Novel Bioinformatical and Statistical Methods for the Analysis of Mass Spectrometry-Based Phosphoproteomic Data

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

Dr. rer. nat.

der Fakultät für

Biologie

an der

Universität Duisburg-Essen

vorgelegt von

Martin Klammer

aus Kremsbrücke, Österreich

Februar 2015

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden in den Abteilungen für Bioinformatik des Zentrums für Medizinische Biotechnologie (ZMB) der Universität Duisburg-Essen und der Evotec (München) GmbH in Martinsried durchgeführt.

1. Gutachter: Prof. Dr. Daniel Hoffmann
2. Gutachter: Prof. Dr. Jürgen Cox

Vorsitzender des Prüfungsausschusses: Prof. Dr. Markus Kaiser

Tag der mündlichen Prüfung: 17. Juni 2015

Contents

List of Abbreviations

ATP	Adenosine triphosphate
AUROC	Area under the receiver operating characteristic
BH	Benjamini-Hochberg correction (confer **FDR**)
CV	Cross validation
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen
EGF	Epidermal growth factor
EGFR	Epidermal growth factor receptor
eIF	Eukaryotic initiation factor
ELISA	Enzyme-linked immunosorbent assay
FBS	Fetal bovine serum
FDA	Food and drug administration
FDG-PET	Fluorodeoxyglucose-positron emission tomography
FDR	False discovery rate
FN	False negative
FP	False positive
FWER	Family-wise error rate
GA	Genetic algorithm

List of Abbreviations

GDP	Guanosine diphosphate
GI_{50}	Half-maximum growth inhibitory concentration
GO	Gene ontology
GR	Global rank test
GTP	Guanosine triphosphate
HCD	Higher-energy collisional dissociation
HPRD	Human protein reference database
IMAC	Immobilized metal affinity chromatography
ITGB4	Integrin beta 4
iTRAQ	Isobaric tags for relative and absolute quantitation
KEGG	Kyoto encyclopedia of genes and genomes
k-NN	k-nearest neighbor
LC	Liquid chromatogrphy
LC-MS/MS	Liquid chromatography tandem mass spectrometry
LIMMA	Linear models of microarrays
LOOCV	Leave-one-out cross validation
MAPK	Mitogen-activated protein kinase
MPI	Max Planck Institute
MR	Mean rank test
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry

List of Abbreviations

mTOR	Mammalian target of rapamycin
mTRAQ	Mass differential tags for relative and absolute quantification
m/z	Mass over charge ratio
NP	Non-deterministic polynomial-time
NSCLC	Non-small cell lung cancer
NSGA-II	Non-dominated sorting genetic algorithm-II
PEP	Posterior error probability
PI3K	Phosphatidylinositol-4,5-bisphosphate 3-kinase
PSA	Prostate-specific antigen
ROC	Receiver operating characteristic
RP	Rank products test
SAM	Significance analysis of microarrays
SCX	Strong cation exchange
SILAC	Stable isotope labeling by amino acids in cell culture
STRING	Search tool for the retrieval of interacting genes/proteins
SVM	Support vector machine
TMT	Tandem mass tag
TPR	True positive rate
ZMB	Zentrums für Medizinische Biotechnologie

List of Figures

List of Figures

List of Tables

Summary

In living cells, reversible protein phosphorylation events propagate signals caused by external stimuli from the plasma membrane to their intracellular destinations. Aberrations in these signaling cascades can lead to diseases such as cancer. To identify and quantify phosphorylation events on a large scale, mass spectrometry (MS) has become the predominant technology. The large amount of data generated by MS requires efficient, tailor-made computational tools in order to draw meaningful biological conclusions.

In this work, four new methods for analyzing MS-based phosphoproteomic data are presented. The first method, called SubExtractor, combines phosphoproteomic data with protein network information to identify differentially regulated subnetworks. The method is based on a Bayesian probabilistic model that accounts for information about both differential regulation and network topology, combined with a genetic algorithm and rigorous significance testing.

The second method, called MeanRank test, is a global one-sample location test, which is based on the mean ranks across replicates, and internally estimates and controls the false discovery rate. The test successfully deals with small numbers of replicates, missing values without the need of imputation, non-normally distributed expression levels, and non-identical distribution of up- and down-regulated features, while its statistical power scales well with the number of replicates.

The third method is a biomarker discovery workflow that aims at identifying a multivariate response prediction biomarker for treatment of non-small cell lung cancer cell lines with the kinase inhibitor dasatinib from phosphoproteomic data (referred to as NSCLC biomarker). An elaborate biomarker workflow based on robust feature selection in combination with a support vector machine (SVM) was designed in order to find a phosphorylation signature that accurately predicts the response to dasatanib.

The fourth method, called Pareto biomarker, extends the previous NSCLC biomarker workflow by optimizing not only one single objective (i.e. best possible separation of responders and non-responders), but also the objectives signature size and relevance (i.e. association of signature proteins with dasatinib's main target). This is achieved by employing a multiobjective optimization algorithm based on the principle of Pareto optimality, which allows for a simultaneous optimization of all three objectives.

These novel data analysis methods were thoroughly validated using experimental data and compared to existing methods. They can be used on their own, or they can be combined into a joint workflow in order to efficiently answer complex biological questions in the field of large-scale *omics* in general and phosphoproteomics in particular.

Zusammenfassung

In lebenden Zellen sind reversible Proteinphosphorylierungen für die Weiterleitung von Signalen externer Stimuli zu deren intrazellulären Bestimmungsorten verantwortlich. Anomalien in solchen Signaltransduktionswegen können zu Krankheiten wie beispielsweise Krebs führen. Um Phosphorylierungsstellen in großem Maßstab zu identifizieren und zu quantifizieren, hat sich die Massenspektrometrie (MS) zur vorherrschenden Technologie entwickelt. Die große Menge an Daten, die von Massenspektrometern generiert wird, erfordert effiziente maßgeschneiderte Computerprogramme, um aussagekräftige biologische Schlüsse ziehen zu können.

In dieser Arbeit werden vier neue Methoden zur Analyse von MS-basierten phosphoproteomischen Daten präsentiert. Die erste Methode, genannt SubExtractor, kombiniert phosphoproteomische Daten mit Proteinnetzwerkinformationen um differentiell regulierte Subnetzwerke zu identifizieren. Die Methode basiert auf einem Bayesschen Wahrscheinlichkeitsmodell, das sowohl Information über die differentielle Regulation der Einzelknoten als auch die Netzwerktopologie berücksichtigt. Das Modell ist kombiniert mit einem genetischen Algorithmus und stringenter Signifikanzanalyse.

Die zweite Methode, genannt MeanRank-Test, ist ein globaler Einstichproben-Lagetest, der auf den mittleren Rängen der Replikate beruht, und die *False Discovery Rate* implizit abschätzt und kontrolliert. Der Test eignet sich für die Anwendung auf Daten mit wenigen Replikate, fehlenden und nicht normalverteilten Werten, sowie nicht gleichverteilter Hoch- und Runterregulation. Gleichzeitig skaliert die Teststärke gut mit der Anzahl an Replikaten.

Die dritte Methode ist ein Arbeitsablauf zur Biomarkeridentifizierung und hat zum Ziel, einen multivariaten Stratifikationsbiomarker aus phosphoproteomischen Daten zu extrahieren, der das Ansprechen von nichtkleinzelligen Bronchialkarzinomzelllinien auf den

Kinaseinhibitor Dasatinib vorhersagt (bezeichnet als NSCLC-Biomarker). Dazu wurde ein ausführlicher Biomarkerarbeitsablauf basierend auf einer robusten *Feature Selection* in Kombination mit *Support Vector Machine*-Klassifizierung erstellt, um eine Phosphorylierungssignatur zu finden, die das Ansprechen auf Dasatinib richtig vorhersagt.

Die vierte Methode, genannt Pareto-Biomarker, erweitert den vorherigen Biomarkerarbeitsablauf, indem nicht nur eine Zielfunktion (d.h. die bestmögliche Trennung von Respondern und Nichtrespondern) optimiert wird, sondern zusätzlich noch die Signaturgröße und Relevanz (d.h. die Verbindung der Signaturproteine mit dem Targetprotein von Dasatinib). Dies wird durch die Verwendung eines multiobjektiven Optimierungsalgorithmus erreicht, der auf dem Prinzip der Pareto-Optimalität beruht und die gleichzeitige Optimierung aller drei Zielfunktionen ermöglicht.

Die hier präsentierten neuen Datenanalysemethoden wurden gründlich mittels experimenteller Daten validiert und mit bereits bestehenden Methoden verglichen. Sie können einzeln verwendet werden, oder man kann sie zu einem gemeinsamen Arbeitsablauf zusammenfügen, um komplexe biologische Fragestellungen in *Omik*-Gebieten im Allgemeinen und Phosphoproteomik im Speziellen zu beantworten.

Chapter 1

Introduction

1.1 Protein phosphorylation

Protein phosphorylation is one of the most important post-translational modifications in a living cell. For uncovering its outstanding biological importance, Edmond H. Fisher and Edwin G. Krebs were awarded the Nobel Prize in Physiology or Medicine "for their discoveries concerning reversible protein phosphorylation as a biological regulatory mechanism" in 1992 [1]. In chemical terms, phosphorylation is the addition of a phosphate group (PO_4^{3-}) to a molecule. Added predominantly to specific serine, threonine or tyrosine residues, a phosphorylation can change the protein's conformation and/or subcellular localization and thus alter its function and activity – either positively (activation) or negatively (inhibition). The class of enzymes catalyzing phosphorylations is called kinases, or more specifically protein kinases, when phosphorylating proteins. Protein phosphorylations are reversible; the process of dephosphorylation is catalyzed by phosphatases. Phosphorylations play a pivotal role in signal transduction, where an external stimulus causes the activation of certain signaling pathways. Such pathways usually start with the stimulation of receptors on the cell surface and communicate the signal along a cascade of kinases to its destination (nucleus, ribosome, proteasome, etc.). Aberrations in these signaling pathways are responsible for many types of diseases, most prominently, cancer.

One of the best studied signaling pathways in cancer is the MAPK (mitogen-activated

[1] http://www.nobelprize.org/nobel_prizes/medicine/laureates/1992

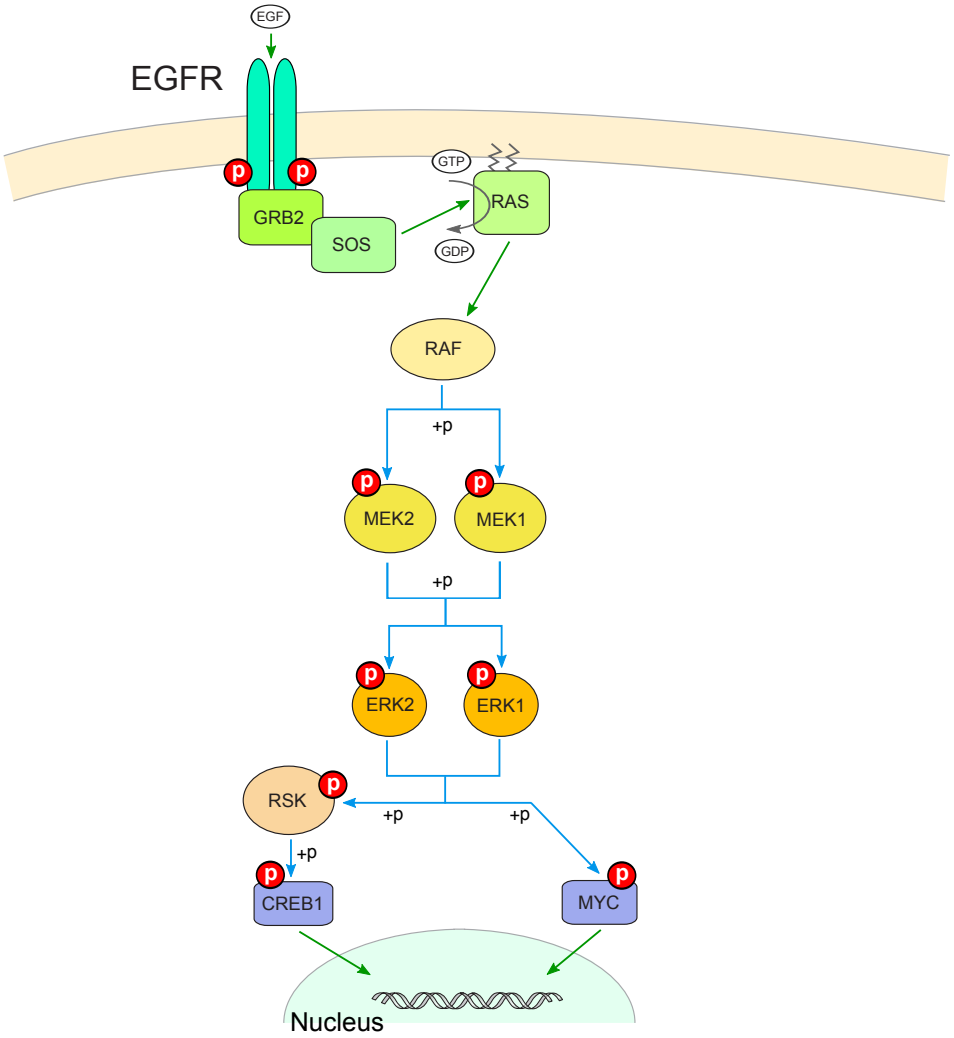

Figure 1.1: MAPK pathway example displaying the RAS-RAF-MEK-ERK cascade. Blue arrows indicate protein phosphorylation events (also displayed by red circles), green arrows other physical protein interactions.

protein kinase) pathway. In healthy cells, the MAPK pathway is responsible for cell differentiation, proliferation and apoptosis; however, mutations in this pathway can easily lead to uncontrolled pathological cell proliferation, that is, cancer. The MAPK pathway consists of different sub-cascades [1], one of them, the RAS-RAF-MEK-ERK cascade, which is found mutated in more than 30% of all tumors [2], is displayed as an example in Figure 1.1. Here, the epidermal growth factor receptor (EGFR; a receptor tyrosine kinase) homodimer is stimulated by an extracellular ligand (the mitogen EGF), and subsequently phosphorylates itself on its cytosolic domain [3] (a process called autophosphorylation).

The docking protein GRB2 is then able to bind to this phosphorylated region, and in turn recruits the guanine nuclear exchange factor SOS to the plasma membrane [4]. Subsequently, SOS activates the membrane bound GTPase RAS by promoting the exchange of the bound guanosinediphosphate (GDP) for guanosine triphosphate (GTP) [5]. Next, RAS activates the protein kinase RAF (a MAP3K – MAP kinase kinase kinase) [6], which phosphorylates and therefore activates MEK1/2 (MAP2K1/2 – MAP kinase kinase 1/2), which in turn phosphorylate ERK1/2 (MAPK3/1 – MAP kinase 3/1) in an activating manner [1]. Both ERK1 and ERK2 then translocate to the nucleus where they directly phosphorylate and activate various transcription factors (e.g. the proto-oncogene protein MYC [7]), or phosphorylate other kinases, such as ribosomal s6 kinases (RSKs), which in turn phosphorylate transcription factors (e.g. the cyclic AMP-responsive element-binding protein 1(CREB1) [8]). The pathway depicted in Figure 1.1 is an exemplified version of the processes taking place in a living cell. In reality, there are various feedback mechanisms and crosstalks between different pathways (e.g. between the MAPK and the mammalian target of rapamycin (mTOR) pathway [9] – another very important pathway in cancer signaling).

In many types of cancers, genetic mutations render protein kinases constitutively (permanently) active. Drugs that specifically target such dysregulated kinases have become increasingly important in cancer therapy. Most of these drugs are either antibodies (e.g. trastuzumab (Herceptin®), the first FDA approved kinase inhibitor targeting HER2/neu [10], a protein of the EGFR family) or small molecules (e.g. dasatinib (Sprycel®), a multi-kinase inhibitor targeting BCR-ABL, the Src-kinase family, c-Kit, ephrin receptors, and PDGFRb [11, 12]. Such targeted drugs are less toxic than traditional chemotherapeutic therapies, but they are not effective in all patients. Thus, to effectively apply these targeted therapies, discrimination between the different patient populations is indispensable in order to determine the optimal treatment for each patient. These therapeutic approaches are referred to as personalized medicine.

1.2 Mass spectrometry-based proteomics

To identify and quantify proteins and their modifications (e.g. phosphorylation) on a large scale, mass spectrometry (MS) has become the predominant technology. By applying latest sample preparation workflows, mass spectrometric equipment and bioinformatics tools, today up to 10,000 proteins [13] or more than 20,000 phosphorylation sites [14] can be routinely identified, and their relative abundance can be compared across different experimental conditions (e.g. drug treated versus untreated cell line samples).

The most widely used proteomics workflow is referred to as bottom-up or shot-gun proteomics. This approach involves enzymatic digestion of complex protein mixtures into peptides, followed by fractionation and MS analysis [15]. In the case of phosphoproteomics, the workflow contains an additional phosphopeptide enrichment step, since phosphorylated peptides are underrepresented in the total cell lysate and therefore enrichment is required to efficiently analyze phosphoproteomes with high coverage by MS-based workflows (see Figure 1.2 for a typical phosphoproteomics workflow).

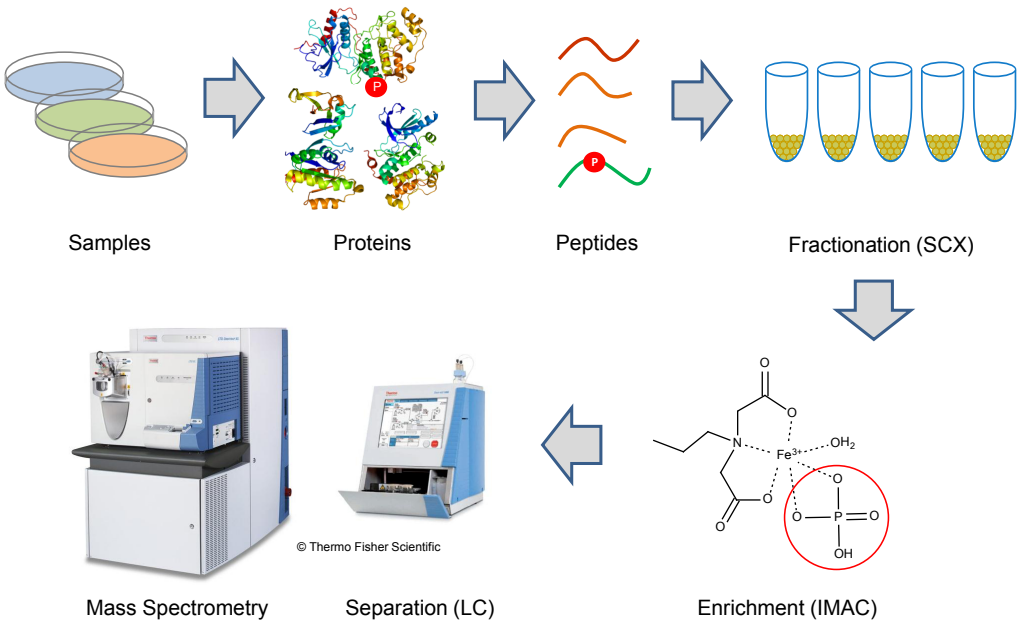

Figure 1.2: Phosphoproteomics workflow displaying the steps from sample to mass spectrometry analysis. Cells are lysed and their proteins are digested. The resulting peptide mixture is fractionated and subjected to phosphopeptide enrichment, followed by LC-MS/MS analysis.

Samples to be analyzed by an MS-based proteomics experiment can be any kind of protein source (e.g. cell lines, primary cells, tissues or body fluids). After cell lysis, the pool of proteins is digested with a protease, usually trypsin. Trypsin has the advantage to generate peptides with suitable lengths and molecular masses for detection in a mass spectrometer. In addition, trypsin generates peptides with C-terminal arginine (Arg) and lysine (Lys) residues (except for the C-terminal peptide, if the protein sequence does not end with Arg/Lys), which can be easily protonated under acidic conditions and thereby support mass spectrometric detection in the positive ion mode.

Next, the peptide pool is fractionated, which represents a first step of complexity reduction. For phosphoproteomics, strong cation exchange (SCX) chromatography has been the method of choice for many years [15]. SCX separates peptides according to their charge states and implicitly performs a phosphopeptide pre-enrichment, as peptides carrying a phosphate have a lower net charge than unmodified peptides. Each of the generated fractions is then subjected to further phosphopeptide enrichment by employing immobilized metal affinity chromatography (IMAC). IMAC is based on the high-affinity binding of phosphates to certain trivalent metal ions (e.g. Fe^{3+}; confer [15] and Figure 1.2) immobilized on solid support beads.

The highly enriched phosphopeptide sample is then subjected to liquid chromatography (LC), which is usually directly coupled (on-line) to the mass spectrometer (LC-MS). In the LC procedure, the sample mixture is forced by a liquid (mobile phase) through a packed column (stationary phase) containing a hydrophobic surface (reverse-phased), leading to adsorption of the peptides to the stationary phase under aqueous conditions. By increasing the concentration of the hydrophobic buffer, a gradient is generated that leads to a fractionated elution of the bound peptides, while hydrophilic peptides elute earlier than hydrophobic ones. The eluting peptides are electrosprayed into the mass spectrometer [16], where they are further analyzed.

Since the nominal masses of peptides are not sufficient for reliable sequence assignment when comparing them to peptide databases, a two-step MS approach is necessary in order to reliably identify peptides and the corresponding protein. In the first step, the mass over charge ratios (m/z) and corresponding intensities of the eluting peptide analytes are measured in a so-called survey scan (MS or MS^1 spectrum). From this scan, the most-

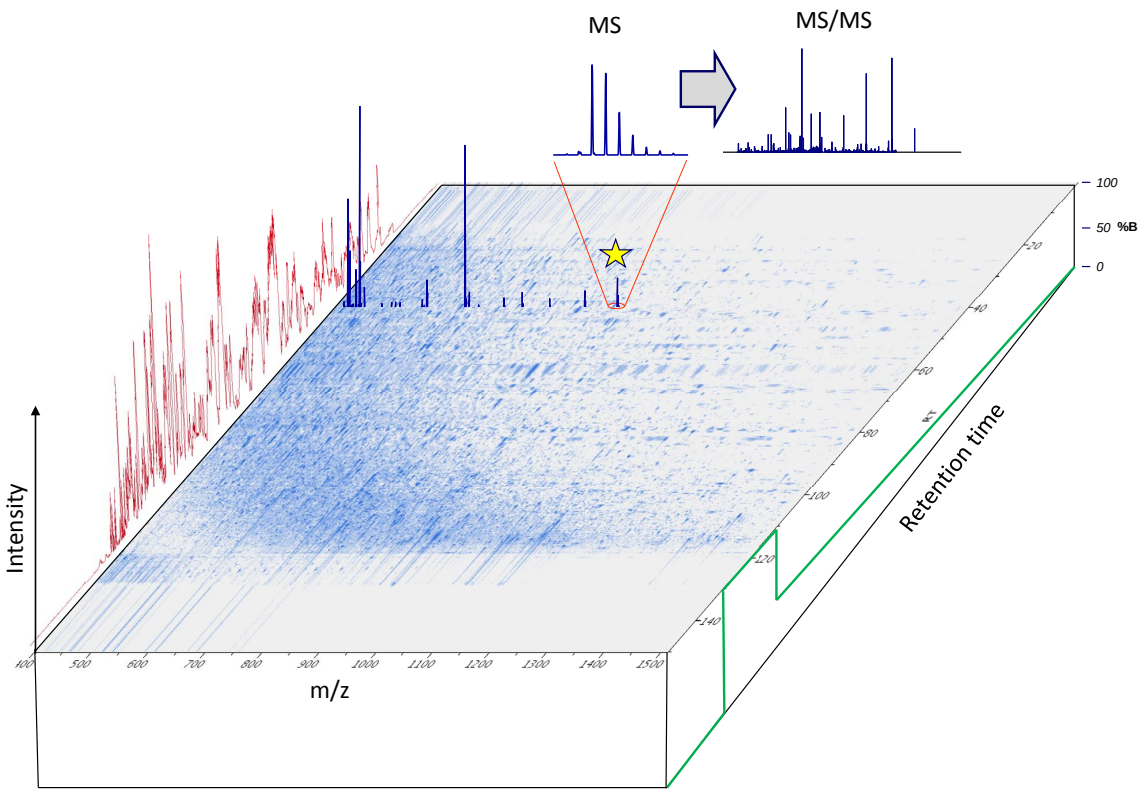

Figure 1.3: 3D spectra example illustrating the vast amount of data generated in an LC-MS/MS experiment. The green line on the right represents the gradient of hydrophobic elution buffer (referred to as buffer B) resulting in the chromatogram painted in red on the left. Approximately once every second a survey scan (MS spectrum) followed by 10 fragment scans (MS/MS spectra) are performed, leading to the complex pattern of measured intensities for tens of thousands of m/z over time. With kind permission of Andreas Tebbe, creator of this picture.

intense peaks (typically 10) are selected, isolated and fragmented as described in detail below and the resulting fragment spectra are recorded (MS/MS or MS^2 spectrum). This cycle is repeated over the entire length of the LC gradient to analyze and identify as many peptides as possible. The vast amount of data generated by this procedure is illustrated as 3D spectral image in Figure 1.3.

To draw meaningful biological conclusions from MS experiments, it is usually not sufficient to only analyze and identify the protein repertoire of the given samples (qualitative

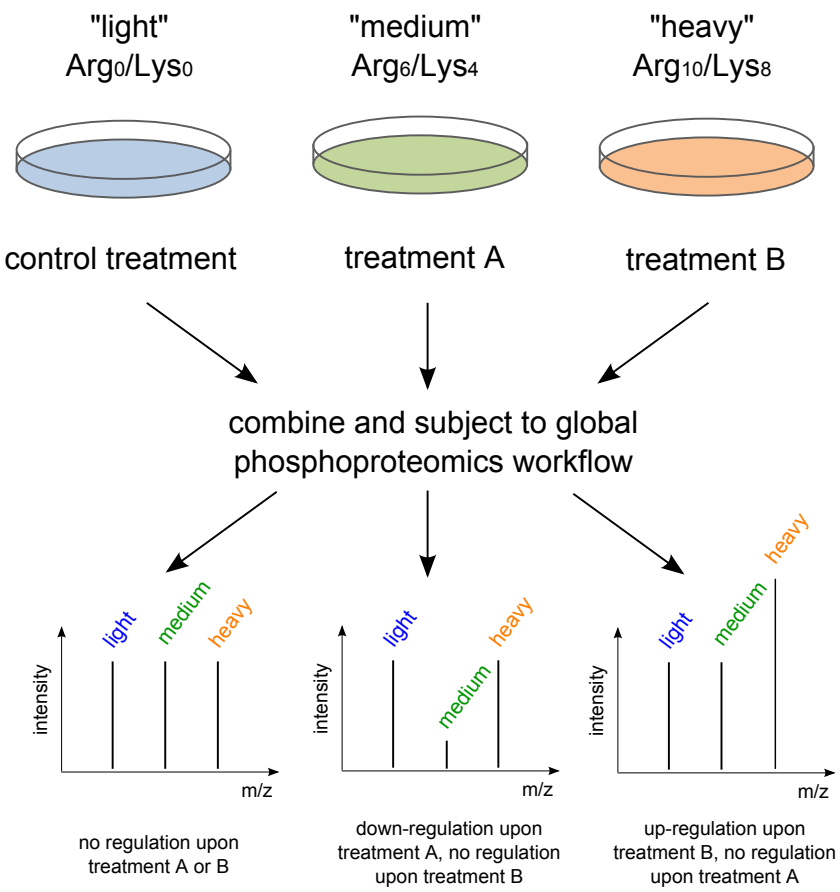

Figure 1.4: SILAC labeling diagram illustrating a triple-SILAC experiment with three different treatments. On the bottom, intensities for one peptide are displayed exemplarily.

analysis), but also to quantitatively compare different proteomes under different conditions (quantitative analysis). However, signal intensities generated by a mass spectrometer are per se not quantitative, that is, they depend on the environment of co-eluting peptides and are thus not directly comparable across different MS runs. In order to accurately quantify differences in protein expression or phosphorylation levels across different experimental conditions, labeling strategies were developed and are now routinely applied. One of the most widely used labeling approaches is Stable Isotope Labeling by Amino acids in Cell culture (SILAC) [17], where growth media of dividing cell lines are supplemented with either light, medium or heavy isotopes of the amino acids arginine and lysine. After the labeling process, the three different cell populations can be treated differently and are subsequently pooled and subjected to the proteomics workflow. Here, another advantage

of using trypsin as proteolytic enzyme takes effect, i.e. cleaving after Arg/Lys ensures the presence of at least one SILAC labeled amino acid needed for quantification (again, with the exception of C-terminal peptides). As the differently labeled peptides from each SILAC growth condition are physico-chemically identical, they co-elute simultaneously into the MS. Nevertheless, they are distinguishable due to their defined mass differences introduced by the metabolically incorporated isotopologues, allowing for a direct and accurate relative quantification between the three treatment conditions. There are various alternatives to SILAC labeling, which are applied if metabolical labeling is not applicable (e.g. mass differential tags for relative and absolute quantification (mTRAQ) [18] for tissue samples), or a higher level of multiplicity is desired (e.g. tandem mass tags (TMT) [19] for time series experiments with up to 10 states). Recent advances in computer algorithms also allow for label-free quantification (e.g. [20]); however quantifications based on labeled peptides are usually more precise [21] and allow for a higher level of multiplexing capacity in such experiments.

Depending on the application, different types of mass spectrometers are preferred. For global proteomics experiments aiming at detecting as many proteins or phosphosites as possible, Orbitrap instruments are well suited. They provide high resolution and high accuracy [23], which are both vital for reliable peptide identification and quantification. The design of a current Orbitrap instrument, the Q Exactive (Thermo Fisher Scientific), is depicted in Figure 1.5. The Q Exactive is a hybrid mass spectrometer that combines a quadrupole used for ion selection with an Orbitrap used for ion detection [22]. The electrospray of ionized peptides from the nanospray source is guided through a set of ion optics (s-lens, flatapole) to the machine's quadrupole. A quadrupole consists of four hyperbolical rods that are set parallel to each other, enabling the filtering of ionized peptides of a certain mass over charge ratio (m/z) based on the stability of their trajectories in the oscillating electric fields that are applied to the rods [24]. In the survey scan (MS^1 spectrum), no filtering is applied, thus the spectrum contains information about the entire effluent of a given time. To perform a fragment spectrum (MS^2 spectrum) of a certain spectral peak, the quadrupole is configured such that only a small mass range around the desired peak's m/z has stable trajectories and can pass through. Once the desired ions have passed the quadrupole, they are guided to the C-trap, where they are collected and

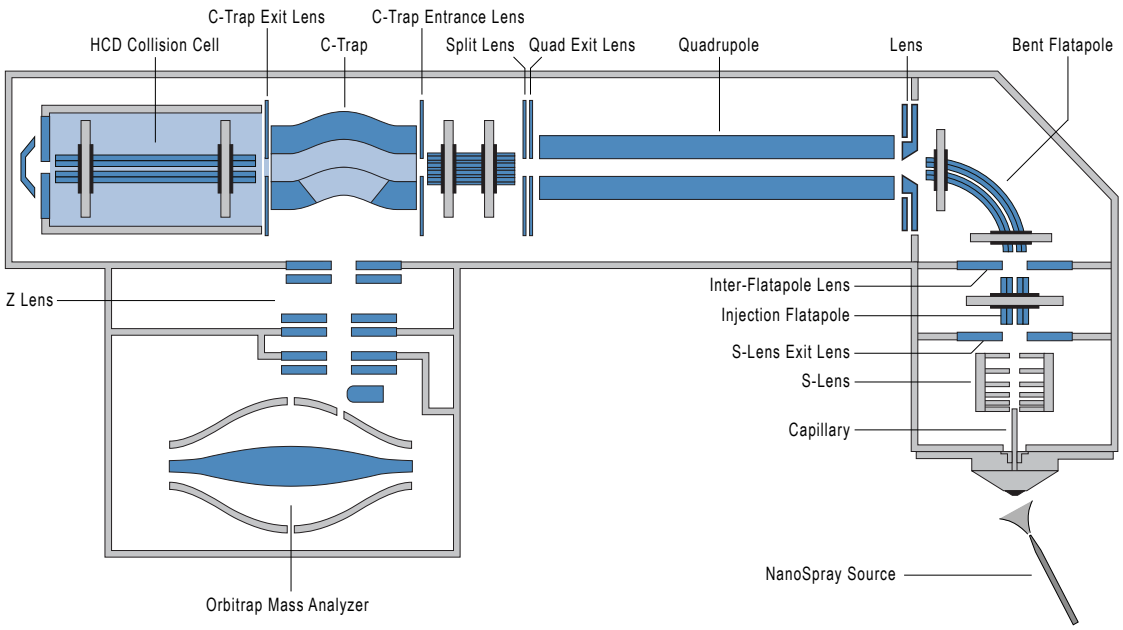

Figure 1.5: Schematic overview of a Q Exactive mass spectrometer. Ions are guided through ion optics (s-lens, flatapole) to the quadrupole, which can filter ions within a specific m/z range. The succeeding C-trap collects ions and injects them to either the HCD collision cell for fragmentation or the Orbitrap for high-resolution spectra generation. Picture adapted from Michalski et al. [22].

subsequently injected into the Orbitrap and higher-energy collisional dissociation (HCD) collision cell, respectively. In survey scan mode (MS^1), the total ion population is injected directly into the Orbitrap that consists of an electrostatic device, into which ions orbit around a spindle-shaped electrode [22, 25]. The image current of the ions' axial motion is detected, and this signal is Fourier-transformed in order to yield high resolution mass spectra. To analyze fragment spectra, the ions of interest are first filtered by the quadrupole and subsequently injected into the HCD cell, where ions are accelerated by an electrical potential and collide with neutral molecules such as nitrogen, resulting in a fragmentation of the peptides [26]. The fragments are then returned to the C-trap, which injects them into the Orbitrap to perform a fragment spectrum scan (MS^2). This is usually repeated for 10 peptides, until the next survey scan is performed.

To extract qualitative and quantitative information from the vast amount of MS spectra, sophisticated software like MaxQuant [27] is used to process the raw spectral files.

MaxQuant identifies peptides by comparing fragment spectra to theoretical fragment ion masses from a sequence database with the peptide search engine Andromeda [28]. Andromeda employs a probabilistic score that is based on the probability that the observed number of matches between the measured and the theoretical fragments occur by chance. For an accurate quantification, MaxQuant integrates peptide ion peaks, isotope clusters and labeled peptide tuples as three-dimensional objects in m/z, elution time and signal intensity space.

1.3 Analysis of phosphoproteomic data

One major interest in the area of phosphoproteomics is to unravel a drug's mode of action. In many cases, pharmaceutical investigators observe phenotypic effects when treating their model organisms with a certain drug; however, they do not know the exact cellular mechanism that governs this effect. Unraveling a drug's mode of action is vital during drug discovery and development, helping to identify new medical applications, suggesting its use in combination therapy, and predicting the responsiveness of patients [29–31]. In the case of phosphoproteomic mode of action analyses, such drugs are mostly small molecules or antibodies targeting one or several kinases [14, 32, 33]. Once a drug interferes with theses kinases and inhibits their activity, they are unable (or at least reduced in their ability) to phosphorylate their downstream targets. As a consequence, entire signaling cascades can be affected by the treatment, eventually resulting in a change of gene transcription, translation, apoptosis or the like. In the biological reality, however, such direct effects are often accompanied by secondary effects, like feedback mechanisms, or, if treatment lasts for hours, even changes in protein expression, which in turn has an influence on phosphorylation changes. Moreover, if the downstream effects of a certain inhibition are off known canonical signaling pathways, customized tools are needed to help uncover the mode of action. In Chapter 2, a method based on protein-protein interaction networks that reports significantly regulated subnetworks is presented. The algorithm, called SubExtractor, employs a Bayesian probabilistic model in combination with a genetic algorithm and rigorous significance testing.

When conducting mass spectrometry-based phosphoproteomic mode of action anal-

yses, the limiting factor regarding both time and cost is MS run time. To keep the experiments affordable, the number of biological replicates is often limited to three, which requires proper statistical methods that are able to deal with data that consist of only few replicates but thousands of features (phosphosites) at the same time. Moreover, mass spectrometers regularly produce - more or less randomly - missing data, i.e. even if the experiments are replicated three times, there is no warranty that every phosphorylation site is quantified three times. In Chapter 3, a global one-sample location test is presented, which is based on the mean ranks of the respective features across biological replicates. The hypothesis test, called MeanRank test, was specifically designed for experiments with few replicates and thousands of features, implicitly handles missing data, and internally controls the false discovery rate.

Another area of interest is the discovery of biomarkers in the field of personalized medicine. In general, there are four different types of biomarkers:

- Response prediction (stratification) markers foretelling the effect of a certain drug treatment on patients (e.g. HER2/neu over-expression for predicting response to treatment with trastuzumab (Herceptin®) [34, 35]).
- Diagnostic markers detecting a disease that already exists (e.g. PSA level for diagnosis of prostate carcinoma [36]).
- Prognostic markers predicting how a disease may develop (e.g. MammaPrint, the first *omics* marker for foretelling the risk that a breast tumor will metastasize [37]).
- Pharmacodynamic markers monitoring pharmacological response (e.g. FDG-PET imaging for monitoring tumor size [38]).

Chapter 4 deals with the identification of a response prediction biomarker for treatment of non-small cell lung cancer (NSCLC) with the kinase inhibitor dasatinib from phosphoproteomic data. This study was the first global and unbiased approach to develop a biomarker based on the differences in the basal phosphorylation levels of cancer cell lines. To this end, an elaborate biomarker workflow based on robust feature selection in combination with a support vector machine (SVM) was designed in order to find a multivariate phosphorylation signature that accurately predicts drug response.

Undoubtedly, the primary aim of a stratification biomarker is to accurately predict

drug response in patients. However, in many cases there are several limitations, for example, an effective application of a predictive phosphorylation signature in the clinical environment may require phosphorylation-specific antibodies, which are not available for all phosphorylation sites. For similar reasons, multivariate markers should not contain too many different features (i.e. phosphosites); and ideally, signature proteins should be meaningful, e.g. by having a connection to the drug's target or mechanism of action. Furthermore, it might be desirable to detect more than one signature in parallel. These requirements are fulfilled by the Pareto biomarker workflow in Chapter 5, where the standard workflow is extended to optimize not only one single objective (i.e. best possible separation), but also the objectives size and relevance, simultaneously. This is achieved by employing the multiobjective optimization algorithm NSGA-II [39] that is based on the principle of Pareto optimality.

Chapter 2

SubExtractor

In this chapter, a new method for identifying differentially regulated subnetworks from phosphoproteomic data is presented. The proposed algorithm, called SubExtractor, combines phosphoproteomic data with protein network information from STRING to identify significantly regulated subnetworks. The method is based on a Bayesian probabilistic model combined with a genetic algorithm and rigorous significance testing.

The content of this chapter was published as:
M. Klammer, K. Godl, A. Tebbe, and C. Schaab. "Identifying differentially regulated subnetworks from phosphoproteomic data." In: *BMC bioinformatics* 11 (2010), p. 351

The author was a key contributor to designing and implementing the algorithm, as well as writing the paper. The wet laboratory work (cell culture and mass spectrometry) were performed by his colleagues at Evotec Munich under the supervision of A. Tebbe.

2.1 Background

Global quantification technologies such as microarray, MS-based proteomics and phosphoproteomics can measure the expression of thousands to tens of thousands of genes, proteins and phopshorylation sites, respectively. Often, a few thousand of them are identified as being significantly differentially regulated, but interpreting these results at a single gene or protein level is a tedious and frequently unsuccessful task. However, by integrating these data with protein-protein interaction networks, it is possible to identify significantly regulated subnetworks that can be interpreted directly in a biological context. Moreover,

identifying regulated entities from often noisy high throughput data should be supported by this kind of integration.

One simple approach for detecting regulated subnetworks could involve distinguishing between significantly regulated and non-regulated phosphosites by applying standard hypothesis testing procedures such as t-statistics or SAM [41] to each phosphosite (the number of data points corresponds to the number of experimental replicates). To avoid too many false positives, one must further apply concepts such as the family-wise error rate (FWER [42]) or the false discovery rate (FDR [43]) for multiple hypothesis testing correction. Subsequently, the resulting list of statistically significant entities can be mapped on pathways or protein-protein interaction networks, and connected subnetworks can be determined. While this procedure may point to regulated subnetworks, it is not an integrated solution, since the significance of each protein solely depends on the data of its own phosphosites, regardless of its interactions with other proteins.

More sophisticated approaches use statistic-based techniques to score subnetworks. In these cases proteins are first mapped onto a protein interaction network, and subsequently high-scoring subnetworks are extracted. Ideker *et al.* [44] use an aggregated z-score of the form

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in S} z_i,$$

where k is the number of nodes in the subnetwork and z_i is the z-score of a single protein in the subnetwork S. High-scoring subnetworks are then found with a simulated annealing approach [45]. Chuang *et al.* [46] presented a method based on the same idea, but with a greedy search algorithm that specifies a seed and adds the best nodes in the neighbourhood until the aggregated score no longer improves. Subsequently, the significance of the resulting subnetworks is assessed based on null distributions estimated from permuted networks. However, neither method accounts for the network topology, i.e. the degree of interconnections between nodes.

Subsequently, Sanguinetti *et al.* [47] introduced a Bayesian probabilistic model that integrates *a priori* network topology information into the analysis of high throughput data. The authors used Gibbs sampling [48] to obtain suitable posterior probabilities and

thus derived subnetworks. A major drawback of this method, however, is the missing significance assessment for the resulting subnetworks.

All methods described above used either only a subset of known protein-protein interactions or KEGG pathways [49] for their assessment. To obtain the most information from such investigations, and considering that canonical pathway databases like KEGG are rather static and contain only a limited number of interactions, it seems natural to use larger and frequently updated protein-protein interaction network databases such as STRING [50] or FunCoup [51].

Here, we introduce a Bayesian probabilistic model that combines local as well as topological information, i.e. information about regulation of a certain node and information about the connectivity with its neighbours. Identification of subnetworks is carried out using a genetic algorithm (GA [52]), followed by performing a significance analysis based on a global rank test [53]. As a special feature, the significance test not only considers subnetworks, but also single nodes that are not part of any larger subnetwork. This makes the proposed method a powerful tool to uncover both differentially regulated subnetworks and differentially regulated single proteins. The performance was assessed on an artificial data set as well as on a comprehensive phosphoproteomics data set.

2.2 Methods

2.2.1 Data pre-processing and z-score calculation

The input of the proposed method is formed by a table with n rows and m columns; n being the number of detected phosphosites and m the number of biological replicates (i.e. MS measurements of experiments using identical settings but conducted independently). Several replicates (at least 3–5) are necessary to reliably identify differential phosphorylations. Each value in this table represents a ratio between the degree of phosphorylation under two conditions (e.g. the extend of phosphorylation of a specific site in cells treated with a drug versus its degree in untreated cells).

Log-transformation is preferred before calculating the z-score, since the distribution of the transformed ratios is closer to normal. Subsequently, the log-ratios x_{ij} of phosphosites

$i = 1, ..., n$ and replicates $j = 1, ..., m$ are further transformed to z-scores (referred to as single z-scores) using the formula:

$$z_{ij} = \frac{x_{ij} - \mu_0}{\hat{\sigma}}, \tag{2.1}$$

where $\mu_0 = 0$, since it is expected that the majority of phosphosites are not differentially regulated and therefore their log-ratios are 0, and $\hat{\sigma}$ the standard deviation across replicates estimated on the entire data set. Further, a combined z-score for each phosphosite over all replicates is calculated as:

$$z_i = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} z_{ij}. \tag{2.2}$$

Not all phosphosites are detected in every experimental replicate. The resulting missing values are simply ignored, so, for example, if three replicates have been conducted and a given phosphosite was only detected in two of them, m is set to 2 for this site and the combined score is calculated based on the two available z-scores.

2.2.2 Protein network preparation

In this work STRING [50] was chosen as the source for protein-protein interactions. STRING is a comprehensive resource that combines a vast number of databases derived in different ways (e.g. experimentally determined interactions, gene neighbourhood data, or data acquired via text mining) and is able to transfer homology information across organisms. Obviously the method presented here is not limited to STRING and can also be used in combination with other protein-protein-interaction databases. Depending on the context of the study databases like HomoMINT [54], HPRD [55], or FunCoup [51] may be preferable.

In STRING, all interactions are assigned with a confidence value ranging from 0 to 1. In order to retain only high confidence interactions, a very conservative cut-off value of 0.995 is used. While this cut-off may seem too high, there is a valid reason for it: some interactions reach very high confidence values (> 0.99), although the evidence is only from text mining, which was considered too weak evidence. Furthermore, analysis of canonical pathways showed that virtually all known interactions pass this high cut-off of

0.995. Applying this cut-off, an interaction network of approximately 10,000 interactions between 2,997 proteins is obtained (STRING version 8.1).

