Investigation of protein expression dynamics in human tumor cells following pharmacological treatment

Insights from wet and dry lab approaches

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Biotech. Sven Nahnsen
aus Heilbronn-Neckargartach

Tübingen
2010

Tag der mündlichen Qualifikation:	20/12/2010
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr.-Ing. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Alfred Nordheim

This thesis is dedicated to the memory of Andreas Bertsch, my dear
friend, whom I thank for a great time together. Andreas
taught me C++ and instilled endless enthusiasm for
the field of computational proteomics.

Acknowledgements

I would like to express my deepest gratitude to my advisors Oliver Kohlbacher and Alfred Nordheim. They gave me the opportunity to undertake this interdisciplinary PhD project. They encouraged and supported me at any time during my graduate studies; their open-mindedness gave me the confidence to pursue this direction.

I am also very grateful to Boris Macek for sharing his enormous knowledge in mass spectrometry and the great scientific support.

Furthermore, I would like to thank all my colleagues at the Proteome Center, namely Alejandro Carpy, Irina Droste-Borel, Ulrike Grammig, Mirita Franz, Karsten Krug, Johannes Madlung, Raphael Otto, Wolfgang Schütz, Nicole Sessler and Silke Wahl, as well as the former PCT members Michael Beller, Inga Buchen, Claudia Fladerer, Stephan Jung and Stuart Penguelly, for the inspiring atmosphere at the PCT.

Likewise I am very thankful for a great time to all my colleagues at the Center for Bioinformatics, namely Sebastian Briesemeister, Magdalena Feldhahn, Nina Fischer, Sandra Gesing, Erhan Kenar, Andreas Kämper, Peter Niermann, Lars Nilse, Marc Röttig, Timo Sachsenberg, Marcel Schumann, Nora Toussaint and Mathias Walzer, as well as former members Thorsten Blum, Nico Pfeifer and Marc Sturm.

Many thanks go Birgit Schittek and the whole Division of Dermatooncology for a fruitful collaboration and for providing the melanoma cells.

The scientific discussions and close interactions alike with all my colleagues made my time in Tübingen very interesting.

I am thankful to David Wojnar for his work on OpenMS and especially for his contributions to the posterior error probability implementation.

Furthermore I would like to thank Seong-Hwan Rho for the philosophical discussions on systems biology, as well as for his help on the analysis of the DIGE data.

Finally, I would like to wholeheartedly thank my parents and my whole family, who patiently followed my scientific endeavor.

Most importantly I would like to thank Katharina, for her endless patience and for her support during all the years I worked towards this thesis. Thank you for making every day so colorful.

In accordance with the standard scientific protocol, I will use the personal pronoun *we* to indicate the reader and the writer, or my scientific collaborators and myself.

Abstract

Cancer, the multifactorial disease, resulting in uncontrolled growth of malignant cells, is the second most frequent cause of death worldwide. Despite enormous growth in knowledge on cancer pathology, efficient medication still remains elusive. In recent years, global profiling approaches are increasingly important tools to study complex biological problems, such as cancer. One emerging profiling technology is proteomics, the continuously growing research branch of (bio)analytical chemistry that studies the entire set of proteins in a biological system, their modifications and interactions. However, a variety of computational and technological challenges in proteomics are still limiting the broad application of the technology in cancer research.

This thesis contributes in three major topics to new methodological approaches for the analysis of proteomics data and to novel insights of the effects of therapeutical treatment in cancer cells. In the first research part, a new method to analyze 2D-Polyacrylamid Gel Electrophoresis (PAGE) proteomics data is introduced. Although the DIGE (Difference Gel Eletrophoresis) technology greatly influenced the quality of 2D-PAGE experiments through the fluorescent labeling of different samples and their common separation in the same 2D gel, the technology is still accompanied with major challenges. In this thesis we provide a solution to one of the major problems, the accurate and automated mapping of protein spots from different DIGE gels. We implemented a novel scoring method and applied a graph-theoretical approach to solve the assignment problem and to ultimately find the protein spots with reproducible regulation on different gels.

The second research section presents a new method for the integration of several database search engines for improved peptide identification. Database search for peptide identification belongs to the cornerstones in the processing of shotgun proteomics data. The underlying algorithms from different search engines produce results that overlap in parts and disagree in others. Here we present a new computational framework that combines results from several search algorithms and thereby shows significant gain in peptide identification rates. Our method relies on the normalization of single engine scores and on a weighted, average-like method to combine the identification results from different engines to a common consensus score. This new approach to peptide identification yields up to 63% more identifications as the single engines alone.

In the last research section we present the application of quantitative shotgun proteomics to an important aspect of cancer research, the study of the influence of kinase inhibitors to the global protein expression. Dynamic quantitation of protein expression after kinase inhibitor treatment using SILAC (Stable Isotope Labeling by Amino Acids in Cell culture) opened new insights to the quantitative and dynamic effects of the two multi-kinase inhibitors, sorafenib and LY294002, on the whole proteome. In these experiments, we were able to identify and quantify more than 5,400 proteins and to investigate the protein expression levels at five different time points, revealing unprecedented insights to the kinetic behavior of the proteome as a function of length of treatment. We could show that for both inhibitors several clusters of proteins show similar regulation following inhibitor treatment. We confirm the known regulation of the mTor pathway by the LY294003 inhibitor and we speculate about the influence of LY294002 to DNA replication. Furthermore, the investigations on the kinetic effects of sorafenib treatment revealed known mechanisms, such as the influence to the Rho and Ras mediated cell cycle progression, but opened also new and interesting hypothesis, such as sorafenib's contribution to autophagy induction. Large scale proteomics datasets provide a wealth of information and new ways to study biological systems on a system-wide level.

Zusammenfassung

Krebs, die multifaktorielle Krankheit bei der sich pathologisch veränderte Zellen unkontrolliert teilen, ist weltweit die zweithäufigste Todesursache. Trotz des enormen Zuwachses an Wissen über die Entstehung von Krebs, bleiben effiziente Therapiemethoden bislang aus. Globale Profilierungsmethoden haben sich als sehr vielversprechende Ansätze für die Untersuchung von komplexen biologischen Problemen, wie Krebs, erwiesen. Eine dieser neuen Methoden ist die Proteomik, der stetig wachsenden Zweig der (bio)analytischen Chemie, welcher die Gesamtheit der Proteine eines biologischen Systems, sowie ihre Modifikationen und Interaktionen erforscht. Eine Vielzahl von bioinformatischen und technologischen Herausforderungen in der Proteomik verhindern jedoch immer noch den breiten Einsatz dieser Technologie in der Krebsforschung. Im Rahmen dieser Dissertation tragen wir zu drei wichtigen Themengebiete der Proteomik und ihrer Anwendung in der Krebsforschung bei. Wir entwickelten neue methodische Ansätze für die Analyse von proteomischen Daten und wendeten proteomische Methoden an, um ein besseres Verständnis zum Mechanismus von therapeutischen Substanzen in Tumorzellen zu gewinnen.

In dem ersten Teil der Forschungsarbeiten stellen wir eine neue Methode für die Analyse von 2D Gel basierten Daten vor. Obwohl die DIGE Technologie durch die Floureszenzmarkierung von verschiedenen Proben und deren gemeinsame Trennung auf einem Gel, einen erheblichen Beitrag zur Verbesserung der Qualität von 2D Gel Experimenten gemacht hat, gibt es nach wie vor noch erhebliche Herausforderungen in der DIGE basierten Proteomanalytik. Diese Dissertation präsentiert eine neue Lösung für eines der größten Probleme der DIGE basierten

Proteomik, der akkurate und automatisierte Abgleich von Proteinspots auf verschiedenen DIGE Gelen. Die Implementierung einer neuen Scoring-Methode und die Anwendung von graph-theoretischen Ansätzen zur Lösung des Zuordnungsproblems erlauben das schnelle Finden von Proteinspots, welche auf verschiedenen Gelen reproduzierbar reguliert sind.

Das zweite Kapitel der Forschungsarbeiten behandelt eine neue Methode zur Integration von mehreren Datenbanksuchmaschinen zur Verbesserung von Peptididentifizierungsraten. Datenbanksuchen zur Identifizierung von Peptiden gehören zu den Eckpfeilern der Datenprozessierung in der Massenspektrometrie-basierten Proteomik. Die verschiedenen Algorithmen, die den unterschiedlichen Suchmaschinen zu Grunde liegen, annotieren einen Teil der Spektren mit den gleichen Sequenzen, aber schlagen oft unterschiedliche Peptide für einen anderen Teil der Spektren vor. Hier präsentieren wir einen neuen algorithmischen Ansatz, der die Ergebnisse verschiedener Suchmaschinen integriert und dabei einen signifikanten Zuwachs an identifizierten Peptiden erzielt. Unsere Methode beruht auf der Normalisierung der Suchresultate der einzelnen Suchmaschinen und wendet dann ein neues, gewichtetes, dem Durchschnitt ähnliches Maß an, um die verschiedenen Suchresultate zu einem gemeinsamen Konsensus-Score zu verbinden. Dabei konnten wir im Vergleich zu den einzelnen Suchmaschinen bis zu 63 % mehr Peptide identifizieren.

In dem folgenden Kapitel präsentieren wir eine Anwendung von Methoden der Massenspektrometrie-basierten Proteomik zu einer wichtigen Fragestellungen in der Krebsforschung. Hierbei wurde die dynamische Veränderung der Proteinexpression in Tumorzellen nach Behandlung mit Kinase-Inhibitoren analysiert. Mit Hilfe der SILAC-Methode wurde die Wirkung der zwei Multi-Kinase-Inhibitoren Sorafenib und LY294002 in humanen Melanomzellen untersucht. In diesen Experimenten gelang es mehr als 5400 Proteine zu identifizieren und zu quantifizieren. Mit unseren experimentellen Untersuchungen in fünf verschiedenen Zeitpunkten erzielten wir eine bisher noch nie dargelegte Einsicht in die Proteinexpressionsdynamik als Funktion der Inhibitionsdauer. Mit Methoden der Clusteranalyse konnten wir zeigen, dass beide Inhibitoren verschiedene Cluster von Proteinen bilden, die in gleicher Weise reguliert sind. Für die Behandlung mit LY294002 konnten wir die bekannte m-Tor Inhibition bestätigen und neue Hypothesen über den Einfluß des Inhibitors auf die DNA Replikation formulieren. In ähnlicher Weise konnten wir durch die Untersuchungen zur Expressionskinetik nach der Behandlung mit Sorafenib bekannte Mechanismen, wie den Einfluß auf die Ras vermittelte Proliferation betätigen, aber unser Datensatz erlaubt auch neue Hypothesen und ermöglicht neue Einblicke, wie den Einfluß von Sorafenib auf die Induktion der Autophagie.

Umfassende proteomische Datensätze bieten eine Fülle an Informationen und neue Wege ein biologisches System als Ganzes zu verstehen.

List of Tables

List of Figures

Contents

Chapter 1

Introduction

Solving computational and technological problems accompanied with proteomics will be an important step towards the availability of proteomics technologies to non-specialized research groups. The generic setup of a proteomics experiment can potentially open a new and fundamentally different view to a biological system. In contrast to gene-centric, reductionist approaches that arose from the advent of molecular biology, proteomic technologies are not necessarily limited to hypothesis-driven investigations of target molecules, but allow to assess most, if not all proteins in a cellular biological system. This is uniquely important for biological questions facing mechanisms that can not easily be reduced to single cellular players. One such systemic and by far not understood mechanism is the pathology of cancer.

Different forms of cancer belong to the major causes of death worldwide. Decades of cancer research contributed to detailed molecular insights into cancer cells, but the current knowledge did so far not allow to cure the disease.

Drugs that are used to treat cancer patients can be assigned to various categories. Among many other ways of treatment, approaches in cancer therapy include the treatment with monoclonal antibodies or small molecular weight kinase inhibitors. Monoclonal antibodies have mono-specificity to antigens on cancer cells and allow the patient's immune system to recognize and to destroy the cancer cells. Kinase inhibitors in contrast, are used to inhibit major signaling proteins that are known to be involved in the progression of cancer. Kinase inhibitors have had remarkable successes (e.g., imatinib [Gleevec] in the management of chronic myeloid leukemia), but these drugs do also come with a variety of side effects and they are usually only effective in a subclass of patients. Current cancer drugs and small molecular weight kinase inhibitors in particular, interact with comparably large set of target molecules. Although it seems theoretically attractive to approach tumors in several ways, this promiscuity makes it hard to know the full spectrum of activity (Branca, 2005) and to predict therapeutical benefit.

It becomes more and more evident that the understanding of the entire complexity of multiplex diseases, such as cancer and the potential of drugs against these diseases need global analyses (Hwang et al., 2009). A global view of all components in a biological system should allow to integrate known mechanisms with new ideas and hypotheses. Research areas such as drug development and evaluation of drug treatment are especially suited for system-wide approaches, as many influences of drug-target interactions remain unseen by more classical approaches. The emergence of system-wide technologies as tools for global investigations in human cells or tissues was mainly initiated by the sequencing of the human genome (Lander et al., 2001; Venter et al., 2001). The human genome sequence revealed approximately three billion base pairs. Large parts of this collection of information remains to be understood. According to the latest information from the *Swissprot* database (Release 2010_10 of 05-Oct-10), the human genome contains 20,258 protein-coding genes. Large parts of the non-coding regions are occupied by putative RNA genes and pseudogenes. Pseudogenes are deoxyribonucleic acid (DNA) regions that are structured as genes, but are not found to be expressed. Although the definite number of genes is still uncertain, the human genome enables various directions of functional research, the study of genetic disorders, the identification of SNPs (single nucleotide polymorphisms), ribonucleic acid (RNA) and protein expression analysis. The sequence of the genome enables studies on the processes and mechanisms in living organisms and more precisely how they arise from the constituent parts (fundamentally atoms and molecules) that make up an organism (Welsh et al., 2006). Genes are transcribed to messenger ribonucleic acid (mRNA) molecules and mRNA molecules may be translated to proteins. This information flow from DNA to protein is known as the central dogma of molecular biology. The dogma reflects a linear relationship between gene activity, mRNA and protein and is still broadly accepted and taught in basic biological courses. However, the scientific literature accumulates evidence that this linear relationship does not allow to entirely explain the phenotype of an organism with its underlying genome. In fact, completely disconnected expression levels of mRNA and protein molecules have been recently found in *Escherichia coli* (*E. coli*) by Taniguchi et al. (2010). Based on single cell assays, Taniguchi et al. (2010) found that large fluctuations in low-abundance proteins, as well as a common extrinsic noise in high-abundance proteins might be responsible for uncorrelated expression levels of mRNA and protein of the same gene. Immediately after a protein is translated by the ribosome, post-translational modifications can be added to the proteins, resulting in a variety of different proteins, which contribute to a high increase in complexity compared to the set of all genes. Counting all different isoforms of proteins originating from the same gene, it is estimated that their might be more than one million different proteins in a human cell (Jensen, 2006). The increased number of proteins compared to the number of genes is due to the events that can occur during protein expression (transcription + translation), such as alternative splicing or post translational modifications. The number of genes that are expressed as proteins at a given time was determined in studies using budding yeast (de Godoy et al., 2008), where whole proteome measurements were confirmed by the analysis of tagged proteins. Such studies allow to estimate the number of expressed proteins in other organisms. According to Cox and Mann (2007), the number of different genes that are expressed to proteins in humans at a given time is approximately 10,000. The entire protein content of a biological system defines networks of protein-protein interactions and these networks enable intracellular communication, gene expression, structure, metabolism, etc. Fraser and Plotkin (2007) suggested to use information on protein-protein interactions to predict cellular phenotypes. Assuming the cell is predominantly defined by its protein content, observed changes in the phenotype need to be reflected in the proteome. These changes may be triggered by different protein expression levels or modulated post-translational modifications of proteins. In general, proteins, as well as their modifications and interactions, most accurately reflect gene function.

In 1994, the term proteome was introduced for the first time by the Australian scientist Marc Wilkins at a congress in Siena, Italy. The word proteome is a blend of genome and protein and refers to all proteins expressed from a given genome at a particular time in a given cell, tissue, species, etc. In 1997, the term 'proteomics' was coined (James, 1997). Proteomics was initially defined as the general study of the proteome and later expanded as the study of the whole set of proteins, protein isoforms and their modifications, as well as the interactions between them, the structural description of proteins and their higher-order complexes, and almost everything 'post-genomic' related to proteins (Tyers and Mann, 2003). Most proteomics approaches rely on the availability of gene and genome sequences, making proteomics a post-genomic branch of the biological sciences.

Traditionally, proteomics has been based on two-dimensional separation of proteins prior to their identification via mass spectrometry (MS). Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) can theoretically resolve several thousands of proteins and provides the great advantage of permitting the analysis of protein isoforms. Thus the technique has a great potential for in-depth functional analysis of differential mechanisms on the protein level. However, since the invention of 2D-PAGE in 1975 (Klose, 1975; O'Farrell, 1975), researchers were struggling with the enormous work load and technical skills that were necessary to achieve satisfactory results. More recently, 2D-PAGE-based proteomics was facilitated by the introduction of Difference Gel Electrophoresis (DIGE), where different proteomes are labeled with fluorescent tags and several proteomes can be quantified on one gel.

Because of some inherent limitations of the 2D-PAGE technology, gel-free methods for peptide or protein separation have gained popularity (Monteoliva and Albar, 2004). Gel-free methods may include fractionation approaches such as isoelectric focusing to reduce sample complexity. After fractionation and enzymatic digestion, the process that uses enzymes to cut proteins to smaller peptides, the complex peptide mixture is subjected to liquid chromatography (LC)-MS analysis. In combination with stable isotopic labeling, such as stable isotope labeling with amino acids in cell culture (SILAC) or more recently also without labels, gel-free approaches can also be performed in a quantitative manner.

Both DIGE and SILAC-based quantitative proteomics experiments were previously demonstrated to be powerful technologies to get quantitative information on global protein expression. Both techniques offer generic workflows that can be applied to interesting questions in biological and biomedical research.

Proteome-wide studies are faced with a high dynamic range in protein expression and need to be able to resolve the enormous complexity. The cope with this analytical challenge, new experimental methods are continuously developed. Whereas the challenges in 2D-PAGE-based proteomics are mostly related to the reproducibility of protein separation on a gel and the matching between gels, the difficulties that are accompanied with gel-free approaches are of a different nature. Extremely complex peptide mixtures are measured in single LC-MS runs, resulting in thousands of peptide and peptide fragment mass spectra. These mass spectra are then mapped to their corresponding peptides and thereby identify their proteins of origin. This mapping is done by database search algorithms that compare experimental to theoretically calculated candidate spectra. Besides the peptide sequencing (identification), mass spectra can also be used to differentially quantify proteins by comparing peptide intensities. In contrast to the 2D-PAGE approach, the identification and quantification using gel-free approaches can be highly automated.

The field that involves computers for the interpretation of raw MS data, is called computational proteomics. Computational proteomics has become an indispensable part in every proteomics experiment. Typically an LC-MS experiment produces large amounts of raw data (up to several hundred gigabytes per experiment) that need to be stored, processed and analyzed. After the analytical steps a successful proteomics experiment requires thorough computational data analysis. Important steps and challenges in LC-MS data analysis include:

- **Peptide identification**. Peptide sequences are assigned to tandem spectra, either without using previous knowledge (*de novo*) or via database search algorithms. Peptide assignment remains a difficult problem. As shown later, the tools available for the peptide assignment have different underlying algorithms and sometimes produce only partly overlapping results.

- **Feature detection and quantitation**. A feature in this context corresponds to all mass spectrometric information that can be assigned to one peptide. This feature information is compared from sample to sample for relative quantitation and it is compared to internal standards to retrieve absolute quantities. Relative quantitation relies either on stable isotopes that allow a comparison within the same experiment or on label-free comparison of features detected in different experiments. The latter approach requires an accurate map alignment.

- **Map alignment**. If several samples are compared without label, they need to be measured in separate experiments. The different experiments need to be mapped to each other in order to compare quantities of the peptide features. The problem of map alignment emerged in proteomics with the introduction of 2D gel electrophoresis in the 1970ths, where protein features, in this case protein spots, from different gels have to be aligned.

Computational proteomics has become indispensable to generate global protein expression profiles. The development in the area of proteomics is by far not finished. Continuously new mass spectrometric instrumentation is launched, permitting a rapidly growing set of biological problems to be addressed. We need reliable, generic, standardized and especially automatable methods for the analysis, storage and interpretation of proteomics data. Algorithms for peptide identification represent one of the most fundamental parts of the whole analysis. Correct peptide identifications are crucial for both protein identification and quantitation. The importance of peptide identification in complex LC-MS data is reflected in the still growing number of available database search algorithms in the scientific literature. Depending on the protein database and the resolution of the mass spectrometer, the number of candidate sequences might be very high. Highly complex samples may result in noisy spectra and low abundant peptide features. This makes the unambiguous assignment of peptide sequences to peptide fragmentation spectra very difficult. Yet it is known that different algorithmic approaches have different advantages and drawbacks. In short, it has become evident that different search engines overlap in parts, but disagree in other parts.

In this thesis, we developed a novel method for the automated analysis of DIGE gels by implementing a new scoring function that relates differentially regulated protein spots from a DIGE gel to intra-gel landmark spots and uses graph-theoretical methods for matching spots on different gels. With this new method we can significantly improve the quality of the alignment of multiple DIGE gels. The new method can be highly automated, thereby it diminishes subjectivity and reduces the analysis time in DIGE-based proteomics.

Furthermore, we present a method for improved peptide identification based on several search engines. Computational peptide identification via database comparison is one of the most important steps in shotgun proteomics. The newly developed method aims on the combination of different known algorithms for peptide identification and uses a similarity measure to account for missing annotations in a subset of search engine results. With this method peptide identification rates can be increased by up to 63 %.

In the experimental sections, we applied those and other methods to an important aspect in cancer research, the global effect of kinase inhibition. The global influences of the multi-kinase inhibitors sorafenib and LY294002 were investigated using DIGE and SILAC in combination with high accuracy mass spectrometry in malignant melanoma cells. The results of these investigations strongly support the promiscuity of the inhibitors. We found a comparably large set of biologically important mechanisms that are modulated upon treatment. These experiments confirm known aspects, such as sorafenib's influence to Ras and Rho mediated cell cycle progression, but open also novel, unprecedented hypotheses to sorafenib's mechanism. One aspect that is suggested by our data is the influence to autophagy induction. The great asset, gained from time resolved quantitative data, is demonstrated by our LY294002 dataset. We can show that metabolic activity is an early response to the inhibition of major signaling pathways by LY294002, whereas expression levels of proteins involved in metabolic activity to normalize to the initial values at later time points. Integrating results from both inhibitors, we can further speculate about the common influence of sorafenib and LY294002 to oxidative phosphorylation in the mitochondrion.

