation changes upon treatment including certain stimuli. The main idea of k-Means clustering is to group a given set of objects into k clusters maximizing the cluster similarity measured in regard to the mean value of the objects in a cluster. It proceeds as follows: First, it randomly selects k objects, each representing a cluster’s center. The remaining objects are assigned to the most similar center out of k centers by the calculated feature distance. It then derives the new mean of each cluster iteratively, until no new cluster assignments can be calculated. FCM is a variant of the K-Means approach and allows membership of data elements in multiple clusters. Thus, FCM offers clustering tolerant to noise by variation of the fuzzification parameter m, which limits the contribution of ill-behaved profiles to the clustering process.

We applied the FCM approach to group profiles reflecting the phosphorylation dynamics upon EGF stimulation (Chapter 4.6.1.1.1). Consequently, each phosphorylated peptide could be assigned to a cluster representing upregulation or downregulation at a certain time point. We found optimal partitioning with six clusters and a fuzzification parameter of two. The corresponding resulting clusters of each identified phosphopeptide are also illustrated in the PHOSIDA online database (Figure 4.17 right panel).

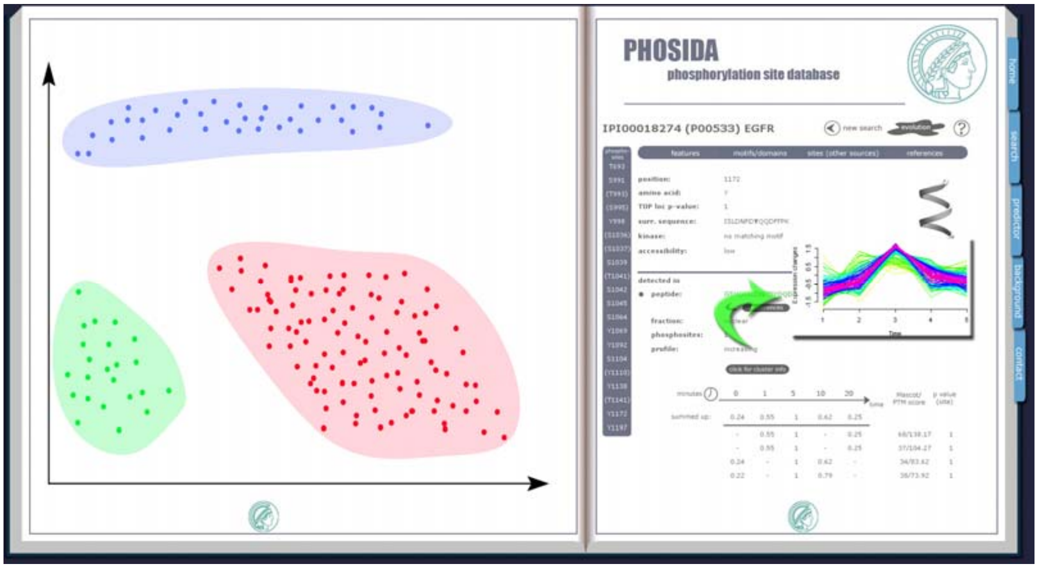


Figure 4.17: Clustering in PHOSIDA

Illustration of three clusters in a two dimensional feature space (left panel) and integration of clusters reflecting phosphorylation dynamics on the basis of quantitative data (right panel) in PHOSIDA.

4.5.3 Classification

Data classification is a two-step process. At first, a model is built on the basis of a set of objects. Each object has certain attribute values, which are transformed into a feature vector space. The objects' attributes are essential to determine dissimilarities between different samples by appropriate distance measures. As the category of each sample is known, the creation of a model describing the differences between classes is named 'supervised learning'. The training samples of known classes are used to build a model described by mathematical formula or decision trees, for instance. To evaluate the accuracy of the learning approach, one usually selects a subset of the training samples. The classifications of these test samples, which are substracted from the training set, are used to test the performance of classification decisions by the learned classifiers.

One usually takes 90% of the specified samples for training and 10% for testing. To avoid scewing the evaluation of the classification performance by random selections, one applies this performance test iteratively (n fold cross validation), where each step comprises another random selection of training and test samples. If the performance of the classification approach is acceptable, one can use the trained model to classify uncategorized future samples.

Hence, classification is very similar to prediction. However, classification is used to predict discrete or nominal values. The species assignment of given organisms is a typical classification problem and the answers are either “dog” or “cat”. In contrast, prediction can be viewed as the construction of a model to assess the (continuous) value ranges of an attribute that a given sample is likely to take on. However, classification and prediction are very similar in their purpose.

Both prediction and classification have numerous applications including selective marketing, medical diagnosis, and protein docking prediction.

We applied a classification approach in order to predict, whether a given protein’s residue is likely to be phosphorylated or not. As consensus sequences are the basis for kinase specific phosphorylation, the surrounding sequence of a given residue is obviously decisive to predict the likeliness to be phosphorylated. With our determined phosphorylation sites from large-scale phosphoproteomics, we trained a support vector machine to classify unlabeled samples (residues) into phosphorylated or unphosphorylated amino acids. The main principle of support vector machines is described in Chapter 7.

We also tried to find additional features besides the raw sequence that enhance to the accuracy of classification. For example, the phosphorylation process suggests that phosphorylation targets (residues) have to be accessible to kinases, thus solvent accessibility is a potential parameter to consider.

As this machine learning approach was applied to various datasets resulting in multiple trained models that enable prediction of phosphorylation sites in various species, its implementation and application is discussed in detail in Chapter 7.

4.6 Phosphoproteome Analysis

4.6.1 Basic Phosphoproteome Analysis

This section describes general features of different phosphoproteomes identified in various species. It comprises the distribution between individually identified phosphosites, the coverage of phosphorylated kinases, and the novelty of identified phosphorylation events. Additionally, the following chapters (4.6.1.1 – 4.6.1.5), which are divided according to the investigated organism, also describe project specific results such as phosphorylation changes at different stages of the cell cycle (Chapter 4.6.1.1.2).

4.6.1.1 *Homo sapiens*

The central organism, in which we are interested in, is our own species. One of the principal ideas of research is to learn more about our own organism and its biological functioning. The discovery of essential processes in our body not only leads to a better understanding of the basic biological principles, but also helps to prevent or cure diseases caused by malfunctions of biological processes. Traditionally, many experiments in the phosphorylation field were conduced outside of cells (*in-vitro*). However, *in-vitro* conditions might not reflect the real events in a living cell. Therefore, most of our experiments are based on *in-vivo* measurements of different cell lines. (Note that this is the biochemical definition of '*in-vivo*'. In biomedicine, *in-vivo* is frequently reserved for animal or human work.) We observed phosphorylation changes in HeLa cells, a human cell line, upon epidermal growth factor stimulation (Chapter 4.6.1.1.1) (Olsen et al., 2006). In addition, we also combined kinase-selective affinity purification with quantitative mass spectrometry to analyze the cell cycle regulation of protein kinases in the same human cell type (Chapter 4.6.1.1.2) (Daub et al., 2008).

4.6.1.1.1 Phosphorylation Dynamics induced by EGF stimulation

As outlined in Chapter 1, the cell constantly receives signals from its surroundings to which it has to respond appropriately. Growth factors, for example, are essential signals as they are capable of stimulating cellular differentiation and cellular proliferation and regulate a variety of cellular processes (Hunter, 2000; Pawson and Nash, 2003). In our study we used integrated phosphoproteomic technology combining phosphopeptide enrichment, high-accuracy identification, and stable isotope labelling by amino acids in cell culture (SILAC) (Ong et al.,

2002) to quantify changes in phosphopeptide levels and to investigate the global *in-vivo* phosphoproteome and its temporal dynamics upon growth-factor stimulation. The epidermal growth factor (EGF) acts by binding to the EGF receptor (EGFR) on the cell surface and stimulating its intrinsic protein-tyrosine kinase activity initiating a signal transduction cascade. This results in a number of biochemical changes ranging from cell proliferation to the increased expression of certain genes including the EGFR.

The application of triple-encoding SILAC for monitoring activation profiles, SCX and TiO_2 chromatography for phosphopeptide enrichment (Gruhler et al., 2005; Larsen et al., 2005), and high-accuracy mass spectrometric characterization allows the investigation of the phosphoproteome in considerable depth. The approach is completely generic for identification of phosphorylation events.

Serum-starved HeLa cells labelled with L-arginine and L-lysine, L-arginine-U-$^{13}C_6{}^{14}N_4$ and L-lysine-$^2H_4$, or L-arginine-U-$^{13}C_6$-$^{15}N_4$ and L-lysine-U-$^{13}C_6$-$^{15}N_2$ were treated with EGF for 0 min, 5 min, and 10 min. A second, identically labelled set of HeLa cells was treated with EGF for 1 min, 5 min, and 20 min. Then cells were combined, lysed and enzymatically digested. After the strong-cation exchange chromatography of digests, TiO_2 enrichment of phosphopeptides was performed (Figure 4.18).

Next, MS2 and MS3 spectra were merged into a single peak-list file and searched against the human IPI database. To establish a cutoff score threshold for a false-positive rate of less than one percent, we performed a MASCOT search against a concatenated target/decoy database (Elias et al., 2005) consisting of a combined forward and reverse version of the IPI human database including known nonhuman contaminants such as porcine trypsin. All spectra and all sequence assignments made by MASCOT (Perkins et al., 1999) were imported into MSQuant. The assignments of individual phosphosphorylation sites were automatically scored using the algorithm implemented in the PHOSIDA upload process (Chapter 3). The identified phosphorylation sites along with additional information including matching kinase motifs and structural constraints were then uploaded to the PHOSIDA database as described in Chapter 4.2.1. In addition, transformed profiles reflecting phosphorylation dynamics upon EGF stimulation were clustered as described in Chapter 4.5.2. We classified the derived clusters into ‘increasing’, ‘decreasing’ and ‘not changing’ and uploaded the clustering assignments to PHOSIDA.

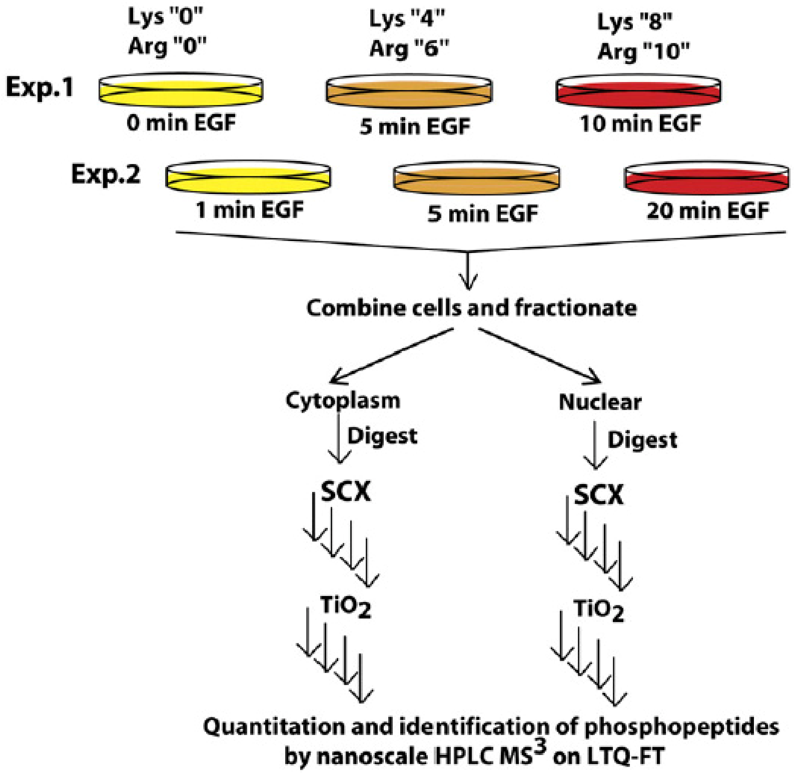


Figure 4.18: Quantitative and Time-Resolved Phosphoproteomics using SILAC

This quantitative, phosphosite-specific approach to detect phosphorylation dynamics upon EGF stimulus on the basis of SILAC-labelling yielded the identification of 6600 phosphorylation sites from 2244 proteins (Olsen et al., 2006).

We grouped potential phosphorylation sites into three categories depending on their PTM localization score and motifs. In the category with highest confidence in localization (class I), the given site had a localization probability for the phospho-group of at least 0.75. In class II, the localization probability is between 0.25 and 0.75, but these sites also had to match a known kinase motif. Class III sites had the same localization probabilities as class II but did not match any of the kinase motifs. According to this categorization, we determined 5674 class I sites, 2256 class II sites, and 1818 class III sites on mainly single phosphorylated peptides (Figure 4.19). In PHOSIDA, identified phosphorylation sites of a given protein of interest, which do not satisfy the class I criteria, are indicated in brackets (Figure 4.11).

We determined the distribution between individually identified sites to be 4901 pS, 670 pT, and 103 pY class I sites (Figure 4.19). Thus, our data set suggests that the distribution of pS, pT, and pY is 86.4%, 11.8%, and 1.8%, respectively.

The proportion of detected phosphoserines and phosphothreonines is in concordance with the one observed in previous studies (Hunter, 2000). However, the percentage of determined phosphotyrosines is much higher (1.8%) than reported previsoulsy (0.05%).

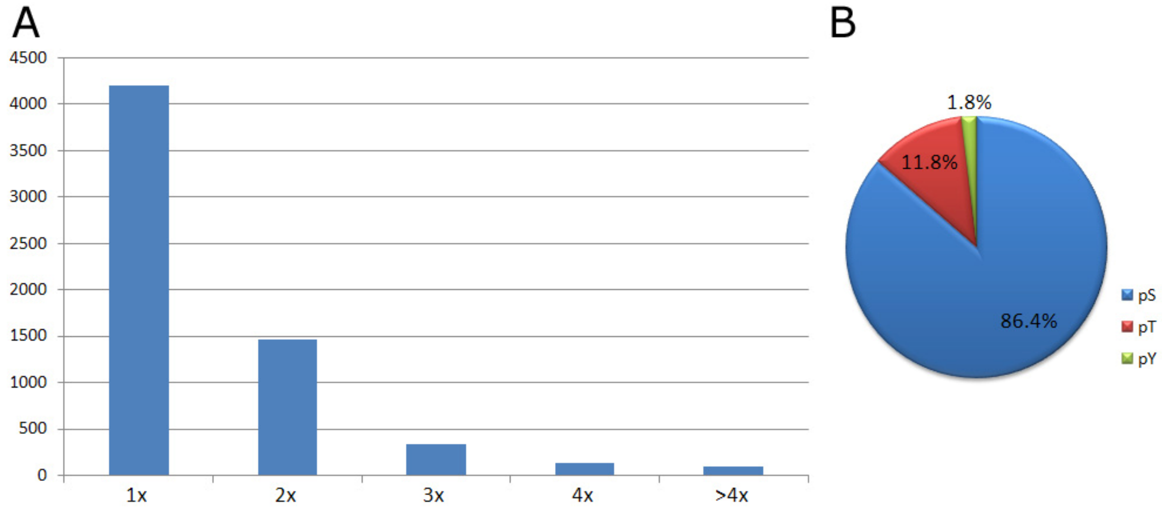
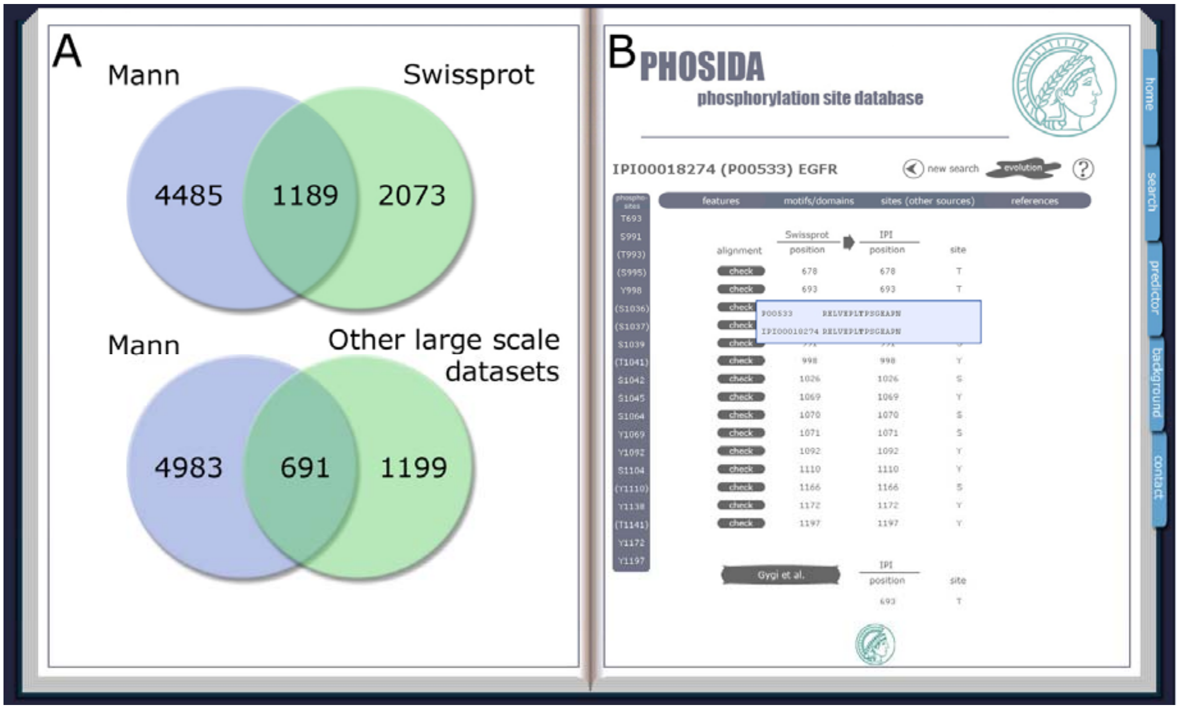


Figure 4.19: (A) Distribution of single, doubly, triply, quadruply and higher phosphorylated peptides. (B) Distribution of phosphorylation sites by amino acid

To determine the novelty of our dataset, we compared it with all annotated human phosphosites in the SwissProt database that were based on experimental data (3262 sites in version 48.0) and also included four previous phosphoproteomes in our analysis.

We found that more than 90% of our sites were novel with respect to SwissProt. In total, 691 (37%) out of 1890 phosphorylation sites from the four previous studies that could be mapped to IPI version 3.13 (Chapter 4.4) were also found in our study. PHOSIDA lists all sites determined from the other large-scale studies or annotated in SwissProt (accessible via the corresponding ‘sites from other sources’ button (Figure 4.20)). As discussed in Chapters 4.2.1.3 and 4.4, all SwissProt entries were mapped to the IPI database via BLAST, in order to ensure accurate comparisons.



IPI00018274 (P00533) EGFR

features | motifs/domains | sites (other sources) | references

phospho-sites: T693, S991, (T993), (S995), Y998, (S1036), (S1037), S1039, (T1041), S1042, S1045, S1064, Y1069, Y1092, S1104, (Y1110), Y1138, (T1141), Y1172, Y1197

alignment	Swissprot position	IPI position	site
check	678	678	T
check	693	693	T
check	[illegible]	[illegible]	[illegible]
check	[illegible]	[illegible]	[illegible]
check	[illegible]	[illegible]	[illegible]
check	998	998	Y
check	1026	1026	S
check	1069	1069	Y
check	1070	1070	S
check	1071	1071	S
check	1092	1092	Y
check	1110	1110	Y
check	1166	1166	S
check	1172	1172	Y
check	1197	1197	Y

```
P00533       RELVEPLTPSGEAPN
IPI00018274 RELVEPLTPSGEAPN
```

Gygi et al.	IPI position	site
	693	T

Figure 4.20: (A) Overlapping phosphorylation sites between our set and SwissProt (top) and the large scale datasets by Gygi and co-workers, Aebersold and co-workers, Stover et al., and Amanchy et al. (bottom); (B) PHOSIDA: Illustration of sites determined by other mass spectrometric approaches

In addition, we investigated the phosphorylation dynamics upon EGF stimulus: EGF signalling begins with activation of the EGF receptor and extends to a cascade of downstream kinases and other effector proteins. We derived four clusters with upregulated phosphopeptides and two with downregulated ones (Chapter 4.5.2). Cluster A, for example, embraced phosphorylation sites that can be classified as signal initiators involved in membrane-proximal signalling events and are enriched in phosphotyrosines. The resulting temporal cluster profiles are illustrated in Figure 4.21. As highlighted in Chapter 4.5.2, the online interface of the PHOSIDA database shows the corresponding clustering of each identified phosphopeptide.

Notably, around 77% of phosphorylated proteins contained at least two peptides that were detected to show different phosphorylation dynamics upon EGF stimulation on the basis of our clustering approach. This suggests that phosphoproteins serve as signal integrators. Interestingly, transcriptional regulators made up a large class of regulated proteins. We identified 26 phosphosphorylated transcription factors, with 33 novel phosphorylation sites showing diverse phosphorylation dynamics.

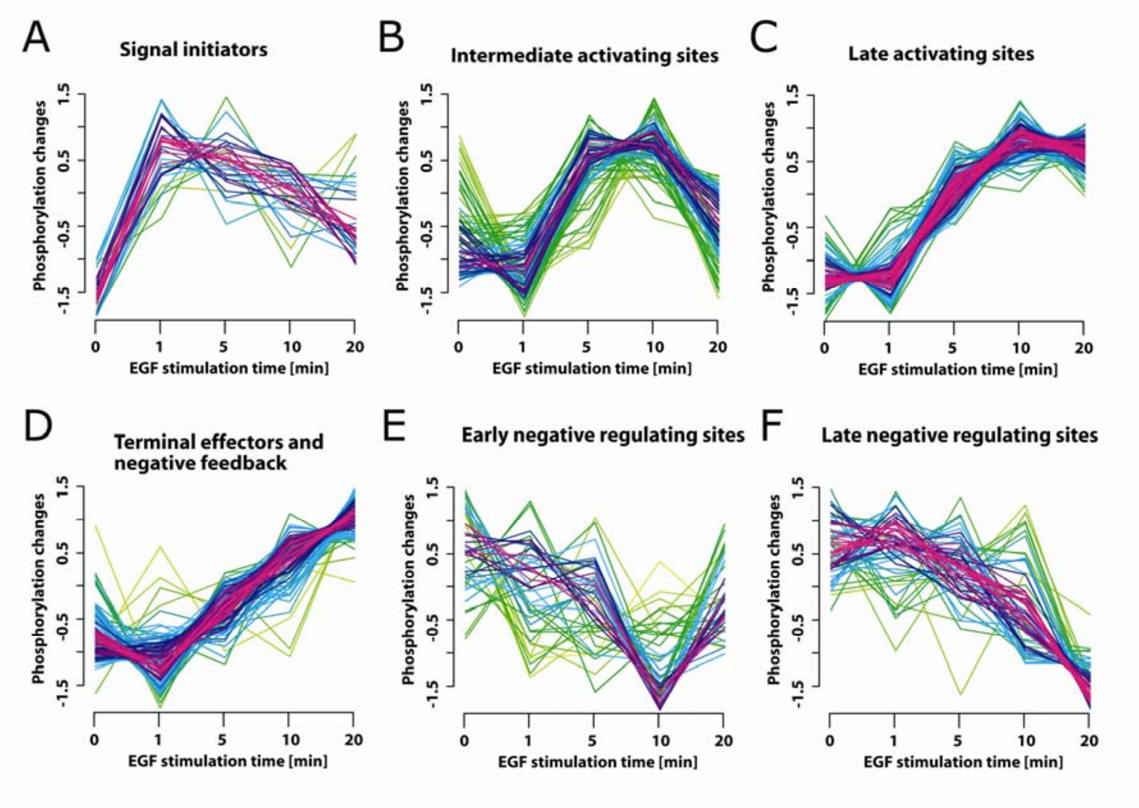


Figure 4.21: Clustering of dynamic phosphorylation profiles.

The y axis is $\log_{10}$ transformed and normalized. Each member (temporal profile) is color coded according to its membership value ranging from close membership (magenta) to distant membership (green) (Olsen et al., 2006).

4.6.1.1.2 Quantitation of the Kinome across the Cell Cycle

As highlighted in Chapter 1, protein kinases are essential regulators of cell signalling that modulate each other's functions and activities through site-specific phosphorylation events (Manning et al., 2002b; Shi et al., 2006). Their low abundances make it difficult to identify them from complete lysates. Thus, to increase the analytical sensitivity for protein kinases, they have to be enriched from total cell extracts prior to MS analysis. We applied an experimental strategy dedicated to enrich phosphorylated peptides from kinases to analyze protein kinase regulation in cell cycle progression (Daub et al., 2008).

The cell cycle comprises the progression of events leading to the replication of the eukaryotic cell. It can be divided into mitosis (M phase) including the nuclear and cytoplasmic division followed by interphase consisting of four phases: During the G_1 phase the cell starts to grow and synthesis of enzymes required for the next phasis – the S phase – is initated. In S-phase DNA is replicated while rates of protein synthesis are slow except for histones, which are needed for packaging of the DNA. During the G_2 phase, the cell prepares for mitosis by producing microtubules, for instance.

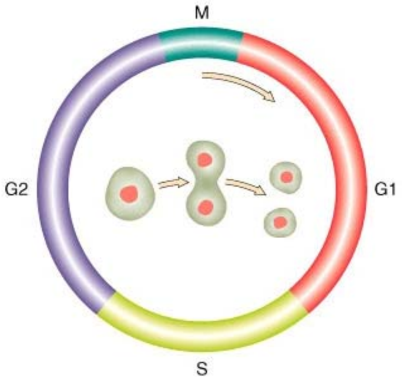


Figure 4.22: Schematic illustration of the cell cycle

In our study, we combined efficient kinase enrichment with quantitative mass spectrometry using SILAC. The basic experimental design is similar to the one applied for the identification of the human phosphoproteome upon EGF stimulation (Chapter 4.6.1.1.1), as it is also based on the same cell type and mass spectrometric technologies including SCX chromatography, TiO_2 peptide enrichment, and the SILAC labelling technique (Daub et al., 2008; Gruhler et al., 2005; Larsen et al., 2005; Ong et al., 2002). The statistical analysis of detected peptides and quantitation were also analogous.

Two populations of HeLa cells were quantitatively labelled by growing them in medium containing either normal arginine and lysine or their heavy isotopic variants. The cells were synchronized in early S phase by a double thymidine block in suspension culture. One of the populations was harvested at this point, whereas cells of the second population were released into a mitotic arrest. Then, pooled lysates from M and S phase cells were loaded onto a series of affinity columns displaying different immobilized kinase inhibitors with distinct kinase binding profiles to enrich protein kinases. We applied both gel electrophoresis followed by tryptic digestion on one of the kinase enriched subfractions and SCX chromatography to the kinase enriched fractions. The resulting peptide fractions were then subjects to phosphopeptide enrichment on TiO_2 beads. The combination of gel-based and gel-free MS separation strategies with phosphopeptide enrichment increases the overall number of detected phosphorylated peptides.

The statistical analysis of assigned peptide sequences and quantitation similar as described above (Chapter 4.6.1.1.1) and employed MASCOT, MSQuant, and various methods provided by the PHOSIDA administration tools (Chapter 4.2.6). This proteomic approach enabled us to quantify protein kinases from S and M phase arrested human cells and to elucidate cell-cycle dependent protein kinase regulation.

We uniquely identified and quantified phosphorylated peptides from 1377 proteins (Daub et al., 2008). The identified peptides harbored 3144 phosphorylation sites (83.5% pS, 14.2% pT, 2.3% pY) (Figure 4.23A). About 14% of all analyzed proteins were protein kinases: 219 different members of the human protein kinase superfamily were detected to be phosphorylated in our study. The phosphorylated kinases embraced 1007 phosphosites that could be assigned to serine (77.5%), threonine (17.2%), and tyrosine (5.3%) residues with high confidence. The vast majority of these detected phosphorylation sites in protein kinases have not been reported earlier.

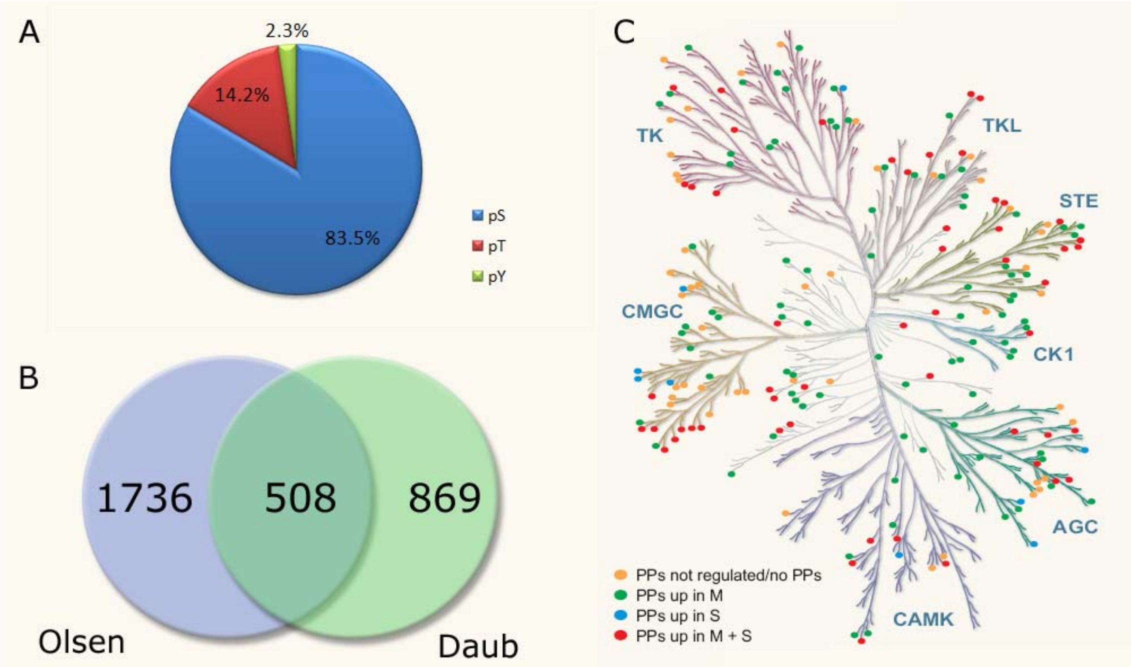
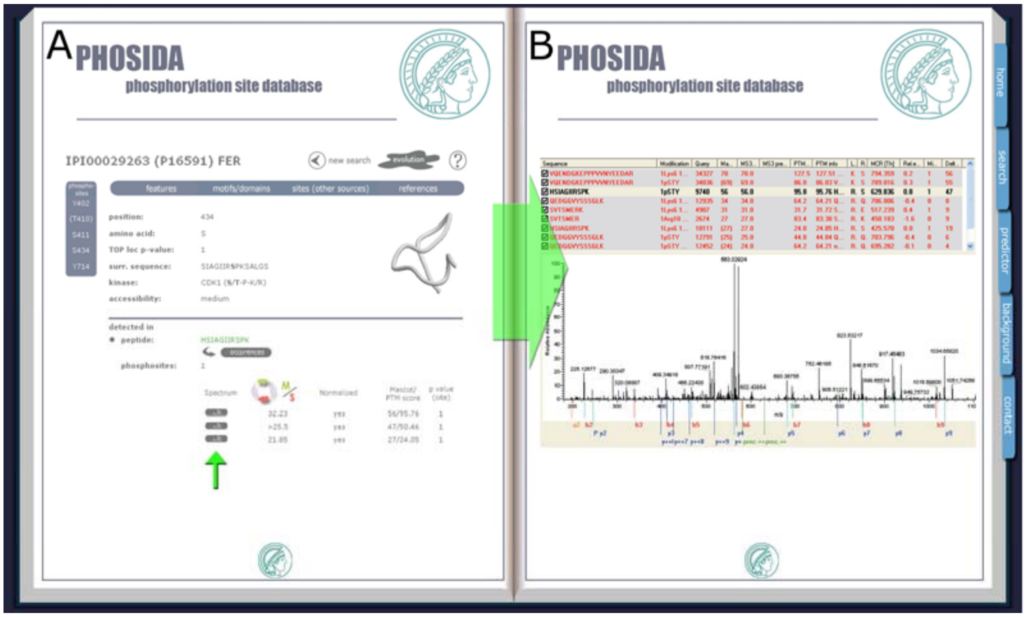


Figure 4.23: (A) Distribution of phosphorylation sites by amino acid. (B) Overlapping phosphoproteins between this study (green) and the previously reported study (blue) (Chapter 4.6.1.1.1). (C) Identified protein kinases marked in the kinome tree as illustrated in Chapter 1. The identification of at least two-fold differentially regulated phosphopeptides (PPs) in M versus S phase derived protein kinases is indicated by different colors.

We determined the overlap between this study and the investigation of the human phosphoproteome upon EGF stimulation (Chapter 4.6.1.1.1) and found that 508 (37%) out of 1377 phosphoproteins were also identified in the other large scale phosphoproteome analysis (Figure 4.23B). An even lower overlap was observed on the site level, as 546 (17%) out of 3144 phosphosites had also been measured in the EGF study. Interestingly, more than half of all kinase phosphopeptides were upregulated at least two-fold in mitotically arrested HeLa cells. In comparison, only 10% showed increased S phase abundance. At the protein level, regulation by factor two or more was observed for less than 10% of all protein kinases. If only SILAC phosphopeptide ratios are considered, apparent changes in phosphorylation could actually be due to a change in protein amount. Therefore, for each phosphorylated peptide the online application of PHOSIDA shows whether the given quantitative data describing the phosphorylation regulation during cell cycle could be normalized by the corresponding protein ratios or not. This information was stored as a special ‘feature’ attribute in the ‘peptides_sub’ relation (Chapter 4.2.1). Overall, 75% of all detected protein kinases contained at least one cell cycle regulated phosphopeptide (greather than two-fold upregulated in S phase or M phase). Strikingly, even for intensely studied cell cycle kinases including PLK1 and CDC2, a large number of new phosphorylation sites were found, demonstrating the high analytical sensitivity of our experimental approach. However, our study also covered a large number of other proteins that were quantitatively evaluated. For example, several regulatory kinase subunits such as different members of the cyclin family showed cell cycle dependent phosphorylation patterns.

In this study the spectra of measured phosphopeptides were also integrated into PHOSIDA because the applied MSQuant version enabled to create and automically save an image file of each spectrum. The corresponding file names can be derived from the MSQuant result files. The corresponding image file names are another ‘feature’ tuple of each entry (peptide object) stored in the ‘peptides_sub’ database relation.

The database storage of associated spectra and the visualization of these spectra enable web users to validate both identification and quantitation of each identified peptide. In PHOSIDA, the corresponding ‘spectrum’ buttons appear at the result page listing all detected peptides that contain the selected phosphorylation site (Figure 4.24). Besides the linkages to spectra, the cell cycle dependent phosphorylation regulation, Mascot scores, PTM scores, and further information are illustrated as discussed in Chapter 4.2.5.



IPI00029263 (P16591) FER

phospho-sites: Y402, (T410), S411, S434, Y714

position: 434
amino acid: S
TOP loc p-value: 1
surr. sequence: SIAGIIRSPKSALGS
kinase: CDK1 (S/T-P-K/R)
accessibility: medium

detected in
- peptide: HSIAGIIRSPK
  phosphosites: 1

Spectrum	M/S	Normalized	Mascot/ PTM score	p value (site)
	32.23	yes	56/95.76	1
	>25.5	yes	47/50.46	1
	21.85	yes	27/24.05	1

Figure 4.24: For large scale phosphorylation studies using MSQuant, PHOSIDA provides linkages (A) to an integrated online spectrum visualizer (B).

4.6.1.2 *Mus musculus*

The mouse is one of the most important model organisms in biology and medicine. It is by far the most commonly used laboratory mammal because of its small size, short reproduction time and short evolutionary distance to human. The genome sequence of this organism suggests a relatively close phylogenetic relationship with human (Bradley, 2002). Thus, it is a good model for a better understanding of basic mammalian biology, human disease and genome evolution. In our study, we investigated the phosphoproteome of the mouse liver using SILAC and high resolution mass spectrometry (Chapter 4.6.1.2.1) (Pan et al., 2008). In addition, we investigated whether our mass spectrometric methods for proteome and phosphoproteome analysis can also be applied to solid tumors (Chapter 4.6.1.2.2) (Zanivan et al., under review). As tumor model, we used mutant mice carrying skin melanomas. Conclusions drawn from these two experiments are applicable to other mammals including human.

4.6.1.2.1 Mouse Liver Phosphoproteome upon Phosphatase Inhibition

The liver is a multifunctional organ, involved in important metabolic functions, synthesis of blood plasma components and detoxification of xenobiotics among many other roles. Liver cancer, liver cirrhosis and insulin resistance of the liver are among the most common diseases associated with malfunctions of the liver. Many diseases are also caused by malfunctioning kinases that aparrently phosphorylate certain cellular substrates. The contrary mechanism, the dephosphorylation of substrates, is carried out by phosphatases (Chapter 1). Thus, phosphorylation regulating the activity of protein substrates is a reversible modification and its level is determined by the interplay of kinases and phosphatases, which add and remove the phosphogroups, respectively. Kinase-substrate specificity is often determined by the amino acid sequence surrounding the phosphosite (kinase motif). Therefore, the surrounding amino acid composition can be used to predict phosphorylation sites *in-silico* as described in Chapter 7. In contrast, phosphatases, especially serine/threonine phosphatases, more commonly rely on their targeting subunits to achieve specificity (Remenyi et al., 2006). Hence, phosphatases are more difficult to study than kinases resulting in a less comprehensive knowledge about phosphatases and their associated substrates. However, phosphatases play key roles in signalling and are frequently involved in diseases. Around 30 protein tyrosine phosphatases have been implicated in cancer, for example. The most common protein phosphatase inhibitors are vanadium compounds. Inhibiting the activity of phosphatases during cell lysis boosts the level of phosphorylation of their substrates.