Subsequently, the phosphoproteomic data is mapped on the network (see upper part of Fig. 1). Before doing so, the list of phosphosites has to be aggregated to a list of proteins, with one z-score per protein and replicate. This is done by simply assigning the values of the phosphosite with the highest combined z-score among all phosphosites of a protein to this protein. Then each protein is mapped on the interaction network, where each node has m single z-scores and the combined z-score. Nodes that do not have a corresponding entry in the phosphoproteomics data set are thought of being not regulated and thus their z-scores are set to 0. On the other hand, proteins on the list that do not occur in the network are added but without any connections in order to give them the chance of being identified as regulated single proteins later on. In the genetic algorithm described below, only nodes in the interaction network will be considered; the set of unconnected nodes will be used again when it comes to significance assessment in the final step of the method.

2.2.3 Bayesian probabilistic model

A probabilistic model that takes into account the above derived z-scores and the network topology was developed. Let $c_i \in \{0, 1\}$ be the latent class variable, with $c_i = 1$ if node i belongs to a differentially regulated subnetwork and $c_i = 0$ if not. Note that the approach can easily be generalized to three classes, if up- and down-regulated subnetworks shall be distinguished. Given the combined z-scores $z_1, \ldots, z_n$ derived from the observations, the posterior probability of the subnetwork configuration $(c_1, \ldots, c_n)$ is

$$p\left(c_1, \ldots, c_n | z_1, \ldots, z_n\right) = \frac{p\left(z_1, \ldots, z_n | c_1, \ldots, c_n\right) p\left(c_1, \ldots, c_n\right)}{p\left(z_1, \ldots, z_n\right)}. \tag{2.3}$$

where the right-hand side is obtained by applying Bayes' theorem. The denominator $p\left(z_1, \ldots, z_n\right)$ does not depend on the c_i and can be ignored when maximizing the posterior probability. Since the observed data of node i are mutually conditionally independent (given the other nodes' class variables) and depend only on the class variable of the node

itself, the conditional probability can be written as

$$p\left(z_1,\ldots,z_n|c_1,\ldots,c_n\right)=\prod_{i=1}^{n}p(z_i|c_i). \tag{2.4}$$

Normal distributions $\mathcal{N}(\mu,\sigma)$ with $\mu=0$ and $\sigma=1$ or $\sigma=\sigma_z$ are assumed:

$$\begin{aligned} p(z_i|c_i=0)&=\mathcal{N}(z_i|0,1)\\ p(z_i|c_i=1)&=\mathcal{N}(z_i|0,\sigma_z^2). \end{aligned} \tag{2.5}$$

The prior probability for the subnetwork configuration $p\left(c_1,\ldots,c_n\right)$ is derived analogously to the derivation of the joint probability distribution from conditional probabilities in Bayesian networks. Let N_i be the set of parents of node i. If the protein interaction network was a directed acyclic graph and the joint distribution fulfilled the Markov condition, the following equality would hold [56]:

$$p\left(c_1,\ldots,c_n\right)=\prod_{i=1}^{n}p\left(c_i|\left(c_j,j\in N_i\right)\right). \tag{2.6}$$

Clearly, protein-protein interaction networks are no directed acyclic graphs. Nevertheless, the prior can be modelled by applying this theorem, if N_i is now defined as the set of neighbours of node i. The conditional probabilities are modelled similarly to [47]:

$$\begin{aligned} p\left(c_i=1|\left(c_j,j\in N_i\right)\right)&=\frac{\alpha+\frac{1}{|N_i|}\sum_{j\in N_i}c_j}{1+2\alpha}\text{ and}\\ p\left(c_i=0|\left(c_j,j\in N_i\right)\right)&=1-p\left(c_i=1|\left(c_j,j\in N_i\right)\right)\text{ or equivalently}\\ p\left(c_i|\left(c_j,j\in N_i\right)\right)&=\frac{\alpha+1-\frac{1}{|N_i|}\sum_{j\in N_i}\left(c_j-c_i\right)^2}{1+2\alpha}, \end{aligned} \tag{2.7}$$

where the parameter α determines the weight of the network structure, and $|N_i|$ is the number of neighbours. For very large α the posterior probability is not influenced by the network structure. Taking the logarithm of Eq. (2.3), inserting above equations, and ignoring the constant summands, the log posterior probability is:

$$\begin{aligned} &\ln p\left(c_1,\ldots,c_n|z_1,\ldots,z_n\right)=\text{const.}\\ &+\sum_{i=1}^{n}\ln\left(\mathcal{N}\left(z_i|0,(1-c_i)+c_i\sigma_z^2\right)\right)+\ln\left(\alpha+1-\frac{1}{|N_i|}\sum_{j\in N_i}\left(c_j-c_i\right)^2\right). \end{aligned} \tag{2.8}$$

The model parameters α and σ_z are fixed. In principle, they could be handled as unknown parameters in the Bayesian model, with the effect that the joint posterior probability would have to be maximized for $(c_1,\ldots,c_n)$, α and σ_z. Since the results turned

out to be rather insensitive to variations in α and σ_z (see *Results and Discussion*), the model and the optimization were simplified by *a priori* fixing of these parameters.

2.2.4 Subnetwork extraction

To maximize the posterior probability, the optimal combination of the nodes' class associations (i.e. whether a protein is part of a regulated subnetwork to be extracted or not) has to be found. Since this problem is NP-hard [44], a heuristic strategy has to be applied. Genetic algorithms (GAs) are particularly well-suited for this kind of binary-valued combinatorical problem, since they are able to find close-to-optimum solutions even in complex scoring landscapes with many local optima (see e.g. [52] for more details). An overview of a standard GA workflow can be found in Supplementary Information A.

To apply a GA to the subnetwork extraction problem, the network has to be encoded into a vector (i.e. an individual's chromosome). Here each node in the network was assigned a consecutive index value that represents the position of this node in the vector. The values in the vector are binary: 1 meaning that the corresponding node is part of a regulated subnetwork, and 0 that it is not (see also Figure 2.1). Initially, values of these binary vectors are randomly generated, one for each of the 1000 individuals used. According to the Bayesian scoring function described above, the fitness of each individual is evaluated and 100 individuals are selected and used for breeding. Selection of these individuals is performed using the tournament selector (cf. [57]), which randomly draws a subset of individuals and then determines the fittest within this subset. By repeating these steps 100 times, the 100 parent individuals are selected. Tournament selection ensures that average-performing individuals also have some chance to reproduce, which reduces the risk of premature convergence. Recombination of the selected individuals is carried out with two-point crossover, that is, the chromosomes of two parents are cut at two identical, random points $c1$ and $c2$, and the genes in the range $[c1, c2]$ are crossed (see also Figure 2.1). Mutation, which is a simple bit flip, occurs with a probability of 0.05. The newly created offspring's fitness is assessed, and the fittest offspring replaces the weakest individual in the parental generation. Then the algorithm continues with the selection of a new set of parents. The algorithm is run for 5000 generations, an empirically determined

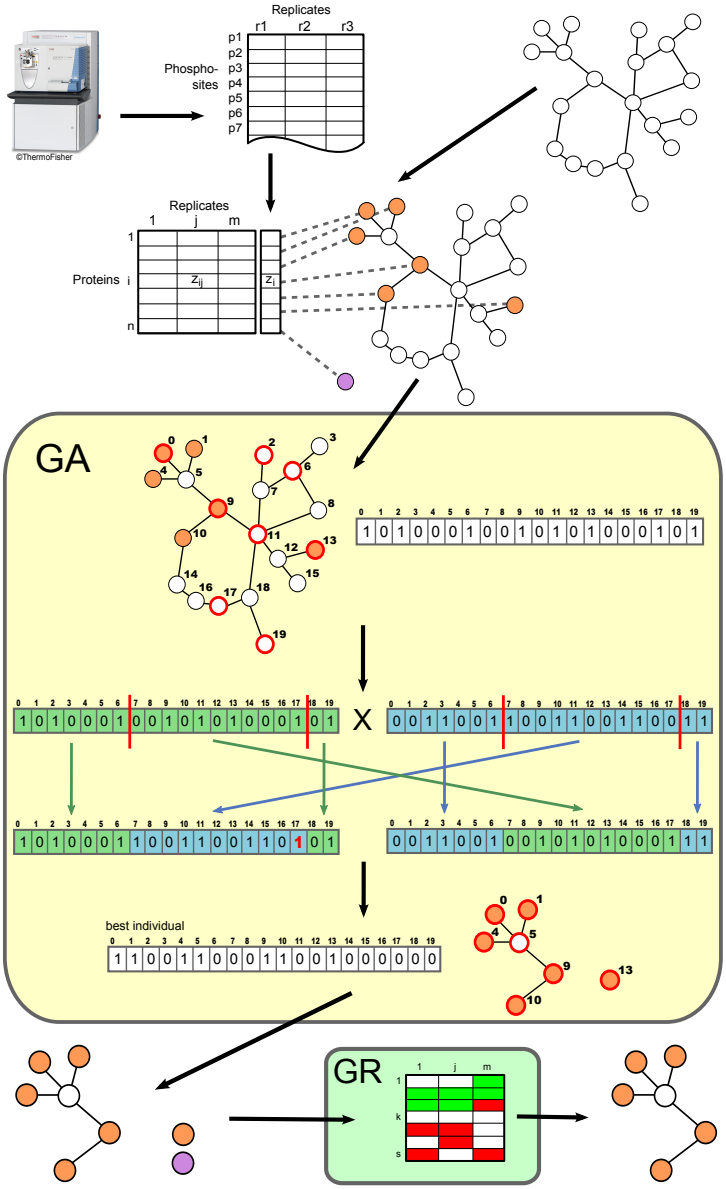

Figure 2.1: Workflow of the subnetwork extraction. First, z-scores are calculated from the phosphoproteomics data and mapped on an interaction network (orange nodes). Proteins that do not occur in the network are stored in a separate list (violet node). For the genetic algorithm (GA) procedure the network is encoded into a binary vector. The GA runs for a defined number of generations and the strongest individual of the final generation encodes for the globally best achievable solution. Finally, the global rank (GR) significance test is performed on both extracted subnetworks and single nodes resulting in a set of significantly regulated subnetworks.

value, from where on no more appreciable improvement is observed. The best solution (represented by the individual with the highest fitness value in the final generation) is then used to extract all subnetworks from the entire network by starting at a given node, checking all neighbours for their class association, and iteratively adding all neighbours that belong to a regulated subnetwork. To avoid cycles, every node is flagged after it has been checked, and if no more neighbours are to be added to the current subnetwork in a certain iteration step, another as yet unchecked node is used as the starting point for the next subnetwork. This is repeated until no unchecked nodes are left, and therefore all subnetworks are detected. The z-score of a subnetwork is then defined as:

$$z_s = \frac{1}{\sqrt{|S_s|}} \sum_{i \in S_s} |z_i| \,, \tag{2.9}$$

where z_i is the combined z-score of a protein as described in (2.2), S_s is the set of proteins in the subnetwork, and $|S_s|$ is its size. The absolute value of z_i is taken, since it is not know *a priori* whether the interaction between two proteins is activating or inhibiting, and therefore this distinction is not made. Rather only the degree of regulation is taken into account. When analysing gene or protein expression data, however, the direction of regulation may be important and should not be ignored. In such cases the signed values can be used. In some cases a subnetwork may contain only one node, which is not an issue, since both, significant subnetworks and single nodes shall be determined anyway.

2.2.5 Significance evaluation

Once regulated subnetworks are extracted, one has to determine their statistical significance. Single nodes (those that could not be mapped on the network but had been detected in the phosphoproteomics experiment) are regarded as subnetworks with only one member and are thus added to the list of subnetworks. The significance test is based on a modified version of the global rank test [53].

The main idea of this method is to identify differentially regulated entities (genes, proteins or subnetworks) not based on hypothesis tests conducted for each entity independently, but rather based on the entire set of entities at once. Under the null hypothesis that entities are neither up- or down-regulated, the authors state the theorem of random ordering, i.e. that no entity can rank consistently high or low across all replicates. On

the contrary, those entities that do consistently rank top or bottom in all replicates are identified as being significantly regulated. The number of identified significant entities will then solely depend on the number that determines how many entities are considered *top* or *bottom* ranked (here denoted as N), e.g. if N is chosen to be a small number, only a few entities or none at all will be among the *top-N* or *bottom-N* across all replicates.

Raising N not only increases the number of identified significant entities, but also the expected number of false positives. As described in [53], this number of false positives can be estimated non-parametrically from the empirical null distribution. The idea for this procedure is that a non-regulated entity has the same probability of ranking *top-N* as ranking *bottom-N*. In other words, under the null hypothesis an entity has the same probability of ranking *top-N* across all replicates (denoted as TTT for three replicates [$R = 3$]) as ranking *bottom-N* across all of them (BBB) or *top-N* in the first two and *bottom-N* in the third (TTB). The same is true for all $2^R = 8$ classes of possible combinations of high and low ranks. Entities in the TTT and BBB classes are differentially regulated, and those in the remaining $2^R - 2 = 6$ classes are not. By dividing the average number of entities in the 6 non-consistently regulated classes by the number of those in one of the regulated classes, for each N the FDR can be estimated (once for up- and once for down-regulated entities). Different values of N can now be tried until the desired FDR level is reached (cf. algorithm in Table 1, line 10 – 19).

For the application to subnetworks the method estimating false positives has to be modified, since the subnetworks' z-scores have non-negative values only, which means that *bottom-N* ranking subnetworks would be the ones with the weakest regulation. To overcome this problem, one first has to introduce another way of counting entities that fall under the non-consistently regulated classes, since the bottom ranked no longer represent differentially regulated entities. In this new counting process, not simply the entities in the non-regulated classes are counted but rather the signs of the replicates' z-scores are alternately changed (cf. algorithm in Table 1, line 5 – 8) and subsequently the number of entities that consistently rank top across all replicates after this transformation are counted (cf. algorithm in Table 1, line 14 – 16). In the case of the TTB class, for example, rather than determining the number of entities ranking *top-N* in the first two replicates and *bottom-N* in the third, the signs of the third replicate's z-scores are flipped and one

Algorithm 2.1 The algorithm for significance evaluation in pseudocode.

$A = z$-transformed phosphoproteomic data (n sites, m replicates)

$STRING$ = STRING interaction data

$origSN$ = list of extract subnetworks from $STRING$ using A

flippedSNs = container for flipped subnetwork lists

for all $s \in$ Cartesian product $\{-1, +1\}^m$ without $\{(-1, \ldots, -1), (+1, \ldots, +1)\}$ **do**

flippedA = multiply values in column $(1, \ldots, i, \ldots, m)$ of A with the value at index i in s

add list of extracted subnetworks from $STRING$ using *flippedA* to *flippedSNs*

end for

$FDR = 1.0$

$N = n$

while $FDR >$ desired FDR cutoff **and** $N > 0$ **do**

$origCount$ = count subnetworks that are among the N most-regulated ones across all replicates in $origSN$

flippedCount = 0

for all flipped lists of subnetworks **in** *flippedSNs* **do**

flippedCount = *flippedCount* + number of subnetworks from list of flipped subnetworks that are among the N most-regulated ones across all replicates

end for

FDR = (*flippedCount* / number of lists in *flippedSNs*) / $origCount$

$N = N - 1$

end while

if $N > 0$ **then**

return list of subnetworks that are among the $N + 1$ most-regulated ones across all replicates in $origSN$

else

return empty list

end if

determines the number of entities now ranking *top-N* across all three replicates (those that are now in the TTT class). Note that both counting methods yield the same results, since it makes no difference whether one counts the number of *bottom-N* entities of a given replicate or the number of sign-flipped *top-N* ones.

The z-score of a subnetwork is as defined in (2.9), where z_i is the combined score over all replicates. To find subnetworks that are top ranked across all replicates z-scores have to be calculated for each replicate separately:

$$z_{sj} = \frac{1}{\sqrt{|S_s|}} \sum_{i \in S_s} z_{ij}, \tag{2.10}$$

where z_{ij} is calculated with equation (2.1). The problem here is that two nodes within a subnetwork – one with a highly positive and one with a highly negative score – would mutually neutralize each other. This effect is undesirable, since the direction of regulation does not matter for the application described here. On the other hand, if the absolute value of z_{ij} was taken, the sign-flipping used to calculate the FDR would have no effect. Thus, a trick is applied: if the sign of a given z_{ij} is in accordance with the z-scores of all replicates (i.e. if it has the same sign as $\sum_{j'} z_{ij'}$), z_{ij} will contribute positively to the score $\hat{z}_{sj}$, if not it will contribute negatively:

$$\hat{z}_{sj} = \frac{1}{\sqrt{|S_s|}} \sum_{i \in S_s} \left(z_{ij} \cdot \operatorname{sgn} \sum_{j'} z_{ij'} \right), \tag{2.11}$$

where sgn is the sign function. This equation is applied in line 12, 15 and 21 of the algorithm in Table 1 to find consistently top ranked subnetworks.

Entities that lack data in one replicate are accepted as differentially regulated, if they rank top in the remaining $m - 1$ replicates. This criterion compensates for missing data, a particular problem in mass spectrometry experiments.

2.2.6 Implementation

Pre-processing, z-score calculation and generation of the artificial data set was performed using Matlab. The SubExtractor algorithm is written in Java using the GA library Jenes (http://jenes.ciselab.org; version 1.2.0) and made available for download online at http://www.kinaxo.de/SubExtractor. Java version 5.0 or higher is required to run the program. Network diagrams were created with Cytoscape [58].

2.3 Results and Discussion

2.3.1 Artificial data

In order to benchmark and assess the proposed method, the algorithm was tested with artificial data. For this purpose scale free networks based on the algorithm described in [59] with 1000 nodes and an average connectivity of approximately 3.5 were generated. Artificial z-scores were produced by sampling values for 969 nodes from a normal distribution with $\mu = 0$ and $\sigma = 1$ representing non-regulated proteins (background distribution); three times for each entity to simulate experimental replicates. The values for the 31 regulated nodes were determined in a two-step procedure. Firstly, the means x were sampled from a normal distribution with $\mu = 0$ and $\sigma = 5$. Secondly, the actual replicate values were generated by drawing three times from a normal distribution with $\mu = x$ and $\sigma = 1$. All 31 regulated nodes are connected with each other forming one regulated subnetwork, which should be extracted by the algorithm as accurately as possible. This data generation process was repeated ten times, resulting in ten artificial data sets.

Different σ_z and α values were used to assess the subnetwork reconstruction. Values of the σ_z parameter ranged from 2.0 to 8.0. The parameter α that determines the weight of the network structure on the entire Bayesian score was varied within a range of 0.01 to 10. Figure 2.2 shows the mean prediction accuracies over all ten artificial data sets at an FDR level of 0.05 (with 100 GA individuals and 3000 GA generations). Not surprisingly, a σ_z value of 5.0 delivers the best results (see Figures 2.2a and 2.2b), which is the same value as used for sampling the regulated nodes. At the same time the graphs show a rather weak dependence on its exact value. Only very small values (e.g. $\sigma_z = 2.0$) lead to a considerable increase of false positive predictions (see Figure 2.2a), which was also expected since such values are already very close to the σ value of the background distribution. For α the best results could be obtained by setting its value between 0.5 and 2.5 (see Figures 2.2c and 2.2d). Lower values cause the model to put too much weight on the network structure, which causes especially weakly regulated nodes that are only connected to strongly regulated ones to be spuriously incorporate into the regulated subnetwork. Higher values, on the other hand, result in under-weighting of the network structure, which in turn causes an incorporation of moderately regulated nodes even if the majority of their

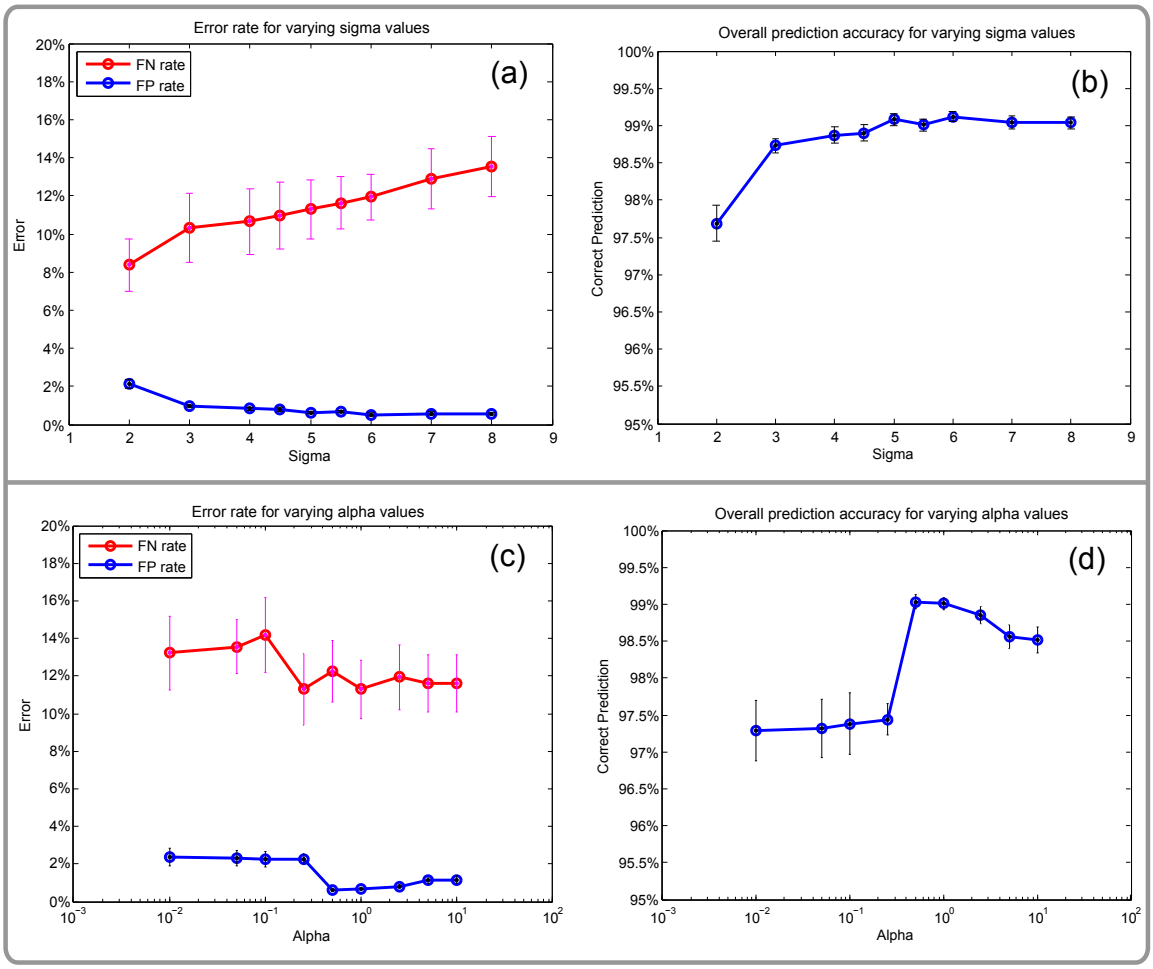

Figure 2.2: SubExtracor's performance on artificial data. Ten artificial data sets were generated to assess the prediction quality of SubExtractor. The top figures (2a and 2b) show the performance for varying σ_z values and a fixed α of 1.0. The figures at the bottom (2c and 2d) depict the mean accuracy for varying α values ranging from 0.01 to 10 and a fixed σ_z of 5.0. Nodes sampled with the background distribution ($\sigma = 1$) are the negatives, those coming from the distribution with $\sigma = 5$ are the positives. The FN rate is defined as $\frac{false\ negatives}{actual\ positives}$, the FP rate as $\frac{false\ positives}{actual\ negatives}$. The overall prediction accuracy is $1 - \frac{false\ negatives + false\ positives}{actual\ negatives + actual\ positives}$. Error bars display the standard error of the mean over the ten generated data sets.

neighbours are not regulated at all. Furthermore, one can clearly see that the results are not sensitive to the exact values of the parameters α and σ_z, which supports the decision to fix them *a priori*. However, the overall prediction accuracy steeply increases between α-values of 0.25 and 0.5 (see Figure 2.2d). This is due to the effect that if a non-regulated

node has only one connection to a well-regulated node (and no other connections) and α is smaller than a critical value α_c, it will be added to the differentially regulated subnetwork, just because of this special connectivity property. To avoid this undesired effect, α has to be chosen

$$\alpha > \alpha_c = \frac{\mathcal{N}(0|0, \sigma_z^2)}{\mathcal{N}(0|0, 1) - \mathcal{N}(0|0, \sigma_z^2)} \tag{2.12}$$

(the derivation of this formula and further explanation can be found in Supplementary Information A). For $\sigma_z = 5.0$ this leads to valid α values of $\alpha > 0.25$, which explains the large number of false positives for values ≤ 0.25 (as depicted in Figure 2.2c).

A detailed graphical view of the α parameter's impact on the prediction results can be seen in Figure 2.3, where the originally regulated network and three examples of networks reconstructed by the method (for a fixed σ_z of 5.0 and alpha set to 0.3, 1 and 5) are depicted. A small value of α just above α_c(Figure 2.3 top right) causes an acquisition of some low regulated nodes (the bright ones within the green circles), since the Bayesian score is mainly influenced by the network structure. On the other hand, one node is lost since it has many connections to non-regulated nodes but only a few to regulated ones (7 and 3, respectively) causing the network to break apart (upper right empty circle). For $\alpha = 0.3$ the algorithm extracts 4 false positive nodes while missing 3 true positives. On the contrary, a high value of $\alpha = 5$ (Figure 2.3 bottom right) causes the algorithm to almost entirely ignore topology information, and thus nodes are incorporated mostly according to their level of regulation. This leads to false positive classification of 5 nodes, of which 4 are fairly well-regulated (i.e. although they were sampled from the background distribution they received a high score by chance), and the fifth one–although not regulated itself–acts as a link to one of the well-regulated false positives. Only one of the true positives was missed. The results for $\alpha = 1$ (Figure 2.3 bottom left) form a good compromise between the previous two settings, as neither of the two score components is over-weighted. This reconstructed network has a lower number of false predictions (3 false positives and 1 false negative), which is a very satisfying result given that many nodes classified as regulated show very moderate regulation (weaker than some nodes from the background distribution).

To demonstrate the advantage of SubExtractor over a method that does not take network information into account, the original global rank test [53] was applied to the

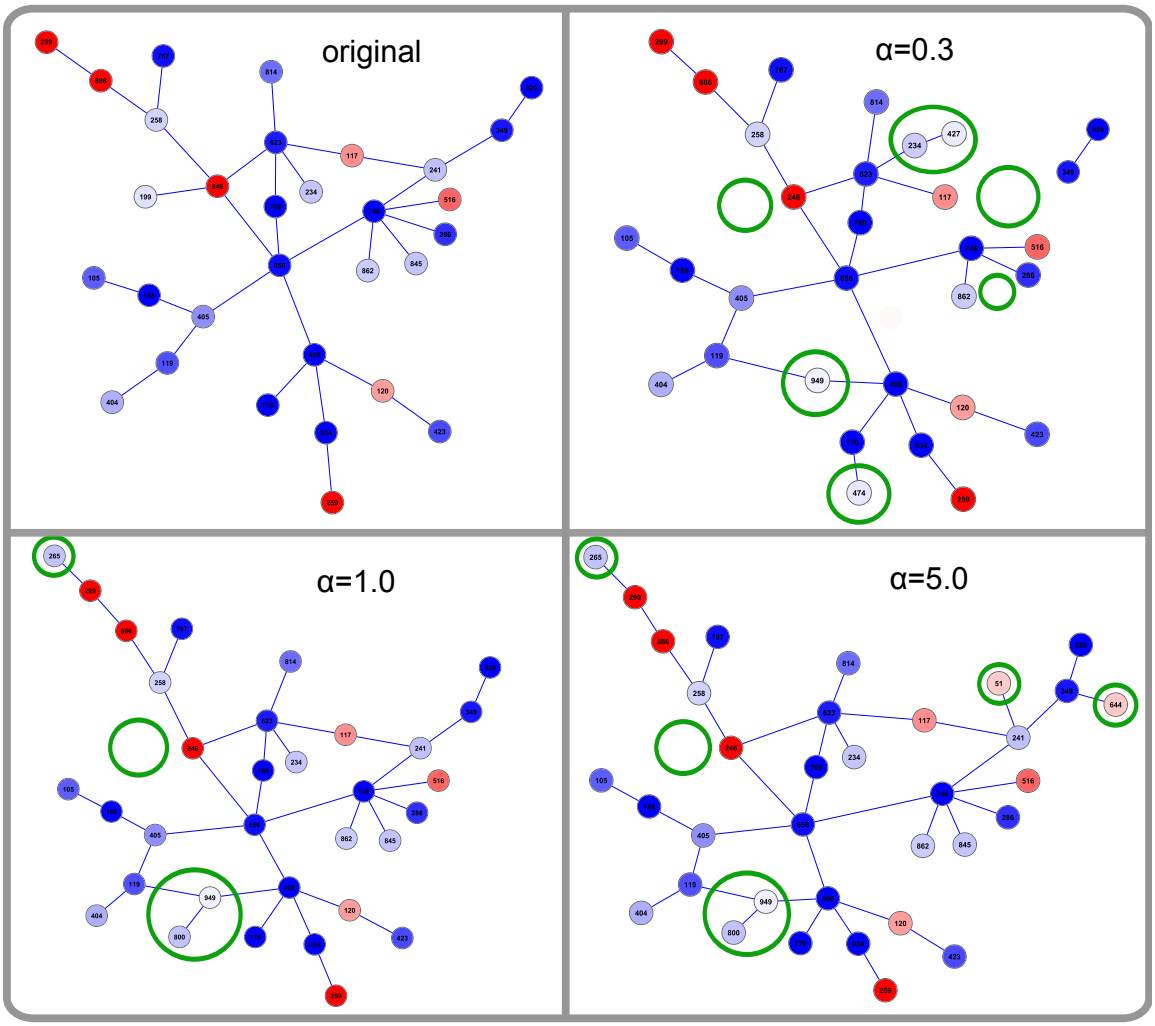

Figure 2.3: Example of subnetwork extraction for one artificial data set. The top left area shows the network of 31 nodes that have been sampled from the normal distribution with $\mu = 0$ and $\sigma = 5$, thus being the regulated ones in the artificial data set containing 1000 nodes in total. The remaining three areas show networks reconstructed by the proposed algorithm using different values of the parameter α. The colouring represents the level of regulation, where down-regulated nodes are coloured blue, up-regulated ones red and non-regulated nodes white (the darker the colour the stronger the regulation). The differences between the original and the reconstructed subnetworks are highlighted by green ellipses.

artificial data sets. The average false negative rate of this method at an FDR level of 0.05 was 29.0%, the average false positive rate was 0.2% (the best results of SubExtractor with $\alpha = 1.0$ and $\sigma_z = 5.0$ were 11.3% and 0.7%, respectively). Although SubExtractor

produces slightly more false positives, the superior capability to detect true positives even if they are only moderately regulated is obvious.

2.3.2 Sorafenib mode of action study

Subsequently the algorithm was applied to a real phosphoproteomics experiment, in which triply SILAC-labeled PC3 cells were incubated with the small molecule kinase inhibitor sorafenib (Nexavar®, Bayer HealthCare) for 30 and 90 minutes, including a control. Proteins were extracted and digested, and phosphopeptides were enriched using SCX-IMAC/TiO_2. High resolution LC-MS/MS data of three biological replicates were processed using MaxQuant [27].

A total of 15,800 class-1 sites (i.e. highly confident phosphosites) on 3,900 unique proteins were detected. Since two time points are not sufficient to perform any sensible time-course analysis, the more time point with the more extreme absolute value of its average log ratios (either $\log \frac{30min}{ctrl}$ or $\log \frac{90min}{ctrl}$) over the three replicates is taken for each phosphosite. Phosphorylation sites were then pre-processed as described in the *Methods* section. Interaction data was taken from STRING version 8.1 [50] and pre-processed as described in *Methods*. The α parameter was set to 1.0, based on the observations made from artificial data. σ_z was estimated by applying the original global rank method [53] to the list of phosphosites and calculating the standard deviation of the resulting differentially regulated sites' combined z-scores, which led to a value of $\sigma_z = 5.5$. Other parameter values were also tested, resulting in very similar networks (data not shown). This supported the findings from the artificial data study, where it has been shown that results are rather insensitive to the exact parameter values.

At an FDR level of 0.05 the proposed algorithm was able to reconstruct 21 significantly regulated subnetworks with 168 nodes in total. Additionally, 225 individual proteins were identified as significantly regulated. A selection of the results are depicted in Figure 2.4. Besides parts of the MAPK pathway, which is known to be affected by sorafenib, the largest network contains a substantial fraction of proteins from the mTOR pathway, which was previously not known to be affected. Subsequent enrichment analyses of the mTOR KEGG pathway confirmed the results of SubExtractor (p-value < 0.005 using Fisher's

exact test; data not shown). In particular, a substantial number of translation initiation factors (eIF's) show regulation of phosphorylation upon sorafenib treatment.

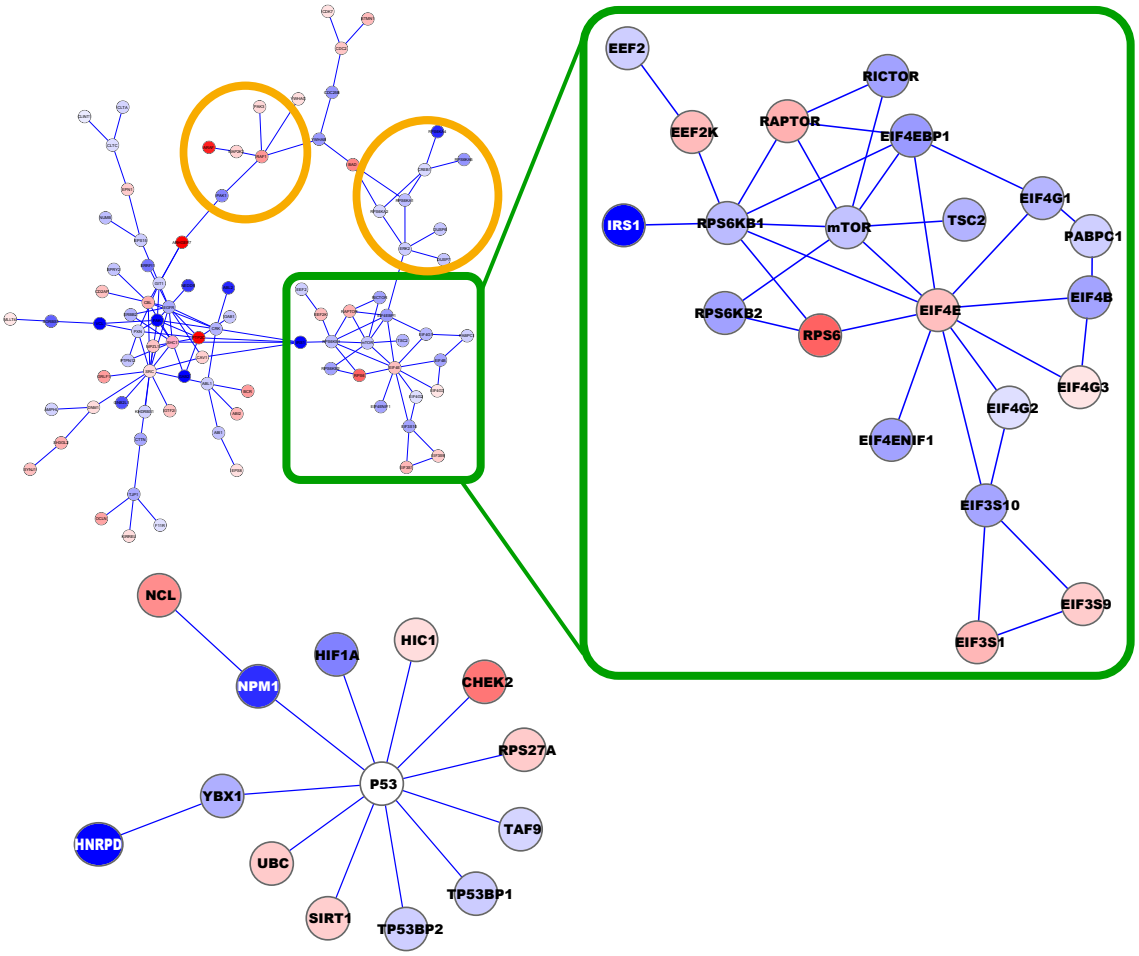

Figure 2.4: Subnetwork extraction for sorafenib mode of action study. The largest two resulting subnetworks are shown (blue nodes are down-regulated, red ones up-regulated). Proteins in the orange circles belong to the MAPK pathway, which is known to be affected by sorafenib. The green rectangle depicts the part of the largest subnetwork that belongs to the mTOR pathway, which has not previously been reported to be affected by sorafenib. The network on the right hand side shows an important strength of the algorithm, i.e. that subnetworks are also reconstructed if the centre node (i.e. the hub) is not detected to be regulated.

Another example in Figure 2.4 depicts a subnetwork centring the tumour suppressor p53. This example shows the strength of the method to reconstruct networks, even if

the hub of the subnetwork is not phosphorylated, not detected, or not regulated. Greedy search methods that grow subnetworks by selecting a seed and iteratively expand it by adding regulated neighbours cannot identify such subnetworks.

The complete result in Cytoscape session file format is provided as Additional File 2, and in Excel format as Additional File 3.

2.3.3 Normal distribution assumption

Both regulated and non-regulated phosphosites were assumed to be normally distributed with different variances (1 and σ_z, respectively). Hence, a mixture model of these two distributions should well describe the experimental data. To further investigate this assumption we created a probability plot, which is used to assess whether data comes from a given distribution. However, the plot (see Supplementary Figure A.2) indicates that a mixture model of standard normal and t location scale distribution (essentially a normal distribution with heavier tails) fits the data better than the mixture of the two normals.

Next, the impact of the different distributions on the SubExtractor results was assessed by modelling the regulated data (cf. Equation 2.5) with a t location scale distribution with the mean parameter set to 0, a variance of σ_z^2 and 6 degrees of freedom (estimated based on the fit above). However, the results of the t-normal mixture model were strikingly similar to those of the normal-normal mixture, suggesting that the slightly better fit of the former does not increase the prediction accuracy (compare Additional Files 2 and 4). Given the simplicity of normal distributions (i.e. in comparison to t distributions no degrees of freedom have to be estimated) and the comparable results, the normal-normal mixture model was considered preferable.

2.3.4 Alternative STRING network preparation

Instead of applying a very conservative cut-off of 0.995 to the combined STRING interaction score, an alternative version was created where the score was re-computed omitting text mining evidence. The computation was performed according to [60], and should avoid very high confidence values that are only due to sometimes doubtable text mining evidences. For the re-computed score the cut-off was set to 0.95, which is still conservative

but increases the number of interactions by 80% and the number of involved proteins by 20%. SubExtractor was then run with this version of network information and the sorafenib data (all parameters were left unchanged). While the general tendency of affected pathways and groups of proteins is very similar, the nodes of the largest network have roughly doubled making it rather complex (see Additional File 5). The decision on which network data file to use is left to the user, as it may depend on the application whether he prefers rather complex but comprehensive networks or smaller networks that are easier to interpret. Both files are available for download at http://www.kinaxo.de/SubExtractor.

2.4 Conclusion

Here, we propose a novel method, SubExtractor, for extracting differentially regulated subnetworks from protein-protein interaction networks based on data from global quantification technologies. The core of the method is formed by a Bayesian probabilistic model that accounts for the regulation of proteins as well as for the network structure. A genetic algorithm was implemented to find the subnetworks that maximize the Bayesian score. Furthermore, a global rank significance test was used to distinguish between significantly regulated subnetworks and those formed by chance.

Although some parts of the method have already been presented elsewhere (cf. *Introduction*), the main advantage of the proposed method is the combination of the three main parts: Bayesian probabilistic model, powerful heuristics in the form of GA and rigorous significance testing. To our knowledge none of the existing methods offer this combination. Additionally, the significances of single nodes (i.e. either proteins that could not be mapped on the interaction network or extracted single-node networks) are also assessed, which makes separate statistics on a protein scope redundant. Using data from the comprehensive STRING database guarantees high reliability of the detected interaction subnetworks.

The method was tested with artificial data sets and showed a high level of reconstruction accuracy. Knowledge from this study was transferred to a mode of action study, where SubExtractor revealed differentially regulated subnetworks from known and novel sorafenib-affected pathways, e.g. the MAPK- and mTOR-pathway, respectively. These reg-

ulated subnetworks led to creating new hypotheses about the mode of action of sorafenib in prostate cancer PC3 cells. Furthermore, the subnetworks may also play an important role in discovering biomarkers. It has been shown [46] that identified markers for class prediction are more reproducible if their identification is based on subnetworks rather than single genes. Generalization of the proposed method for identifying subnetwork markers used for class prediction will be the focus of future work.

Chapter 3

MeanRank test

In this chapter, a new hypothesis test specifically designed for experiments with few replicates and thousands of features is presented. The proposed test, called MeanRank test, is a global one-sample location test, which is based on the mean ranks across replicates. The MeanRank test internally estimates and controls the false discovery rate, and handles missing data without the need of imputation.

The content of this chapter was published as:
M. Klammer, J.N. Dybowski, D. Hoffmann, and C. Schaab. "Identification of significant features by the Global Mean Rank test." In: *PloS one* 9.8 (2014), e104504

The author was a key contributor to designing and implementing the algorithm, as well as writing the paper. The figures in this chapter were created by J.N. Dybowski, the co-first author of this publication.

3.1 Background

Today, omics-technologies are capable of generating vast amounts of data. Typical microarray experiments measure the abundance of thousands of features. With recent advances in the field of mass spectrometry (MS), over 10,000 proteins can currently be measured in cell systems [62], while recent studies identified even more phosphorylation sites through quantitative phosphoproteomics [63–65].

Many of these microarray and proteomics studies include the detection of differentially regulated features as core step in the data analysis. For data with thousands of features,

the false discovery rate (FDR), defined as the expected number of false positive features among those reported as significant, has to be controlled [66]. However, strong control of the FDR reduces the rate of true positive features (TPR) discovered. The problem is often aggravated by experimental designs with small numbers of replicates. Further complications arise from missing data, especially common in MS-based shot-gun proteomics experiments. Microarray technologies often produce non-normally distributed expression levels and non-identical distributions between genes [67].

In principle, single-feature hypothesis tests like Student's t-test or the Wilcoxon rank-sum test can be applied to assess the significance of each feature, if results are corrected for multiple testing, e.g. by Benjamini-Hochberg (BH) [43] or the family-wise error rate (FWER) [68] procedures. However, when applied to data with only few replicates, these approaches are lacking statistical power, due to difficulties in estimating variance. Tusher *et al.* developed the Significance Analysis of Microarrays (SAM) [41], a more sophisticated method based on a modification of the t-statistic. The FDR is controlled by a permutation-based approach and adjusted using an estimate of the fraction of truly unregulated features. Moreover, SAM employs k-nearest-neighbor (k-NN) imputation to replace missing data. A similar approach is taken by empirical Bayes methods. Linear Models of Microarrays (LIMMA), for example, uses a moderated t-statistics, in which the estimated sample variance is shrunk towards a pooled estimate across all features [67].

Recently, methods applying a global approach, rather than determining significance on a feature-by-feature basis, were proposed. These methods take into account the entire dataset at once and thus avoid the difficult task of estimating the variance of each feature. Zhou *et al.* proposed a rank-based, global one-sample location test, which performs very well for small numbers of replicates and internally controls the FDR [53]. However, this global rank test requires features to consistently rank high or low across all replicates. The RankProducts test [69] is based on a similar global approach, but the ranks of each feature are multiplied. The FDR is then estimated numerically using random rank matrices.

The MeanRank test presented here borrows concepts of the GlobalRank and RankProducts tests, but uses a different test statistics and a different method for estimating the null-distribution. In the following, we describe the concept of MeanRank, including its handling of missing data. While we focus on the one-sample case in the main text, exten-

sions to the two-sample case are discussed in Appendix B. The one-sample location test problem is equivalent to the paired-difference test problem for dependent samples. Paired samples are very common in proteomics experiments, which often apply labeling methods such as SILAC or iTRAQ, but also in transcriptomics (e.g. two-color microarray). We then present an extensive simulation study, in which the performance of MeanRank is compared to the previously mentioned tests, the t-test, and the Wilcoxon signed-rank test. In order to demonstrate the value of MeanRank, it is compared to SAM and LIMMA on the 'Ag-Spike' two-color microarray spike-in data set recently published by Zhu *et al.* [70]. Finally, MeanRank and SAM are applied to datasets of two published phosphoproteomics-studies.