This following part of this thesis is divided into five section. The adjacent background section will provide the relevant information on cancer pathology, mass spectrometry and it will also discuss methods and techniques relevant to quantitative proteomics and, computational methods for the analysis mass spectrometric data. The following three parts provide the major lines of research, performed throughout the thesis, namely the novel graph-theoretical approach in DIGE-based proteomics, the new method for consensus peptide identification, and the quantitative shotgun proteomics profiling of inhibitor treatment in malignant melanoma. They all follow the same structure: starting with an introduction and presentation of the main ideas and motivation, then the technical aspects, methods and materials are described, followed by a presentation of the results and each Chapters closes by a detailed discussion of the results. The last Chapter provides a general conclusion of the work and perspectives for open problems in computational and experimental proteomics.

Chapter 2

Background

2.1 Cancer

Cancer is the second most frequent cause of death, after to cardiovascular diseases, in western societies. Despite decades of cancer research, a clear understanding of disease pathology and progression is still missing. Molecular biology has contributed extensively to the discovery of aberrant molecular mechanisms associated with various types of cancer. In a normal (healthy) cell, homeostasis and signal propagation are well coordinated and the cell fate is a well-defined balance between proliferation, survival and apoptosis (the programmed cell death). In tumor cells, however, this fate is shifted in favor of proliferation and survival, whereas apoptosis is impaired.

All these information on cancer pathology formed the understanding of cancer as a multiplex disease involving several complex anomalies. Hanahan and Weinberg (2000) defined six hallmarks of cancer (Fig. 2.1): self-sufficiency in growth signals, insensitivity to growth inhibitory signals, evasion of apoptosis, limitless replicative potential, tissue invasion and metastasis, and sustained angiogenesis (Hanahan and Weinberg, 2000). Along these lines, cancer can not be assigned to a single impairment, but is based on multiple abnormalities. These multiple abnormalities sum up to an enormous complexity that makes cancer research and cancer therapy very challenging.

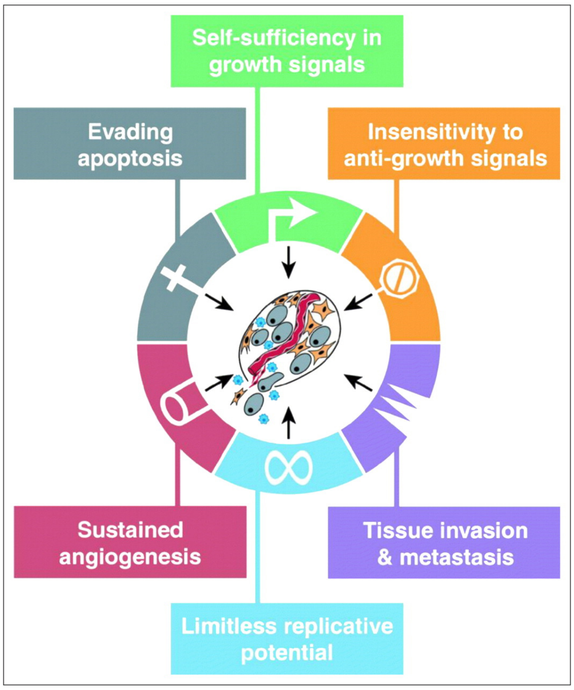

Self-sufficiency in growth signals
Evading apoptosis
Insensitivity to anti-growth signals
Sustained angiogenesis
Tissue invasion & metastasis
Limitless replicative potential

Figure 2.1: Six aberrant cellular mechanisms in cancer [1].

Self-sufficiency in growth signals is an important property of cancer cells. Normal cells require growth signals to proliferate. Cancer cells, however, have acquired the capability to proliferate without those signals. Growth signals are polypeptides that are transmitted from cell to cell. Well-known examples are polypeptides from the FGF family (fibroblast growth factor), the TGF family (transforming growth factor) or PDGF (platelet-derived growth factor). Growth signals are transmitted via transmembrane receptors, through the cytosol to the nucleus where they activate their target genes. Tumor cells can mimic these signals by generating their own growth signals, such as PDGF in glioblastoma (Nistér et al., 1986). In other tumors the expression of transmembrane receptors is up-regulated, which results in hypersensitivity to rather low levels of growth factors (Slamon et al., 1987). An even more complex mechanism, leading to growth factor autonomy, is the modulated expression of proteins that are part of signaling cascades downstream of the receptor, such as the MAPK pathway. For

[1] Reprinted from Hanahan and Weinberg (2000), with permission from Elsevier.

example, the Ras protein, a member of this pathway, has an altered structure in about 25 % of all human tumors (Hanahan and Weinberg, 2000).

Besides the independence of growth signals, cancer cells are frequently also **insensitive to anti-growth signals**. Cellular homeostasis is regulated by a fine interplay between signals promoting growth and their inhibitory counterparts. Theses anti-growth signals are either soluble inhibitors or inhibitors embedded in the extracellular matrix (Hanahan and Weinberg, 2000). A commonly known mechanism of action of anti-growth signals is the transition of cells from a proliferative state in the quiescent G_0 cell cycle state. Anti-growth signals are known to interact with the retinoblastoma protein (pRb). Differentially phosphorylated versions of pRb can alter the function of the E2F transcription factor that is essential for the expression of genes involved in cell cycle progression. If pRb is non-functioning, E2F is released and proliferation is triggered (Weinberg, 1995). Another prominent candidate is the TGF-β protein that coordinates pRb phosphorylation by blocking cyclin:cyclin-dependent kinase (CDK) complexes (Datto et al., 1997).The independence of growth signals and the resistance to inhibitory signals are strong factors for increased life time of cells.

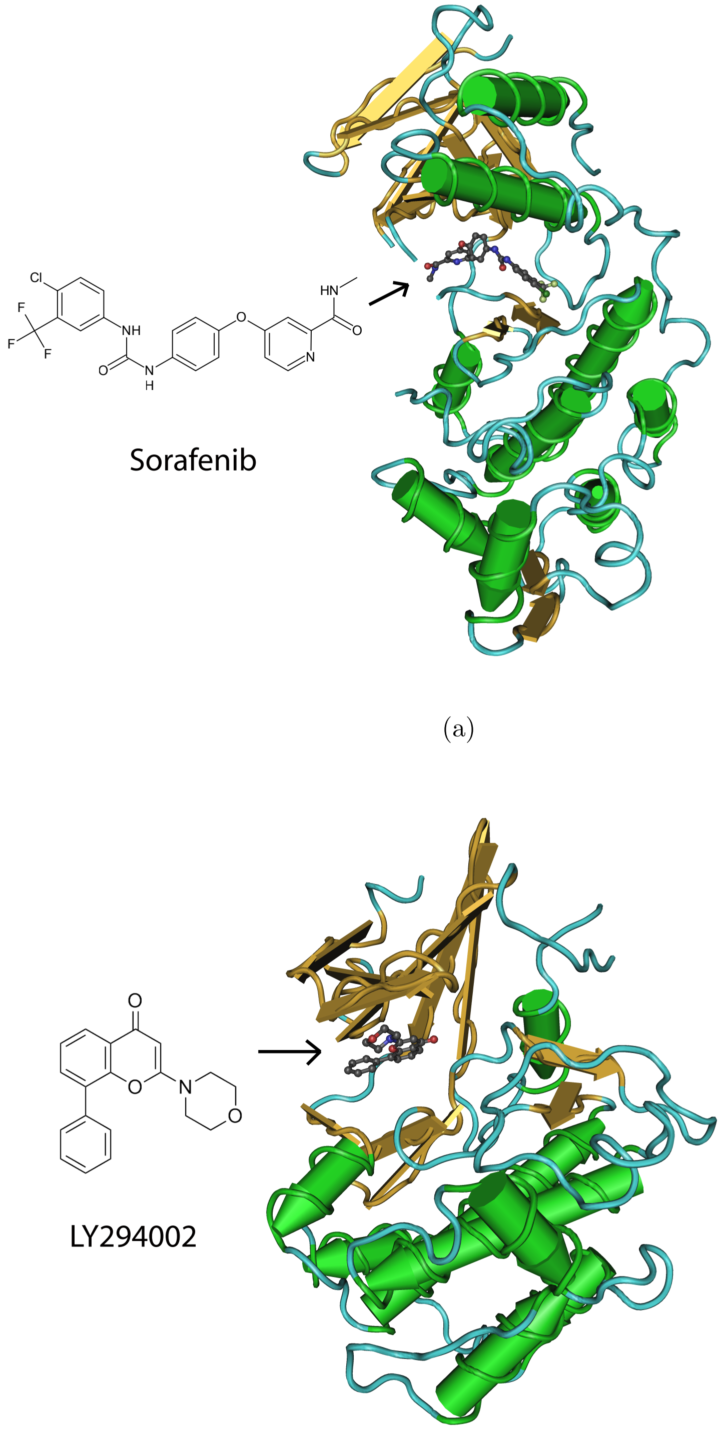
It has been shown that the underlying molecular constituents of cancer cells allow to **evade apoptosis**. The mechanism of apoptosis involves the disruption of cellular membranes, the destruction of intracellular skeletons, the degradation of chromosomes, and the fragmentation of the nucleus (Wyllie et al., 1980). The mitochondrion, where many of the intracellular signals converge, is at the core of this process. Important members of the mitochondrial apoptotic pathway include members of the B-cell lymphoma-2 (BCL-2) family, most prominently the BCL-2-associated X protein (BAX). BAX stimulates the release of cytochrome c from the mitochondrion. The released cytochome c can then activate caspases, triggering cell death.

In contrast to a non pathogenic system, cancer cells have an **unlimited replicative potential**, which is for example expressed by impaired senescence - the loss of the ability to divide. Non-functional pBb and other tumor suppressors, such as p53, can prevent senescence and enable replication.

The two last properties according to Hanahan and Weinberg (2000) concern **angiogenesis and tissue invasion**. Angiogenesis is the process of growing new blood vessels (Weinberg, 2007). Blood vessels are important to support the invading cancer tissue with nutrients and oxygen. Thus formation of new cancer tissue relies on newly formed blood vessels for nutrient and oxygen supply.

2.1.1 Kinase inhibitors for cancer therapy

Many of the deregulated mechanisms described above can be assigned to abnormal activities of kinases, which result in differentially phosphorylated target molecules. These uncontrolled mechanisms will result in dys-regulated gene expression and ultimately in cellular transformation to the cancer phenotype. In 2009, eight kinase inhibitors had been approved for clinical application in the Unites States (Ghoreschi et al., 2009). Examples of FDA-approved kinase inhibitors are sunitinib, sorafenib, imatinib, desatinib, erlotinib, gefitinib and lapatinib. Selectivity has initially been a major attribute for the production of kinase inhibitors. However, recent studies showed that the target spectrum of an ideal kinase inhibitor should include a broad and clearly defined set of kinases (Ghoreschi et al., 2009). This highly demanding task to design the inhibitors in a way that they have an ideal target spectrum is barely fulfilled with current inhibitors. It could be shown that most kinase inhibitors have multiple and widely spread targets (Karaman et al., 2008). Eukaryotic protein kinases change the activity state of their target proteins by adding phosphate groups to serine, threonine, or tyrosine residues. The human genome sequence revealed 518 different protein kinases (Manning et al., 2002). Kinases have a conserved kinase domain, which consists of a five-stranded β-sheet and a single α-helix on the N-terminal lobe connected to a larger C-terminal lobe by a hinge region. The phosphorylation reaction takes place when the kinase binds the protein substrate to a groove formed by the α helical C-lobe. This binding enables the formation of an ATP-binding pocket, where phosphate groups from an ATP molecule are transferred to the hydroxyl groups on the target residues. All kinases share strong structural similarities in these pockets, as they all bind ATP. Kinase inhibitors are designed as competitive antagonists to ATP.



(a)

(b)

Figure 2.2: (A) p38 kinase with bound Sorafenib, (B) Pim-1 kinase with bound LY294002.

The interactions of sorafenib with p38 kinase and the interaction of LY294002 with Pim-1 kinase are shown in Fig. 2.2. p38 is also known as a mitogen-activated kinase. Within both structures the N-terminal β strand and the α helix form the ATP binding pocket, which is now occupied by the inhibitors.

Sorafenib:

Although sorafenib is widely known as a Raf-kinase inhibitor, it has become evident that sorafenib is a promiscuous inhibitor. Besides the inhibition of different isoforms of the Raf kinase, sorafenib is known to inhibit tyrosine kinase receptors, such as EGFR or PDGFR, Flt-3 and c-KIT. Sorafenib, a bi-aryl urea derivative, was shown to interact with Raf-1 (residue 305-648), wtBRAF (residue 409 - 765), V599E-BRAF (residue 409 - 765) in cell-based and biochemical assays (Wilhelm et al., 2004). Sorafenib shows highest affinities to Raf-1 and least to V599E-BRAF (IC_{50}: 6 nM, wt BRAF (22 nM), V599E BRAF (38 nM)). In biochemical assays it inhibits MEK-1 (full length) activity, murine (m)VEGFR-2 (residue 785 - 1367), human VEGFR-2 (KDR kinase domain), mPDGFR-β (residue 560-1098), mVEGFR-3 (residue 818-1363), EGFR (669-1210), Her2/neu (691 - 1255), FGFR-1 (398 - 882) and it has been shown to interact with purified proteins, such as Flt-3 and c-KIT. Wilhelm et al. (2004) also showed that Sorafenib does not influence the activity of IGFR-1, VEGFR-2, c-MET, cdk-1/cyclin B, activated PKB, PKA, LCK, activated c-yes, and pim-1. They also could not observe significant inhibition of MEK-1 and ERK-1 activity. In the work, published by Wilhelm et al. (2004), all experiments were performed at a maximum sorafenib concentration of 10 μM.

In most cell lines, Sorafenib inhibits phosphorylation of ERK1/2, although biochemically it does not interact with ERK. These results suggest that Sorafenib interrupts the MAPK pathway via potent inhibition of RAF. Despite the inhibition of RAF, in some cell lines (e.g., colon cancer cell lines) ERK is still phosphorylated. Suggesting that there are other pathways that lead to ERK phosphorylation as the commonly known Ras-Raf-Mek-Erk pathway (as sorafenib usually inhibits Raf activity).

Sorafenib's affinity to receptor tyrosine kinases is even stronger than the activity to intracellular proteins. EGFR and PDGFR were inhibited at much lower concentrations (20 to 100 nM) compared to the inhibition of the MAPK pathway

(90 to 4000 nM).

LY294002:

LY294002 is known as a potent inhibitor of phosphoinositide 3-kinases. In 1994, it was shown that LY294002 specifically inhibits PI3-kinases (Brunn et al., 1996; Vlahos et al., 1994). The initial paper claimed very high specificity of this inhibitor. However, more recent evidence shows that LY294002 also inhibits the activity of other molecules, such as mTOR (mammalian target of rapamycin) and DNA-PK (DNA-dependent protein kinase 2) (Brunn et al., 1996), CK2 (casein kinase 2) and Pim-1 (Davies et al., 2000). Besides the inhibition of kinases, LY294002 is also involved other mechanisms, such as Ca^{2+} signaling (Tolloczko et al., 2004) or the inhibition of NF-κB activation (Kim et al., 2005). In a chemical proteomics study it was found that the number of interacting proteins for LY294002 is much higher than assumed. 99 proteins were found as interacting partners of LY294002. These include mainly proteins involved in PI kinase activity, such as the PI4 kinases, binding acetylated histones, transferase activity, catalytic activity, binding and kinase activity. LY294002 was shown to influence the expression of metabolite kinases, such as the fructosamine-3 kinase or the phoshofructokinase. Furthermore it was also shown that LY294002 inhibits bromodomain containing proteins (BRDs). BRDs are known to be involved in transcription and LY294002 on the other hand is known for its inhibitory effect on transcription (Gharbi et al., 2007). For cell-based assays previous studies used concentrations ranging from 10 μM to 50 μM of LY294002 for their assays.

2.1.2 Proteome-wide analysis of cancer

The literature on cancer is continuously growing. Integration of data from literature databases is a widely-used strategy for systems biological analyses that aim at capturing the dynamics in cancer cells. However, literature may yield contradictory data on biological systems, which makes data integration very difficult. To cope with the enormous complexity of a living system, molecular biology was for decades focused on single genes and pathways in well-established model systems (cell lines or organisms). With the onset of powerful genomic technologies, this hypothesis-driven approach is being rapidly replaced by global screens of cancer-affected cells or tissues at the genome and proteome level. These data-driven approaches provide an unbiased way of looking at the processes underlying cancer at a global scale. Especially the proteome analysis has proved important in cancer research, as aberrant post-translational modifications (PTMs) of proteins cannot be studied at the genome level. Recent advances in quantitative mass spectrometry have enabled to study global changes in both, protein and PTMs, in space and time. Using proteome wide techniques, protein expression, being a very dynamic process, can be mapped at different time points. This information can be used to interpolate protein expression in parallel for thousands of proteins. This dynamic behavior of biological systems carries important information and needs to be addressed in order to fully understand a biological system.

2.2 Proteomics

2.2.1 General workflows in proteomics

Proteomics is a branch of (bio)analytical chemistry that studies a biological system at the level of gene products (proteins). Two cornerstones of every proteomics experiment are protein identification and protein quantification, which is mainly performed by mass spectrometry. Starting from the biological sample, all proteins need to be extracted and separated from the other content of the cells. A list of three major workflows, accompanied with a more detailed comparison of the two workflows used in this thesis are illustrated in Fig. 2.3.

2D-PAGE-based proteomics is the oldest of the methods presented here. Over 30 years ago 2D gel electrophoresis was used the first time to resolve complex mixtures of proteins (Klose, 1975; O'Farrell, 1975). In 2D-PAGE, proteins are separated according to their pI value and their molecular weight. After separation, the gels are stained and their patterns are compared to patterns from different conditions. After staining and visual inspection of the protein patterns, interesting protein spots are excised, digested and subjected to a tandem mass spectrometer, where a similar identification workflow as described below for shotgun proteomics, leads to the sequence information of the peptides and protein.
Top-down proteomics, in contrast, refers to an approach where short intact proteins are ionized and the m/z values of the entire protein is recorded, as well as fragment masses that result from the fragmentation of the whole protein ion. The major limitation of the top-down approache is the maximal molecular weight of the analytes. Usually only proteins up to 50 kDa can be measured in this way (McLafferty et al., 2007). However, the top-down technology has the potential to analyze multiply modified protein isoforms more accurately as the bottom-up method, as proteolytic digest destroys the mass information of the entire protein.
Shotgun proteomics is named after the shotgun approach for DNA sequencing. In shotgun DNA sequencing long DNA sequences are computationally reconstructed from many short sequencing reads (Marcotte, 2007). Shotgun proteomics identifies proteins from their shorter fragments: peptides. Bottom-up and shot-

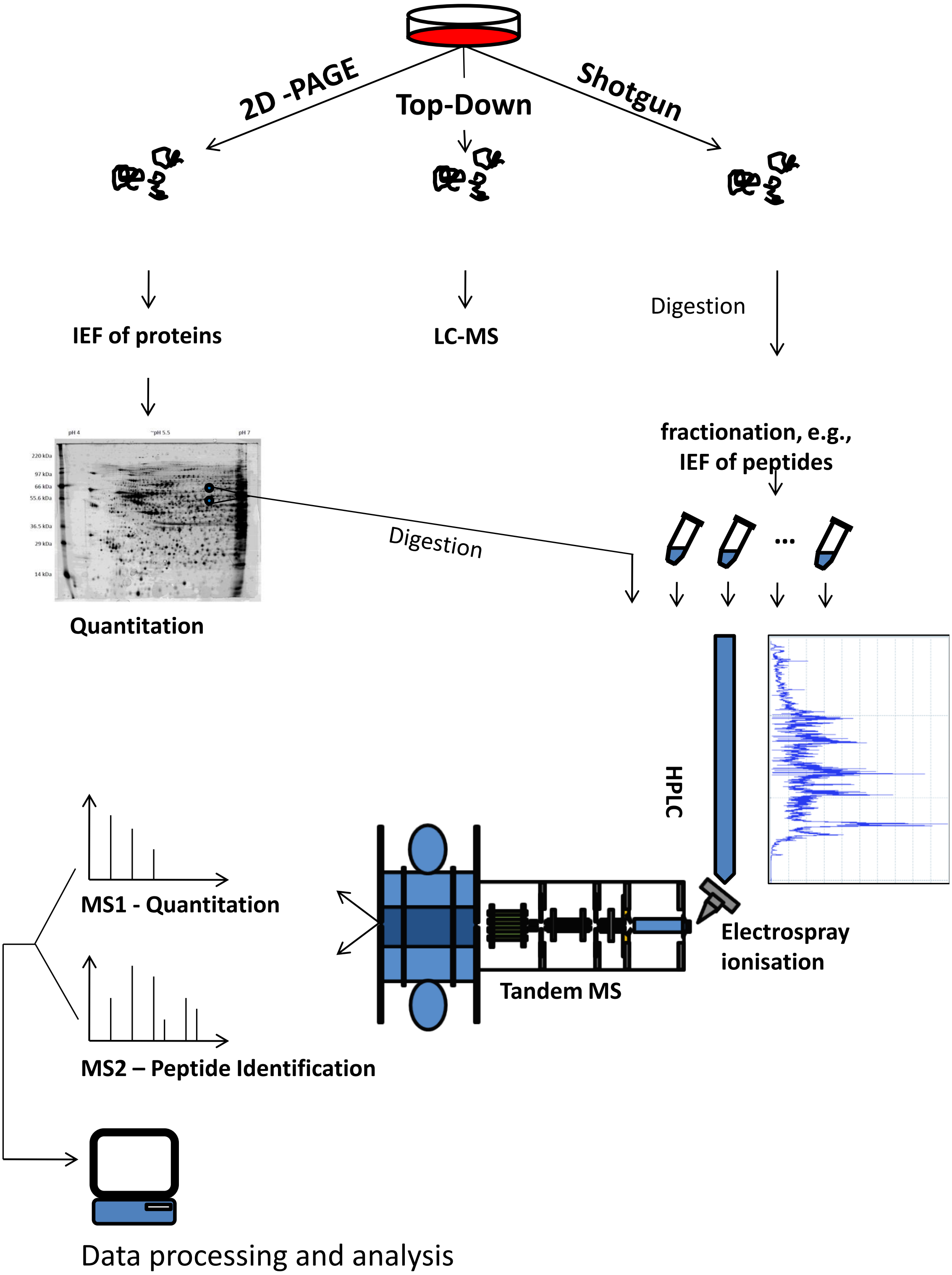


Figure 2.3: The three most commonly used workflows in proteomics. IEF abbreviates isoelectric focusing.

gun proteomics are often used to describe the same workflow. A main step in the sample preparation for shotgun proteomics experiments is the use of proteases to digest proteins to peptides. These proteases usually follow specific cutting rules (motifs or residues), where they cut the protein. In this way, complex mixtures of thousands of proteins are digested to hundreds of thousands of different peptides. These even more complex peptide mixtures are then separated according to physicochemical properties and subjected to mass spectrometry. In the mass spectrometer peptide ions are isolated and fragmented and the mass spectra are recorded. Following the mass spectrometric experiment, all mass spectra are computationally compared to a given database of proteins and the best matching sequence is assigned to each spectrum.