In our study, we SILAC labelled the mouse Hep1-6 cell line, in which one population was treated with a mixture of phosphatase inhibitors (Pan et al., 2008). Thus, resulting quantitative data represented the increase of phosphorylation level caused by phosphatase inhibition on the basis of control versus phosphatase strategy.

We applied an in-depth, quantitative phosphoproteome analysis using high resolution MS-based proteomics to determine phosphorylation sites that are affected by phosphatase inhibition. The experimental set up is again similar to the one applied to human HeLa cells stimulated with EGF (Chapter 4.6.1.1.1). Trypsin digestion, SCX chromatography and phosphopeptide enrichment by TiO_2 beads as used as preliminary steps before MS measurements using LTQ-FT or LTQ-Orbitrap followed by the data integration into PHOSIDA (Gruhler et al., 2005; Larsen et al., 2005; Ong et al., 2002). The main difference is the phosphatase treatment of one population labelled with ‘heavy’ arginine and lysine (Arg10 and Lys8), whereas the other cell population was labelled with ‘light’ arginine and lysine and left untreated.

In total, we sequenced and identified 3430 phosphopeptides from 1808 phosphoproteins. Based on our Posttranslational Modification Scoring algorithm (Chapter 3), we identified 4253 phosphorylation sites with high confidence (class I sites). Out of these unambiguously identified phosphosites, 79.6% were serines, 9.3% were threonines, and 1.8% were tyrosines. The distribution of phosphorylated residues is similar to the one observed in human cells (Chapter 4.6.1.1). In addition, the frequency of singly and multiply phosphorylated peptides was also similar to the one found in human: The majority of phosphopeptides were singly phosphorylated (75%).

We identified 51 phosphorylated transcription factors, 121 phosphorylated protein kinases, and 28 phosphorylated phosphatases. In this project, we also determined the dynamic range of phosphopeptide detection. Figure 4.25A shows that the detected phosphopeptides follow a Gaussian intensity distribution on a logarithmic x-axis reflecting the intensity. It illustrates the numbers of untreated phosphorylated peptides that were measured within the range of a given intensity bin. It also shows that the distribution after phosphatase inhibitor treatment shifted by a factor of two relative to the untreated population. Interestingly, only 27% of the peptides where induced more than two-fold by the phosphatase treatment (Figure 4.25B). Some phosphorylation sites (8%) even decreased after phosphatase inhibitor treatment.

The most severe effects by phosphatase inhibition were observed for tyrosine phosphorylation sites. Overall, 70% of phosphotyrosines were upregulated at least two-fold. For phosphothreonine 41% of the sites were upregulated by this factor and for phosphoserine the number is 26%. This is a surprisingly low number considering that the investigated inhibitors are thought to block most phosphatase activity.

Again, the implemented cross reference between the IPI database used for the assignments of spectra to peptide sequences and the annotation rich SwissProt database (Chapter 4.2.1.3) made it possible to determine the overlap of phosphosites identified in our study and phosphosites reported in SwissProt. In total, we found 864 phosphoproteins in our study that have already been shown to be phosphorylated according to SwissProt annotation (Figure 4.25C). Therefore, 169 proteins were shown to be phosphorylated by our study for the first time. This is a striking overlap, given that there are more than 50000 protein entries in the mouse IPI database. However, our dataset has substantial novelty on the site level, since more than half (1428) of 2590 class I sites, whose assigned proteins are annotated in SwissProt, are novel (Figure 4.25D).

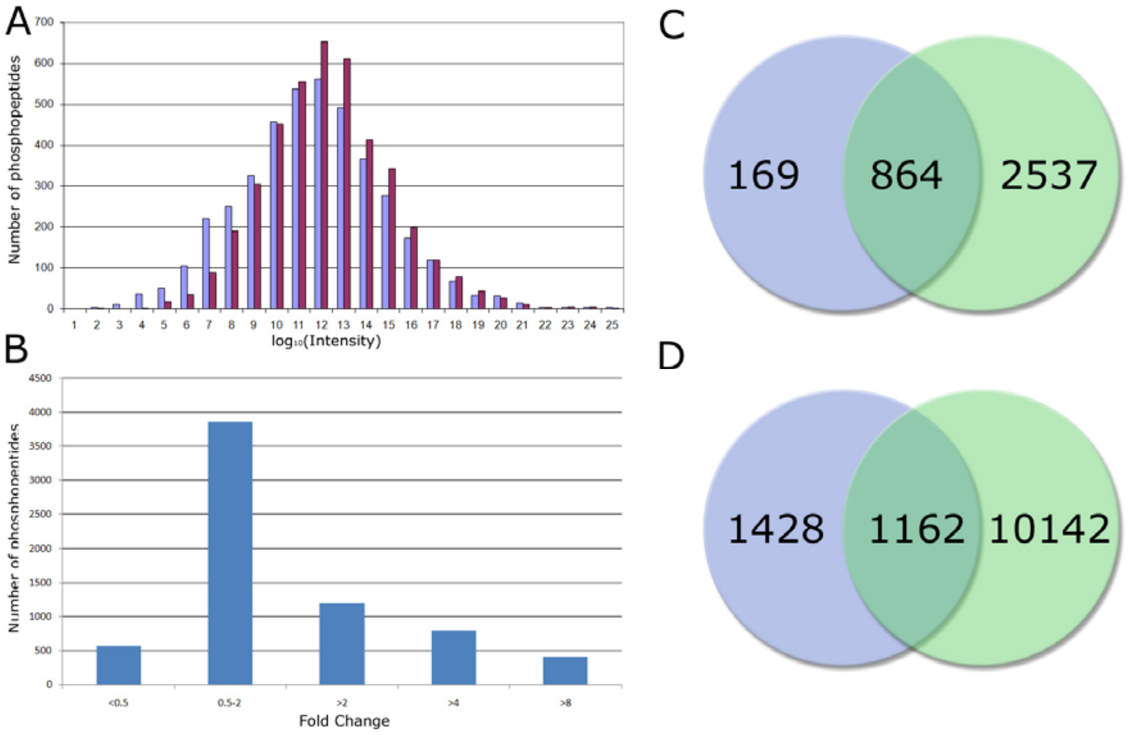


Figure 4.25: Mouse liver phosphoproteome

(A) Number of phosphopeptides identified at certain intensity bins. Both phosphopeptides from the untreated population (light) and phosphopeptides from the phosphatase inhibited population (dark) are Gaussian distributed on a log x scale. (B) Number of phosphopeptides that show a given intensity change after phosphatase inhibition. (C) Overlap of phosphorylated proteins found in this study (blue) and phosphorylated proteins annotated in SwissProt. (D) Overlap of phosphosites identified in our analysis (blue) and phosphosites reported in SwissProt

4.6.1.2.2 Solid Tumor Phosphoproteome

Cancer is often caused by a disregulation of signals and tumors are characterized by multiple aberrations in their signalling machinery (Dhillon et al., 2007; Hanahan and Weinberg, 2000). This results in increased replicative potential, decreased apoptosis, growth factor indepence and metastatic capability. Thus, kinases and phosphatases - as key regulators in signalling - play prominent roles in diseases such as tumor development. It suggests the presence of specific underlying phosphorylation patterns during tumor development. Understanding the molecular mechanisms including phosphorylation events that cause deregulated signalling would help in understanding many aspects of tumorigenesis.

In our study, we applied MS-based proteomics analysis to identify the phosphoproteome as well as the proteome of solid tumors in mice (Zanivan et al., in press). As tumor model we used TG3 mutant mice carrying skin melanomas. These mice ectopically express Grm1, a glutamate receptor, which results in the constitutive activation of the Erk pathway. Consequently, they develop melanomas several months after birth.

For the phosphoproteome analysis we enriched phosphopeptides with strong cation exchange chromatography (SCX) followed by titansphere enrichment or with TiO_2 only (Gruhler et al., 2005; Larsen et al., 2005; Ong et al., 2002). Digested proteins were analyzed using an LTQ-Orbitrap mass spectrometer. Using Mascot, MaxQuant and the PHOSIDA environment (Chapter 4.2.6), the identified phosphorylation sites were uploaded to the PHOSIDA database. In addition, the proteomic data were uploaded to the MAPU database (Chapter 5). The workflow was similar to the EGF signalling study described above but no SILAC quantitation was performed. Furthermore, we also used the detected phosphorylation sites to train a mouse specific phosphosite predictor (Chapter 7).

Because of the close evolutionary relationship to human and because of the fact that Grm1 is also expressed in a subset of human melanomas, this study is also of clinical interest. For the first time, it reveals the phosphorylation pattern of a solid tumor and therefore it might extend our knowledge of underlying deregulated signalling in cancer.

The main purpose of this study was to investigate if advances in instrumentation, algorithms and preparation techniques applied to the other studies make the solid tumor phosphoproteome amenable to such an analysis. Indeed, the analysis of the phosphoproteome of the tumour tissue of TG3 mice proves this point: Combining data from SCX-TiO_2 enrichment and TiO_2 chromatography led to the identification of 5250 phosphopeptides, belonging to 2250 proteins. In total, we identified 5698 class I phosphorylation sites (90% phosphoserines, 9% phosphothreonines, 1% phosphotyrosines). These relative abundances are similar to the ones observed after phosphatase inhibition (Chapter 4.6.1.2.1) and the ones reported for a human cancer cell line (Chapter 4.6.1.1). We also compared the identified phosphoproteome with published gene expression profiling studies of melanoma (Hoek, 2007). Many of these genes were found in our melanoma proteome and phosphoproteome.

The characterization of the functional impact of the phosphorylated proteins was performed by gene ontology analysis. Using Cytoscape, the results from the calculation of over- and underrepresented gene ontology categories describing molecular functions, biological processes and cellular component localization is described in Chapter 4.6.2.

Furthermore, we found evidence for the constitutive activation of the MAPK and mTor signalling pathways in melanoma. It has been reported that these pathways play major roles in the development and progression of melanoma (Lasithiotakis et al., 2008; Meier et al., 2005).

We found phosphorylation sites from the mTor pathway, which regulates protein translation through the phosphorylation of p70 S6 kinase 1 (p70S6K), and eIF-4E binding protein (4EBP1), for example. In addition, we found Tsc2 phosphorylated at Serine 981, which is a target of Akt and induces the translocation of Tsc2 to the cytosol (Dan et al., 2002). This mechanism is thought to be responsible for mTor pathway activation (Cai et al., 2006).

4.6.1.3 *Drosophila melanogaster*

Martin Brookes mentioned in his book about *Drosophila* that ‘a glass of milk and a piece of rotting banana is enough in order to jolly 200 fruitflies along for 14 days’ (Brookes, 2002). This statement describes the relatively easy treatment of flies in the laboratory. Robust viability in laboratory environments and a short generation time of about two weeks along with a lifetime of 50-60 days are substantial arguments for categorizing *Drosophila melanogaster* as a ‘model organism’. Other important arguments for the *Drosophila* modle are the fact that its genome is compact (four chromosomes) and completely sequenced as well as its homology to humans: Around 60% of fly genes show parallels in the human genome (Adams et al., 2000). Many conclusions drawn from observations based on fly cells including those gained from cell lines have turned out to be also valid for human.

Large-scale site specific *Drosophila* phosphoproteome studies were performed in Kc cells by Aebersold et al. (Bodenmiller et al., 2007). Gygi et al. characterized the phosphoproteome of fly embryos (Zhai et al., 2008). Both studies report more than 10000 identified phosphorylation sites indicating that the size of the fly phosphoproteome is comparable to the human phosphoproteome (Chapter 4.6.1.1).

However, the above studies were purely qualitative. Here we applied a functional quantitative phosphoproteomic study in *Drosophila* elucidating the biological impact of the protein tyrosine phosphatase Ptp61F on the fly phosphoproteome using RNA interference. We also established a high quality basal fly phosphoproteome in the process.

To characterize the endogenous phosphorylation sites of the embryonic Drosophila SL2 cell line, in the following named ‘basal phosphoproteome’, we applied SILAC-based quantitative proteomics, where the ‘heavy’ cell population was treated with a phosphatase inhibitor mix while the ‘light’ population was kept untreated. The SILAC-based quantitative strategy comparing endogenous phosphorylation to phosphatase inhibitor enhanced phosphorylation helps in triggering the identification of very low abundant phosphosites that are ‘upregulated’ in response to the phosphatase inhibition. Furthermore, it effects a better identification, as each peptide appears in pairs (heavy and light).

Drosophila is also a very suitable model system for loss-off function studies by RNA interference (RNAi) because of the highly efficient and penetrant RNAi, fewer 'off target' effects compared to mammalian models, as well as the lower degree of functional redundancy compared to higher vertebrates. Ptp61F is an ortholog to the human phosphatase Ptb1b, which is thought to be involved in type 2 diabetes, obesity and cancer. Thus, we extended our quantitative phosphoproteomics approach with RNA interference for the functional analysis of the perturbation caused by Ptp61F knock down. To normalize for expression changes and to elucidate proteomic changes, we also analyzed the proteome after RNAi treatment.

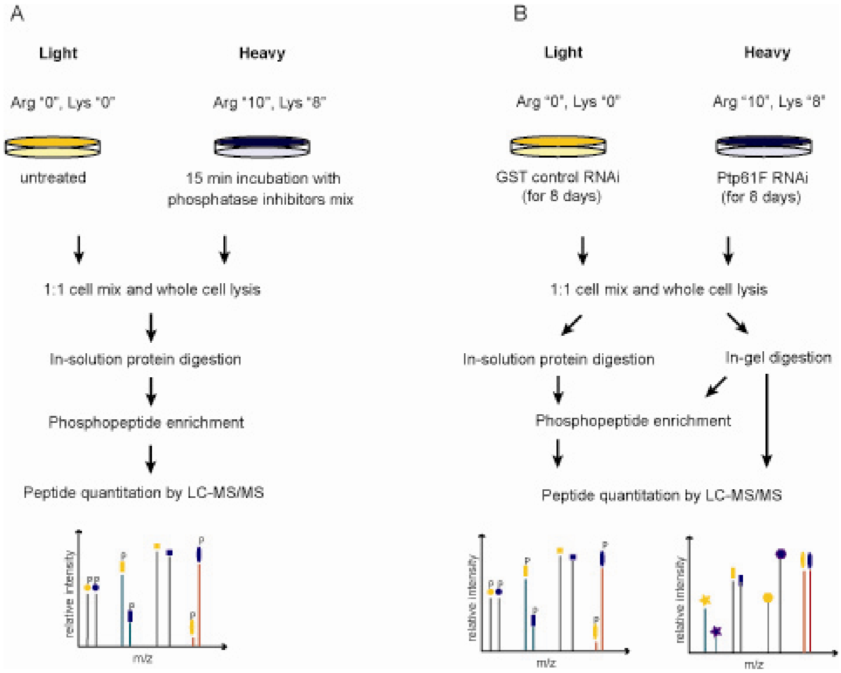

A
Light
Heavy
Arg "0", Lys "0"
Arg "10", Lys "8"
untreated
15 min incubation with phosphatase inhibitors mix
1:1 cell mix and whole cell lysis
In-solution protein digestion
Phosphopeptide enrichment
Peptide quantitation by LC-MS/MS
relative intensity
m/z
B
Light
Heavy
Arg "0", Lys "0"
Arg "10", Lys "8"
GST control RNAi (for 8 days)
Ptp61F RNAi (for 8 days)
1:1 cell mix and whole cell lysis
In-solution protein digestion
In-gel digestion
Phosphopeptide enrichment
Peptide quantitation by LC-MS/MS
relative intensity
m/z
relative intensity
m/z

Figure 4.26: Overview of the analytical workflow used in the study to detect the *Drosophila* phosphoproteome upon phosphatase inhibition (A) and Ptp61F RNAi (B)

The experimental strategy was based on the one applied to human cells (Chapter 4.6.1.1.1) from trypsin digestion to phosphopeptide enrichment by SCX/TiO_2 chromatography and high resolution MS (Figure 4.26). The resulting large-scale data were uploaded to PHOSIDA for further data mining and transformed to the PHOSIDA online database scheme. In addition, we uploaded the perturbated proteome to the MAPU database (Chapter 5).

The application of the SILAC-based phosphoproteomics approach on one heavy cell population treated with a phosphatase inhibitor cocktail while the light population was kept untreated yielded the identification of 6752 phosphorylation sites on 1928 proteins. The percentage of determined tyrosine phosphorylation sites increased to 4.1%. The extension of the experimental design with RNAi interference of the phosphatase Ptp61F led to the identification of 6516 phosphorylation sites on 1952 proteins. We found that the proportion of phosphotyrosines was 1.5% in that experiment. Importantly, phosphorylation dynamics could be normalized by detected proteome changes. Figure 4.27 depicts the plot of normalized phosphorylation changes upon phosphatase knockdown. Phosphorylation sites that are not affected by the treatment are highlighted in gray. Phosphorylation sites that significantly respond to the Ptb61F knockdown are marked in green. The phosphorylation pattern of STAT92E, a known target of the phosphatase, was found to be significantly affected, providing a positive control. The corresponding SILAC pair of the associated phosphorylated peptide is illustrated in Figure 4.27.

Overall, 9749 phosphorylation sites on 2285 proteins were determined with high confidence. The overlaps between our dataset and the large-scale studies by Gygi et al. (Zhai et al., 2008) and Aebersold et al. (Bodenmiller et al., 2007) were 1506 (65.9%) phosphoproteins and 1719 (75.2%) phosphoproteins respectively. In total, 1274 phosphorylated proteins were identified in all three studies, whereas 334 phosphoproteins were exclusively determined in our approach. On the site level, we detected 4691 (48.2%) novel phosphorylation sites, whereas 5051 phosphorylated sites were already covered by the other two studies.

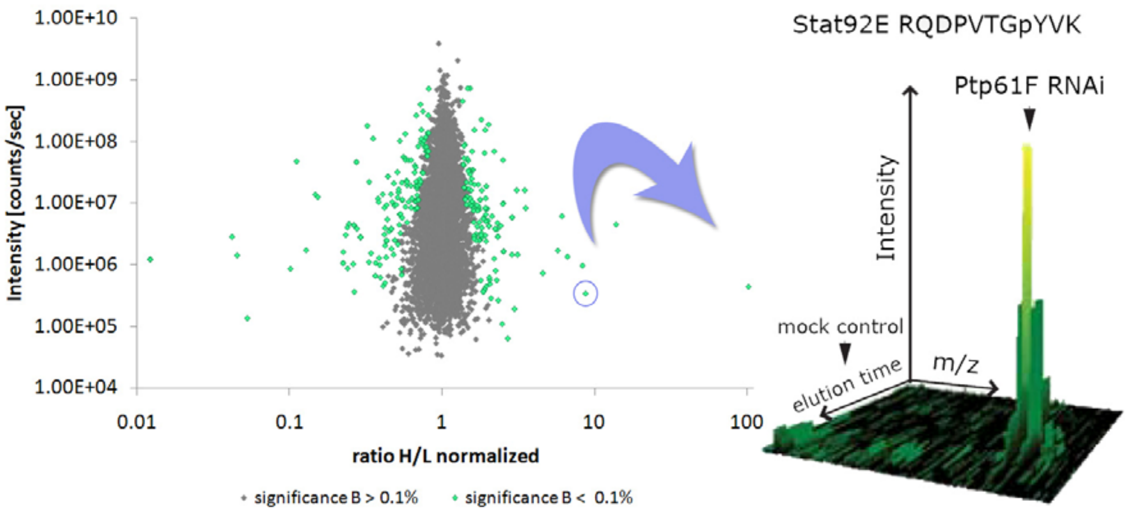


Figure 4.27: Phosphorylation site changes upon phosphatase knockdown

Phosphorylation site changes are normalized by proteome changes and plotted against the measured intensity (left panel). Statistically unaffected phosphorylation sites are indicated in gray, whereas significant phosphorylation changes are marked in green. Significantly phosphorylation up-regulation was observed for Stat92E, for example. The corresponding three-dimensionally represented spectrum is illustrated on the right panel.

4.6.1.4 *Saccharomyces cerevisiae*

Yeast is another widely used model organism and has an important role in industry being involved in bread fermentation and ethanol production. Its genome was the first eukaryotic one to be completely sequenced (Cherry et al., 1998; Williams, 1996). Soon, it became obvious that yeast and human share a substantial number of homologous proteins. Thus, the yeast organism is often used to gain biological insight in the basic functioning of the eukaryotic cell.

Protein phosphorylation is ubiquitous in all eukaryotes including yeast (Ptacek et al., 2005). The application of MS-based proteomics using immobilized metal-affinity chromatography (IMAC) for phosphopeptide enrichment has already proven successful in large-scale yeast phosphoproteomics (Ficarro et al., 2002). We applied the SILAC technology (Ong et al., 2002) to two cell populations with normal or heavy forms of both arginine and lysine (De Godoy et al., under review). After lysis, 1:1 mixing and trypsin digestion, we applied two tinanium dioxide chromatography (TiO_2) strategies to enrich phosphorylated peptides. In this study, the use of SILAC provides a more accurate identification of phosphopeptides, as all peptides are detected by the mass spectrometer as characteristic pairs. The data was searched against a decoy database for estimation of the false positive rate, and peptide identification and validation were based on the MaxQuant software. Again the experimental design is roughly based on the protocol established on the basis of the identification of the human phosphoproteome (Chapter 4.6.1.1.1).

We identified alarge set of *in-vivo* phosphorylation in yeast covering even low abundant transcription factors and a representative set of the kinome (Hunter and Plowman, 1997). This data allows to draw general conclusions about phosphorylation in 'lower' eukaryotes regarding structural constraints, subcellular localization, or the occurrence of kinase motifs (Chapters 4.6.2 – 4.6.3). The evolutionary conservation between the yeast phosphoproteome and phosphoproteomes of 'higher' eukaryotes such as fly (4.6.1.3), mouse (4.6.1.2), and human (4.6.1.1) is especially interesting and is the main subject of Chapter 9.

The 1:1 SILAC labelling of yeast cells combined with titanium dioxide chromatrography and strong cation exchange chromatography yielded the identification of 4160 phosphorylation sites mapping to 1192 proteins (De Godoy et al., under review). As in the other studies on phosphoproteomes of higher eukaryotes (Chapters 4.6.1.1 – 4.6.1.3), we determined phosphorylation events on proteins with less than 1% false positive rate at both peptide and protein levels. The unambiguously identified phosphorylation sites correspond to 3469 phosphoserines (83.2%), 635 phosphothreonines (15.2%) and 66 phosphotyrosines (1.6%). We found that around 500 phosphorylated proteins and 3000 phosphorylation sites detected in our study are novel compared to SwissProt, which nearly doubles the number of sites previously reported. Using the gene ontology annotations integrated in PHOSIDA, we found phosphorylation event on about one third of known yeast transcription factors including low abundant ones.

A further objective of this study was to examine the yeast kinome (Hunter and Plowman, 1997). To retrieve known protein kinases from our phosphorylation set, we used KinBase (Manning et al., 2002b), an open access database that includes kinases from vertebtrates, invertebrates, and unicellular organisms such as yeast. Overall, 45 kinases were revealed to be phosphorylated in our set. Since KinBase reports 124 yeast kinases in total, our set covered all main kinase families representatively (Figure 4.28). The identified phosphorylated kinome includes kinases such as AKT and CKI, which are conserved throughout eukaryotes, but also yeast-specific kinases such as RIM15 and RAN. Besides protein kinases, we also identified various phosphatases and cyclins, which are listed by KinBase.

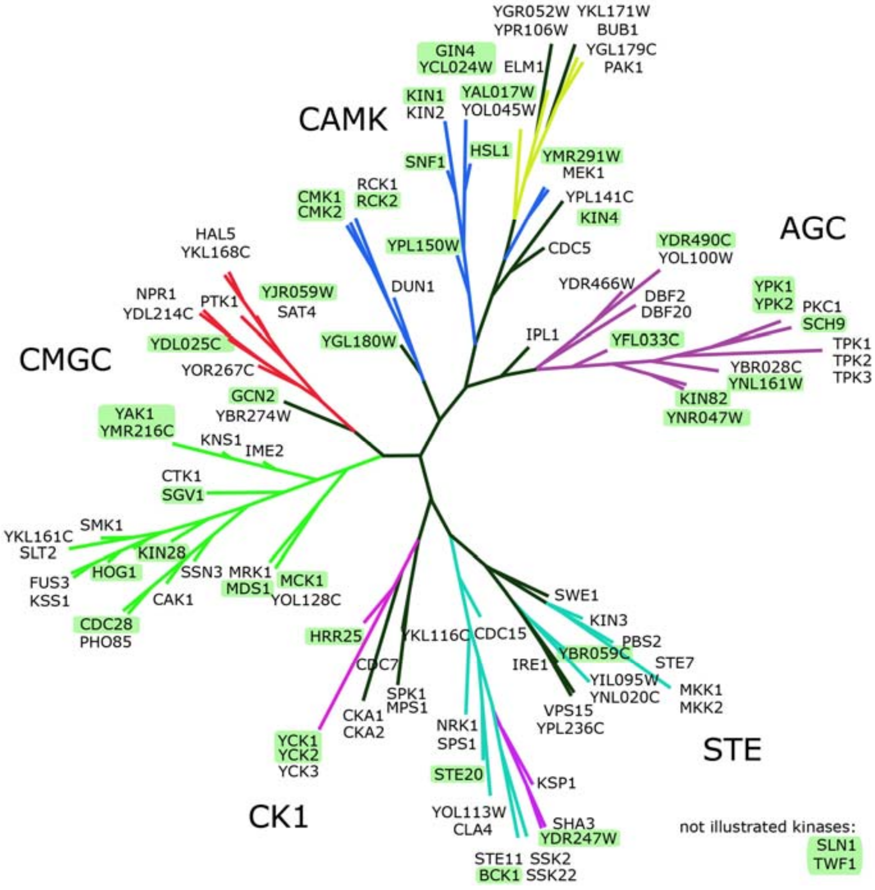


Figure 4.28: Yeast kinome tree

Kinases that we found to be phosphorylated are indicated in green.

4.6.1.5 Prokaryotic Phosphoproteomes

Protein phosphorylation on serine, threonine, and tyrosine is well established as a key regulatory posttranslational modification in eukaryotes, but little is known about its extent and function in prokaryotes. For some time the field of protein phosphorylation held the view that eukaryotes use serine/threonine/tyrosine phosphorylation, whereas bacteria instead use histidine and aspartate phosphorylation, mainly in their two-component systems. However, accumulating evidence has shown that serine/threonine/tyrosine phosphorylation also plays a vital role in bacteria (Deutscher and Saier, 2005). Bacteria possess both kinases and phosphatases that show homologous counterparts in eukaryotes (Kennelly, 2002), but also kinases that lack of any homology throughout the other domains of life, which supports the idea of the occurrence of prokaryotic specific phosphorylation (Mijakovic et al., 2005). As the application of MS-based proteomics to various eukaryotes has proven to be suited for the detection of thousands of phosphorylation events in the eukaryotic cell (Chapters 4.6.1.1 - 4.6.1.4), we used this technology to obtain site-specific, *in-vivo* phosphoproteomes of *Bacillus subtilis*, *Escherichia coli*, and *Lactococcus lactis* (Figure 4.29) (Macek et al., 2008; Macek et al., 2007; Soufi et al., 2008). We even determined the phosphoproteome of *Halobacterium salinarium*, a member of the third domain of life (archaea) (Aivaliotis et al., under review).

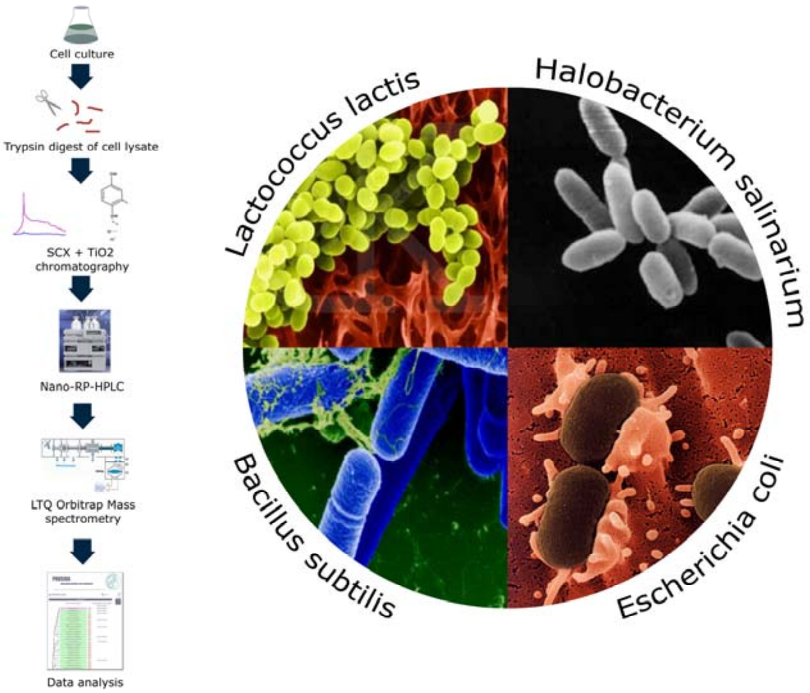


Figure 4.29: Overview of the analytical workflow used to detect prokaryotic phosphoproteomes

Trypsin digestion of the whole cell lysate was followed by enrichment of phosphopeptides using two stages of chromatography (SCX and TiO_2). Phosphopeptides were separated on nano-HPLC, mass-measured and fragmented in the LTQ-Orbitrap mass spectrometer

The first described prokaryotic phosphoproteome was the one of *Bacillus subtilis*, a model Gram-positive bacterium (Macek et al., 2007). In the past, investigation of *B. subtilis* has already made significant contribution to the understanding of fundamental processes such as carbon catabolite regulation and sporulation. In addition, it represents the most intensely studied bacterium regarding phosphorylation. However, before our study, a mere eight phosphorylated proteins have been identified in *B.subtilis* (Wurgler-Murphy et al., 2004). Thus, we intended to detect a more comprehensive set of phosphorylation events in this bacterium. Furthermore, we wanted to investigate a representative member of Gram-negative bacteria (Macek et al., 2008), which can be pathogenic. Their pathogenicity is usually associated with lipopolysaccharides that are constituent parts of the cell wall. The most prominent member of Gram-negative prokaryotes is *Escherichia coli*, which has been the model system that spawned molecular biology. It is commonly found in the intestine of warm-blooded animals, but it is also capable of surviving outside the body. Some strains can also cause food poisoning in humans. The occurrence of phosphorylation in *E.coli* has already been shown: In two-dimensional gel experiments with protein extracts labeled with radioactive phosphorus, more than one hundred phosphorylated protein spots were observed (Cortay et al., 1986). However, most of them were never identified. Nevertheless, two Serine/Threonine kinases, namely the isocitrate dehydrogenase kinase/phosphatase (Oudot et al., 2001) and the YihE kinase (Zheng et al., 2007), have been well characterized. Two tyrosine kinases, Wzc and Etk, point to the possibility of tyrosine phosphorylation in *E.coli* (Grangeasse et al., 2007). The global and site-specific analysis of the *E.coli* phosphoproteome also established that serine/threonine phosphorylation is a general regulatory process and not restricted to eukaryotes.

Furthermore, we investigated the phosphoproteome of the Gram-positive non pathogenic bacterium *Lactococcus lactis* (Soufi et al., 2008), a representative of lactic acid bacteria. *L.lactis* as starter culture is used in the production of more than ten million tons of cheese and it thus crucial in the dairy industry. It is also important for the proper digestion of lactose in human. Thus, we decided to extend our analysis on phosphorylation in bacteria to *Lactococcus lactis*. A phosphorylated serine on position 46 of the phosphocarrier protein HPr presents the only phosphorylation site that had been reported so far (Monedero et al., 2001). There are two known Serine/Threonine kinases, namely the HPr kinase and the eukaryot-like kinase PknB, both lacking known substrates (Bolotin et al., 2001; Monedero et al., 2001). The elucidation of site-specific phosphorylation might be conducive to the optimization for desired functions of this organism in industry and to gain more insight into its physiology.

Finally, we investigated if posttranslational modification by covalent phosphorylation is also found in archaea. Thus, we measured the phosphoproteome of *Halobacterium salinarium* (Aivaliotis et al., under review). This obligate aerobic member of archaea is a halophilic marine Gram-negative organism.

In our study we performed a global, gel-free, and site-specific analysis of the four prokaryotic phosphoproteome using high accuracy mass spectrometry in combination with biochemical enrichment of phosphopeptides from digested cell lysates. Apart from the SILAC labeling method, the very basic experimental concept is similar to the one applied to the detection of the human phosphoproteome upon EGF stimulation (Chapter 4.6.1.1.1). Thus, the basic experimental set up provides an across-the-species protocol for large scale quantitative mass spectrometry analysis of *in-vivo* phosphoproteomes ranging from human (4.6.1.1) to bacteria (4.6.1.5). The underlying experimental design is illustrated in Figure 4.29. Importantly, the integration of identified phosphorylated sites into PHOSIDA is followed by data mining linking specific residues to kinase motifs (Chapter 4.2.1.3), evolutionary conservation (Chapter 4.2.4), protein structure (Chapter 4.2.3) and gene ontology annotations (Chapter 4.2.1.3). Our data and analyses allows not only to derive general patterns regarding phosphorylation in prokaryotes, but also to gain more comprehensive biological insight for specific prokaryotic proteins of interest to individual researchers using the PHOSIDA online database.

The site-specific and global analysis of the *Bacillus subtilis* phosphoproteome resulted in the identification of 103 unique phosphopeptides from 78 proteins (Table 4.1). In total, 78 phosphorylation sites were determined with a probability higher than 75% (class I sites). Among the identified phosphosites, 54 were on serine (69.2%), 16 were on threonine (20.5%), and eight were on tyrosine (10.3%) (Figure 4.30). As expected, we did not detect any histidine or aspartate phosphorylation. Interestingly, the phosphoproteome of *E.coli* showed striking similarity in size and number of detected phosphorylation sites: we measured 105 phosphopeptides from 79 proteins, with 81 class I phosphorylation sites. A total of 55 serines, 19 threonines, and 7 tyrosines were found to be phosphorylated, yielding a Ser/Thr/Tyr phosphorylation ratio of 67.9%, 23.5%, and 8.6%, respectively. The size of the *L.lactis* phosphoproteome was also similar to the ones of *B.subtilis* and *E.coli*. We identified 102 unique phosphopeptides in 63 proteins, with 73 phosphorylation sites. However, the distribution of Ser/Thr/Tyr phosphorylation differed, as we identified 34 phosphoserines (46.5%), 37 phosphothreonines (50.6%), and 2 phosphotyrosines (2.7%). Interestingly, the archaean phosphoproteome was similar in size, as we identified 115 unique phosphopeptides from 69 *H.salinarium* proteins. We determined 81 class I phosphorylation sites, 70 on serine (87%), 10 on threonine (12%), and one on tyrosine (1%).

	Genome Size (ORFs)	Number of Phosphoproteins	Number of Phosphopeptides	Number of Phosphosites
E.coli	4300	79	105	81
B.subtilis	4100	78	103	78
L.lactis	2266	63	102	73
H.salinarium	2821	69	115	81

Table 4.1: Comparison of detected prokaryotic phosphoproteomes

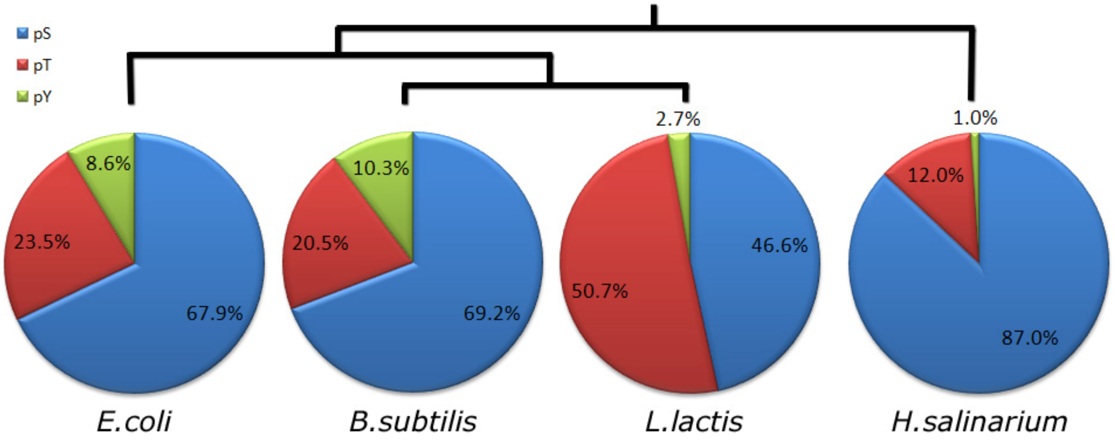


Figure 4.30: Distribution of Ser/Thr/Tyr phosphorylation in the bacteria *E.coli*, *B.subtilis*, and *L.lactis*, and the archaean species *H.salinarium*

In *Bacillus subtilis*, we detected phosphorylation sites on many glycolytic enzymes, including phosphohexose-isomerase, aldolase, triose-phosphate isomerase, glyceraldehydes 3-phosphate dehydrogenase, phosphoglycerate kinase, phosphoglycerate mutase, enolase and pyruvate kinase. In addition, phosphorylation sites were detected on several members of the pentose phosphate pathway. Furthermore, several phosphorylated proteins are involved in DNA metabolism and protein synthesis such as initiation factor IF-1 and elongation factor Ts. Other phosphoproteins were members of the phosphoenolpyruvate-dependent phosphotransferase (PTS) system. A significant overrepresentation of detected phosphoproteins involved in the main pathways of the carbohydrate metabolism was also evident in *E.coli*, as essential enzymes such as pyruvate kinase were phosphorylated. Other phosphoproteins were involved in protein synthesis and the PTS system.