3.2 Methods

3.2.1 MeanRank test

Given a matrix M of R columns (replicates) and N rows (features, e.g. genes, proteins, phosphorylation sites). Let M_{if} be the value of feature f (with $f = 1, ..., N$) in replicate i (with $i = 1, ..., R$). Based on this matrix M, for each replicate i the ranks r_{if} of each feature f within this replicate and across all features can be determined by sorting the values in each replicate. This is in contrast to the Wilcoxon signed-rank test, for which the ranks are calculated across the replicates. Then the mean rank is calculated for each feature across all replicates. Similar to the approach of Zhou *et al.*, [53], the mean rank statistic is motivated by the random ordering theorem, i.e. under the null hypothesis H_0 that no feature is either up- or down-regulated, it is very unlikely that a feature ranks consistently high or low across all replicates. Therefore no extreme (very large or very small) mean rank values can be expected. In contrast to Zhou *et al.*, who require features to rank top or bottom consistently across all replicates, the mean rank statistic may tolerate some moderate outliers.

For simplicity, we will focus on the detection of significantly down-regulated features in the following, but the same approach is applicable for up-regulated features by simply switching the signs of all values. The mean rank test proceeds in these steps:

1. Sort features ascendingly by their values within each replicate

2. Calculate mean rank as

$$\overline{r}_f = \frac{\sum_{i=1}^{R} r_{if}}{R} \tag{3.1}$$

3. Sort values $\overline{r}_f$ ascendingly ($\overline{r}_f^*$) and identify the top n as significantly down-regulated

In case of tied ranks, the values are left in the original order, receiving ascending ranks. The list of significantly regulated features depends on the value of n, which has to be chosen to meet the specified FDR. The FDR is defined as the expected fraction of false positives among the reported positives. Following the approach of Zhou *et al.* [53], we denote $\alpha^0(n)$ the expected number of false positives among the top n features. The FDR is thus

$$\mathrm{FDR}(n) = \frac{\alpha^0(n)}{n}. \tag{3.2}$$

As the true form of the null distribution is not known, we have to estimate a null distribution either parametrically or non-parametrically. For a parametric estimate, we assume that the mean ranks of the null distribution follow a Bates distribution, i.e. the distribution of the mean of statistically independent uniformly distributed random variables. The cumulative distribution function is defined as:

$$\mathrm{F}_{\mathrm{Bates}}(m, x) = \frac{1}{2m!} \sum_{k=0}^{m} (-1)^k \binom{m}{k} (mx - k)^m \mathrm{sgn}(mx - k) \tag{3.3}$$

where m is the number of random variables and x is the mean of the random variables scaled to the interval $(0, 1)$, and $\mathrm{sgn}(a)$ is -1 for $a < 0$, 0 for $a = 0$, and 1 for $a > 0$. The expected number of false positives is then calculated as:

$$\alpha^0(n) = \mathrm{F}_{\mathrm{Bates}}\left(R, \frac{\overline{r}_n^*}{N}\right) \cdot N \tag{3.4}$$

Non-parametric estimation of $\alpha^0(n)$ follows Zhou *et al.*, assuming a non-regulated feature has the same probability of ranking top or bottom [53]. Thus, the null distribution is independent of whether the features are sorted in ascending or descending order, or – analogously – whether the features values have a positive or negative sign. Consequently, $\alpha^0(n)$ can be estimated by alternately flipping the signs of the ratios of the replicates, calculating the flipped mean ranks $\overline{r}_{flipped}$ on this flipped data, and counting the number of values in $\overline{r}_{flipped} < \overline{r}_n^*$ (see pseudo-code in Appendix B).

Missing data

To account for missing data values, which are especially common in MS-based proteomics experiments, the equation in step (3.2) of the algorithm has to be modified to

$$\overline{r}_f = \frac{\sum_{i=1}^{R} r_{if}}{\hat{R}_f}, \tag{3.5}$$

where $\hat{R}_f$ is the number of present data values of the respective feature and $r_{if} = 0$ if the value is missing. It has to be ensured that missing values are not considered in the ranking process and consequently do not receive a rank (they are ignored completely). The FDR estimation has to be modified as well, as there are now features with different numbers of data values in the dataset. Thus, the parametric estimation of false positives has to be modified to

$$\alpha^0(n) = \sum_{i=1}^{R} \mathrm{F}_{\mathrm{Bates}}\left(i, \frac{\overline{r}_n^*}{N}\right) \cdot N_i \tag{3.6}$$

where N_i is the number of features with i data values present. The non-parametric estimation of $\alpha^0(n)$ remains largely unchanged, however, one has to ensure that features unaffected by sign-flipping are excluded. This occurs when sign-flipping is by chance applied only to values that are missing. The resulting feature would be unchanged and receive the same ranks and subsequently mean rank, after flipping.

3.2.2 Simulations

Artificial data was generated by sampling from various distributions. The background distribution (unregulated data) containing 3600 features was drawn from a normal distribution with zero mean ($\mu = 0$) and standard deviation $\sigma = 0.1$; 400 regulated features (80 up- and 320 down-regulated) were sampled from normal distributions with shifted means (shift $\Delta = 0.2$). We investigated the performance with an increasing number of replicates (3 to 15). The described settings were then altered to simulate variable variance by drawing σ from a uniform distribution $\sigma \sim \mathcal{U}(0.05, 0.25)$)in combination with constant regulation strength between features ($\Delta = 0.3$) and variable regulation $\Delta \sim \mathcal{U}(0.2, 0.4)$. Missing data were introduced by randomly discarding 20% of data points while ensuring

that at least two thirds of the data points were present for each feature. In simulations of non-normal data we sampled features from a t-distribution with two degrees of freedom. Whenever imputation of missing values was applied, the k-nearest neighbor ($k = 10$) method was used.

Significance analyses by RankProducts, SAM and LIMMA were performed using the *RankProd* [71], *samr*, and *limma* packages of Bionconductor [72] for R [73], respectively. The global rank method by Zhou *et al* [53] was implemented by the authors. t-test and Wilcoxon signed-rank test p-values were BH corrected for multiple hypothesis testing [43].

3.3 Results and Discussion

3.3.1 Simulated data

In order to evaluate the performance of the MeanRank test and to compared it with various other tests, we performed an extensive simulation study extending the range of scenarios found in comparable publications [53, 74, 75] by including more parameters and wider ranges of replicates and methods. The advantage of simulations is that underlying statistical properties are known and, thus, the performance of different hypothesis tests can be compared under various conditions. In the first set of simulations we assessed the performance of the one-sample location tests for different sampling distribution parameters. Simulation parameters were strength of regulation (Δ), within-feature variance (σ^2) – both of which were either held constant or chosen to be variable – and the presence of missing values. These parameters were combined to generate different simulation scenarios. We calculated the performance for an increasing number of replicates for the respective scenarios. The parameters were deliberately chosen to simulate experiments with hard-to-identify regulated features to investigate the added power over a wide range of additional replicates. With the chosen settings, a true positive rate (TPR) of 1.0 should not be achieved easily.

The simplest simulation setting assumes a constant variance and strength of regulation. Figure 3.1A shows TPR and FDR achieved by the tests when 3,600 unregulated features were sampled with constant $\sigma^2 = 0.01$ and 400 regulated features were sampled

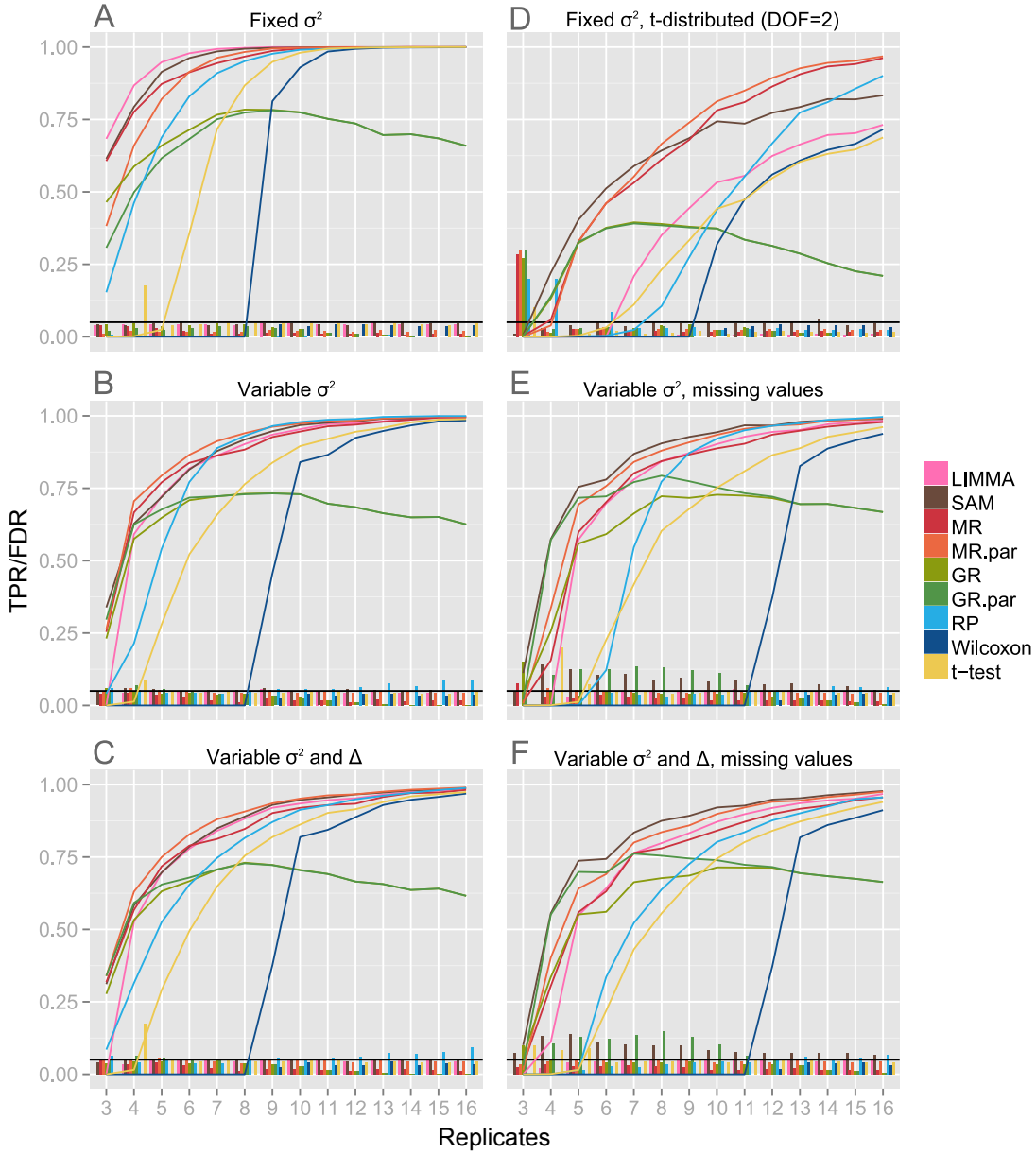

Figure 3.1: Performance on simulated data. Performance plot of one-sample significance tests under different simulation settings. Traces show the true positive rate (TPR) of the respective tests for a given number of replicates. Bars at bottom denote the false discovery rate (FDR). TPR and FDR are averaged over ten independent simulations. All tests were set to control the FDR at 0.05.

with a constant shift $\Delta = 0.2$. The leading method in this setting is LIMMA, followed closely by SAM, and then the non-parametric MeanRank (MR). This top-group clearly

outperforms the other methods. The parametric MeanRank test (MR.par) has a somewhat lower power for data with less than five replicates in this specific simulation setting. The power of the GlobalRank tests (GR and GR.par) does not scale with the number of replicates, but reaches its maximum performance at nine replicates. Additional replicates will even lead to a loss in power. This behavior is expected, because with a growing number of replicates it becomes less likely for a regulated feature to consistently rank top or bottom. Similar to the parametric MeankRank (MR.par), the parametric GlobalRank (GR.par) is less powerful than its non-parametric counterpart for less than five replicates. In contrast to the GlobalRank, the power of the RankProducts (RP) scales well with the number of replicates, but it is less powerful for experiments with small number of replicates. The TPR curves of GlobalRank and RankProducts underline the initial motivation of developing the MeanRank test, i.e. combining the strengths of both tests without inheriting their shortcomings. The t-test shows significant lower TPR, most likely due to variance estimation issues, especially evident at very small number of replicates. As an example of a non-parametric, rank-based test that does not belong to the class of global approaches, we included the Wilcoxon signed-rank test. Because of the discreteness of the test statistics, it is not surprising that a minimum of nine replicates is required to identify any significantly regulated feature after multiple hypothesis testing correction. For eleven or more replicates the TPR approaches the TPR of the other tests beside the GlobalRank tests. All tests correctly control the FDR at the pre-specified level of 0.05.

Next, we investigated the scenario with feature dependent variable variance, which is frequently observed in omics data due to the dependence of the variance on the signal intensity [76]. Overall the tests display a similar behavior as in simulations with constant variance (Figure 3.1B). However, while the overall TPR is slightly lower for most tests with variable σ^2, the parametric MeanRank and GlobalRank tests seem to be largely unaffected. Thus, the discrepancy between the parametric and non-parametric versions, which was observed for small number of replicates, disappears. Furthermore, MeanRank has a slightly higher overall TPR than SAM or LIMMA under these simulation conditions. The small gain in power for the t-test results from features with small variance caused by the variable σ^2 setting.

We then combined the variable variance σ^2 with a variable regulation strength Δ,

reflecting the complex response of systems to perturbations, e.g. of cells to drug treatment. There is a further loss in power across all tests, since some of the regulated features are hidden in the background noise (Figure 3.1C). The parametric MeanRank performs best across all replicate numbers. The non-parametric MeanRank, SAM and LIMMA, exhibit comparable but slightly reduced power. In general, the behavior of all tests is similar to the previous simulation (Figure 3.1B).

When using heavy-tailed distributions, such as a t-distribution, SAM and MeanRank exhibit similar power until up to seven replicates. However, while MeanRank progresses to a TPR of 1.0 for 15 replicates, SAM has by then just reached TPR 0.8 and almost levels off. (Figure 3.1D). The power of LIMMA is considerably reduced compared to the previous scenarios and is comparable to the power of the t-test. The GlobalRank shows particular problems with this setting, achieving a TPR of merely 0.4, before starting to drop. The RankProducts even falls behind the t-test for less than nine replicates.

Missing data are common in technologies such as MS-based shotgun proteomics, thus in the next set of simulations we introduced missing values combined with variable variance σ^2 (Figure 3.1E). It should be noted, that SAM is the only method used that does not handle missing data intrinsically. Instead, it employs a k-NN imputation prior to the actual significance analysis. In terms of power, parametric and non-parametric MeanRank together with SAM and LIMMA delivered the best results. For small numbers of replicates, the power of GlobalRank was comparable to that of MeanRank and SAM. However, SAM and the parametric GlobalRank systematically underestimated the FDR.

We additionally simulated the effect of missing values on data with both variable variance σ^2 and shift Δ (Figure 3.1F). Here, the parametric and non-parametric MeanRank, SAM and LIMMA perform best with respect to the TPR. As in the previous scenario, SAM always underestimated the FDR considerably. In order to investigate whether the violations of FDR threshold observed for SAM were due to imputation, we also applied the other tests to the imputed data (see Supplementary Figure B.2). This resulted in similar behavior: a general violation of the FDR threshold, accompanied by a slightly higher TPR. Although it can be argued, that this is not a problem of SAM *per se*, the inability of handling missing data makes imputation inevitable.

Zhou *et al.* [53] stated that, in contrast to single-feature analysis methods, large num-

bers of features are advantageous for global methods and will lead to increased statistical power. We tested whether this applies to MeanRank, by altering the proportion of regulated and background features for a constant number of six replicates (see Supplementary Figure B.3). The hypothesis was confirmed, revealing that the rank-based tests (MeanRank, GlobalRank, and RankProducts) possess more power when the proportion of regulated to background features is small. The opposite is true for the single-feature-based tests, such as *t*-test, SAM and LIMMA. Despite experiencing a loss of power over an increasing fraction of regulated features, MeanRank always met the desired FDR threshold, while GlobalRank increasingly violated this threshold.

The simulations show that the parametric MeanRank generally had a higher power than the non-parametric version. Thus, we only used the parametric test in the following real data experiments. In the following, we applied parametric MeanRank, SAM and LIMMA, i.e. the tests showing the best performance in the above simulations, to microarray spike-in data and finally to real experimental datasets, for which, of course, the identity of truly regulated features is not known. However, since we showed that all tests with exception of SAM meet the pre-specified FDR in a series of different simulation scenarios, we can judge the performance of the test by evaluating the number of regulated features identified.

3.3.2 Microarray spike-in data

Spike-in datasets are well-suited for the comparison of significance analysis methods, since the identity of truly regulated features is known before-hand. Here, we used the Agilent two-color microarray spike-in dataset ('Ag-Spike') consisting of 1300 differentially expressed and 2500 background cRNAs across 12 replicates [70]. In their study the authors explored different combinations of preprocessing methods (background correction, within-, between-array normalization) in order to identify optimal preprocessing routes for the detection of differentially expressed genes using LIMMA.

We used the published preprocessed data to compare performance of the parametric MeanRank test with that of SAM and LIMMA. Figure 3.2 shows the true positive and false discovery rates of the three methods on the differently preprocessed spike-in data.

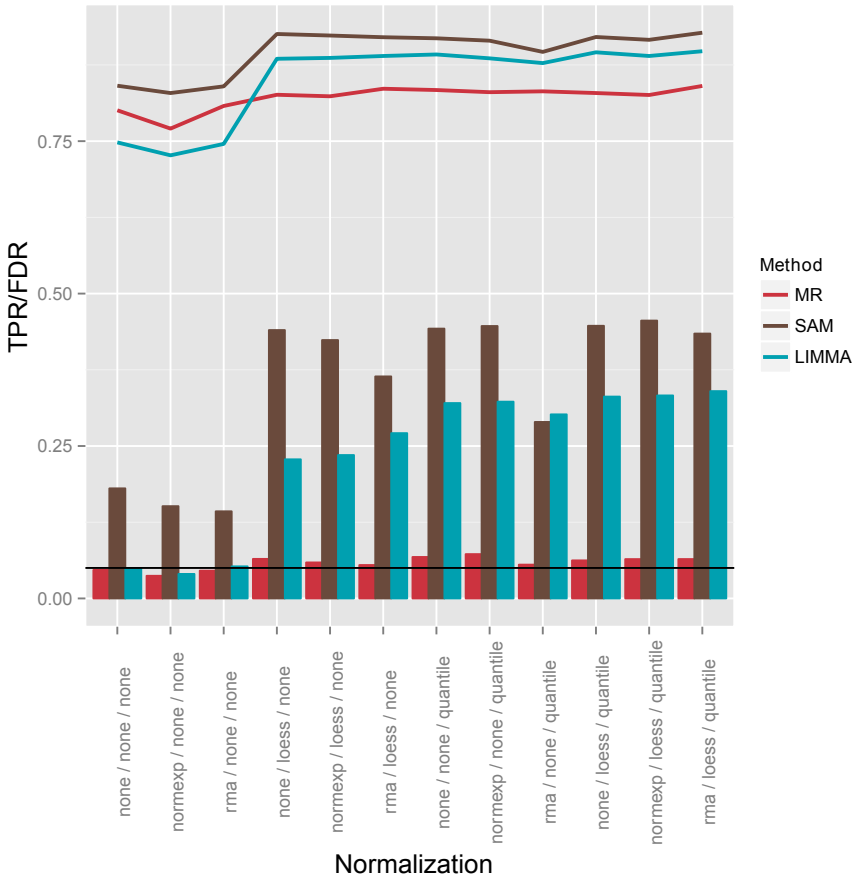

Figure 3.2: Performance on spike-in data. Performance comparison of MeanRank (red), SAM (brown), and LIMMA (cyan) on the 'Ag-Spike' microarray dataset [70]. TPR and FDR shown by lines and bars, respectively. Different combinations of preprocessing investigated by the authors of the original study are shown on the x-axis.

Most notably, the rank-based approach of MeanRank is very robust against changes in preprocessing: $CV_{\text{TPR}} = 0.04\%$ and $CV_{\text{FDR}} = 0.75\%$ compared to SAM ($CV_{\text{TPR}} = 0.21\%$, $CV_{\text{FDR}} = 5.02\%$) or LIMMA ($CV_{\text{TPR}} = 0.54\%$ and $CV_{\text{FDR}} = 6.02\%$). Slight variation is still introduced by methods applying local corrections, thus causing rank alterations (e.g. normalization *loess*). MeanRank on average identifies 2691 positives, 2354 (87%) of which are identified in all twelve preprocessing scenarios. The number of positives identified by SAM (4119) and LIMMA (3246) are higher on average, but clearly more dependent on the preprocessing protocol, with the number of constantly identified features being 2413 (59%) and 1989 (61%), respectively. This behavior is in line with the observations

of Zhu *et al.*, who in a prior study found that the preprocessing protocol has a great impact on the performance of methods for detection of differentially expressed features [77]. The power of MeanRank is comparable to that of SAM and LIMMA, when none or only minimal efforts of normalization are made. Additional preprocessing steps result in greater power for SAM and LIMMA, however at the cost of an under estimated FDR. Zhu *et al.* found that a combination of background correction by *normexp* and within-array normalization using *loess* yields the best result. This measure looks at the true positive and corresponding false positive rates given the absolute value of the test statistic. Hence, the correct estimation of the FDR is not taken into account. Figure 3.3 shows volcano plots of the *normexp*-corrected and *loess*-normalized spike-in data and highlights differentially expressed features as identified by the different tests. The column-like structure of data points on the x-axis reflects the levels of spike-in (see Supplementary Figure B.4). The largest column centered at zero contains features not regulated. SAM and LIMMA, in contrast to the MeanRank test, tend to produce more false positives as the feature variance decreases.

3.3.3 Phosphoproteomics data of *erlotinib*-treated AML cells

We applied MeanRank, SAM and LIMMA to phosphoproteomics data published by Weber *et al.* [78]. The authors of that study performed SILAC-based, large-scale, quantitative mass spectrometry analyses of KG1 acute myeloid leukemia cells treated with the small molecule tyrosine-kinase inhibitor *erlotinib*, which mainly targets the *epidermal growth factor receptor* (EGFR). In their subsequent significance analysis of ratios of *erlotinib versus* control treatment the authors applied the RankProducts test to identify 33 significantly (FDR 0.05) regulated class-I sites (i.e. phosphorylation sites identified with high confidence). Prior to testing, ratios of class-I sites were $\log_{10}$-transformed and subjected to sample-wise median normalization (cf. [32]).

The MeanRank test yielded 57 significantly regulated phosphorylation sites at FDR 0.05, including 24 of the 33 sites published by Weber *et al* (Figure 3.4, Additional File 2). Of the remaining 9 sites, 8 had a local FDR smaller than 0.07, thus missing the significance criterion only marginally. 27 of the additional 33 sites identified by MeanRank

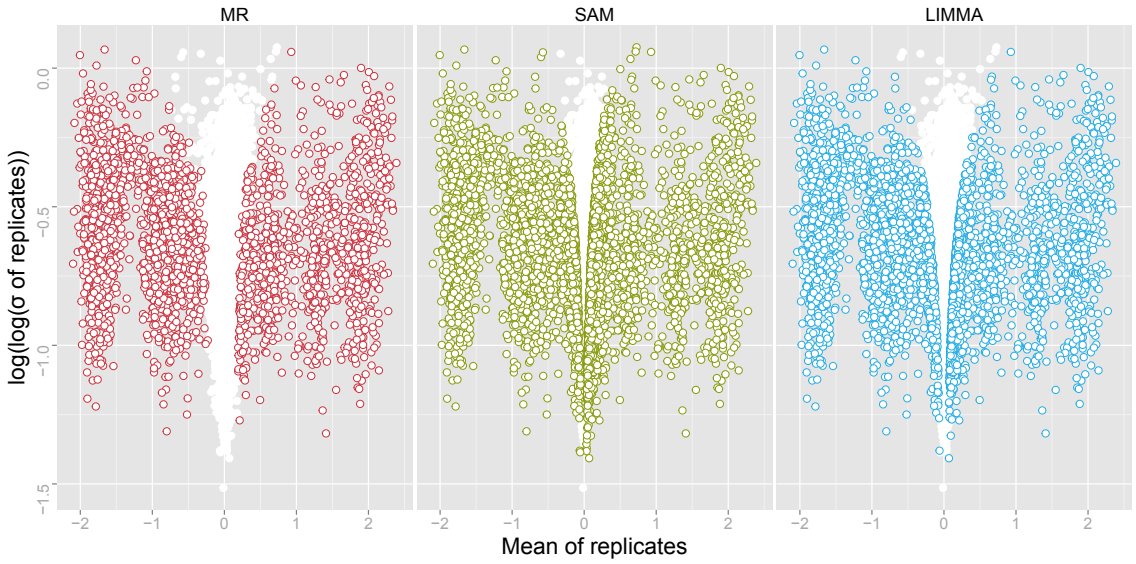

Figure 3.3: Volcano plot of spike-in data. Volcano plot of the 'Ag-Spike' data, background corrected by *normexp* and normalized with *loess*. This combination of preprocessing steps was found to deliver the best performance by the authors of the original study [70]. Genes are represented as points. Non-differentially expressed genes are scattered around Mean=0 on the x-axis. Differentially expressed genes, as identified by the respective methods are colored.

had a missing ratio, emphasizing the tolerance of the test towards incomplete data. SAM identified only 5 sites as significantly regulated, while LIMMA did not identify any significantly regulated phosphorylation sites at all. The sites newly identified by MeanRank are located on 29 different proteins. Most of these proteins are annotated as being involved in the *cell surface receptor signaling pathway* (GO:0007166). Weber *et al.* further found that most site-specific repression of phosphoserines by *erlotinib* occurred on proteins involved in mRNA translation control. Supporting this finding, the MeanRank test also identified several transcription factors (GTF2B, GTF2F1, GTF3C1, DEAF1, and TCF12) to be significantly regulated upon treatment. In addition, we identified 6 additional phosphotyrosines sites. As the primary targets of erlotinib are tyrosine kinases, this significant relative enrichment (Fisher's exact test $p < 9.5 \cdot 10^{-6}$) compared to the proportion of phosphotyrosins in the full dataset supports the findings of MeanRank. One of the sites that has not been identified as significantly regulated in the original paper is Tyr427 on

Figure 3.4: Volcano plot of AML data. Volcano plot of the phosphoproteomic data published by Weber *et al.* [78]. Significantly regulated phosphorylation sites are shown by colored circles as identified by SAM (left), the MeanRank test (right center), and in the original study (right).

the SHC-transforming protein 1 (Shc1). Tyr427 is phosphorylated *in-vitro* by Src kinase and *in-vivo* in EGF-stimulated cells [79]. Phosphorylated Shc1 forms a complex with Grb2 which in turn activates Ras signaling [80]. By down-regulation of Tyr427 on Shc1, erlotinib treatment inhibits the transmission of growth signals to the Ras signaling cascade.

3.3.4 Phosphoproteomics data upon reactivation of Plk1

We investigated the behavior of MeanRank and SAM on data from a second phosphoproteomics study. Here, telomerase-expressing human retinal pigment epithelial (hTERT-RPE) cells expressing an analog-sensitive Plk1 mutant ($Plk1^{as}$) were treated with the bulky kinase inhibitor 3-MB-PP1 [81]. 3-MB-PP1 inhibits the mutant kinase $Plk1^{as}$ harboring an enlarged catalytic pocket, but not wild-type Plk1. This allowed the investigation of downstream effects upon Plk1 reactivation by inhibitor wash-out. The dataset contained four biological replicates with a total of around 20,000 identified phosphorylation sites. In this analysis, we considered only sites with values present in all four replicates in order to avoid having to impute data for SAM analysis. This left around 5,200 phosphosites to

be tested for significant regulation upon inhibitor wash-out. Since SAM requires proper pre-processing, the data were $\log_{10}$-transformed and median normalized (cf. [32]).

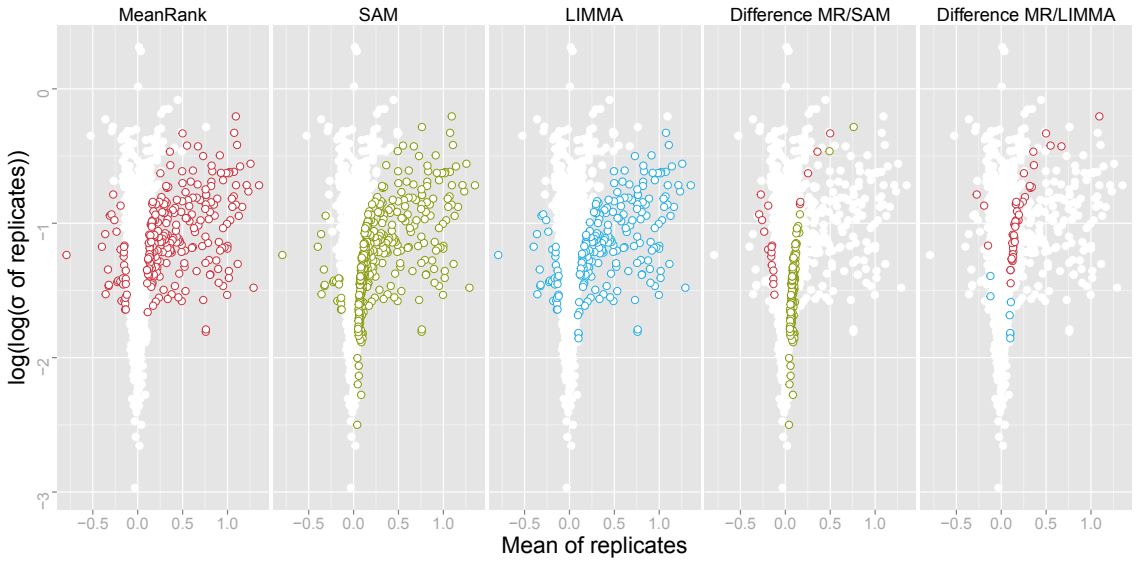

Figure 3.5: Volcano plot of Plk1-kinase-inhibited cells data. Volcano plot of the phosphoproteomic data of cells treated with an Plk1 tyrosine kinase inhibitor *versus* control [81]. Significantly regulated phosphorylation sites shown in colored circles as identified by MeanRank test, SAM, LIMMA (from left). The two rightmost volcano plots shows differences in detected phosphorylation sites by MeanRank/SAM and MeanRank/LIMMA.

While MeanRank identified 313 significantly regulated phosphorylation sites (FDR 0.05), SAM reported a slightly higher number of 359 significant sites for the same FDR level (Additional File 3). The overlap of the reported significant features was 249. SAM identified more significantly up-regulated features than MeanRank, most of which exhibit low variance and low mean regulation (Figure 3.5). SAM found 152 sites that were less than 1.5-fold up-regulated on a linear scale; MeanRank only 45. In contrast, SAM found only 8 sites that were less than 1.5-fold down-regulated (linear scale), while MeanRank reported 61. MeanRank draws a more consistent threshold between significantly up-, and down-regulated features than SAM.

LIMMA reports 229 significantly regulated phosphorylation sites, 225 are also identified by MeanRank. Similar to MeanRank, LIMMA mainly reports sites with mean regulation stronger than ± 1.5-fold on a linear scale. Only 23 sites with down-regulation of

less than 1.5-fold and 18 with an up-regulation of larger than 1.5-fold are reported.

Generally, it appears that SAM puts more emphasis on variance and MeanRank more emphasis on the level of regulation. This is reflected in the shape of the region within the volcanoplot, in which the significant features are located (Figure 3.5): While the significantly regulated features identified by MeanRank can be separated from the background features almost by a straight line, those identified by SAM exhibit a rather curved threshold line. LIMMA behaved similar to MeanRank, however identifying slightly less significantly regulated sites. Since the simulation study suggests that both tests comply with the pre-specified FDR level when applied to four-replicate experiments, it can be argued that at least 95% of the phosphosites reported as being significantly regulated by either test are in fact true positives. While all three tests perform well and have a high overlap, either might be more suitable depending on the application.

3.3.5 Two-sample test

Established methods such as SAM and LIMMA support two-sample comparison experiments. The MeanRank test can be extended to accommodate two-sample comparisons by basically transforming the two-sample into a one-sample problem. To do so, we create a difference matrix by calculating the difference of each possible pair from both groups. Here we assume that the data is log-transformed. The calculations of the mean ranks is then performed on the difference matrix in the same way as for the one-sample test. Since the columns of the difference matrix are not independent anymore, the dependency structure has to be taken into account when estimating the null distribution. We found that although the test generally performs well in terms of power compared to SAM for most cases in our simulations and spike-in microarray data [77] while reliably controlling the FDR, it is very conservative when applied to data with missing values. This can be explained by the way the difference matrix approach exaggerates the relative amount of missing values. The method and simulation setup is described in detail in Appendix B (see also Supplementary Figure B.1).

3.4 Conclusion

The simulations showed that borrowing traits from both the GlobalRank and RankProducts methods strongly improved the power over either of the two tests in all simulated scenarios, while reliably estimating the FDR. All three tests are rank-based and use a global approach rather than testing feature-by-feature. The main differences of the MeankRank test compared to the other two tests are the test statistics and the methods for estimating the distribution under the null-hypothesis. We showed that this improves the power of the test with respect to the RankProducts test for low number of replicates and avoids a drop in power with increasing number of replicates in the case of the GlobalRank test.

While single-feature-based non-parametric tests, such as the Wilcoxon rank-sum or signed-rank tests, require nine or more replicates in order to identify any significant regulated feature at all, this is not the case for global rank-based tests. The fixed σ^2 simulations showed, that the non-parametric MeanRank test identifies more than 60% of the true positives for three replicates.

The parametric and non-parametric MeanRank tests performed comparably to SAM and LIMMA in most simulation scenarios. While SAM and LIMMA performed slightly better in the case of fixed σ^2 simulations, MeanRank had a slightly higher power in the cases with variable σ^2 and both variable σ^2 and Δ.

When introducing missing data, our simulations suggest that SAM tends to underestimate the FDR, since missing values have to be imputed. This naturally raises concerns when applying SAM – and thus imputation – to data resulting from technologies like MS-based shotgun proteomics, regularly producing missing values. The matter is further complicated by the fact that different imputation methods (k-nearest-neighbor, singular value decomposition, multiple imputation, etc.) can deliver deviating results [82]. These aspects have to be considered, when applying SAM to data with missing values, while the MeanRank test offers a convenient way to entirely avoid imputation. However if, under certain conditions, imputation delivered results close to the ground truth, the power of any test would increase. A distinct advantage of the MeanRank test lies in the decoupling of significance testing and imputation procedures, leaving the freedom of choice with the researcher. If the data were not normally distributed but followed a heavy-tailed distri-

bution such as the t-distribution with few degrees of freedom, the one-sample MeanRank test showed a better performance than SAM and in particular LIMMA, especially for experiments with many replicates.

The global nature of the MeanRank test leads to a loss in power when a very large fraction of features is truly regulated. However, several studies suggest that the fraction of differentially regulated features is often lower than 10% [78, 81]. In fact, given such experiments, our simulations show that the MeanRank test has an advantage over single-feature-based tests like SAM and LIMMA. A notable practical advantage of MeanRank over other methods such as SAM is that normalization of samples is not necessary due to its rank-based nature. This is advantageous because normalization can have a direct influence on the results, as was demonstrated by our comparison based in the 'Ag-Spike' data. Here, MeanRank, in contrast to SAM or LIMMA, produced very stable results, independent of the preprocessing steps applied. SAM attempts to determine the proportion π_0 of true null hypotheses in the dataset in order to adjust the false discovery rate [66]. This usually leads to more positive calls; however, the estimation of π_0 is not robust against small variations in the data and depends strongly on the preprocessing applied. Since the FDR estimation of the MeanRank test is rather conservative, an implementation of a similar estimation could help to further improve the test with respect to statistical power. However, we deliberately omitted π_0 estimation because of the described inconsistent behavior also seen in other studies [83].

In summary, the key advantages of the MeanRank test compared to other tests are: a comparable or even superior power in detecting regulated features without underestimation of the FDR, the possibility to analyze data with missing values without the necessity for imputation; the robustness with respect to preprocessing. Although we focused on the one-sample test in the main text, a two-sample version of the test is also available and described in Appendix B. One-sample location tests are particular important for the analysis of proteomics data which often uses labeling methods such as SILAC or iTRAQ, but also for the analysis of two-color microarrays. Furthermore, they can be applied to paired two-sample test problems emerging, for example, if matched tumor and normal tissues are measured across many patients. The MeanRank test is not limited to testing the significance of gene- or protein regulation. As no strong assumptions about the un-

derlying distributions are made for the non-parametric test, inference about statistically significant differences between groups could, in principle, be made for any kind of ordinal features. Furthermore, MeanRank is freely available and can be used by anyone without any restrictions, whereas SAM is patented and requires proper licensing. For most experiments, running the MeanRank test is a matter of seconds, and can be performed on standard computers (see Supplementary Table B.1).

Finally, we would like to emphasize the intuitiveness of our test. MeanRank is easy to understand, easy to implement, does not require any parameter optimization and yields results that are easy to interpret.

Chapter 4

NSCLC biomarker

In this chapter, a phosphorylation signature that predicts the response to treatment with the kinase inhibitor dasatinib in non-small cell lung cancer (NSCLC) cell lines is presented. Quantitative mass spectrometry was used to globally profile the basal phosphoproteome of NSCLC cell lines; the effect of dasatinib on cellular growth was tested against the same cell line panel. An elaborate cross-validation workflow including robust feature selection and support vector classification was developed in order to detect a phosphorylation signature that accurately predicts sensitivity to dasatinib.

The content of this chapter was published as:
M. Klammer, M. Kaminski, A. Zedler, F.S. Oppermann, S. Blencke, S. Marx, S. Müller, A. Tebbe, K. Godl, and C. Schaab. "Phosphosignature Predicts Dasatinib Response in Non-small Cell Lung Cancer." In: *Mol Cell Proteomics* 11.9 (2012), pp. 651–668

The author was a key contributor to designing and implementing the algorithms, as well as writing the paper. The wet laboratory work were performed by his colleagues at Evotec Munich: A. Zedler, S. Blencke and S. Marx performed the cell culture work under the supervision of S. Müller; M. Kaminsky and F.S. Oppermann performed the MS analyses under the supervision of A. Tebbe.

4.1 Background

The introduction of targeted drugs for treating cancer is a major biomedical achievement of the past decade [84, 85] . Since these drugs selectively block molecular pathways that

are typically over-activated in tumor cells, they are more precise and less toxic than traditional chemotherapeutics. However, while many cancer patients benefit from a specific targeted therapy, many others do not. Therefore, predictive molecular markers are needed to confidently predict the patient's response to a specific therapy. Such markers would facilitate therapy personalization, where the selected therapy is based on the molecular profile of the patient.

Predictive tests currently used in the clinic are frequently based on one particular marker that is often linked to the drug's target. A well-known example for a predictive test is assessing HER2/neu overexpression using immunohistochemistry or fluorescent in situ hybridization to predict the response to therapy with trastuzumab (Herceptin®, Roche) (see [34, 35]). However, in some cases the expression or mutational status of the target or other singleton markers might not be sufficient to predict a therapeutic response. Recently, several studies tried to identify molecular signatures comprising multiple markers for response predictions, usually based on gene expression profiling (e.g. [86, 87]. To our knowledge, no study successfully identified a signature from global phosphoproteomic profiles so far.

Recent advances in mass spectrometry, methods for enriching phosphorylated proteins or peptides, and computer algorithms for analysing proteomics data have enabled the application of mass spectrometry-based proteomics to monitor phosphorylation events in a global and unbiased manner. These methods have become sufficiently sensitive and robust to localize and quantify the phosphorylation sites within a peptide sequence [15, 32, 63]. Phosphorylation events are important in signal transduction, where signals caused by external stimuli are transmitted from the cell membrane to the nucleus. Aberrations in these signal transduction pathways are particularly important for understanding the mechanisms of certain diseases, such as cancer, inflammation and diabetes [88, 89].

Approximately 391,000 incidences and 342,000 deaths from lung cancer were estimated in Europe in 2008 [90], accounting for nearly 20% of all cancer deaths in Europe. Approximately 85% of all lung cancer incidences are non-small cell lung cancer (NSCLC) [91]. Dasatinib (Sprycel®, Bristol-Myers Squibb) is a multi-kinase inhibitor targeting BCR-ABL, the Src-kinase family, c-Kit, ephrin receptors, and PDGFRb [11, 12]. It is currently approved for chronic myelogenous leukaemia and Philadelphia chromosome-positive acute

lymphoblastic leukaemia. Recently, dasatinib was clinically evaluated in patients with advanced NSCLC. Dasatinib had modest clinical activity, with only one partial response and twelve stable diseases among thirty patients. Neither Src family kinase activation nor EGFR and Kras mutations could predict the response to dasatinib [92].

In this study we wanted to identify a signature of protein phosphorylation that predicts the response to dasatinib in NSCLC cell lines. In total, 26 NSCLC cell lines were tested for their response to dasatinib. The identical cell lines were profiled in a global, unbiased, phosphoproteomics study and the obtained phosphoproteome profiles were used to assemble a biomarker signature of 12 phosphorylation sites. We evaluated the performance of this signature in a cross-validation set-up and investigated the robustness of the selected predictive features. Finally, we confirmed the predictive power of the signature in an independent set of breast cancer cell lines.

In a recent study, Andersen et al. identified phosphorylation sites predicting response to phosphatidy-linositol 3-kinase (PI3K) inhibitors [93]. Their study differs in two aspects from the study presented here. First, the authors focused on the PI3K and MAPK pathways by immunoprecipitating phosphorylated peptides with antibodies directed against corresponding phospho-motifs. In contrast, we followed an unbiased approach, where no hypothesis about the involved signalling pathways has to be made. Second, the authors first investigated the regulation of phosphorylation sites upon drug treatment in one sensitive cell line, and subsequently confirmed the applicability of one site to response prediction by evaluating its basal phosphorylation in a panel of cell lines. Here, we started directly by investigating the basal phosphoproteome of a panel of sensitive and resistant cell lines.

4.2 Methods

4.2.1 Cell culture

Based on the half-maximum growth inhibitory concentrations (GI_{50}) of dasatinib on a panel of 84 NSCLC cell lines reported in Supplemental Table 5 of Sos et al.[94], 13 cell lines with low and 13 with high GI_{50} values were selected (cf. Supplementary Table

C.1). These 26 cell lines were obtained from the LGC Standards (Wesel, Germany), from the DSMZ - Deutsche Sammlung von Mikroorganismen und Zellkulturen (Braunschweig, Germany), and Roman Thomas' group at the Max Planck Institute for Neurological Research (Cologne, Germany). The six breast cancer cell lines were obtained from the LGC Standards (see Supplementary Table C.1).

All cell lines were cultivated in RPMI1640, 10% foetal bovine serum, 2 mM glutamine, 1 mM sodium pyruvate and penicillin/streptomycin (PAA, Cölbe, Germany). Cells were routinely monitored for mycoplasma infection using the MycoAlert reagents (Lonza, Cologne, Germany). Metabolic labelling of the cell lines was performed using SILAC (stable isotope labelling with amino acids in cell culture [17]). Cells were cultivated in media containing SILAC-RPMI (PAA) and dialysed FBS (Invitrogen, Darmstadt, Germany). L-lysine and L-arginine were replaced by normal L-lysine (Lys-0) and L-arginine (Arg-0), or medium isotope- labelled L-$D_4{}^{14}N_2$-lysine (Lys-4) and L-$^{13}C_6{}^{14}N_4$-arginine (Arg-6), or heavy isotope-labelled L-$^{13}C_6{}^{15}N_2$-lysine (Lys-8) and L-$^{13}C_6{}^{15}N_4$-arginine (Arg-10). Isotope-labelled amino acids were purchased from Cambridge Isotope Laboratories (Andover, MA, USA). Cells were cultivated for a minimum of six doubling times to obtain an incorporation efficiency for the labelled amino acids of at least 95%.

16 NSCLC cell lines were selected as a reference pool: A549, Calu6, H1395, H1437, H1755, H2030, H2052, H2172, H28, H460, HCC827 (obtained from LGC Standards), LCLC103H, LouNH91 (obtained from DSMZ), H322M, HCC2279, HCC2429 (obtained from MPI for Neurological Research). The selected cell lines were grown in SILAC media supplemented with the natural 'light' forms of arginine and lysine. The labelled cells of each cell line were lysed, pooled, aliquoted, and stored at -80°C. In total, 40 aliquots with 12 mg of protein each were generated.

4.2.2 Determination of cellular growth inhibition

Sensitivity of the cell lines for dasatinib was determined by measuring the cellular ATP content after 96 hours of treatment using the CellTiter Glo chemiluminescent viability assay (Promega, Mannheim, Germany). Cells were cultivated in 96-well plates (Greiner, Frickenhausen, Germany) in the presence of dasatinib (LC Laboratories, Woburn, MA,

USA) within a concentration range between 3 nM and 30 μM.

The raw data from the chemiluminometer (FLUOstar OPTIMA, BMG Labtech, Offenburg, Germany) was used to determine the GI_{50} value. First, the background was determined by calculating the median value of the plate's border wells, which contained only growth media. This value was then subtracted from each inner well. Since two experiments were conducted on one 96-well plate with 10 compound concentrations each (0(DMSO), 3 nM, 10 nM, 30 nM, 100 nM, 300 nM, 1μM, 3 μM, 10 μM, 30 μM), three data points per concentration and experiment were available. Ratios representing the percentage of growth inhibition were calculated by dividing each data point coming from a concentration >0 by the median of the DMSO values. A logistic regression was performed to fit a curve to those ratios and compute the GI_{50} value.