2.2.2 2D-PAGE

Traditionally large scale protein analysis has been performed on the basis of two-dimensional gel electrophoresis (2D-PAGE). This method has a high resolution power for intact proteins. Gauss et al. (1999) showed that the method allows resolving almost 9,000 spots at a time. In the first dimension proteins are separated according to their isoelectric point. The isoelectric point of a protein is the pH-value where the net charge of the protein is zero. In the second dimension the proteins are separated with respect to their molecular weight, according to the method described by Laemmli (1970). In 1975, it was independently shown by O'Farrell (1975) and Klose (1975) that isoelectric focusing can be used to separate intact proteins, which was the birth of 2D-PAGE-based protein analytics. To date 2D-PAGE is very well established in many laboratories. Despite the technical development in the field (e.g., immobilized pH gradients), there are several remaining drawbacks in 2D-PAGE, such as low reproducibility, limited dynamic range, high sensitivity to contaminants and most importantly a lack of automation methods. Nearly all steps of this workflow and especially the analysis of the 2D images require skilled user interventions. Due to this strong need for

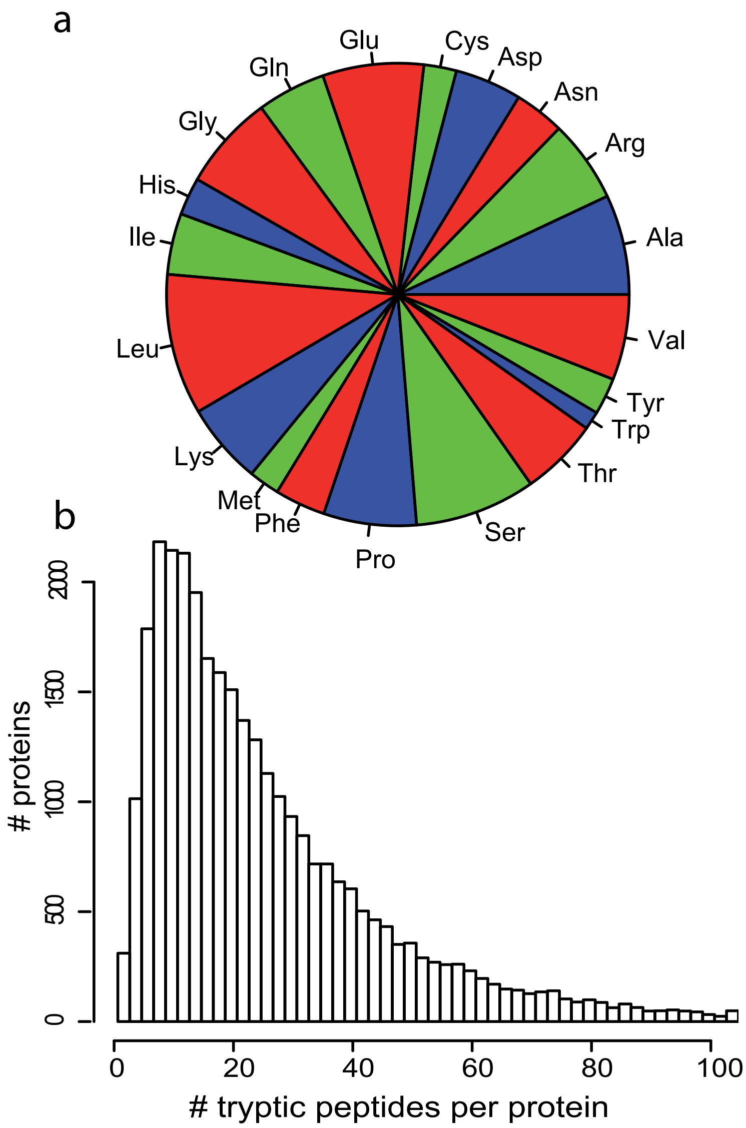


Figure 2.4: (a) Amino acid abundance in the human proteome, (b) Number of tryptic peptides per protein in the human proteome.

manual handling, the acquisition of reproducible results by 2D gel electrophoresis is a very demanding task.

2.2.3 Shotgun proteomics

In shotgun proteomics the proteins are immediately digested to peptides by a protease. Two-dimensional gels and top-down approaches quantify intact proteins. While 2D-PAGE is still frequently used, inherent analytical problems of this technique made shotgun proteomics the method of choice, especially for system-wide global analyses. In shotgun proteomics the mixture of proteins is typically cleaved by proteases. The most commonly used protease is trypsin. Trypsin cuts after K (lysine) and R (arginine), except if the following amino acid is P (proline). Fig. 2.4 shows the result of an *in silico* tryptic digestion of the human IPI database (Kersey et al., 2004), in version 3.64. The average length of a tryptic peptide from this database is 9.14 amino acids. Considering only peptides with at least six amino acids, which would correspond to the minimal peptide length that is considered to provide specific protein identification, the average length increases to 15.2 amino acids. (A) shows the abundance of the 20 amino acids in all sequences of the human proteome and (B) shows the distribution of the human proteins as a function of their theoretically calculated number of tryptic peptides. These data show that lysine and arginine are moderately abundant amino acids. 5.6 % of all amino acids are lysine residues and 5.7 % of all amino acids are ariginine residues. In fact the trypsin cutting definition leads to a high number of specific peptides that match only one protein. The human IPI database, in version 3.64, contains 84,295 protein entries. Even in this database with a lot of isoform information tryptic peptides from 74,925 proteins can be found that fit only to one protein. Besides trypsin, Lys-C (cuts C terminal of lysine) or Asp-N (cuts N-terminal of aspartic acid) are other examples for frequently used proteases in proteomics.

2.2.4 Chromatographic separation

Chromatographic separation is the most commonly used method to reduce the complexity of peptide mixtures. The term high performance liquid chromatography (HPLC) refers to a liquid chromatography (LC) experiment with relatively high pressures. During liquid chromatography, peptides are subjected to a column and the separation is based on different physicochemical properties of the linear amino acid chains. The stationary phase of an LC separation consists of particles which interact with the analytes (peptides) and the mobile phase can be an aqueous solution that is usually a mixture of water and organic solvent, such as acetonitrile (ACN). After loading the peptide mixture onto the column, the analytes are eluted from the column in an increasing concentration gradient of the mobile phase. The mobile phase has different physicochemical properties as the stationary phase and by changing the concentration fraction of the mobile phase, due to their hydrophobicity properties, different peptides will have different affinities to either the mobile or the stationary phase and thus they will elute at different times. Liquid chromatography can be done in various ways. The most important methods for proteomics approaches are ion exchange chromatography and reversed-phase chromatography. In reversed-phase chromatography the stationary phase usually consists of linear unpolar carbon chains, such as $C_{18}H_{37}$ (C18) chains. During an LC run the composition of the mobile phase is continuously changed by increasing the percentage of organic solvent. With this setup hydrophilic peptides will elute earlier than the more hydrophobic peptides.

2.2.5 Ionization

The analysis of biomolecules by mass spectrometric methods had only become possible through the invention of soft ionization methods. MALDI (Matrix-Assisted-Laser-Desorption-Ionization) (Karas and Hillenkamp, 1988; Tanaka et al., 1988) and ESI (electrospray ionization)(Whitehouse et al., 1985) are the major ionization methods for biomolecules. Both technologies were awarded Nobel prizes in 2002: John Fenn was honored for his work on electrospray ionization and Koichi Tanaka was given the prize for the invention of MALDI. The principle of MALDI relies on a laser beam that triggers the ionization of analytes when fired onto a solid matrix of small organic molecules wherein the analytes are embedded. In ESI, the analytes in aqueous phase are ionized by pushing through a very small charged metal capillary at atmospheric pressure. The metal capillary points to the orifice of the mass spectrometer. A high voltage between the end of the thin capillary and the orifice of the mass spectrometer induces the formation of charged droplets (carrying analyte molecules) and a high orifice temperature induces extensive solvent evaporation. Thereby ions are formed through charge-charge repulsion of the droplets that emerge from the capillary. This repulsion takes place when the droplet reaches its Rayleigh limit - the maximal number of equally oriented charges that can be carried by a droplet. This process, called Coulomb fission, continues until analyte ions are free of solvent. The charged ions are then accelerated towards the onset of the mass spectrometer. Fig. 2.5 shows the principle of electrospray ionization.

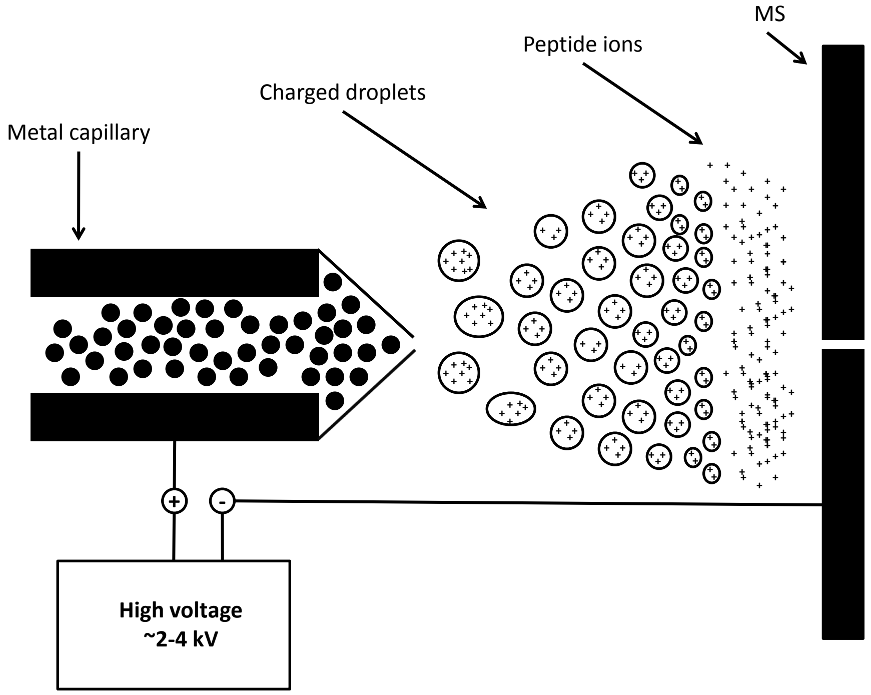


Figure 2.5: Electrospray ionization (ESI) process.

2.2.6 Mass spectrometry

A mass spectrometer measures the mass-to-charge (m/z) ratio of ions in the gas phase. This analysis can be performed by various mass analyzers. Although the principles of operation can be very different, all mass spectrometers consist of an ion source, a mass analyzer and a detector. The most commonly used mass analyzers in proteomics are time-of-flight (TOF), quadrupole and since recently also orbitrap analyzers. TOF instruments measure the time an ion takes to fly until it reaches the detector. Quadrupole mass analyzers use oscillating electrical fields to analyze ion trajectories. Other mass spectrometers have sector field, Fourier transform ion cyclotron or orbitrap mass analyzers. In early times of MS based proteomics, m/z values of peptides, detected by single MS measurements, were used to identify proteins. This approach is called peptide mass fingerprinting (PMF). However, some peptides, may have the same molecular weight while having different sequences. The emergence of tandem MS (MS/MS) enabled the accurate identification of peptides and proteins. This is usually performed by tandem mass spectrometers. Tandem mass spectrometry can be performed *in time*, using the same mass analyzer to record the peptide masses (precursor ions) as well as the fragment masses from the peptides after fragmentation in the collision cell or *in space*, if one mass analyzers detects the precursor ions and a second mass analyzer detects the fragment masses. If two mass analyzers are used and they are based on different types of analyzers (e.g., linear ion trap and orbitrap), the setup is called hybrid mass spectrometer. There are various types of mass analyzers that may be combined to hybrid MS instruments and used for tandem MS. In this thesis mainly the LTQ-Orbitrap mass spectrometer was used. We will now discuss details of tandem mass spectrometry at the example of the LTQ-Orbitrap. The LTQ-Orbitrap is a hybrid mass spectrometer, consisting of a linear ion trap (LTQ), a radio frequency (RF)-only (curved) C-trap and an orbitrap mass analyzer. The principle of tandem mass spectrometry in the LTQ-Orbitrap starts with the collection of ions in the linear ion trap, followed by axial injection of the ions into the C-trap. From the C-trap the ions are injected into the orbitrap, where an electro static potential allows accumulating ions while they are axial oscillating and rotating around a central electrode. A scheme of the LTQ-Orbitrap is shown in Fig. 2.6

Linear ion trap

Ions, generated by the electrospray ionization (ESI) source are transferred through the ion transfer capillary and injected into the first ion guide, consisting of a square quadrupole. Therefore, all ions have to pass the skimmer, which acts as a vacuum baffle. An RF voltage applied to the rods of the quadrupole pushes the

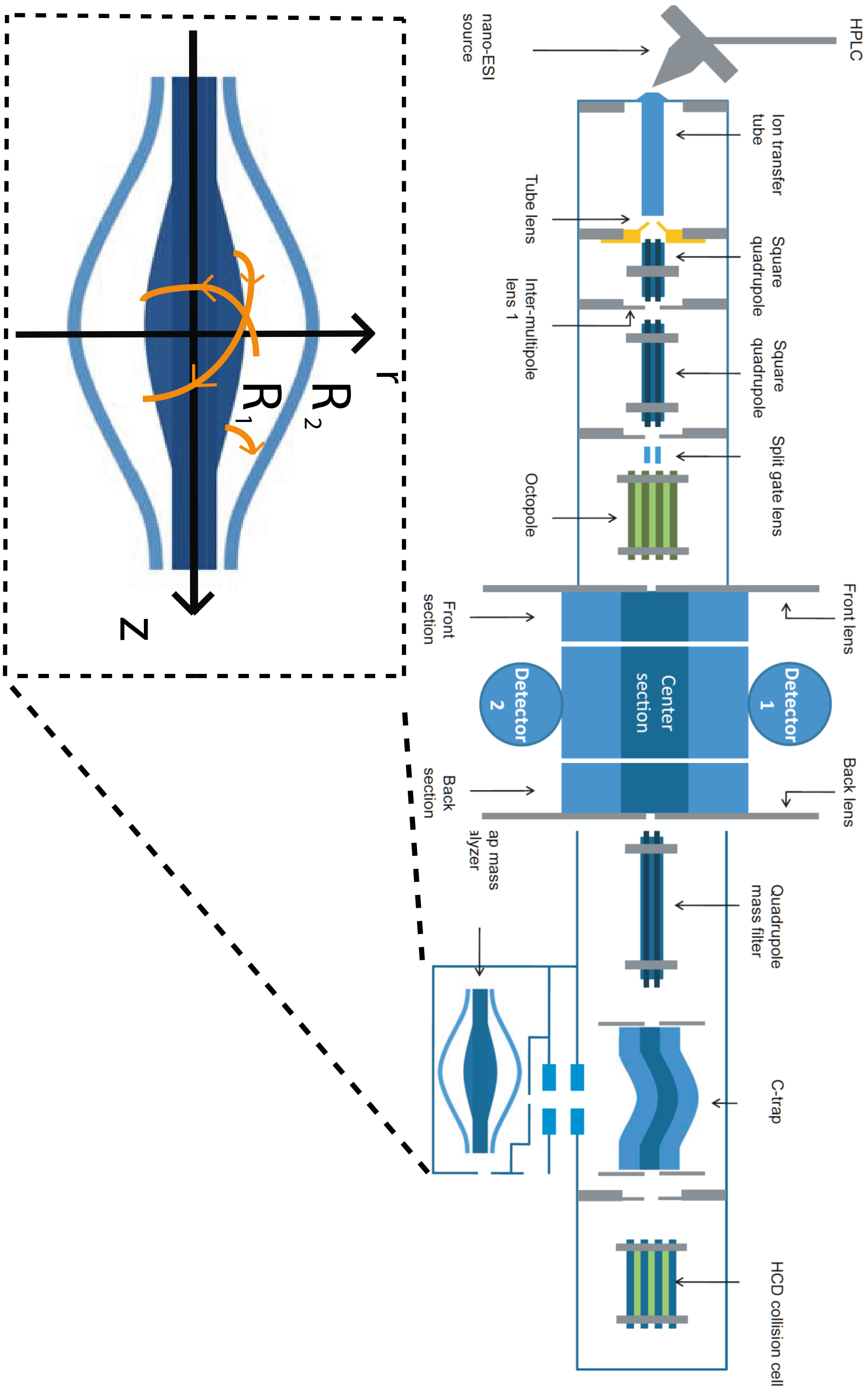


Figure 2.6: Schematic overview of an LTQ-Orbitrap XL mass spectrometer. Zoom-in illustration of an ion trajectory in the orbitrap analyzer.

ions along the axis of the quadrupole. The kinetic energy of the ions is regulated by an offset voltage applied from ground to the quadrupole (LTQ-Manual, 2006). The next ion guide consists of the second square quadrupole, which serves as a further guidance of the ions. The last ion guide before the mass analyzer consists of an octopole with similar functions as the previous quadrupoles. After the main ion guide the ions are injected into the center section of the linear ion trap, which is surrounded by the front and back lenses. The front and back lenses mainly serve as conductance limits. Specific potentials are applied in the center section for ion storage. Helium is injected into the center section through a gap between the quadrupole rods. During collision-induced-dissociation (CID), the kinetic energy of ions is increased and through the higher-energy collisions of parent ions they will dissociate to one or more product ions. To detect the ions stored in the linear ion trap electron multipliers are used. An electron multiplier consists of a cathode and an anode. The anode collects the electrons produced by the cathode and the electron flow can be measured. This current is proportional to the number of ions ejected from the center section.

C-trap

On the way from the linear ion trap to the orbitrap, ions move through the nitrogen filled curved linear trap (C-trap). The nitrogen is used for lowering the kinetic energy of the ions injected from the linear ion trap. From the C-trap ions are rapidly pulsed into the orbitrap. This rapid pulsing results in very low time gaps (10^{-7} s) between ions with the same m/z values.

Orbitrap mass analyzer

The main principles of the orbitrap analyzers can be based on technical details concerning the *Kingdon* trap and the Fourier transform ion cyclotron measurement (Hu et al., 2005). The Kingdom trap has a simple architecture consisting of a thin-wire central electrode (inner electrode) that is surrounded a by coaxial cylindrical (outer) electrode. The modified outer electrode of the *Kingdon* trap (modified by Knight (1981)) leads to a symmetric potential with an electrostatic potential Φ given by

$$\Phi = A\left(z^2 - \frac{r^2}{2} + B\ ln(r)\right)$$

Where r and z are cylindrical coordinates and A and B correspond to constants related to the electrode geometry and the applied voltages. This new configuration corresponds to the orbitrap analyzer, as now the m/z values can be converted from the frequency of harmonic ion oscillations along the z-axis (Perry et al., 2008), similar to the transformation calculated for ion cyclotron measurements. The precise definition of the potential Φ for orbitrap mass analyzers is given by

$$\Phi(r,z) = \frac{k}{2}(z^2 - \frac{r^2}{2}) + \frac{k}{2}(R_m)^2 ln\left[\frac{r}{R_m}\right] + C$$

Here C is a constant, k the field curvature and R_m is the characteristic radius. The shape of these axially symmetrical electrodes is given by the cylindrical coordinate z.

$$z_{1,2}(r) = \sqrt{\frac{r^2}{2} - \frac{R_{1,2}^2}{2} + (R_m)^2 ln\left[\frac{R_{1,2}}{r}\right]}$$

z_1 is this case would correspond to central electrode and z_2 denotes the outer electrode. $z = 0$ is the plane of symmetry and $R_{1,2}$ are the maximum radii of the corresponding electrodes (Makarov, 2000). By calculating ion trajectories for ions with mass m and charge q and by defining energy characteristics for the ion motion in the z-axis as outlined in Makarov (2000), the equation of motion along z can be explained by the harmonic oscillator

$$z(t) = z_0 cos(\omega t) + \sqrt{\frac{2E_z}{k}} sin(\omega t)$$

With

$$\omega = \sqrt{\left(\frac{k\ q}{m}\right)}$$

this allows to determine the mass to charge ratio, here m/q.

2.2.7 Peptide fragmentation

Peptide fragmentation is one of the most fundamental principles in proteomics, as all subsequent analyses rely on the peptide fragment spectra. Various types of fragmentation methods have been implemented in recent years. Electron capture dissociation (ECD), electron transfer dissociation (ETD) and collision-induced-dissociation (CID) are the most frequently used methods. ETD and ECD rely both on the reduction of peptide ions, which reduces the charge of the positively charged ions and induces fragmentation of the unstable radical ions. These methods do not work on singly charged ions, as the electron reduces the charge by one. In both methods, peptides tend to fragment at the Cα-NH bond, resulting predominately in c and z ions. ETD of ECD are often used in the analysis of post-translational modification (PTM), as the PTM remains attached to the peptide ion (Zubarev, 2004). In contrast to ETD or ECD, CID fragmentation is based on collisions of the peptide ions with inert gases (e.g., helium or nitrogen) at very low energies (eV range). The cleavage of the peptide at the peptide bond, generating b and y ions, makes CID fragment spectra comparable easy to interpret. The upper part of Fig. 2.7 shows the different types of ions that can be generated by fragmentation methods and the lower part shows b and y ions that are predominately produced by CID. R_i symbolizes the amino acids side chain residues.

2.2.8 Identification of tandem mass spectra

An essential step in proteomics is the identification of peptides from their tandem mass spectra. Computational analysis of tandem MS spectra relies on the presence of distinct ion series. Since the experimental part of this thesis employs CID fragmentation only, we will only discuss CID spectra in this section. CID spectra are dominated by b and y ions. The lower part of Fig. 2.7 shows typical product ions of CID fragmentation. Tandem spectra contain series of masses that help to assign the peptide sequence. Proteomics datasets usually contain thousands of tandem spectra and in most of the cases the spectra are not easy to identify due to noise and interferences with other masses. Manual evaluation of all spectra recorded in a large scale experiment is thus not feasible. Research efforts for the identification of tandem spectra are far-ranging. Two main areas are *de novo* methods for the assignment of sequences and database retrieval algorithms.