The functional distribution of phosphorylation events detected in *L.lactis* was similar to the ones of the other bacteria, as the majority of glycolytic enzymes were found to be phosphorylated. Aminoacyl-tRNA phosphorylated proteins and ribosomal proteins in *L.lactis* were also phosphorylated. Even the more distantly related archaean organism, *Halobacterium salinarium*, showed a majority of phosphorylated proteins that play essential roles in a variety of metabolic pathways such as carbohydrate metabolism, amino acid metabolism, and nucleotide metabolism. Although the annotation of the *H.salinarium* proteome is not as comprehensive as the ones for the investigated bacteria, corresponding functions of determined phosphoproteins could be estimated by homology searches to other prokaryotes as implemented in PHOSIDA (Chapter 4.2.4). Figure 4.32 illustrates two phases of glycolysis and indicates phosphorylated enzymes, which were determined in our studies.

It was important to exclude the possibility of spurious detection of phosphopeptides of eukaryotic origin, which might have been present in the reagents used in sample preparation. For this purpose, we BLASTed all detected phosphopeptides against the complete NCBI protein database. This analysis resulted in only three *L.lactis* phosphopeptides with identical sequence and therefore mass as in eukaryotic proteins. Furthermore, given the starting amount of the *L.lactis* cell lysate, the probability of detection of eukaryotic phosphopeptide contaminants, even for these three peptide cases in *L.lactis*, is extremely low.

On the basis of two-directional BLAST runs (Chapter 4.2.4), we also determined the overlaps between the bacterial phosphoproteomes: Despite a relatively high conservation (overlap) on the functional protein level, on the level of phosphorylation sites the conservation was less pronounced (Figure 4.31). There are only a few identical phosphorylation sites detected in all prokaryotic species. More details about the conservation of phosphorylation events in prokaryotes are described in Chapter 9.

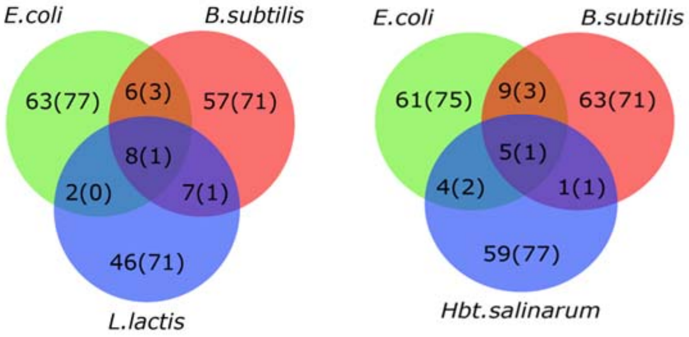


Figure 4.31: Phosphoprotein and phosphosite (in brackets) overlap in *E.coli*, *B.subtilis*, *L.lactis* (left panel), and *H.salinarium* (right panel)

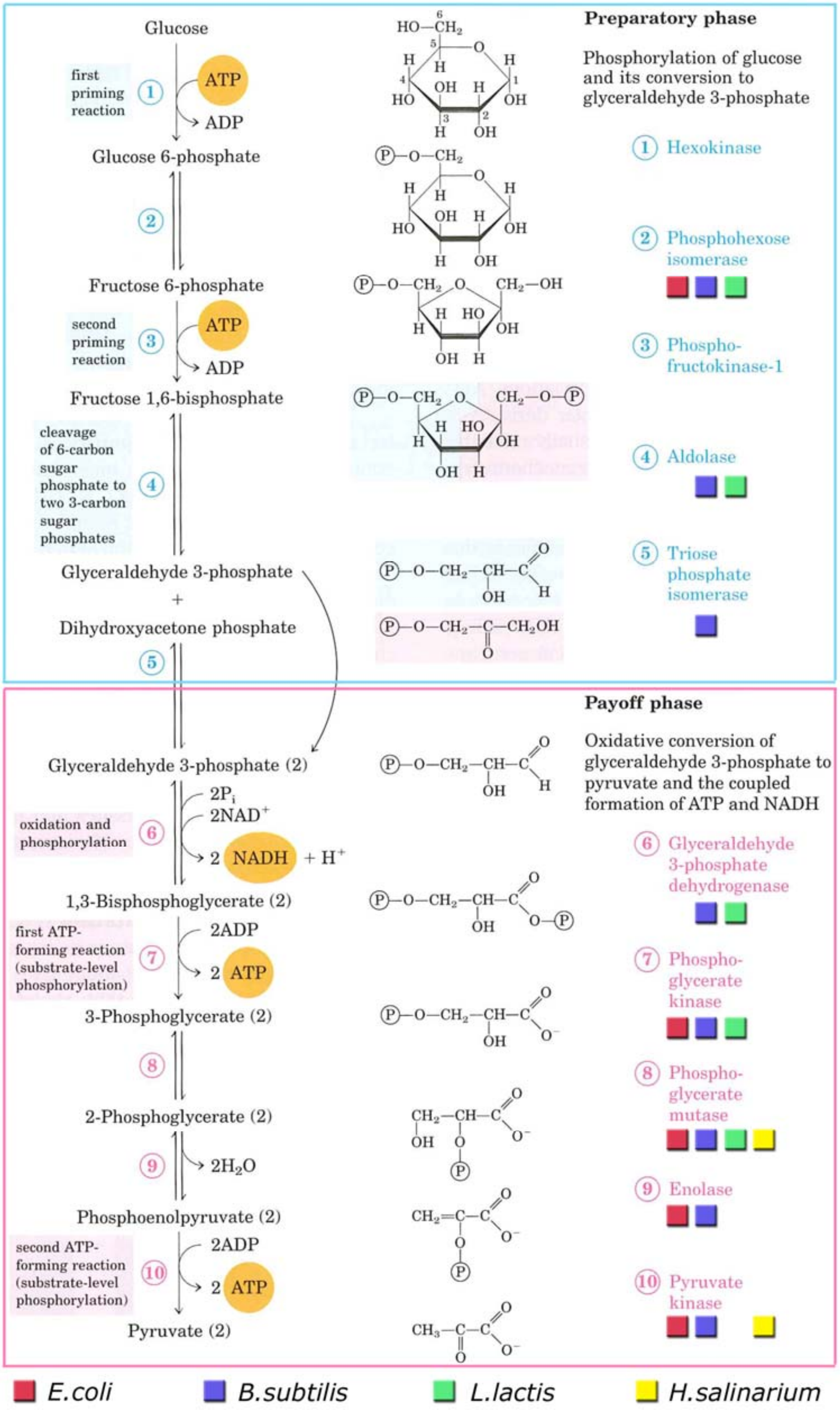

Preparatory phase
Phosphorylation of glucose and its conversion to glyceraldehyde 3-phosphate
Glucose
first priming reaction
① ATP ADP
Glucose 6-phosphate
②
Fructose 6-phosphate
second priming reaction
③ ATP ADP
Fructose 1,6-bisphosphate
cleavage of 6-carbon sugar phosphate to two 3-carbon sugar phosphates
④
Glyceraldehyde 3-phosphate
+
Dihydroxyacetone phosphate
⑤
① Hexokinase
② Phosphohexose isomerase
③ Phospho-fructokinase-1
④ Aldolase
⑤ Triose phosphate isomerase
Payoff phase
Oxidative conversion of glyceraldehyde 3-phosphate to pyruvate and the coupled formation of ATP and NADH
Glyceraldehyde 3-phosphate (2)
oxidation and phosphorylation
⑥ 2Pi 2NAD+ 2 NADH + H+
1,3-Bisphosphoglycerate (2)
first ATP-forming reaction (substrate-level phosphorylation)
⑦ 2ADP 2 ATP
3-Phosphoglycerate (2)
⑧
2-Phosphoglycerate (2)
⑨ 2H2O
Phosphoenolpyruvate (2)
second ATP-forming reaction (substrate-level phosphorylation)
⑩ 2ADP 2 ATP
Pyruvate (2)
⑥ Glyceraldehyde 3-phosphate dehydrogenase
⑦ Phospho-glycerate kinase
⑧ Phospho-glycerate mutase
⑨ Enolase
⑩ Pyruvate kinase
🟥 E.coli
🟦 B.subtilis
🟩 L.lactis
🟨 H.salinarium

Figure 4.32: Schematic illustration of the glycolysis pathway (Nelson and Cox, 2008)

Enzymes that we determined to be phosphorylated in *E.coli* (red), *B.subtilis* (blue), *L.lactis* (green), and *H.salinarium* (yellow) are marked accordingly.

4.6.2 Gene Ontology Analysis

As described in Chapter 4.2.1.3, PHOSIDA stores annotation data ranging from determined domain structures to known active sites. The PHOSIDA administration tool enables mining of the data and extraction of knowledge from the data. One of the available methods is the automated set up of Cytoscape runs (Chapter 4.5.1), which search for significantly overpresented gene ontology annotations in the given phosphodataset. Here we analyze the functional distribution of the phosphoproteomes from the model species described above.

We found that around half of the phosphorylation events in human cells (Chapter 4.6.1.1) occurred on nuclear proteins, whereas only one third of all proteins in the database were assigned as nuclear by GO (Figure 4.33). Based on the hypergeometric test along with Benjamini & Hochberg False Discovery Rate correction, this represents a significant enrichment of the phosphoproteome in the nucleus. This tendency was also observed in the phosphoproteomes of other organisms (Chapters 4.6.1.2 – 4.6.1.4): In *D.melanogaster*, 42% of the identified phosphoproteins are located in the nucleus whereas only 20.8% of all proteins annotated in FlyBase are localized in the nuclear section. As expected, proteins annotated as extracellular were significantly underrepresented in the phosphoproteome. In humans, a mere 3% of the determined phosphoproteins are annotated to be localized in the extracellular space, whereas 11% of human proteins in general are localized to this compartment. Although there is evidence of a mitochondrial phosphoproteome, proteins annotated as mitochondrial by GO were underrepresented: In fly, for example, 2.6% of phosphorylated proteins were detected in mitochondria. In comparison, 6.7% of all FlyBase proteins are located in mitochondria.

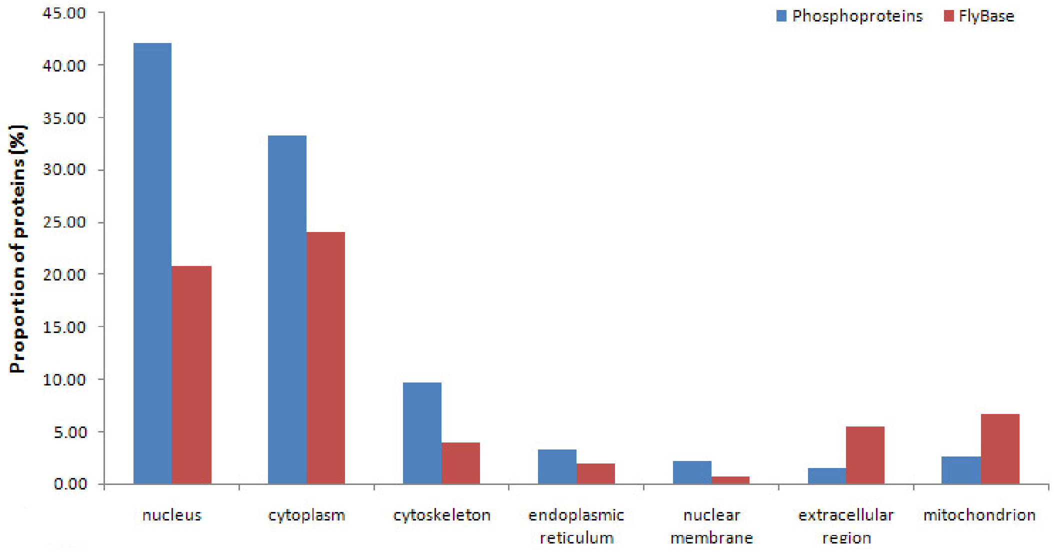


Figure 4.33: Gene ontology component analysis of the *D.melanogaster* phosphoproteome

Regarding the functional impact and biological processes associated with proteins, we found evidence for a significant overrepresentation of cell signalling functions: As expected, kinase activity, ATPase activity, receptor signalling protein activity, transcription regulator activity, and translation regulator activity were all found to be highly significantly overrepresented functions in the measured phosphorylated eukaryotic proteins (Figure 4.34). The observations relating to the over- and underrepresentation of cellular component localizations and biological functions were similar in all investigated eukaryotic species.

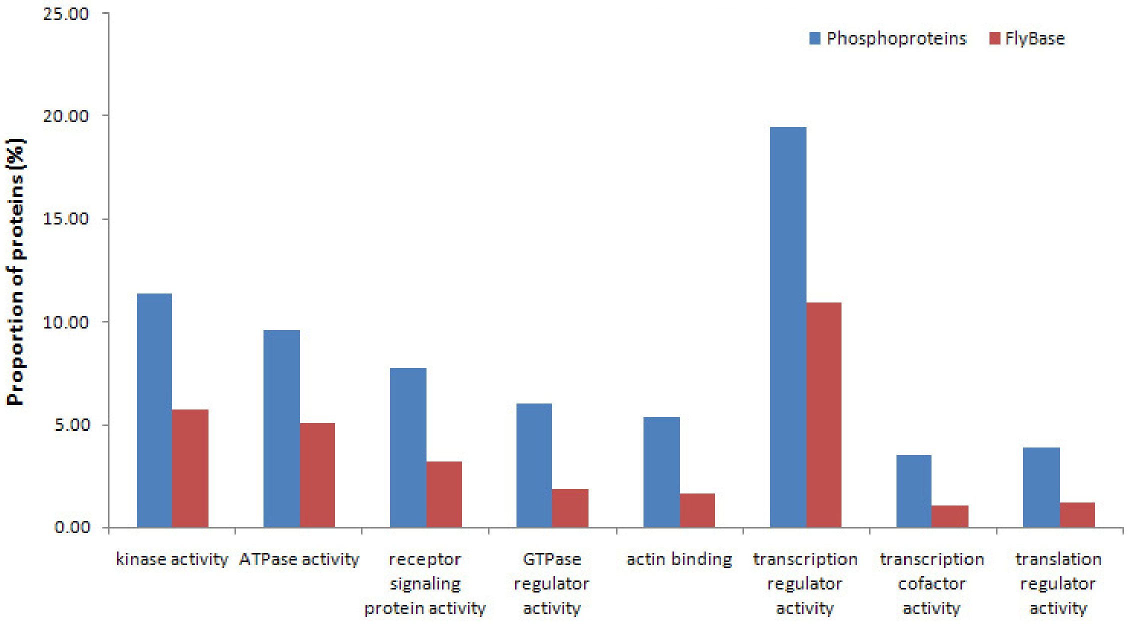


Figure 4.34: Gene ontology function analysis of the *D.melanogaster* phosphoproteome

As there is virtually no gene ontology annotation for bacteria, we used the Blast2GO tool (Conesa et al., 2005) to extract the GO terms for prokaryotic proteins from their closest GO-annotated orthologs in the SwissProt database. In this way, we obtained information on biological process for 60 out of 78 phosphorylated *Bacillus subtilis* proteins. In addition, we derived information on cellular localization for 26 phosphorylated proteins. Phosphoproteins were found to be present in all compartments of the bacterial cell (Figure 4.35) and distribute among a wide variety of metabolic and regulatory enzymes. In concordance with the observations described in Chapter 4.6.1.5, a GO enrichment analysis against the entire proteome of *B.subtilis* showed that protein phosphorylation is statistically overrepresented among enzymes involved in the main pathways of carbohydrate metabolism, DNA metabolism, protein synthesis and phosphoenolpyruvate-dependent phosphotransferase system (PTS). These results were also true for the phosphoproteomes of the other prokaryotes.

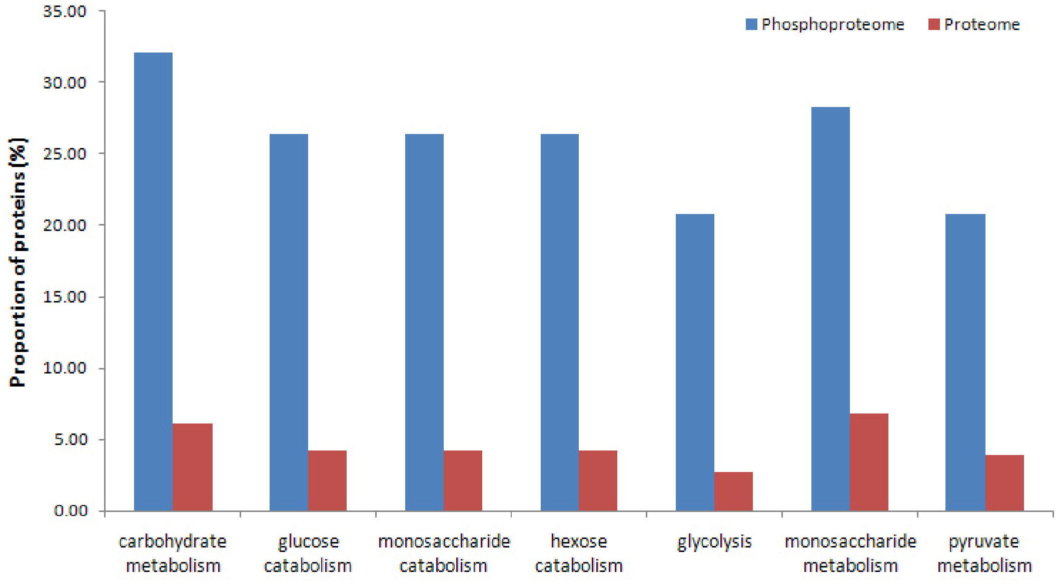


Figure 4.35: Gene ontology biological process analysis of the *B.subtilis* phosphoproteome

4.6.3 Sequence Motif Analysis

We next wished to infer the possible kinases responsible for the phosphoproteome using kinase motifs and statistical test. We employed the χ^2 test via the PHOSIDA administration tool as described in Chapter 4.4, to check whether phosphorylation sites identified in a given project match significantly with known human kinase motifs integrated into PHOSIDA (Chapter 4.2.2). We estimated the statistical chance for each kinase motif to match with a given phosphosite according to the amino acid composition of the motif and the relative frequencies of each amino acid composition in the entire proteome of the investigated organism. We found that phosphorylation sites of the mouse proteome (Chapter 4.6.1.2) matched significantly with most of the known human kinase motifs with only a few exceptions such as the motif of the NEK6 kinase. As an example, the number of mouse phosphosites that matched with motifs of the protein kinase A (PKA) was ten times higher than one would expect by chance. Significantly overrepresented matches with human kinase motifs were also observed in phosphoproteomes of eukaryotes that are more distantly related to human: For phosphosites identified in fly cells (Chapter 4.6.1.3), the CDK1 motif p[ST]-P-X-[KR] was enriched six-fold, for instance. Even in yeast, the consensus sequence of the CK2 kinase motif was enriched by a factor of three. However, as expected, kinases that are not present in yeast, such as EGFR or ALK, did not show a significant overrepresentation of candidate substrates in the yeast phosphoproteome. Table 4.2 lists investigated human kinase motifs along with the number of observed and expected yeast phosphorylation sites that matched with the given kinase motif.

motif	kinase	class 1 (observed)	class 1 (expected)	class 1 (chi-square)
R.p[ST]	PKA	189	172.7	1.61
R[RK].p[ST]	PKA	91	20.3	247.52
KR..p[ST]	PKA	37	12.6	47.4
S..p[ST]	CK1	678	351.3	334.05
[ST]...pS	CK1	673	496.3	73.97
p[ST]..E	CK2	786	250	1228.29
pS...S	GSK3	445	300.5	76.4
p[ST]P.[RK]	CDK2	61	19.9	85.32
R..p[ST]	CAMK2	315	172.7	122.71
R..p[ST]V	CAMK2	19	9.6	9.23
P.p[ST]P	ERK	46	7.4	201.73
V.p[ST]P	ERK	25	9.5	25.35
PEp[ST]P	ERK	6	0.5	60.51
R[RST].p[ST].[ST]	AKT	35	5	180.23
R.R..p[ST]	AKT	33	7.7	83.29
R..p[ST].R	PKC	2	7.7	4.23
[LVI].[RK]..p[ST]	PKD	132	99	11.29
[IEV]pY[EG][EDPN][IVL]	LCK	0	0.1	0.1
[IVL]pY..[PF]	ABL	1	0.9	0.01
[ED]..pY..[DEAGST]	SRC	7	2.2	10.98
pY..[ILVM]	ALK	3	11.4	8.13
[DPSAEN].pY[VLDEINP]	EGFR	11	7.9	1.46
p[ST]P.[KR]	CDK1	61	19.9	85.32
p[ST]P[KR]	CDK1	52	19.9	52.05
[RK].p[ST][ILV]	Aurora	65	99	11.98
[RKN]R.p[ST][MILV]	Aurora-A	19	7.3	18.79
[DE].p[ST][VILM].[DE]	PLK	28	13.8	14.66
[ED].p[ST][FLIYWVM]	PLK1	105	155.3	16.97
L..p[ST]	NEK6	174	372.3	116.83
L.R..p[ST]	CHK1/2	40	16.6	33.13
[MILV].[RK]..p[ST]	CHK1	143	108.6	11.21
F..Fp[ST][FY]	PDK1	0	0.6	0.6
[FLM][RK][RK]p[ST]	NIMA	4	8.7	2.54

Table 4.2: Number of observed versus expected yeast phosphorylation sites that matched with human kinase motifs. A chi-square value larger than six is equivalent to a p value of 0.01.

To confirm the significant overrepresentation of human kinase motifs without any a priori information we used Motif-X an iterative approach to derive significantly overrepresented motifs from large-scale datasets as described in Chapter 4.5.1. The PHOSIDA administration tool created query sets by pre-aligning all clearly identified phosphorylation sites along with their surrounding sequence of 12 residues. A probability p-value of less than 0.0001 was considered significant. In addition, a minimum occurrence of 20 of the sequence pattern in the phosphodata was required to derive a significant consensus sequence.

The application of this unbiased statistical approach led to the same outcome as above: Extracted overrepresented amino acid compositions around phosphorylated residues of various eukaryotes were similar to known human kinase motifs. For example, the second most significant consensus sequence in the fly phosphoproteome was the CDK1 motif, which was also proven to be significantly enriched according to the χ^2 test. To expand the consensus sequence comparison, we applied the Motif-X approach to phosphosites detected in human HeLa cells exposed to EGF stimulation. In total, 20 significant sequence motifs matched exactly with those derived from the *Drosophila* phosphoproteome set. Other extracted motifs were similar in composition between human and fly, but varied only in one amino acid position. Figure 4.36 shows examples of motif logos that were found to be significantly overrepresented in the human phosphoproteome and in the fly phosphoproteome. In contrast, consensus sequences derived from the yeast phosphoproteome were found to be more organism-specific, as the overlap with consensus sequences of higher eukaryotes including human, mouse and fly was relatively low.

We also used the χ^2 test to test whether phosphorylation sites determined in prokaryotic cells matched significantly with human kinase motifs. However, we did not find evidence for any significantly overrepresented eukaryotic kinase motifs in bacteria. The application of Motif-X also did not yield any significant sequence motif from prokaryotic phosphoproteomes.

As highlighted in Chapter 4.2.5, the online application of PHOSIDA lists all matching kinase motifs for a given phosphorylation site. The display of matching kinase motifs enables web users to explore possible kinases responsible for any phosphorylation site of interest.

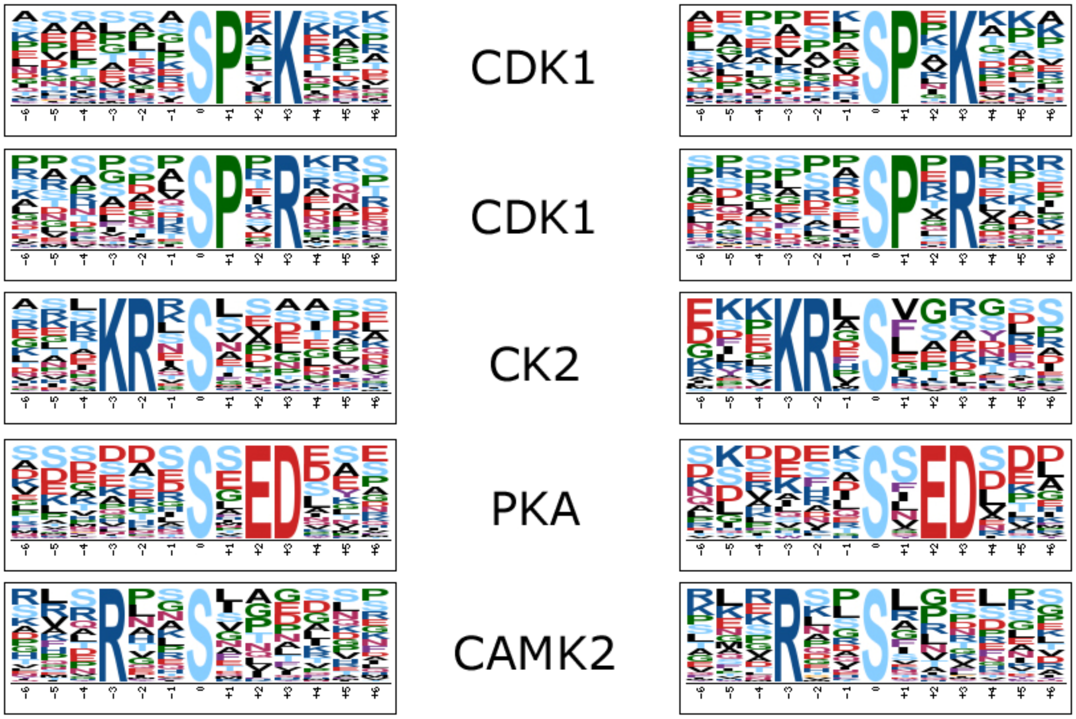
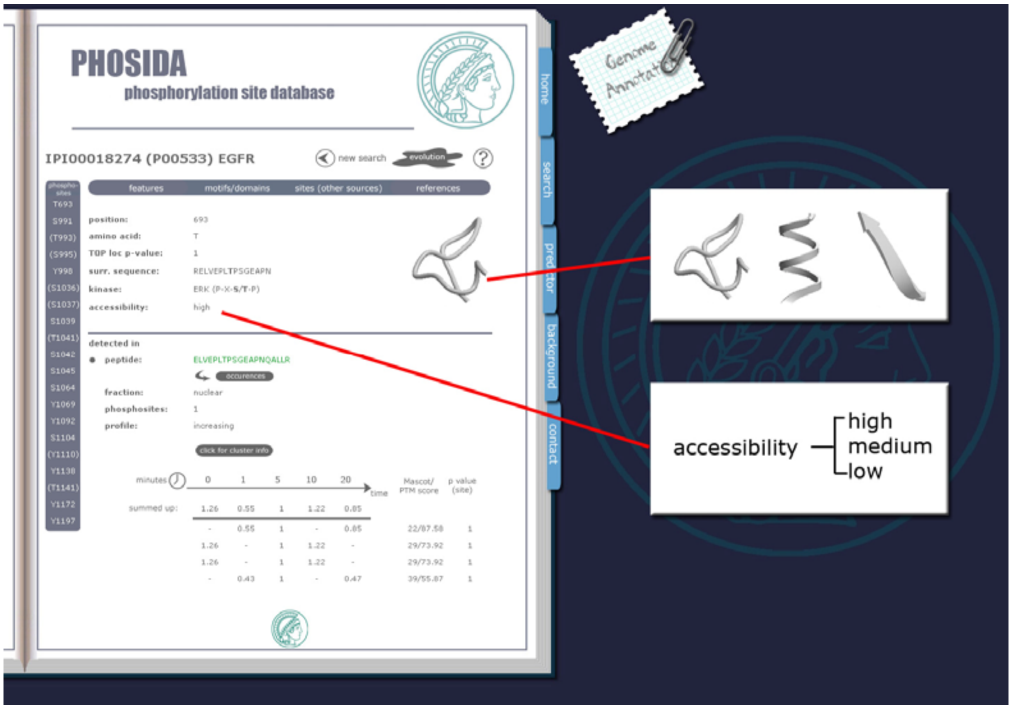


Figure 4.36: Consensus sequences identified in the fly phosphoproteome (left panel) and human phosphoproteome (right panel). Data were calculated in with identical methods.

4.6.4 Structural Constraints on Phosphorylation Sites

Previous anectodatal observations had already suggested that phosphorylation sites are mainly located in parts of proteins without regular structure (Iakoucheva et al., 2004). To verify this observation on the basis of our large-scale and unbiased studies and to enable users to investigate the structural context of each phosphorylation site of interest, we made use of the secondary structure and solvent accessibility predictions integrated in PHOSIDA (Chapter 4.2.3). As shown in Figure 4.37, the structural attributes of each phosphorylation site are visualized in PHOSIDA.



IPI00018274 (P00533) EGFR

position: 693
amino acid: T
TOP loc p-value: 1
surr. sequence: RELVEPLTPSGEAPN
kinase: ERK (P-X-S/T-P)
accessibility: high

detected in

- peptide: ELVEPLTPSGEAPNQALLR
  fraction: nuclear
  phosphosites: 1
  profile: increasing

minutes	0	1	5	10	20	Mascot/ PTM score	p value (site)
summed up:	1.26	0.55	1	1.22	0.85		
	-	0.55	1	-	0.85	22/87.58	1
	1.26	-	1	1.22	-	29/73.92	1
	1.26	-	1	1.22	-	29/73.92	1
	-	0.43	1	-	0.47	39/55.87	1

Figure 4.37: Predicted secondary structures and solvent accessibilities of identified phosphorylation sites as illustrated in PHOSIDA

To determine the overall accessibility at the protein level, we compared identified human phosphoproteins (Chapter 4.6.1.1.1) with random proteins from SwissProt. We found that phosphoproteins as a group have significantly higher accessibilities than a set of randomly selected proteins (t-test: $\sigma = 0$). This means that all residues that occur in phosphoproteins show a higher accessibility on average than all residues in non-phosphorylated proteins. Phosphoproteins, on average, are longer than the average of the database; thus, this effect is not caused by a smaller surface to volume ratio.

Furthermore, global analyses on all eukaryotic phosphoproteomes ranging from yeast to human showed that the accessibilities of phosphoserine, phosphothreonine and phosphotyrosine are significantly higher than the ones of non-phosphorylated serines, threonines or tyrosines. Non-phosphorylated residues were taken from phosphoproteins, excluding bias due to protein selection (Figure 4.38).

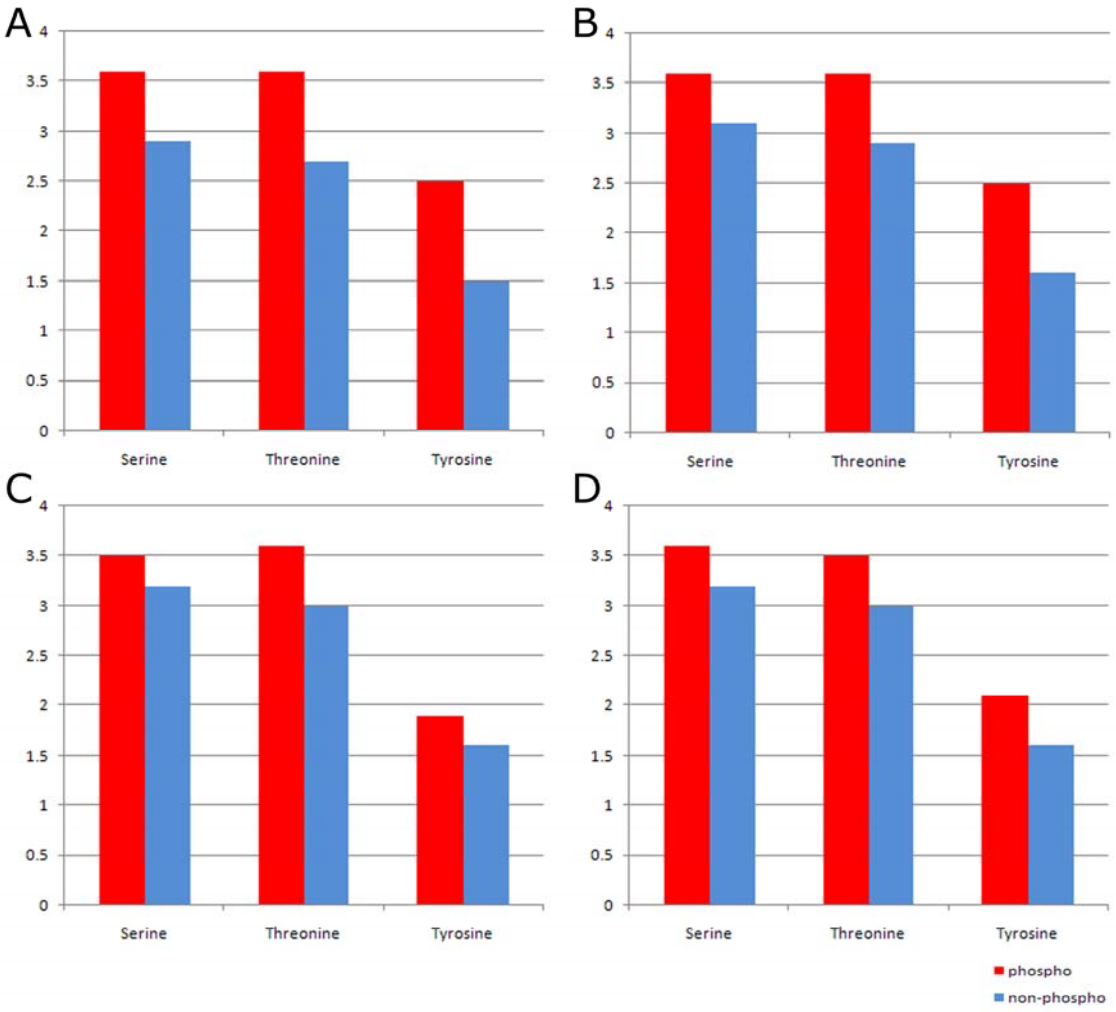


Figure 4.38: Accessibilities of phosphorylation sites as calculated by SABLE

The relative accessibility prediction assigns a value between 0 (fully buried) and 9 (fully exposed) to each residue. Accessibility is significantly higher than for their non-phosphorylated counterparts in the same proteins in all phosphoproteomes of eukaryotes (A: *S.cerevisiae*, B: *D.melanogaster*, C: *M.musculus*, D: *H.sapiens*) and for all phosphorylatable residules.

The high accessibility of phosphorylation sites suggests that they are largely localized in hinges and loops, since these structural elements are at the protein surface. In fact, this is the case to a striking degree for pS (yeast: 91%, fly: 93%, mouse: 93%, human: 93%), as well as for pT (yeast: 92%, fly: 92%, mouse: 92%, human: 88.5%). pY (yeast: 75%, fly: 78%, mouse: 78%, human: 67.3%) is also predominantly found in these regions (Figure 4.39). To confirm the generality of these observations, we mapped identified *in-vivo* phosphorylation sites to three-dimensional coordinates for phosphoproteins with a solved structure in the Protein Data Bank (Berman et al., 2000). As is apparent from the structures, the phosphogroups were always located in highly accessible parts of the proteins (Figure 4.40). In many cases, the structure around the phosphosites was even so flexible that it could not be determined at all.

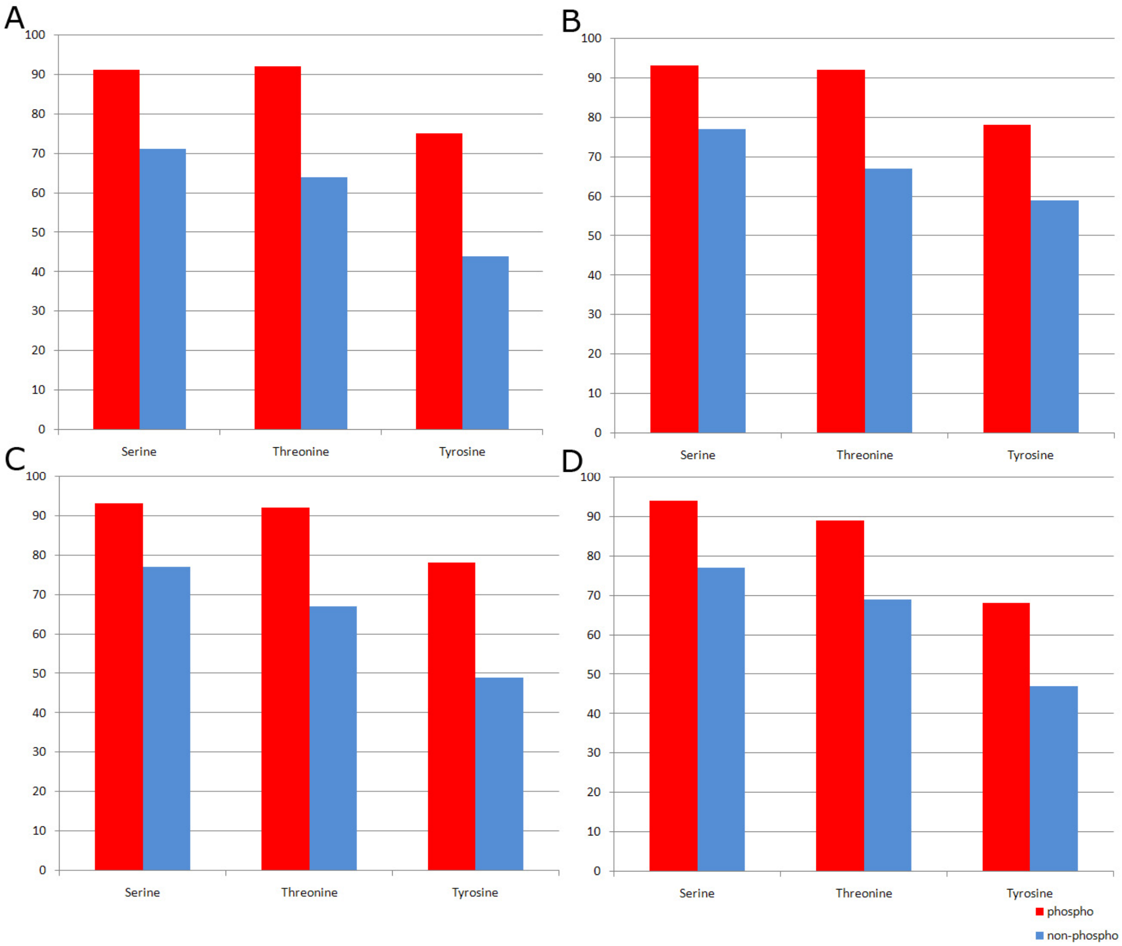


Figure 4.39: Proportion of phosphorylation sites located in loops and hinges as determined by SABLE
In each case (A: *S.cerevisiae*, B: *D.melanogaster*, C: *M.musculus*, D: *H.sapiens*), phosphosites are significantly more frequently located in flexible regions (loops, hinges).