4.2.3 Classification into sensitive/resistant

The calculated GI_{50} values of the 26 selected cell lines were compared with the values reported in [94]. Although the correlation between the two sets was strong (Pearson correlation = 0.50, p = 0.009 on logged GI_{50}s), a few cell lines showed inconsistent behaviour. By setting the threshold to discriminate between sensitive and resistant cells to a GI_{50} value of 1 μM, seven cell lines were classified inconsistently (5 were resistant in the reference paper, but sensitive in this study, 2 vice versa). Consequently, these cell lines were excluded from the workflow that aims at finding a predictive phospho-signature.

4.2.4 Phosphoproteomics workflow

Responsive and non-responsive cell lines were grown in medium or heavy SILAC media and after washing twice with ice-cold PBS the cells were lysed directly on the plates by the addition of ice-cold lysis buffer (8 M urea, 50 mM Tris pH 8.2, 5 mM EDTA, 5 mM EGTA, SIGMA HALT Phosphatase Inhibitor Mix, ROCHE Complete Protease Inhibitor Mix). After sonication cell debris was sedimented by centrifugation and the protein concentration was determined by Bradford assays. Equal protein amounts of the reference cell culture mix and a medium and heavy labelled cell line (7 mg protein each) were mixed as depicted in Supplementary Figure C.2 and subsequently subjected to reduction

(20 mM DTT, 30 min 37°C) and alkylation (50 mM iodoacetamide, 30 min RT) prior to proteolytic cleavage. Then 80 µg of LysC (Wako) was added for 4 h followed by a 4-times dilution with 50 mM Tris pH 8.2. Proteolytic cleavage was continued by the addition of 120 µg of trypsin (Promega) overnight. The peptide mixtures were acidified by addition of TFA to a final concentration of 0.5% and subsequently desalted via C18 SephPack columns (Waters). Peptides were eluted with 50% ACN and dried under vacuum. For a first separation of phosphorylated and non-phosphorylated peptides, the dried peptide powder was reconstituted in 1 ml SCX buffer A (5 mM K_2HPO_4, pH 2.7, 30% ACN) and loaded onto a polysulphoethyl column (9.4 x 250 mm, PolyLC) using an ÄKTA Purifier chromatography system equipped with a fraction collector. The peptides were separated by a linear gradient to 25% SCX buffer B (buffer A supplemented with 500 mM KCl) over 40 min at flow rate of 3 ml/min. Twenty fractions (12 ml each) were collected across the gradient.

Prior to IMAC enrichment the solvent of the SCX-fractions was removed by lyophilisation. Dried peptides were reconstituted in 1 ml of 0.1% TFA and desalted by using C18 reversed phase cartridges (Waters). The bound peptides were eluted with 50% ACN, 0.5% HOAc and the peptides were lyophilized again. Dried peptides were reconstituted in 40% ACN, 25 mM formic acid and phosphopeptides were captured using PhosSelect (Sigma) according to the manufacturer's instructions. Eluted phosphopeptides were subjected to mass spectrometric analysis.

4.2.5 LC-MS/MS Analysis

Mass spectrometric analysis was carried out by on-line nanoLC-MS/MS. The sample was loaded directly by an Agilent 1200 nanoflow system (Agilent Technologies) on a 15 cm fused silica emitter (New Objective) packed in-house with reversed phase material (Reprusil-Pur C18-AQ, 3 µm, Dr. Maisch GmbH) at a flow of 500 nl/min. The bound peptides were eluted by a gradient from 2% to 40% of solvent B (80% ACN, 0.5% HOAc) at a flow of 200 nl/min and sprayed directly into a LTQ-Orbitrap XL or LTQ-Orbitrap Discovery mass spectrometer (Thermo Fischer Scientific) at a spray voltage of 2 kV applying a nanoelectrospray ion source (ProxeonBiosystems). The mass spectrometer was

operated in the positive ion mode and a data dependent switch between MS and MS/MS acquisition. To improve mass accuracy in the MS mode, the lock-mass option was enabled. Full scans were acquired in the orbitrap at a resolution R = 60,000 (Orbitrap XL) or 30,000 (Orbitrap Discovery) and a target value of 1,000,000 ions. The five most intense ions detected in the MS were selected for collision induced dissociation in the LTQ at a target value of 5000. The resulting fragmentation spectra were also recorded in the linear ion trap. To improve complete dissociation of phophopeptides, the multi-stage activation option was enabled applying additional dissociation energy on potential neutral loss fragments (precursor minus 98, 49 and 32.7 Thompson). Ions that were once selected for data dependent acquisition were 90 sec dynamically excluded for further fragmentation.

4.2.6 MaxQuant analysis

The raw mass spectral data was processed using the MaxQuant software (version 1.1.1.25) [27] applying the Andromeda search engine for peptide and protein identification. The human UNIPROT database (version: 57.12) was used comprising 110,595 database entries including the UNIPROT splice variants database. The minimal peptide length was set to 6 amino acids, trypsin was selected as proteolytic enzyme and maximally 3 missed cleavage sites were allowed. Carbamidomethylation of cysteines was selected as fixed modification, whereas methionine oxidation, N-terminal protein acetylation and phosphorylation of serine, threonine and tyrosine residues were considered as variable modifications. As MaxQuant automatically extracts isotopic SILAC peptide triplets, the corresponding isotopic forms of lysine and arginine were automatically selected. The maximal mass deviation of precursor and fragment masses was set to 20 ppm and 0.5 Da before internal mass recalibration by MaxQuant. A false discovery rate (FDR) of 0.01 was selected for proteins and peptides and a posterior error probability (PEP) below or equal to 0.1 for each MS/MS spectrum was required. The MaxQuant results were uploaded to the MaxQB database [95] for further analysis.

4.2.7 Data pre-processing

Data from MaxQuant's PhosphoSTY table were the data source for identifying a predictive phospho-signature. Each entry in this table describes one specific phosphosite along with information about its localisation, confidence and regulation. The regulation of a phosphosite is provided as ratio of the site's abundance between each cell line and the super-SILAC standard. MaxQuant already provides normalized ratios, which were used in this study. There are two coefficients that account for the reliability of identification and localization of a phosphosite, i.e. Localization Probability and Score Diff. Sites that satisfy the constraints Localization Probability ≥ 0.75 and Score Diff ≥ 5 were considered to be sufficiently reliable (class-I sites). Furthermore, sites that are flagged as Reverse or Contaminant hits were also excluded. All phosphosites that fulfill both requirements (class-I, no contaminant/reverse) were subjected to further analysis. The identification and quantification data on the class-I sites, as well as the fragment spectra of the best localization evidence are accessible in Additional Files 2-5 (Appendix C).

4.2.8 Analysis of differential phosphorylation sites

Significance analysis

After preprocessing the data, a Wilcoxon rank-sum test was applied to find differentially abundant phosphorylation sites between sensitive and resistant cell lines. For this analysis only phosphosites with values in at least two thirds of the experiments in each group were considered (i.e. at least 8 of 11 sensitive and 6 of 8 resistant data points had to be present). Subsequently, the p-values reported by the Wilcoxon rank-sum test were corrected for multiple hypotheses testing by applying Benjamini-Hochberg FDR correction [43].

Enrichment analysis

To analyze whether proteins harboring differentially abundant phosphorylation sites are enriched in certain GO terms [96] or KEGG pathways [49], FatiScan enrichment analysis [97] was applied. In brief, FatiScan performs a segmentation test, which checks for asymmetrical distribution of biological labels (e.g. GO terms, KEGG pathways) associated with proteins in a ranked list. For this purpose, the phosphorylation sites were sorted according

to their q-values and the algorithm was set-up to search for a possible enrichment in the low-q-value area of this ranked list. The analysis was performed via the Babelomics web interface (http://babelomics.bioinfo.cipf.es/, version 4.2).

Detection of significantly different subnetworks

In order to visualize and interpret the data in a network context, the SubExtractor algorithm was applied [40] (see also Chapter 2. In brief, SubExtractor combines phosphoproteomic data with protein-protein interaction data via a Bayesian probabilistic model. Regulated subnetworks are found with a genetic algorithm and subsequent significance evaluation based on the global rank test [53]. The STRING database version 8.3 [50] was used as source for protein-protein interactions. It was preprocessed to contain only human interactions with a confidence score larger than 0.9 without considering text mining evidences. The algorithm's parameters were set to $\alpha = 0.5$ and $\sigma = 5.0$, and subnetworks with an FDR smaller than 0.1 were reported.

To calculate z-scores required as input for the algorithm, pair-wise phosphorylation abundance differences between sensitive and resistant cell lines had to be computed first. Since the number of experiments in the two groups are not balanced (11 and 8, respectively), sampling with replacement was applied to the smaller group (i.e. it was sampled 11 times from the 8 experiments ensuring that each experiment was chosen at least once). Subsequently, the pair-wise differences could be computed along with the estimated global standard deviation as suggested in [40]; and finally the z-scores were calculated.

4.2.9 Identification and evaluation of phospho-signature

Cross validation

The data set containing $N = 19$ objects was split into two parts, one containing data of one cell line, and the other containing the data of the remaining $N - 1$ cell lines. The larger part was then used for training a predictor (training set) and the smaller one for testing this predictor (test set). By alternating the cell lines that made up the training set, each cell line was used once for testing. Each of the N cross validation steps included missing data imputation, feature selection, predictor training and predictor testing (see

also Supplementary Figure C.3). A phosphosite was only considered as a potential feature if it had training data values in at least two thirds of the experiments in each class (e.g. if the training set contained data from 10 sensitive and 8 insensitive cell lines, at least 7 and 6 training data points had to be present, respectively). Since this criterion uses the class-labels, the features have to be filtered within the cv-loop. This further means, that the filtered features may be different in each cv-step.

Data imputation

For each phosphosite and class the mean and standard deviation was computed and the missing values were filled by sampling from the resulting normal distribution. This procedure was only applied to the training data, since the test data should be handled as if the class association was unknown. Nevertheless, test data can also contain missing values. If so, the mean of the corresponding two group means was imputed, which is an unbiased way of replacing the missing value that does not involve information about the test sample's class association. Geometrically speaking, the imputed test sample value is located exactly halfway between the two class means, which should minimize its influence on the prediction process.

Feature selection

In this study, a simple Wilcoxon rank-sum test in combination with the ensemble feature selection method [98] was used. Since the Wilcoxon test often delivers identical p-values due to its rank-based nature, ties were broken by preferring features that have a larger difference in their two classes' medians. The core idea of the ensemble method is that robust features should still rank among the best if the dataset is slightly modified. For this purpose, different samplings of the training data were generated by drawing (with replacement) 50 different bootstrap samples (i.e. if the training set consists of 10 sensitive and 8 resistant cell lines, one randomly draws 10 and 8 times with replacement from the respective set to get one bootstrap sample). The Wilcoxon rank-sum test is applied to each sample, and thus a diverse set of feature rankings is generated. The ranks of each feature were then averaged across all bootstrap runs and sorted in descending order according to this meta-ranking. Subsequently, the k best features were used to train and test the

predictor. By varying $k = 1 \ldots 200$ and assessing the prediction accuracy and area under the receiver operator curve (AUROC), one can find the optimal number of features.

Support Vector Machine training

Once a set of features has been selected, and the training and test data have been modified to include only those features (i.e. 'reduced' sets), a SVM with linear kernel (see e.g. [99]) can be trained. Besides the kernel function, an SVM has a parameter C, which controls the trade-off between margin maximization and training error minimization, if the hyper plane cannot perfectly separate the two classes. The default value of C=1 was used throughout the analysis. First, the SVM was trained with the training data. Subsequently, the class association of the test data was predicted with the trained SVM. The result of this prediction is the probability of the test sample belonging to either of the two classes (the closer the test data is to the decision boundary, the less confident the prediction is). The class prediction with the larger probability was then taken and compared to the actual class association. In this way, correct predictions were counted across all cross validation steps.

Area under the receiver operating characteristic curve

To calculate the area under the receiver operating characteristic curve (AUROC), the separating hyper-plane of a trained SVM was shifted by introducing cost matrices. For example, by shifting the hyperplane towards the group of sensitive training samples, it becomes more likely for a test sample to be classified as resistant. Ultimately, this shifting leads to the extreme that every test sample is classified as resistant, which means that all resistant test samples have been classified correctly (true negative rate = 1 and false positive rate = 0, given that the resistant ones are the negatives) and all sensitive test samples wrongly (true positive rate = 0). The exact opposite is true if the separating hyperplane is shifted towards the resistant group. Thus, by applying different cost values, one can control the degree of shifting, calculate the respective true positive rates and false positive rates, and compute the resulting area under the curve by means of the trapezoidal rule (see Appendix C for an example).

Random seeds

For the imputation of missing values, a random number generator is needed to sample values from a normal distribution. Different seeds of the random generator will produce different imputation data. To avoid a bias of the data towards the seeding, the entire cross validation procedure was repeated five times using different random number generator seeds. The prediction accuracies, AUROC values and global feature rankings for different numbers of selected features (k) were averaged over the five CV runs and used for the final selection of the phospho-signature.

Data Normalization

Among the fraction of non-phosphorylated peptides, 15 peptides had values in at least two thirds of the experiments and a standard deviation < 0.1 (log10 scale). Eight of them were from ribosomal proteins, which are expected to be constantly expressed. Thus, for each experiment the median of the corresponding eight ratios was computed and used as an alternative normalization approach (by subtracting the median from each phosphosite's non-MaxQuant-normalized logarithmic ratio).

Final predictor construction

When selecting the final set of phosphosites (phospho-signature) to be used for the prediction of future samples, the optimal number of features was determined in a CV loop. This is essentially the same as the inner loop in the quality assessment process (see also Supplementary Figure C.4). Therefore, after running the cross validation process five times with different random number generator seeds, we obtained the following results: A 200x5 prediction result matrix (200 being the rows, 5 the columns) containing the number of correct CV predictions for $k = 1 \ldots 200$ selected features (i.e. k best ranking in each CV step) across the 5 random seeds; a 200x5 AUROC matrix containing the corresponding area under the ROC curve values; and a 25,020x19x5 rank matrix holding the rank of each feature in each CV step across the 5 random seed runs (features that were not subjected to imputation/feature selection due to too many missing values received the rank maxRank+1, where maxRank is the number of features that were subjected to

imputation/feature selection).

The primary criterion for selecting the best subset of features was the number of correct predictions. For this purpose the values in the prediction matrix and AUROC matrix were row-averaged, leading to a vector of 200 average correct predictions and area under the curve values. Within this vector the indices (numbers of features) that lead to the best number of correct predictions were determined. Among those the one index that had the highest AUROC value was selected as best performing feature number, which was twelve. Next, the final feature rank was determined by averaging first over the third and subsequently over the second dimension of the rank matrix. The resulting vector of length 25,020 containing the average rank of each feature was sorted in ascending order and the 12 top-ranked were selected. These were the phosphosites described in Table 4.2.

The twelve selected final features were then used to train the final predictor. However, since these features also contained missing values, imputation had to be performed first. The original sampling should reflect the variance within each feature and class, which is crucial for the quality of a feature. Since the best features had already been selected at this stage, sampling can influence the feature weights in the final predictor only. We used the mean of each feature and class for replacing missing values in the dataset for the final predictor. Alternatively, we could use the same sampling approach as above, and then aggregate the resulting predictors by, for example, averaging the classification score. The differences in these two alternatives are only marginal (Supplementary Figure C.9). Finally, a SVM based on the predictive 12-site phospho-signature (again with linear kernel and C=1) was trained and can now be applied to the classification of new samples.

4.2.10 Quantitative Western-Blot Analysis

For protein detection in human lung cancer cell lines, exponentially growing cells from 15 cm dishes were used. After cell lysis 80 µg of total protein was separated on 4−12% Bis-Tris NuPAGE gels (Invitrogen) for the detection of integrin β4 or on 7.5% Tris-Glycine gels (Biorad Mini PROTEAN) for the detection of tankyrase 1-binding protein (TNKS1BP1). Proteins were transferred overnight to 0.2 µm nitrocellulose membranes and probed with the appropriate antibodies in LI-COR Odyssey blocking buffer. All primary antibod-

ies were used in 1:1000 dilutions: anti-integrin β4 antibody [M126] (ab29042, Abcam); anti-TNKS1BP1 (SAB4503414; Sigma Aldrich); anti-actin (I-19) (sc-1616-R, Santa Cruz Biotech). Actin served as a loading control. Following primary antibody incubation, membranes were probed with IRDye 800CW conjugated goat anti-mouse IgG (H+L9 (LI-COR #926-32210), dilution 1:15000 for the detection of integrin β4; or IRDye 800 conjugated affinity purified anti-rabbit IgG, (611-732-127; Rockland), dilution 1:20000, for the detection of TNKS1BP1 and actin; or DyLight 800 conjugated affinity purified anti-rabbit IgG (H+L) (611-145-122; Rockland), dilution 1:50000 for the detection of actin. Signals were detected at 800 nm using the LI-COR Odyssey infrared system.

4.3 Results and Discussion

4.3.1 Confirmation of dasatinib sensitivity

Based on the half-maximum growth inhibitory concentration (GI_{50}) of dasatinib reported previously [94], 13 sensitive and 13 resistant NSCLC cell lines were pre-selected. For these 26 cell lines we repeated viability assays to verify the reported GI_{50} values. We chose the median GI_{50} as classification threshold, so that depending on the GI_{50} the cell lines were assigned to sensitive (GI_{50} <1 µM) and resistant (GI_{50} >1 µM) classes. For 19 out of 26 cell lines the assignment was consistent. For 7 cell lines the assignment based on the sensitivity determined here differed from that reported previously [94]. By using only the cell lines, for which the sensitivity could be reproduced in two different labs, we maximize the reproducibility of the cell line assignment and therewith the robustness of the predictive signature. The other cell lines were therefore excluded from the training set (see Supplementary Table C.1 for GI_{50} values). The remaining 19 cell lines (11 sensitive and 8 resistant) were used to identify a predictive phospho-signature. The peak dasatinib plasma concentration (C_{max}) obtained in a phase II trial in patients with advanced NSCLC is 124 ± 59 ng/mL [92]. The corresponding molarity is below the classification threshold chosen above. However, only the GI_{50} values of two cell lines, HCC4006 and H322M, are marginally higher than the average peak plasma concentration.

4.3.2 Identification of differentially phosphorylated proteins

To quantitatively compare the cell lines to be analyzed, we isotopically labelled sensitive and resistant NSCLC cell lines using stable isotope labelling by amino acid in cell culture (SILAC; [17]). The sensitive cell lines were grown in SILAC media supplemented with the medium forms of arginine and lysine (Arg6/Lys4), whereas the resistant cell lines were grown in heavy media (Arg10/Lys8, see Supplementary Table C.2 for experimental pairing scheme). A Super-SILAC reference [100] was generated by mixing protein lysates of 16 randomly selected cell lines in unlabelled (light, Arg0/Lys0) media. The Super-SILAC reference serves as a spike-in standard, enabling accurate cross-sample comparison (see Supplementary Figure C.2). Equal protein amounts of the Super-SILAC reference, a sensitive, and a resistant cell line were mixed and subsequently subjected to a global, quantitative phosphoproteomics workflow using strong cation exchange chromatography (SCX) and immobilised metal ion affinity chromatography (IMAC) followed by liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis (see Methods for details). In total, 37,747 phosphosites were identified in the 26 profiled cell lines. 88% of all quantified phosphorylation sites had a cell line to Super-SILAC ratio <4-fold, which allowed for accurate quantification of phosphorylation changes between the analyzed cell lines. From the 37,747 identified phosphorylation sites, 25,020 were rated as class-I sites, i.e. sites that could be identified with high localization confidence [63]. Only these sites were used in the following analyses. The frequency distribution of the phosphorylated residues (serine: 83.2%, threonine: 15.3%, tyrosine: 1.5%) is similar to the frequency distribution observed by Olsen et al. [63].

We first tried to identify proteins that are differentially phosphorylated between the sensitive and resistant cell lines. To this end, the Wilcoxon rank sum-test was applied to the set of phosphosites with data values in at least two thirds of the experiments (leading to 4457 valid sites with approximately 11% missing values on average). Indeed, 58 phosphosites were significantly regulated between the group of 11 sensitive and 8 resistant cell lines at a false discovery rate (FDR) of 10% (see Table 4.1). The regulated sites reside on 41 unique proteins. Most of the regulated sites (53 or 91%) are stronger phosphorylated in sensitive cell lines. Only 5 (9%) sites are stronger phosphorylated in resistant cell lines.

Table 4.1: Significantly different phosphorylation sites. Median diff: median difference of log10 ratios between sensitive and resistant classes; q-value: FDR-corrected Wilxocon rank-sum p-value.

Accession	Gene Name	Site	Median diff	q-value
A8K556	GPCR5A	S345	0.872	0.047
Q6ZSZ5	ARHGEF18	S1101	0.419	0.047
Q13177	PAK2	S141	0.315	0.047
Q15149-2	PLEC1	S42	0.334	0.047
Q9C0C2	TNKS1BP1	S429	0.968	0.047
P16144-2	ITGB4	S1424	1.406	0.055
P16144-2	ITGB4	S1387	0.992	0.055
Q6ZSZ5	ARHGEF18	S1103	0.345	0.055
Q3KQU3	MAP7D1	S116	0.514	0.055
Q86SQ0	LL5B	S212	0.721	0.055
Q8IVF2	AHNAK2	S2657	0.909	0.055
Q92614	KIAA0216	S1970	0.657	0.055
Q9Y2U5	MAP3K2	S153	0.494	0.055
P49792	RGP3	T799	-0.261	0.055
B2R5W6	MAPRE3	T164	0.447	0.055
B8QGS6	PKP2	S151	0.742	0.079
B4DIK2	NUP153	S338	-0.317	0.079
Q13177	PAK2	S2	0.234	0.079
O15231-3	ZNF185	S469	1.662	0.080
O43399-2	TPD52L2	S141	0.563	0.080
Q14573	ITPR3	S916	0.782	0.080
Q676U5	APG16L	S269	0.725	0.080
Q86SQ0	LL5B	S513	0.606	0.080
B8QGS6	PKP2	S154	0.688	0.082
P16144-2	ITGB4	S1445	1.473	0.082
P16144-2	ITGB4	S1448	1.544	0.082
P16144-2	ITGB4	S1069	1.236	0.082
A6NDI6	FNBP1L	S490	0.239	0.082

Table 4.1: Significantly different phosphorylation sites (continued).

Accession	Gene Name	Site	Median diff	q-value
A8K1D2	LASP1	S146	0.366	0.082
A8K7M3	SEPT10	S451	1.015	0.082
A9UF02	BCR/ABL	S459	0.270	0.082
B3KSZ4	GATAD2B	S129	-0.181	0.082
D6W4Y8	ASAP2	S701	0.430	0.082
O60303	KIAA0556	S691	0.618	0.082
Q52LW3	ARHGAP29	S1019	1.340	0.082
Q8WUF5	IASPP	S102	0.528	0.082
Q9UQB8-5	BAIAP2	S509	1.197	0.082
P16144-2	ITGB4	T1385	0.937	0.082
B8QGS6	PKP2	S155	0.854	0.083
O15231-3	ZNF185	S466	1.560	0.083
P23528	CFL	S156	0.445	0.083
Q13439	GOLGA4	S78	0.468	0.083
Q8N4C8	MINK	S699	0.486	0.083
Q14573	ITPR3	S934	0.788	0.086
Q9BY89	KIAA1671	S1800	0.422	0.086
B8QGS6	PKP2	S251	0.655	0.096
B2RBM8	ADNP	S769	0.793	0.096
Q8NEY8	HSPC206	S133	-0.213	0.096
D3DXE9	BAZ1B	S1468	-0.217	0.096
P28066	PSMA5	S16	0.430	0.096
Q53EP0	FAD104	S208	0.391	0.096
Q6ZRV2	FAM83H	S870	1.049	0.096
Q6ZRV2	FAM83H	S936	1.026	0.096
Q6ZRV2	FAM83H	S785	0.795	0.096
Q86SQ0	LL5B	S415	0.631	0.096
Q86YV5	SGK223	S696	0.716	0.096
Q8TDM6	DLG5	S264	0.518	0.096
Q3KQU3	MAP7D1	T118	0.396	0.096

For three known dasatinib targets, Bcr-Abl, EphA2, and Lyn [11, 12], we could detect phosphosites that were quantified in at least two thirds of the experiments. The phosphorylations of EphA2 and Lyn cannot differentiate between the sensitive and resistant groups (see Supplementary Figure C.6). Only the site S459 on the breakpoint cluster region protein (Bcr) is differentially phosphorylated (see Table 4.1 and Supplementary Figure C.6).

We next investigated whether any KEGG pathway or Gene Ontology term is enriched in the set of proteins with differential phosphosites. The list of proteins ordered by the Wilcoxon rank sum-test statistic of their most significant phosphosite were analysed with FatiScan [97]. Only the KEGG pathway 'Regulation of actin cytoskeleton' (hsa04810) is significantly enriched at an FDR of 5%. Many of the significantly regulated phosphosites are located on proteins involved in this pathway. A similar analysis revealed that 40 terms of the biological process and the molecular function gene ontologies are significantly enriched (see Supplementary Table C.6). Many of them relate to very generic and not surprising terms, like 'kinase activity' (GO:0016301) or 'signal transduction' (GO:0007165). However, a few of them are more specific, like 'Ras protein signal transduction' (GO:0007265) and 'Rho protein signal transduction' (GO:0007266) in the biological process ontology, and 'cytoskeletal protein binding' (GO:0008092) and 'actin binding' (GO:0003779) in the molecular function ontology.

As a next step, we applied the SubExtractor algorithm [40] to the phosphoproteomic data. SubExtractor detects significantly regulated sub-networks in the STRING protein-protein interaction network [50]. The tool combines local as well as topological information, i.e. information about the regulation of a certain node (represented by the protein's strongest regulated phosphorylation site) and information about the connectivity with its neighbours. The largest sub-network that has been identified by SubExtractor (Figure 4.1) clustered around the EGF receptor, with most of the proteins again being stronger phosphorylated in the sensitive cells. The largest subnetwork comprises many proteins involved in cell-adhesion and actin cytoskeleton organization, such as ajuba (JUB), catenin α1 (CTNNA1) and δ1 (CTNND1), ephrin type-A receptor 2 (EPHA2), brain-specific angiogenesis inhibitor 1-associated protein 2 (BAIAP2), integrin β4 (ITGB4), and plectin (PLEC1).

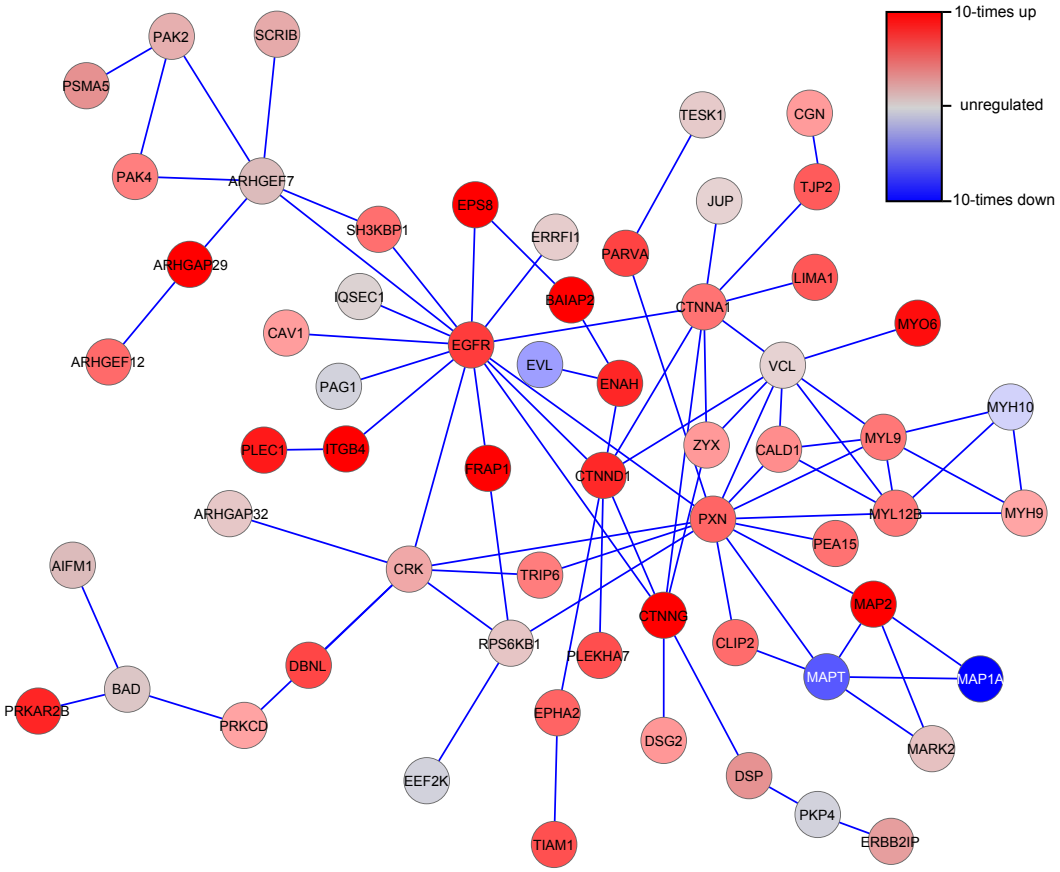

Figure 4.1: Protein-protein interaction subnetwork showing differential phosphorylation in sensitive and resistant cells. The subnetworks were identified using the SubExtractor algorithm. Only the largest network is shown. Red (blue) nodes are stronger (weaker) phosphorylated in sensitive than in resistant cells.

4.3.3 Identification of a predictive phospho-signature

Following the general workflow for detecting phospho-signatures (Figure 4.2), a predictive phospho-signature was identified and its accuracy was estimated by cross validation (CV) based on the cell line dataset (19 valid cell lines). Feature selection was applied within each CV loop to reduce dimensionality of the data and thus avoid overfitting the resulting predictor. We used a Wilcoxon rank-sum test combined with the ensemble method [98] for selecting the phosphosites used for the signatures. The number of phosphosites is optimized in an inner leave-one-out cross-validation loop. The phosphosites were used to train

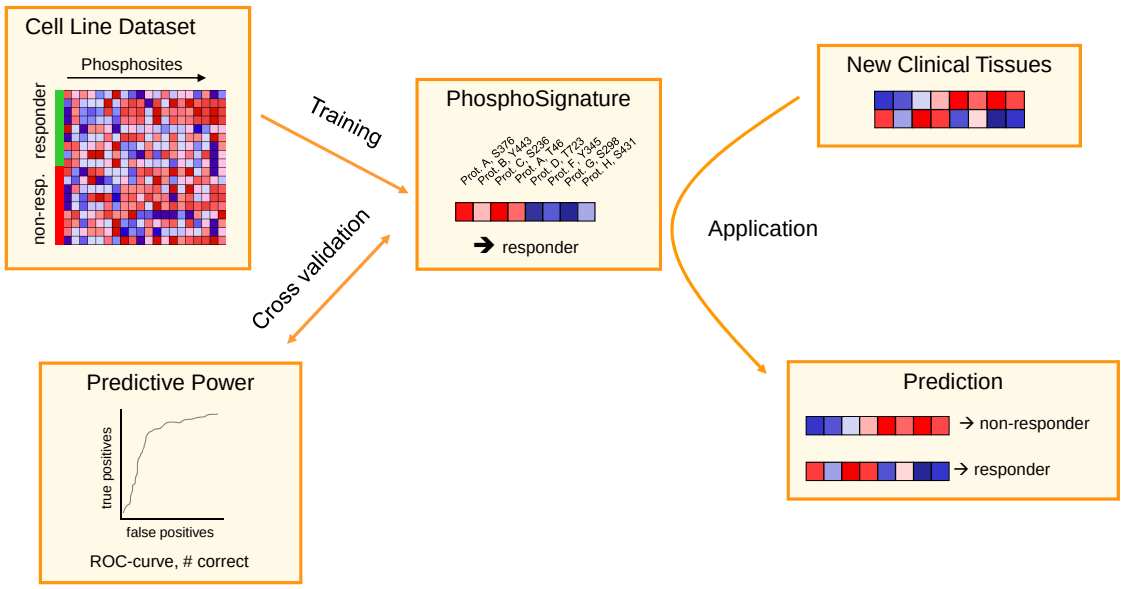

Figure 4.2: The general workflow of phospho-biomarker classification. First, a predictive phospho-signature is identified based on phospho-profiles of sensitive and resistant cell lines using the cross validation approach (described in detail in the text). Once this signature has been identified, it can be applied to new samples to predict the response of the donor to the respective drug.

a support-vector machine (SVM) with linear kernel, which was chosen as the predictor, since it offers state-of-the-art prediction quality and has been successfully applied several times to biological data (e.g. see [29, 101, 102]). SVMs separate two classes by a hyper plane, such that the margin between the classes becomes as wide as possible (e.g. [99]).

The final phospho-signature comprises twelve phosphosites (Table 4.2) located on nine different proteins. The phosphorylation degrees of the twelve identified sites strongly separate the class of sensitive and resistant cell lines (Figure 4.3). All of them are stronger phosphorylated in the sensitive cell lines. The five highest ranked phosphosites show approximately 10-fold differences in their medians. The differences between the 25th and 75th percentiles are still approximately 5-fold. Interestingly four of the highest ranked phosphosites are located on the same protein, integrin β4 (ITGB4 or CD104). The second highest ranked phosphosite is located on the brain-specific angiogenesis inhibitor 1-associated protein 2 (BAIAP2). Further we identified phosphosites that are located on the G-protein coupled receptor family C group 5 member A (GPCRC5A), the inosi-

Table 4.2: Phosphorylation sites of the final phospho-signature. Avg rank: the average rank of the feature across all cross validation steps; Median diff: median difference of log10 ratios between sensitive and resistant classes; #Rank $\leq$ 12: the number of times the feature was among the 12 best across all CV steps; SV weight: the importance of the feature in the SVM predictor (the larger the absolute weight, the more important).

Accession	Gene name	Site	Avg rank	Median diff	#Rank $\leq$12	SV weight
P16144-2	ITGB4	S1448	2.716	1.544	18	-0.386
Q9UQB8-5	BAIAP2	S509	3.611	1.197	18	-0.311
P16144-2	ITGB4	S1387	4.337	0.992	19	-0.155
P16144-2	ITGB4	T1385	5.716	0.937	18	-0.275
P16144-2	ITGB4	S1069	7.937	1.236	13	-0.076
A8K556	GPCR5A	S345	9.632	0.872	16	-0.174
Q14573	ITPR3	S916	14.168	0.782	8	-0.205
Q9C0C2	TNKS1BP1	S429	15.032	0.968	1	-0.159
Q6ZSZ5	ARHGEF18	S1101	16.874	0.419	0	-0.188
Q8WUF5	IASPP	S102	17.516	0.528	7	-0.145
Q676U5	APG16L	S269	18.190	0.725	13	-0.240
O43399-2	TPD52L2	S141	18.274	0.563	8	-0.155

tol 1,4,5-triphosphate receptor type 3 (ITPR3), the 192kDa tankyrase-1-binding protein (TNKS1BP1), the Rho guanine nucleotide exchange factor 18 (ARHGEF8), the RelA-associated inhibitor (IASPP), the autophagy-related protein 16-1 (APG16L), and the tumor protein D54 (TPD52L2).

4.3.4 Sensitivity and specificity of the phospho-signature

To determine the prediction performance, leave-one-out cross validation (LOOCV) was applied. It has been shown that CV, including LOOCV, estimates the true prediction performance accurately and shows a low bias [103]. Since not all phosphosites discriminate well between sensitive and resistant cell lines, feature selection is applied in each CV step, which selects a defined subset of predictive phosphosites. First the features are ranked

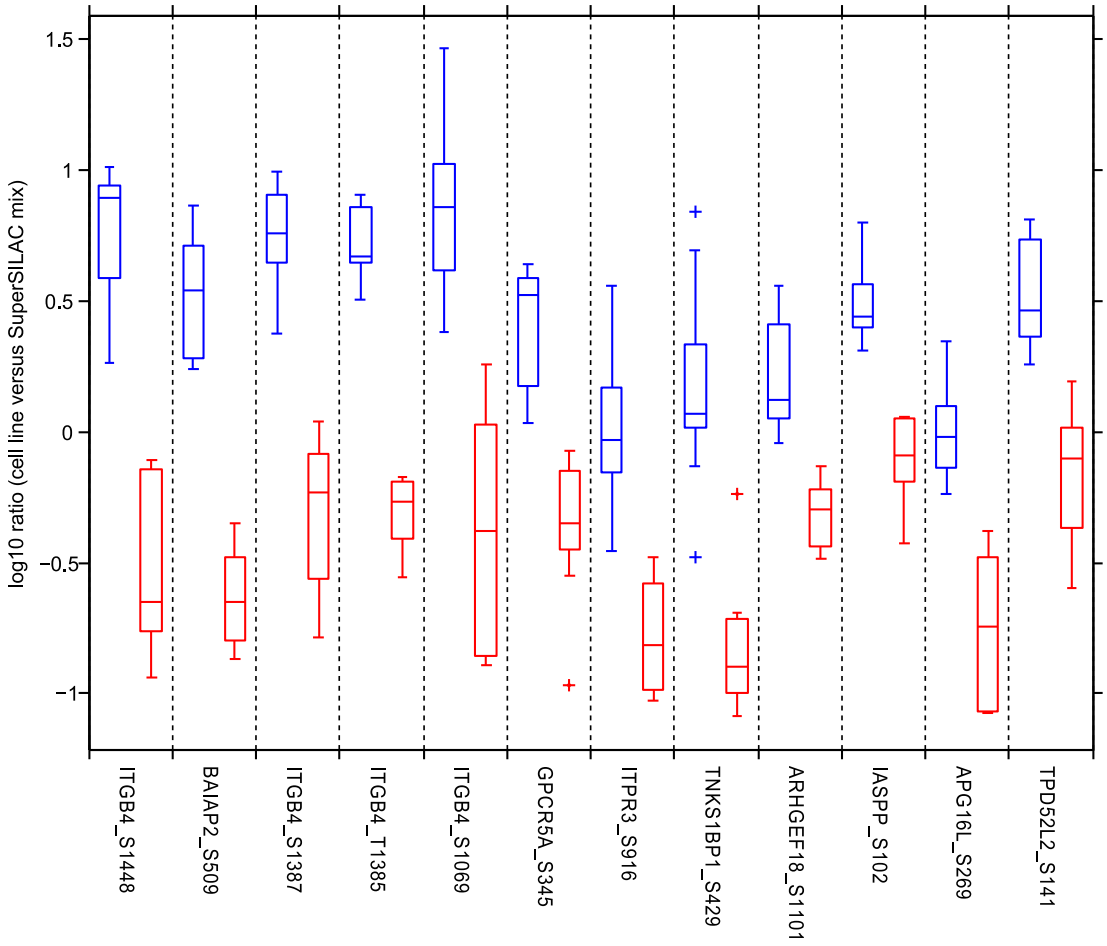

Figure 4.3: Final phospho-signature consisting of 12 phosphosites. Each pair of boxes corresponds to one phosphosite. The blue (red) box represents the sensitive (resistant) cell lines. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are marked individually with crosses.

according to their discriminative power, and then the optimal number of top-ranking features is determined by an inner parameter optimization cross validation. In this inner CV procedure, different numbers of top-ranking features ($k = 1 \ldots 200$) are used, and their respective performance is assessed. The smallest number of features leading to the best prediction quality in the inner CV loop is then applied to the feature selection in the outer cross validation loop (see also Supplementary Figure C.3). Subsequently, a SVM predictor is trained on the reduced training data (reduced in the sense of containing only

features that passed the feature selection criteria) and tested with the reduced test data. It is important to note that the test sample is used neither for optimising the number of features nor for selecting the features within cross validation. Furthermore, the pre-processing steps and classification workflow were fixed before acquiring the NSCLC data. Otherwise, the prediction accuracy would be overestimated.

Missing data are a common phenomenon in shotgun proteomics. Although the quantitative information (i.e. SILAC peaks) of a peptide may be present in the MS spectrum, at least one of the SILAC peaks has to be selected for fragmentation. In this case, the resulting fragment spectrum is used to identify the corresponding peptide. Since the selection of peptides for fragmentation is data-dependent, a certain peptide may be selected in some MS runs but not in others. Therefore, a missing value does not necessarily mean that the corresponding phospho-peptide was not present. This is particularly true when applying the Super-SILAC approach like in this study.

Since many machine learning techniques (SVMs among them) cannot handle missing values, they were replaced by estimated values that were randomly sampled from the respective empirical distribution. As a consequence, the entire assessment was carried out five times with different seeds for the random number generator used for imputation, leading to five distinct prediction results. The five results were strikingly similar, as can be expected from a robust set of features, i.e. four times only one cell line was misclassified (HCC78), and once two were falsely classified (HCC78 and HCC827), which leads to a prediction accuracy of 94% and an area under the receiver operating characteristic curve (AUROC) of 0.92 (Figure 4.4A). Each circle in Figure 4.4A shows the averaged predicted outcome of this cell line when all other cell lines were used as training data. A sensitive cell line is predicted correctly if the SVM predictor assigns a negative value; vice versa for a resistant cell line. The larger the distance to the separating hyperplane (i.e. the distance from 0 in the plot), the more confident the prediction is. It can be clearly seen that 18 of 19 cell lines were predicted correctly by cross validation.

For the final predictor, the workflow was carried out with only one CV loop, corresponding to the inner loop during the prediction quality assessment (see Supplementary Figure C.4). This resulted in identifying a predictive phospho-signature containing 12 phosphosites. Interestingly, the average number of selected features within the

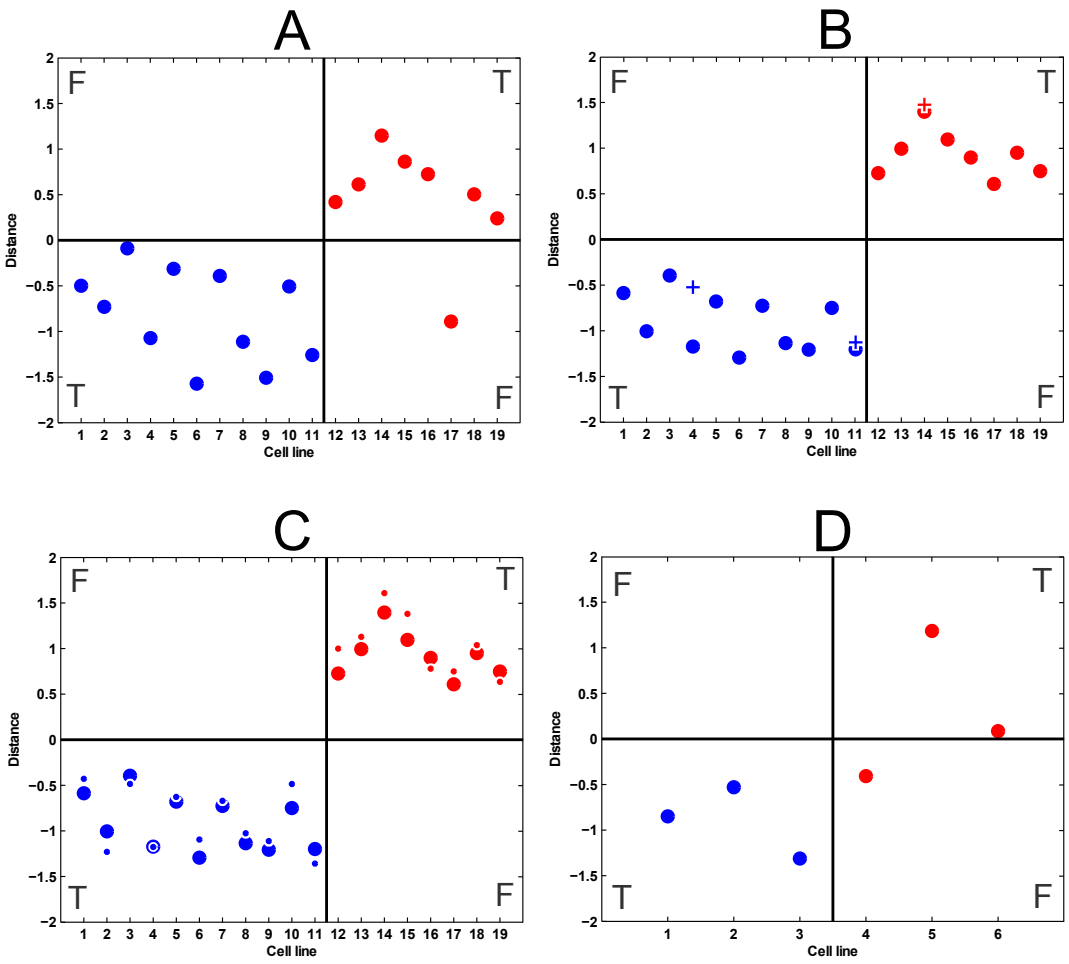

Figure 4.4: Classification results represented by distance to the respective SVM's separating hyperplane. The cell lines in A, B and C are: 1 LouNH91, 2 H1648, 3 HCC827, 4 H322M, 5 H2030, 6 HCC2279, 7 HCC366, 8 HCC4006, 9 H1666, 10 PC9, 11 H2009, 12 H460, 13 Calu6, 14 H2077, 15 H1395, 16 H2172, 17 HCC78, 18 H157, 19 H520; in D: 1 BT-20, 2 MDA-MB-231, 3 HCC1937, 4 MDA-MB-468, 5 BT-549, 6 MCF7. Sensitive cell lines (blue) are predicted correctly if they get a negative value; resistant ones (red) if they are positive. (A) The results of the prediction quality assessment. (B) Prediction results of the final predictor when applied to the same date as used for training (circles) along with the results for the label switch experiments (crosses). (C) Prediction results of the final predictor when applied to the same data as used for training (circles), along with the results for the same data when normalized by the selected set of ribosomal proteins (dots). (D) Prediction results of the final predictor when applied to the breast cancer samples.