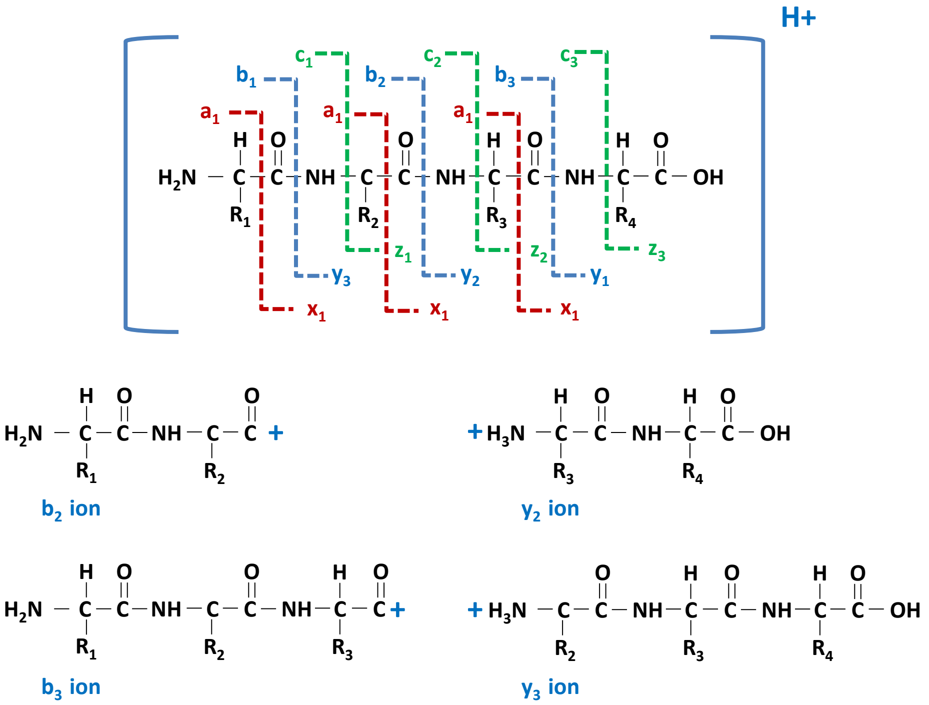


Figure 2.7: Fragment ion nomenclature in tandem mass spectrometry.

De novo sequencing

De novo in this case refers to calculations that make predictions on the sequence without extrinsic data. *De novo* methods are of particular interest if the task is to find novel proteins, amino acid mutations, and studying the proteome before the genome. Among *de novo* algorithms there are several different approaches. Graph-based methods are very popular in the *de novo*-approach. The peaks from a tandem spectrum are considered nodes on a directed graph, nodes are connected by edges if the mass difference between them equals the mass of any amino acid (DiMaggio and Floudas, 2007). A peptide sequence corresponds to a path through this graph. *De novo* methods, however, rely on very clean and complete spectra. Experimental spectra often have missing peaks or contain a lot of noise, which makes *de novo* assignments unreliable. Furthermore *de novo* identification relies on high mass accuracy, as the ambiguity for amino acid masses in low accuracy data complicates the identification.

Database retrieval algorithms

Alternative methods that do not assign sequences *de novo*, but rely on knowledge about the protein that is expected to be expressed are summarized as database retrieval algorithms. Common to all database retrieval algorithms is the comparison of experimental to theoretically calculated spectra, as outlined in the general schema for database search in Fig. 2.8. This comparison is performed by various tools, so-called search engines. The most frequently used search engines are the commercial tools Phenyx (Colinge et al., 2003), Sequest (Eng et al., 1994), and Mascot (Perkins et al., 1999). Popular non-commercial solutions include X!Tandem (Craig and Beavis, 2004), OMSSA (Geer et al., 2004) or InsPect (Tanner et al., 2005). The methods used by these algorithms are very diverse, but rely fundamentally on the comparison of the experimental spectrum to a set of candidate spectra that are calculated from a database. The size of this set of candidates has a strong influence on the performance of the particular algorithms. An important parameter for the calculation of the list of candidates is the mass error window. Novel mass spectrometers allow measuring very accurate

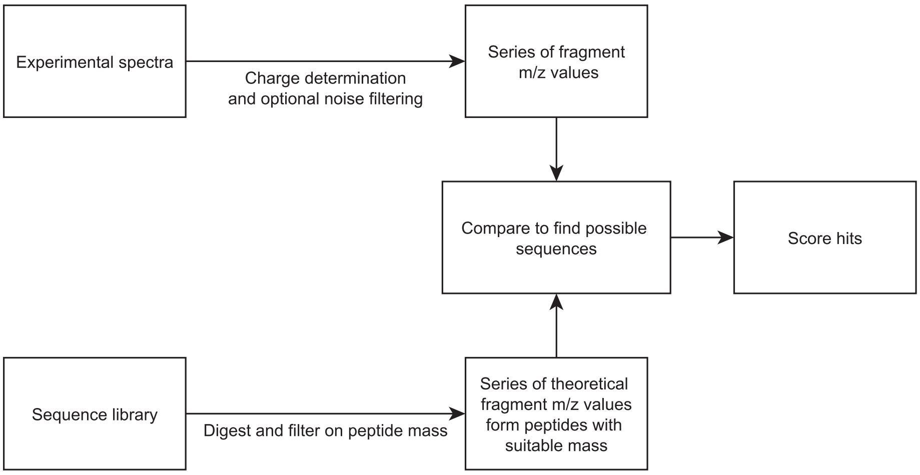

Experimental spectra
Charge determination and optional noise filtering
Series of fragment m/z values
Compare to find possible sequences
Score hits
Sequence library
Digest and filter on peptide mass
Series of theoretical fragment m/z values form peptides with suitable mass

Figure 2.8: A general workflow of database search approaches.

masses, with mass errors at 1 ppm (parts per million) or lower. With such high accuracy the number of peptides from a given proteome that have masses that fall in such a window is rather low. With low accuracy instruments this window is comparatively big. The great asset of high accuracy mass spectrometry is visualized in Fig. 2.9. At a low mass tolerance the number of candidates in the search space is significantly reduced, which implies a decreased likelihood for false positive peptide identifications. For all peptides that fall in such a window the theoretical tandem spectra are calculated. Fig. 2.10 shows an example of an experimental spectrum annotated with the corresponding ion ladder and the theoretically calculated spectrum. Mascot, X!Tandem and OMSSA are the three search engines relevant to the remainder of this thesis. We will thus discuss the algorithms underlying these search engines in more detail.

X!Tandem

The algorithmic procedure underlying the X!Tandem approach is a dot product of the experimental and the theoretical spectra (Fenyo and Beavis, 2003). A theoretical tandem mass spectrum is encoded as an n-dimensional vector $\mathbf{t}$. The set of all theoretical candidate spectra is given by $\mathbf{K}$ and defines the search space. $\mathbf{t} \in \mathbf{K}$ contains n entries t_i of either 1 or 0. n is the parent ion mass divided by the accuracy of the mass spectrometer. Thus n corresponds to all technically distinguishable peaks. If the mass i can occur for the peptide in consideration, then $p_i = 1$, otherwise $p_i = 0$. The theoretical spectrum t is thus given by

$$\mathbf{t} = (t_0, ..., t_n)$$

The experimental spectrum is encoded in the same way as the theoretical spectrum, using an n-dimensional vector $\mathbf{e}$.

$$\mathbf{e} = (e_0, ..., e_n) \text{ with } 0 \leq e_i \leq 100 \; \forall \; i$$

The values e_i are normalized product ion intensities. $\mathbf{e}$ is now compared to a all theoretical candidates $\mathbf{t} \in \mathbf{K}$ and a correlation score x is calculated for each

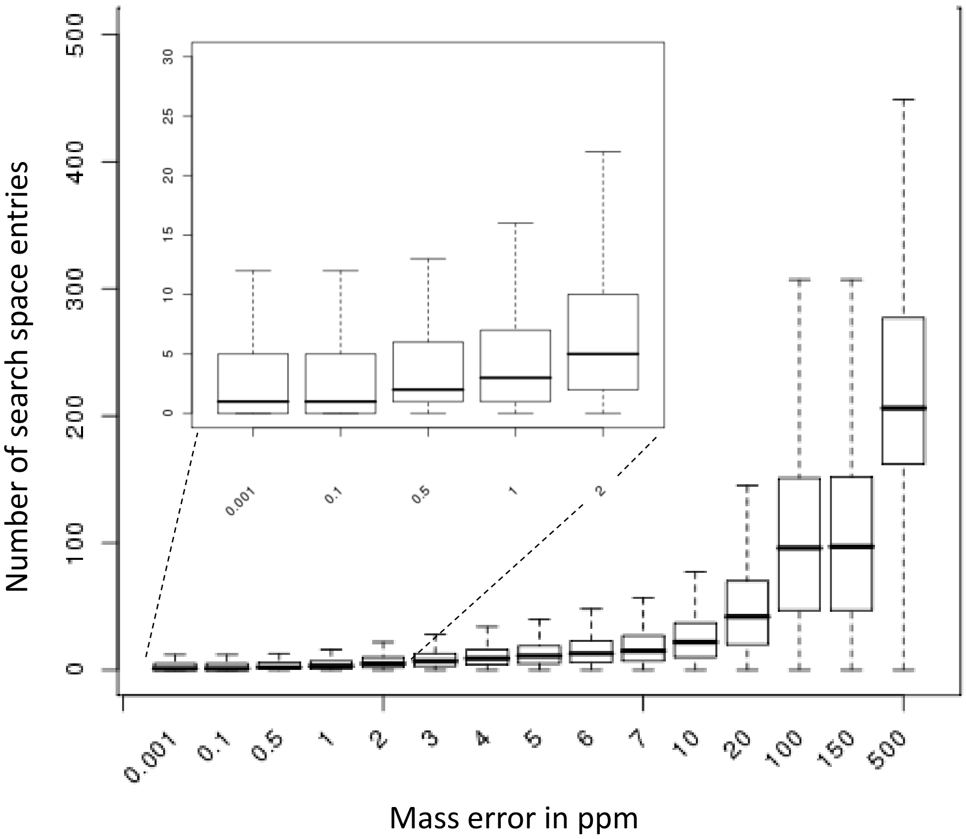


Figure 2.9: The number of candidates in the search space as a function of mass tolerance. Calculation were performed on a 12,804 entries database from *B. subtilis.*

comparison of a theoretical spectrum to the experimental spectrum.

$$x(\mathbf{e}, \mathbf{t}) = \langle \mathbf{e}, \mathbf{t} \rangle = \sum_{i=0}^{n} e_i t_i \tag{2.1}$$

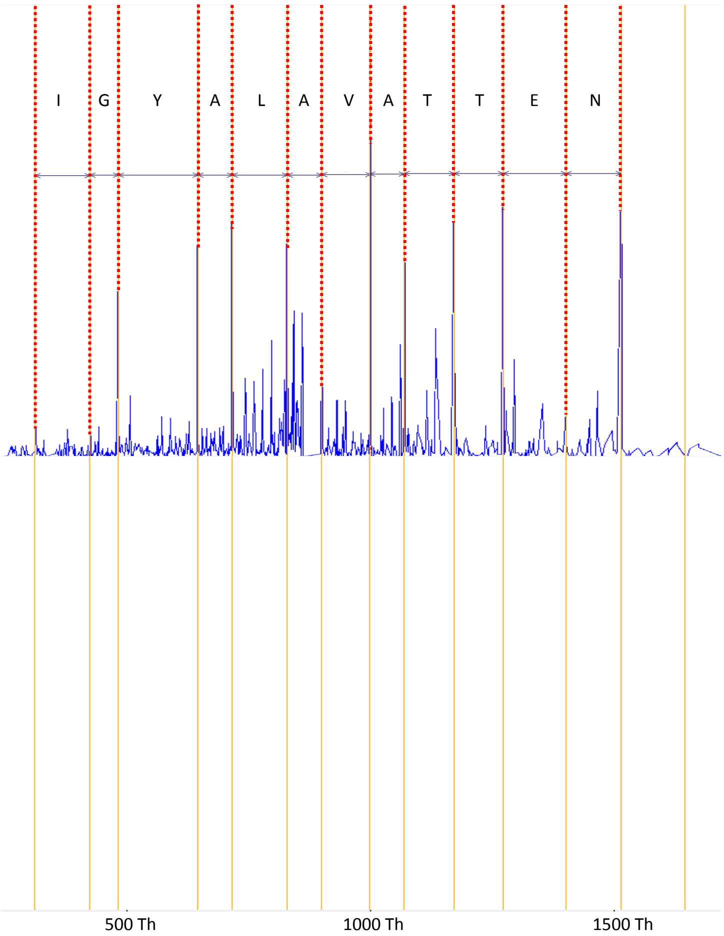


Figure 2.10: Experimental spectrum (top) with the corresponding theoretical spectrum (down). The peptide sequence is LMNETTAVALAYGIYK, a peptide from an HSP70 (heat-shock-protein 70) isoform.

The X!Tandem hyperscore is calculated on the basis of the dot product.

$$h(\mathbf{e}, \mathbf{t}) = \left(\sum_{i=0}^{n} e_i t_i \right) N_b! N_y! \tag{2.2}$$

Where N_b and N_y correspond to number of assigned b and y ions, respectively. The hyperscore assumes an underlying hypergeometric distribution for the number of matches N_b and N_y. Due to the additional factors, the usage of the hyperscore accounts more strongly for matchings with an increased number of mapping product ions, compared to the simple dot product. The peptide with the top hyperscore is now assigned to the spectrum. In addition to the hyperscore X!Tandem also assigns on E-value to each score. The E-value for a given peptide hit simply indicates the probability that any hyperscore equal or higher might be assigned by chance.

The implementation of X!Tandem (Craig and Beavis, 2004) routinely uses an additional refinement function. The refinement is a secondary run that automatically can identify peptides with various modifications, missed cleavages and even polymorphisms. This special feature is realized by consecutive identification runs. X!Tandem is designed in a way that it initially identifies peptides with a small user-defined set of parameters and enzymatic specificities with the algorithmic approach as outlined above. As a next step a new database is constructed. This new database only consists of proteins that were significantly identified (E-value ≤ 0.1). Although this search strategy seemingly provides a high advantage compared to other approaches, there are a lot of unresolved statistical issues accompanied with the X!Tandem's refinement function, for example X!Tandem does not construct decoy hits in the refined database, which prohibits the assessment of the target-decoy-based false discovery rate.

Mascot

Mascot uses a probability-based search algorithm (Perkins et al., 1999), however, the authors never published details on their algorithm. The Mascot score is based on the Mowse score, which was an early algorithm for peptide mass fingerprinting (Pappin et al., 1993).

As in most scoring algorithms, the first step of the Mowse score calculates peptide masses for all theoretically possible peptide candidates. This calculation allows defining the search space. With the information on all theoretical peptide masses a matrix $\mathbf{F}$ is created. Each row in $\mathbf{F}$ corresponds to a bin of 100 Da in peptide mass and the columns correspond to bins of 10 kDa of intact protein mass. Every measured mass entry can now be assigned to elements in $\mathbf{F}$. After column-wise normalization steps by the largest entry per column, F is transformed into the Mowse factor matrix $\mathbf{M}$. $\mathbf{M}$ allows assigning a score to each experimental precursor mass and serves thus as an experimental mass list pre-processing. The principle of the Mowse score is incorporated in the Mascot tandem MS search algorithm. The final Mascot score involves the selection of two fragment ion types, where most fragment matches are observed, and a probability-based score is computed on the basis of these two fragment types only (Colinge and Bennett, 2007). It is calculated as $-10 \times log_{10}(P)$, where P corresponds to the absolute probability of the observed match to be a random event. An additional parameter reported by Mascot is the E-value. The E-value for a hit can be considered as a score. It indicates the expected number of random hits that might be assigned to the spectrum with probabilities as good or better than the given hit. An E-value of 1.0 is interpreted in a way that one peptide sequence with a score equal to or better than the hit being scored can be seen simply by chance.

OMSSA

The OMSSA (Open Mass Spectrometry Search Algorithm) algorithm (Geer et al., 2004) calculates an E-value based on the information on how many product ions, calculated from peptides in the search space, can randomly hit the tandem spectrum. The OMSSA scoring is based on the consideration of random matches to m/z values. The distribution of those random matches allows assigning significance values (probabilities) of hits for a given spectrum. The probability of the hit being random is calculated, where a low probability implies a significant hit. The overlap of fragment ions for each candidate in the search window is calculated separately for the different charge states. Starting with charge state 1+ as follows: Let s be the smallest measured product ion mass and let h be the highest measured product ion mass, then there can theoretically be $\frac{h-s}{2t}$ possible matches, if the mass tolerance is given by t. Furthermore, if the precursor weight is given by m, then there can be $\frac{k(h-s)}{m}$ calculated product ions, if k is the total number of calculated m/z values. This number of calculated m/z values needs to be matched to e experimental product ions. The OMSSA algorithm assumes a Poisson process for the distribution of the number of matches, which is further supported by Sadygov and Yates (2003). Poisson distributions are used in random processes where the average number of success is much lower than the possible number (Geer et al., 2004). These assumptions lead to the mean

$$\mu_1 = (\frac{2t}{h-s})(\frac{k(h-s)}{m})e = \frac{2tke}{m}$$

for the Poisson distribution given by

$$P(x,\mu) = \frac{\mu^x}{x!}e^{-\mu}$$

where x is the number of matches. μ is separately calculated for the different charge states. With the assumption of the Poisson distribution of the number of matches, OMSSA calculates an E-value, according which the hits are ranked. If a theoretical spectrum is compared to a calculated spectrum, then the probability that this matching is not random is given by

$$\sum_{x=0}^{y-1} P(x,\mu)$$

where y is the number of product ion matches. Then the probability that the search against N theoretical spectra is random, is obviously given by

$$1 - (\sum_{x=0}^{y-1} P(x,\mu))^N$$

and the E-value is calculated by

$$E(y,\mu) = N(1 - (\sum_{x=0}^{y-1} P(x,\mu))^N)$$

The interpretation of the OMSSA E-values is identical to that of X!Tandem and Mascot E-values.

Statistical assessment of identification results

A very common method to estimate the statistical significance for tandem MS search results is the calculation of false discovery rates. The false discovery rates (FDR) in a general setup is defined as the expected ratio of the number of false positive (FP) instances to the number of all positive (P) instances that have a score s above any given threshold t.

$$FDR = \frac{FP_t}{P_t} = \frac{FP_t}{FP_t + TP_t}$$

Whereas TP corresponds to the number of true positives. Usually it is impossible to calculate this ratio, as false and true positives cannot be discriminated easily. To estimate appropriate values for the FDR, the target-decoy database search strategy has become a best practice for the statistical assessment of database search results (Elias and Gygi, 2007). Peptides are searched against a concatenated database, consisting of the usual forward database and a reversed (randomized or shuffled) version of the original database. For the estimation of the denominator all peptide spectrum matches (PSMs) with $s > t$ are counted. The enumerator counts all PSMs that are given a sequence from a decoy protein Since the FDR can be smaller for an instance i with score $s(i)$ as for an instance j with score $s(j)$, despite the fact that $s(i) < s(j)$, the notion of q-values was introduced by Storey and Tibshirani (2003). A q-value for instance i is the minimum FDR that would be necessary to accept this instance as a true positive, given the acceptance threshold t.

$$q(i) = \min_{t<s(i)}(FDR(t))$$

The application of the q-value measure to tandem MS search results was first introduced by Käll et al. (2008a). The q-value of a PSM simply corresponds to the smallest PSM at which the peptide would be accepted. Peptides accepted at FDR α refers to all peptides that were identified with q-values $\leq \alpha$.

2.3 Quantitative proteomics

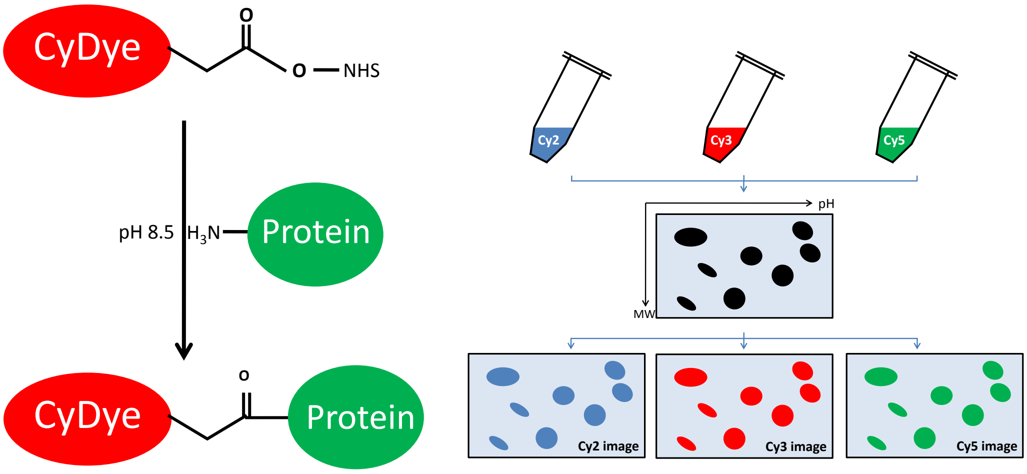
Mass spectrometry-based proteomics has so far been described as a powerful tool for the qualitative assessment of huge sets of proteins. In most biological questions, however, it is essential to not only have the information of what protein is expressed, but also to what extent it is expressed. The field of quantitative proteomics can fill this gap. Quantitative in this context can either mean absolute or relative quantitation. Absolute quantitation requires the calibration with samples of known concentration. Relative quantitation can be achieved by direct comparison of protein signals from two consecutive experiments or by the comparison of differentially labeled peptides that are analyzed in the same experiment.

2.3.1 Gel-based quantitation

Gel-based quantitation relies on the 2D separation with respect to the isoelectric point and the molecular weight of the proteins. After the separation of the proteins on a 2D gel, the gels are stained, scanned and software tools are used to detect the protein spots. These spots are then matched across different gels in the experiment and the different intensities of the protein spots are used for the quantitative comparison. Due to various drawbacks in conventional 2D gel electrophoresis, such as spot detection and spot matching, the gel-based quantitation remains difficult. Recent advances in 2D-PAGE include Difference Gel Electrophoresis (DIGE) (Unlü et al., 1997). The DIGE method allows accurate quantitation of relative protein abundances on one gel, since up to three samples can be separated on the same gel (2.11(b)). For this purpose CyDyes are used to label the protein samples previous to their mixed separation on a 2D-PAGE. In the DIGE setup three different CyDyes are used, Cy3 and Cy5 are predominately used to label the protein samples, and in most experiments Cy2 is used as an internal standard. The internal standard can consist of all samples that are used in the experiments. In this way the normalization procedure of ratios is facilitated. All CyDyes have an NHS ester reactive group and can covalently attach to the ϵ group of lysine; this reaction requires accurate pH adjustment to 8.5, as shown in 2.11(a). After the separation of the proteins in the two dimension the gels are scanned with specific filters that allow the separated detection of the different CyDyes. The images are then overlaid and as the same proteins from the different samples migrated exactly to the same point, the matching of those DIGE maps is trivial. However, this great advantage of CyDye labeling of different protein samples diminishes of the biological experiments consists of more than three samples. If the sample size increases, protein samples have to separated on different gels and for comparison, images from different gels have to be matched, which is still a challenging task (Faergestad et al., 2007).