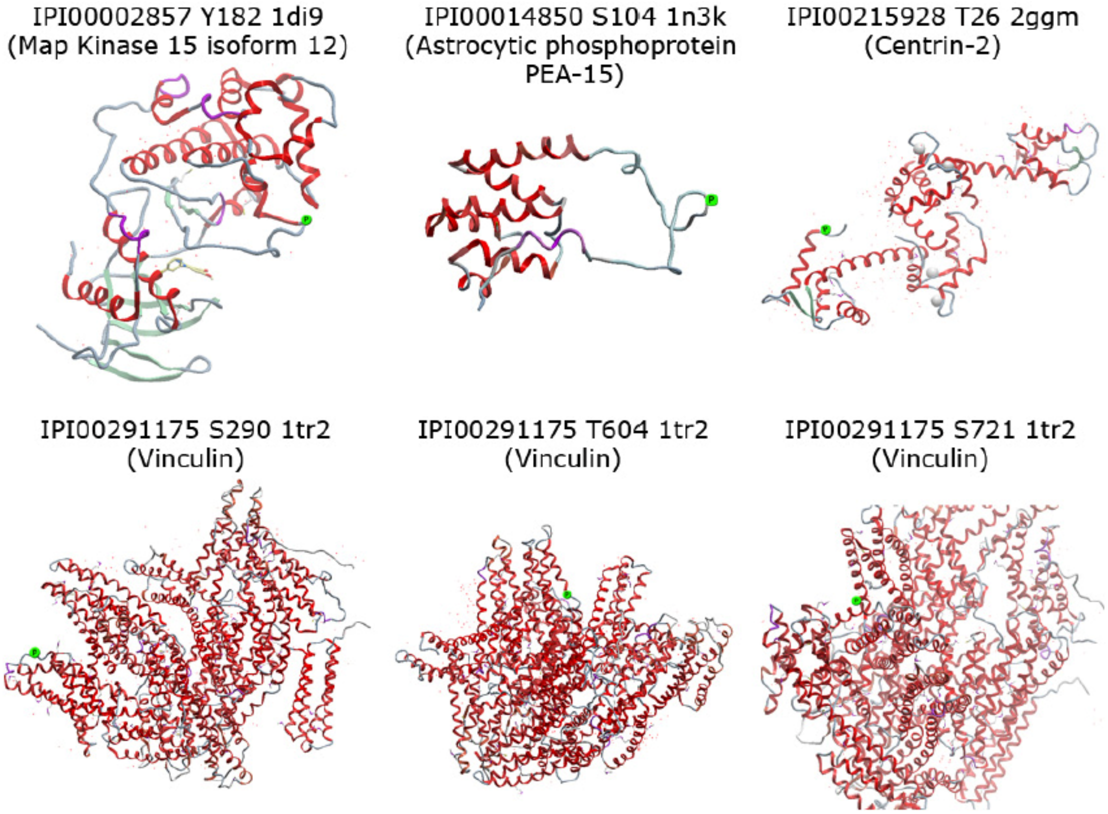


Figure 4.40: Example PDB structures of phosphoproteins (phosphosites marked in green)

4.7 Discussion

Our group has developed a strategy combining SILAC for encoding phosphorylation changes, SCX and TiO_2 chromatography for phosphopeptide enrichment, and high-accuracy mass spectrometric characterization. We applied this strategy to a several model organisms in different biological contexts ranging from EGF stimulation (Chapter 4.6.1.1.1) to phosphatase inhibition (Chapter 4.6.1.2.1) and perturbation by phosphatase RNAi knockdown (Chapter 4.6.1.3). We even applied this strategy to the determination of different prokaryotic species (Chapter 4.6.1.5). The detailed implementation of the mass spectrometric approach was somewhat different among the specified projects: For the identification of the yeast phosphoproteome, for example, we applied SILAC 1:1 labeling meaning that there is no biological difference between the two (heavy and light) populations. For the detection of prokaryotic phosphoproteomes we did not apply the SILAC technology at all. Nevertheless the main workflow of the described strategy was basically the same in each large scale study.

The approach is completely generic for identification of phosphorylation events in signalling pathways.

Identification of numerous phosphorylation sites on kinases and other low-abundance regulatory proteins demonstrates that the technology can probe the *in-vivo* phosphoproteome in considerable depth.

As a large proportion of cellular proteins are phosphorylated and the phosphoproteome is therefore very large and complex, the investigation of various *in-vivo* phosphoproteomes requires consistent data management and user friendly open access interface to retrieve data. In addition, the determination of thousands of phosphorylation sites requires a strategy to derive knowledge from the raw data. These requirements motivated the conception of PHOSIDA, the phosphorylation site database. On the basis of mySQL, C# and the ASP.NET technology (Chapter 2), we created a comprehensive database management system, which embraces the upload of experimental data, followed by the automated application of a range of mining methods. The entire workflow presents a 'Knowledge Discovery from Databases' (KDD) process, one of the most important methods in database technology (Chapter 4.1).

The large scale study of the human phosphoproteome upon EGF stimulation (Chapter 4.6.1.1.1) showed that only a small subset of phosphorylation sites are regulated in response to a stimulus. The observation that individual phosphosites on a protein are typically regulated differently suggests that proteins generally serve as integrating platforms for a variety of incoming signals. Therefore global investigations of phosphorylation events have to be site specific and there is a need for algorithms that assign phosphorylation sites to given spectra with statistical rigor (Chapter 3). This pioneering study showed that detailed and time-resolved information about numerous signalling events controlled by phosphorylation can be obtained by modern phosphoproteomics. About 90% of our phospohorylation sites were novel both compared to SwissProt and to other large scale studies. Taken together, our data suggested that, despite several decades of research into phosphorylation, most *in-vivo* phosphorylation sites have still not been detected.

The focus of the second major study of the human phosphoproteome described in this thesis was the investigation of cell cycle dependent phosphorylation regulation of protein kinases (Chapter 4.6.1.1.2). Here, we established a phosphoproteomics strategy that combines SILAC based mass spectrometry as described above with efficient kinase enrichment. This approach led to the identification of more than 1000 phosphorylation sites on protein kinases, most of which have not been described previously. We found more than half of all phosphopeptides on kinases significantly upregulated in mitotic cells pointing to wide-spread regulation of the kinome in mitotic cells. Interestingly, we determined novel cell cycle dependent regulation by phosphorylation even for the most intensely studied kinases. This approach has potential applications in drug research, as kinases that are potentially cell cycle dependently de-regulated in tumours represent prime targets for anti-cancer drugs.

Our generic phosphoproteomics strategy also proved to be successful in the mouse model. We used the SILAC technology to quantify basal phosphorylation against upregulated phosphorylation after applying a cocktail of phosphatase inhibitors (Chapter 4.6.1.2.1). Employing phosphatase inhibitors resulted in a boost of low level phosphorylation sites and made them more likely to be sequenced and identified. Again, more than half of the identified sites were novel suggesting that the determination of the mouse phosphoproteome is also far from complete. For phosphotyrosine, inhibition was effective and the majority of pY sites were strongly increased upon treatment. However, there was no evidence for a strong increase of the phosphorylation level of serines and threonines. The majority of pS and pT was unaffected by the inhibitors. One plausible reason for this observation could be a specificity of the applied inhibitors for only small classes of phosphatases.

Cancer is predominantly a genetic disease and genome projects have already shown that more than hundred protein kinases are involved in human cancer. Mutations in the genome often lead to deregulated protein kinase activity in cancer, primarily constitutive activation. We used mutant mice as a skin tumor model to prove that our established mass spectrometry strategy is applicable to solid tumor analysis (Chapter 4.6.1.2.2). To study the global phosphoproteome of solid tumor tissue we used SCX-TiO_2 and multiple TiO_2 incubation which allowed mapping the position of more than 5000 phosphorylation sites in melanoma tissue with confidence. We found phosphosites from many pathways directly or indirectly involved in cancer, for example, in the mTor pathway, which regulates protein translation. The coverage of known melanoma associated phosphorylation sites in our pilot study indicates that the approach is well suited for the analysis of the tumor tissue phosphoproteome.

The increase of identified phosphotyrosines found in the phosphatase inhibitor study in mouse was again observed in fly cells (Chpater 4.3.3): In total, 4.1% phosphorylated tyrosines were determined after phosphatase inhibitor treatment. This observation supports a more efficient inhibition of phosphotyrosine phosphatase compared to serine/threonine phosphatases. Overall, more than 6700 phosphorylation sites were found in fly cells for the phosphatase inhibitor experiment. The described experimental design was then extended by the knockdown of the Ptp61F phosphatase, the homolog of the human Ptb1B phosphatase, important in Type II diabetes. As proof of principle, we showed that the phosphoproteome can be analyzed quantitatively in response to knock down of a single regulator. The RNA interference approach revealed a comparable number of 6515 phosphorylation sites on 1952 proteins. Together, we detected nearly 10000 phosphorylation sites in *D. melanogaster*. Around half of all determined phosphorylation sites proved to be novel compared to other large-scale studies. From the technological point of view, the study also showed that it is required and feasible to normalize phosphorylation dynamics by measured proteome changes, in order to derive quantitative data that are exclusively caused by phosphorylation changes rather than protein changes upon the specified treatment. Besides the integration into PHOSIDA, the large number of phosphorylated sites allowed the implementation of the first fly specific phosphosite predictor (Chapter 7). In addition, the fly phosphoproteome provided invaluable data for the evolutionary analysis of phosphorylation (Chapter 9).

The application of our MS strategy to yeast yielded the identification of more than 4000 phosphorylation sites (Chapter 4.6.1.4). Surprisingly, we found evidence for phosphorylation events on 66 tyrosine residues. This was unexpected, as very little is known about tyrosine phosphorylation in yeast. However, even though the entire data set has a false positive rate of less than 1% on the protein and peptide levels, it is possible that the false positive rate for a subset of the data is different. In any case, the application of SILAC labelling means that all peptides are detected by the mass spectrometer as characteristic pairs or doublets that can be analyzed separately, and their sequencing in both forms increases the chance of true identification. Thus, we also detected low abundant proteins ranging from transcription factors to kinases. The project not only represents a pioneer study for the application of quantitative phosphoproteomics in yeast, but also contributes a large number of novel phosphorylation sites to the annotation of posttranslational modifications in yeast.

Finally, we determined, for the first time, the *in-vivo* and site-specific bacterial and archaea phosphoproteomes (Chapter 4.6.1.5), choosing the model organisms *E.coli*, *B.subtilis*, *L.lactis*, and *H.salinarium*. The number of identified phosphorylation events in prokaryotic cells is orders of magnitudes lower than that of eukaryotes. We sequenced between 73 (*L.lactis*) and 81 (*E.coli*) phosphorylation sites. Most of them are found on glycolytic and tricarboxylic acid cycle enzymes and members of the phosphoenolpyruvate-dependent phosphotransferase system. Despite their phylogenetic distance, phosphoproteomes of the investigated prokaryotes are similar in size, classes of phosphorylated proteins, and pS/pT/pY distribution.

All measured phosphoproteomes were analyzed for GeneOntology overrepresentation using the implemented mining methods that link with open source applications such as Cytoscape. We found that the most significantly overrepresented biological functions of eukaryotic phosphorylated proteins are associated with binding to targets ranging from ATP to transcription factors. As expected, kinase binding activity is significantly overrepresented in all eukaryotic phosphoproteomes. Functions that are related to general kinase activities, translational activation, and transcriptional regulation also proved to be significantly overrepresented. In contrast, mitochondria and secreted proteins proved to be significantly underrepresented in the phosphoproteomes. In prokaryotic phosphoproteomes, phosphoproteins involved in the main pathways of carbohydrate metabolism, DNA metabolism, protein synthesis and phosphoenolpyruvate-dependent phosphotransferase system (PTS) are significantly overrepresented.

As each phosphorylated site identified in a given species must be the substrate of one or more kinases, we matched our sites to the known substrate specifities of 33 human kinases through motif analysis. We used human kinase motifs because the ones of other eukaryotes such as mouse, fly and yeast are not known and kinase substrates are generally assumed to be well conserved throughout higher eukaryotes. Using the PHOSIDA administration tool, the application of statistical methods such as the χ^2 test and Motif-X indeed showed that phosphosites detected in eukaryotic cells match significantly to most of the known human kinase motifs. This observation verifies the high degree of conservation of kinases and their signalling pathways ranging from CDKs to ERK. Our results are concordance with a previous study of Manning et al. (Manning et al., 2002a): They compared the kinomes of various eukaryotes with known human kinases and came to the conclusion that eukaroytes share several kinase families involved in functions such as immunity, neuro-specifc functions and the cell cycle. However, the yeast phosphoproteome proved to be more distinct from the ones of higher eukaryotes. For individual reachers, the inclusion of matching motifs in the web application of PHOSIDA allows the estimation of kinase correspondences of any given substrate.

As eukaryote-like kinases have been found in bacteria, we wondered if human kinase motifs matched amino acid sequences surrounding the identified phosphorylation in the investigated prokaryotes. Although 17 phosphosites found in *B.subtilis* matched the target motifs for eukaryotic casein kinases CK1 and CK2, this distribution corresponded to expected frequencies of these motifs obtained by chance. Although it has been shown by previous studies that bacteria possess kinases and phosphatases that structurally resemble their eukaryotic counterparts (Kennelly, 2002) we could not find evidence for any significantly overrepresented consensus sequences. However, this observation does not imply that there are no serine/threonine protein kinases in bacteria with consensus substrate motifs. It rather suggests that the spectrum of substrates phosphorylated by bacterial protein kinases is not as large as that in eukaryotes perhaps indicating a more specific kinase-substrate association in prokaryotes. This is not unexpected given the relatively low number of around 80 measured phosphorylation events in prokaryotic cells in comparison to eukaryotic phosphoproteomes each comprising more than 10000 phosphorylation events.

On the basis of predicted secondary structures and solvent accessibilities integrated into PHOSIDA, we found that phosphorylation events are not distributed along the whole protein structure but are instead constrained to sites of high accessibility and structural flexibility. Particularly in the case of serine and threonine, phosphorylation is almost completely restricted to loops and hinges. Tyrosine is found to some degree in regular secondary structure elements but overall phosphotyrosines are very likely to be in flexible regions as well. Mechanistically, localization of phosphorylation in flexible regions of the protein is advantageous as it provides access for the kinase to substrate, which needs to be positioned into the active site. Furthermore, functional consequences of the phosphorylation in many cases also depend on the flexibility of the phosphorylated sequence, such as when loops are repositioned after phosphorylation or when the phosphorylated loop participates in a protein-protein interaction. However, it is important to emphasize that the secondary structural analysis was based on predictive methods rather than experimental data. Nevertheless, it stands to reason that the large size of the dataset should compensate for statistical errors caused by the prediction algorithm.

The evolutionary sections of PHOSIDA also provide insights into the evolution of phosphorylation. Main results and conclusions are discussed in Chapter 9. Furthermore, we implemented a phosphorylation site predictor that makes it possible to find putative novel phosphorylation sites that have not been experimentally identified. The concept of the predictor is described in Chapter 7.

To make the data freely and efficiently available to the community, we also implemented an online application that allows the retrieval of the phosphoproteomic data (http://www.phosida.com) (Chapter 4.2.5). The concept of an online phosphorylation site database is, of course, not a novel one. PhosphoSite (Hornbeck et al., 2004) and Phospho.ELM (Diella et al., 2004) are already established databases containing phosphorylation sites from the literature. In contrast to those efforts, the aim of PHOSIDA is to include only very high quality input data as well as quantitative information such as regulation after stimuli or perturbation after phosphatase inhibition. Additionally, we take into account structures and evolutionary data across a variety of species, in order to integrate biological context into the database and to quantify constraints of phosphorylation on a proteome-wide scale. With a total of 289 phosphoprotein entries and 313 reported phosphorylation sites from four prokaryotic species (*B.subtilis*, *E.coli*, *L.lactis* and *H.salinarium*), PHOSIDA is currently the largest open source database of prokaryotic Ser/Thr/Tyr phosphorylation. Thus, PHOSIDA provides a rich environment for the biologist wishing to analyze phosphorylation events of proteins of interest.

Chapter 5

MAPU 2.0: Max-Planck Unified Proteome Database

The MAPU 2.0 database contains proteomes of organelles, tissues and cell type (Gnad et al., in press). It allows the organism-specific retrieval of proteomic data obtained by high accuracy MS-based proteomics. The combination and update of various experiments on the basis of the same underlying database version make it possible to obtain an overall idea about the tissue-specific or organelle-specific localization of any protein of interest. In addition, the new release of the MAPU database addresses mass spectrometry specific problems including ambiguous peptide-to-protein assignments. Furthermore it provides insight into general features on the protein level ranging from gene ontology classification to SwissProt annotation. Moreover, the derived proteomic data are used to annotate the genomes. MAPU 2.0 is available on line at http://www.mapuproteome.com.

5.1 Introduction

The mapping of various proteomes having potential diagnostic utility presents one of the fundamental challenges of MS-based proteomics. Besides biotechnological problems including biochemical purification of organelles, the consolidated database management of various identified mapped proteomes is another challenge that proteomic research has to face. The MAPU 2.0 database provides a comprehensive proteome information system consisting of data integration and combination of various large-scale proteomic assays and inclusion of protein annotations from other databases (Gnad et al., in press). To allow the peptide-based retrieval of quantitatively evaluated proteomic data, we changed the basic concept of the previous version of the MAPU database completely compared to the original release of MAPU. The main modifications are the combination of various proteomic sub-databases, the employment of another programming language (C#), the addressing of MS specific problems including peptide-to-protein assignments, the inclusion of additional large-scale proteomic datasets, the detailed cross-reference to SwissProt annotations, and the new web design.

Moreover, as the number of sequenced genomes increases rapidly, the integration of biological information on the genome sequence becomes imperative (Curwen et al., 2004;

Olason, 2005). Thus, it is important to map large-scale data derived from MS-based proteomics to the genome sequence. The EnsEMBL project provides an excellent system to integrate any kind of data that contributes to the annotation of the genome (Birney et al., 2004). Therefore, we mapped the generated proteomic data to the genome and used the Distributed Annotation System (DAS) to vizualize key features such as the localization in specific cell types for each identified gene transcript.

5.2 Implementation of MAPU 2.0

The initial content and the original format of MAPU have been described in Zhang et al. (Zhang et al., 2007). The general format of the database has changed drastically, as the previous database version was divided up into several sub-databases, each containing a discrete proteomic dataset. The new version (MAPU 2.0; (Gnad et al., in press)) unifies all sub-databases by re-assigning the determined peptides along with their corresponding data of each experiment to proteins entries of an updated database version. This allows the organism-specific retrieval of various cell type and organelle associated proteomic data:

The user can query the database organism-specifically by protein name, protein description, gene symbol, accession number in the database used for identification (such as the International Protein Index (IPI)), SwissProt accession identifier, protein sequence or peptide sequence (Figure 5.1, left panel).

If more than one protein entry match with the submitted query string, MAPU 2.0 will list all relevant proteins and mark the ones that show peptides determined in specified sub-proteomes in red (Figure 5.1, middle panel). Clicking on one of the red high-lighted entries leads to the result page (Figure 5.1, right panel). If there is only one match to the query, the web user will be guided directly to the result page of the protein. The left panel of the resulting web page displays all investigated cell types and tissues that have been explored. If the given protein was detected in a specific project, the corresponding button is highlighted (Figure 5.1, right panel). Otherwise, the image of the given tissue or cell type is illustrated in very light colors indicating the absence of the specified protein of interest.

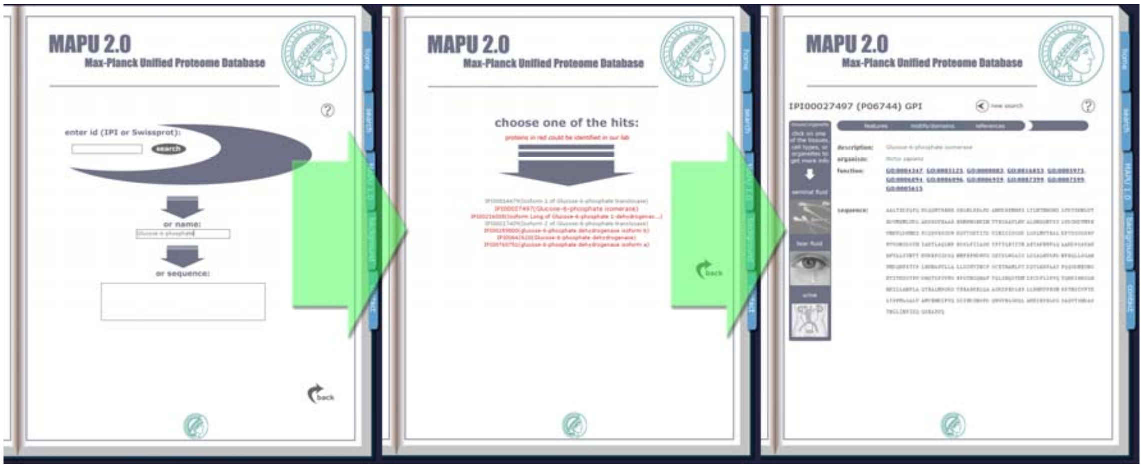
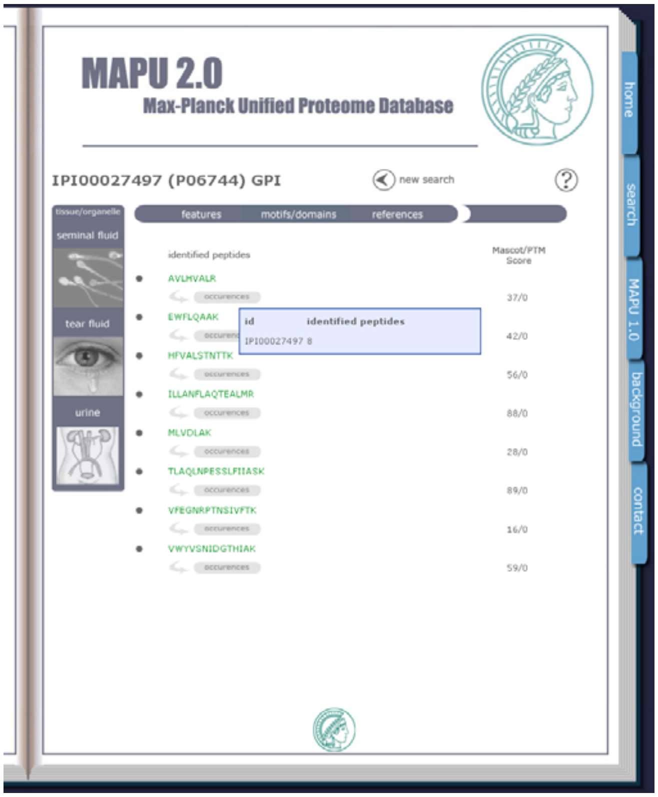


Figure 5.1: The Max-Planck Unified Proteome Database 2.0

The web user can search for any protein of interest via accession numbers, gene symbols, gene name, protein description, peptide sequence, or protein name. The final result page illustrates the occurrence of the specified protein in certain tissues or organelles along with general annotations.

Clicking on one of the buttons on the left panel results in the complete listing of all peptides that have been measured in the selected cell type along with associated data such as Mascot scores or PTM scores (Figure 5.2).



IPI00027497 (P06744) GPI

identified peptides	Mascot/PTM Score
AVLHVALR	37/0
EWFLQAAK	42/0
HFVALSTNTTK	56/0
ILLANFLAQTEALMR	88/0
MLVDLAK	28/0
TLAQLNPESSLFIIASK	89/0
VFEGNRPTNSIVFTK	16/0
VWYVSNIDGTHIAK	59/0

id	identified peptides
IPI00027497	8

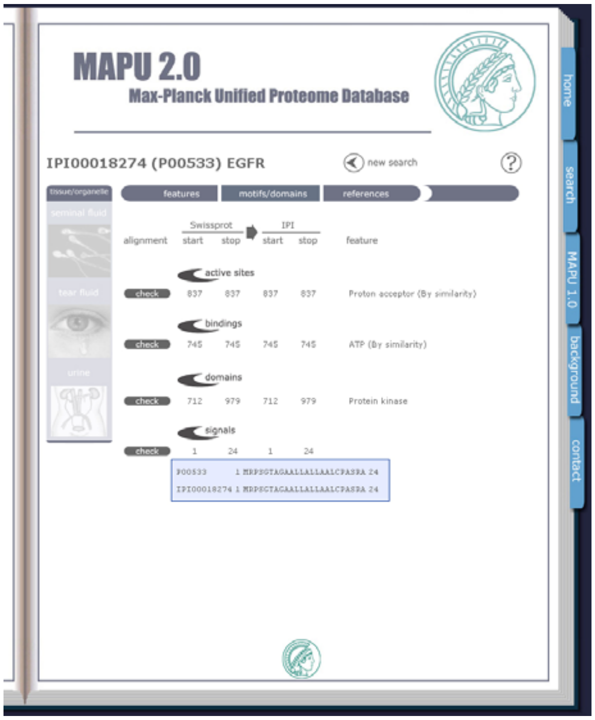
Figure 5.2: Listing of peptides that were identified in a given tissue in MAPU 2.0

The coloration of the illustrated peptide sequences indicates the uniqueness regarding the protein assignment

The peptide-to-protein assignment represents one of the main problems of MS data, since a peptide might occur in several proteins, usually isoforms or truncated versions of the gene. Multiple incidences of a peptide sequence can lead to ambiguous protein assignments. This can partially be resolved by noting that it is more likely that a given peptide sequence corresponds to the candidate protein that shows the highest number of peptides in total. MAPU addresses this issue by color highlighting the listed peptides: Green indicates that the peptide sequence is found exclusively in the selected protein of interest, whereas blue indicates that there is another protein entry that contains the peptide and shows the same number of identified peptides in total. Red points to the occurrence of another protein that shows a higher number of detected peptides in total and thus represents the more likely protein present in the sample. If one points the computer mouse to one of the corresponding 'occurrences' buttons, a blue colored box will pop up showing all protein entries that contain the given peptide along with the total number of containing peptides that have been identified (Figure 5.2). This fundamental principle of visualizing the ambiguity of protein assignments is also used in PHOSIDA – the phosphorylation site database (Chapter 4).

If the experimental design of a given project also focused on the organellar localizations of proteins, all organelles, in which the protein of interest was detected, are listed.

In addition to the illustration of associated cell types and organelles along with the measured peptides, general information about the protein is provided: Besides protein descriptions and full protein sequence, the corresponding GO identifiers are listed and they link to the Gene Ontology web site reporting full descriptions of the selected annotation. Furthermore, the annotations to each instance include the PubMed references and general features such as active sites, motifs, domains, or signaling sites derived from SwissProt (Figure 5.3). Since there may be several entries covering various isoforms or splice variants that corresponds to one SwissProt entry, we aligned the protein sequence of each SwissProt instance with the one of the corresponding entry of the database that was used for identification, which is usually the IPI database. We used BLASTP to align the protein sequences. The main purpose of this extensive alignment approach is to derive the exact sequence positions of relevant protein features that are annotated in SwissProt within the protein sequences of the entry of the other database.



IPI00018274 (P00533) EGFR

alignment	Swissprot start	Swissprot stop	IPI start	IPI stop	feature
active sites					
check	837	837	837	837	Proton acceptor (By similarity)
bindings					
check	745	745	745	745	ATP (By similarity)
domains					
check	712	979	712	979	Protein kinase
signals					
check	1	24	1	24	

```
P00533     1 MRPSGTAGAALLALLAALCPASRA 24
IPI00018274 1 MRPSGTAGAALLALLAALCPASRA 24
```

Figure 5.3: Protein annotations in MAPU 2.0 based on SwissProt cross-references

If the experiment is quantitative the median quantitative data of all measured assigned peptides are taken to describe the quantitation of the protein (provided by MaxQuant output).

Moreover, a further essential difference to the previous database version is the underlying programming language. The new release is exclusively based on C# and the ASP.NET technology, in order to have a shared class library, which is also used for the implementation of PHOSIDA (Chapter 4).

Furthermore, the concepts and web applications of MAPU 2.0 and PHOSIDA are very similar. This presents a great advantage for researchers that use both our in-house proteomic database (MAPU) as well as the phosphorylation site database (PHOSIDA). The similar web design also promotes the idea to have a corportate design of our group.

Additionally, each displayed web page includes a question mark button that directs to the help section of MAPU 2.0 describing the format of the current page or exemplifying the web application guideline. These help sections are also available via the ‘background’ section of MAPU 2.0. They contain general descriptions of the experimental designs of various projects, for instance. To allow the retrieval of sub-databases that could not be established in the new concept, a link to the old database version is provided. This is the case for the organellar database as well as the red blood database, as both datasets are exclusively protein-based and therefore cannot be mapped to MAPU 2.0 due to the lack of peptide information.

Next, we wished to use the proteomic data to annotate the genome. We extracted all measured peptides of each proteomic dataset and reassigned the given peptide sequences to gene transcripts that are annotated in the EnsEMBL database. We linked our in-house proteomic databases with the genome database in an efficient manner via the DAS/Proserver System (Finn et al., 2007). The basic concepts are explained in Chapter 8.

5.3 Discussion and Future Directions

The previous version of MAPU was not integrated and only listed the proteome results in a project specific manner. This made it impossible to query the presence or absence of a protein of interest in all proteomics projects undertakine in the group. We have therefore completely redesigned the MAPU database and it now combines all available proteomic sub-databases via the organism-specific reassignments of peptide sequences to the same underlying species databases. Thus, the format of the MAPU database has completely changed, since the previous version was protein based, whereas the new database release is peptide based.

In addition, in MAPU 2.0 we have addressed MS-specific problems such as ambiguous peptide-to-protein assignments by straightforward approaches such as color highlighting of given peptide sequences. Furthermore, MAPU 2.0 dynamically recalculated of quantitative protein data that are assigned to proteins on the basis of the individual peptide quantitation values.

Moreover, we switched to C# and ASP.NET as the underlying programming technology, in order to establish a corporate web concept and class libraries shared with PHOSIDA, which is focussed on the management of identified phosphorylation sites. In addition, we used the proteomic data that are integrated in MAPU 2.0 to annotate the genome via the DAS technology provided by the EnsEMBL project.

The success of the MS-based proteomic technology is a significant challenge for bioinformatics resources. Thus, we aim to manage and combine the available proteomic data generated in our department in an efficient manner. We intend to improve the underlying concept of MAPU continuously with the help of feedback and suggestions by the web users of the database. One of our major future goals is the provisions of more detailed validation reports of measured proteomes. This could be realized by the display of spectra images of each identified peptide sequence, for example. Besides the solution of MS-specific problems, we intend to extend our proteome database by the inclusion of measurements of additional proteomes on the basis of different tissues, cell types and organisms.

Chapter 6

SEBIDA – Sex Bias Database

In sexually reproducing species, males and females differ in many morphological and behavioural traits. Because sex-specific chromosomes such as the Y chromosome are typically highly heterochromatic and contain few genes, almost all intersexual differences arise through the differential expression of genes that are physically present in both sexes. With the advent of microarray technologies, it has become possible to detect such sexual dimorphism in gene expression on a genome-wide scale. For example, one of the first applications of microarrays in *Drosophila melanogaster* was to quantify expression differences between males and females. Since then, numerous studies have compared gene expression between the sexes in various insect species (Gibson et al., 2004; Hahn and Lanzaro, 2005; Parisi et al., 2004; Parisi et al., 2003; Ranz et al., 2003; Stolc et al., 2004).

In addition to their obvious interest for developmental biologists studying sexual differentiation, genes with sex-biased expression are also of great interest to evolutionary biologists. This is because they may be enriched for adaptively evolving genes that are subject to forces such as sexual selection or intersexual co-evolution. It is well documented that sex-biased genes, particularly those with a male expression bias, tend to evolve rapidly in both expression level and DNA/protein sequence and there is growing evidence that much of this rapid evolution may be attributable to positive selection (Ellegren and Parsch, 2007). These results are in keeping with the main findings of my Master's thesis that focused on the inter- and intra-species evolutionary analysis of sex biased genes in *Drosophila* and *Anopheles gambiae*.

To perform meta-analyses on various studies comparing male and female gene expression, we established Sebida (sex bias database) (Gnad and Parsch, 2006), a database that integrates results from multiple microarray studies comparing male versus female gene expression levels. In addition to the ratio of male to female expression for each gene, Sebida provides information useful for evolutionary studies, including measures of recombination, codon bias and interspecific divergence. The design of an online database was already subject to my 'diploma study'. However, during my PhD study we have finished the main modules that manage the web application and the underlying mySQL database. Moreover, we have added further data comprising various microarray data sets that contain male versus female gene expression levels of various insect species (Goldman and Arbeitman, 2007; McIntyre et al., 2006; Wayne et al., 2007). The results of the additional datasets are consistent with the outcomes of my previous on those datasets that have been considered in my Master's thesis. Sebida is available on line at http://www.sebida.com.

6.1 Introduction

Sexual dimorphism is the systematic difference between individuals of different sex in the same species. At the most basic level, sexual dimorphism is most evident in primary sexual characteristics defined as the different reproductive organs of male and female. These differences are often referred to as sex-dichotomous differences. They are completely specific to one sex or the other like the uterus, for instance. In comparison, phallic size is a sex-dimorphic difference. The sexes of many species also differ in secondary sexual characters that are not directly related to reproduction such as size, coloration, or behaviour (Figure 6.1). In mammals, the males are larger than the females whereas to the opposite is true in spiders, for instance. Other examples are parts of the body that are used in the struggle for dominance over other males such as tusks, antlers, or horns. Some cases of sexual dimorphism are so striking that males and females were originally taken to be members of entirely different species. For example, male eclectus parrots are green with an orange beak in contrast to scarlet female parrots with a black beak. In most cases, it is the male that shows extravagant or exaggerated secondary sexual characteristics.

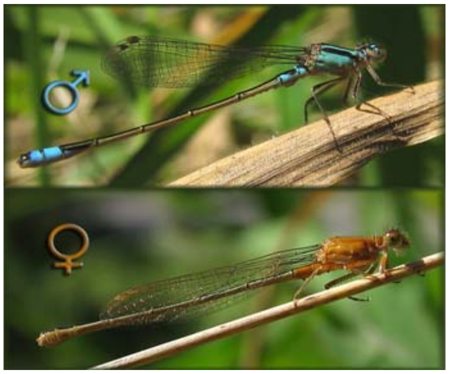


Figure 6.1: Sexual dimorphism in damselflies (www.treknature.com)

Sexual selection was Darwin's solution to the problem of why conspicuous, and apparently non-adaptive traits such as the bright colors, horns, and displays of males of many species have evolved. He proposed two forms of sexual selection: contest between males for access to females ("intrasexual selection") and female choice of some male phenotype over others ("intersexual selection") (Futuyma, 1998; Ridley, 2003).

Sexual selection exists because females produce few large gametes and males produce many small gametes. This creates an automatic conflict between the reproductive strategies of the sexes: a male can mate with many females, and often suffers little reduction in fitness if he should mate with an inappropriate female, whereas all a female's eggs can be fertilized by a single male, and fitness can be significantly lowered by inappropriate matings.

Females are a limiting resource for males competing for mates, but males are not a limiting resource for females. Because a male is capable of multiple matings, variation in mating success is generally greater among males than among females and indeed is a measure of the intensity of sexual selection. In many animals, males engage in contests that determine which will gain access to females or to resources to which females are attracted. Therefore visual or vocal signals play important roles in the competition. The males of many mammals possess weapons such as horns or tusks that inflict injury (Figure 6.2). Consequently, sexual selection by male contest supports the directional selection for greater size, weaponry, or display features.

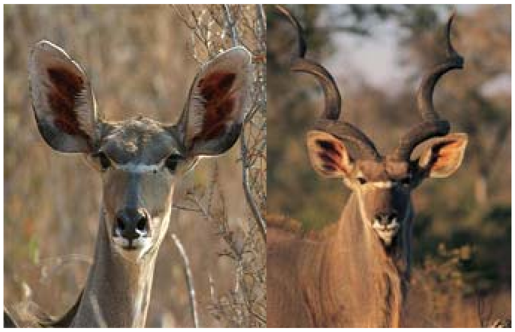


Figure 6.2: *Tragelaphus strepsiceros*: a male kudus has conspicuous antlers (right) in contrast to a female cudus (left) (www.exto.nl, www.africantravelinc.com)

In addition, females mate preferentially with males that have larger, more intense, or more exaggerated characteristics such as color patterns, ornaments, vocalizations, or display behaviors (Figure 6.3).

In summary, differential selection pressure between the sexes has been postulated to explain the substantial between-sex differences observed in morphology, physiology, and behavior, indicating the existence of different optimal sex-dependent phenotypes. Especially traits that are involved in male reproduction tend to evolve fast.

Figure 6.3: Sexual dimorphism is ubiquitous among higher eukaryotes

Male competition (left) and the female's preference for conspicuous male phenotypes such as the peacock's trait of males (right) present possible solutions to the rise of sexual dimorphism (www.classicescape.com, www.ellentroutzoo.com)

While the evolutionary apects of sexual dimorphism have been extensively studied, the molecular mechanisms are much less clear. Increasing evidence suggests that molecular mechanisms associated with sex and reproduction change substantially faster than those more narrowly restricted to survival. In order to obtain gene expression levels in males and females, high-throughput and large-scale technologies are required. This leads to the mMicroarray technology is one of the results of the astonishing development in biology in recent years. It has been developed for studying the regulation of thousands of genes. Studies of gene expression during the life cycle of *Drosophila melanogaster* have found that, for sexually mature males and females, a substantial fraction of the *Drosophila* transcriptome displays sex-dependent regulation. The enormously large amount of accquired data requires smart and efficient storage and management. Therefore, database systems become indispensable along with data mining algorithms that find valid patterns in the data.