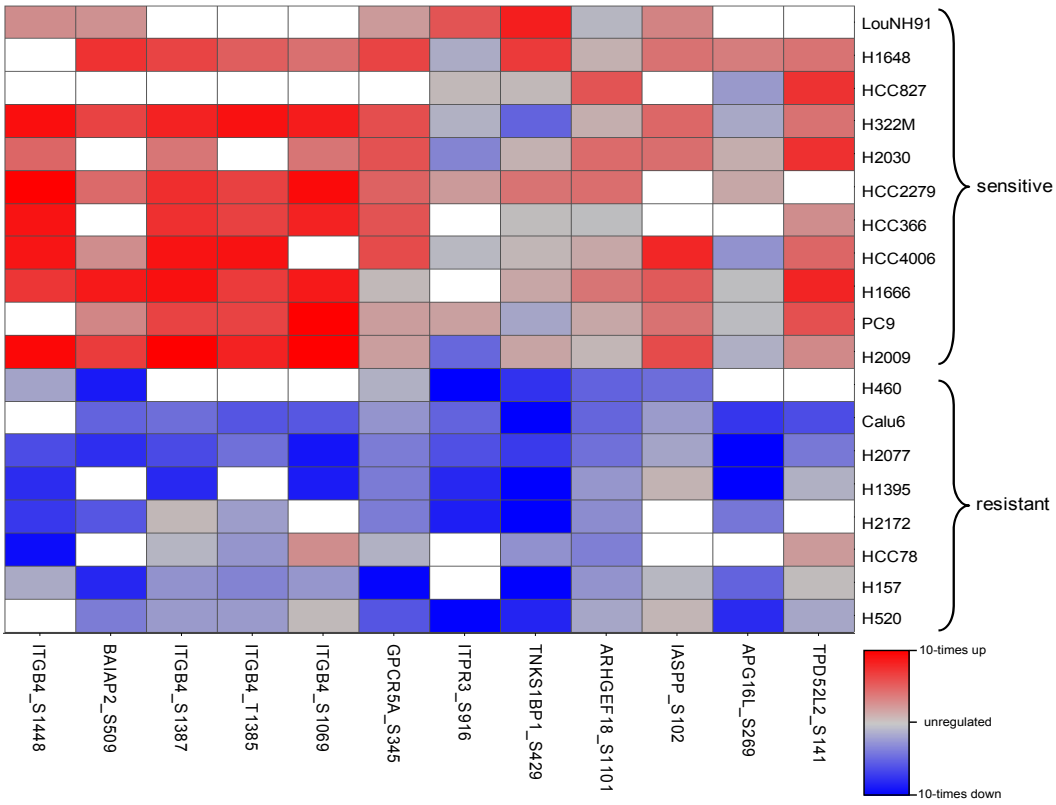

Figure 4.5: Heat map of the final 12 selected phosphorylation sites. Rows are the 19 cell lines that were used to identify the phospho-signature (the upper 11 are sensitive, the lower 8 resistant), columns are the phosphosites ordered by their importance ranks (left is the best). Red indicates up-, blue down-regulation, grey no regulation. Missing values are coloured white.

inner parameter optimization loop during the prediction quality assessment was also approximately 12, which further supported the robustness of the selected set of phosphosites. The sites are listed in Table 4.2) sorted by their global feature ranks, and depicted as a heat map in Figure 4.5 (see also Supplementary Table C.3 for more details and Supplementary Table C.4 for observed ratios). With an increasing number of features the prediction accuracy also increased, until it saturated at 12 features (see Supplementary Figure C.5). Additional features did not improve the prediction accuracy.

These results show that a predictive phospho-signature can be identified from phosphoproteomics data. However the question remains, whether the identified signature is specific to dasatinib or whether it also works for other substances not related to dasa-

tinib. As a first step to answer this question, we applied the prediction quality assessment workflow to randomized class labels. Strikingly, the prediction accuracy was only 51% (AUROC=0.53), which is almost exactly what one would expect if predicting the classes by chance. Thus, a predictive signature cannot be found for arbitrary class associations. As a next step, we investigated whether the classification scores of the final predictor correlate with the cell doubling times of untreated cell lines. The classification score corresponds to the distance from the SVM classification hyperplane and can be interpreted as the confidence in correct classification. In particular, the score is negative (positive) if the sample is predicted as being sensitive (resistant). The cell doubling times range from 25 to 55 hours (confer Supplementary Table C.1). A Pearson correlation coefficient of -0.08 (p-value 0.79) indicates that the doubling times are not associated with the classification. In contrast, the correlation between classification scores and GI_{50} values of dasatinib is significant (0.81, $p = 2.6\text{E-}6$). Finally, we sought to show whether the dasatinib signature is predictive for other substances. The small molecule sorafenib (Nexavar®, Bayer) is a multi-kinase inhibitor targeting the Raf/Mek/Erk and the VEGFR pathway. The correlation between the doubling times and GI_{50} values of sorafenib [94] is -0.05 (p-value 0.83). Taking these results together, we could demonstrate that the identified phospho-signature is specific for predicting response to treatment with dasatinib.

4.3.5 Robustness of the phospho-signature

A good feature and consequently a good set of features should be robust to small variations in the data. Only when slight changes in the data composition still lead to correct predictions, is the biomarker reliably applicable to samples not used for training. Therefore, robustness already plays a crucial role in the process of feature selection. First, a robust feature is chosen frequently by the feature selection method across all cross validation steps. Second, within each cross validation step, slight variations in the training data should also result in the constant selection of robust features.

To identify such robust phosphosites, we applied the Wilcoxon rank-sum test in combination with the ensemble feature selection method [98] to get a feature ranking in each CV step. The average of these ranks across all CV iterations for the signature's 12 fea-

tures along with the number of times each of them was ranked under the first 12 positions are listed in Table 4.2). The best features turned out to be very stable, e.g. the top four have an average rank smaller than 6 and were among the 12 best more than 90% of all iterations. The importance of these features is also indicated by their high weight in the SVM. Overall, 7 features are among the 12 best in more than two thirds of the iterations, and only 2 in less than one third.

To ensure that the SILAC labelling procedure of cell lines has no effect on the results, label switch experiments were performed, where originally medium-labelled cell lines were now labelled with heavy amino acids and vice versa. The classification results of the final predictor applied to these experiments are depicted in Figure 4.4B. For two of the three label switched samples, the prediction is virtually identical to the original data (circles and crosses on position 11 and 14; Figure 4.4B). In the case of the position 4 (H322M), the difference is somewhat larger, but the corresponding label switch experiment still classifies it correctly.

Since phosphosites in this study are detected in a global and unbiased way, we applied global normalization strategy during the discovery phase. However, when the phosphosignature is applied in the clinic, a method that specifically measures the phosphosites of the signature in a robust and cheap way is more likely to be used (see Appendix C for how SVM predictor can be adapted to use data from other methods). Such targeted methods could be either based on phospho-specific antibodies (e.g. immunohistochemistry or ELISA based assays) or targeted mass spectrometry methods such as multiple reaction monitoring (MRM [104, 105]). Since a global normalisation strategy is not applicable to targeted methods, it is necessary to develop an alternative. We focused on non-phosphorylated peptides that showed a very low variance across all cell lines' regulation data regardless of whether the cell line was sensitive or resistant. Although the phosphoproteomic workflow is designed to specifically enrich for phosphorylated peptides, a significant fraction of non-phosphorylated peptides is still present. In this study, a normalization factor based on a set of non-phosphorylated ribosomal proteins exhibiting low variance across all cell lines proved useful (see Supplementary Table C.5 for normalization data). The classification results of the ribosomal protein normalized data are depicted in Figure 4.4C, which shows that the prediction quality is essentially as good as for the

globally normalized data the predictor was trained on.

4.3.6 Signature validation in breast cancer cells

To test whether the phospho-signature is also applicable to other cancer types, we selected 3 sensitive and 3 resistant breast cancer cell lines. Again, GI_{50} values were also determined in-house and compared to the previously reported values [86]. This time, all data were consistent (confer Supplementary Table C.1) and the 6 breast cancer cell lines were subjected to our global phosphoproteomics workflow (see Supplementary Table C.4 for data).

Subsequently, the cell lines were classified with the SVM predictor trained on the set of NCSLC cell lines. Strikingly, 5 of the 6 breast cancer cell lines could be classified correctly (Figure 4.4D); only one resistant sample was wrongly predicted to be sensitive (MDA-MB-468). These findings indicate that the proposed phospho-signature is also predictive for dasatinib sensitivity in other cancer types.

4.3.7 Integrin β4 expression as a surrogate marker

Four of the highest ranked predictive phosphosites reside on the protein Integrin β4 (ITGB4, see Table 4.2)). Since we did enrich for phosphorylated peptides and did not measure the abundance of the non-phosphorylated peptides or the total protein, it is principally impossible to distinguish between differences in the phosphorylation degree and differences in the expression of the corresponding protein. However, in case of ITGB4 it is likely that the differences in the phosphorylation of the four sites are caused by differences in the abundance of the protein itself. To prove that the expression of this protein is indeed different in the two classes of the NSCLC cell lines, we performed quantitative western blots using antibodies against the total protein of ITGB4 and 182 kDa tankyrase-1-binding protein (TNKS1BP1). We selected TNKS1BP1 as one of the eight proteins, for which only one phosphosite was identified as predictive feature. Whereas TNKS1BP1 is present in almost all cell lines and its expression shows no correlation with the sensitivity of the cell line to dasatinib, ITGB4 can be detected in 8 sensitive cell lines, but only in 2 resistant cell lines (see Figure 4.6A). This is confirmed by quantitative analysis of

three replicate experiments (see Figure 4.6B and 4.6C). The background-corrected signals of ITGB4 correlate with the phosphorylation degree measured by mass spectrometry (Pearson correlation 0.88, $p = 2\text{E-}6$). The signals of most resistant cell lines are low, while strong signals can be determined in the sensitive cell lines. This clearly shows that expression of ITGB4 is also predictive and that it can be used as surrogate marker instead of its phosphorylation. Indeed, if choosing the average of the median signals in each group as classification threshold, all resistant and 8 sensitive cell lines would be correctly classified, whereas 3 sensitive cell lines would be falsely classified as resistant. Nevertheless, the prediction accuracy of ITGB4 expression (84%) is not as high as the accuracy of the full phospho-signature (94%). In contrast, the signals for total TNKS1BP1 do not correlate with sensitivity, although its phosphorylation is predictive.

4.3.8 Expression of integrin β4 in lung and breast cancer tissues

We demonstrated that the signature consisting of 12 phosphorylation sites and the expression of ITGB4 is predictive in NSCLC and breast cancer cell lines. To explore, whether ITGB4 is also expressed in cancer tissues, we examined immunohistochemistry images of several cancer tissue slices. The Human Protein Atlas [106] systematically analyses the human proteome in cell lines, normal tissues and cancer tissues using antibodies. In particular, it contains a number of immunohistochemistry images of cancer tissues stained with an antibody (CAB005258) against total protein of ITGB4. Five lung cancer samples (42%) are negative, whereas seven samples show weak to strong expression of ITGB4 (see Supplementary Figure C.7). Similarly, six breast cancer samples (50%) are negative, whereas six samples show weak expression (see Supplementary Figure C.8). In summary, we could show that the expression of ITGB4 can be used as surrogate marker for its phosphorylation. The marker is measurable by immunohistochemistry in clinical tissue samples and it is present in a sub-population of approximately 50% of the investigated cancer tissues.

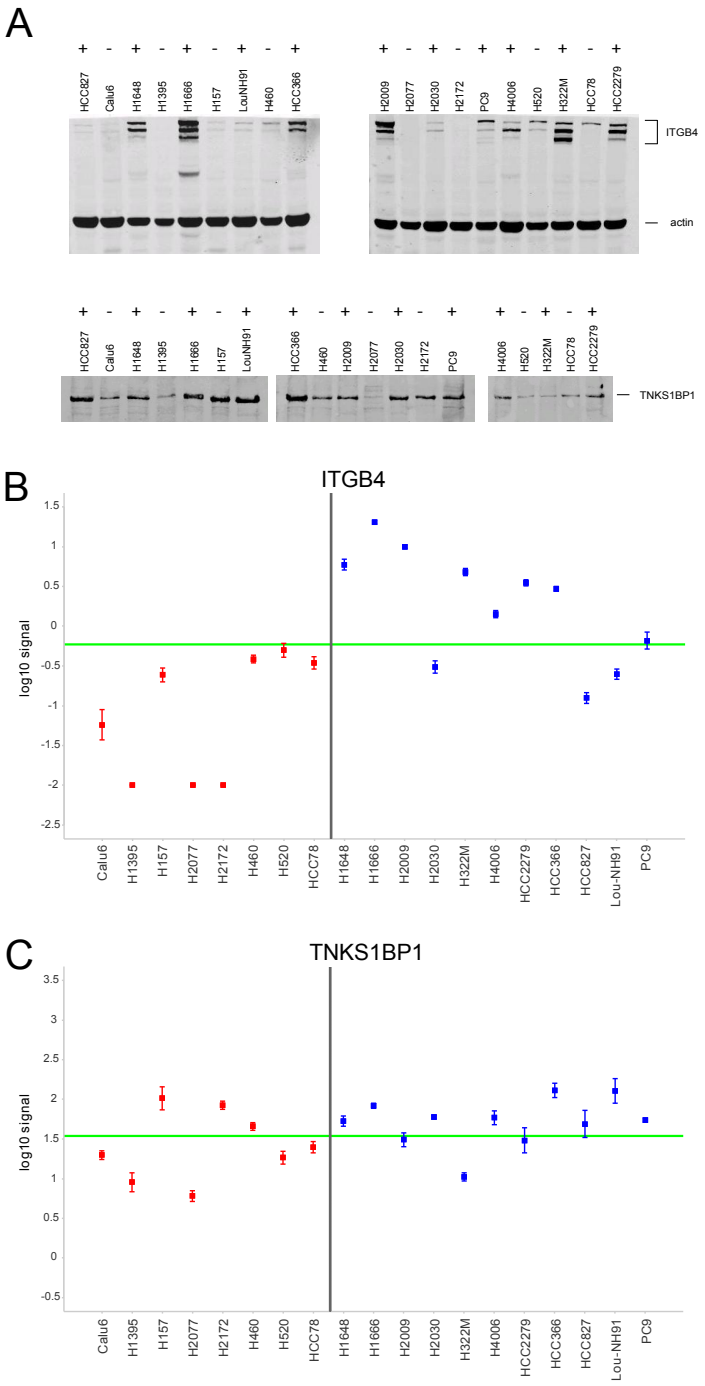

Figure 4.6: Western blots of ITGB 4 and TNKS1BP1 in NSCLC cell lines. A: Western blot images for one replicate. The sensitivity to dasatinib treatment is indicated by +/-. B: Quantitative readout for ITGB4 in resistant (red) and sensitive (blue) cell lines. The error bars represent the standard error across three replicates; the green line is the average of the class medians. C: Quantitative readout for TNKS1BP1.

4.4 Conclusion

This study shows that the identification of response prediction markers from global and unbiased quantitative phosphoproteomics experiments in a preclinical setting is possible. Detection of a few ten thousands of phosphorylation sites across a panel of cancer cell lines is feasible. The use of a pool of cell lines as a common reference enabled the accurate quantification of the detected sites. The accuracy and reproducibility of the phosphoproteomic workflow was demonstrated in label switch experiments. Measuring protein phosphorylation levels allowed us to monitor over-activation and repression of disease-specific signalling pathways. Since kinase inhibitors, such as small molecules and monoclonal antibodies interfere with signal transduction pathways, we hypothesised that determining the basal activity of these pathways will allow predicting a response to therapy with such an inhibitor.

We identified 58 phosphosites that are differentially abundant between sensitive and resistant cell lines. Enrichment analysis of gene ontology terms and KEGG pathways as well as subnetwork analysis show that many of the differentially phosphorylated proteins are involved in cell adhesion and cytoskeleton organization, where most phosphorylations are higher in the sensitive group. Interestingly, it has been shown that dasatinib inhibits migration and invasion of various solid tumors through inhibition of the Src-kinase [107–109], which is one of the main targets of dasatinib [11, 12]. We thus hypothesize that cells, in which pathways related to cell adhesion and cytoskeleton organization are over-activated, respond to a treatment with dasatinib. Src is a non-receptor tyrosine-protein kinase. That none of the differentially phosphorylated residues is a tyrosine, does not contradict the hypothesis, since we studied the basal phosphoproteome of untreated cells. Proteins that are causal for resistance to Src-inhibition may be located down- or up-stream of the direct Src-kinase substrates in the signalling cascades.

We showed that a phospho-signature consisting of only 12 phosphorylation sites it sufficient to predict the response from the basal phosphoproteome of a cultured cell. The predictor model was based on a support vector machine with linear kernel. We validated the accuracy of the prediction in a leave-one-out cross-validation procedure. 18 out of 19 cell lines could be classified correctly. The obtained prediction accuracy was 94%, the area

under the curve was 92%.

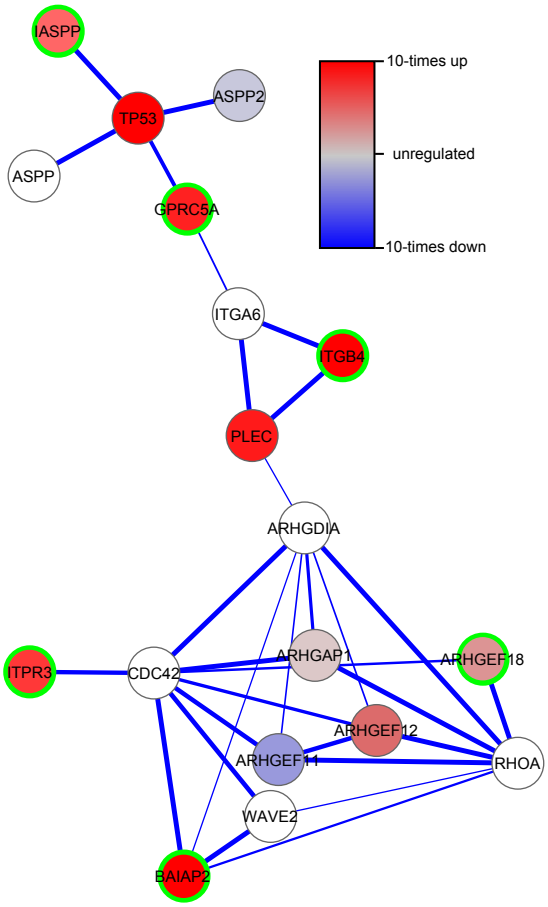

Figure 4.7: Protein-protein interaction network that shows the relationship between six of nine predictive signature proteins (marked with green border). The network was obtained using STRING.

The 12 phosphorylation sites were located on 9 different proteins (see Table 4.2) and Figure 4.7). Four of the phosphorylation sites are located on Integrin β4 (ITGB4 or CD104). In general, integrins mediate cell-matrix or cell-cell adhesion and are involved in transducing signals to regulate transcription and cell growth. The subunit β4 associates with α6 and the resulting integrin α6β4 is a receptor for the laminin family of extracellular matrix proteins. Integrin β4 is linked to various signalling pathways such as the MAPK, PI3K-Akt, and Src-Fak pathways [110–112]. Furthermore, expression of α6β4 is associated with poor patient prognosis in various cancers [113–115]. According to the PhosphoSite database [116] the sites S1457 and S1518 were detected in previous mass spectrometry

based proteomics experiments, but to our knowledge the functions for none of the four sites have been described so far. All four sites are stronger phosphorylated in sensitive cells than in resistant cells.

Besides the integrin β4 phosphorylations, the signature comprised eight additional phospho sites on eight other proteins. Like integrin β4, the brain-specific angiogenesis inhibitor 1-associated protein 2 (BAIAP2) and the Rho guanine nucleotide exchange factor 18 (ARHGEF18) are involved in regulating the actin cytoskeleton. BAIAP2 (also called insulin receptor substrate p53, IRSp53) serves as an adaptor link-ing a Ras-related protein Rac1 with a Wiskott-Aldrich syndrome protein family member 2 (WAVE2). The recruitment of WAVE2 induces Cdc42 and the formation of filopodia [117, 118]. ARGHEF18 acts a guanine nucleotide exchange factor for the GTPases RhoA and Rac1 [119, 120]. Activation of RhoA induces actin stress fibres and cell rounding.

The RelA-associated inhibitor (PPP1R13L, also called inhibitor of ASPP protein, IASPP) and the G-protein coupled receptor family C group 5 member A (GPRC5A, also called retinoic acid-induced protein 3, RAI3) are functionally connected to the tumor suppressor p53. PPP1R13L binds to p53 and inhibits its activation by ASPP1 and ASPP2 [121]. On the other hand, p53 was demonstrated to bind to the promoter of GPRC5A and thereby negatively regulates its expression [122].

The tumor suppressor p53 is associated with at least two signature proteins. At the same time, p53 is inactivated by mutations in a large proportion of tumor cell lines. We therefore investigated whether the p53-status alone is predictive of a response to dasatinib. According to the IARC TP53 database [123], 6 out of 7 sensitive and 3 out of 5 non-sensitive cell lines have a mutation in the p53 protein (7 cell lines were not listed, see also Supplementary Table C.1). Since the functional effect is not known for all mutations, we assumed that any mutation, apart from neutral or silent mutations, is functionally relevant. The null-hypothesis that sensitivity to treatment with dasatinib does not differ between p53-mutated and p53-wildtype cell lines cannot be rejected (Fisher's exact test p-value is 0.52). Therefore, the mutation status of p53 is not a good predictor of dasatinib sensitivity.

Although, based on the current literature, a direct link cannot be made between the other four proteins inositol 1,4,5-triphosphate receptor type 3 (ITPR3), 182 kDa

tankyrase-1-binding protein (TNKS1BP1), autophagy-related protein 16-1 (APG16L), tumor protein D54 (TPD52L2), and the main dasatinib targets, the fact that their phosphorylation correlates with the treatment response supports their use in the predictive model.

From the discussion above it is clear, that many of the signature proteins are related to each other. Indeed, when mapping the nine proteins to the STRING protein-protein interaction network [50], we revealed one network involving six signature and few additional proteins (Figure 4.7). Phosphorylation sites for most of the proteins in this network are less abundant in the resistant cell lines than in the sensitive cell lines.

The difference in phosphorylation of a specific site between two cell lines may be due to a difference in either expression of the corresponding protein, or the degree of phosphorylation of this site, or a combination of both. Since we did not investigate the protein expression, we cannot distinguish between the three possibilities. However, as long as the abundance of a certain phosphorylated peptide consistently differs between sensitive and resistant cell lines, the cause for its difference is not important for its use as a predictive biomarker. In case of ITGB4, we could indeed show that its protein expression is also predictive. Contrary, the protein expression of TNKS1BP1 does not differentiate between sensitive and resistant cell lines. The study also showed that the predictor identified from a panel of NSCLC-cell lines can be used in other cancer cell lines. 5 out of 6 breast cancer cell lines were correctly predicted (prediction accuracy 83%). Only one resistant cell line (MDA-MB-468) was predicted to be sensitive.

A few markers for dasatinib have been suggested in the literature or are already applied in the clinic. For example, Huang et al. [86] identified a predictive six-gene model from gene expression profiles. Obviously, the phosphorylation grade may be largely independent of the mRNA expression level. Nevertheless, we investigated whether the phosphorylation sites on the corresponding proteins are also predictive. We detected phosphorylation sites on five of the six proteins: EPHA2, CAV1, CAV2, ANXA1, and PTRF. Although, the phosphorylation tends to be high in sensitive cell lines and low in resistant cell lines, the relationship is not as sound as for the markers identified in this study. All sites are not significantly different between the two classes. As an example, Supplementary Figure C.6 shows three sites on the Ephrin type-A receptor (EPHA2). Additionally the

tyrosin phosphorylations p-Src(Y418), p-BCR-ABL(Y412), p-Crkl(Y207), p-Pax(Y31), p-Fak(Y576) have been described as pharmacodynamic markers for dasatinib in mouse experiments and in clinical trials [124–126]. These markers are modulated after treatment with dasatinib and their basal levels do not necessarily differentiate between sensitive and resistant subjects. Nevertheless, we were interested in their behaviour across the untreated cell lines. We could detect the phosphorylation site Y418 of Src in five cell lines, but could not identify any relationship to the sensitivity of these cell lines. The site ABL(Y412) on the fusion protein BCR-ABL was not detected. However, a different site BCR(S459) was detected in almost all cell lines and is significantly modulated between the sensitive and resistant group (see Table 4.1 and Supplementary Figure C.6).

We demonstrated our method for the identification of a predictive phospho-signature in a set of NSCLC and breast cancer cell lines. The application to cultured cells has a number of advantages: the cell population is very homogenous; sample amounts from cell lines are not limited; experiments are easily reproducible; and the drug's efficacy can be experimentally determined. However, whether the signature or parts of the signature are also predictive in clinical samples has to be shown in future studies with clinical samples. Instead of applying shotgun phosphoproteomics, it is possible to apply targeted detection methods, such as immunological methods, or the mass-spectrometry-based multiple-reaction-monitoring method [127]. These methods allow the quantification of marker phosphosites of relatively low sample amounts and can be applied to large number of samples. Since fresh-frozen tissues are rare, the translation of our results to the clinic requires the analysis of formalin-fixed and paraffin-embedded (FFPE) tissues. It has been assumed, that the cross-linking of proteins prevents a proteomic analysis. Recently, it could be shown that proteins can be effectively extracted from FFPE samples and that the proteins and phosphorylations are quantitatively preserved compared to fresh-frozen tissues [128–130].

As an alternative, we demonstrated that the expression of ITGB4 can be used as surrogate marker for its phosphorylation. The marker is measurable by immunohistochemistry in clinical tissue samples and it is present in a sub-population of approximately 50% of the investigated cancer tissues.

In this study, the phosphorylation data were globally normalized, assuming that the

overall phosphoproteome is fairly well conserved between the different cell lines. However, this strategy is no longer applicable to targeted detection of the selected phosphosites, since all measured phosphosites will be regulated. We proposed an alternative normalization strategy using the expression of eight non-regulated ribosomal proteins. It could be demonstrated that the prediction of sensitivity using the phospho-signature is stable for the application of the alternative normalization strategy.

In summary, the identified phospho-signature consisting of twelve phosphorylation sites is highly predictive for the sensitivity to treatment with dasatinib in NSCLC cell lines as well as breast cancer cell lines. The results suggest that the phosphorylations of integrin β4 as well as eight further proteins are candidate biomarkers for predicting response in solid tumors to dasatinib and potentially to other Src family kinase inhibitors. That many of the signature proteins have related function and are connected in a protein-protein interaction network, further supports the generalizability of the predictive signature.

In this study we proposed a general method for identifying response prediction biomarkers based on a phosphorylation signature. The method is hypothesis-free insofar as the investigated phosphorylation sites do not have to be preselected, and no assumptions about the mechanism of action of the therapeutic drug have to be made. The basis of the method is the global quantitative phosphoproteomic analysis of baseline samples. While we demonstrated that the method permits identifying a highly predictive phosphorylation signature for response to dasatinib treatment in NSCLC cell lines, it can be assumed that the method can also be applied to other drugs, particularly other kinase inhibitors, and to other tumor types.

Chapter 5

Pareto biomarker

In this chapter, the biomarker workflow presented in Chapter 4 is extended to optimize not only one single objective (i.e. best possible separation of responder and non-responder), but also the objectives signature size and relevance (i.e. association of signature proteins with dasatinib's main target). This is achieved by employing a multiobjective optimization algorithm based on the principle of Pareto optimality, which allows for an optimization of all three objectives in parallel.

The content of this chapter was submitted for publication as:
M. Klammer, J.N. Dybowski, D. Hoffmann, and C. Schaab. "Pareto Optimization Identifies Diverse Set of Phosphorylation Signatures Predicting Response to Treatment with Dasatinib". In: *PLoS one* 10.6 (2015), e0128542

The author was a key contributor to designing and implementing the algorithm, as well as writing the paper.

5.1 Background

Targeted drugs, such as kinase inhibitors, are extensively studied as promising agents either alone or in combination with other agents for treating cancer. Unfortunately, only subsets of patients usually respond to targeted therapeutic interventions. Tests that can predict whether patients will benefit from these therapies are therefore desired companions of targeted drugs. Many, if not all response prediction tests currently used in clinical practice are based on markers directly linked to the disease-relevant drug target. However,

singleton markers measuring the expression or mutation status of a drug target may often not be sufficient to predict response. For example, it has recently been shown that the success of predicting how melanomas respond to targeted therapies by genotyping alone may be limited [132].

Therefore, several studies have focussed on identifying molecular signatures comprising multiple markers for response prediction. Predominantely, these signatures were identified using transcriptomics data (for example [86, 87]). In recent years, advances in sample processing, mass spectrometry, and computer algorithms for the analyses of proteomics data have enabled the application of mass spectrometry-based proteomics in order to monitor phosphorylation events in a global and unbiased manner [15, 32, 63]. These methods have become sufficiently sensitive and robust to identify and quantify thousands of phosphorylation sites in a single experiment. Multivariate markers based on the phosphorylation status of certain sets of proteins – here referred to as phospho-signature – can predict the clinical response, as they link therapy outcome to the most predictive phosphorylation events in the context of signal transduction therapy. This has been demonstrated in two recent studies, where phosphoproteomics data was used to identify predictive multivariate markers for the multi-kinase inhibitor dasatinib [65] and the FLT3 inhibitor quizartinib [133].

Previous studies have focused on the identification of *one* single multivariate marker signature that was optimized for prediction accuracy. Here, we investigate a method that allows for the incorporation of additional objectives that are optimized simultaneously, and enables the identification of *several* predictive markers. Such objectives can, for instance, be related to the annotations of protein markers (e.g. localization, function), to technical properties (e.g. size of the signature), or to network information (e.g. proximity of markers to drug target). It has been shown recently that the inclusion of annotated biological information, but not the method category (e.g. support vector machine, random forest, etc.) or handling of missing data, significantly improved prediction accuracy in a study analyzing 44 drug sensitivity prediction algorithms [134]. More specifically, it has also been demonstrated that adding network information can improve prediction accuracy or at least improve the robustness of feature selection (e.g. [135]). These methods have in common that the network information is factored in by modifying the objective function (e.g.

network-based support vector machines [136]) or the rank order for filter-based feature selection (e.g. NetRank [137, 138]). However, instead of optimizing a combined objective function, we choose to optimize multiple objectives in parallel using principles of multi-objective or, specifically, Pareto optimization [139]. Multi-objective optimization methods return a set of optima, the so-called Pareto front, instead of a single optimum solution. In case of selection of predictive biomarkers, these solutions differ in their composition of selected features and in the degree to which different objectives are optimized. If necessary to limit the number of marker candidates, the researcher can apply *post hoc* weighting of objectives.

In biomedical research, Pareto optimization has been mainly applied to design of small molecules [140] and peptide sequences [141]. More recently however, it has also been applied to selection of features. For example, Rajapakse and Mundra optimized features for multi-class classification by decomposing the over-all objective to multiple objectives for each pair of classes [142]. Xue et al. complemented the objective of classification accuracy with minimizing the size of the signature [143]. In this study, we generalize the idea of applying Pareto optimization to the problem of selecting predictive marker signatures by optimizing not only the prediction accuracy and the size of the signature, but also the biological relevance of the selected features. Here, the biological relevance is defined by the proximity of features to the respective drug target as derived from protein-protein interaction networks. In principle, all obtained solutions on the Pareto front can be evaluated and tested in validation experiments. However, since in practice the Pareto front consists of several dozens of solutions, we propose to cluster these solutions in feature space and investigate a much smaller number of cluster centroids. We apply the proposed method to the identification of multivariate phosphorylation signatures that predict response to dasatinib in non-small cell lung cancer and breast cancer cell lines using the phosphoproteomic data generated by Klammer et al. [65].

5.2 Methods

5.2.1 Data

The training data comprising the class-I phoshorylation site ratios of 19 NSCLC cell lines relative to a SuperSILAC spike-in were obtained from Supplemental Table 3 of Klammer et al. [65]. The validation data for 6 breast cancer cell lines were taken from Supplemental Table 4 of the same source. Detailed information about the generation of both datasets is provided in the main article of Klammer et al. [65]. In brief, the dataset contains more than 25,000 class-I phosphorylation sites (i.e. sites with high localization confidence), contaminant and reverse database hits were removed, and the normalized ratios (cell line versus SuperSILAC) were log10-transformed.

5.2.2 Pareto objective functions

Three objectives were considered: signature size, separation and relevance.

Signature size: This objective score is defined by the number of phosphosites in a given signature. The score is to be minimized.

Separation: This objective focuses on the generalization of the marker. Not only should a good marker separate the training data, but also unseen data. Thus, an inner leave-one-out cross validation was performed by employing a support vector machine with linear kernel and cost parameter $C = 1$. For each test sample, its distance of to the SVM hyperplane was computed and the posterior class probability was calculated from a sigmoid model $p_i = \frac{1}{1+\exp(Af_i+B)}$, where p_i is the probability of the sample being resistant and f_i the SVM output of the respective training data [144]. Parameter A was determined by optimizing a regularized maximum likelihood problem (see [144] for details); parameter B was fixed to 0, so that points on the separating hyperplane are assigned a probability of 0.5. The separation objective corresponds to minimization of the negative minimal probability distance $-\min_i(c_i(\frac{1}{2} - p_i) + \frac{1}{2})$, where c_i is the actual class of the cell line (sensitive = 1, resistant = -1).

Relevance: This objective deals with the relevance of a signature with respect to the drug target. Here, the mean distance of the signature's proteins to the target of the investigated drug in a protein-protein interaction network (STRING) was calculated. To this end, we calculated an adjacency matrix using all interactions with a interaction confidence larger than 0.9 (STRING version 9.05 [145]). The remaining edges with interaction confidence scores s ranging from 0.9 to 0.999 were transformed into a penalty score using the equation $\rho_i = \frac{1}{-\log_{10}(1-s_i)}$, ranging from 0.33 to 1. The function was chosen to get more pronounced differences between higher and lower confidence scores. Subsequently, the shortest path between each protein in the signature and the drug target based on the penalties ρ was calculated with the Dijkstra algorithm [146] and the mean distance of all signature proteins was used as objective score.

5.2.3 Pareto optimization

For the detection of the Pareto front, we applied the NSGA-II algorithm. NSGA-II is a fast, elitist multi-objective genetic algorithm [39] that employs the principle of Pareto optimality. In brief, the algorithm works as follows: First, a random parent population P_0 was generated that consists of $N = 200$ chromosomes. The representation of the chromosome is binary, i.e. a feature that is part of the signature is represented by 1, a feature that is not part of the signature by 0. The chromosomes were randomly initialized with 10% of features set to 1. Next, a fitness value was assigned to each chromosome, which represents the Pareto front the individual was located on. A fitness value of 1 stands for a solution on the first Pareto front, 2 for a solution on the second Parteo front, and so forth.

Subsequently, 5-way-tournament selection, single-point crossover ($p = 0.8$) and bit flip mutation ($p = 0.02$) were performed to generate the first offspring generation Q_0 ($N = 200$). To compile the next parent generation P_1, the individuals of P_0 and Q_0 were combined and the individuals were sorted according to non-domination (Pareto front $1 \ldots F$) and within each front according to the crowding distance (favors individuals that have a large distance to their neighbors, see [39] for details). Finally, the top N individuals were chosen to become P_1, which ensured elitism.

This procedure was repeated for G generations, where G is a number determined at runtime, at which the solutions on the first Pareto front do not change for 200 generations in feature space.

5.2.4 Biomarker discovery workflow

In order to detect multiple signatures based on the 19 NSCLC samples, the phosphorylation sites were first pre-filtered for missing data, i.e. only class-I sites with at least 2/3 of ratios present in each group (responder and non-responder) were considered for further analyses. Next, the 100 sites that discriminated best between responders and non-responders according to the MeanRank [61] test were selected, while ensuring that the mean difference of the features between the two groups was at least 4-fold and only one phosphosite per protein was included. This pre-selection is necessary to reduce the complexity of the subsequent Pareto optimization. The 100 top-ranking features were subjected to the NSGA-II algorithm, which aims at detecting Pareto-optimal solutions based on the three objective functions (size, separation, relevance).

After convergence, the results were filtered for solutions that were located on the first Pareto front. Since many of them were very similar and only differ in very few features, hierarchical clustering with Ward's method was applied on the binary solution vectors to detect clusters of similar features. Subsequently, the solutions that had the smallest Euclidean distance to the cluster centroids were taken as final Pareto signatures. If more than one solution had the smallest distance, the one with the better separation score was preferred.

5.2.5 Biomarker validation

For each Pareto signature, a support vector machine with linear kernel and cost parameter $C = 1$ was trained. These SVMs are the final predictors and can be used to predict new samples. To validate the signatures, we used phosphorylation site data of the six breast cancer samples. Prior to prediction, missing values were imputed by the mean of the training data class means (for details see [65]). Subsequently, the responsiveness of the six samples was predicted with each of the final predictors.

5.3 Results and Discussion

The main goal of response prediction biomarker studies is the identification of molecular signatures that separate the group of responders from the group of non-responders well and thus enable an accurate prediction of drug response. However, there are further qualities that characterize a successful biomarker. For example, a marker should consist of a manageable number of features (i.e. genes or proteins) in order to allow testing through methods applied in clinical routine such as quantitative PCR or ELISA. Furthermore, the features should be biologically relevant, for instance, by being connected to the drug's target or mechanism of action. In the proposed Pareto biomarker workflow, these three objectives - separation, signature size and relevance - are optimized in parallel (for definitions of objectives see section Pareto objective functions in Materials and Methods).

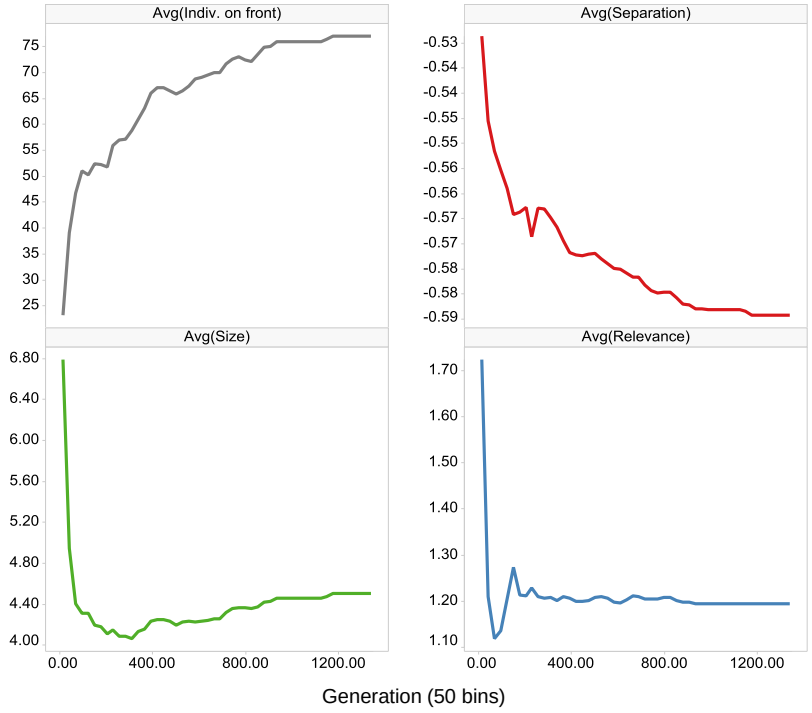

Figure 5.1: Evolution of the number of individuals (solutions) on the Pareto front and for the three objectives (separation, size and relevance), as generated by the NSGA-II algorithm [39]. The objectives are averaged across the solutions on the Pareto front. The number of generations is binned and the average of each bin is displayed on the y-axis.

5.3.1 Pareto biomarker workflow

To this end, a multiobjective optimization algorithm (MOA) was incorporated into our established biomarker discovery workflow [65], allowing the simultaneous optimization of all three objectives. Most MOAs employ the principle of Pareto optimality, which aims at detecting solutions that are not dominated by other solutions. At any given iteration, non-dominated solutions are defined such that there exist no other solutions that have a better or equal score in *all* objectives and a strictly better score in *at least* one objective. All non-dominated solutions (Pareto points) together form the Pareto front (see also Figure D.1), which is optimized during each iteration. Of the many MOA algorithms available (e.g. PAES [147], PESA [148], SPEA2 [149], NSGA-II [39] or SMS-EMOA [150]), we found the NSGA-II algorithm [39] most suitable for our Pareto biomarker workflow, as it shows fast convergence, is efficient and well tested [151, 152].

In a previous study, we used quantitative mass-spectrometry to globally profile the basal phosphoproteome of a panel of 19 non-small cell lung cancer (NSCLC) cell lines [65]. The effect of the kinase inhibitor dasatinb on cellular growth was tested against the same panel. Using the phosphoproteome data, we identified a phosphorylation signature consisting of 12 phosphorylation sites on 9 different proteins (referred to as original signature). The signature accurately predicted response to treatment with dasatinib in the NSCLC cell lines used for training and in an independent validation panel of breast cancer cell lines.

Here, we investigated whether the Pareto biomarker workflow could confirm the original signature and/or identify additional multivariate predictive phosphorylation signatures when applied to the same data set. In particular, these signatures should not only maximize class separation, but also the two additional objectives *signature size* and *relevance*. We hypothesize that a marker protein is more relevant if it is closely related to the drug target (e.g. through interaction). Although this might not always be the case, we think that this is a good assumption on average. Since this is only one out of three objectives to be optimized, signatures that are not connected to the drug's target may still be identified and are not discarded. More specifically, we define the relevance score of a signature as the average distance of the signature's proteins to dasatinib's main tar-

get in solid tumors, the Src kinase (SRC), as it has been shown that dasatinib inhibits migration and invasion of various solid tumors through inhibition of SRC [107–109]. All three scores are defined such that smaller values are better. Thus, all three objectives are to be minimized (see Materials and Methods for details).

From the 4,457 phosphorylation sites quantified in at least 2/3 of the samples in each class (responders and non-responders), we selected the 100 sites that discriminated best between responders and non-responders according to the MeanRank test [61], while ensuring that the mean difference between the two groups was at least 4-fold and only one phosphosite per protein was taken. This pre-selection was performed to reduce the complexity of the subsequent Pareto optimization. The algorithm terminated after 1353 generations, at which point the results on the first Pareto front had not changed for 200 generations (Figure 5.1). While the number of solutions on the Pareto front constantly increased, the three objectives (i.e. separation, size and relevance) were minimized with respect to Pareto optimality. As can be deduced from the graphs of the three objectives, the size and relevance criteria are rather easy to optimize, as they exhibit a steep decline at the beginning of the optimization process and reach the global minimum early on. Optimization of the separation criterion took longer and its decrease in the later stages was accompanied with an increase of the size objective, while the relevance criterion remained stable. In essence, small signatures that have a short distance to the drug target in the STRING protein-protein interaction (PPI) are readily discovered. It is, however, harder to find those that additionally separate the groups well and, essentially, good separation comes at the cost of larger signatures.

After termination, 77 solutions were located on the Pareto front. Solutions with a separation score ≥ -0.6 were removed ($N = 24$), as the Pareto approach also found very small and biologically relevant solutions with poor separation. This is an inherent feature of Pareto optimization, and removing undesired solutions is common practice (see e.g. [153]). The remaining 53 solutions contained 35 different phosphorylation sites. One site, S1148 on integrin β4 (ITGB4), was part of all but three solutions.