2.3.2 Gel-free quantitation

Gel-free quantitation has some distinct advantages over 2D gel-based quantitation. 2D gel-based analyses in general are biased towards high-abundance proteins



(a)

(b)

Figure 2.11: (a)CyDye labeling takes place on the ϵ of lysine residues. (b) The DIGE workflow allows to mix samples previous to separation.

and it is barely possible to resolve membrane proteins, due to their biochemical properties. There are two fundamentally different approaches to quantitation in the mass spectrometer, methods using peptide labeling as well as label-free methods. Both methods rely on the quantitative nature of signal detection in the mass spectrometer. Fig. 2.12 illustrates the most commonly used methods for gel-free quantitative proteomics. The very left column on Fig. 2.12 shows the stage of the proteomics experiment. The workflow starts with the cells or tissue and is followed by the extraction of the sub-proteomes, the proteolytic digestion, the mass spectrometric measurement, and finally the data analysis as the last step.

Metabolic labeling

The earliest possible state where samples can be combined is before the extraction of proteins. Metabolically labeled proteins are produced in conditions where cells have been grown on media that contained only specific isotopes for some amino acid or media that contain only specific isoforms of atoms (e.g. $^{15}\mathbf{N}$ labeling). The procedure that uses stable isotopes of amino acids has first been published by Ong et al. (2002) and is known as stable isotope labeling with amino acids in cell

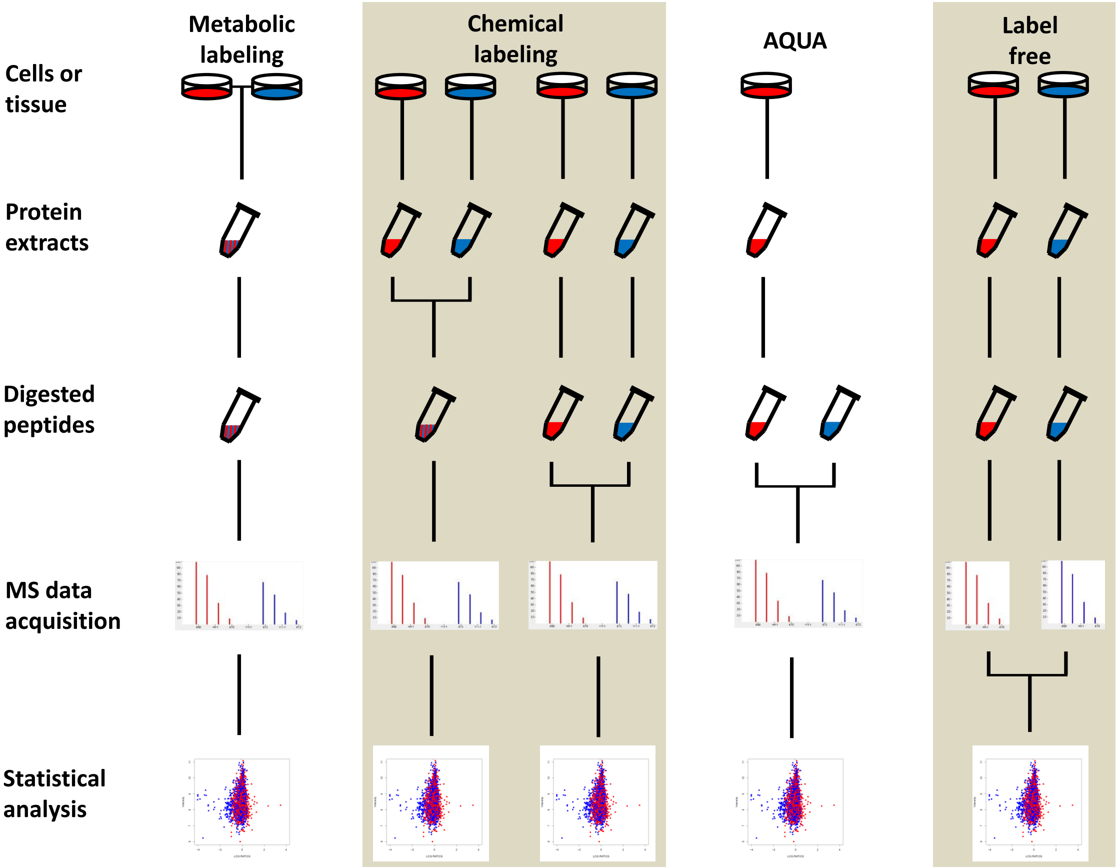


Figure 2.12: Comparison of different quantitative mass spectrometric workflows.

culture (SILAC). The SILAC strategy has major advantages over other quantitation methods. The method allows combining the samples at a very early stage and early mixing of samples avoids any systematic error during sample preparation on one sample only. In theory, every amino acid could be used as a SILAC label. However, most commonly Arg and Lys are used, as those labels allow the quantitation of every tryptic peptide. Lys4 ($\mathbf{^{2}H_4}$), Lys8 ($\mathbf{^{13}C_6^{15}N_2}$), Arg6 ($\mathbf{^{13}C_6}$) and Arg10 ($\mathbf{^{13}C_6^{15}N_4}$) are common examples for stable isotopes of Arg and Lys. Combinations of those labels allow a routine comparison of three different samples and theoretically the comparison of five samples is possible (Molina et al., 2009).

Chemical labeling

Chemical labeling summarizes methods where either the whole protein is labeled by chemical reactions or the labeling is introduced on the peptide level. In 1999, Isotope-Coded Affinity Tags (ICAT) (Gygi et al., 1999) were introduced as the first quantitative labeling methods in MS-based proteomics. In an ICAT experiment two different mass tags, differing by eight Da, are used. The ICAT tags covalently bind to cysteines in peptides. Cysteine affinity purification is used to extract the ICAT-labeled peptides from whole cell lysates. Following this purification, every peptide is labeled with tags on their cysteines and mass difference of the ICAT tags allows a relative quantitation of peptides.

Another approach for chemical labeling is the iTRAQ method (Ross et al., 2004). iTRAQ labeling is also a labeling method that is done on the peptide basis. In a single MS mode the differentially labelled versions of a peptide are indistinguishable. However, in tandem MS mode (in which peptides are isolated and fragmented) each tag generates a unique reporter ion. Protein quantitation is then achieved by comparing the intensities of the four reporter ions in the MS/MS spectra (Shadforth et al., 2005). Using 4-plex iTRAQ, four reporter masses of 114.1, 115.1, 116.1 and 117.1 Da are used. The reporter ions accompanied with four different balancer masses, result in a total mass of 145 Da for all iTRAQ labels. The labels are covalently bound to the N-terminus of different peptide populations and after fragmentation the reporter ion intensities allow to differentially quantify up to four different samples in a single MS run.

AQUA

The AQUA (absolute quantification) strategy was introduced in 2003 by Gerber et al. (2003). AQUA peptides are synthesized heavy isoforms of peptides in the complex sample. By spiking these AQUA peptides in known amounts to the sample, absolute peptide concentrations can be derived in an MS experiment for a selective set of peptides. The quantitation of the detected peptide pairs is done in the same way as for metabolic labeling strategies. A calibration line, drawn from different experiments, allows the determination of absolute quantities.

Label-free quantitation

Label-free methods have recently gained popularity due to advances computational proteomics (Grossmann et al., 2010; Schulz-Trieglaff et al., 2007). In such experiments, signal intensities of the same peptide features are directly compared in different LC-MS runs. This comparison relies on the accurate determination of the peptide signal intensities, as well as on the accurate mapping of those signals across maps. The success of a label-free quantitation relies on the alignment of several complex LC-MS maps.

2.4 Computational proteomics

In every proteomics experiments the goal is to identify and possibly to quantify all proteins in a biological systems. As already introduced, a variety of mass spectrometric methods can be used to acquire the qualitative and and quantitative information on the proteome sample. The analysis of these data is logically divided in two parts. The first, here defined as *data pre-processing*, comprises all steps that are necessary to annotate the mass spectra with peptide sequence information and optionally with relative expression rates or absolute concentrations. This pre-processing leads to interpretable data and is the basis for the next analysis step, where these data are subjected to statistical analysis and to automated biological interpretation. The second step has often been termed *downstream analysis*, (Kumar and Mann, 2009).

2.4.1 Data pre-processing

MS-based proteomics produces highly complex and large data sets. This sheer amount of data requires elaborate computational analysis methods. To this end, we used mainly two platforms. For the analysis and implementation in the ConsensusID project, the C++ software framework OpenMS was used and in the more experimentally orientated SILAC project the MaxQuant software suite (Cox and Mann, 2008) was used for data pre-processing.

OpenMS

The OpenMS project was initiated in 2003 by the Division for Simulation of Biological Systems of Tübingen University and the Algorithmic Bioinformatics group at the Freie Universität Berlin. OpenMS has originally been designed as a framework of data structures and algorithms for the development of software tools that can be used for the analysis of complex LC-MS data. Due to the library concept of OpenMS (Sturm et al., 2008), it allows rapid development of new algorithmic approaches. Besides the infrastructure for software development, OpenMS also provides a modular structured collection of tools, the TOPP tools (Kohlbacher et al., 2007). These tools can easily be combined into analysis pipelines. All identifications for the ConsensusID project were performed with the search engine adapters from the OpenMS TOPP tool collection and the ConsensusID implementation was also realized in the OpenMS library.

MaxQuant

In contrast to OpenMS, the MaxQuant (Cox and Mann, 2008) software is a 'ready to use' platform for the analysis of high-accuracy mass spectrometric data. MaxQuant is especially well suited for the quantitation of SILAC pairs. It uses Mascot for database identification of the peptide spectra and a decoy database for the assignment of false discovery rates.

2.4.2 Downstream analysis

Data normalization

Data normalization is an important procedure in the comparison of different experiments. Peptide ratios that were recorded in separate LC-MS runs need to be comparable. This is essential to get a time profile over all time points. A well-known and very simple method for data normalization is the Z transformation.

Z transformation has previously been applied to transcriptomics data (Cheadle et al., 2003), as well as to proteomics data (Olsen et al., 2006). The raw peptide or protein SILAC ratio is log-transformed (with the natural logarithm). The log ratios are then used for the calculation of Z scores. Z scores are calculated by subtracting the average ratio for every protein across all time points from the actual ratio at a given time point, and dividing that result by the standard deviation of all measured ratios for the given time point across all time points. The Z score for any ratio r in an experiment is calculated as follows:

$$Z_{score}(r_i) = \frac{r_i - \mu}{\sigma}$$

where μ is the mean ratio, calculated from all time points for a given peptide/protein and σ is the standard deviation for this peptide/protein, respectively.

Fuzzy c-means clustering

The fuzzy c-means algorithm is an unsupervised clustering method. It is widely used for pattern recognition in multivariate datasets. The term fuzzy implies that objects are assigned to several clusters. An important read-out of this algorithm is the membership degree, which corresponds to the probability that a given object is assigned to a given cluster (Bezdek, 1981). In the following u_{ik} will be the membership degree of the object x_i for the cluster c_k. The algorithm aims to maximize the objective function $J(U, K)$.

$$J(U, V) = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{ik}^{m} d_{ik}^{2}$$

where d_{ik}^2 corresponds to the squared Euclidean distance between the points x_i and the center of the clusters v_i and is defined as follows,

$$d_{ik}^2 = (x_k - v_i)^T (x_k - v_i)$$

The partition matrix M corresponds to all membership degrees for all clusters. The parameter $m > 1$ is called fuzzification parameter. Note that

$$u_{ik} \rightarrow \frac{1}{C} \text{ for } m \rightarrow \ \infty \ \forall \ i, k$$

and

$$u_{ik} \rightarrow \begin{cases} 0 \\ 1 \end{cases} \quad \text{for } m \rightarrow \ 1 \ \forall \ i, k$$

where C is the number of clusters. The objects x_i will be assigned equally well to each cluster, if m is chosen too large and the objects x_i will be assigned to one cluster only, if $m = 1$. The general idea of the fuzzy c-means clustering aims to assign data points to clusters in a way that d_{ik}^2 is minimal with the following constraints.

$$\text{(i)} \quad \sum_{i=1}^{C} u_{ik} = 1 \ , \ \forall \ k$$

The sum of the cluster membership is 1 for every object x_i and

$$\text{(ii)} \quad \sum_{k=1}^{N} u_{ik} > 0 \ , \ \forall \ i$$

And the clusters have to be non-empty. The *Langrangian* function is used to find the minima of the objective function.

$$J(U, V, \lambda) = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{ik}^m d_{ik}^2 - \sum_{z=1}^{N} \lambda_z \sum_{i=1}^{C} u_{ik} - 1$$

GO and KEGG databases

The Gene Ontology (GO) (Ashburner, 2000) database is a hierarchical database containing annotations for the biological process, molecular function and the cellular compartment of a large set of known proteins. The hierarchical structure of the database assigns several annotations to a protein, starting from global properties, such as metabolic process, to more detailed descriptions, such as catabolic processes of amino acids. The consistency of the annotations is controlled by using a controlled vocabulary (CV) of terms. The Kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa and Goto, 2000) is structured in a similar way as the GO database and assigns pathway information to every known protein from the human genome. The KEGG database collects information for every annotated gene that allows the description of higher-order systemic behaviors of the cell and the organism from these information. As all proteins in these databases are annotated with a CV of terms, a given dataset (e.g., a experimentally determined set of differentially expressed proteins) can be tested for the enrichment of database annotations, using the background of all annotated proteins (reference set).

Statistical testing

To determine the significance of the enriched categories an hypergeometric test (Rivals et al., 2007) and the Benjamini-Hochberg false discovery rate correction (Benjamini and Hochberg, 1995) were used. In the example below (Tbl. 2.1) GO category denotes the number of proteins that carry a given GO annotation. We assume there are n proteins grouped in a cluster and the total number of proteins in the reference dataset (e.g., the GO database) is N. Furthermore, among the N proteins in the reference set there are m proteins that belong to the GO category. In the given cluster there are k proteins associated with the GO category. We further assume the distribution of k follows a hypergeometric distribution. The hypergeometric distribution indicates the exact probability p of observing this k proteins with the GO annotation 'GO category' in the experimental set of n proteins.

Table 2.1: We hypothesize that the proportion of proteins with the GO annotation 'GO category' is higher in the set of proteins that clustered together than in the reference proteome.

	cluster	**reference**
$\in$ **GO category**	k	m
$\notin$ **GO category**	$n-k$	$N-m$
total	n	N

$$f(k; N, m, n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

The hypergeometric test calculates the probability to observe exactly k or more proteins.

$$F(X \geq k) = \sum_{i=k}^{m} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$$

Enrichment analysis

Cytoscape (Shannon et al., 2003) and the plugin ClueGo (Bindea et al., 2009) were used to determine significantly enriched categories of biological processes and KEGG pathways. For the visualization of the enrichment results the cytoscape network representation was used.

Cytoscape representation

The GO categories are represented as nodes in the enrichment graph. Two nodes are connected if they share proteins. The connection between two nodes is derived by applying κ statistics and a κ value threshold of 0.3 to accept interconnectivity (Bindea et al., 2009). The κ value is expressed by the strength of the connecting line. If three or more nodes are connected they are assigned to groups and the node with the most significant p-value determines the name of the group. For the overall representation it was required that the corrected group p-value is equal or smaller to 0.05. Fig. 2.13 shows a typical graph-based representation of enriched GO categories. GO:01 and GO:11 are GO categories that contain either the same set of proteins or at least a subset of proteins from each other. GO:03 contains only proteins $pro_{i+2}, ..., pro_n$ and GO:02 contains proteins pro_i and pro_{i+1}. The group name for group one and group two will be given by the single categories with the most significant p-value for each group separately. GO:04 contains none

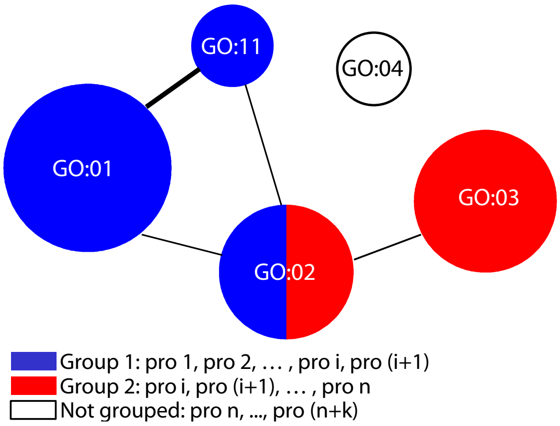

GO:11
GO:04
GO:01
GO:03
GO:02
Group 1: pro 1, pro 2, … , pro i, pro (i+1)
Group 2: pro i, pro (i+1), … , pro n
Not grouped: pro n, ..., pro (n+k)

Figure 2.13: Network based visualization of enriched biological processes.

of the proteins that are contained in the categories GO:01 - GO:03, therefore it is not grouped with any of the other categories.

Chapter 3

Analysis of protein expression dynamics using 2D-DIGE

Proteins that are differentially expressed between two cellular states are of particular interest for the analysis in every proteomics experiment, since these proteins allow to draw conclusions about the perturbation that has been introduced to one of the states. Two-dimensional Difference Gel Electrophoresis (DIGE) is a valuable tool for such comparative analyses. The technology involves the labeling of lysine residues in the different cellular protein extractions with one of three different fluorescent dyes, Cy2, Cy3 and Cy5 (CyDyes). The differentially labeled protein samples are then mixed and separated on the same gel. The CyDye labels have similar weights and pI values, resulting in the same migration position of the proteins separated on the same gel. Images, taken from the different CyDye labels, allow a direct comparison of the fluorescence intensities. However, for thorough statistical analysis, replicate analyses on different gels have to be performed. Fig. 3.1 shows a suitable experimental setup of an DIGE analysis, involving only two different PAGEs (PolyAcrylamide Gel Electrophoresis). As demonstrated in this sketch, the matching of the images acquired from the same gel is trivial, due to identical protein spot coordinates. However, the migration differences for samples, run on different gels, confront the researcher with the same matching problem as for classical 2D-PAGE without CyDye labels. To date, there are no experimental solution to cope with distortions due to protein migration in different PAGEs. For that reason DIGE-based proteomics needs accurate software solutions for the alignment of multiple DIGE maps.

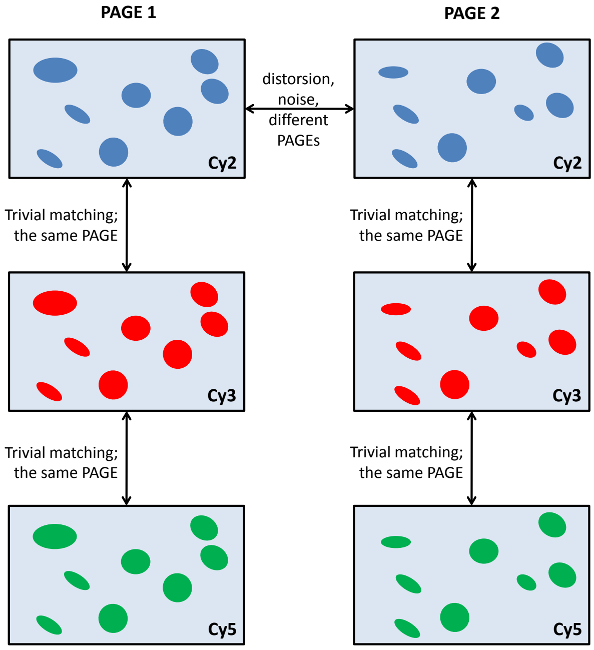

PAGE 1
PAGE 2
Cy2
Cy2
distorsion, noise, different PAGEs
Trivial matching; the same PAGE
Trivial matching; the same PAGE
Cy3
Cy3
Trivial matching; the same PAGE
Trivial matching; the same PAGE
Cy5
Cy5

Figure 3.1: This experiments involves six different proteome samples that are separated on two different DIGE gels.

3.1 DIGE map alignment

Methods from graph theory offer generic frameworks that can be used to model a variety of problems. Those methods have been applied to various fields of bioinformatics, such as modeling of metabolic networks or biochemical pathways (Sirava et al., 2002), analysis and visualization of expression data (Gentleman et al., 2004; Shannon et al., 2003). Matching approaches for maps generated by 2D-PAGE include pixel-based-warping methods as implemented in the Progenesis SameSpots software (Faergestad et al., 2007) and spot-based warping methods as implemented in the Proteomweaver software (Bio-Rad, CA, USA). Recently a graph theoretical framework was also applied to the analysis of 2D-PAGE (two-dimensional gel electrophoresis) (Peres et al., 2008). In the following sections, we will introduce a novel algorithm, called GBD (Graph Based Dewarping). This algorithm implements a geometric matching using a solution to the maximum weight matching problem in complete bipartite graphs. In contrast to previous algorithms, the GBD method reduces the complexity of the alignment problem by aligning only the differentially regulated protein spots, making use of the unique advantage of the DIGE method, which allows easy selection of differentially regulated protein spots on every DIGE gel. The gel positions of the differentially regulated protein spots are normalized to a set of landmarks on the each gel. The identity of theses landmarks on the different gels is always validated by MS identifications. This intra-gel normalization method corrects the distortion that is one of the main problems in comparing several 2D gel maps. The reduction of the matching problem to differentially regulated proteins accompanied with the intra-gel normalization outperforms other methods, such as pixel- or spot-based-dewarping in terms of accuracy and run time.

3.1.1 General idea

A graph $G = (V, E)$ is a collection of vertices V and edges E. G is bipartite if there exist partitions $V = X \cup Y$ with $X \cap Y = \emptyset$ and $E \subseteq X \times Y$ and G is complete if every pair of vertices $(v_i, v_j) \in V$ is connected. G is called a complete bipartite graph if every vertex from X is connected to every vertex from Y. Fig. 3.2 shows a bipartite and a complete bipartite graph.

A matching of G is a subset $M \subseteq E$. $|M|$ is the number of edges in M. M is a maximum matching $\Leftrightarrow \nexists M'$ with $|M'| > |M|$ (Cormen et al., 2001). A bipartite graph is called weighted $\Leftrightarrow$ each edge (i, j) has a weight $\omega(i, j)$. $\omega(M) = \Sigma_{e \in E} \omega(e)$ is called the weight of the matching M (Cormen et al., 2001).