To perform meta-analyses on different studies comparing male and female gene expression, we established Sebida (sex bias database), a database that integrates results from multiple microarray studies comparing male versus female gene expression. For each gene, Sebida provides information about the ratio of male to female expression and further data that are useful for evolutionary studies such as measures of recombination, codon bias and interspecific divergence. Furthermore, it contains a detailed summary section that describes the main findings of the analyses on sex biased genes.

If it is possible to study differential gene expression underlying the faster evolution of male biased genes with microarrays at the transcript level, it should also be possible to study differential regulation of the proteome by quantitative MS. This generation of male versus female protein expression data represents a potentially very interesting ongoing project in our proteomics laboratory and this project will also be based on SEBIDA.

6.2 Implementation of SEBIDA

Ratios of gene expression levels, recombination rates (Hey and Kliman, 2002), codon bias estimations (Ikemura, 1981; Wright, 1990) and further evolutionary data such as d_N/d_S ratios (see Chapter 9) were integrated into the database, which is implemented in mySQL. The initial data integration modules that have been used in my Master's thesis were mainly in Java. In contrast, the upload and normalisation of very recent datasets are in C#. To compress the data structure, we joined different database relations into one organism-specific comprehensive database relation that stores a multitude of information for each gene. Each tuple is identified by its primary key such as the FlyBase identifier, for example. In contrast to the initial database scheme, the resulting capacious tables include 'null' attributes, if a certain feature is not reported (e.g., the gene was not identified in a certain study). In the previous database schema, the request of this missing data tuple would have resulted into an empty join of several database relations. Besides HTML as the established markup language, we used PHP as the underlying programming language to generate dynamic web pages.

For *Drosophila melanogaster*, Sebida includes male versus female gene expression data using eight different microarray platforms. Five of them (Gibson et al., 2004; Parisi et al., 2004; Parisi et al., 2003; Ranz et al., 2003; Stolc et al., 2004) have already been subject to my Master's study. We have added another three data sets recently (Goldman and Arbeitman, 2007; McIntyre et al., 2006; Wayne et al., 2007). Furthermore, we integrated microarray datasets for *Drosophila simulans* (Ranz et al., 2003) and *Anopheles gambiae* (Hahn and Lanzaro, 2005). Moreover, the additional inclusion of strain or body component specific expression levels provides even more insight into the occurrence of sex bias (Dorus et al., 2006; Mikhaylova et al., 2008).

The web user can search for male versus female expression for any gene of interest via gene symbol, gene name, gene description, EnsEMBL identifier, FlyBase accession, or Affymetrix number (Figure 6.4). The resulting web page illustrates the microarray data including corresponding significance p-values, d_N/d_S ratios and further measures useful for evolutionary analysis. For some studies, the displayed gene expression levels are separated according to the investigated strains or dissected body sections.

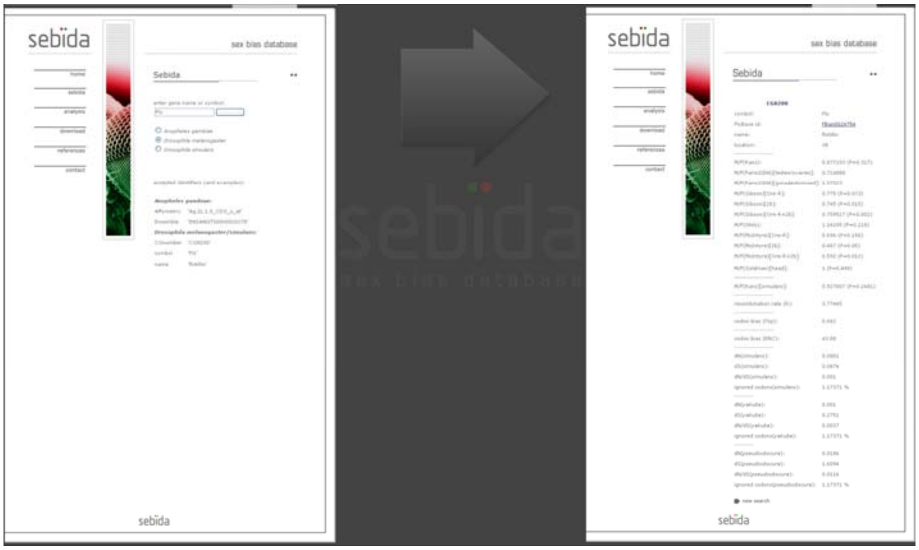


Figure 6.4: SEBIDA – Sex Bias Database

Searching for a gene of interest (left panel) yields a comprehensive report (right panel) about male versus female gene expression and further information providing insight into evolutionary relationships.

Besides the listings of gene specific data that contributes to a better understanding of sex bias along with evolutionary constraints, we integrated a comprehensive analysis section in SEBIDA (Figure 6.5). It describes the main findings of the study on sex biased genes including the observation that male biased genes evolve rapidly and therefore have less orthologous proteins than female biased and unbiased genes. The comprehensive analysis on evolutionary patterns relating to sex bias was subject of my diploma study and they can be looked up via the web application of Sebida, which presents one of the first projects during my PhD study.

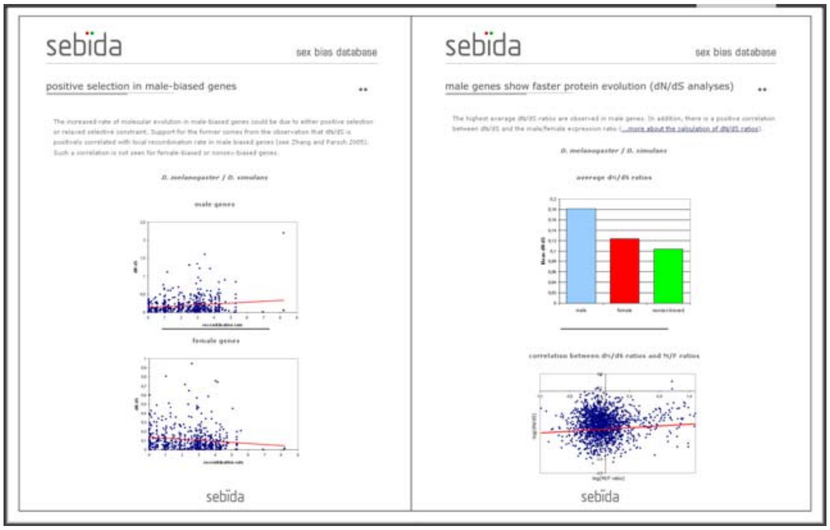

sebida
sex bias database
positive selection in male-biased genes
sebida
sebida
sex bias database
male genes show faster protein evolution (dN/dS analyses)
sebida

Figure 6.5: The analysis section of SEBIDA provides insight into the main findings of the evolutionary analysis of sex biased genes

6.3 Discussion and Future Directions

The development of an online database that focuses on the storage of data related to sex bias makes it possible to perform meta-analyses on various studies comparing male and female gene expression and to derive general patterns relating to difference between sexes. Besides the retrieval of male versus female gene expression data on the basis of microarray technologies, further information relating to evolution has been added. As new large-scale studies are performed every year, Sebida has to be updated and administrated.

To date, the addition of datasets was mainly done by specific ad hoc programs written in Java or C#. Therefore, one major goal is to implement administration tools similar to the ones of PHOSIDA (Chapter 4) and MAPU 2.0 (Chapter 5), in order allow the automated upload of new data and updates of the database. Another goal is the inclusion of high-throughput data on sex bias in species other than insects. The inclusion of other organisms would allow generatizing the observations made in Drosophila.

Finally, the application of mass spectrometry based proteomics is particularly interesting, as one could show that observations on the transcript level are in keeping with findings on the protein level. A MS-based generation of male versus female expressions on the protein level would make a considerable contribution to the sex bias database. In fact, we intend to determine sex-specific protein expression differences in *Drosophila melanogaster* using SILAC in the near future.

Chapter 7

Phosphorylation Site Prediction

The major focus of machine learning is to design algorithms that allow computers to learn (Mitchell, 1997). The general idea is to derive patterns and rules from extensive datasets. In the case of prediction methods, the resulting rules can then be used to classify a given set of new data. Here we took advantage of the large-scale datasets of measured phosphoproteomes (Chapter 4.3), we took advantage of the large number of *in vivo* phosphosites to create a phosphosite predictor in PHOSIDA. The work on this predictor was carried out in collaboration with Shubin Ren in our department.

7.1 Rationale

One aspect of learning is to deduce rules on the basis of given instances. As massive datasets such as MS-based measurements of dynamic proteomes exceed the capacity of human learning ability, the application of computer based machine learning approaches becomes indispensible (Chapter 4.5). Machine learning has a wide spectrum of applications ranging from object recognition to the classification of DNA sequences. Concerning the posttranslational modifications of proteins, various algorithms have already been applied to the prediction of phosphorylation sites. For example, the prediction system Netphos (Nielsen et al., 1999) is based on neural networks, whereas Scansite (Obenauer et al., 2003) uses a profile method to predict phosphorylation events. We used our large-scale studies to construct a phosphorylation site predictor on the basis of a support vector machine (SVM) (Gnad et al., 2007). The basic idea of SVMs is to transform observed features of a given instance into a vector based feature space (Noble, 2006). Each dimension of this feature space presents a certain attribute. Then, after the transformation of a multitude of positive and negative instances (such as phosphorylated and non-phosphorylated residues) into the vector space, a 'maximum margin hyperplane' is created (Figure 7.1). This hyperplane is intended to separate the two datasets. If one intends to classify a new instance, the given sample has to be transformed into the feature space and categorized depending on the vector localization relating to the separating hyperplane. To estimate the accuracy of the SVM, one usually uses 90% of the classified dataset to train the SVM, whereas 10% of the given data is used to test the SVM. We took advantage of the large number of *in vivo* phosphosites from various species to create an organism-specific phosphosite predictor in PHOSIDA.

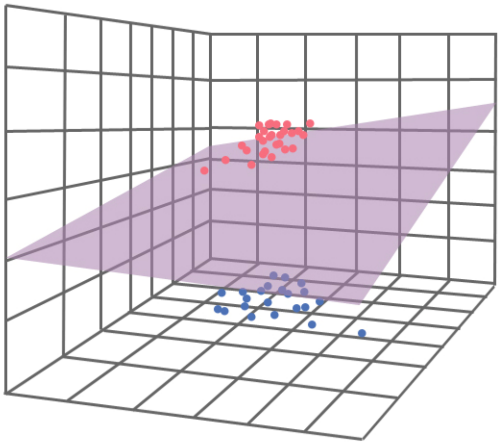


Figure 7.1: Feature space

A maximum margin hyperplane (magenta) separates two distinct datasets that were transformed into a hyperdimensional space reflecting certain features of each instance.

7.2 Implementation of the Support Vector Machine

The large-scale study on general patterns relating to phosphorylation events demonstrates that phosphorylated proteins are highly conserved throughout all phylogenetic kingdoms (Chapter 9). This observation suggests that proteins that undergo posttranslational modifications present functionally important key players of cell signalling processes and therefore have to be preserved in evolution. In addition, a higher conservation of phosphorylated residues in comparison to their non-phosphorylated counterparts was revealed throughout higher eukaryotes. Besides these outcomes on the evolutionary preservation, we have noted the predominant localizations of phosphorylation sites in loops and turns on protein surfaces (Chapter 4.6.4). This finding illustrates the structural constraints of phosphorylated residues to be accessible to certain targets such as kinases or other interacting proteins. Therefore, we used these outcomes on general features of phosphorylation sites to fill the high-dimensional feature space on which support vector machines act.

As shown in the study, phosphoserines, phosphothreonines and phosphotyrosines show the same general patterns relating to protein structure and conservation, but each to a different extent. Therefore, we applied the machine learning approach separately to each organism-specific set of pS, pT and pY sites. To create a negative set of the same size, we randomly chose sites from proteins that were not present in the phosphoset. The positive and negative datasets were split into a training set (90%) and a test set (10%). SVMs attempt to partition true from false sites by separating them in a high dimensional vector space with the help of hyperplanes and kernel functions (see Chapter 7.1). A few sites out of the negative set may turn out to be phosphorylation sites in future experiments. This problem was addressed by optimizing the 'C parameter' of the SVM, which controls the softness of the margin. We optimized the parameters C and σ by varying them from 2^{-10} to 2^{10} in multiplicative steps of two and chose the best combination of both parameters out of the 21 × 21 possibilities. The optimization was based on a five-fold cross validation on the training set. To determine the importance of each feature in the accuracy of phosphosite prediction, we created various sets, which contain different information for each phosphosite:

Set A: The primary sequence comprising the site and its 12 surrounding residues
Set B: The surrounding primary sequence and the predicted secondary structure of the site
Set C: The surrounding primary sequence and the predicted accessibility in addition to the secondary structure of the site
Set D: The surrounding primary sequence, the conservation of the phosphosite in mammals and the protein conservation throughout several eukaryotes
Set E: The surrounding primary sequence, the accessibility of the phosphosite and secondary structure as well as its conservation in mammals, and the protein conservation

This resulted in 260 to 274 dimensions that represent the features of each phosphosite. We investigated several common kernel functions and found that the radial basis function (RBF) turned out to be the most powerful compared to linear, polynomial and sigmoid Kernel functions. We optimized parameters C and σ, the width of the Gaussians used as the RBFs, and trained the optimal model for each set of each phospho amino acid).

We employed the machine learning approach to each organism-specific large-scale phosphorylation data set. Thus, we applied the method to the human (4731 pS, 664 pT, 107 pY), mouse (3733 pS, 437 pT, 83 pY), fly (7756 pS, 1427 pT, 325 pY), yeast (3320 pS, 562 pT, 48 pY), archaean and bacterial phosphoproteomes separately, in order to construct organism-specific phosphosite predictors that are trained on high-accuracy data.

7.3 Results

7.3.1 *Homo sapiens* specific Phosphosite Predictor

We found that the accuracy of the prediction based on the primary sequence was already very high: in the case of phosphoserines, 89.85% were predicted correctly in the test set as were 74.24% of the phosphothreonines. The accuracy of the prediction increased to 90.17% for pS and 77.27% for pT by adding structural information (sets b and c). For serines, the accessibility was slightly more important than the secondary structure information, whereas for threonines, the opposite was the case. The additional dimensions reflecting the conservation of the site and of the entire protein (set d) increased the accuracy to 90.70% (pS) and 81.06% (pT). By combining structural and evolutionary information (set e), we found that 91.75% in the serine set and 81.06% in threonine set were predicted correctly. The accuracy of the prediction of phosphotyrosines increased from 66.67% to 76.19% when including the structural and conservational information. However, that increase is not significant due to the fact that there were only around 100 phosphotyrosines sites.

The recall reflects the proportion of true positives to the sum of true positives and false negatives, whereas the precision describes the number of true positives out of all predicted positives. As outlined in Figure 7.2, the precision-recall curve of set e is slightly better than that of set a, indicating that the inclusion of evolutionary and structural information increased the recall and precision of the prediction to a minor degree.

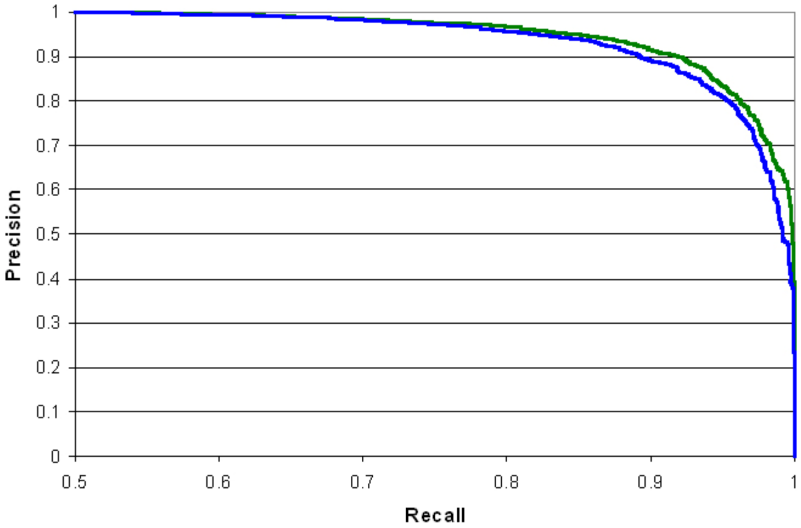


Figure 7.2: Human phosphorylation site prediction

Precision-Recall Curve reflecting the accuracy of the prediction of phosphorylated serines in human on the basis of a support vector machine trained by the surrounding primary sequence (blue) and evolutionary as well as structural constraints (green)

The prediction accuracies of phosphorylated serines and phosphorylated threonines in human are depicted in Figure 7.3 a.

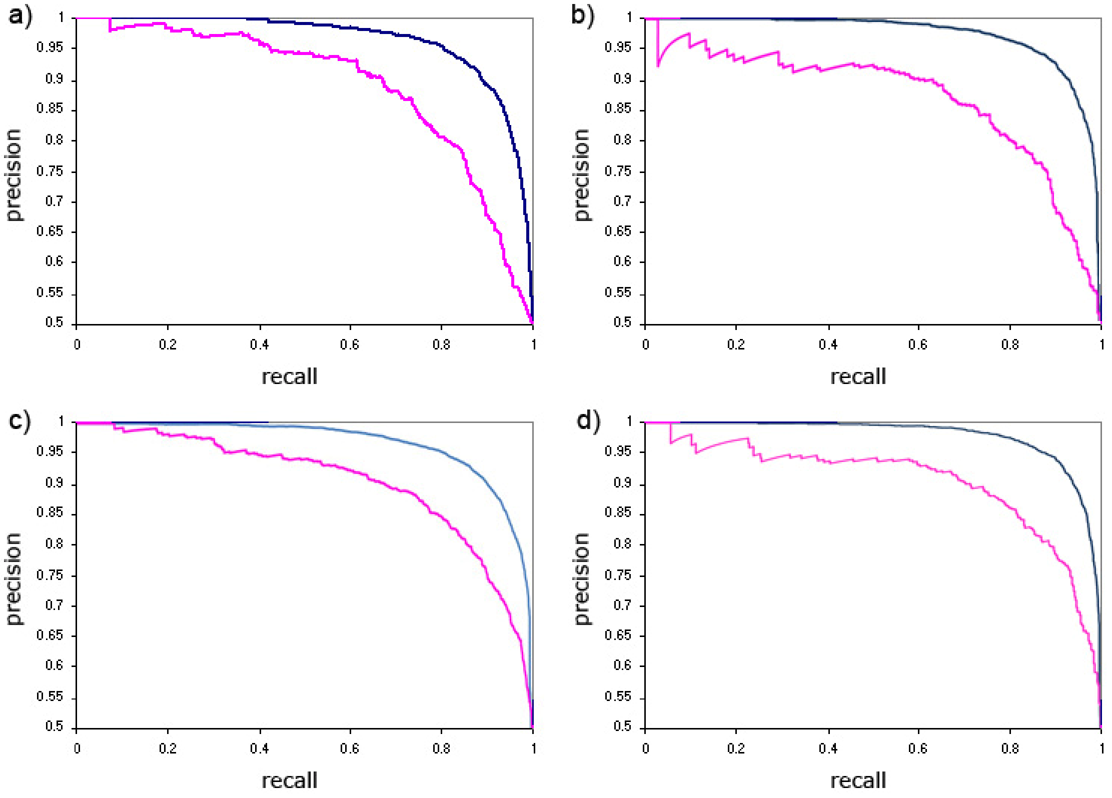


Figure 7.3: Phosphorylation site prediction

Precision Recall curves reflecting the prediction accuracies of phosphoserines (blue) and phosphothreonines (magenta) in human (a), mouse (b), fly (c) and yeast (d).

7.3.2 *Mus musculus* specific Phosphosite Predictor

We trained the support vector machine (SVM) separately on unambiguously identified phosphorylation sites (3733 pS, 437 pT, 83 pY). The essential feature of each phosphorylation site that was used as input for this machine learning approach was the raw sequence, as the main finding of the prediction of human phosphosites showed that the addition of structural and evolutionary information increases the performance of the prediction only slightly (Chapter 7.3.1). In the case of phosphoserines, 88% were predicted correctly in the test set as were 78% of the phosphothreonines. The accuracy of predicting phosphorylated tyrosines was also very high (73%), but lacks statistical significance due to the low number of sites. Figure 7.3 b depicts the accurracy of the prediction of mouse phosphosites. It is comparable to the one of predicting human phosphosites.

7.3.3 *Drosophila melanogaster* specific Phosphosite Predictor

We trained the SVM on 7756 pS, 1427 pT and 325 pY along with their surrounding sequences. We found that 89.81% in the serine set and 81.05% in the threonine set were predicted correctly. The accuracy of the prediction of phosphotyrosines was 63%. The corresponding precision recall curve is illustrated in Figure 7.3c.

7.3.4 *Saccharomyces cerevisiae* specific Phosphosite Predictor

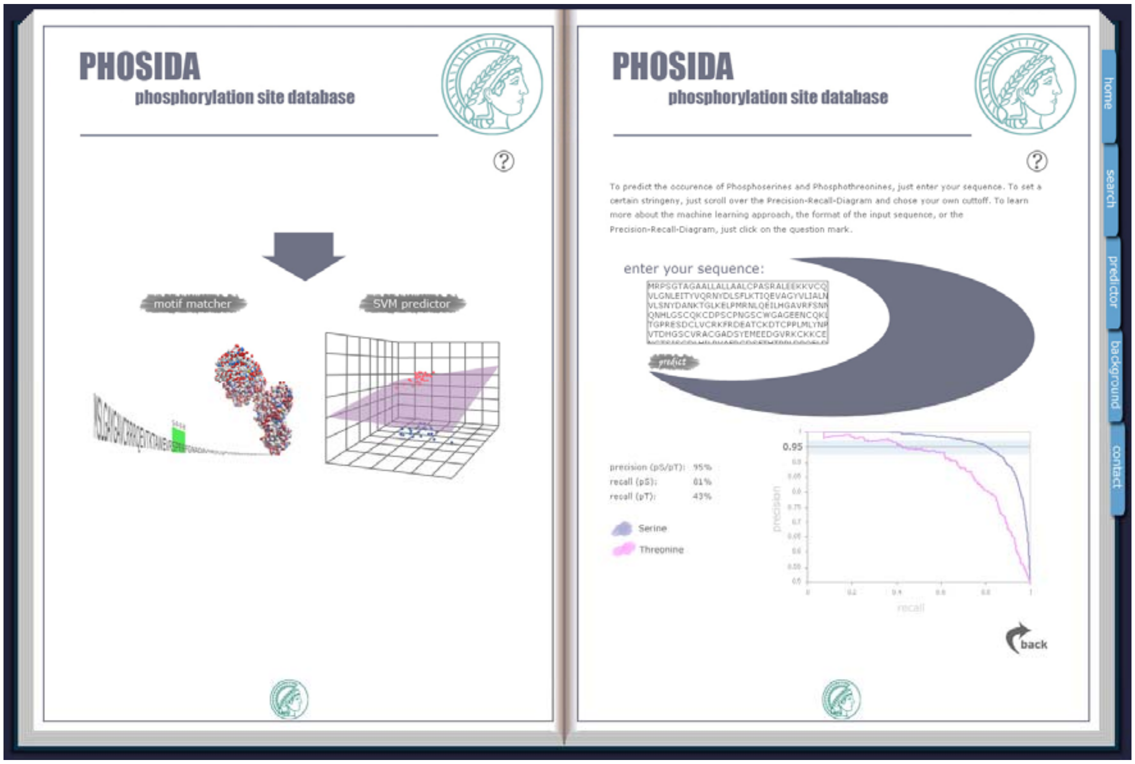
We applied the machine learning approach to 3320 pS, 562 pT, and 48 pY separately. As was the case in the human phosphoproteome, the inclusion of structural and evolutionary information increased the accuracy of prediction only slightly. However the performance of prediction proved to be already very high without such additional information. Therefore, the support vector machine was exclusively trained on the primary sequence comprising the site and its 12 surrounding residues. In total, 92% phosphoserines were predicted correctly in the test set. A high accuracy was also observed for the prediction of phosphothreonines (87%) and phosphotyrosines (66%). Figure 7.3 d shows the Precision-Recall curves for phosphoserine and phosphothreonine prediction in yeast. Due to the low number of phosphotyrosines, a reflection of the performance of the hardly reliable prediction of phosphotyrosines is not demonstrated.

7.3.5 Prokaryotes specific Phosphosite Predictor

Using the phosphoproteomes of various prokaryotes such as *Escherichia coli*, *Lactococcus lactis*, *Bacillus subtilis* and *Halobacterium salinarium*, we tried to train the support vector machine on the basis of the primary sequences surrounding the prokaryotic phosphorylation sites. However, in contrast to the accuracy of the predictions of eukaryotic phosphorylation sites (Chapters 7.3.1 – 7.3.4), the performance of the prokaryotic specific phosphorylation site predictor was very poor and close to random. This could eiter be due to the low number of training sites (100 fold less than in eukaryotes) or it could reflect a different mode of substrate specificity of prokaryot vs. eukaryote kinases.

7.4 Integration of organism-specific Phosphosite Predictors in PHOSIDA

To enable researchers to predict the occurrences of phosphorylation sites on any protein of interest, we created an online version of the predictor in PHOSIDA. It transfers all candidate serines and threonines of any inserted amino acid sequence into the feature space by transforming the site along with the surrounding sequence into a high dimensional vector. The SVM predicts the chance for each residue to be phosphorylated according to the vector orientation in the trained model along with the derived maximum margin hyperplane that separates phosphorylated and non-phosphorylated residues. The web application allows to set a desired cutoff directly on the given organism-specific precision recall curve (Figure 7.4)



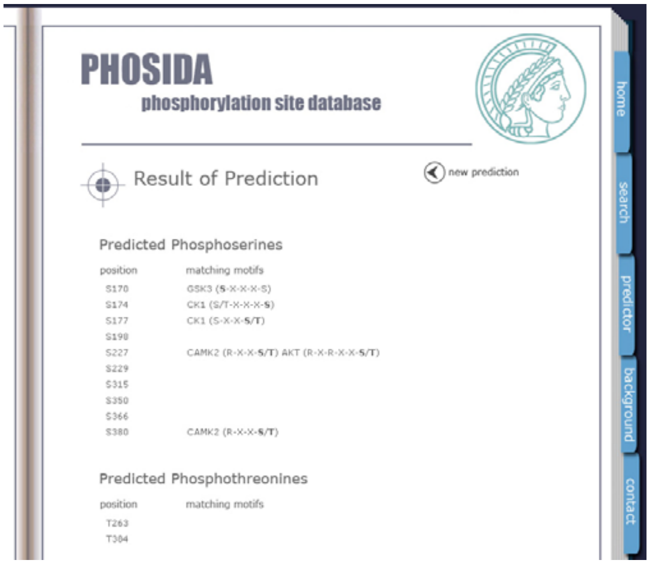
To predict the occurence of Phosphoserines and Phosphothreonines, just enter your sequence. To set a certain stringeny, just scroll over the Precision-Recall-Diagram and chose your own cuttoff. To learn more about the machine learning approach, the format of the input sequence, or the Precision-Recall-Diagram, just click on the question mark.

```
MRPSGTAGAALLALLAALCPASRALEEKKVCQ
VLGNLEITYVQRNYDLSFLKTIQEVAGYVLIALN
VLSNYDANKTGLKELPMRNLQEILHGAVRFSNN
QNHLGSCQKCDPSCPNGSCWGAGEENCQKL
TGPRESDCLVCRKFRDEATCKDTCPPLMLYNP
VTDHGSCVRACGADSYEMEEDGVRKCKKCE
```

precision (pS/pT): 95%
recall (pS): 61%
recall (pT): 43%

Figure 7.4 Integration of the phosphorylation site predictor in PHOSIDA

Each predicted phosphorylated residue is checked for matching with any eukaryotic kinase motif that is included in PHOSIDA. Thus, the listing of all predicted phosphosites also includes all matching kinase motifs, in order to suggest the kinase affiliation (Figure 7.5).



PHOSIDA
phosphorylation site database

Result of Prediction

Predicted Phosphoserines

position	matching motifs
S170	GSK3 (**S**-X-X-X-S)
S174	CK1 (S/T-X-X-X-**S**)
S177	CK1 (S-X-X-**S/T**)
S198	
S227	CAMK2 (R-X-X-**S/T**) AKT (R-X-R-X-X-**S/T**)
S229	
S315	
S350	
S366	
S380	CAMK2 (R-X-X-**S/T**)

Predicted Phosphothreonines

position	matching motifs
T263	
T304	

Figure 7.5: Result of phosphorylation site prediction in PHOSDA

Each predicted phosphosite is tested for matching with known eukaryotic kinase motif to assess the kinase potentially phosphorylating this site.

7.5 Discussion

The organism-specific PHOSIDA phosphorylation site predictor makes it possible to find putative novel phosphorylation sites that have not yet been experimentally identified in yeast, fly, mouse, or human. While experimental data, especially quantitative data, are the ‘gold standard’, predicting novel phosphosites and matching kinase motifs on proteins of interest should be valuable for the design of biological experiments or for predicting a protein’s role in a pathway. Furthermore, once predictors are trained, these prediction methods are basically ‘free’. We provide an innovative method for setting a desired level of precision and recall. For example, for mutagenesis experiments one may want to set the precision very high, and for rationalizing the function of a protein in a pathway one may want to set it relatively low. Thus, in the absence of experimental data, the prediction of novel phosphosites can be taken as the first step in an experimental design to uncover the function of a protein of interest and to elucidate its involvement in signalling cascades.

As new phosphorylation data are integrated to PHOSIDA our SVM will also be updated, leading to increasingly accurate predictions.

Chapter 8

Genome Annotation

The genome is the most comprehensive and fundamental biological resource. It encodes all possible proteins and comprises the entire hereditary information. However, the derivation of coding regions in the nucleotide sequence of the genome is not trivial. Current methods for gene prediction provide useful information but are still limited (Brent, 2007). Furthermore, it is hardly possible to predict all features of the genome from its sequence alone. Thus, the integration and validation of mass spectrometry derived experimental data in a genomic context is expected to contribute to the annotation of the genome and to the identification of genes for which there was no previous experimental information.

8.1 Rationale

The genome encodes the whole hereditary information of an organism. Its fundamental unit is the DNA comprising both genes and non-coding sequences. The first bacterial genome to be completed was that of *Haemophilus influenzae* in 1995 (Fleischmann et al., 1995). Seven years later, the Human Genome Project provided the complete genetic blueprint of a human being by sequencing the whole genome. At the present time, the database GenBank (Benson et al., 2008) contains nucleotide sequences for more than 240000 named organisms obtained primarily through submissions from large-scale sequencing projects. In total, around 2000 eukaryotic genomes have been completely sequenced until now.

These comprehensive high-throughput sequencing efforts establish a basis for the large-scale detection of the encoded proteome powering an organism's life. However, a complete annotation and understanding of the genome requires experimental evidence ranging from the primary observation that a genomic sequence encodes a protein to the measurement of specific features such as residues that are phosphorylated and thereby essential for the regulation of certain biological processes.

As the application of high-accuracy MS technologies potentially enables the measurement of nearly complete proteomes of given cells *in-vivo* along with certain features such as phosphorylation, it is obvious that this large-scale data represent a very valuable source to annotate the genome. Therefore, we endevoured to assign measured MS sequencing information and associated information on posttranslational modifications to the genome. The simplest and fastest way is to map peptide sequences, which have been identified via the established approach of searching MS-information in amino acid sequence databases, to the genome. With the detected peptide sequences in hand, we reassigned all peptides to gene transcripts that are annotated in the genomic database EnsEMBL (Birney et al., 2004). The reassignment of sequence stretches to genes allows the usage of proteomic data to annotate the genome via the DAS/Proserver technology (Finn et al., 2007) in EnsEMBL. In addition, we added extra genome annotation sections in our proteomic databases. Thus, the genome database and the proteomic in-house databases PHOSIDA (Chapter 4) and MAPU (Chapter 5) are linked, so that proteomic data is mapped as features to the genome sequence.

8.2 Mapping Proteomic Data to the Genome

8.2.1 Assignment of MS peptide data to Genes annotated in EnsEMBL

We wished to use the proteomic data to annotate the genome. Thus, we extracted all measured peptides of each proteomic dataset and reassigned the given peptide sequences to genes that are annotated in the EnsEMBL database (Figure 8.1). If a specified peptide matches with sequences of more than one gene transcripts, we assigned the peptide to the one transcript that shows the highest number of matching peptides within the associated experiment. Therefore, the peptide-to-gene transcript assignments result into one-to-one relationships reducing potential redundancy. The reassignment of all detected peptides of various projects to EnsEMBL gene transcripts contributes to the compilation of a new database instance stored on the same web servers that manage the actual MAPU 2.0 proteome database and the phosphorylation site database PHOSIDA.

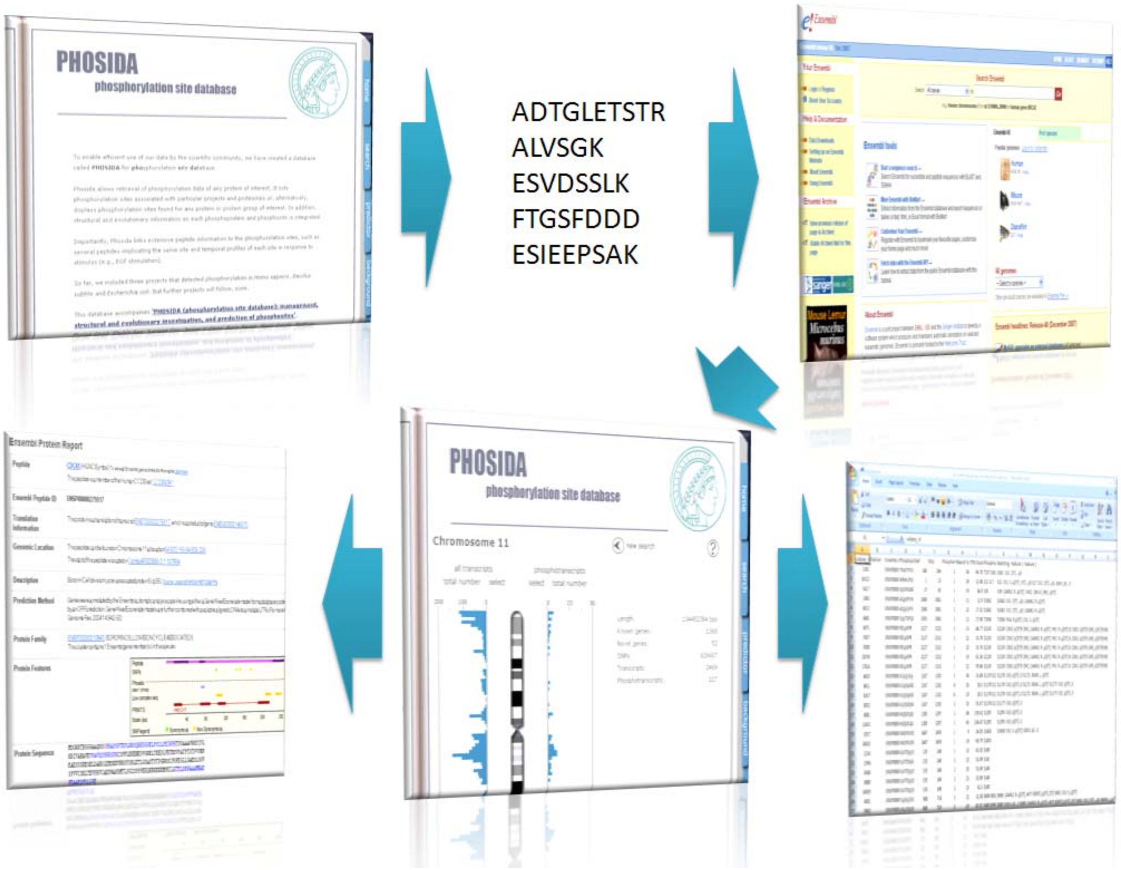


Figure 8.1: Basic genome annotation concept

8.2.2 PHOSIDA and MAPU as Annotation Source in EnsEMBL

The Distributed Annotation System (DAS) allows the visualization of layers of annotation data for a given gene's sequence and thereby makes it possible to gain an overview of the features of that sequence. It presents an excellent technology to integrate annotation data from multiple sources into a simple graphical view in EnsEMBL.

We used Proserver, a Perl-based and standalone DAS Server. At the top of the ProServer architecture (Figure 8.2) is a daemon executable positioned between requests and the resulting code. ProServer comes bundled with modules for data stores ranging from flat file to MySQL. The major method is the source adapter that comprises the data retrieval methods. It had to be adjusted to the data structure of PHOSIDA and MAPU 2.0. Furthermore, it defines the view illustration of the requested data. Its superclass handles the transformation of data to XML. Moreover, the ProServer configuration files had to be set up. The architecture of ProServer is described in detail in Finn et al. (Finn et al., 2007).

Besides the set up of the ProServer, we had to establish Cygwin, a linux-like environment for Windows. It provides a dynamic link library (DLL) acting as a LINUX API emulation layer.