Figure 5.2A shows a series of three-dimensional plots of the Pareto front. The front has the shape of a stretched canvas attracted by the origin, which represents an ideal but infeasible point. Figure 5.2B depicts 2D projections of the 53 Pareto front solutions in

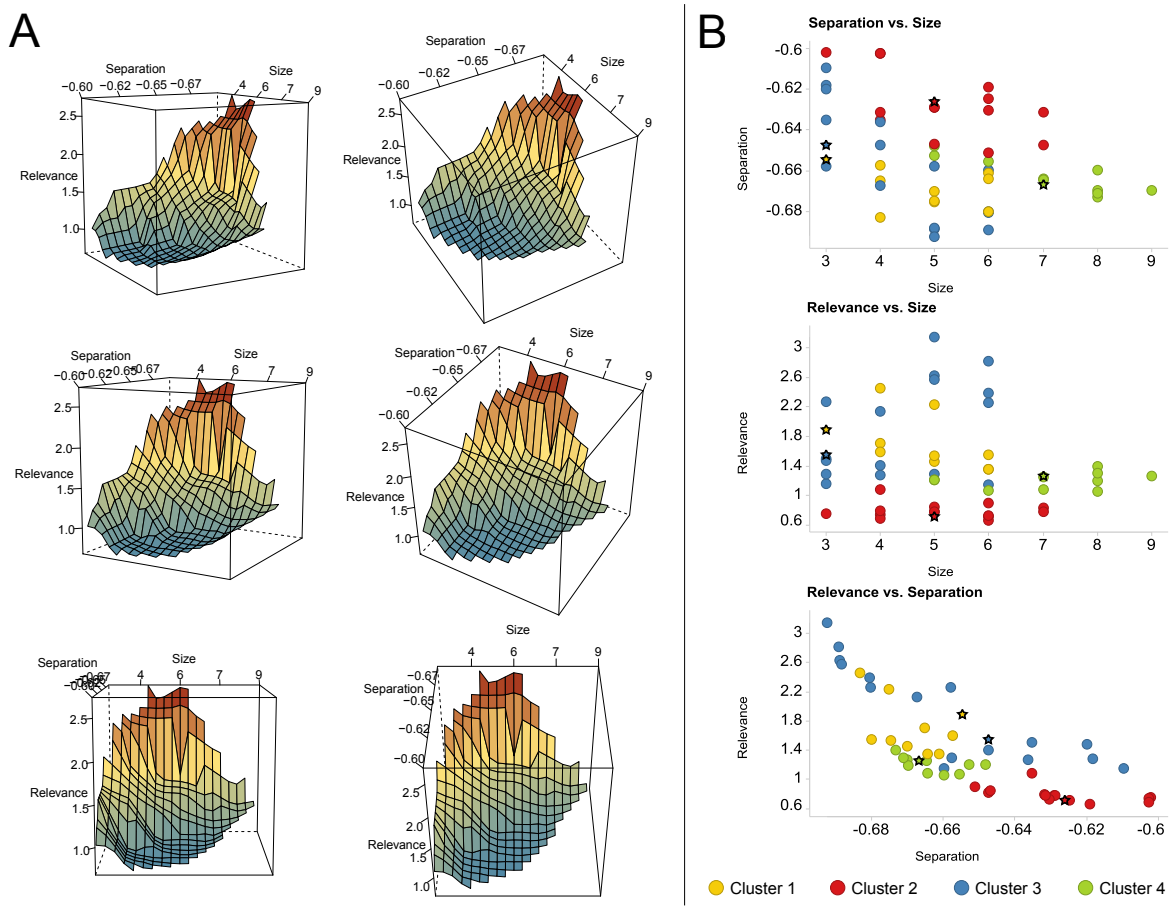

Figure 5.2: 3D plots of the Pareto front (A) and 2D projections (B). (A) The different panels are views of the Pareto front when rotated around the relevance-axis (with two different viewing angles in each column). Coloring indicates relevance score, from blue (low) to red (high). Since better solutions are smaller in all three dimensions, the optimal point is the origin in the lower background, i.e. the only hidden vertex in the plots. (B) 2D projections of the solutions on the Pareto front. Solutions are colored according to their assignment to four clusters. Stars mark the solutions closest to the respective cluster centroid that were selected as final Pareto signatures.

objective space. The top panel, relating size and separation, shows that smaller signatures lead to less pronounced separation and illustrates our initial motivation for identifying multivariate markers. Therefore, the lower left corner in the plot, where ideal solutions for the two respective objectives are expected, is not populated. However, there are also no large signatures in the area of the best-separating solutions (< -0.68). This is due to the third objective, the relevance criterion, as it becomes harder to identify features

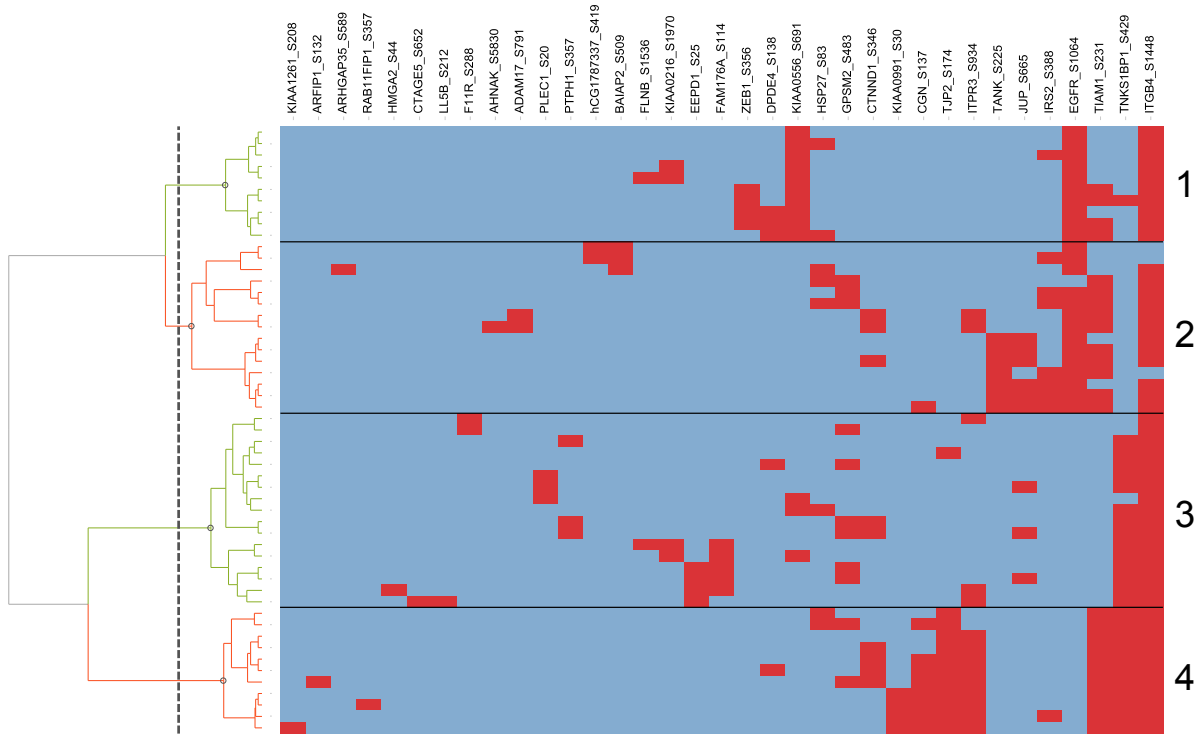

Figure 5.3: Hierarchical clustering of the 53 accepted solutions on the Pareto front in feature space. In each row, red areas represent features (phosphosites) that are part of the corresponding solution. The solutions were subdivided into four clusters according to the row dendogram on the left. Cluster numbers are indicated on the right.

that all interact directly or indirectly with the target. As mentioned before, the task of finding small and biologically relevant solutions is achieved more easily, as can be seen in the center panel of Figure 5.2B. Solutions are found in the lower left area, but not in the lower right. The bottom panel of Figure 5.2B depicts the relationship between separation and relevance. This projection of the Pareto front has a curved shape, revealing the compromise between good separation and biologically meaningful features, as not all well-discriminating phosphosites are also related to the drug target.

5.3.2 Pareto signatures

Each of the identified solutions on the Pareto front is optimal in the sense that none of them are dominated by any other solution. Therefore, each solution could be evaluated individually. Here we took another approach and investigated whether solutions can be

reduced by clustering according to their similarity while retaining discriminatory power. To this end, we hierarchically clustered the solution in features space using the Ward method and obtained four major clusters (see Figure 5.3). For each of these clusters, the feature with the smallest Euclidean distance to the respective cluster centroid was selected as so-called *Pareto signature* for further analysis (see Figure 5.2B).

In order to compare the original 12-phosphosite signature with the Pareto signatures, we calculated its objective values: $size = 12$, separation $= -0.60$, relevance $= 1.63$ (see also Table 5.1). Note, that the original signature was optimized with respect to prediction accuracy only, and the feature selection method did not explicitly optimize the separation criterion as defined here (see Materials and Methods). Figure 5.4A shows the PPI network of the original marker, where solid lines indicate the shortest path from each signature phosphoprotein (blue) to SRC (red), which is dasatinib's main target in solid tumors. The phosphorylation sites of the signature are listed in Table 5.2. Some of the signature proteins ITGB4, ARHGEF18 and BAIAP2 are closely related to SRC, while others (e.g. ATG16L1 and TNK1SBP1) have larger distances in the PPI network. TPD52L2 and GPRC5A have no connection to SRC at all. In the original publication [65], the selected features were used to train a support vector machine (SVM) with linear kernel. The signature and the corresponding predictor were then validated by application to six independent breast cancer cell lines, which had not been used for feature selection or SVM training. In the case of the original signature, five out of six cell lines were predicted correctly with an average probability distance to the hyperplane of 0.13 (calculated as $\frac{1}{N}\sum_{i=1}^{N}(0.5-p_i)c_i$, where N is the number of tested cell lines, p_i the prediction probability for the cell line to be resistant, and c_i the actual class of the cell line (sensitive $= 1$, resistant $= -1$)).

The first Pareto signature (Pareto1, Figure 5.4B) contains only three phosphoproteins - ITGB4 (integrin β4) S1448, EGFR (epidermal growth factor receptor) S1064 and KIAA0556 (uncharacterized protein KIAA0556) S691 – for details see Table 5.2. While the separation and size objective scores are better than those of the original signature (see Table 5.1), the relevance score is slightly worse, which is due to the uncharacterized protein KIAA0556 that lacks functional annotation and therefore has no connection with SRC. The prediction accuracy on the validation set is comparable to that of the original

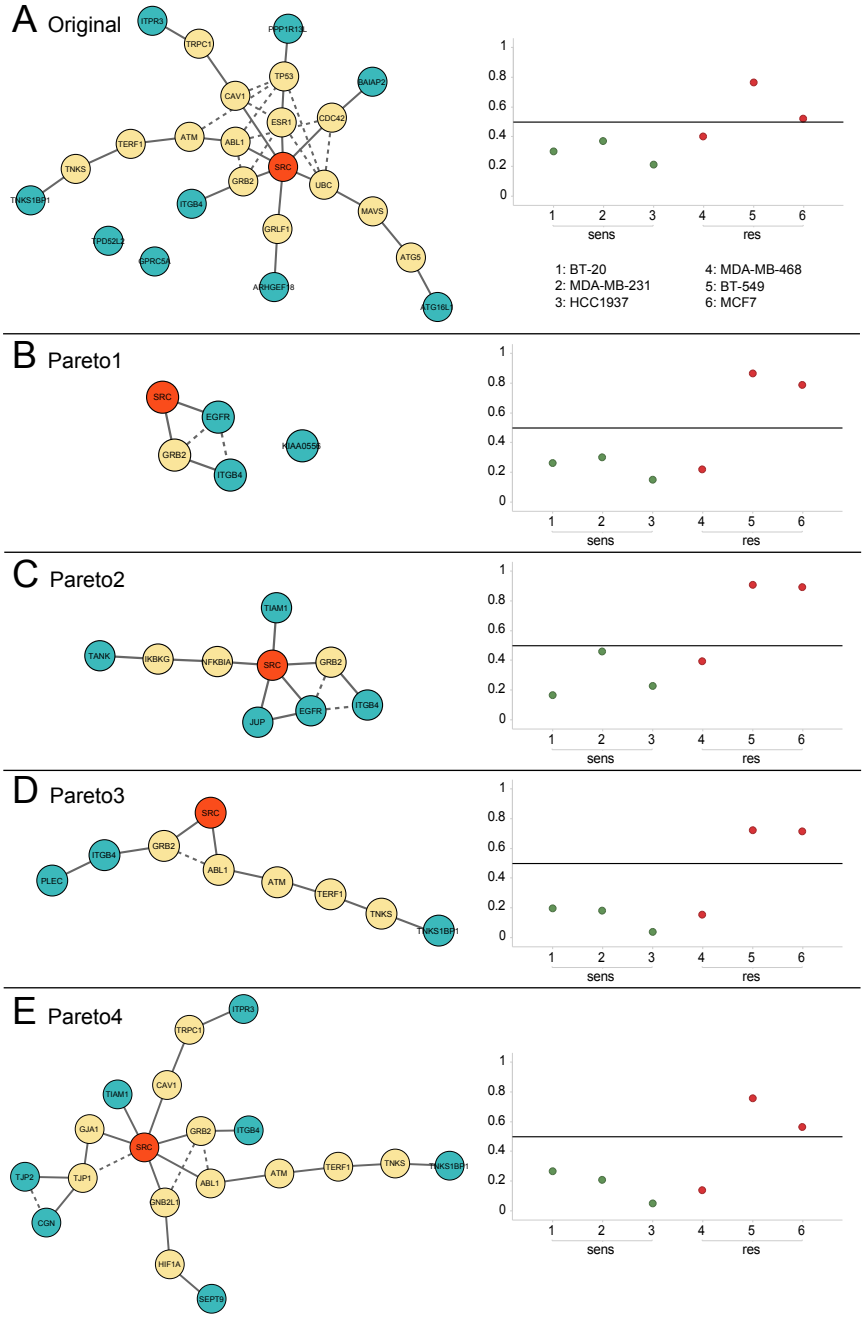

Figure 5.4: Network of the original (A) and the four Pareto signatures (B-E) are shown in the left column along with the prediction results for the breast cancer validation set (right column). Blue nodes in the network are signature proteins, solid lines represent the shortest path to SRC, dashed lines indicate additional high-confidence interactions. Sensitive (1–3) and resistant (4–6) breast cancer cell lines were considered to be predicted correctly if they received a probability (y-axis) of less than 0.5 and more than 0.5, respectively.

signature, however, the average probability distance to the SVM hyperplane is slightly higher and thus better. Phosphorylation site S1448 on ITGB4 is one of the best separators in the data set and is also part of the original signature. ITGB4 is linked to the Src-Fak pathway [112] and is associated with poor patient prognosis [113–115]. The EGF receptor can be phosphorylated by the Src kinase [154], and is therefore directly linked to SRC in the protein-protein interaction network (STRING confidence score of 0.999).

The signature Pareto2 (Figure 5.4C), contains the same phosphosites on ITGB4 and EGFR, and additionally TANK (TRAF family member-associated NF-kappa-B activator) S225, TIAM1 (T-lymphoma invasion and metastasis-inducing protein 1) S231 and JUP (Junction plakoglobin) S665 – see also Table 5.2. This signature has a particularly good relevance score (cf. Table 5.1), which is also visible in the PPI network, where 4 out of 5 proteins are closely connected to SRC. The performance on the validation data is comparable to that of Pareto1.

The third Pareto signature (Pareto3, Figure 5.4D) is another small signature containing sites S1448 on ITGB4, S20 on PLEC1 (Plectin) and S429 on TNKS1BP1 (182 kDa tankyrase-1-binding protein). S429 on TNKS1BP1 is also part of the original signature, together with S1448 on ITGB4. TNKS1BP1 has a rather large distance to SRC, leading to a mediocre relevance score. The other scores are identical to those of Pareto1, the second 3-phosphosite signature (cf. Table 5.1).

Finally, the largest Pareto signature (Pareto4, Figure 5.4E), contains ITGB4 S1448, TNK1SBP1 S429, TJP2 (Tight junction protein ZO-2) S174, CGN (Cingulin) S137, SEPT9 (Septin-9) S30, TIAM1 S231 and ITPR3 (Inositol 1,4,5-trisphosphate receptor type 3) S934. Again, the sites on ITGB4 and TNKS1BP1 are those that are part of the original signature. ITPR3 appears in the original signature with a different phopshosite (S916).

Taken together, Pareto markers are consistently smaller than the original marker, while three of four also have better separation and relevance scores. The prediction accuracy on the validation set is identical for all investigated signatures, however, the average probability distance to the separating SVM hyperplane is slightly higher for the Pareto signatures, suggesting that the Pareto signatures are more robust when being applied to other classes of related tumor cell lines.

Table 5.1: Objective scores (smaller are better), prediction accuracy and average probability distance for the validation data (larger are better).

Signature	Size	Separation	Relevance	Validation accuracy	Validation distance
Original	12	-0.60	1.63	5/6	0.13
Pareto1	3	-0.65	1.88	5/6	0.19
Pareto2	5	-0.63	0.72	5/6	0.22
Pareto3	3	-0.65	1.55	5/6	0.19
Pareto4	7	-0.67	1.26	5/6	0.15

5.4 Conclusion

We and others have previously shown that the identification of response prediction markers from phosphoproteomics experiments in pre-clinical or clinical settings is possible [65, 93, 133]. These studies sought to identify single signatures of phosphorylation sites maximizing the separation on the data used for training. Here, we investigated the idea of integrating additional objectives, such as the relevance with respect to the drug target or the size of a signature, into the feature selection process. We applied the multi-objective genetic algorithm NSGA-II [39] to the identification of Pareto-optimal solutions for the prediction of response of NSCLC cell lines to treatment with dasatinib. Beside separability, we used the proximity of markers to the main drug target – the Src kinase – and the size of the signature as objectives for optimization.

In total, the algorithm identifies 77 Pareto-optimal solutions, i.e. solutions that are not dominated by any other solution. Each solution corresponds to a phosphorylation signature that can be used for response prediction. 53 of them had a sufficiently good separation score and were considered in the following analysis. Clustering of these solutions in feature-space revealed four groups of solutions with similar sets of phosphorylation sites. We used the solution closest to the centroid of each cluster as representatives of the four Pareto signatures. All four signatures predicted the response of six breast cancer cell lines that were not used for training with good accuracy (83%). The same accuracy was also reached by the original 12-marker signature identified in Klammer et al. [65].

Table 5.2: Phosphorylation sites of the final signatures. Sites/proteins in bold are part of the original signature [65].

Signature	Accession	Gene name	Site
Original	P16144-2	ITGB4	S1448
	Q9UQB8-5	BAIAP2	S509
	P16144-2	TGB4	S1387
	P16144-2	TGB4	T1385
	P16144-2	TGB4	S1069
	A8K556	GPCR5A	S345
	Q14573	ITPR3	S916
	Q9C0C2	TNKS1BP1	S429
	Q6ZSZ5	ARHGEF18	S1101
	Q8WUF5	PPP1R13L	S102
	Q676U5	APG16L	S269
	O43399-2	TPD52L2	S141
Pareto1	**P16144-2**	**ITGB4**	**S1448**
	A9CB80	EGFR	S1064
	O60303	KIAA0556	S691
Pareto2	B2R7S3	TANK	S225
	P16144-2	**ITGB4**	**S1448**
	A9CB80	EGFR	S1064
	Q13009	TIAM1	S231
	P14923	JUP	S665
Pareto3	**P16144-2**	**ITGB4**	**S1448**
	Q15149-4	PLEC1	S20
	Q9C0C2	**TNKS1BP1**	**S429**
Pareto4	**P16144-2**	**ITGB4**	**S1448**
	Q9C0C2	**TNKS1BP1**	**S429**
	Q9UDY2	TJP2	S174
	B9EK46	CGN	S137
	Q9UHD8	SEPT9	S30
	Q13009	TIAM1	S231
	Q14573	**ITPR3**	S934

The phosphorylation site S1448 on ITGB4 (Uniprot accession: P16144-2; or, equivalently, S1518 in the canonical sequence P16144-1) is the central feature in all four Pareto signatures. This is in accordance with the results from Klammer et al. [65], where the same site was ranked first in the robust feature selection approach. Furthermore, ITGB4 is closely linked to dasatinib's main target SRC via the adapter protein GRB2. Thus, the Pareto marker approach is consistent with the outstanding role of ITGB4 in predicting dasatinib response in cancer cells.

The four Pareto signatures are characterized by properties that correspond to the objectives used for optimization. Signature 1 and 3 are relatively small with only three phosphorylation sites each. On the other hand, signature 4 is larger (7 sites), but has the best separation. Finally, signature 2 shows the best relevance score, meaning that its marker proteins are interacting with the drug target SRC either directly or through intermediate proteins. Surprisingly, while its separation on the training data is the smallest of all four signatures, it yields the highest separation on the breast cancer cell lines that were used for validation. This hints at the importance of incorporating network information in general and the relationship to the drug target in particular for the selection of predictive features.

Here, we optimized the selected features with respect to the objectives separation, size, and relevance. Naturally, the proposed method can be applied to other objectives. For example, it may be sensible to include the detectability of marker phosphorylations in immunoassays, the localization of the marker proteins (e.g. cell membrane, nucleus, or cytosol), or the extend of knowledge about the proteins (e.g. number of PubMed abstracts).

Aside from the possibility of incorporating multiple objectives into the selection of the biomarker signatures, an even more important advantage of the approach presented here is the identification of several independent signatures instead of only one. These signatures can be evaluated post-hoc using additional criteria before a final signature or a set of a few signatures is selected for further validation experiments.

In summary, we presented a general method for identifying a set of biomarker signatures from high-dimensional data such as proteomics, phosphoproteomics, or transcriptomics data. Besides optimizing the separation between two classes, the method allows

the consideration of additional objectives. In particular, we showed that the relation of the marker proteins to the drug target in a protein-protein interaction network can improve the robustness of the prediction when applied to new samples.

Chapter 6

Conclusions and Outlook

In this thesis, four new methods for analyzing mass spectrometry-based phosphoproteomic data were presented. The first two methods (SubExtractor and MeanRank test) were designed to help uncovering a drug's mode of action, but are also applicable to target identification or biomarker discovery projects. The SubExtractor algorithm (Chapter 2) aims at discovering significantly regulated subnetworks from protein-protein interaction databases by employing a Bayesian probabilistic model in combination with a genetic algorithm and stringent significance evaluation. SubExtractor has been part of Evotec Munich's phosphoproteome analysis platform from the beginning, and has aided the interpretation of numerous mode-of-action studies (e.g. [33], [14]) as well as biomarker projects (e.g. [65], Chapter 4). Since the first development and publication of the algorithm, some enhancements and modifications have been made. The hypothesis test to determine the significance of the resulting subnetworks was changed from the global rank test [53] to the MeanRank test, as the latter shows superior performance under virtually all conditions (confer Chapter 3). Furthermore, the inclusion of the MeanRank test now allows for a comparison between two groups (2-sample test), which is useful when a SuperSILAC or label-free approach is used for quantification of phosphosite abundances.The repertoire of interaction databases has been expanded from the protein-protein interaction database STRING to more specific kinase-substrate interaction databases like PhosphoSitePlus [116], PhosphELM [155] and NetPhos [156].

The MeanRank test (Chapter 3) focuses on detecting significantly regulated feature from noisy and sparse high-throughput data with only few replicates. Its big advantages

are the rank-based nature that does not require any a priori distribution assumptions, its tolerance regarding missing values, and its high statistical power that scales well with the number of available replicates. The test is especially valuable when only few replicates are available and the fraction of truly regulated features is around or below 10%, which is usually the case, when samples are treated with a kinase inhibitor. In such scenarios, the MeanRank test outperforms well-established methods such as SAM and LIMMA. The test has been incorporated into the standard proteome and phosphoproteome analysis workflow at Evotec Munich, where it is now routinely applied to analyze customer projects as well as research projects (e.g. [21]).

The latter two chapters of this thesis (NSCLC biomarker and Pareto biomarker) aim at finding multivariate phosphorylation signatures for predicting the response of non-small cell lung cancer (NSCLC) cell lines to the kinase inhibitor dasatinib. The earlier method described in Chapter 4 (NSCLC biomarker) was the first published global and unbiased approach to develop a biomarker based on the differences in the basal phosphorylation levels of cancer cell lines. The presented algorithm employs a robust feature selection method in combination with support vector machine (SVM) classification in order to identify a predictive set of phosphorylation sites. Applying the fully cross-validated workflow led to a correct prediction of 18 out of 19 samples. The final signature consists of 12 phosphorylation sites, and was able to correctly predict dasatinib sensitivity of 5 out of 6 samples from an independent breast cancer validation set. In a later study, a similar workflow was applied to uncover a phosphorylation signature for predicting the response of acute myeloid leukemia (AML) patient samples to the FLT3 inhibitor quizartinib (AC220) [133]. Here, 11 out of 12 bone marrow samples could be correctly predicted in the cross-validation procedure. By applying the final signature of 5 phosphorylation sites to an independent validation set, 7 out of 9 predictions were correct.

The biomarker workflow described in Chapter 4 focuses on the identification of a signature that separates the two groups (responders and non-responders) well. While this was the only objective in this study, the Pareto marker workflow presented in Chapter 5 adds additional objectives to the feature selection process, i.e. feature relevance and signature size. To this end, the multiobjective Pareto optimization algorithm NSGA-II was incorporated into the biomarker workflow, resulting in a set of four Pareto signatures after

filtering and clustering of the Pareto-optimal solutions. All of these four signatures were smaller than the original signature reported in Chapter 4 (ranging from 3 to 7; originally 12) and three of them were on average closer related to the main target of dasatinib, the SRC kinase. At the same time, the prediction performances of all Pareto signatures on the validation set were as good before, and the key protein of the original signature – integrin beta 4 (ITGB4) – was also present in all of them.

Concluding, the novel methods described in this thesis help to better understand the processes underlying drug treatment and support the development of response prediction biomarkers in the field of large-scale *omics* in general and phosphoproteomics in particular.

Appendix A

Supplementary Information to Chapter 2

A.1 Introduction to Genetic Algorithms (GAs)

GAs mimic the process of biological evolution. The primary component of the GA is the individual, which contains exactly one chromosome and one fitness value. A chromosome in GA language is a vector of values (in the simplest way in binary form) representing one distinct solution for the optimization problem. The fitness value determines the quality of the corresponding solution encoded by the chromosome. Depending on the underlying fitness function it is desirable to either maximize or minimize the fitness value.

A typical GA has at least tens or hundreds of different individuals with different chromosomes. As the algorithm evolves, individuals are selected according to their fitness value and bred using crossover and mutation operators to create new offspring and thus new solutions to the problem. Subsequently, some weak individuals (i.e. individuals with low fitness value) from the parental generation are replaced by strong offspring individuals and the process starts over again. According to the building block hypothesis, small areas with superior fitness on different chromosomes are thus iteratively combined into longer ones, leading to a steady increase in fitness (not necessarily for each individual but at the level of the entire population). Random mutations reduce the risk of getting trapped in a local optimum. The general workflow for a GA is depicted in Supplementary Figure A.1.

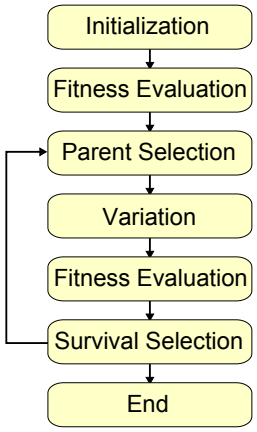

Figure A.1: Schematic GA workflow. First, the individuals' chromosomes are initialized. Then, their fitnesses are evaluated and parents for the first reproduction are selected. Subsequently, variation, i.e. recombination of the parents' chromosomes and mutation, takes place. This is followed by the fitness evaluation of the newly created individuals and subsequent survival selection. In this step low-performing individuals of the parental generation are replaced by high-performing offspring. Steps 3 to 6 are repeated until a certain termination condition (e.g. number of generations or satisfying solution) is fulfilled.

A.2 Lower bound for parameter α

As described in the *Artificial data* subsection of the main article, a too small value for α will lead to incorporation of unregulated nodes if their only connection is to a well-regulated one. To avoid this, the α value should be chosen such that an unregulated node with only one well-regulated neighbour always gets a higher score when it is flagged as inactive, i.e. not part of an differentially regulated subnetwork. More formally, this requirement can be expressed based on Equation 2.8 in the main article with the equation

$$\ln\left(\mathcal{N}(0|0,1)\right) + \ln\left(\alpha + 0\right) > \ln\left(\mathcal{N}(0|0,\sigma_z^2)\right) + \ln\left(\alpha + 1\right). \tag{A.1}$$

Solving for α leads to

$$\alpha > \alpha_c = \frac{1}{\frac{\mathcal{N}(0|0,1)}{\mathcal{N}(0|0,\sigma_z^2)} - 1}, \tag{A.2}$$

or equivalently

$$\alpha_c = \frac{\mathcal{N}(0|0,\sigma_z^2)}{\mathcal{N}(0|0,1) - \mathcal{N}(0|0,\sigma_z^2)}. \tag{A.3}$$

Equation A.3 is then used to calculate the lower bound of reasonable values for α. Some examples for varying σ_z values are:

$$\sigma_z = 3\text{: } \alpha_c = 0.5$$

$$\sigma_z = 5\text{: } \alpha_c = 0.25$$

$$\sigma_z = 10\text{: } \alpha_c = 0.11$$

A.3 Supplementary Figures

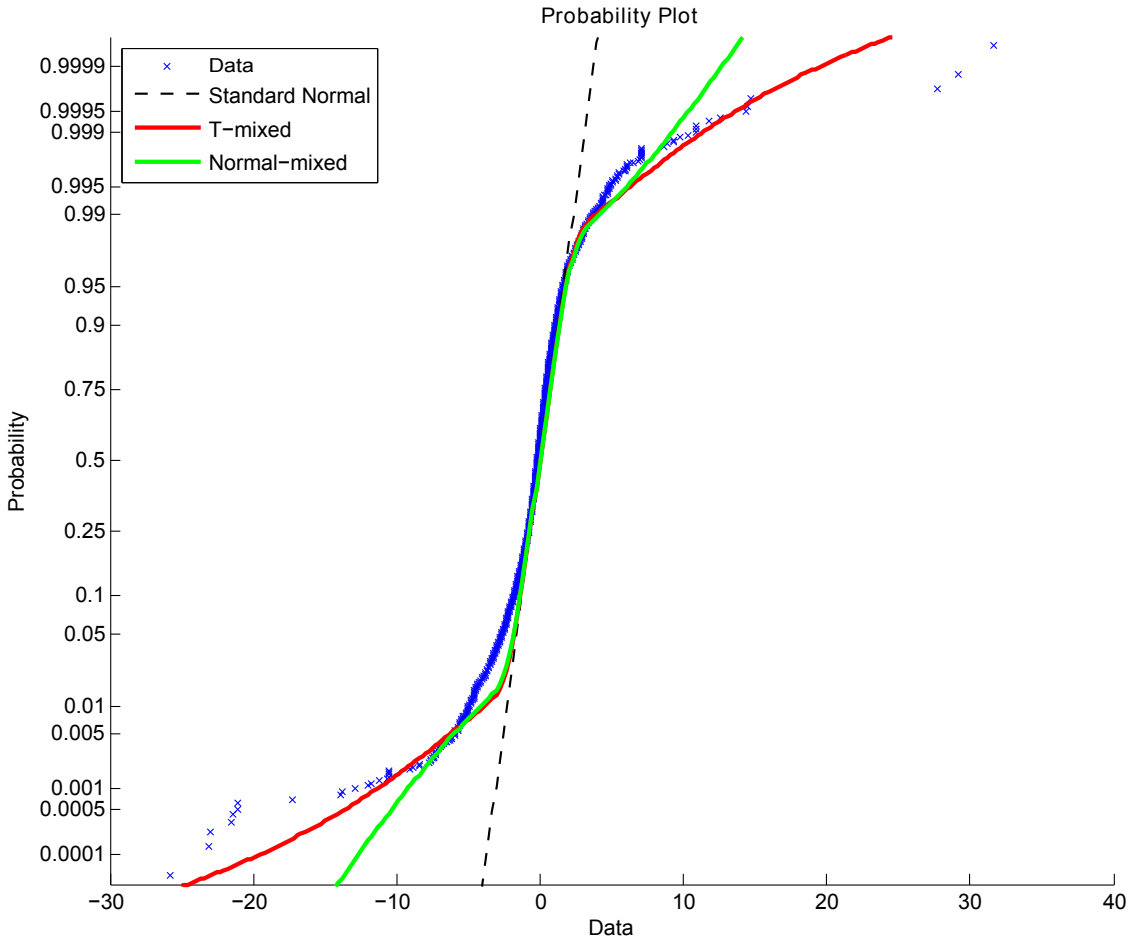

Figure A.2: Different distributions and their fit to the sorafenib data. *Normal-mixed* is a mixture model of two normal distributions; *t-mixed* is a mixture of a normal and t location scale distribution.

A.4 Additional Files

Additional Files 2-5 can be found at:

http://www.biomedcentral.com/1471-2105/11/351/additional

Appendix B

Supplementary Information to Chapter 3

B.1 Pseudocode of one-sample MeanRank algorithm

Algorithm B.1 Calculation of α^0, the expected number of false discoveries.

M = matrix holding features and replicates

$\overline{r}_{sorted}$ = vector of sorted mean ranks

α^0_{sum} = vector of length N

for all $s \in$ Cartesian product $\{-1,+1\}^R$ without$\{(-1,\ldots,-1),(+1,\ldots,+1)\}$ **do**

  $M_{flipped}$ = multiply values in columns $(i = 1,\ldots,R)$ of M with values of s_i

  $\overline{r}$ = mean ranks of$M_{flipped}$

  for $i = 1,\ldots,N$ **do**

    $\alpha^0_{sum}(i)+=$ count $\overline{r}_{flipped} < \overline{r}_{i,sorted}$

  end for

end for

$\alpha^0 = \alpha^0_{sum}$ / number of flips

B.2 Two-sample Mean Rank test

The two-sample version of the proposed test is similar to the one-sample case, with only a few modifications. For the two-sample case, the two input matrices, M_1 and M_2, hold the data from two different groups, e.g. treated and untreated. Both must have the same number of features N, but the number of replicates may be different (R_1 and R_2, respectively). As the aim of two-sample tests is to find differentially regulated features between two groups, we create a difference matrix prior to step 1. in Equation 3.1 in the main article. This difference matrix contains all possible $R_1 \cdot R_2$ pair-wise differences between the two data matrices. Note, that often values should be log-transformed to achieve a symmetric distribution of differences. The ranks are then calculated on the difference matrix and steps 2. and 3. of Equation 3.1 are performed using this $R_1 \cdot R_2$ matrix of difference ranks.

The Bates distribution cannot be used for the parametric estimation of α^0, as the $R_1 \cdot R_2$ columns of the difference rank matrix are not independent. Thus, the null distribution has to be determined numerically. This is done by generating two random data matrices (sampled from a standard normal distribution) with R_1 and R_2 columns, respectively, and very large N ($\geq 100,000$). The difference rank matrix is then calculated as described above, and the empirical distribution of the resulting mean rank values is determined, which can then be used instead of $\mathrm{F}_{\mathrm{Bates}}$ function to estimate α^0.

The non-parametric estimation of α^0 has to be modified for the two-sample test, as well. Instead of performing sign flipping to estimate false positives, we now randomize the group association and calculate α^0 accordingly.

Additional simulations were carried out to assess the performance of the two-sample MeanRank, and compare it to the two-sample versions of SAM, LIMMA, RankProducts and the t-test. A two-sample version of GlobalRank does not exist. In line with the results of the previous one-sample simulations, MeanRank and SAM performed better than RankProducts and the t-test. In simulations with normally distributed data and no missing values MeanRank and SAM showed comparable power and met the FDR level; however, both were outperformed by LIMMA (see Supplementary Figure B.1A-C). In the case of simulation data sampled from a non-normal (Student's-t) distribution, the

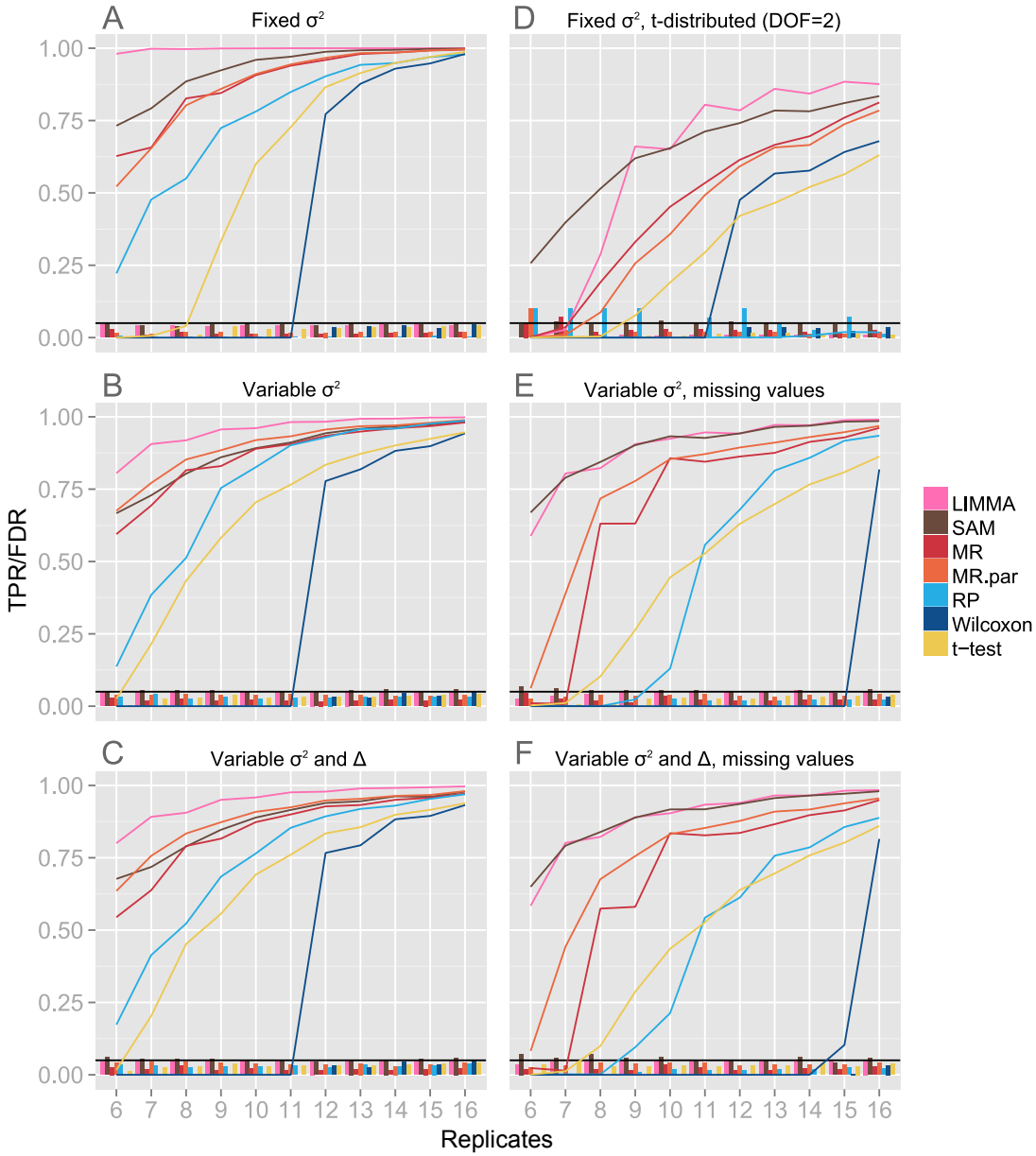

Figure B.1: Performance of two-sample tests on simulated data. Performance plot of two-sample significance tests under different simulation settings. Traces show the true positive rate (TPR) of the respective tests for a given number of replicates. Bars denote the false discovery rate (FDR). TPR and FDR are averaged over ten independent simulations. All tests were set to control the FDR at 0.05.

TPR of the parametric MeanRank test drops below the non-parametric version, as the parametric FDR estimation involves the assumption of normal distributions. Both MRs

performed better than RankProducts and the t-test, but worse than SAM and LIMMA. The introduction of missing values also led to a drop in power for MeanRank. This can be explained by the way the difference-rank matrix is calculated. Since each subtraction involving a missing value again produces a missing value, the proportion of missings in the difference-rank matrix is larger than in the initial data matrices M_1 and M_2.

B.3 Supplementary Figures

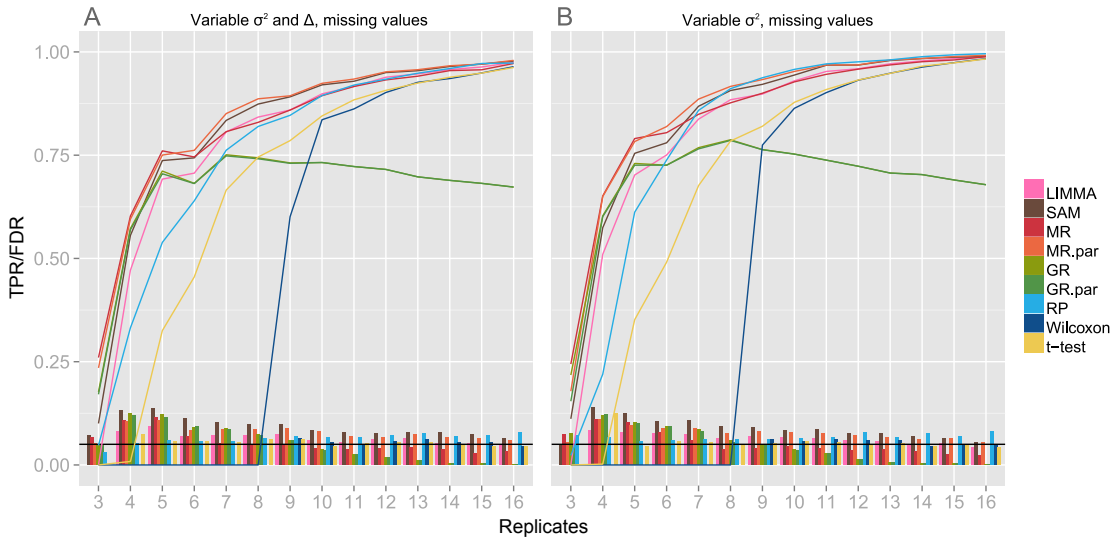

Figure B.2: Performance on simulated data using imputation. Performance plot of tests for one-sample simulation data with missing data imputed by k-nearest-neighbor (k-NN) with $k = 10$.

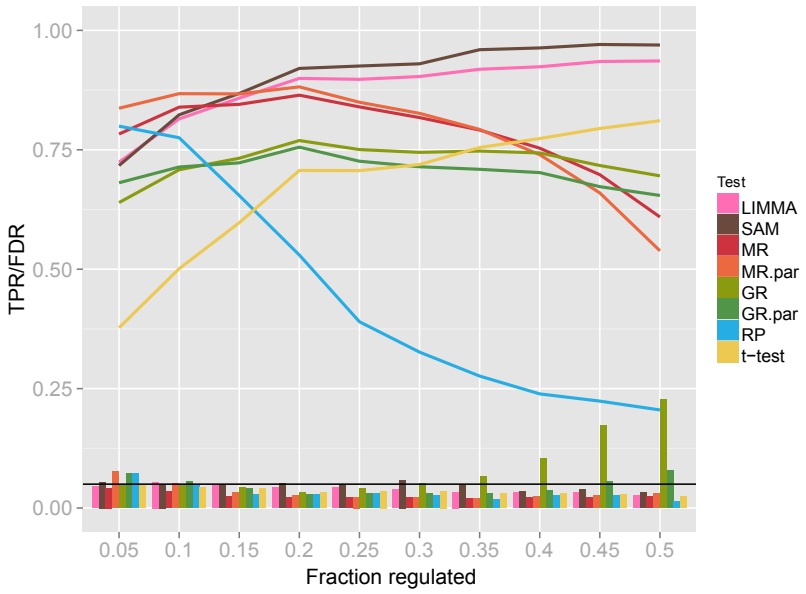

Figure B.3: Performance for different fractions of regulated and unregulated features. Performance with fixed number of replicates ($R = 6$), over a varying fraction of regulated features to background features.

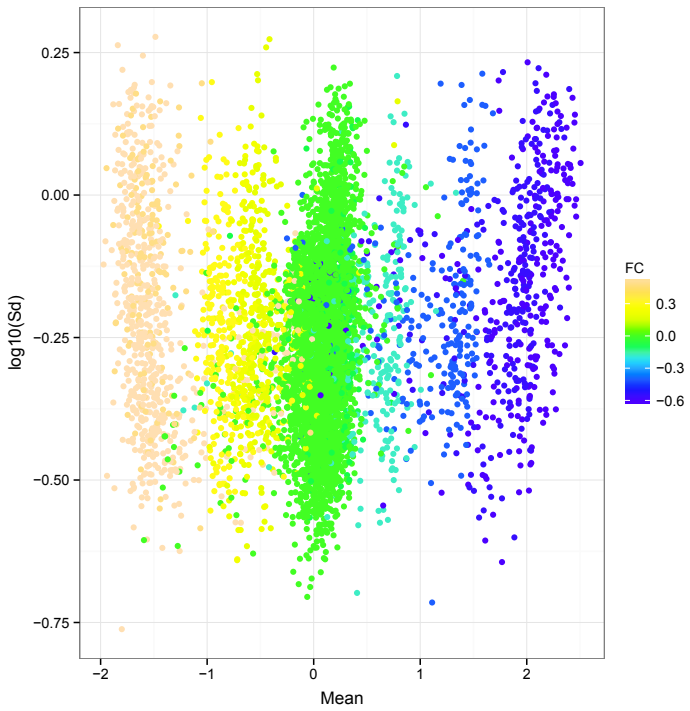

Figure B.4: Volcano plot highlighting spike-in concentrations. Volcano plot of the 'Ag-Spike' data, colored by fold-change of spike-in.