The Hungarian algorithm

Let $G = (X \cup Y, E)$ be a complete weighted bipartite graph, where partitions

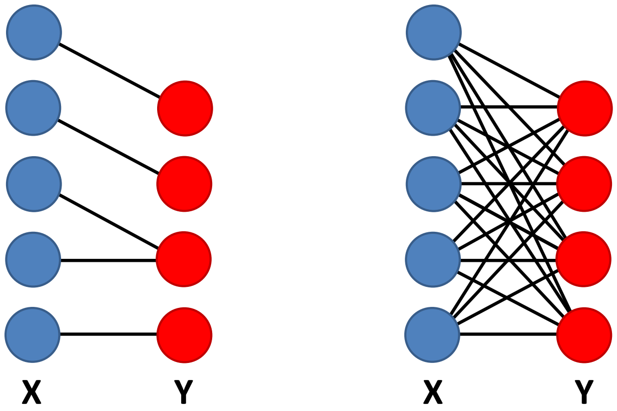


Figure 3.2: A bipartite (left) and a complete bipartite graph (right).

X and Y correspond to the differentially regulated spots on DIGE maps from two different biological replicates. Each vertex $x \in X$ (differentially regulated proteins on DIGE map X) has an edge to each vertex $y \in Y$. To calculate weights for these edges, a two-step procedure is necessary. First, an n-dimensional vector is assigned to each differentially regulated protein (this is done on each map separately). This vector contains the absolute distances to n landmarks. An example for $n = 3$ is shown in Fig. 3.3. $l_1, ..., l_3$ are landmarks and S_1 is a differentially regulated protein spot. For each pair of landmarks we can calculate the distance $\mathrm{y}_{1,i,j}$ to S_1, e.g. $\vec{d}(S_1, l_1, l_2) = y_{1,1,2}$ and $\vec{d}(S_1, l_2, l_3) = y_{1,2,3}$. Using n landmarks to determine the intra-gel position of m regulated protein spots results in the matrix G_k for gel k.

$$G_k = (\sum_{i=1}^{n-1} i \times m) = \begin{pmatrix} y_{1,1,S_1} & \cdots & y_{1,1,S_m} \\ \cdots & \cdots & \cdots \\ y_{n,n-1,S_1} & \cdots & y_{n,n-1,S_m} \end{pmatrix}$$

Second the complete bipartite graph $B_{k,l} = G_k \times G_l$ is constructed by considering two matrices, G_k and G_l. The weights of the edges are calculated as follows.

$$\omega_{i,j} = \vec{d}(P_{1,i}, P_{2,j}) = \min_{r,q} |y_{r,q,1} - y_{r,q,2}| \tag{3.1}$$

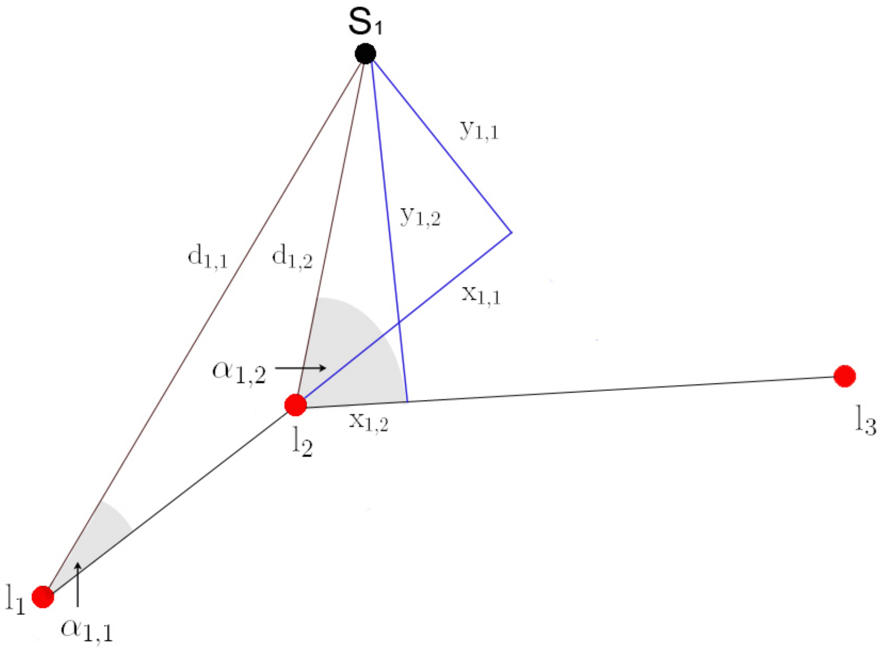


Figure 3.3: Exemplified construction of intra-gel distances.

This results in the bipartite graph $B_{k,l} = G_k \times G_l$

$$B_{k,l} = G_k \times G_l = \begin{pmatrix} \omega_{1,1} & \dots & \omega_{1,j} \\ \dots & \dots & \dots \\ \omega_{i,1} & \dots & \omega_{i,j} \end{pmatrix}$$

The assignment problem is to find a min-weight matching (find the pairs of vertices that have the smallest distance) in $B_{k,l}$. This matching results in pairs of protein spots that are reproducibly regulated in both maps. The Hungarian algorithm solves this problem in polynomial time (Kuhn, 2005). Since the theoretical investigations on the performance of the Hungarian algorithm in 1957 by James Munkres (Munkres, 1957), it is also known as the Munkres algorithm. The Munkres algorithm consists of the following six steps:

- ***STEP 1****: For each row of $B_{k,l}$, find the smallest element and subtract it from every element in its row. Go to Step 2.*

- ***STEP 2***: *Find a zero (Z) in the resulting matrix. If there is no starred zero in its row or column, star Z. Repeat for each element in the matrix. Go to Step 3.*

- ***STEP 3***: *Cover each column containing a starred zero. If K columns are covered, the starred zeros describe a complete set of unique assignments. In this case, go to DONE, otherwise, go to Step 4.*

- ***STEP 4***: *Find a non covered zero and prime it. If there is no starred zero in the row containing this primed zero, go to Step 5. Otherwise, cover this row and uncover the column containing the starred zero. Continue in this manner until there are no uncovered zeros left. Save the smallest uncovered value and go to Step 6.*

- ***STEP 5***: *Construct a series of alternating primed and starred zeros as follows. Let Z0 represent the uncovered primed zero found in Step 4. Let Z1 denote the starred zero in the column of Z0 (if any). Let Z2 denote the primed zero in the row of Z1 (there will always be one). Continue until the series terminates at a primed zero that has no starred zero in its column. Unstar each starred zero of the series, star each primed zero of the series, erase all primes and uncover every line in the matrix. Return to Step 3.*

- ***STEP 6***: *Add the value found in Step 4 to every element of each covered row, and subtract it from every element of each uncovered column. Return to Step 4 without altering any stars, primes, or covered lines.*

- ***DONE***: *Assignment pairs are indicated by the positions of the starred zeros in the cost matrix. If C(i,j) is a starred zero, then the element associated with row i is assigned to the element associated with column j.*

This pair-wise procedure is done for each pair of DIGE maps. The next step is a simple comparison of all pairs in an experiment. If a protein spot is always assigned to the same corresponding spot on different maps, this protein is differentially regulated. This procedures is outlined in Fig. 3.4. For n experiments the assignment problem is solved $\frac{n(n+1)}{2}$ times.

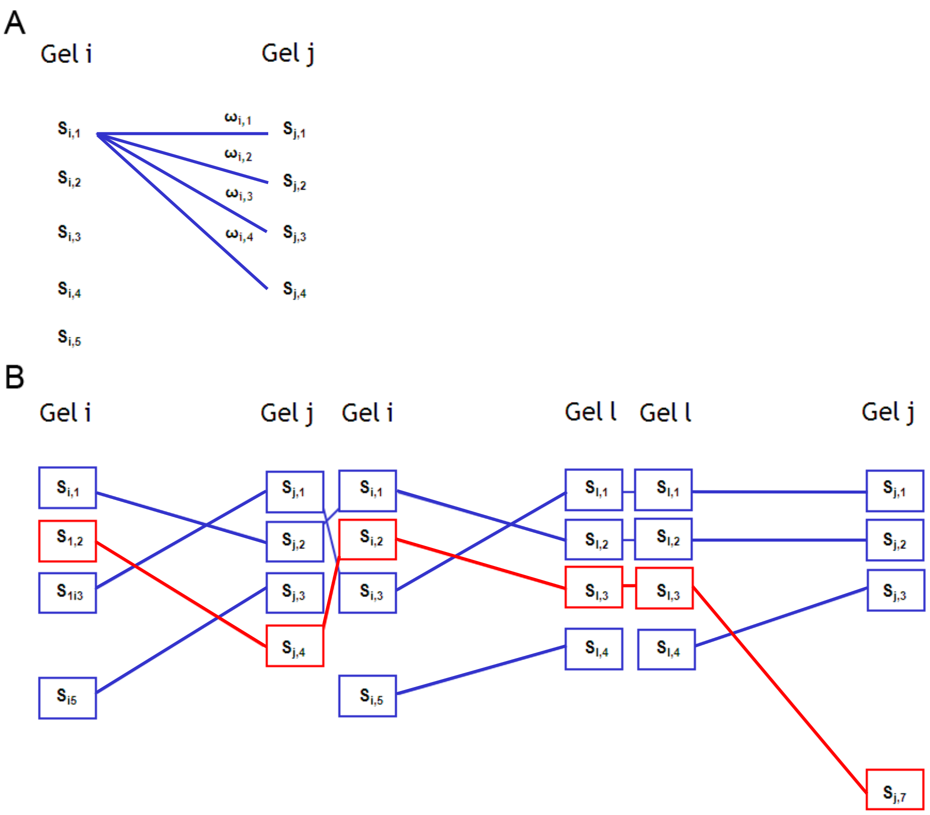


Figure 3.4: For each pair of DIGE gels the assignment is solved separately and the following deconvolution of the assignment network finds the same spots on all gels.

3.1.2 Results

GBD: A novel algorithm for DIGE map alignment

In a subset of gels we thoroughly evaluated the GBD performance in comparison to other software solutions:

To identify the protein spots that are reproducibly regulated we performed DIGE analysis from three experiments with independent biological material. Additionally, the whole DIGE data analysis was performed using two different commercial software solutions. Proteomweaver (Biorad) and SameSpots (Progenesis) were used as common commercial tools. We usually used the spot detection algorithms from the Proteomweaver software. Our matching procedure allows to handle large numbers of incorrectly detected spots, since only the reproducibly differentially regulated spots are further analyzed. For both commercial tools automatic intergel spot matching was not possible. Repeatedly spots needed to be matched between gels by the user. This very time-consuming and highly subjective step could be overcome by our approach. The advantages of the GBD workflow in comparison to the Proteomweaver matching is shown in Fig. 3.5. Using the automated pipeline provided by the commercial tools, such as Proteomweaver, the matched maps contain many regions where the matching did not work, as shown in the zoom regions of Fig. 3.5. Regions where spots are not matched result in singleton signals, unless they are curated manually. Next, we evaluated the time needed for the analysis of the three replicated experiments. For the commercial software tools these values are averaged estimates, as this task highly depends on the expertise of the user. In Tbl. 3.1 we compare the estimated time that is needed to perform the analysis with Proteomweaver, SameSpots or GBD. t_{detect} corresponds to the averaged time the user needs to adjust spot detection parameters. t_{intra} is the time needed for the intra-gel matching of different images from the same gel. t_{inter} is the matching of biological replicates that are separated on different gels. Note that GBD does not perform spot detection, however the spot detection parameters for the individual software do not have to be optimized. We used the default settings from the Proteomweaver software for spot detection. Incorrectly detected spots will be eliminated in the GBD workflow, as they will not be biologically reproducible. As shown in Tbl. 3.1 the analysis time is strongly reduced by our method. Time for spot detection is significantly

Table 3.1: Different processing times for DIGE gel analysis.

	Proteomweaver	SameSpots	GBD
t_{detect}	60 min	30 min	**5 min**
t_{intra}	5 min	5 min	**5 min**
t_{inter}	60 min	90 min	**5 min**

reduced. The time for intra-gel matching is not influenced, whereas we can offer tremendous improvements for the time necessary for inter-gel matching. For the comparative analysis we used five landmarks. Using Proteomweaver without any manual intervention less than 20 % of the proteins, assigned as differentially regulated, were correctly matched across experiments. The same analysis with SameSpots revealed that approximately 40 % of the regulated protein spots were correctly matched between gels and we found that 95 % of the protein spots that were assigned by GBD were found to be truly regulated in all experiments. Each matching result has been manually validated. The results obtained for differ-

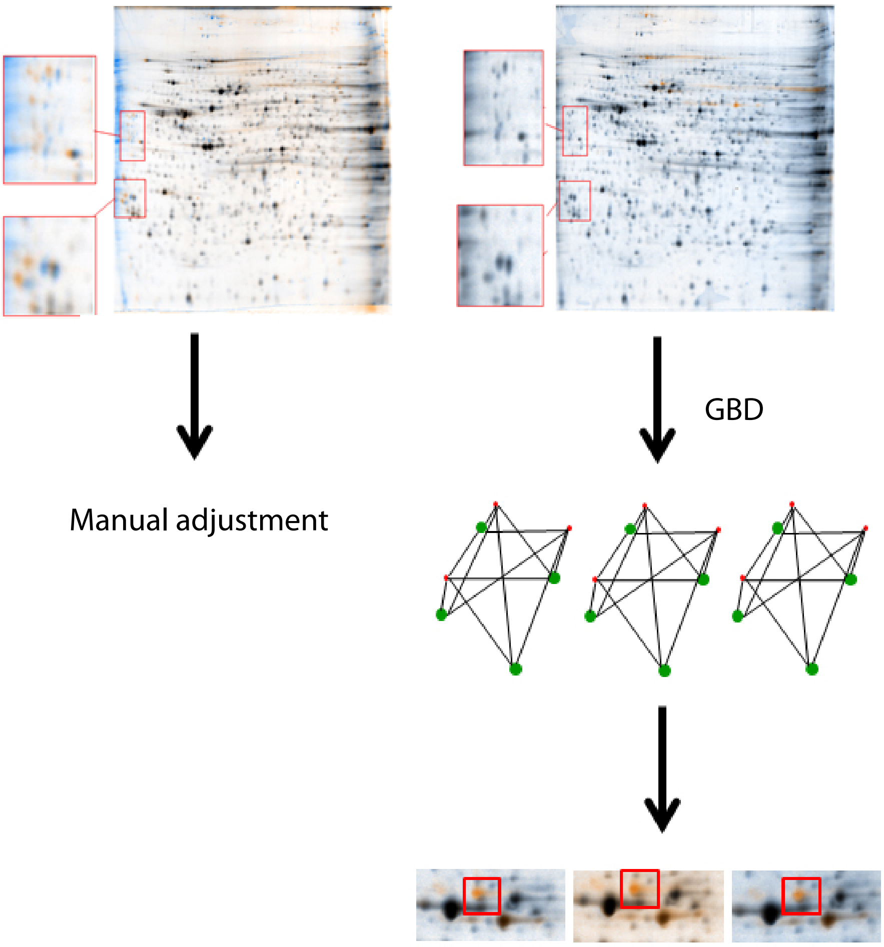


Figure 3.5: Automated matching performed by the Proteomweaver software (left) and the GBD workflow (right).

entially regulated proteins after three hours of treatment with the multi-kinase inhibitor sorafenib are shown in Fig. 3.5. Analyzing a single experiment usually leads to high numbers of singleton spots (Fig. 3.6 A). Two experiments of the same system with independent biological material reduces the number of singleton spots by more than 43 %. The third replication experiment introduces further reduction in the number of singleton spots by 75 %. The right part of figure 4 shows the final results. We observed 23 proteins as singletons in the controls (0) following sorafenib treatment, the expression of two proteins is switched on (unique in treated) (1), one is up-regulated (+) and three proteins are down-regulated (-). We accepted a protein spot to be regulated, if the fold changes were greater than two.

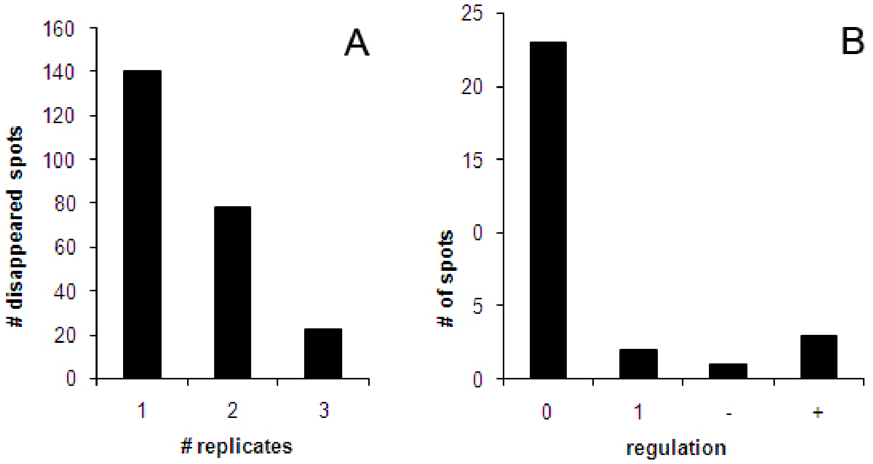


Figure 3.6: A: GBD analysis. B: Regulated protein spots.

3.1.3 Discussion

We present a method that enables high-throughput and improves sensitivity in DIGE-based proteomics. Our method is based on the assignment of geometrical distances of each differentially regulated protein spot to a set of landmarks. These intra-gel characteristics are highly reproducible and can thus serve to formulate the assignment problem for a maximum weight matching. This problem is well-known in theoretical computer science and can be efficiently solved by the Hungarian algorithm. The analysis of DIGE-based proteomics experiments starts with the spot detection and the assignment of expression ratios for the proteomes that have been separated in one gel. Spot detection is the first crucial part in the classical workflow for the analysis of DIGE gels. Stringent parameters for spot detection, allowing only the detection of clearly defined regions, tend to lose a lot of important protein spots, whereas less stringent parameters collect highly noisy data, that might not correspond to protein spots. If spot detection is followed by GBD analysis, the spot detection parameters can be chosen less stringent and consequently the spot detection software can detect higher numbers of protein spots. The use of replication experiments serves as a filter to distinguish noise from real protein data. As finding the appropriate spot detection parameters is a time consuming task for classical matching methods, GBD allows reducing subjectivity and increases the throughput by simply collecting as many candidate spots as possible. The time for intra-gel matching does not improve, as the intra-gel-matching is simply an overlay between images. In our case study we could significantly reduce the number of singleton spots per gel. Singleton spots are spots that are only found in one of the proteomes that are separated on one DIGE gel. As the complete down-regulation of a protein has strong consequences for the respective biological system, this needs careful investigation. The analysis of replication experiments showed that most of these singleton spots do not reflect true biological regulation, but indicate a lack of reproducibility in the single experiment. Our suggested method serves as a platform for high-throughput analysis of DIGE gels and we could show that the method allows a faster analysis of an DIGE-based proteomics experiment and significantly reduces the need for manual intervention. An additional research direction could include the automated selection of the landmark spots and thereby allowing the complete analysis without manual intervention. Alternatively, the experimental setup can be designed in a way that a set of proteins that separate well on a the 2D map is spiked into the protein mix previous to the separation on the 2D gel. By knowing the exact position of the spiked proteins, theses can automatically be used as accurate landmarks, without the need of MS confirmation on every gel.

3.2 Proteomic profiling of signaling cascade inhibitors using DIGE

The following section describes an application of DIGE-based proteomics profiling of global protein expression in melanoma cells following treatment with sorafenib. For the sorafenib treatment three time points were used. The results will be presented and discussed. A similar approach with an orthogonal, gel-free, proteomics platform will be described in Chapter 5.

3.2.1 Material and Methods

Cell lysis and protein extraction

Before lysis the cell pellets were washed twice with with ice-cold PBS by centrifuging at 12,000 g for 4 min at 4°C. The washing steps were crucial to avoid culture medium in the samples. Then the cell pellet was resuspended in lysis buffer (30 mM Tris, 7 M urea, 2 M thiourea, 4 % (w/v) CHAPS at pH 8.5). The cells are incubated with the lysis buffer on ice for 30 min. To separate cell debris from the protein solution the cell lysate is centrifuged at 4°C for 20 min. The supernatant contains the extracted proteins. At this point the pH of the protein solution is still greater than 8.0. Protein concentration was determined using the Bradford method (Bradford, 1976).

Preparation of CyDye labeling

The protein samples are labeled using the fluorescent Cy-Dyes, the N-hydroxysuccinimidyl ester forms of Cy2, Cy3 and Cy5. For each gel one protein sample (50 μg total protein) is labeled with either Cy3 or Cy5. Label swap was performed in order to avoid dye specific labeling artifacts. Cy2 was used to label 50 μg of the pooled standard, which consisted of equal proportions of all samples used in the analysis. 50 μg protein extract is labeled with 400 pmol of CyDye dissolved in DMF (Dimethylformamide). For the labeling reaction samples are incubated for 30 min on ice in the dark. 1 mM lysine in water was used to stop the reaction.

Protein separation

After the proteins have been CyDye labeled an equal amount of buffer solution (8 M urea, 4 % CHAPS (w/v), 130 mM DTT, 2 % (v/v) pharmalyteT̂M 3-10) is added and the solution is incubated on ice for 10 minutes. All protein samples are combined and separated on one gel. The total volume of the sample is now adjusted to 450 μl with rehydration buffer (8 M urea, 4 % (w/v) CHAPS, 1 % (v/v) pharmalyte 3-10). The whole sample is loaded onto a 24 cm immobolized pH gradient (IPG) strip with a pH gradient ranging from pH 4 to pH 7 (the IPG strips were purchased from Biorad).

- *Isoelectric focusing on 24 cm strips with immobilized pH gradient from 4 to 7*
  The isoelectric focusing was performed in the dark using a Biorad Protean IEF Cell with the following steps (Tbl. 3.2):

Table 3.2: Isoelectric focusing steps used for all 2D gel analyses.

Step	Voltage (V)	Time
1	0-100	1 min
2	100	120 min
3	100-1,000	20 min
4	1,000	30 min
5	1,000-4,000	60 min
6	4,000	30 min
7	4,000-10,000	60 min
8	10,000	70,000 Vh

- *Reduction and alkylation* After the first dimension separation, the pH strips are incubated for 15 min in 10 mg/ml DTT (dissolved in 10 ml 6 M urea, 4 % SDS, 0.05 M Tris pH 8,8, 30 % glycerol), followed by a 15 min incubation in 40 mg/ml iodoacetamide (IAA).

- *Polyacrylamide gel electrophoresis*
  The second dimension separation was a 12 % SDS-PAGE in the dark with the following steps (Tbl. 3.3):

Table 3.3: Running conditions for PAGE used for all 2D gel analyses.