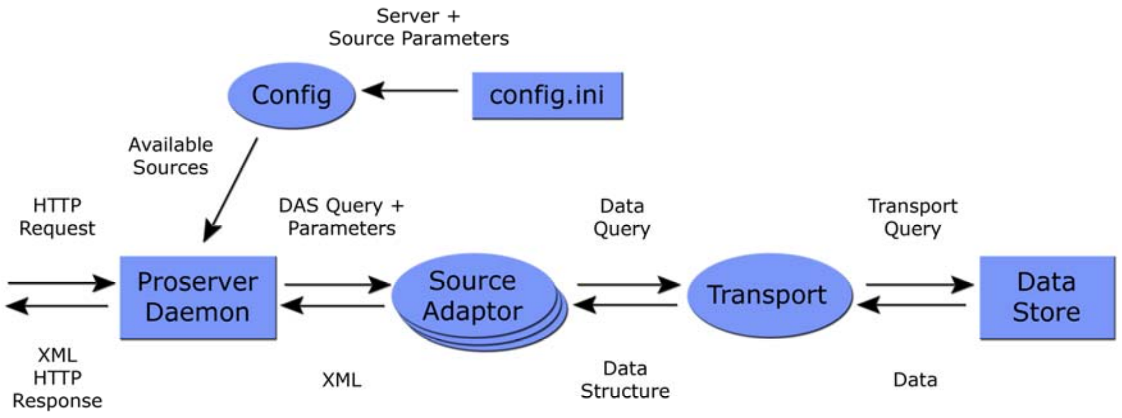
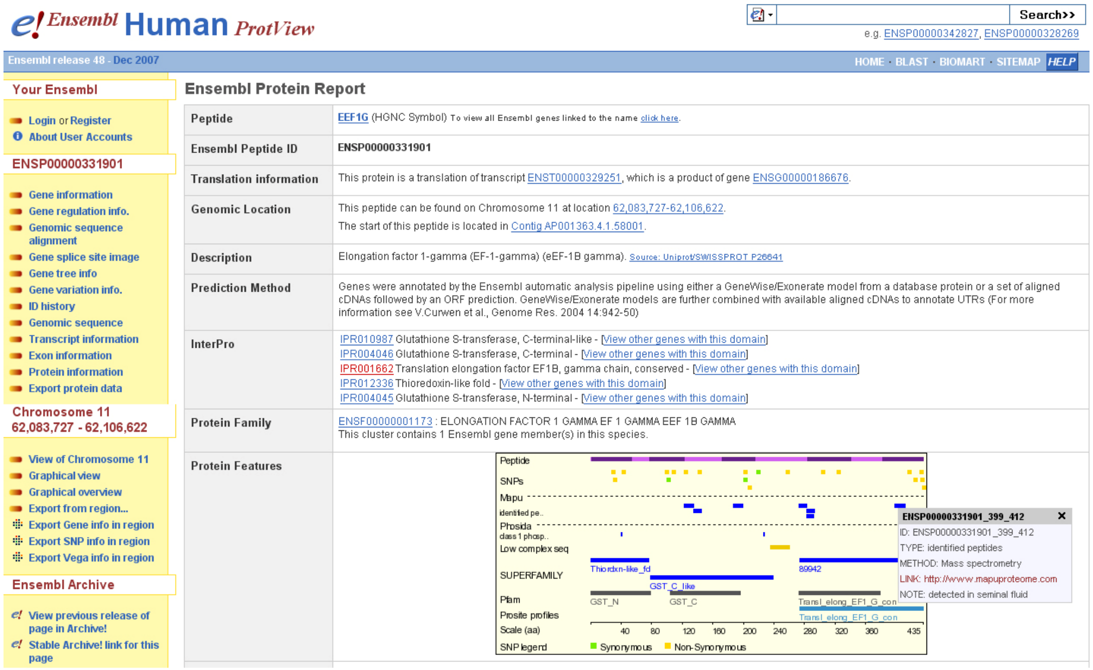


Figure 8.2: Architecture of the ProServer technology

After the installation and set-up of the DAS environment on our servers, web users are able to obtain the gene related data gained by mass spectrometry technologies. For each gene transcript, one layer shows all detected peptides stored in the MAPU 2.0 DAS source (Figure 8.3). Clicking on one of the illustrated peptides yields a report of all the cell types in which the selected peptide has been measured. In addition to the MAPU 2.0 DAS source, the established PHOSIDA DAS source provides all phosphorylation sites that have been unambiguously identified (Class 1 sites), but also phosphosites that lack of precise identification within the phosphorylated peptide sequence due to limited fragmentation (ambiguous PTM localization).

The aggregate view of all displayed features of the genome sequence enables researchers to obtain a summary of the genes' sequence characteristics and can already lead to insights or hypotheses into the biological function of the gene.

The background sections of PHOSIDA and MAPU 2.0 contain detailed guide lines about the set up of our DAS sources in EnsEMBL.



Ensembl Protein Report

Peptide	EEF1G (HGNC Symbol) To view all Ensembl genes linked to the name click here.
Ensembl Peptide ID	ENSP00000331901
Translation information	This protein is a translation of transcript ENST00000329251, which is a product of gene ENSG00000186676.
Genomic Location	This peptide can be found on Chromosome 11 at location 62,083,727-62,106,622. The start of this peptide is located in Contig AP001363.4.1.58001.
Description	Elongation factor 1-gamma (EF-1-gamma) (eEF-1B gamma). Source: Uniprot/SWISSPROT P26641
Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise/Exonerate model from a database protein or a set of aligned cDNAs followed by an ORF prediction. GeneWise/Exonerate models are further combined with available aligned cDNAs to annotate UTRs (For more information see V.Curwen et al., Genome Res. 2004 14:942-50)
InterPro	IPR010987 Glutathione S-transferase, C-terminal-like - [View other genes with this domain] IPR004046 Glutathione S-transferase, C-terminal - [View other genes with this domain] IPR001662 Translation elongation factor EF1B, gamma chain, conserved - [View other genes with this domain] IPR012336 Thioredoxin-like fold - [View other genes with this domain] IPR004045 Glutathione S-transferase, N-terminal - [View other genes with this domain]
Protein Family	ENSF00000001173 : ELONGATION FACTOR 1 GAMMA EF 1 GAMMA EEF 1B GAMMA This cluster contains 1 Ensembl gene member(s) in this species.
Protein Features	

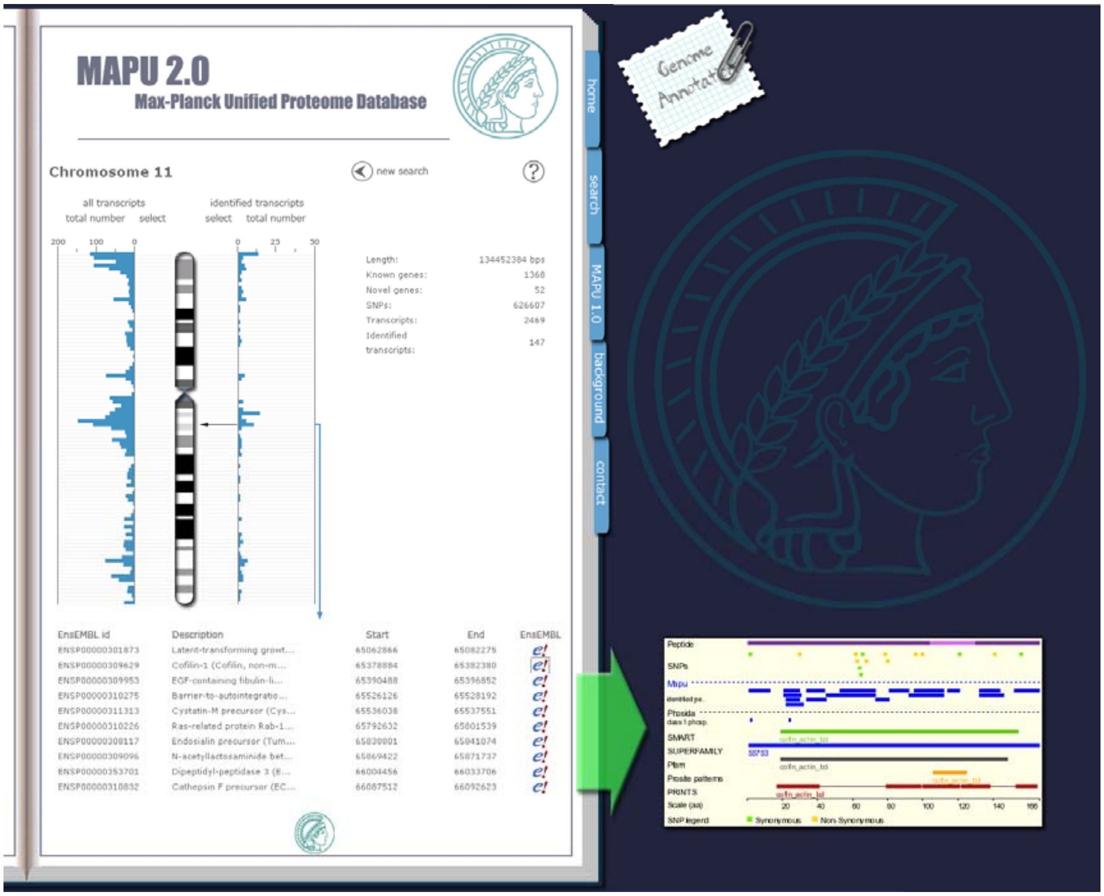
ENSP00000331901_399_412
ID: ENSP00000331901_399_412
TYPE: identified peptides
METHOD: Mass spectrometry
LINK: http://www.mapuproteome.com
NOTE: detected in seminal fluid

Figure 8.3: EnsEMBL gene report including the PHOSIDA and MAPU 2.0 DAS protocols

8.2.3 Representation of Genome Annotations in PHOSIDA and MAPU 2.0

The genome annotation section is accessible via the ‘notepad’ button located right next to the main ‘web book’ of MAPU 2.0 and PHOSIDA (Figure 8.4). At first, the user is required to select a species of interest. Then, the karyotype of the selected species is illustrated along with a link that connects to the EnsEMBL genome annotation webpage. Clicking on one of the displayed chromosomes shows a more detailed image along with general information such as length of the chromosome, number of known and predicted genes, number of single nucleotide polymorphisms (SNPs), and number of gene transcripts. Besides these annotations derived from the EnsEMBL database, the number of gene transcripts that have been identified in the proteomic data is posed.

Furthermore, each chromosome is divided up into 93 bins: On the left side, the number of transcripts that are annotated in EnsEMBL is displayed. Clicking on one of the bin boxes pops up the EnsEMBL web page showing a detailed view of the selected chromosomal region. On the right side the number of transcripts that have been detected in any of the uploaded proteomics projects is illustrated for each bin. Clicking on one of these bin buttons lists all identified gene transcripts along with the descriptions of the corresponding genes and their exact localizations on the chromosome. Furthermore, a link is provided for each gene transcript that connects to the EnsEMBL homepage displaying the full annotation of the given transcript. In addition to the general annotation of the given gene transcript, the popped up EnsEMBL page will show all peptides that have been identified by proteomics via the MAPU 2.0 DAS source and all detected phosphorylation sites via the PHOSIDA DAS source (see Chapter 8.2.2).



Chromosome 11

Length:	134452384 bps
Known genes:	1368
Novel genes:	52
SNPs:	626607
Transcripts:	2469
Identified transcripts:	147

EnsEMBL id	Description	Start	End	EnsEMBL
ENSP00000301873	Latent-transforming growt...	65062866	65082275	e!
ENSP00000309629	Cofilin-1 (Cofilin, non-m...	65378884	65382380	e!
ENSP00000309953	EGF-containing fibulin-li...	65390488	65396852	e!
ENSP00000310275	Barrier-to-autointegratio...	65526126	65528192	e!
ENSP00000311313	Cystatin-M precursor (Cys...	65536038	65537551	e!
ENSP00000310226	Ras-related protein Rab-1...	65792632	65801539	e!
ENSP00000308117	Endosialin precursor (Tum...	65830001	65841074	e!
ENSP00000309096	N-acetyllactosaminide bet...	65869422	65871737	e!
ENSP00000353701	Dipeptidyl-peptidase 3 (E...	66004456	66033706	e!
ENSP00000310832	Cathepsin F precursor (EC...	66087512	66092623	e!

Figure 8.4: Genome annotation section of MAPU 2.0

8.3 Results

The main finding of the genome annotation approach is that nearly all peptide sequences that were identified on the basis of an underlying protein database could be assigned to genes annotated in EnsEMBL (Table 8.1). On average, less than 1% of peptide sequences did not match with any translated gene transcript sequence. In the case of the fly phosphoproteome and the yeast phosphoproteome, all peptide sequences could be assigned to translated gene transcripts annotated in EnsEMBL.

Another outcome of this method was that most of the assigned genes are known, whereas less than one percent of identified genes encoding phosphoproteins are annotated in EnsEMBL on the basis of predictive methods lacking of experimental evidence. In the case of the human phosphoproteome identified in cancer cells upon EGF stimulus (Chapter 4.6.1.1.1), a mere 0.4% of determined genes were novel (predicted). In the case of yeast and fly, we did not detect any genes that are classified as 'novel' in EnsEMBL.

Furthermore, we did not find evidence for significant biased distributions of identified phosphotranscripts encoded on the plus or minus DNA strand. The localizations of detected genes on the chromosomes also showed a uniform pattern. When analyzing the occurrences of detected genes on the different chromosomes (Figure 8.5), we noticed that none of the phosphopeptides mapped to the Y-chromosome. This is a good positive control, as the HeLa cell line is female and therefore should not express any genes from the Y-chromosome.

Another finding of this study was that the number of assigned genes is lower than the number of determined proteins. For example, in total, 2200 proteins were associated to 1982 genes in the case of the investigated human phosphoproteome. This is due to the fact that one gene can give rise to several distinct proteins that are distinguishable to MS.

The main results of the genome annotation approach applied to eukaryotic phosphoproteomes are shown in Table 8.1.

Category	Human (4.3.1.1)	Human (4.3.1.2)	Mouse (4.3.2.1)	Mouse (4.3.2.2)	Fly (4.3.3)	Yeast (4.3.4)
Phosphoproteins	2200	1377	1808	2250	2285	1192
Genes	1982	1303	1729	2181	2280	1190
Known genes	1974	1300	1719	2171	2280	1192
Novel genes	8	3	10	10	0	0
Genes located on the + DNA strand	997	656	887	1105	1165	603
Genes located on the – DNA strand	985	647	842	1076	1135	587
Phosphopeptides	5569	3898	3430	5250	8777	3000
Phosphopeptides assigned to genes	5460	3835	3378	5207	8777	3000

Table 8.1: Genome annotation using identified phosphoproteomes

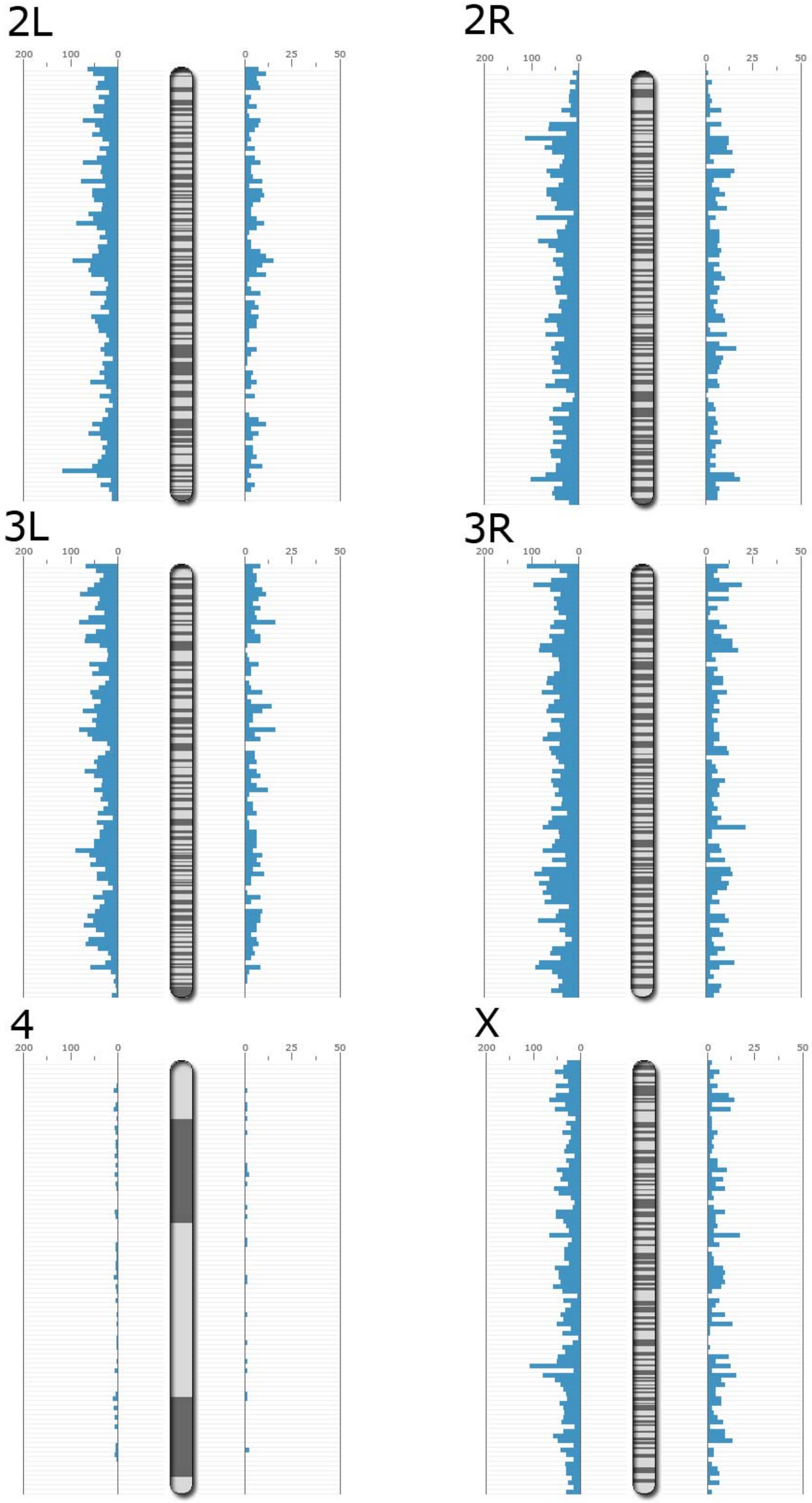


Figure 8.5: Annotation of the fly genome

The distribution of genes that encode phosphorylated proteins (right) on fly chromosomes is similar to the one of all genes (left)

8.4 Discussion

It is not surprising that almost all human and mouse phosphopeptides could be assigned to their genes, as the protein database used for mapping the spectra to peptide sequences was the IPI database comprising all translated gene transcripts of the EnsEMBL database. In the case of yeast and fly, the corresponding databases used for identification (SGD and FlyBase respectively) are completely integrated into EnsEMBL, so that all peptide sequences could be mapped. The observation that 2200 human phosphoproteins correspond to 1982 genes reflects the fact that each gene can potentially encode various transcripts and isoforms of the protein product. The chromosome localization of genes that encode phosphoproteins does not significantly deviate from the localization of all genes. This was expected, as there is no plausible reason for a preferred localization of phosphogenes on a certain DNA strand or chromosome, as is the case for sex biased genes (Chapter 6). For the human phosphoproteome, the absence of measured proteins, whose genes are on the Y chromosome, is related to the experimental design, as we used HeLa cells. The HeLa cell line was derived from cercival cancer cells taken from a woman named Henrietta Lacks. Thus, the derivation of gene products, whose origins are located on the Y chromosome, would point to contamination by a male mass spectrometrist such as Jesper Olsen, for example. Therefore, this genome annotation approach can also be used for quality control.

Besides general conclusions regarding the number of detected gene transcripts on certain regions of the genome, the genome annotation approach makes it possible to integrate large scale MS based proteomic data into the genome database EnsEMBL. The compilation of general gene annotation extended by proteomic data on the basis of the DAS technology enables biologists to visualize a variety of gene features. Moreover, the linkage between our proteomic databases and the genome database allows the discovery of other patterns relating to phosphorylation. For example, below we use annotation data included in the EnsEMBL Compara database to elucidate the evolution of phosphorylation.

Chapter 9

The Evolution of Phosphorylation

As described in Chapter 4.2.4, we integrated evolutionary conservation as another dimension of biological information of the phosphoproteome into PHOSIDA. Phylogenetic relationships and global sequence alignments of homologous proteins elucidate the conservation of given phosphorylated proteins and phosphorylated sites of interest. It also enables the analysis of the evolution of phosphorylation from a global point of view. For this purpose we either used protein-protein alignments of phospho datasets obtained via the automated PHOSIDA pipeline, or the comprehensive evolutionary information that is provided by the EnsEMBL Compara database. In order to use the Compara database, we made use of the mapping of phosphopeptides to genes in PHOSIDA (Chapter 8).

9.1 Rationale

Evolution is a change in the inherited traits of a population from one generation to the next (Futuyma, 1998; Ridley, 2003). These traits can be classified as the ultimate effects of all proteins that are encoded by genes. DNA contains the genetic instructions and therefore presents the long-term storage of genetic information, which is passed on by reproduction. However, mutations in specific regions of the DNA can change the encoded traits or even create novel traits. If the resulting changes have a negative effect on the chance of survival or decreases the chance to reproduce, the genetic alteration is sorted out. This phenomenon is defined as ‘negative selection’. In contrast, if the changes in the DNA have a positive impact on the probability of survival or reproduction this is ‘positive selection’. Over many generations, adaptions result from the genetic preservation of positively selected traits that are advantageous in a given environment. In contrast, ‘genetic drift’ causes random changes in the frequency of traits in a population. Therefore, natural selection and genetic drift present the predominant forces that drive the evolution of species via mutations.

Evolution not only advances the design and development of traits within one species, but also causes the generation of new species. This evolutionary process by which new biological species arise is termed ‘speciation’. The main cause of speciation is geographic isolation. This evolutionary process has yielded into a great variety of species over billions of years. With the availability of completely sequenced genomes of various species, one can compare the genomic sequences between species, and therefore derive their evolutionary relationship and suggest their phylogenetic division. The overall phylogenetic relationship of all species yields the tree of life (Figure 9.1) (Ciccarelli et al., 2006). Evolution can be analyzed on different levels ranging from the evolution of entire genomes of species as a whole to the preservation of specified protein sequence segments. For example, the intent of my Master's study was to analyze the evolution of sex biased genes integrated in the sex bias database SEBIDA (Chapter 6). On the basis of extensive conservation analyses on the DNA level, we came to the conclusion that male biased genes evolve faster in evolution than female biased genes. The biological reasons for this observation are the phenomena of 'female choice' and 'male competition'. This project illustrated that evolutionary analysis on the basis of bioinformatics methods including extensive sequence alignment approaches enables derivation of patterns regarding the functional impact of proteins and their preservation over time.

Here, we intended to study the evolution of phosphorylation. Although conservation of specific sites is often taken to imply biological importance, relatively little is known about the evolutionary constraints on the phosphoproteome. We investigated these constraints on three levels: conservation of phosphoproteins, regions surrounding the site and the phosphosite itself. We used the phylogenetic relationships derived from two-directional BLAST searches and pairwise global alignments created via the Needleman-Wunsch algorithm, which are integrated into the PHOSIDA database, to study the evolution of phosphorylation. In addition, we linked the PHOSIDA database containing gene assignments of phosphorylation sites with the EnsEMBL Compara database, which contains a very large amount of evolutionary information at the DNA level. Combining evolutionary annotation data from EnsEMBL with phosphoproteomic data managed by PHOSIDA enables the investigation of the evolution of phosphorylation at the genome sequence level. Conservation studies at the DNA level are even more comprehensive than at the protein sequence level, as one can calculate synonymous and non-synonymous changes of the gene sequence This makes it possible to learn whether a given gene is positively or negatively selected by comparing evolutionary rates of synomymous and non-synomymous changes in the coding sequence.

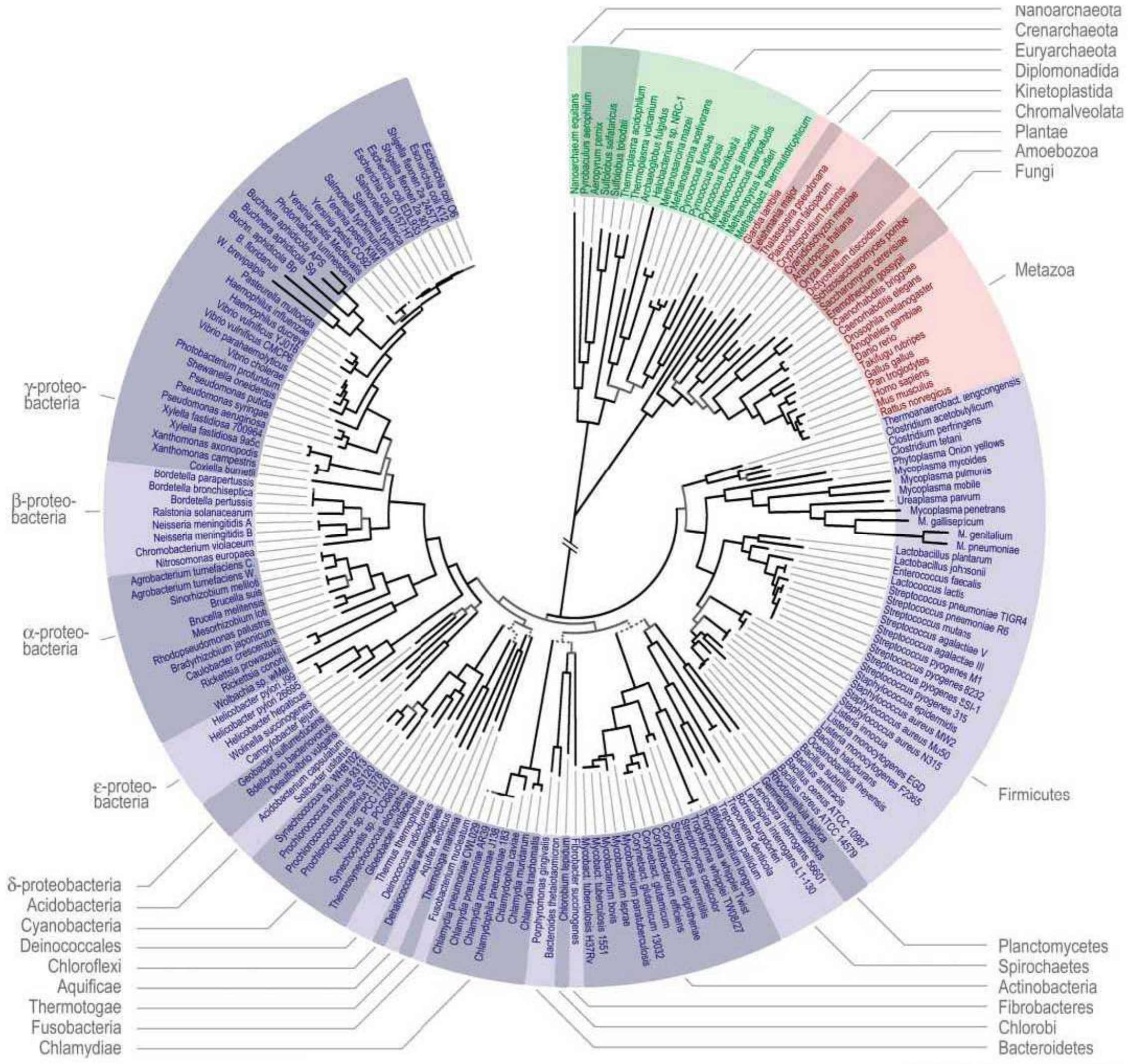


Figure 9.1: Tree of life (Ciccarelli et al., 2006)

9.2 Derivation of Phylogenetic Relationships and Global Alignments

The evolutionary analysis is divided up into two parts (Figure 9.2): The first part describes the evolutionary analysis of phosphorylation events at the DNA level.

The EnsEMBL Compara database contains whole genome alignments, ortholog predictions, paralog predictions, and various phylogenetic parameters describing the inter-species and intra-species homology. It is freely accessible via user accounts that allow complete access to all associated database relations. Hence, we integrated the EnsEMBL Compara database access along with standard queries on required database relations into the employed C# class library. However, some database queries would require complex workarounds. The retrieval of cDNA alignments between gene transcripts would require the composition of certain segments of the whole genome alignments containing associated exon regions, for instance. Such complex queries cannot be formulated by mySQL commands alone. Hence, EnsEMBL also provides access to a Perl API application, which embraces a multitude of methods that allow the formulation of more complex queries on the database. Therefore, we also embedded required methods of the EnsEMBL Perl API into the PHOSIDA analysis pipeline. The integrated access to the genome database enabled us to retrieve comprehensive evolutionary information about genes encoding proteins that we found to be phosphorylated. The evolutionary analysis of phosphorylation at the gene level is fully automated and accessible via the PHOSIDA administration tool (Chapter 4.2.6).

In the second part, we integrated an evolutionary conservation section into PHOSIDA via a self-implemented pipeline (Chapter 4.2.4). We determined homologous proteins to all phosphoproteins across 70 species from *E.coli* to human. The homology search was performed against protein databases of 53 bacteria, nine archaea, and eight eukaryotes. These databases were retrieved from SwissProt in the case of archaea and bacteria. The yeast proteome was downloaded from SGD, *D.melanogaster* from FlyBase and other eukaryotic sequences from IPI. We defined proteins to be homologous when the resulting E-values were lower than 10^{-5}. For homologous proteins, we used a bidirectional BLASTP approach to distinguish between paralogs and orthologs (O'Brien et al., 2005). PHOSIDA displays the results of the homology search using an approximate phylogeny of all investigated species (Chapter 4.5). In addition, we created global alignments between each phosphoprotein and its corresponding interspecific homolog via the Needleman-Wunsch algorithm. As the length of alignments presents a further criterion for homology besides bidirectional significance, web users are able to check the global alignments along with the proportion of identities and to estimate the degree of homology by themselves.

For the global evolutionary study, we implemented various analyses that require only the project identifier of the given experiment and different parameters such as a minimum length of the pairwise alignment for defining homology. Overall, the integration of bidirectionally derived phylogenetic relationships and global alignments between phosphoproteins and homologs allows testing protein, kinase motif and phosphosite conservation on line for any phosphoprotein of interest. Additionally, it enables analysis of the evolutionary constraints on the phosphoproteome on different levels on the basis of protein-protein alignments from a global point of view.

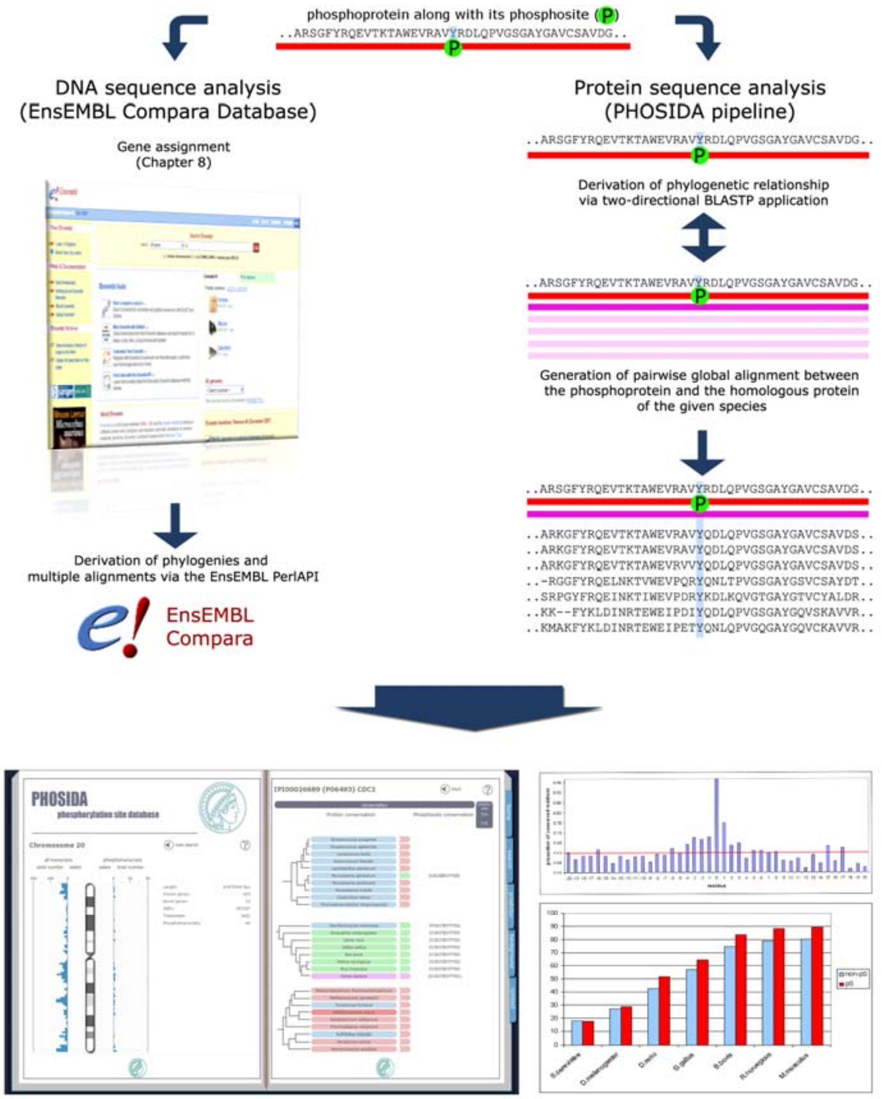

phosphoprotein along with its phosphosite (P)
..ARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVCSAVDG..
P
DNA sequence analysis
(EnsEMBL Compara Database)
Gene assignment
(Chapter 8)
Derivation of phylogenies and
multiple alignments via the EnsEMBL PerlAPI
EnsEMBL
Compara
Protein sequence analysis
(PHOSIDA pipeline)
..ARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVCSAVDG..
Derivation of phylogenetic relationship
via two-directional BLASTP application
..ARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVCSAVDG..
Generation of pairwise global alignment between
the phosphoprotein and the homologous protein
of the given species
..ARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVCSAVDG..
..ARKGFYRQEVTKTAWEVRAVYQDLQPVGSGAYGAVCSAVDS..
..ARKGFYRQEVTKTAWEVRAVYQDLQPVGSGAYGAVCSAVDS..
..ARKGFYRQEVTKTAWEVRVVYQDLQPVGSGAYGAVCSAVDS..
..-RGGFYRQELNKTVWEVPQRYQNLTPVGSGAYGSVCSAYDT..
..SRPGYFRQEINKTIWEVPDRYKDLKQVGTGAYGTVCYALDR..
..KK--FYKLDINRTEWEIPDIYQDLQPVGSGAYGQVSKAVVR..
..KMAKFYKLDINRTEWEIPETYQNLQPVGQGAYGQVCKAVVR..
PHOSIDA
phosphorylation site database
Chromosome 20
IPI00026689 (P06493) CDC2
S.cerevisiae
D.melanogaster
D.rerio
G.gallus
B.taurus
R.norvegicus
M.musculus
non-pS
pS
0
10
20
30
40
50
60
70
80
90
100

Figure 9.2: Investigation of evolutionary constraints of the phosphoproteome

On the one hand, inter-species and intra-species phylogenetic relationships were derived along with global cDNA alignments using the EnsEMBL Compara database (left panel). The comprehensive annotation describing the evolution and conservation of genes required the assignment of proteomic data to the genome. On the other hand, we derived homology relationships between phosphorylated proteins and proteins of other species via bidirectional BLAST alignments (right panel). To obtain global alignments of homologous proteins, we applied the Needleman-Wunsch algorithm. The integration of evolutionary information of phosphoproteins allows gaining insight into the conservation of any protein of interest on three levels ranging from the protein conservation as a whole to the phosphosite conservation. In addition, it enables the analysis of evolutionary constraints of the phosphoproteome from a global point of view.

9.3 Results

On the basis of the EnsEMBL Compara database, we explored the conservation of 1982 human genes, which encode proteins that we found to be phosphorylated in our study (Olsen et al., 2006). As shown in Figure 9.3a, phosphorylated gene transcripts (proteins) have a higher proportion of homologs, which are classified as 'one to one orthologs' by EnsEMBL, in comparison to the entire human proteome. We found that 65% of human genes, which encode non-phosphorylated proteins, were orthologous to genes in *Canis familiaris* (dog) in comparison to 85% of the phosphoset, for instance. For the comparison set, we took only genes into account that have been experimentally proven to code proteins. In EnsEMBL, experimentally verified proteins are classified as 'known' in the corresponding database relation and are therefore easily retrievable. In contrast, predicted genes, which lack of any experimental evidence, are defined as 'novel' in the EnsEMBL Compara database.
In addition, we also examined the conservation of the kinase enriched human phospho dataset (Chapter 4.6.1.1.2). The genome annotation approach yielded 1303 genes encoding phosphorylated proteins (Chapter 8). Genes encoding proteins whose phosphorylation dynamics could be measured in different cell cycle phases also proved to be more conserved than other human genes that are annotated in the EnsEMBL database. Interestingly, in lower eukaryotes the phosphoproteome measured in different cell cycle phases proved to be more conserved than the one identified in cells exposed to EGF stimulation (Figure 9.3a).

Additionally, we explored the conservation of the mouse phosphoproteome using the dataset of 1729 genes encoding proteins that we detected in liver cells exposed to phosphatase inhibition and the dataset of 2181 genes encoding proteins identified in mouse melanoma tissue. As observed in the case of human, the main finding of the conservation analysis at the protein level was that the identified mouse phosphoproteome showed significantly more orthologs throughout 36 other eukaryotes from rat to yeast (Figure 9.3b). In the case of *Loxodonta africana* (elephant), for example, around 70% of all phosphoproteins in both datasets had orthologs in comparison to 53% of all other mouse proteins. Interestingly, in the case of more distantly related species including several fishes, insects, worm and yeast, we found evidence for a higher conservation of the phosphoproteome measured in liver cells upon phosphatase inhibition compared to the conservation of the phosphoproteome identified in mouse melanoma cells (Figure 9.3b).