B.4 Supplementary Tables

	R replicates			
	R=5		R=30	
N features	Parametric	Non-parametric	Parametric	Non-parametric
$N = 1.000$	$<$ 1s / $<$ 1MB	$<$ 1s / $<$ 1MB	3s / $<$ 1MB	2s / $<$ 1MB
$N = 10.000$	8s / $<$ 1MB	20s / 1MB	45s / 2MB	184s / 18MB

Table B.1: Computational performance of the MeanRank test. Computation time and memory usage shown in seconds and megabytes, respectively. Measurements were performed on a single core of an Intel i5 2400, with 3.1 GHz.

B.5 Additional Files

Additional Files 2-3 can be found at:
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0104504

Appendix C

Supplementary Information to Chapter 4

C.1 Cost matrix example

The example in Figure C.1A shows how the introduction of cost matrices influences the support vector classification. The figure shows a classification example that aims at separating red stars from blue crosses. Each class contains 10 samples with two features. The values of both features were sampled from normal distributions (N(1,1) and N(-1,1) for crosses and stars, respectively). The black line represents the separating hyperplane of the SVM classification with linear kernel (parameter C=1), when no explicit cost matrix is applied (i.e. the cost of misclassifying a star is the same as the cost for misclassifying a cross). One can clearly see that the data is not linearly separable, which leads to one misclassified cross and one misclassified star. The red line shows the hyperplane when the cost for the false classification of stars is twice as high as the cost for star misclassification. As a result, the separating hyperplane is shifted towards the cloud of red stars, but the classification result is still the same. By increasing the cost factor of cross misclassification to ten times the cost of star misclassification, the hyperplane (blue line) is shifted further and all crosses are classified correctly. However, instead of one falsely predicted star there are now four. Finally, when using a cost factor of 200 (see purple line), all samples would be classified as crosses leading to ten wrongly predicted stars.

This shifting of the hyperplane can be used to calculate the receiver operating characteristic (ROC) curve and the area under it. A ROC curve based on the four different cost matrices above would look like Figure C.1B (assuming that the crosses are the positives and the stars the negatives in the ROC statistics). The point at (1.0|1.0) corresponds to the purple hyperplane, where all crosses are classified correctly and all stars wrongly; the point at (0.4|1.0) to the blue discrimination line, where all crosses are classified correctly and 4 stars are falsely predicted as positives; the point at (0.1|0.9) to both the red and black hyperplane, where 9 crosses are classified correctly and one star wrongly as positive; and finally one more point at (0|0) that is not depicted in Figure C.1A but represents the extreme when all samples are assumed to be negatives (stars), which can be considered the opposite of the purple discrimination line. Finally, the area under the curve can be computed, which is 0.93 in this example.

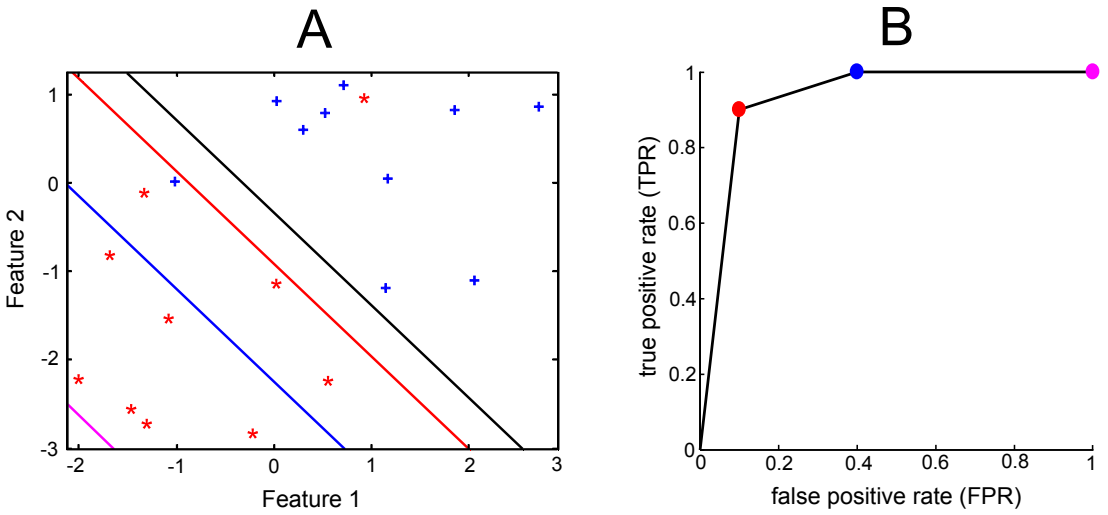

Figure C.1: Classification example using a linear SVM with different cost matrices (A), and the corresponding ROC curve (B).

C.2 Details on SVM prediction

The decision function of the SVM classification is given by

$$f(\vec{x}) = \text{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i k\left(\vec{x_i}, \vec{x}\right) + b\right), \tag{C.1}$$

where m is the number of training samples (cell lines), y_i the class label of the i^{th} training sample (-1 or 1 for sensitive and resistant cell lines, respectively), α_i the respective Lagrange multiplier, $\vec{x_i}$ a vector of length f (f being the number of selected features) holding the ratios of the i^{th} training sample, $\vec{x}$ a vector of length f holding the ratios of the test sample, and b the bias (i.e. the translation of the hyperplane with respect to the origin). $k(\vec{x_i}, \vec{x})$ is called a kernel, i.e. a function that characterizes the similarity of two vectors. Equation C.1 can be rewritten as

$$f(\vec{x}) = \text{sgn}\Big(k\left(\vec{w}, \vec{x}\right) + b\Big), \tag{C.2}$$

with the weight vector $\vec{w}$, whose elements represent the importance (influence) of the corresponding features, defined as $\vec{w} = \sum_{i=1}^{m} \alpha_i y_i \vec{x_i}$. In the case of the linear SVM, the kernel function is defined as the dot product of the two vectors, which leads to the linear decision function

$$f(\vec{x}) = \text{sgn}\left(\sum_{j=1}^{f} w_j x_j + b\right). \tag{C.3}$$

So far, changes in the phosphorylation level were represented by ratios, which can be expressed as $x = S - S_{ref}$, where S is the signal of the phosphosite in the corresponding cell line and S_{ref} the signal of the site in the reference cell line pool. Here, the signal is defined as log intensity of the corresponding phosphosite. For data produced by other methods such as multiple reaction monitoring or ELISA, where the quantitative data are represented by intensities, one can still make predictions with the proposed phosphosignature, but the decision function (Equation C.2) has to be modified to

$$f(\vec{S}) = \text{sgn}\Big(k\big(\vec{w}, \vec{S}\big) + \underbrace{b - k\big(\vec{w}, \vec{S}_{ref}\big)}_{\tilde{b}}\Big). \tag{C.4}$$

Note, that only the bias term has to be modified while the weight vector $\vec{w}$ stays the same. In geometrical terms, the orientation of the hyperplane does not change, but is translated to the new position. In the case of the linear SVM the decision function thus changes to

$$f(\vec{S}) = \mathrm{sgn}\left(\sum_{j=1}^{f} w_j S_j + \tilde{b}\right). \tag{C.5}$$

C.3 Supplementary Figures

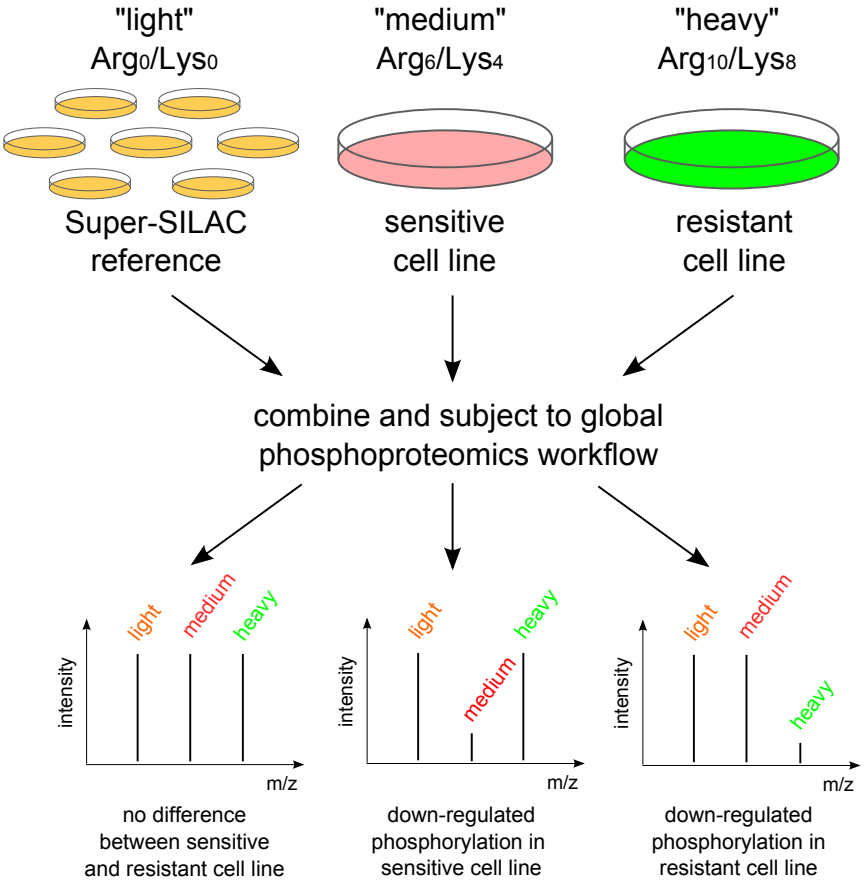

Figure C.2: SILAC labelling diagram. The scheme illustrates how isotopic labelling enables relative quantification of phosphorylation amounts via a spike-in reference (Super-SILAC).

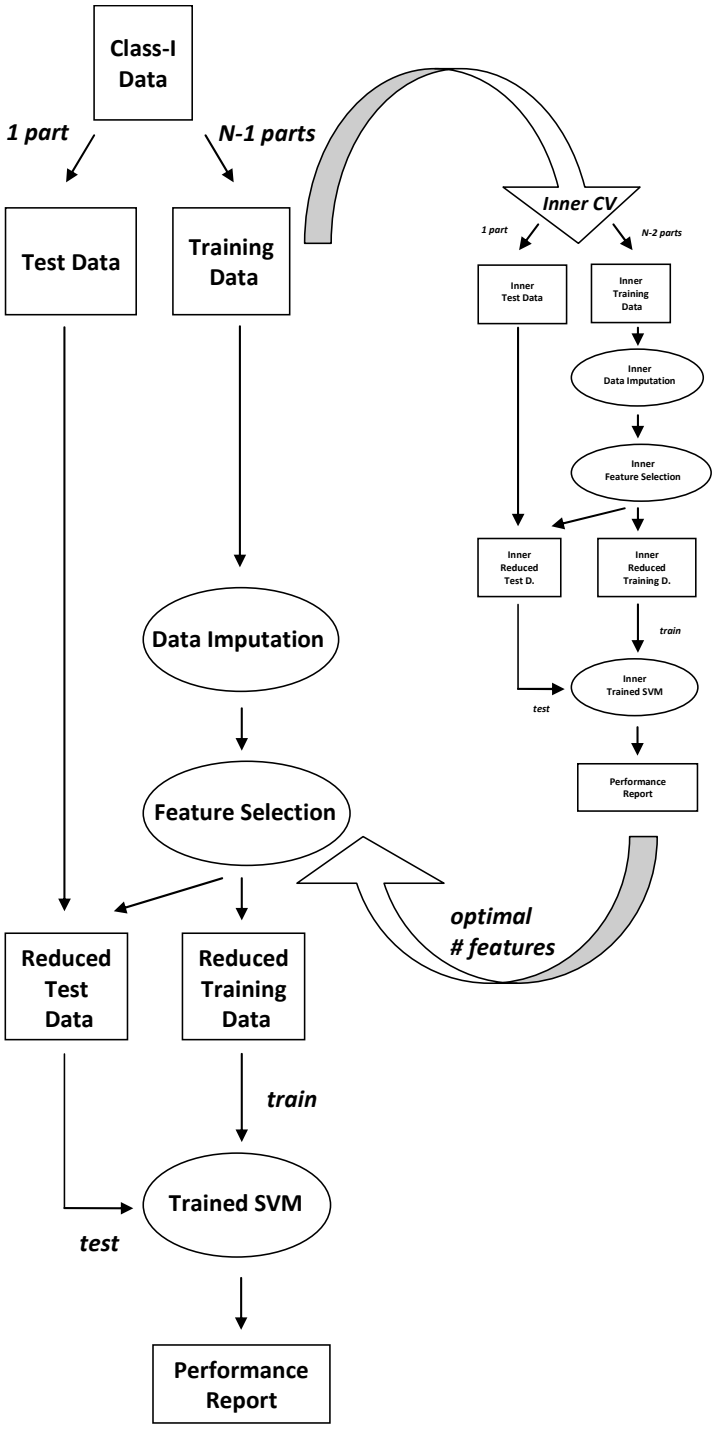

Figure C.3: Workflow diagram for prediction quality assessment, where two cross validation loops are applied. In the inner CV loop the optimal number of features is determined. This number is then used in the feature selection process in the outer CV loop. Subsequently, an SVM is trained and tested with the respective data sets. The prediction results in each outer CV loop are combined and the prediction accuracy is calculated.

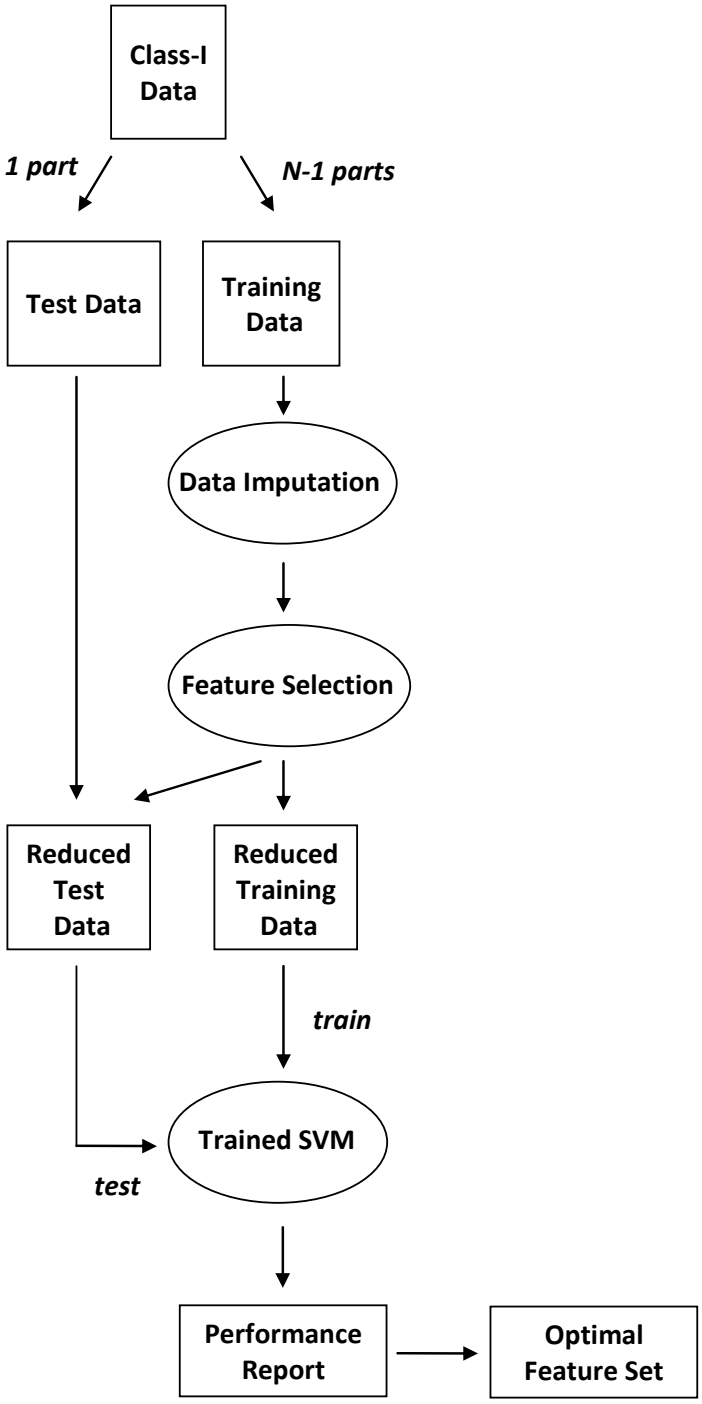

Figure C.4: Workflow diagram for finding the final phospho-signature. The workflow corresponds to one inner CV loop in Figure C.3 resulting in the optimal set of features, which is then used to train the final SVM predictor.

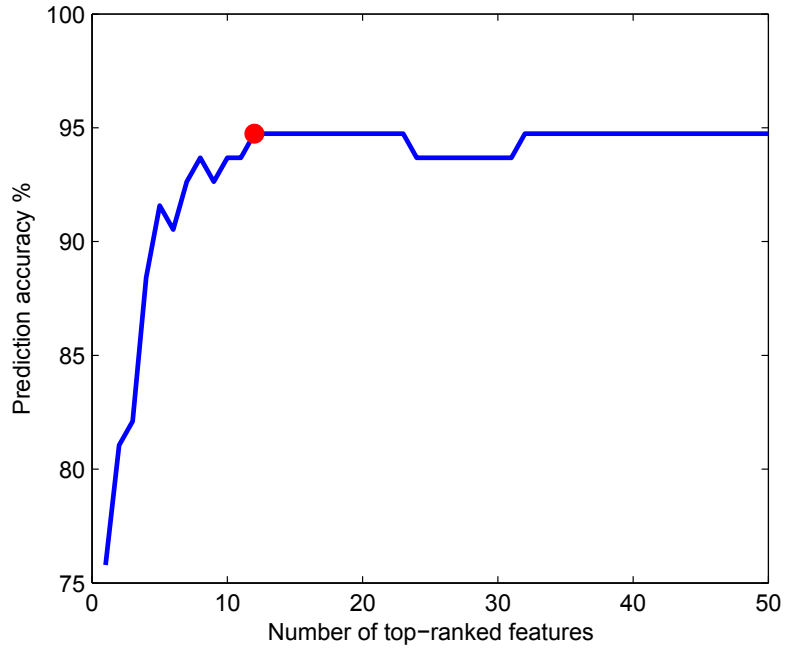

Figure C.5: The prediction accuracy depending on the number of top-ranked features incorporated into the phospho-signature. While the accuracy increased with the first few features, it reached its maximum at 12 features (circle), where it saturated.

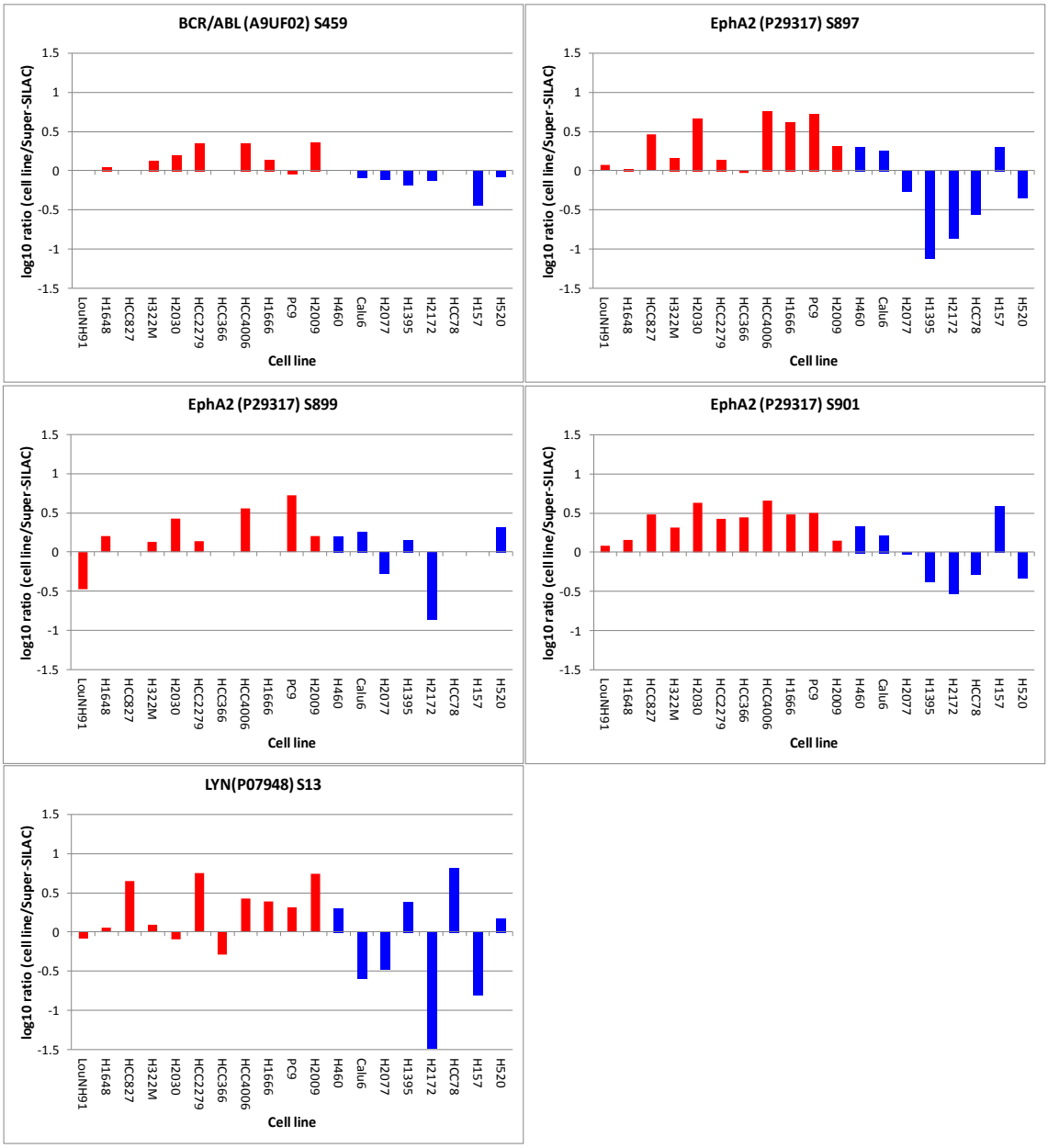

Figure C.6: Bar charts of log10 ratios(cell line/Super-SILAC) of phosphorylation sites on tyrosine kinases quantified in at least two thirds of experiments.

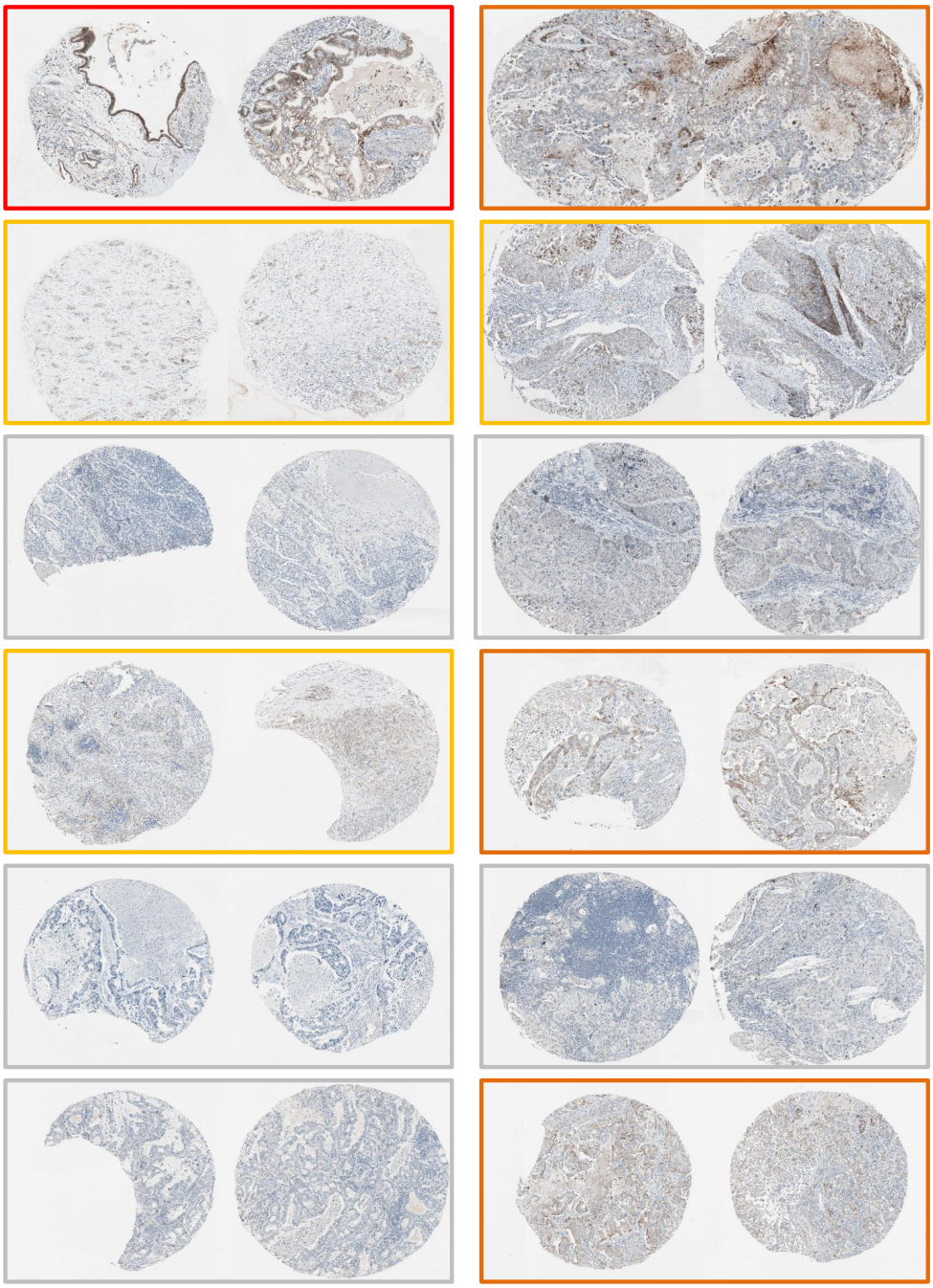

Figure C.7: Immunohistochemical staining of ITGB4 in lung cancer tissue from the Human Protein Atlas. A red border indicates heavy staining, orange moderate, yellow weak and grey no staining.

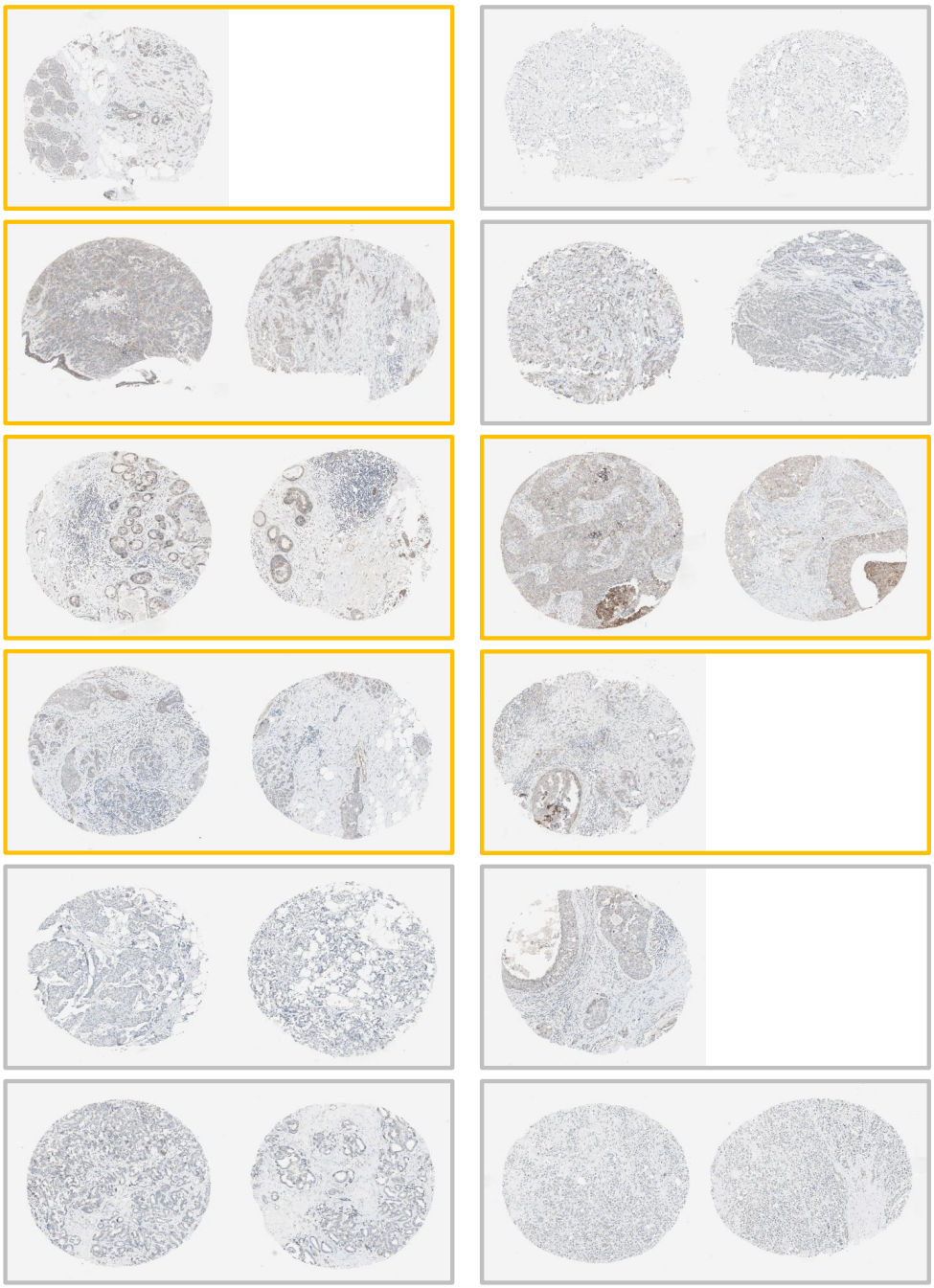

Figure C.8: Immunohistochemical staining of ITGB4 in breast cancer tissue from the Human Protein Atlas. A yellow border indicates weak and grey no staining.

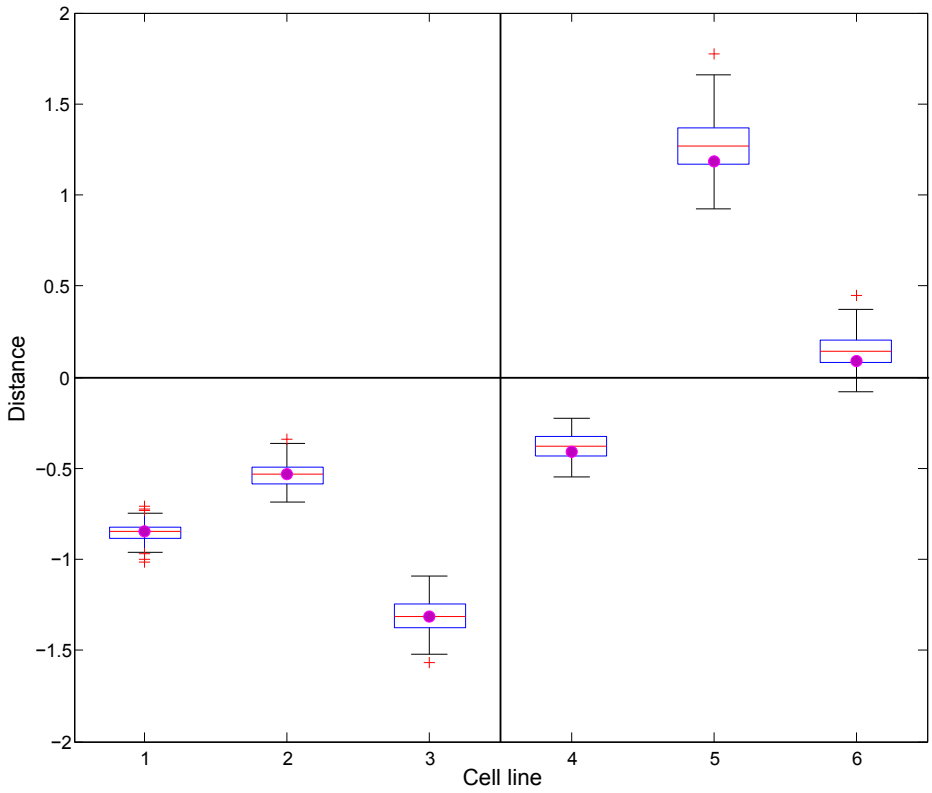

Figure C.9: Effect of the imputation method for the final predictor when applied to the breast cancer samples. Purple dots indicate the classification results with the predictor trained on the mean-imputed NSCLC data, the box plots show the results for the imputation based on 100 samplings from the respective normal distribution of each feature and class.

C.4 Supplementary Tables

Table C.1: Cell line information

Cell line	Indication	Origin	Supplier number	GI_{50}(µM) dasatinib literature[1]	GI_{50}(µM) dasatinib this paper	Class	Valid[2]	TP53 status[3]	Doubling time (h)	GI_{50}(µM) sorafenib [4]literature
Calu6	NSCLC	ATCC	HTB-56	22.54	2.8	-	YES	MUT	25	30
H1395	NSCLC	ATCC	CRL-5868	31.12	4.7	-	YES	WT	50	7.24
H1568	NSCLC	ATCC	CRL-5876	0.8975	5.44	+	no	–	59	6.46
H157	NSCLC	MPI[5]	–	10.54	2.63	-	YES	MUT	25	6.61
H1648	NSCLC	ATCC	CRL-5882	0.0593	0.079	+	YES	MUT	50	6.03
H1666	NSCLC	ATCC	CRL-5885	0.175	0.076	+	YES	WT	30	30
H2009	NSCLC	ATCC	CRL-5911	0.7465	0.085	+	YES	MUT	50	11.09
H2030	NSCLC	ATCC	CRL-5914	0.1183	0.022	+	YES	MUT	25	7.76
H2077	NSCLC	MPI	–	10.07	4.75	-	YES	–	50	5.37
H2172	NSCLC	ATCC	CRL-5930	16.71	5.85	-	YES	–	50	–
H2887	NSCLC	MPI	–	11.3	0.176	-	no	–	40	13.65
H322	NSCLC	MPI	–	0.2588	2.1	+	no	MUT	–	5.43
H460	NSCLC	ATCC	HTB-177	24.16	3.9	-	YES	WT	25	30
HCC827	NSCLC	ATCC	CRL-2868	0.1456	0.033	+	YES	–	43	5.25
H520	NSCLC	ATCC	HTB-182	11.56	1.43	-	YES	MUT	42	4.84
H647	NSCLC	ATCC	CRL-5834	12.39	0.016	-	no	MUT	–	12.45
HCC1359	NSCLC	MPI	–	11.3	0.52	-	no	MUT	30	11.89
LCLC103H	NSCLC	DSMZ	ACC 384	13.9	0.08	-	no	MUT	–	9.66
LouNH91	NSCLC	DSMZ	ACC 393	0.113	0.068	+	YES	–	55	4.68
HCC366	NSCLC	DSMZ	ACC 492	0.482	0.017	+	YES	–	53	6.03
HCC4006	NSCLC	ATCC	CRL-2871	0.8376	0.95	+	YES	–	–	6.46
HCC78	NSCLC	DSMZ	ACC 563	13.9	17.05	-	YES	–	–	11.09
H322M	NSCLC	MPI	–	0.0819	0.311	+	YES	MUT	–	14.13
HOP62	NSCLC	MPI	–	12.76	0.014	-	no	MUT	–	9.44
HCC2279	NSCLC	MPI	–	0.139	0.045	+	YES	MUT	–	12.45
PC9	NSCLC	MPI	–	0.4603	0.02	+	YES	MUT	25	15.85
BT-20	Breast c.	ATCC	HTB-19	0.1652	0.497	+	YES	MUT	–	–
BT-549	Breast c.	ATCC	HTB-122	9.0576	1.71	-	YES	MUT	–	–
MDA-MB-468	Breast c.	ATCC	HTB-132	7.1258	2.8	-	YES	MUT	–	–
MDA-MB-231	Breast c.	ATCC	HTB-26	0.0095	0.036	+	YES	MUT	–	–
MCF7	Breast c.	ATCC	HTB-22	>9.524	3.27	-	YES	WT	–	–
HCC1937	Breast c.	ATCC	CRL-2336	0.07	0.082	+	YES	MUT	–	–

[1] NSCLC data from Sos, *et al.* [94], breast cancer data from Huang, *et al.* [86]

[2] Whether the GI_{50} values from the literature and this paper agree

[3] According to the IARC TP53 database [123] version R15

[4] Data from Sos, *et al.* [94]

[5] Max Planck Institute for Neurological Research (Cologne, Germany)

Table C.2: Mass spectrometric pairing scheme

Exp. number	Group medium	Group heavy	Cell line light	Cell line medium	Cell line heavy
1	+	-	CELLMIX	LouNH91	H460
2	+	-	CELLMIX	H1648	Calu6
3	+	-	CELLMIX	HCC827	[6]LCLC103H
4	+	-	CELLMIX	H322M	H2077
5	+	-	CELLMIX	H2030	H1395
6	+	-	CELLMIX	HCC2279	H2172
7	+	-	CELLMIX	H1568[6]	[6]H647
8	+	-	CELLMIX	H322[6]	[6]HOP62
9	+	-	CELLMIX	HCC366	HCC78
10	+	-	CELLMIX	HCC4006	[6]HCC1359
11	+	-	CELLMIX	H1666	H157
12	+	-	CELLMIX	PC9	H520
13	+	-	CELLMIX	H2009	[6]H2887
14[7]	-	+	CELLMIX	H2077	H322M
15[8]	-	+	CELLMIX	H2887[6]	H2009
16	+	-	CELLMIX	BT-20	MDA-MB-468
17	-	+	CELLMIX	BT-549	MDA-MB231
18	+	-	CELLMIX	HCC1937	MCF7

Table C.3: Additional phosphorylation site information

Gene Name	Site[9]	Canonical Uniprot id[10]	Canonical site[11]	All Uniprot ids[12]	Known [13]site
ITGB4	S1448	P16144	S1518	A0AVL6;B7ZLD5;B7ZLD8;Q0VF97;Q59H46;P16144-2/-4;P16144	YES
BAIAP2	S509	Q9UQB8-5	S509	Q9UQB8-5;B3KPV9	no
ITGB4	S1387	P16144	S1457	A0AVL6;B7ZLD5;B7ZLD8;Q0VF97;Q59H46;P16144-2/-3/-4;P16144	YES
ITGB4	T1385	P16144	T1455	A0AVL6;B7ZLD5;B7ZLD8;Q0VF97;Q59H46;P16144-2/-3/-4;P16144	no
ITGB4	S1069	P16144	S1069	A0AVL6;B7ZLD5;B7ZLD8;Q0VF97;Q59H46;P16144-2/-3/-4;P16144	[14]YES
GPCR5A	S345	Q8NFJ5	S345	A8K556;Q8NFJ5	YES
ITPR3	S916	Q14573	S916	Q14573;Q59ES2;A6H8K3	YES
TNKS1BP1	S429	Q9C0C2	S429	Q9C0C2;B3KXS7	YES
ARHGEF18	S1101	Q6ZSZ5	S1101	Q6ZSZ5;B5ME81;D6W646;Q6ZSZ5-2/-3;A8MV62	YES
IASPP	S102	Q8WUF5	S102	Q8WUF5;Q6ZNZ8	YES
APG16L	S269	Q676U5	S269	Q676U5;Q676U5-3/-4;Q17RG0;Q53SV2	YES
TPD52L2	S141	O43399	S161	O43399;Q6FGS1;Q53GA0;B4DDV4;O43399-2;Q68E05;B4DPJ6	YES

[6]GI50 value inconsistent with the one reported in Sos, *et al.* [94]; cell line was not used in analysis

[7]Label switch of experiment 4

[8]Label switch of experiment 13

[9]As reported throughout the paper

[10]The main Uniprot entry of the corresponding protein

[11]The position in the canonical Uniprot entry

[12]All Uniprot accession numbers from which the corresponding phosphopeptide could originate

[13]According to PhosphoSitePlus (www.phosphosite.org) accessed on 6thAugust 2011

[14]Detected in mouse only

Table C.4: Log10 ratios of cell lines versus SuperSILAC mix

Cell line	Indication	Class	ITGB4 S1448	BAIAP2 S509	ITGB4 S1387	ITGB4 T1385	ITGB4 S1069	GPCR5A S345	ITPR3 S916	TNKS1BP1 S429	ARHGEF18 S1101	IASPP S102	APG16L S269	TPD52L2 S141
LouNH91	NSCLC	+	0.265	0.239				0.192	0.560	0.840	-0.042	0.312		
H1648	NSCLC	+		0.735	0.643	0.507	0.412	0.644	-0.103	0.693	0.074	0.402	0.345	0.393
HCC827	NSCLC	+							0.033	0.032	0.558		-0.194	0.734
H322M	NSCLC	+	0.926	0.645	0.819	0.909	0.852	0.588	-0.070	-0.479	0.085	0.456	-0.118	0.399
H2030	NSCLC	+	0.463		0.377		0.383	0.567	-0.305	0.070	0.439	0.421	0.089	0.746
HCC2279	NSCLC	+	1.012	0.442	0.758	0.656	0.943	0.484	0.194	0.396	0.422		0.124	
HCC366	NSCLC	+	0.896		0.746	0.655	0.818	0.562		0.011	-0.008			0.259
HCC4006	NSCLC	+	0.890	0.261	0.900	0.903		0.603	-0.034	0.044	0.121	0.799	-0.235	0.461
H1666	NSCLC	+	0.717	0.865	0.913	0.690	0.865	0.032		0.130	0.386	0.529	-0.008	0.810
PC9	NSCLC	+		0.296	0.644	0.644	1.101	0.173	0.160	-0.132	0.123	0.399	-0.021	0.580
H2009	NSCLC	+	0.962	0.685	0.996	0.820	1.466	0.172	-0.456	0.138	0.047	0.605	-0.082	0.279
H460	NSCLC	-	-0.142	-0.866				-0.073	-1.025	-0.736	-0.484	-0.429		
Calu6	NSCLC	-		-0.477	-0.421	-0.554	-0.544	-0.223	-0.479	-0.998	-0.467	-0.188	-0.716	-0.597
H2077	NSCLC	-	-0.597	-0.757	-0.609	-0.410	-0.892	-0.349	-0.579	-0.692	-0.411	-0.139	-1.069	-0.367
H1395	NSCLC	-	-0.765		-0.787		-0.857	-0.353	-0.792	-1.086	-0.211	0.058	-1.077	-0.077
H2172	NSCLC	-	-0.705	-0.549	0.042	-0.174		-0.350	-0.839	-0.998	-0.263		-0.381	
HCC78	NSCLC	-	-0.936		-0.049	-0.218	0.257	-0.071		-0.239	-0.334			0.192
H157	NSCLC	-	-0.109	-0.797	-0.233	-0.310	-0.211	-0.971		-0.990	-0.226	-0.040	-0.478	0.017
H520	NSCLC	-		-0.348	-0.189	-0.189	0.029	-0.552	-0.986	-0.807	-0.129	0.051	-0.776	-0.127
BT-20	Breast c.	+		0.585	0.457	0.575	0.478	0.135	0.083	0.295		0.668	-0.008	0.911
MDA-MB-231	Breast c.	+	0.580		0.403	0.432	0.738	0.243	0.547	-0.114			-0.188	-0.431
HCC1937	Breast c.	+	0.495	0.555	0.723	0.685	0.834	-0.487	0.648	0.807	0.092	0.098	0.569	1.252
MDA-MB-468	Breast c.	-		-0.163	0.160	0.290	-0.055	-0.181	-0.147	0.747		0.316	0.327	0.634
BT-549	Breast c.	-	-0.934		-1.239	-1.428	-0.622	-0.296	-0.642	0.153			-0.009	-0.494
MCF7	Breast c.	-	-0.471	0.305	-0.181	-0.114	-0.795	0.127	-0.059	0.231	-0.531	0.177	0.049	0.586

Table C.5: Log10 ratios (cell line versus SuperSILAC) of the non-modified ribosomal peptides used for the alternative normalization

Peptide Seq.	FNADEFEDMVAEK	FTPGTFTNQIQAAFREPR	HGSLGFLPR	HMYHSLYLK	ILDSVGIEADDDRLNK	NIEDVIAQGIGK	TIAECLADELINAAK	VCTLAIIDPGDSDIIR
Name	RPL10	RPSA	RPL3	RPL19	RPLP2	RPLP2	RPS5	RPL30
Uniprot Id	P27635	P08865	P39023	P84098	P05387	P05387	P46782	P62888
LouNH1	0.247	0.243	0.273		0.080	0.157	0.255	0.161
H1648	0.282	0.238	0.220		0.320	0.257	0.255	0.287
HCC827	0.182	0.146			0.056	0.147	0.180	0.130
H322M			0.277	0.306	0.177			
H2030		0.196	0.264	0.308	0.181	0.250		
HCC2279	0.270	0.063	0.219	0.154	0.307	0.232	0.132	0.218
HCC366	0.238	0.158	0.140	0.148	0.295	0.221	0.220	0.259
HCC4006					0.301	0.161		
H1666	0.095	0.121	0.138	0.149	0.151	0.125	0.112	0.147
PC9	0.208	0.143	0.228	0.257	0.218	0.272	0.276	0.282
H2009	0.087	0.174	0.103	0.096	0.223	0.021		
H460	0.426	0.417	0.471		0.252	0.353	0.461	0.361
Calu6	0.340	0.278	0.303		0.397	0.383	0.330	0.339
H2077			0.445	0.456	0.409			
H1395		0.068	0.241	0.344	0.264	0.249		
H2172	0.214	0.198	0.173	0.182	0.302	0.238	0.194	0.149
HCC7	0.203	0.167	0.243	0.280	0.240	0.228	0.211	0.193
H157	0.200	0.189	0.145	0.211	0.210	0.100	0.129	0.111
H520	0.149	0.198	0.205	0.250	0.221	0.217	0.225	0.192

Table C.6: Significantly enriched GO terms

GO term	GO id	Category[15]	q-value[16]
cell communication	GO:0007154	GOBP	2.5E-06
signal transduction	GO:0007165	GOBP	2.5E-06
signal transducer activity	GO:0004871	GOMF	5.1E-04
cytoskeletal protein binding	GO:0008092	GOMF	5.1E-04
small GTPase regulator activity	GO:0005083	GOMF	5.5E-04
regulation of cellular process	GO:0050794	GOBP	7.2E-04
GTPase regulator activity	GO:0030695	GOMF	1.1E-03
protein kinase activity	GO:0004672	GOMF	2.0E-03
protein serine/threonine kinase activity	GO:0004674	GOMF	2.9E-03
protein tyrosine kinase activity	GO:0004713	GOMF	2.9E-03
receptor activity	GO:0004872	GOMF	2.9E-03
phosphotransferase activity, alcohol group as acceptor	GO:0016773	GOMF	3.6E-03
lipid binding	GO:0008289	GOMF	3.6E-03
kinase activity	GO:0016301	GOMF	3.7E-03
actin binding	GO:0003779	GOMF	3.7E-03
zinc ion binding	GO:0008270	GOMF	1.2E-02
Ras protein signal transduction	GO:0007265	GOBP	1.2E-02
protein amino acid phosphorylation	GO:0006468	GOBP	1.2E-02
regulation of cell communication	GO:0010646	GOBP	1.3E-02
transferase activity, transferring phosphorus-containing groups	GO:0016772	GOMF	1.3E-02
regulation of signal transduction	GO:0009966	GOBP	1.7E-02
GTPase activator activity	GO:0005096	GOMF	2.0E-02
cell surface receptor linked signal transduction	GO:0007166	GOBP	2.1E-02
intracellular signaling cascade	GO:0007242	GOBP	2.2E-02
enzyme activator activity	GO:0008047	GOMF	2.2E-02
vesicle-mediated transport	GO:0016192	GOBP	2.4E-02
Rho protein signal transduction	GO:0007266	GOBP	2.5E-02
transport	GO:0006810	GOBP	2.5E-02
establishment of localization	GO:0051234	GOBP	2.5E-02
amine binding	GO:0043176	GOMF	2.7E-02
epidermal cell differentiation	GO:0009913	GOBP	2.9E-02
phosphorylation	GO:0016310	GOBP	2.9E-02
transmembrane receptor activity	GO:0004888	GOMF	3.1E-02
cytoskeleton organization	GO:0007010	GOBP	3.5E-02
locomotory behavior	GO:0007626	GOBP	3.5E-02
transition metal ion binding	GO:0046914	GOMF	3.8E-02
phosphate metabolic process	GO:0006796	GOBP	4.0E-02
Rho guanyl-nucleotide exchange factor activity	GO:0005089	GOMF	4.3E-02
actin filament binding	GO:0051015	GOMF	4.6E-02
post-translational protein modification	GO:0043687	GOBP	4.7E-02

C.5 Additional Files

Additional Files 2-5 can be found at:

http://www.mcponline.org/content/11/9/651/suppl/DC1

[15]GOBP: biological process, GOMF: melecular function

[16]Adjusted p-value

Appendix D

Supplementary Information to Chapter 5

D.1 Supplementary Figures

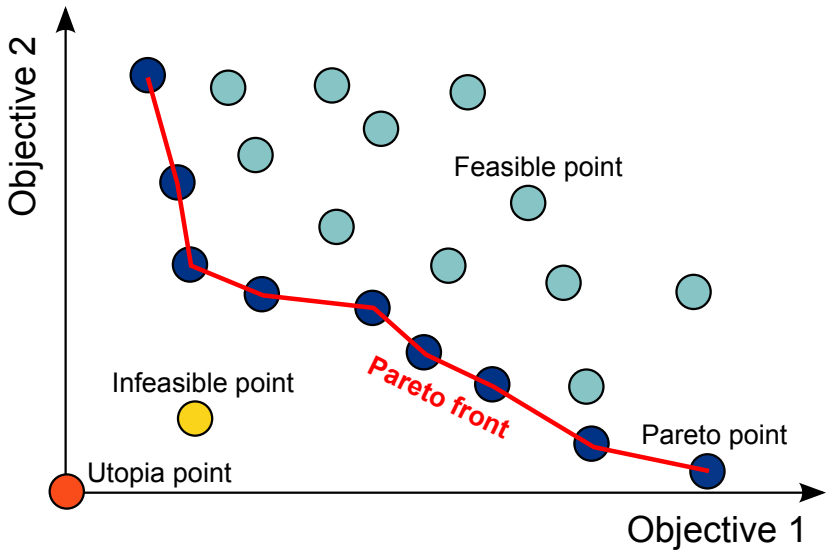

Figure D.1: Example of a Pareto front in a minimization problem. The plot shows different solutions of a toy example. Blue points are feasible solutions, where those that are not dominated by any other solution are referred to as Pareto points (dark-blue). Together they form the Pareto front. The points in the lower left area represent solutions that are desired but not feasible (yellow/red).