Step	Voltage (V)	Current (mA)	Time
1	300	25	12 - 20 h
2	300	2	up to 20 h

Scanning of gel images

The gels are scanned three times on a FLA-5100 fluorescent scanner (FujiFilm). Three different wavelengths are required for scanning of the different labels. The Cy2 images are scanned with a wavelength of 488 nm (emission filter settings: 520 nm BP 40), 532 nm (emission filter settings: 580 nm BP 30) is used for the Cy3 label and the 635 nm (emission filter settings: 670 nm BP 30) is necessary for Cy5. After scanning, the gels are post-stained with silver to enable spot picking.

Silver post-staining

- Fixation of proteins: Proteins are fixed using a fixation solution (40 % EtOH, 10 % acedic acid. This step is necessary to remove all substances on the gel that show high affinity to silver (e.g., Tris, SDS) (Poland et al., 2005). The gel is incubated in this solution for 30 min.
- Sensitization of proteins: To enhance the silver binding to the proteins, the gels are incubated for 30 min in the sensitization solution (0.66 % Na-thiosulfate, 22.6 % Na-acetat disolved in 100 % ethanol).
- The gel is washed 3 times for 5 min in ddH_2O.
- Staining: The gel is stained with silver staining solution (0.5 % silver nitrate in ddH_2O) for 30 min.
- The gel is washed for 5 min in ddH_2O.
- Development. The gel is incubated in the development solution (2.5 % Na-carbonate, 0.05 % formaldehyde (37 %), 0.025 % of freshly prepared 10 %-Na-thiosulate solution, in ddH_2O) until protein spots get visible.

- Sopping: The reaction stopped by adding a 0.5 % glycine solution.
- Storing: After washing with ddH_2O the gel can be stored.

Image analysis

The gel images are analyzed with the ProteomWeaver software and the GBD method as described above.

Spot peaking

The spots of interest are excised form the gels and in-gel digested by trypsin, following the protocol as described by Resch et al. (2006).

Mass spectrometric analysis

Proteins of interest were analyzed on a QSTAR instrument, as described by Resch et al. (2006).

STRING database

The STRING database contains known interactions between proteins (Snel et al., 2000). The information embedded in this database includes different sources, such as genomic context, data from high throughput experiments, co-expression and previous knowledge from literature databases. This information can be used to find interacting partners in a dataset.

3.2.2 Results

We used DIGE in combination with the GBD analysis method to profile the differential protein expression in the human melanoma cell line 451Lu. For this experiment the cells were treated with either DMSO or 13 μM sorafenib for three, six and twelve hours. For one DIGE experiment either the control (DMSO) or the treated extract was labeled with Cy3 or Cy5 respectively. The Cy2 label was used to label the pool of extracts from all samples. For each time point and treatment three independent biological samples were used. A typical DIGE map that was stained with silver after the CyDye images were taken is shown in Fig. 3.5. For the GBD analysis five unambiguous protein spots were selected

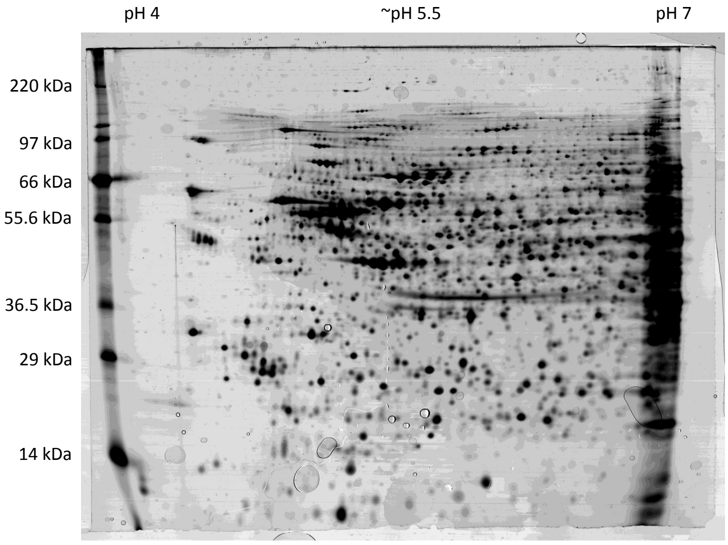


Figure 3.7: Whole proteome separation of 451Lu cells.

manually as landmarks (Tbl. 3.4).

Table 3.4: Landmarks for GBD analysis. These proteins were found to be unambiguously found in all maps. The protein ID was confirmed by tandem mass spectrometry.

Landmark	Protein ID	MW (in kDa)
1	PSME3_HUMAN	29,506
2	TIM50_HUMAN	39,646
3	KGP2_HUMAN	87,432
4	UAP1_HUMAN	58,769
5	IMMT_HUMAN	40,491

Differential regulation

For every experiment 83 protein spots were quantified. These 83 protein spots resulted in 116 protein identification. Not all proteins could be unambiguously identified by our identification method. A heat map of the protein ratios across all time points is shown in Fig. 3.8. Heat maps can be used for the graphical representation of large-scale expression data. Values from any input matrix are represented by colors. The color index from green to red in this heat map corresponds to log_2 ratios calculated from the CyDye intensities. The color index on the left to the heat map indicates which spots contained more than one protein. In spot 16 four different proteins were found: FZD6, OBSCN, NPHP3 and FBXL18. These proteins are marked in yellow. The heat map analysis revealed mainly three sets of proteins. Proteins that are not regulated at any time point after sorafenib treatment, proteins that are up-regulated or proteins that are down-regulated. The ratio of non regulated proteins to up-regulated proteins to down-regulated proteins is approximately 36:39:25. With the help of the STRING database (Snel et al., 2000), known interactions of the differentially regulated proteins were examined. The circles in Fig. 3.9 correspond to proteins, found as regulated. Red color symbolizes up and green color down-regulation. This analysis revealed three distinct interaction maps of proteins. The edges in between the protein nodes in Fig. 3.9 are known interactions, whereas the magenta edges are the most reliable interactions, as they are confirmed by experiments. In Fig. 3.9 Annexin V (ANXA5) is connected via a magenta edge to γ-actin (ACTG1). This interaction has been confirmed by (Tzima et al., 2000). In this study it was shown that annexin V specifically binds to γ-actin and not to other actin isoforms, as β-actin. This interaction analysis reveals that proteins found to be regulated form an interaction map that can be assigned to apoptosis and cellular localization, as well as network of interacting proteins that can be assigned to RNA splicing, translation, and the proteasome.

Differentially regulated isoforms of β-actin and vimentin

Two important areas of the DIGE map are highlighted in Fig. 3.10. The zoom area on the left shows four different isoforms of vimentin. The isoforms of vimentin are not only shifted with respect to their isoelectric point, but also in their molecular weight. The analysis revealed four different isoforms of vimentin with different sensitivities to different lengths of sorafenib treatment. The more acidic and shorter isoforms are regulated stronger as the less acidic and longer isoforms. A similar observation was made for the actin isoforms. In contrast to

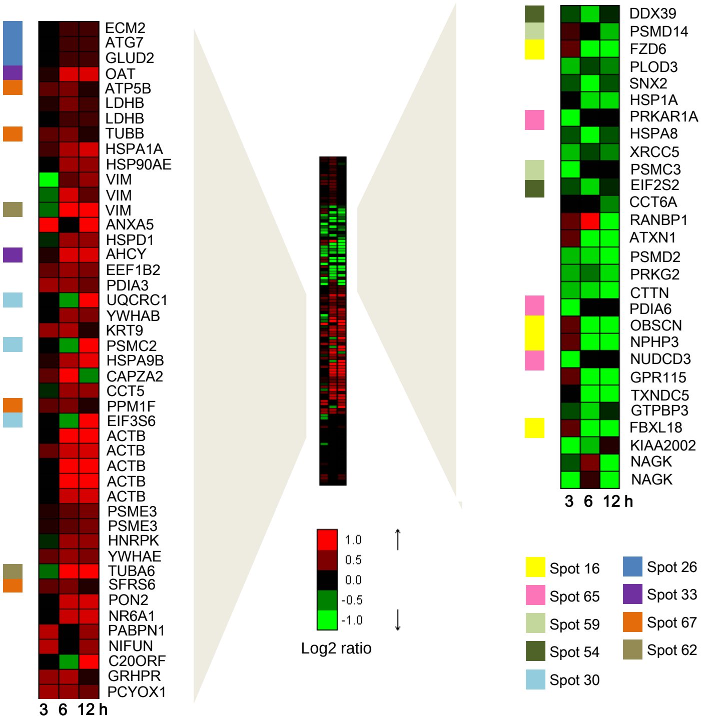


Figure 3.8: This heat map shows up- and down-regulated proteins. The heat map colors indicate the expression ratios, whereas the colors in lower right corner group proteins the are unambiguously assigned to several protein spots.

vimentin, only isoforms with equal molecular weights are observed. The shift is only observed in the isoelectric point of the proteins. The ratios of the more basic isoforms are higher, the longer the cells are treated with the inhibitor.

3.2.3 Discussion

Differential regulation

2D-PAGE in combination with the DIGE labeling method offers several advantages over conventional 2D-PAGE. Differential labeling before protein separation allows separating multiple proteomes in one 2D gel, resulting in a perfect matching for those proteome images and the inclusion of internal standards allows a better statistical assessment, since all ratios can be normalized to this standard.

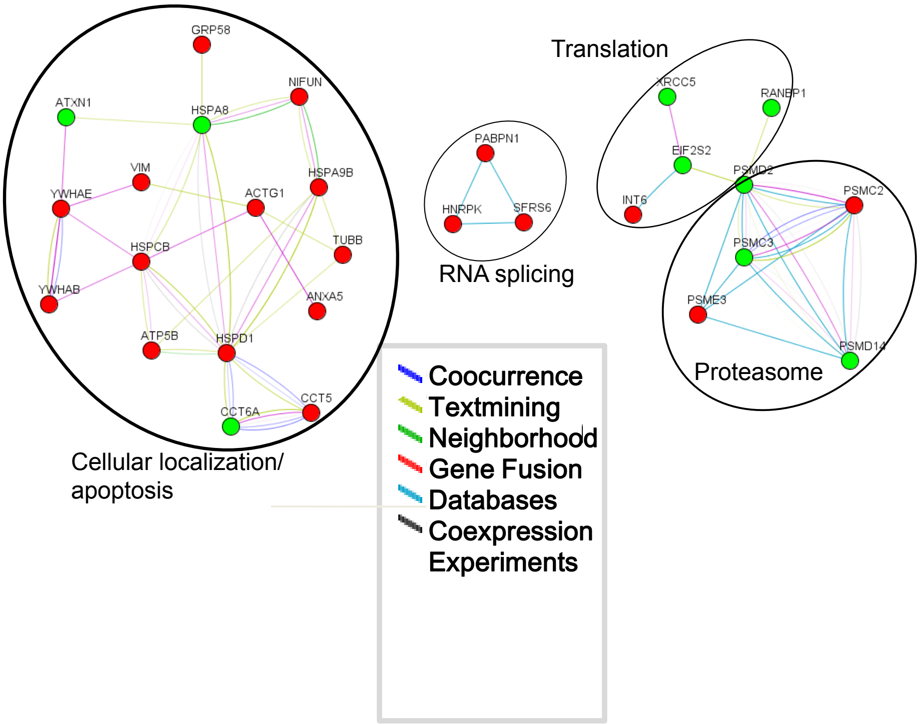


Figure 3.9: Regulated proteins that have known interactions. This analysis was performed using the STRING database.

Using our new map alignment method we identified 83 protein spots across three different time points and at least three independent biological replicates for each time point. Due to the complexity of the proteome it was not always possible to assign proteins unambiguously to the protein spots. The complexity of the proteome is by far higher than the proteins detected by scanning the 2D images. Proteins might hide under the spots that were actually detected and only the LC-MS analysis of the excised region allows finding those proteins. This phenomenon is even more significant if 2D gels are run on broader pH ranges. The separation in the first dimension can also be done on a pH range from 3 to 10. Using broad pH windows results worse resolution of the protein spots. For this and other reasons the pH range pH 4 to pH 7 was chosen for this study. It can be seen in Fig. 3.7 that the protein spot density is reduced below pH 4.5. The resolution of very basic proteins on a 2D map has traditionally been very difficult

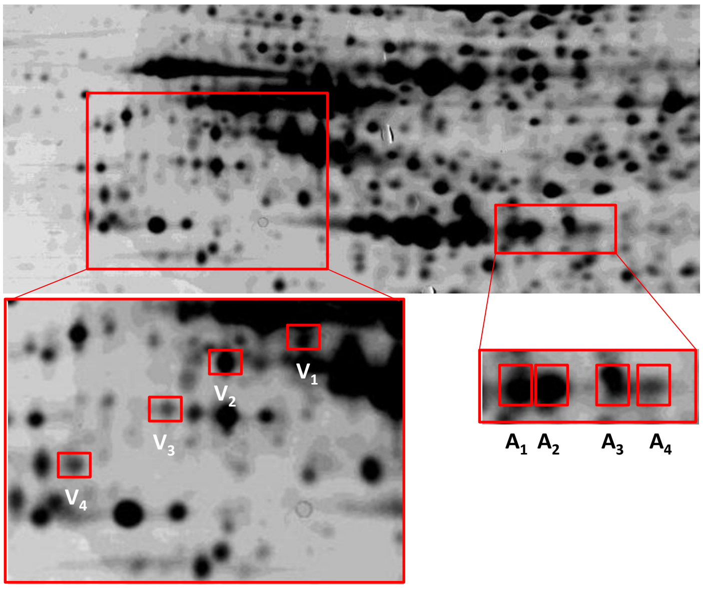


Figure 3.10: Detection of differentially regulated isoforms. The protein identification for every protein spot was done using tandem mass spectrometry.

(Yamaguchi and Pfeiffer, 1999). We found 64 protein spots to be differentially regulated. Sorafenib is known to inhibit major signaling pathways, such as the MAPK pathway. As most signaling events converge in the nucleus to regulate gene expression, the number of differentially expressed proteins following the inhibition of signaling, should be much higher than this. 2D-PAGE is prone to detect only the most abundant proteins, which might not be under regulation of the inhibited kinases. Some of the proteins that were found to be differentially expressed, were previously known as interaction partners. All annotations from the STRING database were used to investigate proteins for their known interactions among each other. If protein interactions can only be based on text mining results, these interactions are not as reliable as interactions based on experimental evidence. The STRING analysis revealed three functional units of proteins that have annotated interactions. The cellular localization group includes cytoskeletal proteins that are connected by edges that are based on experimental evidence.

Furthermore the connected proteins were regulated in the same direction. These results, in combination with sorafenib's influence on β-actin and vimentin isoforms, suggest the modulation of cytoskeletal and/or filamentous proteins as a contribution to the sorafenib-induced cell death.

Differentially regulated isoforms of β-actin and vimentin

The shift in the isoelectric point that is observed for the four actin isoforms might be due to phosphorylation events. Similar shifts have been observed for phosphorylated isoforms of cofilin (Moriyama et al., 1996). The actin spots that are at a higher pH range should correspond to less phosphorylated isoforms. The intensity of these isoforms increases with the length of treatment, suggesting that sorafenib has an influence on the phosphorylation of β-actin. It remains elusive whether this influence is a direct interaction or it is mediated via intermediate kinases. For vimentin similar observations were made. The acidic forms of vimentin may correspond to phosphorylated isoforms. The intensities of these isoforms were decreased as a function of treatment length. This also suggests that sorafenib treatment not only influences β-actin, but also vimentin phosphorylation. Along these lines, it is known that vimentin phosphorylation is strongly influenced by the p21 activated kinase (PAK), which is downstream of the cell division control protein 42 (CDC42) and and RAC, two major Rho GTPases. PAK was shown to phosphorylate vimentin on multiple serine residues (Goto et al., 2002). Rho-kinases, in general, are putative regulators of vimentin filament organization downstream of Rho, furthermore from Goto et al. (2002) it is known that PAK, being downstream of Cdc42/Rac, may regulate vimentin filament reorganization through vimentin phosphorylation. PAK and Rho-kinase were shown to phosphorylate and activate LIM-kinase. This activation of LIM-kinase leads to more stable filamentous actin structures through an inhibition of cofilin by phosphorylation. Therefore, PAK and Rho-kinase may phosphorylate some common targets (LIM-kinase, vimentin, etc.), leading to cytoskeletal rearrangements, such as actin filament stabilization and IF (intermediate filament) reorganization (Goto et al., 2002). The destabilization of cytoskeletal rearrangements in contrast might be correlated to the cell-death-inducing effects of sorafenib. The observed vimentin spots were also shifted with respect to their molecular weight. It is consistently observed that the acidic isoforms have also undergone proteolytic cleavage. Cleaved vimentin has been known as being accompanied with apoptosis. Vimentin is known to be cleaved by a caspase-3/7-like protease during apoptosis and subsequently by caspase-6 at additional sites (Yang et al., 2005). Furthermore, Yang et al. (2005) speculate that orchestrated cleavage of vimentin can be an initiation process for dramatic reorganization of the cytoskeleton that are characteristic for apoptotic cell death. This would imply essential roles of vimentin in apoptosis and its regulation. The hypothesis of differential phosphorylation needs further confirmation. The tandem MS spectra from the excised protein spot did not deliver a clear answer. Gel-free shotgun proteomics experiments allow targeting of phosphopeptides. Such workflows would be promising future projects to confirm these hypotheses.

Chapter 4

Probabilistic consensus scoring

4.1 Introduction

Chapter 2 introduced different methods used for peptide assignment of tandem MS spectra. Although the methods rely on the same underlying basic principle, namely the fragmentation into defined parts (e.g. b- and y-ions), the implementations differ strongly. The methods used by these algorithms are diverse. A detailed description of additional algorithms can be found in a recent review (Nesvizhskii et al., 2007). Typically, all database search algorithms for spectral assignments produce a list of peptides that are ranked according to their scores. Due to shortcomings of the scoring, the first sequence in the list does not necessarily correspond to the correct identification, but might be just a random hit. In fact, there are many cases where the correct sequence is not even contained in the result list of the search engine. In addition, there is a high variation of suggested peptide candidates in the search results from different search engines (Kapp et al., 2005). Scores produced by search engines are often difficult to interpret and to compare. A common approach to normalize search engine scores, using the target-decoy search strategy, has already been introduced in the background chapter. Besides the q-value approach there have been several other approaches for converting the search engine scores into more reliable numbers, e.g., probabilities. Keller et al. (2002a) offered one of the first statistical approaches for converting Sequest scores into probabilities. Their algorithm is based on maximum likelihood estimation of empirically assumed probability distributions and an Expectation Maximization (EM) (Dempster et al., 1977) framework to assess optimal parameters for the mixture model deconvolution. The PeptideProphet method has been widely accepted for statistical assessment and recently been extended to support not only Sequest but also Mascot and X!Tandem results (Choi and Nesvizhskii, 2008). Essentially, the null hypothesis in the mixture model approach is the same as in the target-decoy q-value approach. All peptide scores that fall into the distribution described by the decoy results or into the first component in the mixture model, respectively, are assigned to the null hypothesis, *random chance identification*.

Despite all the effort put into the conversion of intransparent scores from search engines into probabilities, or the assignment of false discovery rates (FDR) (Käll et al., 2008a), the divergence of results from different engines remains largely unused. According to Kapp et al. (2005), only one third of all peptides in an experiment are identified by all engines. Simple consensus identification by voting can enhance the reliability of the identification, but at the cost of a lower number of identified peptides. Combination of the results of search engines becomes difficult, however, if peptides are not scored by all search engines, i.e., if the candidate sequence is not reported by all search engines.

4.1.1 Related work

There have been several approaches describing methods for the combination of different search engine results. In 2008, Searle et al. (2008) suggested a combination method based on Mascot, X!Tandem and Sequest scores. This approach is implemented in the commercial software Scaffold . Peptide probabilities are individually estimated for the search engines. To combine the results from the different search engines, an agreement score is used to account for differences in the significance of the individual scores, if the same peptide was assigned by several engines. Another search engine combiner is PepArML (Edwards et al., 2009). PepArML uses unsupervised machine learning to account for both, the statistical significance and the combination of the scores from the different engines. This approach relies on an iterative process of a random forest learning method. A peptide is assigned a vector of 27 features, containing general information, e.g., m/z and retention time values, and search engine specific information, such as the raw engine score. Another machine learning based tool is iProphet (Shteynberg et al., 2008), which is integrated in the Trans-Proteomic Pipeline (TPP). PeptideProphet (Keller et al., 2002a) builds the basis of iProphet, which performs additional EM estimation to derive a common probability for multiple search runs.

Common to all approaches aiming at combining different tandem MS search runs is the analysis of the results from the same search conducted in parallel using different search engines. This is usually done under the general premise that search engine agreement is correlated with the correctness of peptide identifications (Tharakan et al., 2010). All approaches show an improvement over the single engines, but are in most cases designed for specific search engines and/or instrument types. Furthermore the complex feature encodings for some of the algorithms described above make it very hard to subsequently interpret the search results. Another approach that has recently gained attention in the post processing of tandem MS search results is the 'multi-pass analysis', as introduced by Tharakan et al. (2010). Multi-pass methods combine multiple searches from one engine, by guiding the selection of spectra, parameters, and sequences in subsequent searches based on previous search results. Compared to searches with multiple search engines, multi-pass strategies have a reduced run time, as additional search runs are usually performed on a subset of spectra or against a reduced database. In Chapter 2 a commonly used multi-pass strategy has already been introduced at the example of X!Tandem. The refinement function of X!Tandem automatically constructs a new database, using proteins that have already been identified in a primary run. This database is then searched with an increased number of potential modifications, allowing more missed cleavages and semi-tryptic peptides (peptides where only one terminus corresponds to the trypsin cleavage definition).

4.1.2 General idea

Here, we describe a new method to integrate results from various search runs. These runs can either originate from multiple search engines or they can be multiple runs with the same search engine. The algorithm consists of two parts: First, mixture modeling is applied to convert search engine scores into probabilities. This step is essentially used to normalize the scoring from the different engines. Second, missing scores for peptide sequences in the search engine output are estimated. This score estimation is based on sequence similarity. The measure of sequence similarity between two peptide candidates correlates with the measure of fragment ion similarity. Applying both parts, probability scores are obtained for all search engines and each peptide candidate. This enables a combination of scores into a single consensus score. The performance of the method is evaluated on mixtures of known proteins, that have been measured on different instruments, as well as on a complex mixture, resulting from a whole *Escherichia coli* proteome digest. Besides increased confidence in the peptide identifications, using the novel *consensus scoring*, peptide identification rates can be significantly improved. Using four test datasets, acquired on four different MS platforms, this new method shows significant advantage compared to the performance of single engines. The identification rates using the *consensus scoring* improved consistently over the identification rates of the individual search engines at any q-value cut-off on datasets using Orbitrap, FT Ultra, and LCQ instruments. The target-decoy database approach is used as a significance measure of the final peptide scores.