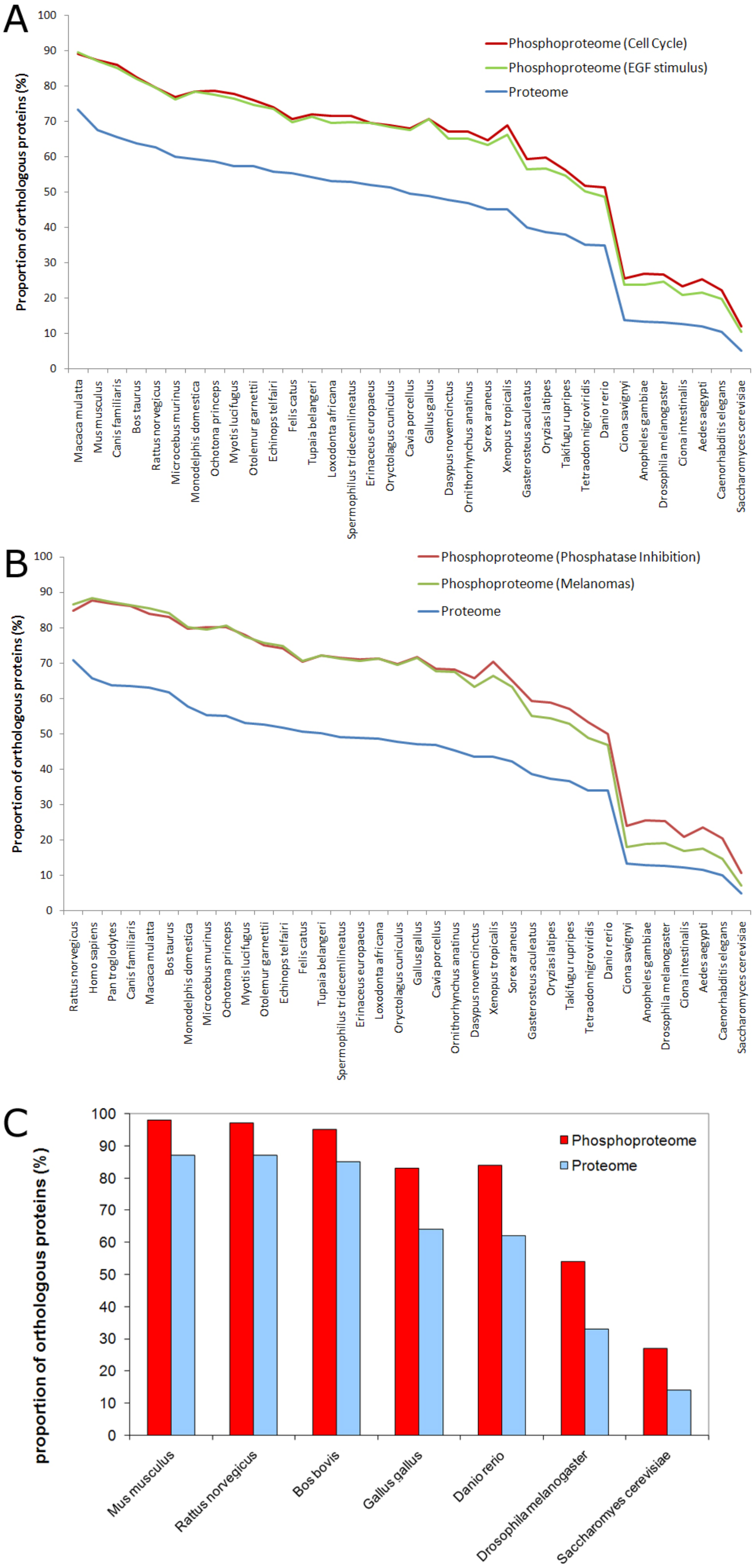


Figure 9.3: Proportions of phosphoproteins with orthologs

Moreover, we checked the conservation of *Drosophila melanogaster* genes (Chapter 4.6.1.3) and *Saccharomyces cerevisiae* genes (Chapter 4.6.1.4) that encode phosphorylated proteins. Similar to higher eukaryotes, we found evidence for a higher conservation of the phosphoproteome in both cases. Overall, 71% of identified phosphorylated proteins in fly showed orthologous proteins in mosquito in comparison to 51% of non-phosphorylated proteins.

To confirm the generality of this observation on the basis of the phylogenetic relationships derived from the two-directional BLASTP approach, we investigated human phosphoproteins (Chapter 4.6.1.1.1) that had an exact sequence match in the SwissProt database. This resulted in a set of 1044 human phosphoproteins. As is apparent from Figure 9.3c, phosphorylated proteins have a higher proportion of two-directionally conserved interspecific homologs (χ^2 test, $p = 0$) in comparison to the entire human proteome (complete human SwissProt database), presumably reflecting regulatory functions that are preferentially conserved during evolution. For example, in the case of *Danio rerio* alignments, we observed that 63% of all human proteins had orthologs in comparison to 84% of the phospho proteins.

Next, we wanted to measure the selective pressure on phosphorylated proteins during evolution. The ratio of non-synonymous to synonymous divergence (d_N and d_S, respectively) indicates whether a given gene is positively or negatively selected. Low d_N/d_S ratios (smaller than one) suggest negative selection implying that there is a high selective pressure in evolution to keep the specified protein unmodified and to select out any mutations changing the amino acid composition. High d_N/d_S levels point to positive selection, as non-synonymous changes were favoured and retained by evolution. Synonymous and non-synonymous changes can only be calculated on the basis of DNA sequences in the coding region, as nucleotide mutations that do not affect the amino acid translation (synonymous change) cannot be derived from protein sequences. The EnsEMBL Compara database provides d_N and d_S values for all coding sequence alignments between homologous genes. The interpretation of d_N/d_S ratios is only reasonable in the case of genes that are orthologous to each other between closely related species. Coding sequence alignments of homologous genes originating from very distantly related species potentially contain multiple silent mutations or diverged to such an extent that the comparison of synonymous and non-synonymous changes does not make sense anymore.

Therefore, we retrieved the d_N/d_S values derived from alignments between human coding genes and genes from closely related species ranging from chimp to mouse. Interestingly, in each case, the median d_N/d_S ratios were significantly lower for genes encoding phosphoproteins in comparison to genes encoding non-phosphorylated proteins (Figure 9.4). The median ratio of non-synonymous to synonymous divergence between human genes and their homologous genes in chimp was 0.29 in comparison to 0.19 in the case of genes encoding phosphoproteins. These findings were in concordance with observations from phosphoproteomes of other species. For example, the median d_N/d_S value derived from alignments between fly genes and orthologous mosquito genes was 0.33 compared to 0.25 in the case of genes coding phosphoproteins.

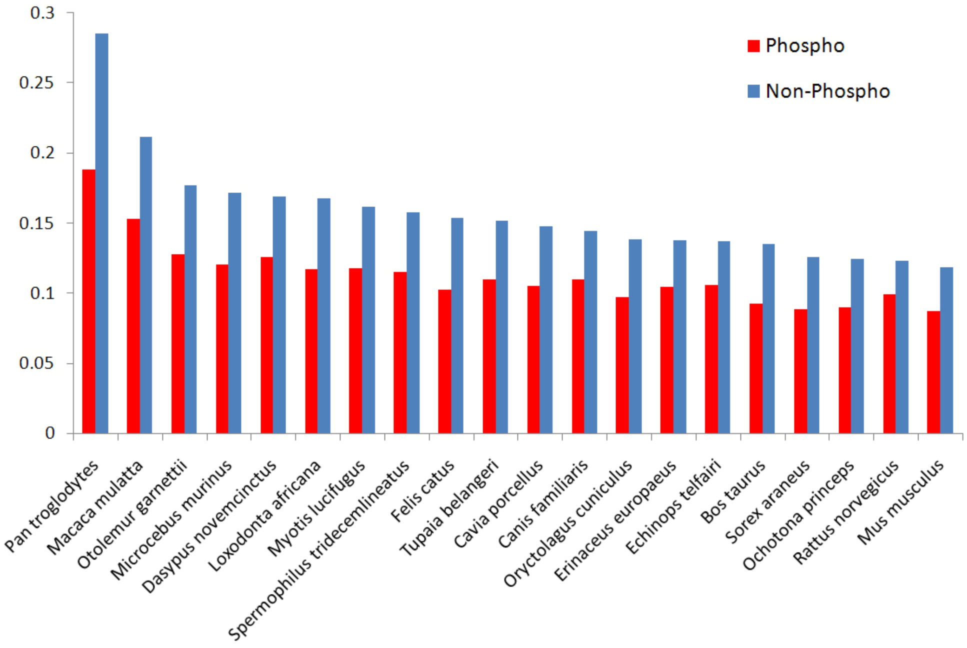


Figure 9.4: Median d_N/d_S ratios of human genes that encode phosphorylated (red) and non-phosphorylated proteins (blue) reflecting evolutionary selective pressure

To investigate the high conservation of the phosphoproteome on the intra-species level, we analyzed the degree of paralogy in phosphorylated proteins versus non-phosphorylated proteins. Paralogous genes are indicated as ‘within-species paralog’ in the EnsEMBL database. Thus, the section of the PHOSIDA analysis pipeline that examines the conservation of the phosphoproteome contains embedded queries that links the phosphorylation site database with the Compara database and estimates the proportion of genes that show intra-species paralogy. Except for the yeast phosphoproteome, the proportions of paralogous phosphoproteins were lower in all the phosphoproteomes in PHOSIDA. In fly (Chapter 4.6.1.3), 27% of the identified phosphoproteins were paralogous to another fly gene in comparison to 45% in the case of non-phosphorylated proteins (Figure 9.5). The higher proportion of paralogous non-phosphorylated proteins was also evident in higher eukaryotes, but to a minor degree: Overall, 61% of phosphorylated proteins identified in mouse cells exposed to phosphatase inhibition (Chapter 4.6.1.2.1), for example, proved to have at least one homolog within mouse. In comparison, 58% of non-phosphorylated proteins showed paralogy.

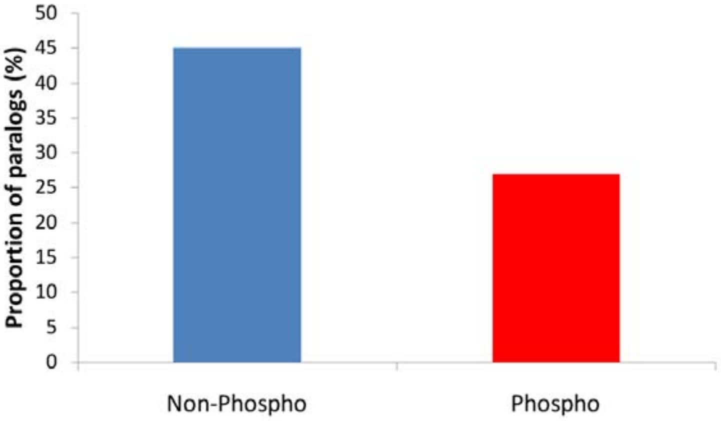


Figure 9.5: Proportion of paralogous phosphoproteins (red) and non-phosphorylated proteins (blue) in fly

Next we investigated the conservation of regions containing phosphorylation sites on the basis of global protein-protein BLASTP alignments for orthologous human phosphoproteins. We found that the average identity in the 40 amino acid window surrounding the aligned phosphorylation sites is lower for each eukaryotic species compared to the entire protein identity. This effect is most pronounced for serine and threonine (Figure 9.6a, b). This observation is in concordance with the finding that phosphorylation sites occur predominantly in loop and hinge regions on the surface of the phosphoprotein (Chapter 4.6.4), as the protein sequence of highly accessible parts of the protein evolve fast (Branden, 1999). However, it is also surprising considering the assumption that phosphorylation sites along with the surrounding kinase motif should be highly conserved to fulfil their functional roles in cell signalling. Therefore, we examined the conservation of the region surrounding phosphorylation sites in more detail: we plotted the conservation of amino acids amino- and carboxyl-terminal to the phosphorylation site for the three phosphorylation sites and for all species. As a typical example, Figures 9.6c and 9.6d show the case of serine and threonine in zebrafish. The figure reveals a symmetric region immediately adjacent to the phosphosite, in which conservation is higher than in the surrounding region. The length of this region is about

-5 to +5 amino acids for both serine and threonine and agrees well with the size of published kinase motifs. Hence, in the evolutionary section of PHOSIDA, the surrounding region of -6 to +6 amino acids is shown, in order to check the conservation of matching motifs (Chapter 4.6.3).

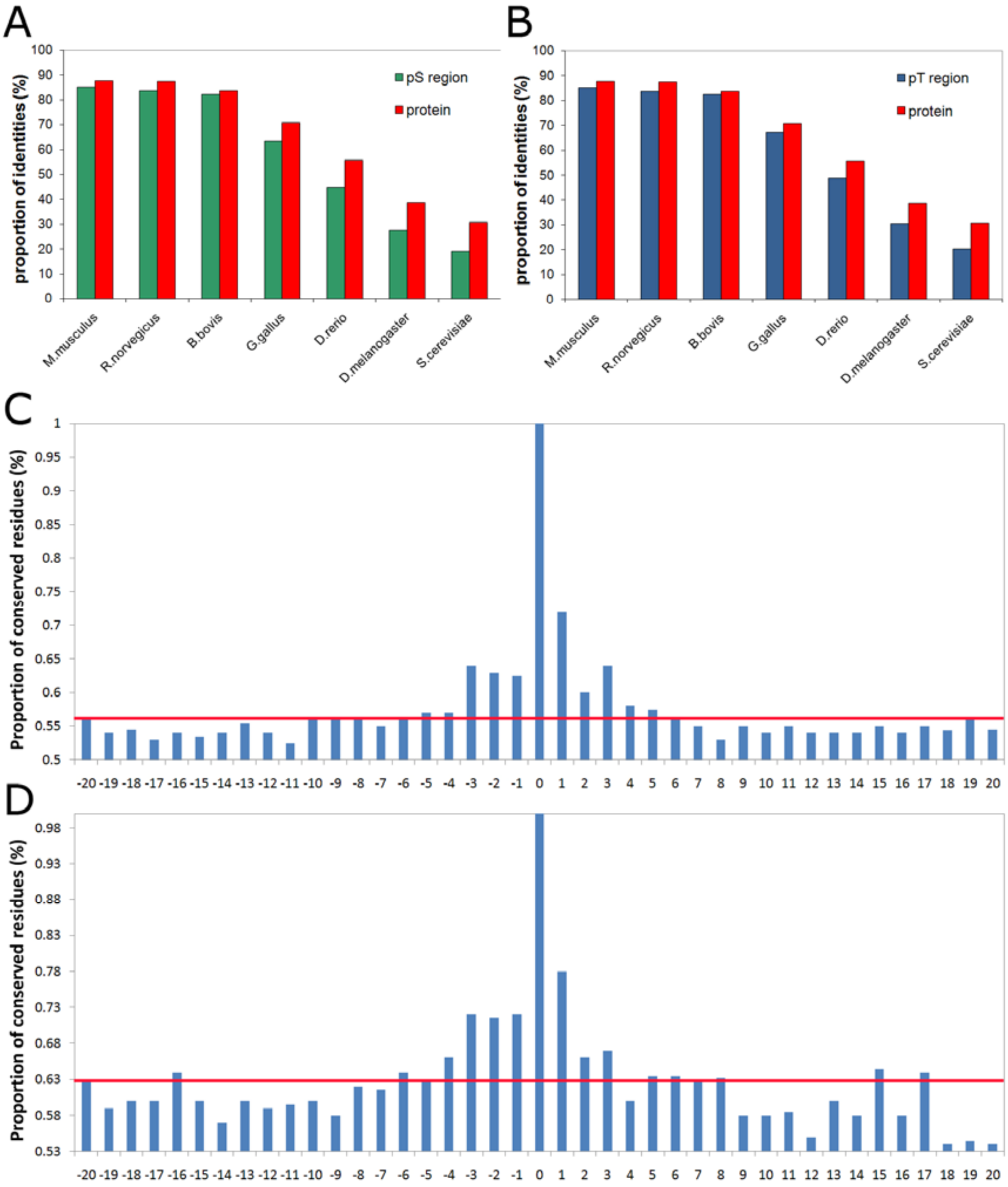


Figure 9.6: Conservation of the sequence region surrounding phosphorylation sites

The average identity of 40 amino acids surrounding phosphosites (red) proved to be less conserved than the average identity of the whole global alignment (A, B). However, the very close region (+/- 5 amino acids) surrounding phosphoserines (C) and phosphothreonines (D) show elevated sequence identity. Bars represent the proportion of identical residues in zebrafish orthologs of human phosphoproteins. The red line is the average identity in the region -20 to +20 amino acids surrounding the phosphosite.

Overall, these data suggest that the surrounding sequence regions may diverge to such an extent that the structural effect (fast sequence evolution) effectively competes with the constraining pressure of function (slow sequence evolution). In order to correctly assess the degree of conservation of phosphosites, it is therefore important to take the structural effect – fast evolution of loop regions – into account. We did this by choosing only sites located in loop regions according to SABLE predictions for the comparison set, which should isolate the functional, evolutionary constraints on the phosphosite itself. We checked the conservation of triplets encoding phosphosites in comparison to triplets that encode non-phosphorylated counterparts in phosphorylated proteins throughout 36 eukaryotes on the basis of cDNA alignments as provided by EnsEMBL. The main finding of the DNA conservation analysis of phosphoserines and phosphothreonines identified in human cells exposed to EGF stimulation (Chapter 4.6.1.1.1) was that human phosphorylation sites are more conserved throughout higher eukaryotes than their non-phosphorylated counterparts (Figure 9.7). Overall, 97% of phosphoserines were found to be conserved in chimp DNA alignments in comparison to 92% of non-phosphorylated serines, for example. In the case of rat, 70% of identified phosphothreonines were conserved in comparison to 61% of non-phosphorylated threonines. However, human phosphosites were not significantly higher conserved in lower eukaryotes such as worm and yeast. Due their low number, it was not possible to find any significant patterns regarding the conservation of highly accessible phosphotyrosines in DNA alignments of orthologous proteins.

The evolutionary study on the basis of two-directional BLASTP searches and Needlemann-Wunsch protein-protein alignments led to the same outcome: The overall conservation of human phosphorylation sites in orthologous eukaryotic proteins (Chapter 4.6.1.1.1) is shown in Figure 9.8a-d. The average amino identity for all human phosphoproteins with orthologs ranges from greater than 80% in mammals to about 25% in yeast based on Needleman-Wunsch alignments (Figure 9.8a). Figure 9.8b compares the conservation of phosphoserines that occur in loops with all non-phosphoserines that occur in loops in the same proteins. As observed in EnsEMBL alignments of coding genes, in all investigated vertebtrates, phosphoserine is significantly more conserved than serine (χ^2-test: $p = 0$). In *Drosophila* the effect is still observable, but is not significant ($p = 0.33$). In yeast this is not the case. As shown in Figure 9.7, these findings are in concordance with the results from the evolutionary study on DNA-DNA alignments. Threonine yields a similar result to serine, but this amino acid is generally less conserved. Tyrosine tends to occur in more conserved regions of the protein as mentioned above.

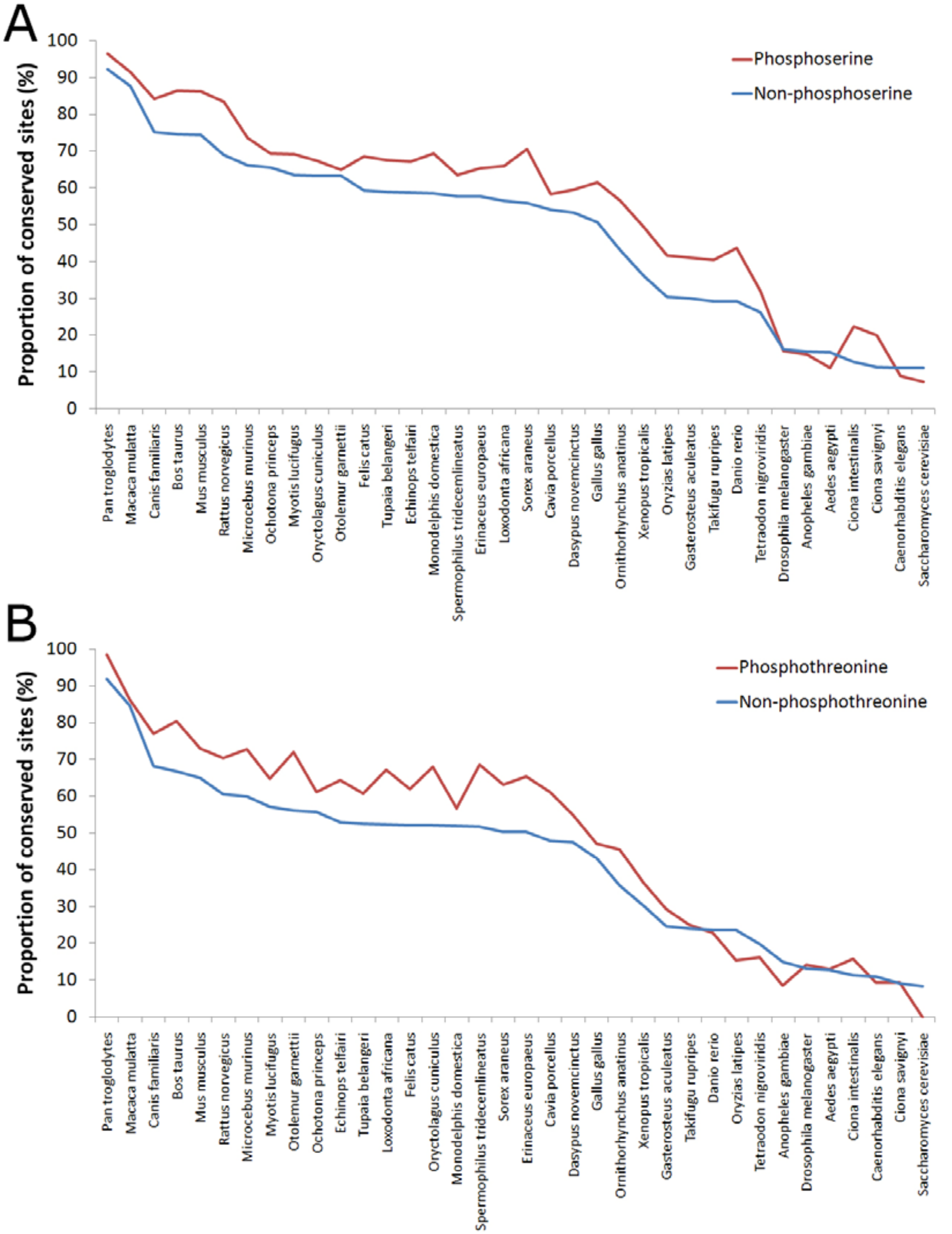


Figure 9.7 Conservation of phosphoserine (A) and phosphothreonine (B) identified in human HeLa cells exposed to EGF stimulation on the basis of cDNA alignments

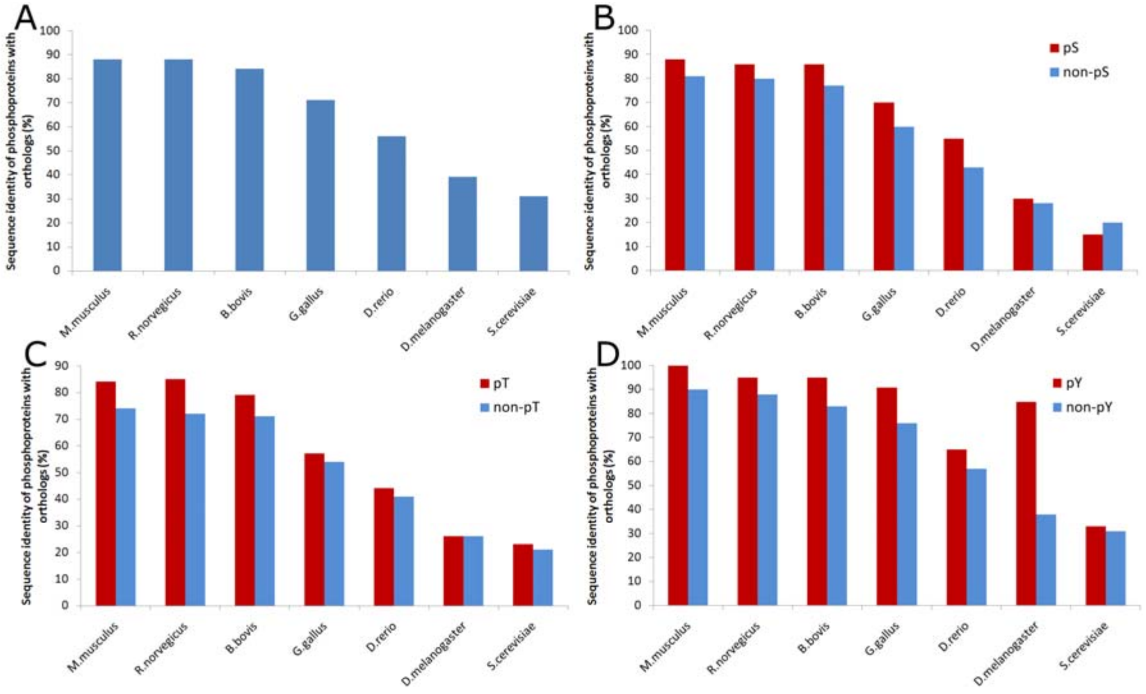


Figure 9.8: Conservation of human phosphosites on the basis of protein alignments

Analogously, phosphorylation sites that were identified in the human kinase enriched set (Chapter 4.6.1.1.2) were also found to be more conserved than their non-phosphorylated counterparts throughout higher eukaryotes on the basis of EnsEMBL cDNA alignments (data not shown). The same tendency was also observed in both mouse phosphoproteome datasets (Chapters 4.6.1.2). For example, 59% of triplets encoding serines that were phosphorylated in mouse cells exposed to phosphatase inhibition were conserved in orthologous cat genes. In comparison, 54% of triplets that encode unmodified serines in the same genes are conserved in cat. Phosphothreonines (57%) also proved to be more conserved than non-phosphorylated threonines (46%) in coding DNA alignments between phosphorylated mouse transcripts and cat orthologs.

As is apparent from Figure 9.9, the high conservation of residues phosphorylated in mouse cells was even apparent in the case of very closely related organisms: Overall, 86% phosphoserines were found to be conserved in rat, for instance, in comparison to 83% non-phosphorylated serines. Phosphothreonines also more conserved in rat (phospho: 89%, non-phospho: 79%). Again, the numbers of detected phosphorylated tyrosines that occur in loop regions on the protein surface and show an corresponding ortholog in another species were too few to derive any significant patterns relating to their overall conservation. Figure 9.9

illustrates the conservation of phosphoserines and phosphothreonines identified in mouse cells upon phosphatase inhibition.

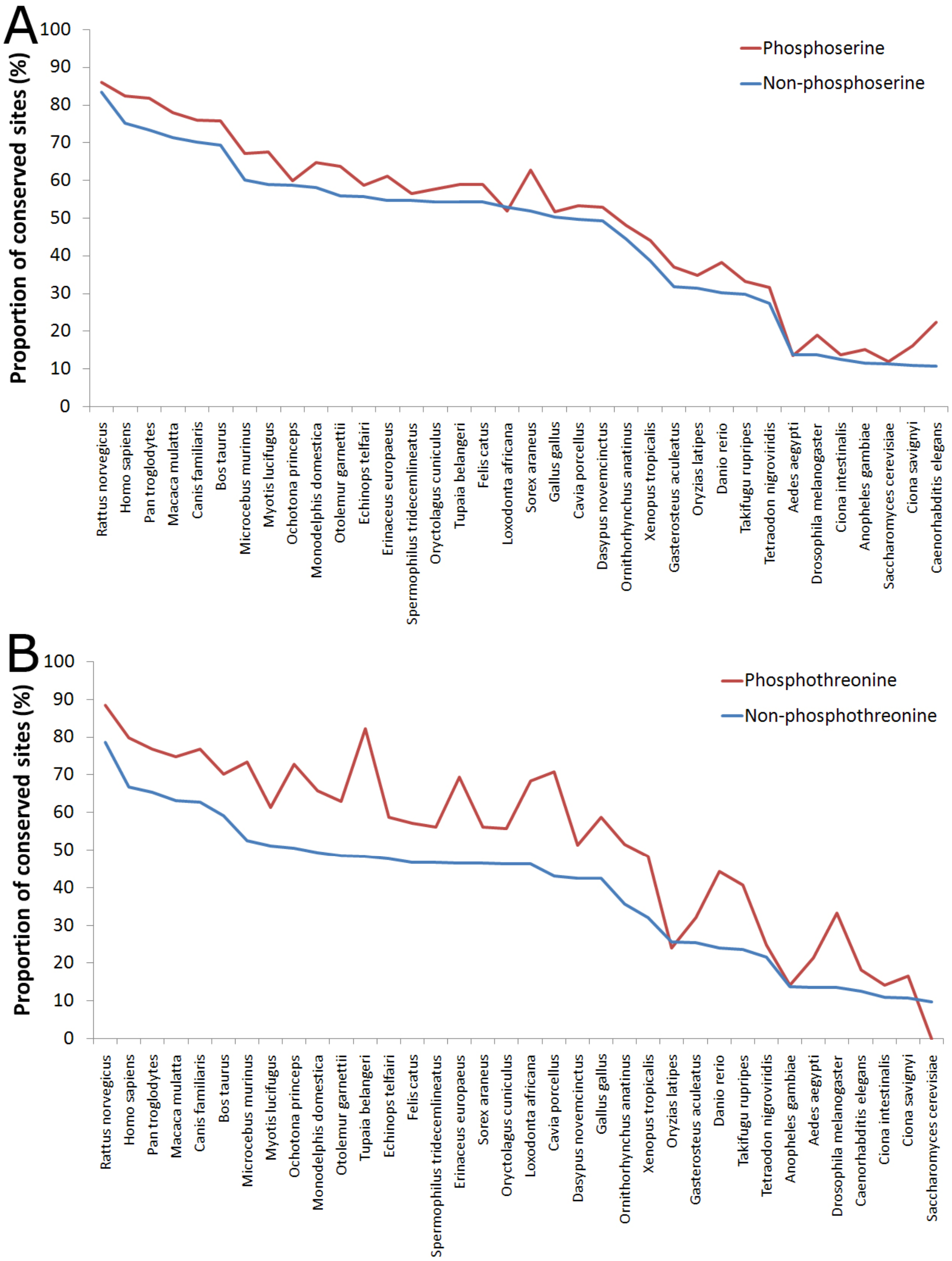


Figure 9.9 Conservation of phosphoserine (A) and phosphothreonine (B) identified in mouse cells exposed to phosphatase inhibition

The high conservation of phosphorylated residues was also evident in the fly phosphoproteome. However, phosphosites identified in yeast cells did not show a significantly higher evolutionary conservation (data not shown).

Moreover, to check the d_N/d_S ratios on the site level, we concatenated all triplets encoding phosphorylated residues, so that a single sequence containing all triplets was created. The translated version of the resulting DNA sequence should therefore yield serine-only, threonine-only or tyrosine-only containing protein sequence. All triplets that are aligned to the given triplets in orthologous genes of the specified species built up the second sequence. This yields a pairwise DNA alignment, whose synonymous and non-synonymous diverge can be calculated on the basis of the Nei and Gojobori method (Nei and Gojobori, 1986; Zhang et al., 2006), for example (Figure 9.10 upper panel). Throughout higher eukaryotes including human, mouse and fly, we found that d_N/d_S ratios derived from aligned phosphoserines and phosphothreonines were significantly lower than d_N/d_S ratios derived from aligned non-phosphorylated counterparts. Figure 9.11 shows the d_N/d_S ratios of aligned human phosphoserines versus d_N/d_S ratios of aligned non-phosphorylated serine triplets. This illustration is representative for the d_N/d_S distribution of phosphoserines and phosphothreonines versus their non-phosphorylated counterparts in each identified eukaryotic phosphoproteome. However, it was not possible to derive synonymous and non-synonymous divergence from alignments between yeast phosphoproteins and their orthologs, as yeast and higher eukaryotes are too distantly related.

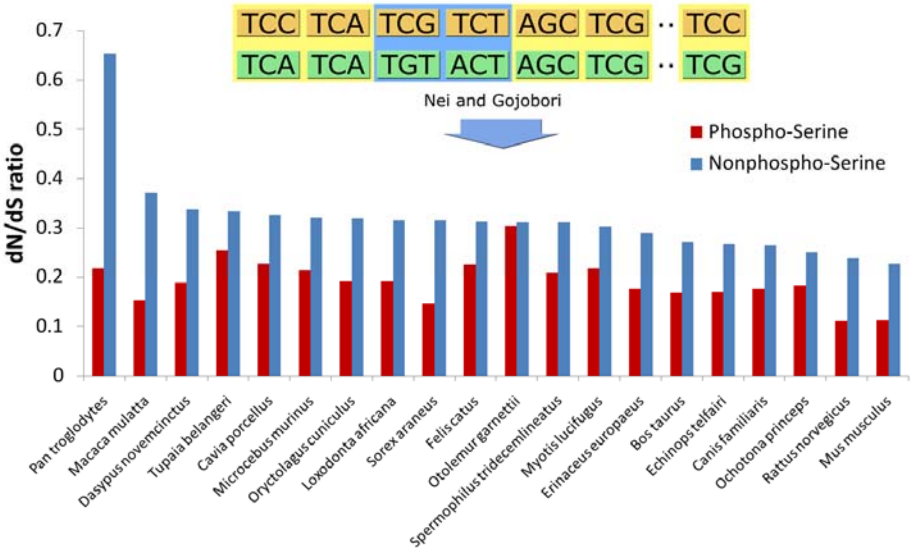


Figure 9.10 Derivation of synonymous (yellow) and non-synonymous changes (blue) between phosphosites encoding triplets (brown) and their aligned triplets (green) in orthologous genes of a given species

Triplets that are preserved in evolution without any mutations are also highlighted in yellow. Overall, aligned human phosphoserines showed lower d_N/d_S ratios than their non-phosphorylated counterparts (bottom).

Surprisingly, the evolutionary constraints on eukaryotic phosphoproteomes were also evident in the prokaryotic phosphoproteomes of *B. subtilis*, *E. coli*, *L. lactis* and *H. salinarium*. In each case, identified phosphoproteins showed more orthologs than non-phosphorylated proteins throughout all domains of life on the basis of the two-directional BLASTP method. In Eubacteria, for example, 42% of *B. subtilis* phosphoproteins were conserved in contrast to about 25% of the nonphosphorylated proteins (Figure 9.11a). In Archaea, the conservation at the phosphoproteome level was 21%, around twice as high as at the proteome level. In eukaryotes, the conservation at the phosphoproteome level was 30%, whereas at the proteome it was 14%. Even on the site level, phosphorylated residues were more conserved throughout bacteria, eukaryotes and archaea than their non-phosphorylated counterparts. In the *B. subtilis* dataset, 41% of phosphoserines were conserved throughout bacteria in comparison to 34% of non-phosphorylated serines (Figure 9.11b) and 47% of phosphothreonines were conserved in comparison to 39% of nonphosphorylated threonines. Because of their low number, we could not draw any statistically significant conclusions about phosphorylated tyrosines.

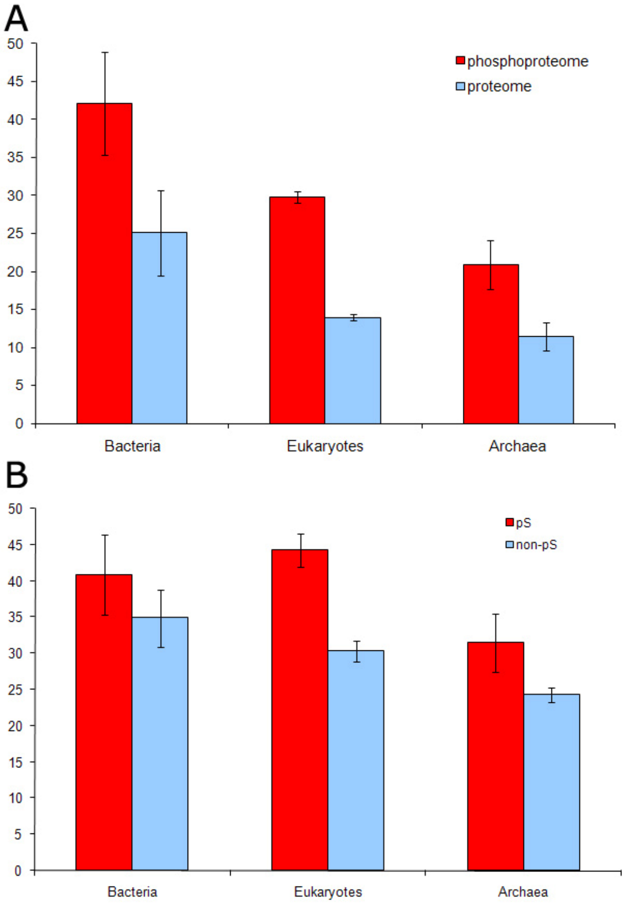


Figure 9.11 Evolutionary Conservation of the *B.subtilis* phosphoproteome

The conservation of phosphoproteins (A) and phosphoserines (B) is reported as the average conservation in all tested species from each domain of life.

9.4 Discussion

The aim of PHOSIDA is to provide a rich environment to the biologist wishing to analyze phosphorylation events of any protein of interest. Thus, PHOSIDA includes not only very high quality input as well as quantitative information, but it also integrates biological context to quantify constraints of phosphorylation on a proteome-wide scale.