References

[1] R. Seger and E.G. Krebs. "The MAPK signaling cascade." In: *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 9 (1995), pp. 726–35 (cit. on pp. 2, 3).

[2] G. Hatzivassiliou et al. "RAF inhibitors prime wild-type RAF to activate the MAPK pathway and enhance growth." In: *Nature* 464 (2010), pp. 431–435 (cit. on p. 2).

[3] N. Jura et al. "Mechanism for Activation of the EGF Receptor Catalytic Domain by the Juxtamembrane Segment". In: *Cell* 137 (2009), pp. 1293–1307 (cit. on p. 2).

[4] M. Rozakis-Adcock et al. "The SH2 and SH3 domains of mammalian Grb2 couple the EGF receptor to the Ras activator mSos1." In: *Nature* 363 (1993), pp. 83–85 (cit. on p. 3).

[5] P. Chardin et al. "Human Sos1: a guanine nucleotide exchange factor for Ras that binds to GRB2." In: *Science* 260 (1993), pp. 1338–1343 (cit. on p. 3).

[6] J. Avruch et al. "Ras activation of the Raf kinase: tyrosine kinase recruitment of the MAP kinase cascade." In: *Recent progress in hormone research* 56 (2001), pp. 127–155 (cit. on p. 3).

[7] R. Sears et al. "Multiple Ras-dependent phosphorylation pathways regulate Myc protein stability". In: *Genes and Development* 14 (2000), pp. 2501–2514 (cit. on p. 3).

[8] D. De Cesare et al. "Rsk-2 activity is necessary for epidermal growth factor-induced phosphorylation of CREB protein and transcription of c-fos gene." In: *Proceedings*

of the National Academy of Sciences of the United States of America 95 (1998), pp. 12202–12207 (cit. on p. 3).

[9] M.C. Mendoza, E.E. Er, and J. Blenis. "The Ras-ERK and PI3K-mTOR pathways: Cross-talk and compensation". In: *Trends in Biochemical Sciences* 36 (2011), pp. 320–328 (cit. on p. 3).

[10] C.A. Hudis. "Trastuzumab — Mechanism of Action and Use in Clinical Practice". In: *The New England journal of medicine* (2007), pp. 39–51 (cit. on p. 3).

[11] K. Sharma et al. "Proteomics strategy for quantitative protein interaction profiling in cell extracts." In: *Nature methods* 6 (2009), pp. 741–744 (cit. on pp. 3, 54, 70, 83).

[12] M. Bantscheff et al. "Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors." In: *Nature biotechnology* 25 (2007), pp. 1035–1044 (cit. on pp. 3, 54, 70, 83).

[13] N. Nagaraj et al. "Deep proteome and transcriptome mapping of a human cancer cell line". In: *Molecular Systems Biology* 7 (2011) (cit. on p. 4).

[14] F.S. Oppermann et al. "Comparison of SILAC and mTRAQ quantification for phosphoproteomics on a quadrupole orbitrap mass spectrometer". In: *Journal of Proteome Research* 12 (2013), pp. 4089–4100 (cit. on pp. 4, 10, 107, 155).

[15] B. Macek, M. Mann, and J.V. Olsen. "Global and site-specific quantitative phosphoproteomics: principles and applications". In: *Annu. Rev. Pharmacol. Toxicol.* 49 (2009), pp. 199–221 (cit. on pp. 4, 5, 54, 90).

[16] J.B. Fenn et al. "Electrospray ionization for mass spectrometry of large biomolecules." In: *Science* 246 (1989), pp. 64–71 (cit. on p. 5).

[17] S.E. Ong et al. "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." In: *Molecular & cellular proteomics : MCP* 1 (2002), pp. 376–386 (cit. on pp. 7, 56, 67).

[18] L.V. DeSouza et al. “Absolute quantification of potential cancer markers in clinical tissue homogenates using multiple reaction monitoring on a hybrid triple quadrupole/linear ion trap tandem mass spectrometer”. In: *Analytical Chemistry* 81 (2009), pp. 3462–3470 (cit. on p. 8).

[19] A. Thompson et al. “Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS”. In: *Analytical Chemistry* 75 (2003), pp. 1895–1904 (cit. on p. 8).

[20] J. Cox et al. “MaxLFQ allows accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction”. In: *Molecular & Cellular Proteomics* (2014), p. M113.031591 (cit. on p. 8).

[21] A. Tebbe et al. “Systematic evaluation of label-free and super-SILAC quantification for proteome expression analysis”. In: *Rapid Commun. Mass Spectrom.* 29.9 (2015), pp. 795–801 (cit. on pp. 8, 108, 155).

[22] A. Michalski et al. “Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer.” In: *Molecular & cellular proteomics : MCP* 10 (2011), p. M111.011015 (cit. on pp. 8, 9).

[23] A. Makarov et al. “Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer”. In: *Analytical Chemistry* 78 (2006), pp. 2113–2120 (cit. on p. 8).

[24] E. De Hoffmann and V. Stroobant. *Mass Spectrometry Principles and Applications*. John Wiley and Sons, 2007, pp. 43–55 (cit. on p. 8).

[25] A. Makarov. “Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis”. In: *Analytical Chemistry* 72 (2000), pp. 1156–1162 (cit. on p. 9).

[26] J.V. Olsen et al. “Higher-energy C-trap dissociation for peptide modification analysis.” In: *Nature methods* 4 (2007), pp. 709–712 (cit. on p. 9).

[27] J. Cox and M. Mann. “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.” In: *Nat Biotechnol* 26.12 (Dec. 2008), pp. 1367–1372 (cit. on pp. 9, 29, 59).

[28] J. Cox et al. “Andromeda: A peptide search engine integrated into the MaxQuant environment”. In: *Journal of Proteome Research* 10 (2011), pp. 1794–1805 (cit. on p. 10).

[29] B. Hutter et al. “Prediction of mechanisms of action of antibacterial compounds by gene expression profiling”. In: *Antimicrob. Agents Chemother.* 48 (Aug. 2004), pp. 2838–2844 (cit. on pp. 10, 72).

[30] Y.P. Lim. “Mining the tumor phosphoproteome for cancer markers”. In: *Clin. Cancer Res.* 11 (May 2005), pp. 3163–3169 (cit. on p. 10).

[31] P.H. Huang et al. “Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma”. In: *Proc. Natl. Acad. Sci. U.S.A.* 104 (July 2007), pp. 12867–12872 (cit. on p. 10).

[32] C. Schaab. “Analysis of phosphoproteomics data.” In: *Methods Mol Biol* 696 (2011), pp. 41–57 (cit. on pp. 10, 45, 48, 54, 90).

[33] S. Weigand et al. “Global quantitative phosphoproteome analysis of human tumor xenografts treated with a CD44 antagonist”. In: *Cancer Research* 72 (2012), pp. 4329–4339 (cit. on pp. 10, 107, 156).

[34] M.A. Cobleigh et al. “Multinational study of the efficacy and safety of humanized anti-HER2 monoclonal antibody in women who have HER2-overexpressing metastatic breast cancer that has progressed after chemotherapy for metastatic disease”. In: *Journal of Clinical Oncology* 17 (1999), pp. 2639–2648 (cit. on pp. 11, 54).

[35] J.S. Ross and J.A. Fletcher. “HER-2/neu (c-erb-B2) gene and protein in breast cancer”. In: *American Journal of Clinical Pathology* 112 (1999) (cit. on pp. 11, 54).

[36] W.J. Catalona et al. “Measurement of prostate-specific antigen in serum as a screening test for prostate cancer.” In: *The New England journal of medicine* 324 (1991), pp. 1156–1161 (cit. on p. 11).

[37] L.J. van ’t Veer et al. “Gene expression profiling predicts clinical outcome of breast cancer.” In: *Nature* 415 (2002), pp. 530–536 (cit. on p. 11).

[38] L.P. Adler et al. “Noninvasive grading of musculoskeletal tumors using PET.” In: *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* 32 (1991), pp. 1508–1512 (cit. on p. 11).

[39] Kalyanmoy Deb et al. “A fast and elitist multiobjective genetic algorithm: NSGA-II”. In: *IEEE Transactions on Evolutionary Computation* 6 (2002), pp. 182–197 (cit. on pp. 12, 93, 95, 96, 103).

[40] M. Klammer et al. “Identifying differentially regulated subnetworks from phosphoproteomic data.” In: *BMC bioinformatics* 11 (2010), p. 351 (cit. on pp. 13, 61, 70, 155).

[41] V.G. Tusher, R. Tibshirani, and G. Chu. “Significance analysis of microarrays applied to the ionizing radiation response.” In: *Proc Natl Acad Sci U S A* 98.9 (Apr. 2001), pp. 5116–5121 (cit. on pp. 14, 35).

[42] C.E. Bonferroni. “Teoria statistica delle classi e calcolo delle prababilita”. In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenzi* 9 (1936), pp. 3–62 (cit. on p. 14).

[43] Y. Benjamini and Y. Hochberg. “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing”. In: *J R Statist. Soc. B* 57 (1995), pp. 289–300 (cit. on pp. 14, 35, 39, 60).

[44] T. Ideker et al. “Discovering regulatory and signalling circuits in molecular interaction networks”. In: *Bioinformatics* 18 Suppl 1 (2002), pp. 233–240 (cit. on pp. 14, 19).

[45] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. “Optimization by Simulated Annealing”. In: *Science* 220 (May 1983), pp. 671–680 (cit. on p. 14).

[46] H.Y. Chuang et al. “Network-based classification of breast cancer metastasis”. In: *Mol. Syst. Biol.* 3 (2007), p. 140 (cit. on pp. 14, 33).

[47] G. Sanguinetti, J. Noirel, and P.C. Wright. “MMG: a probabilistic tool to identify submodules of metabolic pathways”. In: *Bioinformatics* 24 (Apr. 2008), pp. 1078–1084 (cit. on pp. 14, 18).

[48] A. Gelman, J. B. Carlin, and H. S. Stern. *Bayesian data analysis*. Chapman and Hall/CRC, 2004 (cit. on p. 14).

[49] M. Kanehisa and S. Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic Acids Res.* 28 (Jan. 2000), pp. 27–30 (cit. on pp. 15, 60).

[50] L.J. Jensen et al. "STRING 8–a global view on proteins and their functional interactions in 630 organisms". In: *Nucleic Acids Res.* 37 (Jan. 2009), pp. D412–416 (cit. on pp. 15, 16, 29, 61, 70, 86).

[51] A. Alexeyenko and E. L. Sonnhammer. "Global networks of functional coupling in eukaryotes from comprehensive data integration". In: *Genome Res.* 19 (June 2009), pp. 1107–1116 (cit. on pp. 15, 16).

[52] D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989 (cit. on pp. 15, 19).

[53] Y. Zhou et al. "A global approach to identify differentially expressed genes in cDNA (two-color) microarray experiments". In: *Bioinformatics* 23 (Aug. 2007), pp. 2073–2079 (cit. on pp. 15, 21, 22, 27, 29, 35–37, 39, 42, 61, 107).

[54] M. Persico et al. "HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms." In: *BMC bioinformatics* 6 Suppl 4 (Dec. 2005), S21 (cit. on p. 16).

[55] T.S. Keshava Prasad et al. "Human Protein Reference Database–2009 update." In: *Nucleic acids research* 37.Database issue (Jan. 2009), pp. D767–72 (cit. on p. 16).

[56] R.E. Neapolitan. *Learning Bayesian Networks*. Pearson Prentice Hall, 2004 (cit. on p. 18).

[57] D.E. Goldberg and K. Deb. "A comparative analysis of selection schemes used in genetic algorithms". In: *Foundations of Genetic Algorithms*. Morgan Kaufmann, 1991, pp. 69–93 (cit. on p. 19).

[58] P. Shannon et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome Res.* 13 (Nov. 2003), pp. 2498–2504 (cit. on p. 24).

[59] A.L. Barabasi and R. Albert. “Emergence of scaling in random networks”. In: *Science* 286 (Oct. 1999), pp. 509–512 (cit. on p. 25).

[60] C. von Mering et al. “STRING: known and predicted protein-protein associations, integrated and transferred across organisms.” In: *Nucleic acids research* 33.Database issue (Jan. 2005), pp. D433–7 (cit. on p. 31).

[61] M. Klammer et al. “Identification of significant features by the Global Mean Rank test.” In: *PloS one* 9.8 (2014), e104504 (cit. on pp. 34, 94, 97, 155).

[62] T. Geiger et al. “Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins.” In: *Mol Cell Proteomics* 11.3 (Mar. 2012), p. M111.014050 (cit. on p. 34).

[63] J.V. Olsen et al. “Global, in vivo, and site-specific phosphorylation dynamics in signaling networks.” In: *Cell* 127.3 (Nov. 2006), pp. 635–648 (cit. on pp. 34, 54, 67, 90).

[64] E.L. Huttlin et al. “A tissue-specific atlas of mouse protein phosphorylation and expression.” In: *Cell* 143.7 (Dec. 2010), pp. 1174–1189 (cit. on p. 34).

[65] M. Klammer et al. “Phosphosignature Predicts Dasatinib Response in Non-small Cell Lung Cancer.” In: *Mol Cell Proteomics* 11.9 (2012), pp. 651–668 (cit. on pp. 34, 53, 90–92, 94, 96, 100, 103–105, 107, 155).

[66] J.D. Storey. “A direct approach to false discovery rates”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498 (cit. on pp. 35, 51).

[67] G.K. Smyth. “Linear models and empirical bayes methods for assessing differential expression in microarray experiments.” In: *Stat Appl Genet Mol Biol* 3 (2004), Article3 (cit. on p. 35).

[68] Y. Hochberg and A.C. Tamhane. *Multiple Comparison Procedures*. John Wiley and Sons, 1987 (cit. on p. 35).

[69] R. Breitling et al. “Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.” In: *FEBS Letters* 573.1-3 (2004), pp. 83–92 (cit. on p. 35).

[70] Q. Zhu, J.C. Miecznikowski, and M.S. Halfon. "A wholly defined Agilent microarray spike-in dataset." In: *Bioinformatics* 27.9 (May 2011), pp. 1284–1289 (cit. on pp. 36, 43, 44, 46).

[71] F. Hong et al. "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis." In: *Bioinformatics* 22.22 (Nov. 2006), pp. 2825–2827 (cit. on p. 39).

[72] R.C. Gentleman et al. "Bioconductor: open software development for computational biology and bioinformatics." In: *Genome Biol* 5.10 (2004), R80 (cit. on p. 39).

[73] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2012 (cit. on p. 39).

[74] N. Jain et al. "Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays." In: *Bioinformatics* 19.15 (Oct. 2003), pp. 1945–1951 (cit. on p. 39).

[75] J. Cao et al. "Bayesian optimal discovery procedure for simultaneous significance testing." In: *BMC Bioinformatics* 10 (2009), p. 5 (cit. on p. 39).

[76] I. Eidhammer et al. *Computational methods for mass spectrometry proteomics*. John Wiley & Sons, 2007 (cit. on p. 41).

[77] Q. Zhu, J.C. Miecznikowski, and M.S. Halfon. "Preferred analysis methods for Affymetrix GeneChips. II. An expanded, balanced, wholly-defined spike-in dataset." In: *BMC Bioinformatics* 11 (2010), p. 285 (cit. on pp. 45, 49).

[78] C. Weber, T.B. Schreiber, and H. Daub. "Dual phosphoproteomics and chemical proteomics analysis of erlotinib and gefitinib interference in acute myeloid leukemia cells." In: *J Proteomics* 75.4 (Feb. 2012), pp. 1343–1356 (cit. on pp. 45, 47, 51).

[79] P. van der Geer et al. "The Shc adaptor protein is highly phosphorylated at conserved, twin tyrosine residues (Y239/240) that mediate protein-protein interactions". In: *Curr Biol.* 6.11 (1996), pp. 1435–1444 (cit. on p. 47).

[80] A.E. Salcini et al. “Formation of Shc-Grb2 complexes is necessary to induce neoplastic transformation by overexpression of Shc proteins”. In: *Oncogene* 9.10 (1994), pp. 2827–2836 (cit. on p. 47).

[81] F.S. Oppermann et al. “Combination of chemical genetics and phosphoproteomics for kinase signaling analysis enables confident identification of cellular downstream targets.” In: *Mol Cell Proteomics* 11.4 (Apr. 2012), O111.012351 (cit. on pp. 47, 48, 51).

[82] A.W. Liew, N. Law, and H. Yan. “Missing value imputation for gene expression data: computational techniques to recover missing data from available information.” In: *Brief Bioinform* 12.5 (Sept. 2011), pp. 498–513 (cit. on p. 50).

[83] S. Zhang. “A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance.” In: *BMC Bioinformatics* 8 (2007), p. 230 (cit. on p. 51).

[84] J.A. Katzel, M P Fanucchi, and Z Li. “Recent advances of novel targeted therapy in non-small cell lung cancer”. In: *J Hematol Oncol* 2 (2009), p. 2 (cit. on p. 53).

[85] J.M. Reichert and V.E. Valge-Archer. “Development trends for monoclonal antibody cancer therapeutics.” In: *Nat. Rev. Drug Discov.* 6 (2007), pp. 349–56 (cit. on p. 53).

[86] F. Huang et al. “Identification of candidate molecular markers predicting sensitivity in solid tumors to dasatinib: Rationale for patient selection”. In: *Cancer Research* 67 (2007), pp. 2226–2238 (cit. on pp. 54, 80, 86, 90, 131).

[87] H.K. Dressman et al. “An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer”. In: *Journal of Clinical Oncology* 25 (2007), pp. 517–525 (cit. on pp. 54, 90).

[88] P. Blume-Jensen and T. Hunter. “Oncogenic kinase signalling.” In: *Nature* 411 (2001), pp. 355–365 (cit. on p. 54).

[89] B. Kaminska. “MAPK signalling pathways as molecular targets for anti-inflammatory therapy–from molecular mechanisms to therapeutic benefits.” In: *Biochimica et biophysica acta* 1754 (2005), pp. 253–262 (cit. on p. 54).

[90] J. Ferlay, D. M. Parkin, and E. Steliarova-Foucher. “Estimates of cancer incidence and mortality in Europe in 2008”. In: *European Journal of Cancer* 46 (2010), pp. 765–781 (cit. on p. 54).

[91] A. Jemal et al. “Cancer statistics, 2008”. In: *CA: a cancer journal for clinicians* 58.2 (2008), pp. 71–96 (cit. on p. 54).

[92] F.M. Johnson et al. “Phase II study of dasatinib in patients with advanced non-small-cell lung cancer.” In: *Journal of clinical oncology* 28 (2010), pp. 4609–4615 (cit. on pp. 55, 66).

[93] J.N. Andersen et al. “Pathway-based identification of biomarkers for targeted therapeutics: personalized oncology with PI3K pathway inhibitors.” In: *Science translational medicine* 2 (2010), 43ra55 (cit. on pp. 55, 103).

[94] M.L. Sos et al. “Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions.” In: *The Journal of clinical investigation* 119 (2009), pp. 1727–40 (cit. on pp. 55, 57, 66, 78, 131, 132).

[95] C. Schaab et al. “Analysis of high accuracy, quantitative proteomics data in the MaxQB database.” In: *Molecular & cellular proteomics : MCP* 11 (2012), p. M111.014068 (cit. on p. 59).

[96] M.A. Harris et al. “The Gene Ontology (GO) database and informatics resource.” In: *Nucleic acids research* 32 (2004), pp. D258–D261 (cit. on p. 60).

[97] F. Al-Shahrour et al. “From genes to functional classes in the study of biological systems.” In: *BMC bioinformatics* 8 (2007), p. 114 (cit. on pp. 60, 70).

[98] T. Abeel et al. “Robust biomarker identification for cancer diagnosis with ensemble feature selection methods”. In: *Bioinformatics* 26 (2009), pp. 392–398 (cit. on pp. 62, 71, 78).

[99] B. Schölkopf and A.J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, 2002 (cit. on pp. 63, 72).

[100] T. Geiger et al. “Super-SILAC mix for quantitative proteomics of human tumor tissue.” In: *Nature methods* 7 (2010), pp. 383–385 (cit. on p. 67).

[101] S. Ramaswamy et al. “Multiclass cancer diagnosis using tumor gene expression signatures”. In: *Proc Natl Acad Sci U S A* 98 (2001), pp. 15149–15154 (cit. on p. 72).

[102] O. Thuerigen et al. “Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary breast cancer”. In: *Journal of Clinical Oncology* 24 (2006), pp. 1839–1845 (cit. on p. 72).

[103] A.M. Molinaro, R. Simon, and R.M Pfeiffer. “Prediction error estimation: a comparison of resampling methods.” In: *Bioinformatics* 21 (2005), pp. 3301–3307 (cit. on p. 73).

[104] V. Lange et al. “Selected reaction monitoring for quantitative proteomics: a tutorial.” In: *Molecular systems biology* 4 (2008), p. 222 (cit. on p. 79).

[105] R. Hüttenhain et al. “Perspectives of targeted mass spectrometry for protein biomarker verification”. In: *Current Opinion in Chemical Biology* 13 (2009), pp. 518–525 (cit. on p. 79).

[106] M. Uhlen et al. “Towards a knowledge-based Human Protein Atlas.” In: *Nature biotechnology* 28 (2010), pp. 1248–1250 (cit. on p. 81).

[107] R. Buettner et al. “Inhibition of Src family kinases with dasatinib blocks migration and invasion of human melanoma cells.” In: *Molecular cancer research : MCR* 6 (2008), pp. 1766–1774 (cit. on pp. 83, 97).

[108] A.C. Shor et al. “Dasatinib inhibits migration and invasion in diverse human sarcoma cell lines and induces apoptosis in bone sarcoma cells dependent on Src kinase for survival”. In: *Cancer Research* 67 (2007), pp. 2800–2808 (cit. on pp. 83, 97).

[109] F.M. Johnson et al. “Dasatinib (BMS-354825) tyrosine kinase inhibitor suppresses invasion and induces cell cycle arrest and apoptosis of head and neck squamous cell carcinoma and non-small cell lung cancer cells”. In: *Clin Cancer Res* 11 (2005), pp. 6924–6932 (cit. on pp. 83, 97).

[110] M. Dans et al. “Tyrosine phosphorylation of the beta 4 integrin cytoplasmic domain mediates Shc signaling extracellular signal-regulated kinase and antagonizes formation of hemidesmosomes”. In: *Journal of Biological Chemistry* 276 (2001), pp. 1494–1502 (cit. on p. 84).

[111] J. Chung et al. “Integrin ($\alpha 6\beta 4$) regulation of eIF-4E activity and VEGF translation: A survival mechanism for carcinoma cells”. In: *Journal of Cell Biology* 158 (2002), pp. 165–174 (cit. on p. 84).

[112] U. Dutta and L.M. Shaw. “A key tyrosine (Y1494) in the β4 integrin regulates multiple signaling pathways important for tumor development and progression”. In: *Cancer Research* 68 (2008), pp. 8779–8787 (cit. on pp. 84, 102).

[113] E. Tagliabue et al. “Prognostic value of alpha 6 beta 4 integrin expression in breast carcinomas is affected by laminin production from tumor cells”. In: *Clin. Cancer Res.* 4.2 (Feb. 1998), pp. 407–410 (cit. on pp. 84, 102).

[114] S. Lu et al. “Analysis of integrin beta4 expression in human breast cancer: association with basal-like tumors and prognostic significance.” In: *Clinical cancer research : an official journal of the American Association for Cancer Research* 14 (2008), pp. 1050–1058 (cit. on pp. 84, 102).

[115] C. Van Waes et al. “The A9 antigen associated with aggressive human squamous carcinoma is structurally and functionally similar to the newly defined integrin $\alpha 6\beta 4$”. In: *Cancer Research* 51 (1991), pp. 2395–2402 (cit. on pp. 84, 102).

[116] P.V. Hornbeck et al. “PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation”. In: *Proteomics* 4 (2004), pp. 1551–1561 (cit. on pp. 84, 107).

[117] H. Miki et al. “IRSp53 is an essential intermediate between Rac and WAVE in the regulation of membrane ruffling.” In: *Nature* 408 (2000), pp. 732–735 (cit. on p. 85).

[118] A. Yamagishi et al. “A Novel Actin Bundling/Filopodium-forming Domain Conserved in Insulin Receptor Tyrosine Kinase Substrate p53 and Missing in Metas-

tasis Protein". In: *Journal of Biological Chemistry* 279 (2004), pp. 14929–14936 (cit. on p. 85).

[119] A. Blomquist et al. "Identification and characterization of a novel Rho-specific guanine nucleotide exchange factor". In: *Biochem J* 352 Pt 2 (2000), pp. 319–325 (cit. on p. 85).

[120] J. Niu et al. "G Protein betagamma subunits stimulate p114RhoGEF, a guanine nucleotide exchange factor for RhoA and Rac1: regulation of cell shape and reactive oxygen species production." In: *Circulation research* 93 (2003), pp. 848–856 (cit. on p. 85).

[121] D. Bergamaschi et al. "iASPP preferentially binds p53 proline-rich region and modulates apoptotic function of codon 72-polymorphic p53." In: *Nature genetics* 38 (2006), pp. 1133–1141 (cit. on p. 85).

[122] Q. Wu et al. "Integrative genomics revealed RAI3 is a cell growth-promoting gene and a novel P53 transcriptional target". In: *Journal of Biological Chemistry* 280 (2005), pp. 12935–12943 (cit. on p. 85).

[123] A. Petitjean et al. "Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database". In: *Human Mutation* 28 (2007), pp. 622–629 (cit. on pp. 85, 131).

[124] F.R. Luo et al. "Dasatinib (BMS-354825) pharmacokinetics and pharmacodynamic biomarkers in animal models predict optimal clinical exposure". In: *Clinical Cancer Research* 12 (2006), pp. 7180–7186 (cit. on p. 87).

[125] F.R. Luo et al. "Identification and validation of phospho-SRC, a novel and potential pharmacodynamic biomarker for dasatinib (SPRYCEL™), a multi-targeted kinase inhibitor". In: *Cancer Chemotherapy and Pharmacology* 62 (2008), pp. 1065–1074 (cit. on p. 87).

[126] C.I. Herold et al. "Phase II trial of dasatinib in patients with metastatic breast cancer using real-time pharmacodynamic tissue biomarkers of Src inhibition to escalate dosing". In: *Clinical Cancer Research*. Vol. 17. 2011, pp. 6061–6070 (cit. on p. 87).

[127] N.R. Kitteringham et al. “Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics”. In: *Journal of Chromatography B* 877 (2009), pp. 1229–1239 (cit. on p. 87).

[128] P. Ostasiewicz et al. “Proteome, phosphoproteome, and N-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry”. In: *Journal of Proteome Research* 9 (2010), pp. 3688–3700 (cit. on p. 87).

[129] A. Gámez-Pozo et al. “Protein phosphorylation analysis in archival clinical cancer samples by shotgun and targeted proteomics approaches.” In: *Molecular bioSystems* 7.8 (Aug. 2011), pp. 2368–74 (cit. on p. 87).

[130] D. Berg et al. “Use of formalin-fixed and paraffin-embedded tissues for diagnosis and therapy in routine clinical settings”. In: *Methods Mol.Biol.* 785 (2011), pp. 109–122 (cit. on p. 87).

[131] M. Klammer et al. “Pareto Optimization Identifies Diverse Set of Phosphorylation Signatures Predicting Response to Treatment with Dasatinib”. In: *PLoS one* 10.6 (2015), e0128542 (cit. on pp. 89, 156).

[132] T. Passeron et al. “Signalling and chemosensitivity assays in melanoma: Is mutated status a prerequisite for targeted therapy?” In: *Experimental Dermatology* 20.12 (Dec. 2011), pp. 1030–1032 (cit. on p. 90).

[133] C. Schaab et al. “Global phosphoproteome analysis of human bone marrow reveals predictive phosphorylation markers for the treatment of acute myeloid leukemia with quizartinib.” In: *Leukemia* 28.3 (2014), pp. 716–9 (cit. on pp. 90, 103, 108, 156).

[134] J.C. Costello et al. “A community effort to assess and improve drug sensitivity prediction algorithms.” In: *Nature biotechnology* 32 (2014), pp. 1202–1212 (cit. on p. 90).

[135] Y. Cun and H. Fröhlich. “Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches

using prior knowledge on protein-protein interactions". In: *BMC Bioinformatics* 13.1 (Jan. 2012), p. 69 (cit. on p. 90).

[136] Y. Zhu, X. Shen, and W. Pan. "Network-based support vector machine for classification of microarray samples." In: *BMC bioinformatics* 10 Suppl 1 (Jan. 2009), S21 (cit. on p. 91).

[137] J. Roy et al. "Network information improves cancer outcome prediction." In: *Briefings in bioinformatics* 15.4 (July 2014), pp. 612–625 (cit. on p. 91).

[138] C. Winter et al. "Google goes cancer: Improving outcome prediction for cancer patients by network-based ranking of marker genes". In: *PLoS Computational Biology* 8.5 (Jan. 2012), e1002511 (cit. on p. 91).

[139] K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. Chichester, West Sussex, England: John Wiley & Sons, Ltd., 2001 (cit. on p. 91).

[140] C. Nicolaou and N. Brown. "Multi-objective optimization methods in drug design". In: *Drug Discovery Today: Technologies* 10.3 (Sept. 2013), e427–e435 (cit. on p. 91).

[141] W. Gronwald, T. Hohm, and D. Hoffmann. "Evolutionary Pareto-optimization of stably folding peptides." In: *BMC bioinformatics* 9.1 (Jan. 2008), p. 109 (cit. on p. 91).

[142] J.C. Rajapakse and P.A. Mundra. "Multiclass gene selection using Pareto-fronts." In: *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 10.1 (2013), pp. 87–97 (cit. on p. 91).

[143] B. Xue, M. Zhang, and W.N. Browne. "Particle swarm optimization for feature selection in classification: a multi-objective approach." In: *IEEE transactions on cybernetics* 43.6 (Dec. 2013), pp. 1656–71 (cit. on p. 91).

[144] H.T. Lin, C.J. Lin, and R.C. Weng. "A note on Platt's probabilistic outputs for support vector machines". In: *Machine Learning* 68 (2007), pp. 267–276 (cit. on p. 92).

[145] A. Franceschini et al. "STRING v9.1: Protein-protein interaction networks, with increased coverage and integration". In: *Nucleic Acids Research* 41 (2013) (cit. on p. 93).

[146] E.W. Dijkstra. “A note on two problems in connexion with graphs”. In: *Numerische Mathematik* 1 (1959), pp. 269–271 (cit. on p. 93).

[147] J. Knowles and D. Corne. “The Pareto archived evolution strategy: A new baseline algorithm for Pareto multiobjective optimisation”. In: *Proceedings of the 1999 Congress on Evolutionary Computation, CEC 1999*. Vol. 1. 1999, pp. 98–105 (cit. on p. 96).

[148] D.W. Corne, J.D. Knowles, and M.J. Oates. “The Pareto Envelope-based Selection Algorithm for Multiobjective Optimization”. In: *Decision Analysis* 1917 (2000), pp. 839–848 (cit. on p. 96).

[149] M. Zitzler E.and Laumanns and L. Thiele. “SPEA2: Improving the Strength Pareto Evolutionary Algorithm”. In: *Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems*. 2001, pp. 95–100 (cit. on p. 96).

[150] N. Beume, B. Naujoks, and M. Emmerich. “SMS-EMOA: Multiobjective selection based on dominated hypervolume”. In: *European Journal of Operational Research* 181 (2007), pp. 1653–1669 (cit. on p. 96).

[151] A. Konak, D.W. Coit, and A.E. Smith. “Multi-objective optimization using genetic algorithms: A tutorial”. In: *Reliability Engineering & System Safety* 91 (2006), pp. 992–1007 (cit. on p. 96).

[152] C.A. Coello. “Evolutionary multi-objective optimization: a historical view of the field”. In: *IEEE computational intelligence magazine* (2006), pp. 28–36 (cit. on p. 96).

[153] C.A. Nicolaou, Joannis Apostolakis, and C.S. Pattichis. “De novo drug design using multiobjective evolutionary graphs”. In: *Journal of Chemical Information and Modeling* 49 (2009), pp. 295–307 (cit. on p. 97).

[154] B.M. Chung et al. “The role of cooperativity with Src in oncogenic transformation mediated by non-small cell lung cancer-associated EGF receptor mutants.” In: *Oncogene* 28 (2009), pp. 1821–1832 (cit. on p. 102).

[155] H. Dinkel et al. “Phospho.ELM: A database of phosphorylation sites-update 2011”. In: *Nucleic Acids Research* 39 (2011) (cit. on p. 107).

[156] N. Blom, S. Gammeltoft, and S. Brunak. “Sequence and structure-based prediction of eukaryotic protein phosphorylation sites.” In: *Journal of molecular biology* 294 (1999), pp. 1351–1362 (cit. on p. 107).

[157] M. Klammer, S. Roopra, and E.L. Sonnhammer. “jSquid: A Java applet for graphical on-line network exploration”. In: *Bioinformatics* 24 (2008), pp. 1467–1468 (cit. on p. 155).

[158] M. Klammer et al. “MetaTM - a consensus method for transmembrane protein topology prediction.” In: *BMC bioinformatics* 10 (2009), p. 314 (cit. on p. 155).

Acknowledgements

Over the past years I have received support and encouragement from many people. Christoph Schaab has been a great mentor and a sheer endless source of knowledge. Thank you for giving me the opportunity to pursue my dissertation ambitions in the bioinformatics group at Kinaxo/Evotec Munich.

I would like to thank my doctoral supervisor, Daniel Hoffmann, for his guidance, understanding and valuable input despite the geographical distance. As an external doctoral student, I couldn't have wished for a better supervisor.

I am also grateful to Andreas Jenne, former CEO of Kinaxo, for sharing Christoph's support and making my dissertation possible in the first place.

Thanks to Nikolaj Dyboski for establishing the connection with Daniel, and - together with Manuela Machatti - for being extraordinary colleagues and friends, during office hours and beyond.

I am grateful to Andreas Tebbe and Felix Oppermann for always finding the time to share their mass spec knowledge with me, and for all those nights out talking non-science stuff.

I would also like to thank all other colleagues at Evotec Munich - virtually all of you have contributed in some way to this work.

Finally, I want to thank Michaela for her constant support and sharing her life with me.

List of Publications

First author publications

M. Klammer, S. Roopra, and E.L. Sonnhammer. "jSquid: A Java applet for graphical on-line network exploration". In: *Bioinformatics* 24 (2008), pp. 1467–1468 [†]

M. Klammer, D.N. Messina, T. Schmitt, and E.L. Sonnhammer. "MetaTM - a consensus method for transmembrane protein topology prediction." In: *BMC bioinformatics* 10 (2009), p. 314

M. Klammer, K. Godl, A. Tebbe, and C. Schaab. "Identifying differentially regulated subnetworks from phosphoproteomic data." In: *BMC bioinformatics* 11 (2010), p. 351

M. Klammer, M. Kaminski, A. Zedler, F.S. Oppermann, S. Blencke, S. Marx, S. Müller, A. Tebbe, K. Godl, and C. Schaab. "Phosphosignature Predicts Dasatinib Response in Non-small Cell Lung Cancer." In: *Mol Cell Proteomics* 11.9 (2012), pp. 651–668

M. Klammer, J.N. Dybowski, D. Hoffmann, and C. Schaab. "Identification of significant features by the Global Mean Rank test." In: *PloS one* 9.8 (2014), e104504 [‡]

F.S. Oppermann, M. Klammer, C. Bobe, J. Cox, C. Schaab, A. Tebbe, and H. Daub. "Comparison of SILAC and mTRAQ quantification for phosphoproteomics on a quadrupole orbitrap mass spectrometer". In: *Journal of Proteome Research* 12 (2013), pp. 4089–4100 [§]

A. Tebbe, M. Klammer, S. Sighart, C. Schaab, and H. Daub. "Systematic evaluation of label-free and super-SILAC quantification for proteome expression analysis". In: *Rapid Commun. Mass Spectrom.* 29.9 (2015), pp. 795–801 [¶]

[†]Shared first author with S. Roopra
[‡]Shared first author with J.N. Dybowski
[§]Shared first author with F.S. Oppermann
[¶]Shared first author with A. Tebbe

M. Klammer, J.N. Dybowski, D. Hoffmann, and C. Schaab. “Pareto Optimization Identifies Diverse Set of Phosphorylation Signatures Predicting Response to Treatment with Dasatinib”. In: *PLoS one* 10.6 (2015), e0128542

Other publications

S. Weigand, F. Herting, D. Maisel, A. Nopora, E. Voss, C. Schaab, M. Klammer, and A. Tebbe. “Global quantitative phosphoproteome analysis of human tumor xenografts treated with a CD44 antagonist”. In: *Cancer Research* 72 (2012), pp. 4329–4339

C. Schaab, F.S. Oppermann, M. Klammer, H. Pfeifer, A. Tebbe, T. Oellerich, J. Krauter, M. Levis, A.E. Perl, H. Daub, B. Steffen, K. Godl, and H. Serve. “Global phosphoproteome analysis of human bone marrow reveals predictive phosphorylation markers for the treatment of acute myeloid leukemia with quizartinib.” In: *Leukemia* 28.3 (2014), pp. 716–9

Curriculum Vitæ

For reasons of confidentiality, the curriculum vitæ is not included in the online version of this work.

For reasons of confidentiality, the curriculum vitæ is not included in the online version of this work.

Declarations

Erklärung:

Hiermit erkläre ich, gem. § 6 Abs. (2) f) der Promotionsordnung der Fakultäten für Biologie und Geografie, Chemie und Mathematik zur Erlangung der Dr. rer. nat., dass ich das Arbeitgebiet, dem das Thema "Bioinformatical and Statistical Analysis of Mass Spectrometry-Based Phosphoproteomic Data" zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von Martin Klammer befürworte und die Betreuung auch im Falle eines Weggangs, wenn nicht wichtige Gründe dem entgegenstehen, weiterführen werde.

Essen, den _____ _____

Unterschrift eines Mitgliedes der Univ. Duisburg-Essen

Erklärung:

Hiermit erkläre ich, gem. § 7 Abs. (2) c) + e) der Promotionsordnung Fakultäten für Biologie und Geografie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient habe.

München, den _____ _____

Unterschrift des Doktoranden

Erklärung:

Hiermit erkläre ich, gem. § 7 Abs. (2) d) + f) der Promotionsordnung der Fakultäten für Biologie und Geografie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe und dass diese Arbeit von keiner anderen Fakultät/Fachbereich abgelehnt worden ist.

München, den _____ _____

Unterschrift des Doktoranden