Analyses using the evolutionary sections of PHOSIDA show that phosphoproteins have more orthologs than non-phosphorylated proteins on the basis of the two-directional BLASTP approach (Chapter 4.2.4). These results are in keeping with the evolutionary analysis of phosphorylated and non-phosphorylated gene transcripts on the basis of the cross-reference to the EnsEMBL Compara database, which provides detailed information about phylogenetic relationships and homologies between 37 eukaryotes. The genome annotation approach (Chapter 8) was a requisite to link our phosphodata with the comprehensive annotation of the genome including conservation as provided by EnsEMBL. The high conservation of phosphoproteins probably reflects important and conserved functional roles of proteins with this post-translational modification. However, we emphasize that our datasets might be biased towards abundant proteins, although we found evidence for good coverage of very low abundant proteins including various transcription factors in our proteomic studies. We tried to reduce this potential effect by selecting only non-phosphorylated proteins for the comparison sets that are classified as ‘known’ by the EnsEMBL database meaning that the occurrence of the given protein is experimentally validated. Nevertheless, the bias for detecting very abundant phosphoproteins cannot be excluded with absolute certainty.

Interestingly, phosphorylated proteins of the same species that were identified in different cell lines and tissues after various treatments also differed to some extent with respect to conservation: In lower eukaryotes, for example, phosphoproteins detected in human cells upon EGF stimulus were less conserved than human phosphoproteins identified in different cell cycle phases, because of the absence of the EGF receptor in eukaryotes that are very distantly related to human. In the case of the mouse phosphoproteome, the lower conservation of phosphorylated proteins determined in skin melanomas in comparison to phosphorylated proteins determined in the liver may be tissue function related (liver having a more basic and conserved role than skin).

As consequence of the location of phosphorylation sites in loops and hinges (Chapter 4.6.4), the sequence regions around phosphorylation sites evolve faster than the rest of the protein except for the amino acids making up the kinase motif: the region of about five amino acids around the phosphorylation site is more conserved than the surrounding sequence context. This finding illustrates the evolutionary constraint that the amino acid composition framing the kinase motif around the phosphosite has to be preserved. Otherwise, the substrate would lose its kinase affinity, which would negatively affect the associated signaling cascade.

Our analysis of the global DNA alignments of orthologs in 37 eukaryotes shows that phosphorylation sites are more conserved than non-phosphosites of the same proteins throughout higher eukaryotes including human, mouse and fly.

On the site level, we found that phosphorylation sites are more highly conserved throughout higher eukaryotes than their non-phosphorylated counterparts. In contrast, yeast phosphorylation sites were not highly conserved with respect to higher eukaryotes. These observations are in concordance with the findings of Manning et al. (Manning et al., 2002a): They showed that most of the known human kinases evolved after the divergence between yeast and higher eukaryotes including fly. Therefore there is a considerable number of yeast specific kinases and eukaryotic kinases that are absent in yeast. These results are in close agreement with ours, obtained on the basis of global Needle protein alignments of inter-species homologs in seven eukaryotes.

However, we only checked the residue conservation of given phosphorylation sites. There is no direct experimental evidence that a conserved serine, threonine or tyrosine is also phosphorylated in the orthologous protein in most cases. The overlaps between determined phosphoproteomes of different species are comparable with the overlaps between phosphoproteomes identified in different experiments on the same species (Pan, 2008 in press). These results indicate that the identification of the entire phosphoproteome of a given species is far from being complete. However, we assume that most of the conserved residues, which were found to be phosphorylated in at least one of the specified species, are also phosphorylated in the other organism considering that the surrounding kinase motif is also conserved.

Regardless of the fact that conserved amino acids that are phosphorylated in one species might not be phosphorylated in another other species, the occurrence of residues that are phosphorylated in the human system, for example, but not in very closely related species including chimp and mouse points to ‘background phosphorylation’. Although a substantial proportion of phosphorylation sites are more highly conserved than their non-phosphorylated counterparts, there is a considerable number of phosphosites that are not preserved in evolution. This might reflect the occurrence of sequence regions that build up a kinase motif by chance, but whose phosphorylation by the corresponding kinase by an ‘innocent bystander mechanism’ does not have any detrimental effect on the biological system. Thus, the given phosphorylation site would be lost after DNA mutation, as there is no selective pressure to keep the residue. The occurrence of sequence stretches that make up a kinase motif by chance is also very likely given the fact that there are many unspecific kinase motifs such as the CKII motif pS-X-X-E, for example.

We also showed that the inclusion of evolutionary constraints on the phosphoproteome could slightly increase the performance of the in-silico predictor (Chapter 7).

Chapter 10

Summary and Future Directions

Proteomics as a relatively new 'post-genomic' science focuses on the large scale determination of the functional protein network in the cell. We applied mass spectrometry based proteomics to determine proteomes and phosphoproteomes in different cell types including tumor cells and liver cells and of various organisms ranging from *Escherichia coli* to human. Using SILAC, we investigated protein and phosphorylation changes *in-vivo* upon different treatments including phosphatase inhibition and growth factor stimulation. The determination of thousands of proteins that contain posttranslationally phosphorylated residues demands description, storage, management and recovery of the obtained data. For this purpose, we created the phosphorylation site database PHOSIDA (http://www.phosida.com) (Chapter 4). Its purpose is not only to make the obtained large-scale data public to the scientific community, but also to mine the data and to derive general patterns relating to phosphorylation events. By quantitative proteomics, we show that regulation through posttranslational modifications takes place on the site level rather than the level of the entire phosphoprotein. For example, many proteins contained phosphorylation sites that were differently regulated upon epidermal growth factor stimulation. This demonstrates the necessity to establish methods that extend the common approach of matching spectra to peptide sequences. We created a probability based algorithm to detect phosphorylation events on the site level (Chapter 3). Using a specified cutoff with respect to the localization probability of phosphorylation sites ($p > 0.75$), we tested the accuracy of our method using manually verified high confidence data of a previous phosphoproteomic study. We found that more than 90% of determined phosphorylation sites were correctly localized within the peptide sequence.

We extended the common proteomics workflow ranging from cell preparation to matching the measured spectra to protein sequences by the application of the 'Knowledge Discovery in Databases' (KDD) process to extract knowledge from the obtained large-scale data. For example, we found that only a small subset of phosphorylation sites was regulated upon growth factor stimulation. All quantitative phosphoproteomic studies showed that regulation through phosphorylation was most apparent for tyrosine residues. Our data sets suggest that the distribution of pS, pT, and pY is around 85%, 13%, and 2% on average. We also observed that the number of phosphorylation events in prokaryotic cells is considerably different from the one observed in eukaryotes. In fly, for example, we determined more than 10,000 *in-vivo* phosphorylation sites on even very low abundant proteins including kinases and transcription factors. In comparison, we did not detected more than 100 phosphorylation events in any prokaryotic cell.

The comparison of our phosphoproteomic datasets with large-scale data from other studies, which were also integrated into PHOSIDA, underlined the novelty of our high accuracy data. Overall, around 80% of determined phosphorylation sites of each study were novel. Thus the determination of phosphoproteomes is far from being complete.

Using statistical tests that are integrated into the PHOSIDA environment, we found that phosphorylation events are distributed over all cellular compartments. Some compartments such as mitochondria, however, were underrepresented, whereas phosphorylation events in the nucleus were overrepresented. On the basis of integrated secondary structure and solvent accessibility predictions, we found that phosphorylation sites were predominantly located in loops and hinges on the surface of the protein. We also found evidence for significantly overrepresented consensus sequences that surround eukaryotic phosphorylation sites and make up kinase motifs. In contrast, we could not derive any significant motif from prokaryotic phosphosites. Besides mining methods that derive general patterns regarding function, cell compartment localization, structural constraints, consensus sequences and further categories, we investigated the evolution of phosphorylation (Chapter 9). The high conservation of phosphorylation throughout higher eukaryotes on the protein level as well as on the site level underlines the functional impact of phosphorylated proteins, which play key roles in signalling and therefore have to be preserved in evolution. In this regard, the yeast phosphoproteome presents an outlier, as yeast phosphorylation sites were not significantly more conserved than their non-phosphorylated counterparts. This observation is in agreement with the fact that many kinases evolved after the speciation event that separated yeast from higher eukaryotes. In addition, a non-negligable proportion of amino acids that are phosphorylated in human, but not conserved in mouse, point to background phosphorylation that does not have any functional impact on the underlying system and therefore no selective pressure.

Furthermore, the PHOSIDA knowledge discovery pipeline also includes a phosphorylation site predictor on the basis of a support vector machine (Chapter 7). The accuracy of predicting phosphorylated serines on the basis of the raw sequence was higher than 90% for each investigated eukaryotic organism.

The inclusion of various high confidence large scale data obtained from high accuracy quantitative phosphoproteomic studies along with a phosphorylation site predictor make PHOSIDA a rich environment to the biologist wishing to analyze phosphorylation events of proteins of interest. Moreover, the automated analysis pipeline based on the KDD process enables us to derive various patterns relating to phosphorylation.

We also constructed a proteome database, termed ‘Max-Planck Unified Proteome Database’ (MAPU) that includes proteomes of different organelles, tissues and cell types (Chapter 5). Obtained proteomic data were also mapped to the genome (Chapter 8). The reassignment of identified peptide sequences to corresponding genes allows not only the assignment of important protein features including phosphorylation to the coding genome sequences but also the experimental validation of predicted genes. Using the DAS technology, we linked our proteome database with the genome database EnsEMBL. Finally, the update and extension of the sex bias database SEBIDA was a further intent of my PhD study (Chapter 6).

We intend to extend the phosphorylation site database by the inclusion of other posttranslational modifications such as acetylation, for example. It will be interesting, whether the general constraints observed in phosphorylation events can also be found in other posttranslational modifications using the KDD process. In addition, we wish to establish the first machine learning approach that is capable of predicting acetylation events, on the basis of the raw sequence.

Another goal is to integrate the evolutionary annotation provided by the EnsEMBL Compara database into the PHOSIDA web application. This will enable the web users to study the evolutionary conservation of any given phosphorylated protein throughout 36 eukaryotes. Currently, the evolutionary section of the PHOSIDA online application is restricted to phylogenetic information throughout seven eukaryotes on the basis of our self-coded pipeline. Furthermore, we intend to link our proteome databases with other online environments such as PRIDE and Peptide Atlas, in order to establish a broad proteomic data network.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F.*, et al.* (2000). The genome sequence of Drosophila melanogaster. Science (New York, NY *287*, 2185-2195.

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198-207.

Aivaliotis, M., Macek, B., Gnad, F., Reichelt, P., Mann, M., and Oesterhelt, D. (under review). Ser/Thr/Tyr protein phosphorylation in archaea, the third domain of life. PLOS One.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Amanchy, R., Kalume, D.E., Iwahori, A., Zhong, J., and Pandey, A. (2005). Phosphoproteome analysis of HeLa cells using stable isotope labeling with amino acids in cell culture (SILAC). J Proteome Res *4*, 1661-1671.

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

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.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Bairoch, A., and Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res *24*, 21-25.

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

Begg, C.E., Connolly, T. (2004). Database Systems: A Practical Approach to Design, Implementation and Management (Addison-Wesley Longman).

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2008). GenBank. Nucleic Acids Res *36*, D25-30.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T.*, et al.* (2004). An overview of Ensembl. Genome research *14*, 925-928.

Bodenmiller, B., Malmstrom, J., Gerrits, B., Campbell, D., Lam, H., Schmidt, A., Rinner, O., Mueller, L.N., Shannon, P.T., Pedrioli, P.G.*, et al.* (2007). PhosphoPep--a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. Mol Syst Biol *3*, 139.

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., Ehrlich, S.D., and Sorokin, A. (2001). The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1403. Genome Res *11*, 731-753.

Bradley, A. (2002). Mining the mouse genome. Nature *420*, 512-514.

Branden, C., Tooze, J. (1999). Introduction to Protein Structure (Taylor & Francis).

Brent, M.R. (2007). How does eukaryotic gene prediction work? Nat Biotechnol *25*, 883-885.

Brookes, M. (2002). Fly: The Unsung Hero of Twentieth Century Science (Phoenix).

Cai, S.L., Tee, A.R., Short, J.D., Bergeron, J.M., Kim, J., Shen, J., Guo, R., Johnson, C.L., Kiguchi, K., and Walker, C.L. (2006). Activity of TSC2 is inhibited by AKT-mediated phosphorylation and membrane partitioning. J Cell Biol *173*, 279-289.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res *32*, D262-266.

Cazzulino, D., Aprea, V.G., Greenwood, J. (2004). Beginning Visual Web Programming in C#: From Novice to Professional.

Chen, W.G., and White, F.M. (2004). Proteomic analysis of cellular signaling. Expert Rev Proteomics *1*, 343-354.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M.*, et al.* (1998). SGD: Saccharomyces Genome Database. Nucleic Acids Res *26*, 73-79.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science (New York, NY *311*, 1283-1287.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21*, 3674-3676.

Cortay, J.C., Rieul, C., Duclos, B., and Cozzone, A.J. (1986). Characterization of the phosphoproteins of Escherichia coli cells by electrophoretic analysis. Eur J Biochem *159*, 227-237.

Cox, J., and Mann, M. (2007). Is proteomics the new genomics? Cell *130*, 395-398.

Curwen, V., Eyras, E., Andrews, T.D., Clarke, L., Mongin, E., Searle, S.M., and Clamp, M. (2004). The Ensembl automatic gene annotation system. Genome Res *14*, 942-950.

Dan, H.C., Sun, M., Yang, L., Feldman, R.I., Sui, X.M., Ou, C.C., Nellist, M., Yeung, R.S., Halley, D.J., Nicosia, S.V.*, et al.* (2002). Phosphatidylinositol 3-kinase/Akt pathway regulates tuberous sclerosis tumor suppressor complex by phosphorylation of tuberin. J Biol Chem *277*, 35364-35370.

Date, C.J. (2003). An Introduction to Database Systems (Addison Wesley).

Daub, H., Olsen, J.V., Bairlein, M., Gnad, F., Oppermann, F.S., Korner, R., Greff, Z., Keri, G., Stemmann, O., and Mann, M. (2008). Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. Molecular cell *31*, 438-448.

De Godoy, L., Gnad, F., Olsen, J.V., Ren, S., and Mann, M. (under review). Global analysis of in vivo protein phosphorylation in yeast. Molecular cell.

Desiere, F., Deutsch, E.W., Nesvizhskii, A.I., Mallick, P., King, N.L., Eng, J.K., Aderem, A., Boyle, R., Brunner, E., Donohoe, S.*, et al.* (2005). Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol *6*, R9.

Deutscher, J., and Saier, M.H., Jr. (2005). Ser/Thr/Tyr protein phosphorylation in bacteria - for long time neglected, now well established. J Mol Microbiol Biotechnol *9*, 125-131.

Dhillon, A.S., Hagan, S., Rath, O., and Kolch, W. (2007). MAP kinase signalling pathways in cancer. Oncogene *26*, 3279-3290.

Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N., and Gibson, T.J. (2004). Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics *5*, 79.

Dorus, S., Busby, S.A., Gerike, U., Shabanowitz, J., Hunt, D.F., and Karr, T.L. (2006). Genomic and functional evolution of the Drosophila melanogaster sperm proteome. Nat Genet *38*, 1440-1445.

Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., and Obradovic, Z. (2002). Intrinsic disorder and protein function. Biochemistry *41*, 6573-6582.

Dzeroski, S., and Lavrac, N. (2007). Relational Data Mining (Springer).

Elias, J.E., Haas, W., Faherty, B.K., and Gygi, S.P. (2005). Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Methods *2*, 667-675.

Ellegren, H., and Parsch, J. (2007). The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet *8*, 689-698.

Ester, M., Sander J. (2000). Knowledge Discovery in Databases: Techniken und Anwendungen (Springer).

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. Science (New York, NY *246*, 64-71.

Fermin, D., Allen, B.B., Blackwell, T.W., Menon, R., Adamski, M., Xu, Y., Ulintz, P., Omenn, G.S., and States, D.J. (2006). Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol *7*, R35.

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

Finn, R.D., Stalker, J.W., Jackson, D.K., Kulesha, E., Clements, J., and Pettett, R. (2007). ProServer: a simple, extensible Perl DAS server. Bioinformatics *23*, 1568-1570.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., *et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science (New York, NY *269*, 496-512.

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., *et al.* (2008). Ensembl 2008. Nucleic Acids Res *36*, D707-714.

Futschik, M.E., and Carlisle, B. (2005). Noise-robust soft clustering of gene expression time-course data. J Bioinform Comput Biol *3*, 965-988.

Futuyma, D.J. (1998). Evolutionary Biology (Sinauer Associates).

Gennick, J. (2006). SQL (O'Reilly).

Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin, S., and Wayne, M. (2004). Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. Genetics *167*, 1791-1799.

Gnad, F., Oroshi, M., Birney, E., and Mann, M. (in press). MAPU 2.0: High accuracy proteomes mapped to genomes. Nucleic Acids Research.

Gnad, F., and Parsch, J. (2006). Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. Bioinformatics *22*, 2577-2579.

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

Goldman, T.D., and Arbeitman, M.N. (2007). Genomic and functional studies of Drosophila sex hierarchy regulated gene expression in adult head and nervous system tissues. PLoS Genet *3*, e216.

Gomperts, B.D., Kramer, I.M., Tatham P.E.R. (2004). Signal Transduction (Academic Press).

Goodman, D. (2006). Dynamic HTML: The Definitive Reference (O'Reilly).

Gorg, A., Weiss, W., and Dunn, M.J. (2004). Current two-dimensional electrophoresis technology for proteomics. Proteomics *4*, 3665-3685.

Grangeasse, C., Cozzone, A.J., Deutscher, J., and Mijakovic, I. (2007). Tyrosine phosphorylation: an emerging regulatory device of bacterial physiology. Trends Biochem Sci *32*, 86-94.

Griffiths, I., Flanders, J., Sells, C. (2003). Mastering Visual Studio .NET (O'Reilly).

Gross, J.H. (2004). Mass spectrometry (Springer).

Gruhler, A., Olsen, J.V., Mohammed, S., Mortensen, P., Faergeman, N.J., Mann, M., and Jensen, O.N. (2005). Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. Mol Cell Proteomics *4*, 310-327.

Grumbling, G., and Strelets, V. (2006). FlyBase: anatomical data, images and queries. Nucleic Acids Res *34*, D484-488.

Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc Natl Acad Sci U S A *97*, 9390-9395.
Hager, J.W., and Le Blanc, J.C. (2003). High-performance liquid chromatography-tandem mass spectrometry with a new quadrupole/linear ion trap instrument. J Chromatogr A *1020*, 3-9.
Hahn, M.W., and Lanzaro, G.C. (2005). Female-biased gene expression in the malaria mosquito Anopheles gambiae. Curr Biol *15*, R192-193.
Hamilton, B., MacDonald, M. (2003). ADO.NET in a Nutshell (O'Reilly).
Han, J., Kamber, M. (2000). Data Mining. Concepts and Techniques. In (Morgan Kaufmann).
Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57-70.
Hancock, J.T. (2005). Cell Signalling (Oxford University Press).
Heuer, A., Saake, G. (2000). Datenbanken: Konzepte und Sprachen (MITP).
Hey, J., and Kliman, R.M. (2002). Interactions between natural selection, recombination and gene density in the genes of Drosophila. Genetics *160*, 595-608.
Hoek, K.S. (2007). DNA microarray analyses of melanoma gene expression: a decade in the mines. Pigment Cell Res *20*, 466-484.
Hornbeck, P.V., Chabra, I., Kornhauser, J.M., Skrzypek, E., and Zhang, B. (2004). PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics *4*, 1551-1561.
Hunter, T. (1987). A thousand and one protein kinases. Cell *50*, 823-829.
Hunter, T. (2000). Signaling--2000 and beyond. Cell *100*, 113-127.
Hunter, T., and Plowman, G.D. (1997). The protein kinases of budding yeast: six score and more. Trends Biochem Sci *22*, 18-22.
Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res *32*, 1037-1049.
Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol *151*, 389-409.
Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L.*, et al.* (2005). Reactome: a knowledgebase of biological pathways. Nucleic Acids Res *33*, D428-432.
Karas, M., and Hillenkamp, F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Analytical chemistry *60*, 2299-2301.
Kennelly, P.J. (2002). Protein kinases and protein phosphatases in prokaryotes: a genomic perspective. FEMS Microbiol Lett *206*, 1-8.
Kersey, P.J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004). The International Protein Index: an integrated database for proteomics experiments. Proteomics *4*, 1985-1988.
Kienitz, H. (1968). Einführung in Massenspektrometrie (Verlag Chemie).
Kirkness, E.F., and Kerlavage, A.R. (1997). The TIGR human cDNA database. Methods Mol Biol *69*, 261-268.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.
Larsen, M.R., Thingholm, T.E., Jensen, O.N., Roepstorff, P., and Jorgensen, T.J. (2005). Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. Mol Cell Proteomics *4*, 873-886.

Lasithiotakis, K.G., Sinnberg, T.W., Schittek, B., Flaherty, K.T., Kulms, D., Maczey, E., Garbe, C., and Meier, F.E. (2008). Combined inhibition of MAPK and mTOR signaling inhibits growth, induces cell death, and abrogates invasive growth of melanoma cells. J Invest Dermatol *128*, 2013-2023.

Liberty, J. (2005a). Programming C# (O'Reilly).

Liberty, J., Huwritz, D. (2005b). Programming ASP.NET: Building Web Applications and Services (O'Reilly).

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

Macek, B., Mijakovic, I., Olsen, J.V., Gnad, F., Kumar, C., Jensen, P.R., and Mann, M. (2007). The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis. Mol Cell Proteomics *6*, 697-707.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics *21*, 3448-3449.

Mann, M., and Kelleher, N.L. (2008). Special Feature: Precision proteomics: The case for high resolution and high mass accuracy. Proc Natl Acad Sci U S A.

Manning, G., Plowman, G.D., Hunter, T., and Sudarsanam, S. (2002a). Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci *27*, 514-520.

Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002b). The protein kinase complement of the human genome. Science (New York, NY *298*, 1912-1934.

Marshall, A.G., Hendrickson, C.L., and Jackson, G.S. (1998). Fourier transform ion cyclotron resonance mass spectrometry: a primer. Mass Spectrom Rev *17*, 1-35.

Martin, S.E., Shabanowitz, J., Hunt, D.F., and Marto, J.A. (2000). Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. Anal Chem *72*, 4266-4274.

McIntyre, L.M., Bono, L.M., Genissel, A., Westerman, R., Junk, D., Telonis-Scott, M., Harshman, L., Wayne, M.L., Kopp, A., and Nuzhdin, S.V. (2006). Sex-specific expression of alternative transcripts in Drosophila. Genome Biol *7*, R79.

Meier, F., Schittek, B., Busch, S., Garbe, C., Smalley, K., Satyamoorthy, K., Li, G., and Herlyn, M. (2005). The RAS/RAF/MEK/ERK and PI3K/AKT signaling pathways present molecular targets for the effective treatment of advanced melanoma. Front Biosci *10*, 2986-3001.

Meyer, E.A. (2006). CSS: The Definitive Guide (O'Reilly).

Mijakovic, I., Petranovic, D., Bottini, N., Deutscher, J., and Ruhdal Jensen, P. (2005). Protein-tyrosine phosphorylation in Bacillus subtilis. J Mol Microbiol Biotechnol *9*, 189-197.

Mikhaylova, L.M., Nguyen, K., and Nurminsky, D.I. (2008). Analysis of the Drosophila melanogaster testes transcriptome reveals coordinate regulation of paralogous genes. Genetics *179*, 305-315.

Mitchell, T.M. (1997). Machine Learning (McGrawHill).

Monedero, V., Kuipers, O.P., Jamet, E., and Deutscher, J. (2001). Regulatory functions of serine-46-phosphorylated HPr in Lactococcus lactis. J Bacteriol *183*, 3391-3398.

Mumby, M., and Brekken, D. (2005). Phosphoproteomics: new insights into cellular signaling. Genome Biol *6*, 230.

Musciano, C., Kennedy, B. (2006). HTML & XHTML: The Definitive Guide (O'Reilly).

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443-453.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol *3*, 418-426.

Nelson, D.L., and Cox, M.M. (2008). Principles of Biochemistry (Lehninger).

Nielsen, H., Brunak, S., and von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng *12*, 3-9.
Noble, W.S. (2006). What is a support vector machine? Nat Biotechnol *24*, 1565-1567.
O'Brien, K.P., Remm, M., and Sonnhammer, E.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res *33*, D476-480.
Obenauer, J.C., Cantley, L.C., and Yaffe, M.B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res *31*, 3635-3641.
Olason, P.I. (2005). Integrating protein annotation resources through the Distributed Annotation System. Nucleic Acids Res *33*, W468-470.
Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell *127*, 635-648.
Olsen, J.V., and Mann, M. (2004). Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. Proc Natl Acad Sci U S A *101*, 13417-13422.
Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics *1*, 376-386.
Ong, S.E., and Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol *1*, 252-262.
Oudot, C., Cortay, J.C., Blanchet, C., Laporte, D.C., Di Pietro, A., Cozzone, A.J., and Jault, J.M. (2001). The "catalytic" triad of isocitrate dehydrogenase kinase/phosphatase from E. coli and its relationship with that found in eukaryotic protein kinases. Biochemistry *40*, 3047-3055.
Pan, C., Gnad, F., Olsen, J.V., and Mann, M. (2008). Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors. Proteomics.
Parisi, M., Nuttall, R., Edwards, P., Minor, J., Naiman, D., Lu, J., Doctolero, M., Vainer, M., Chan, C., Malley, J.*, et al.* (2004). A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogaster adults. Genome Biol *5*, R40.
Parisi, M., Nuttall, R., Naiman, D., Bouffard, G., Malley, J., Andrews, J., Eastman, S., and Oliver, B. (2003). Paucity of genes on the Drosophila X chromosome showing male-biased expression. Science (New York, NY *299*, 697-700.
Pawson, T., and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. Science (New York, NY *300*, 445-452.
Pawson, T., and Scott, J.D. (2005). Protein phosphorylation in signaling--50 years and counting. Trends Biochem Sci *30*, 286-290.
Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551-3567.
Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R.*, et al.* (2005). Global analysis of protein phosphorylation in yeast. Nature *438*, 679-684.
Rabilloud, T. (2002). Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. Proteomics *2*, 3-10.
Ramakrishnan, R., Gehrke, J. (2003). Database management systems. In (McGraw-Hill).
Ranz, J.M., Castillo-Davis, C.I., Meiklejohn, C.D., and Hartl, D.L. (2003). Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science (New York, NY *300*, 1742-1745.
Reese, G., Yarger, J.Y., King T. (2002). Managing and Using MySQL (O'Reilly).

Remenyi, A., Good, M.C., and Lim, W.A. (2006). Docking interactions in protein kinase and phosphatase networks. Curr Opin Struct Biol *16*, 676-685.
Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*, 276-277.
Ridley, M. (2003). Evolution (Blackwell Publ).
Rush, J., Moritz, A., Lee, K.A., Guo, A., Goss, V.L., Spek, E.J., Zhang, H., Zha, X.M., Polakiewicz, R.D., and Comb, M.J. (2005). Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. Nat Biotechnol *23*, 94-101.
Sadygov, R.G., Cociorva, D., and Yates, J.R., 3rd (2004). Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat Methods *1*, 195-202.
Salomon, A.R., Ficarro, S.B., Brill, L.M., Brinker, A., Phung, Q.T., Ericson, C., Sauer, K., Brock, A., Horn, D.M., Schultz, P.G.*, et al.* (2003). Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry. Proc Natl Acad Sci U S A *100*, 443-448.
Schlessinger, J. (2000). Cell signaling by receptor tyrosine kinases. Cell *103*, 211-225.
Schwartz, D., and Gygi, S.P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat Biotechnol *23*, 1391-1398.
Schwartz, J.C., Senko, M.W., and Syka, J.E. (2002). A two-dimensional quadrupole ion trap mass spectrometer. J Am Soc Mass Spectrom *13*, 659-669.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.
Shi, Z., Resing, K.A., and Ahn, N.G. (2006). Networks for the allosteric control of protein kinases. Curr Opin Struct Biol *16*, 686-692.
Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195-197.
Soufi, B., Gnad, F., Jensen, P.R., Petranovic, D., Mann, M., Mijakovic, I., and Macek, B. (2008). The Ser/Thr/Tyr phosphoproteome of Lactococcus lactis IL1403 reveals multiply phosphorylated proteins. Proteomics.
Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M.F., Rifkin, S.A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P.E.*, et al.* (2004). A gene expression map for the euchromatic genome of Drosophila melanogaster. Science (New York, NY *306*, 655-660.
Thelemann, A., Petti, F., Griffin, G., Iwata, K., Hunt, T., Settinari, T., Fenyo, D., Gibson, N., and Haley, J.D. (2005). Phosphotyrosine signaling networks in epidermal growth factor receptor overexpressing squamous carcinoma cells. Mol Cell Proteomics *4*, 356-376.
Thingholm, T.E., Jensen, O.N., Robinson, P.J., and Larsen, M.R. (2008). SIMAC (sequential elution from IMAC), a phosphoproteomics strategy for the rapid separation of monophosphorylated from multiply phosphorylated peptides. Mol Cell Proteomics *7*, 661-671.
Valaskovic, G.A., Kelleher, N.L., and McLafferty, F.W. (1996). Attomole protein characterization by capillary electrophoresis-mass spectrometry. Science (New York, NY *273*, 1199-1202.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). The sequence of the human genome. Science (New York, NY *291*, 1304-1351.
Wagner, M., Adamczak, R., Porollo, A., and Meller, J. (2005). Linear regression models for solvent accessibility prediction in proteins. J Comput Biol *12*, 355-369.
Wayne, M.L., Telonis-Scott, M., Bono, L.M., Harshman, L., Kopp, A., Nuzhdin, S.V., and McIntyre, L.M. (2007). Simpler mode of inheritance of transcriptional variation in male Drosophila melanogaster. Proc Natl Acad Sci U S A *104*, 18577-18582.

Williams, N. (1996). Yeast genome sequence ferments new research. Science (New York, NY *272*, 481.

Witten, I.H., Frank, E. (1999). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (Morgan Kaufmann).

Witten, I.H., Frank, E. (2005). Data Mining. Practical Machine Learning Tools and Techniques (Morgan Kaufmann).

Wright, F. (1990). The 'effective number of codons' used in a gene. Gene *87*, 23-29.

Wurgler-Murphy, S.M., King, D.M., and Kennelly, P.J. (2004). The Phosphorylation Site Database: A guide to the serine-, threonine-, and/or tyrosine-phosphorylated proteins in prokaryotic organisms. Proteomics *4*, 1562-1570.

Zanivan, S., Gnad, F., Wickstroem, S., Geiger, T., Macek, B., Cox, J., Faessler, R., and Mann, M. (in press). Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry. Journal of Proteome Research.

Zanivan, S., Gnad, F., Wickstroem, S., Geiger, T., Macek, B., Cox, J., Faessler, R., and Mann, M. (under review). Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry. Journal of Proteome Research.

Zhai, B., Villen, J., Beausoleil, S.A., Mintseris, J., and Gygi, S.P. (2008). Phosphoproteome analysis of Drosophila melanogaster embryos. J Proteome Res *7*, 1675-1682.

Zhang, Y., Zhang, Y., Adachi, J., Olsen, J.V., Shi, R., de Souza, G., Pasini, E., Foster, L.J., Macek, B., Zougman, A., *et al.* (2007). MAPU: Max-Planck Unified database of organellar, cellular, tissue and body fluid proteomes. Nucleic Acids Res *35*, D771-779.

Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Wong, G.K., and Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics *4*, 259-263.

Zheng, J., He, C., Singh, V.K., Martin, N.L., and Jia, Z. (2007). Crystal structure of a novel prokaryotic Ser/Thr kinase and its implication in the Cpx stress response pathway. Mol Microbiol *63*, 1360-1371.

Acknowledgement

I am deeply grateful to my doctoral advisor, Matthias Mann, for introducing me to this line of research and for providing boundless opportunities. Although I just graduated from university, even in the early stages of my PhD study he already had high confidence in me and involved me in many fantastic projects. On the one hand, I learned a lot from him ranging from writing a manuscript to the set up of a scientific study. He encouraged me to give a talk at the 'Interaction Proteome' Conference in Mallorca and to gain experience at the European Bioinformatics Institute in Cambridge UK. On the other hand, he allowed me to integrate my own ideas in all studies. I hope that I could justify the confidence he has placed in me. He inspired me to continue scientific work.

Furthermore, I wish to thank the bioinformatics group including Juergen Cox, Guoqing Li, and Chanchal Kumar for fruitful discussions. I am very thankful to Shubin Ren for his support in setting up the phosphorylation site predictor. I am particularly thankful to Mario Oroshi for his technical support. He also became one of my best friends. Moreover, I thank Jesper Olsen, Boris Macek, Maximiliane Hilger, Sara Zanivan, Cuiping Pan, Karlheinz Mann, Lyris De Godoy, Blagoy Blagoev, Henrik Daub and Felix Oppermann for their collaborative work on phosphoproteomic studies and their confidence in me. I am also thankful to Tine Klitmoeller and Theresa Schneider for organisational help. I am grateful to all members of our department for helpful discussions and a nice time. In addition, I wish to thank John Parsch for fruitful discussions about evolution and for the support in updating the sex bias database SEBIDA.

Research in the PANDA group of Ewan Birney at the European Bioinformatics Institute (EBI) has been supported by a Marie Curie fellowship from the European Commission. I am grateful to Ewan Birney for inviting me to work in his group and for his thoughtful advice. I thank Andrew Jenkinson and Eugene Kulesha for support in the setup of the EnsEMBL DAS sources. I am also grateful to Markus Fritz, Albert Vilella Bertan and Michael Schuster for insightful comments and suggestions.

I am very thankful to my family and friends for their support and encouragement. Finally, I wish to thank the doctoral candidate admission board for their time and review.

CURRICULUM VITAE

PERSONAL DETAILS

Name	Florian Gnad
Address	Wilramstrasse 53 81669 Munich
Telephone	0049 89 8578 2296 (Office)
Date of Birth	23rd May 1981
Marital Status	unmarried

EDUCATION

EMBL – European Bioinformatics Institute, Cambridge UK
Marie Curie Fellowship (1/2008 – 6/2008)
Project: Genome Annotation using Mass Spectrometry derived Data
Advisor: Ewan Birney

Max-Planck-Institute of Biochemistry, Martinsried (near Munich)
Ph.D. study in Bioinformatics (since 11/2005)
Thesis: Bioinformatics of Phosphoproteomics
Advisor: Prof. Matthias Mann

Ludwig-Maximilians-Universität / Technische Universität, Munich
Diploma in Bioinformatics (10/2001 – 10/2005) (passed with distinction)
Thesis: Microarray Data Analysis of Sex Biased Genes in *Drosophila melanogaster*
Advisor: Prof. John Parsch

EXPERIENCE

Internship, 02/2005 – 04/2005
Ludwig-Maximilians-Universität Munich (Biocenter, Dep. of Evolutionary Genomics)

Internship, 10/2003 – 03/2004
Technische Universität Munich (Dep. of Bioinformatics)

Internship, 10/2002 – 03/2003
Ludwig-Maximilians-Universität Munich (Botanical Institute)

Internship, 03/2002 – 04/2002
Ludwig-Maximilians-Universität Munich (Dep. of Bioinformatics)

Student assistant, 02/2002 – 08/2002
Micromet (Biotechnology Company)

SKILLS AND QUALIFICATIONS

Programming ability in C#, Java, ASP.NET, PHP, SQL
Fluent in German and English
Qualification in Latin (passed with distinction)

OUTSIDE INTERESTS AND ACTIVITIES

Soccer, Strength Training, Reading, Travelling, Computer games

PUBLICATIONS

F. Gnad*, M. Oroshi, E. Birney, M. Mann (accepted for publication), *MAPU 2.0: a database of proteomes mapped to the genomes*, Nucleic Acids Research.

S. Zanivan, F. Gnad, S. Wickström, T. Geiger, B. Macek, J. Cox, R. Fässler, M. Mann (accepted for publication), *Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry*, Journal of Proteome Research.

C. Pan, F. Gnad, J.V. Olsen, M. Mann (2008)
Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors, Proteomics [Epub ahead of print].

H. Daub, J.V. Olsen, M. Bairlein, F. Gnad*, F.S. Oppermann, R. Körner, Z. Greff, G. Keri, O. Stemmann, M. Mann (2008), *Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle*, Mol Cell 31(3):438-48.

B. Soufi, F. Gnad*, P.R. Jensen, D. Petranovic, M. Mann, I. Mijakovic, B. Macek (2008), *The Ser/Thr/Tyr phosphoproteome of Lactococcus lactis IL1403 reveals multiply phosphorylated proteins*, Proteomics 8(17):3486-93.

K. Mann, J.V. Olsen, B. Macek, F. Gnad, M. Mann (2008), *Identification of new chicken egg proteins by mass spectrometry-based proteomic analysis*, World's Poultry Science Journal 64: 209-218.

Human Proteinpedia enables sharing of human protein data (2008), Nature Biotechnology 26(2): 164-7.

F. Gnad* and M. Mann (2008). *PHOSIDA – Ressource für Phosphorylierungsstellen in verschiedenen Spezies*, Laborwelt 1/2008: 23-26 (in German).

F. Gnad*, S. Ren, J. Cox, J.V. Olsen, B. Macek, M. Oroshi, M. Mann (2007). *PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites*, Genome Biology 8(11): R250.

B. Macek, F. Gnad*, B. Soufi, C. Kumar, J.V. Olsen, I. Mijakovic, M. Mann (2007). *Phosphoproteome analysis of E.coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation*, Mol Cell Proteomics 7(2), 299-307.

B. Macek, I. Mijakovic, J.V. Olsen, F. Gnad, C. Kumar, P.R. Jensen, M. Mann (2007). *The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis*, Mol Cell Proteomics 6(4), 697-707.

K. Mann, J.V. Olsen, B. Macek, F. Gnad, M. Mann (2007). *Phosphoproteins of the chicken eggshell calcified layer*, Proteomics 7(1), 106-15.

J.V. Olsen, B. Blagoev, F. Gnad*, B. Macek, C. Kumar, P. Mortensen, M. Mann (2006). *Global, in-vivo and site-specific phosphorylation dynamics in signalling networks*, Cell 127, 635-648.

F. Gnad*, J. Parsch (2006). *Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression*, Bioinformatics 22: 2577-2579.

* first author or shared first author