4.2 Methods

The overlap between search results from different engines is rather poor and a large fraction of all peptides do not appear in the results from all search engines. Therefore, combining different engines holds the promise to increase sensitivity and specificity in peptide identification by tandem mass spectrometry. Furthermore, if spectra have been correctly identified by search engine k, but not by the others, these spectra should at least be assigned to sequences with similarity to the correct peptide, if the spectral quality is good enough to trust the identification by engine k.

The peptides that are correctly identified by a given search engine are most frequently ranked in the first place by all three search engines (Fig. 4.1). the numbers of additional peptides (add. peptides) that were identified by the different search engines are also visualized in Fig. 4.1. Additional peptides (add. peptides) in this context refer to identified peptides from target sequences (18 proteins), but with unexpected modification or cleavages. For X!Tandem the add. peptides are due to the refinement mode that was enables for this experiment and OMSSA/ Mascot allow the assignment of peptides from the N-terminus of proteins where the N-terminal methionine is cleaved.

Our strategy for the combination of results from several search engines for peptide identification via tandem mass spectrometry relies on the similarity of peptide sequences. *Peptide similarity scoring* is applied in cases of sequences missing in result lists of search engines. If sequence s that has been assigned to a spectrum by at least one search engine is not contained in the list provided by search engine k, the score for this sequence is imputed. Therefore all hits suggested by search engine k are aligned with sequence s to find the sequence with highest similarity to s. This sequence, accompanied with its similarity to s is used as a substitute. For the global pairwise sequence alignment, the Needleman-Wunsch algorithm is applied. With this approach and the weighted average-like (as described later) combination method, the most similar hit is used as a replacement sequence, but the method accounts for the uncertainty by multiplying its score by the proportion of sequence similarity between the target sequence and the replacement sequence. The influence of the score is thus reduced proportionally to the sequence similarity. This method allows assigning scores to each peptide sequence per search engine and ultimately to combine these scores to a *consensus score*. The Needleman-Wunsch algorithm, like most alignment algorithms, penalizes gaps that need to be introduced in the alignment. The penalization of gaps is implemented with two parameters. δ is the cost for opening a gap and ε is the extension penalty.

It is widely accepted to use mixture modeling for the conversion of search engine scores into probabilities (Nesvizhskii et al., 2007). The mixture modeling

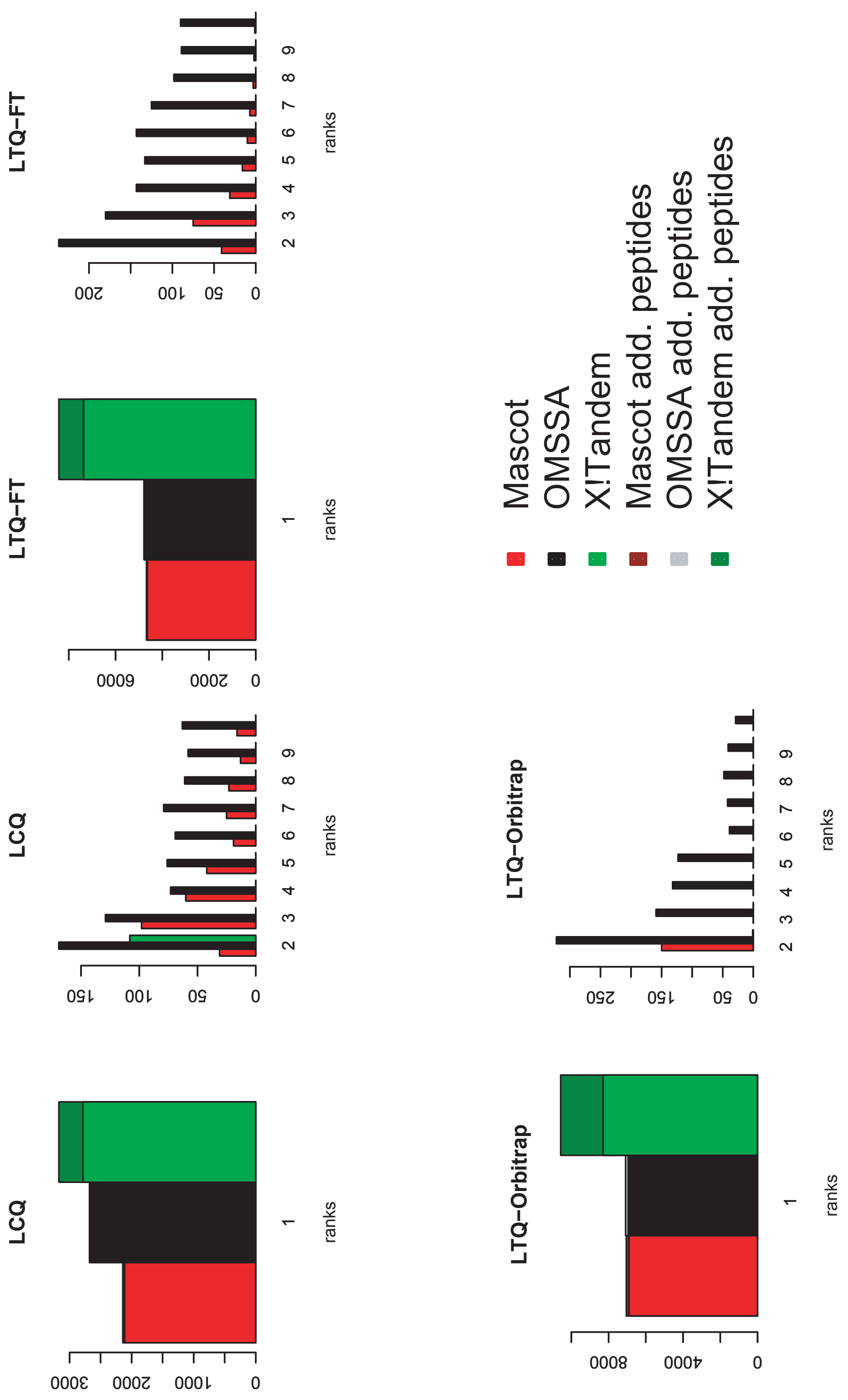


Figure 4.1: Correct peptides on different ranks. This data is based on the 18 protein mix datasets.

approach is implemented in OpenMS, our own C++ software framework. An earlier version of the mixture model was implemented in the statistical software R (R Development Core Team, 2008). Both implementations are freely available. Scores for peptide sequences not appearing in one search engine result list, but in the other are imputed by peptide similarity scoring. For the combination of the search results the similarity-weighted average score is calculated for each peptide.

4.2.1 Mixture modeling and Expectation Maximization

Mixture models are applied in various areas that range from biology, over physics to economics (Leisch, 2004). Usually such models are used to determine the group affiliation of observations in large datasets. The Expectation Maximization (EM) algorithm (Dempster et al., 1977) is a very popular framework for the estimation of mixture models. In the following we will formally introduce finite mixture models and the EM algorithm.
Given a set of observation $\{y_1, ..., y_n\}$. The set is divided into subpopulations and each subpopulation i is assigned the probability density function f_i.

$$f_i(y_j) = P(\text{observation } y_j \text{ is observed in subpopulation } i)$$

Weights π_i are assigned to each subpopulation i.

Definition 1. *Let $Y = \{y_1, ..., y_n\}$ be a set of N independent observations. Y consists of k different components with the weights π_i. Each component i is assigned a probability distribution f_i. Then*

$$f(Y) = \sum_{i=1}^{k} \pi_i f_i \ \ (Y)$$

is called the probability mixture model of Y. f is a convex combination of the probability distributions f_i. If k is finite, then f is called finite mixture model.

In a finite mixture model, each component i is modeled by its own set of parameters Θ_i. Mixture models provide a complex probability distribution that can be used to cluster data (Bishop, 2007). An easy example for mixture models is the mixture of Gaussian distributions

$$p(x) = \sum_{j=1}^{k} \pi_j \ N\left(x|\mu_j, \sigma_j^2\right)$$

Definition 2. *Given two jointly distributed random variables X and Z, the marginal distribution of X, $p(x)$ is the probability distribution P of X averaging over information on Z.*

$$p(X=x)=\sum_{y} P\left(X=x,Z=z\right)=\sum_{y} P\left(X=x|Z=z\right)\ P\left(Z=z\right)$$

The Corollary 1 in the appendix shows that the marginal distribution of x,

$$p(x)=\sum_{z}\ p(x|z)=\sum_{j=1}^{k}\pi_j N(x|\mu_j,\sigma_j^2)$$

is a mixture of Gaussian distributions with parameters μ_j and σ_j and $\forall\ x_n\ \exists\ z_n$. z_n is called the latent variable of x. For $z_k=1$ we can write

$$p(z_k=1|x)=\frac{p(z_k=1)\ p(x|z_k=1)}{\sum_{j=1}^{k} p(z_j=1)\ p(x|z_j=1)}$$

$$\Leftrightarrow p(z_k=1|x)=\frac{\pi_k\ N(x|\mu_k,\sigma_k^2)}{\sum_{j=1}^{k}\pi_j\ N(x|\mu_j,\sigma_j^2)}=\gamma(z_k)$$

π_k are called the prior probability for component k, whereas $\gamma(z_k)$ is called the posterior probability for the latent variable z_k. $\gamma(z_k)$ can also be interpreted as the responsibility that the component k takes for *explaining* the observation x. Using the Expectation Maximization algorithms, the E-step evaluates the responsibilities for initial guesses of the paramters Θ_i, for each distribution. And the M-step re-estimates the parameters to maximize the log-likelihood function.

Given the mixture model $f(x)$, the parameters of the distribution are re-estimated using the current reponisibilities $\gamma(z_k)$.

$$\mu_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

$$\sigma_k^{new}=\sqrt{\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\mu_k^{new})(\mathbf{x}_n-\mu_k^{new})^T}$$

$$\pi_k^{new}=\frac{N_k}{N}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

within a maximum likelihood framework the new parameters are used to recalculate the log-likelihood function.

$$ln\ p(X|\pi, \mu, \sigma^2) = \sum_{n=1}^{N} ln \sum_{j=1}^{K} \pi_j N(x_n|\mu_j, \sigma_j^2)$$

The E- and M-steps are repeated until there is no improvement for the log-likelihood function anymore. The mixture modeling framework, as well as the Expectation Maximization algorithm are applied to the conversion of intransparent search engine scores to error probability scores, indicating the probability that a given score falls into the distribution of false identifications. For each search engine, we consider n spectra. The scores from engine k

$$\mathbf{x_k} = (x_{k_1}, \ldots, x_{k_n})$$

can be modeled as n independent and identically distributed (i.i.d.) random variables. The distribution of these scores is modeled by a two-component mixture model with the function f given by

$$f(x; \Theta_1; \Theta_2) = \pi f_1(x, \Theta_1) + (1 - \pi) f_2(x, \Theta_2) \tag{4.1}$$

where π corresponds to the prior probability of the scores being incorrect. Here, incorrect means that the spectrum is assigned to an incorrect peptide sequence. The functions f_1 and f_2 are the densities for incorrectly and correctly assigned sequences, respectively. The parameters Θ_1 and Θ_2 are used to specify the exact shape of the densities. The function f_1 is modeled as the density of a Gumbel distribution. The use of extreme value distributions as a model for the function f_1 has been introduced as a generic method for the statistical assessment of peptide-spectrum matching scores (Fenyo and Beavis, 2003) and successfully applied to X!Tandem and Mascot searches (Choi and Nesvizhskii, 2008; Searle et al., 2008). An extreme value distribution is a natural candidate for modeling maximized scores (tailing to the higher score regions) from incorrectly assigned sequences. The function f_2 is modeled as a Gaussian density. To perform the estimation of the parameters, an EM framework was implemented. The Expectation step (E-step) comprises the estimation of posterior probabilities, as formalized in equation (4.2), using initial guesses for $\hat{\pi}$, $\hat{\Theta}_1$ and $\hat{\Theta}_2$. This step is followed by the Maximization step (M-step), where the estimated posterior probabilities are used to refit the distributions f_i. With this iteration, the log-likelihood function (4.3) is maximized and the algorithm stops when there is no improvement of the log likelihood function anymore.

$$\hat{p}_i(x) = \frac{\hat{\pi}_i f_i(x, \hat{\Theta}_i)}{\hat{\pi}_1 f_1(x, \hat{\Theta}_1) + \hat{\pi}_2 f_2(x, \hat{\Theta}_2)}, i \in \{1, 2\} \tag{4.2}$$

$$logL = \sum_{i=1}^{n} log(\hat{\pi}_1 f_1(x_i; \hat{\Theta}_1) + (1 - \hat{\pi}_1 f_2(x_i; \hat{\Theta}_2))) \tag{4.3}$$

where Θ_1 is the set of all parameters for the probability distribution f_1. f_1 corresponds in our case to an extreme value distribution with the location parameter α and the scale parameter β. Θ_2 includes the parameters for the f_2 function, which are the mean μ and the variance σ^2 for a Gaussian distribution. Initial parameters for our model are found by employing an ordinary Gaussian mixture model (two Gaussian distributions), as implemented in the flexmix function (Leisch, 2004) to the scores for each search engine. This method allows accurate conversion of search engine scores into probabilities. The input for the mixture modeling is given by discriminate scores from the search engine output. For all search engines the discriminant scores are the negative common logarithms of the search engines' E-values. To ensure a fair and comparable competition of all search engines for the same set of peptides X!Tandem was used with and without enabling the refinement function. X!Tandem searches with the refinement function include modified peptide sequences or sequences that include polymorphisms. Those sequences cannot be expected from OMSSA and Mascot searches. However the set of peptides that are identified by X!Tandem without refinement is directly comparable to the results form the other searches.

4.2.2 Consensus scoring

Different ways of calculating consensus scores for peptides candidates are evaluated. A typical workflow for generating *consensus scores* on the basis of three single search engines and the similarity-based consensus measures (peptide sequence similarity or SPC) is described in Fig. 4.2. In the example shown in Fig. 4.2 Mascot and X!Tandem suggest the same sequence as their top hit, whereas the same sequence is only ranked forth by OMSSA. α and β denote the similarities between sequences if not the same sequence is suggest by the different engines.

In the following the different combination methods will be discussed in detail. Results will be evaluated based on peptide sequence similarity using different scoring matrices, spectral counter, and average scoring methods.

Peptide sequence similarity scoring

Comparing search results from different engines, it can be observed that for some spectra peptide sequences occur only in a subset of the search engine results. The method presented here solves this problem by imputing missing scores based on similarity. This method ensures that there is an (estimated) value for each peptide occurring in any of the search engine lists. Tandem MS spectra characterize peptide sequences based on their fragmentation patterns. Peptide search engines use this information on peptide fragmentation, e.g., y and b ions for CID (collision-induced dissociation) (Steen and Mann, 2004). Sequences that contain isobaric amino acids, such as I and L, are not distinguishable based on their tandem MS spectra. The chemical similarity of peptide sequences is thus used to determine the probability that a *missing* peptide can be assigned to a spectrum. The converted score of the peptide showing the highest degree of similarity is multiplied by its similarity in order to assign values to the missing peptide. As a similarity measure, global sequence alignments with the PAM substitution matrix based on the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). The alignment score for two peptide sequences p_i and p_j is then normalized as follows:

$$\mathrm{sim}(p_i, p_j) = \max \begin{cases} \frac{\mathrm{score}(p_i, p_j)}{\min(\mathrm{score}(p_i, p_i), \mathrm{score}(p_j, p_j))} \\ 0 \end{cases}$$

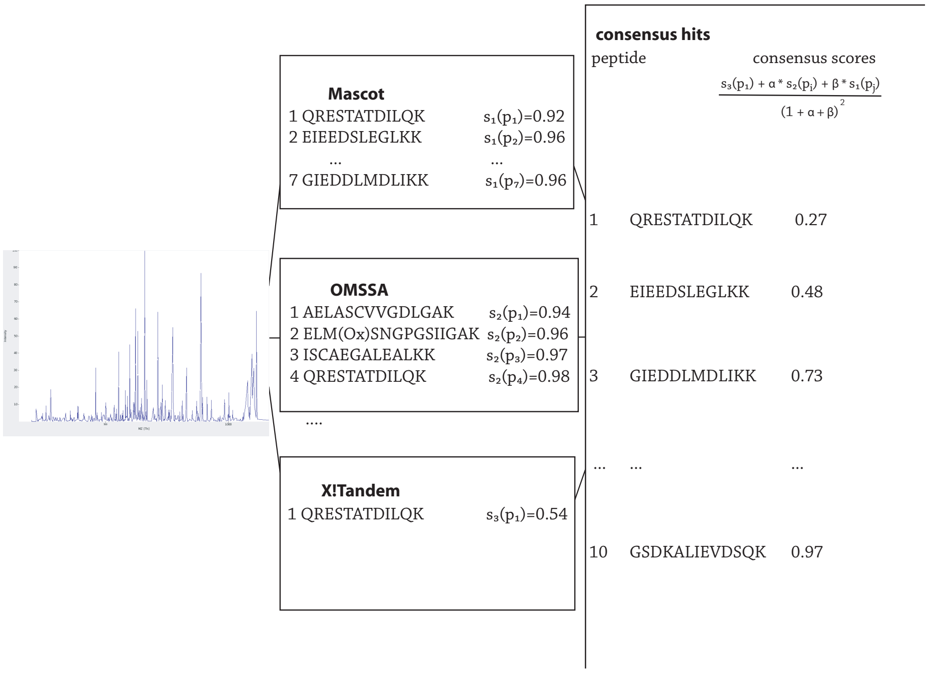

consensus hits
peptide
consensus scores
s₃(p₁) + α * s₂(pⱼ) + β * s₁(pⱼ)
(1 + α + β)²
Mascot
1 QRESTATDILQK s₁(p₁)=0.92
2 EIEEDSLEGLKK s₁(p₂)=0.96
...
7 GIEDDLMDLIKK s₁(p₇)=0.96
OMSSA
1 AELASCVVGDLGAK s₂(p₁)=0.94
2 ELM(Ox)SNGPGSIIGAK s₂(p₂)=0.96
3 ISCAEGALEALKK s₂(p₃)=0.97
4 QRESTATDILQK s₂(p₄)=0.98
....
X!Tandem
1 QRESTATDILQK s₃(p₁)=0.54
1 QRESTATDILQK 0.27
2 EIEEDSLEGLKK 0.48
3 GIEDDLMDLIKK 0.73
... ...
10 GSDKALIEVDSQK 0.97

Figure 4.2: Three search engines assign peptide sequences to a given spectrum. This spectrum is taken from the *E. coli* dataset (measured on an LTQ-Orbitrap) and corresponds to a doubly charged ion with RT: 2233.105 s and MZ: 695.3646 Th. Mascot and X!Tandem list the same peptide sequence as the top hit. The consensus score is calculated by the weighted combination of the real and estimated scores for the given sequences.

The similarity-based consensus score is calculated as follows,

$$\text{Similarity}_e(p_i) = \frac{s_e(p_i) + \sum\limits_{k \in \{\text{E}\} \setminus \{e\}} \hat{s}_k(p_i)}{\left(1 + \sum\limits_{k \in \{\text{E}\} \setminus \{e\}} \frac{\hat{s}(k)}{s_k(p_i)}\right)^2} \tag{4.4}$$

$$\text{with} \quad \hat{s}_k(p_i) = \begin{cases} s_k(p_i), & \text{if } sim(p_i, p_j) = 1 \\ s_k(p_j) \cdot sim(p_i, p_j), & \text{with}\, p_j = \underset{j \in \text{ list of engine}\, k}{\text{argmax}} \; sim(p_i, p_j), \text{otherwise} \end{cases}$$

For a given spectrum all candidate peptide sequences p_i are used. The consensus scores for all peptide sequences are calculated for each search engine e separately. E corresponds to the set of search engines in use. The consensus score for the peptide sequence p_i assigned by engine e for the spectrum S is calculated. The list of candidates for spectrum S consists of a ranked list of all consensus results. As shown later, the PAM30MS matrix was found to be the most suitable substitution matrix. The PAM30MS was previously introduced and intended for cross-species proteomics (Huang et al., 2001). This matrix is constructed and modified from the PAM30 matrix to account for Ile/Leu and Gln/Lys ambiguities associated with determination of peptide sequences using tandem mass spectrometry. Details on the PAM30MS substitution matrix can be found in the appendix.

Spectral counter (SPC) scoring

As a comparable measure for peptide candidate similarity, the theoretical b and y ion series of two sequences are compared. In a similar way as outlined above for the sequence similarity, the SPC similarity measure calculates the ion series similarity. At least two fragment masses are requested to be within a given mass tolerance window, which was set to 0.5 Da. The normalized ion series similarity is then calculated as follows:

$$\mathrm{spc}(p_i, p_j) = \frac{\text{number of overlapping fragment ions (b and y ions)}}{\text{2 * length of smallest sequence}}$$

The final consensus score is calculated in the same way as described for the peptide sequence similarity. The SPC method used the number of overlapping theoretical fragments as a measure of similarity.

$$\mathrm{SPC}_e(p_i) = \frac{s_e(p_i) + \sum\limits_{k \in \{\mathrm{E}\} \setminus \{e\}} \hat{s}_k(p_i)}{\left(1 + \sum\limits_{k \in \{\mathrm{E}\} \setminus \{e\}} \frac{\hat{s}(k)}{s_k(p_i)}\right)^2}$$

$$\text{with} \quad \hat{s}_k(p_i) = \begin{cases} s_k(p_i), & \text{if } spc(p_i, p_j) = 1 \\ s_k(p_j) \cdot spc(p_i, p_j), & \text{with}\, p_j = \underset{j \in \text{ list of engine}\, k}{\text{argmax}} \quad spc(p_i, p_j), \text{otherwise} \end{cases}$$

Note that for $p_i = p_j$ we obtain for both similarity measures $\text{sim}(p_i, p_j) = \text{spc}(p_i, p_j) = 1$.

Average scoring

The Average method calculates the average score, if the the same peptide is suggested by several scores.

$$\text{Average}_e(p_i) = \begin{cases} \frac{s_e(p_i) + \sum\limits_{k \in \{\text{E}\} \setminus \{e\}} s_k(p_i)}{L} \\ s_e(p_i), \text{ if } p_i \text{ is only assigned by engine e} \end{cases}$$

Datasets

To properly assess the performance of the method, different datasets of known protein mixtures were used. The first dataset is the ISB dataset (Keller et al., 2002b) (denoted as ISB1), which is a mixture of 18 proteins acquired on an LCQ DECA XP instrument (ThermoFinnigan, San Jose, CA). Additionally, three data sets from the newer ISB collection (denoted as ISB2) (Klimek et al., 2008) were used. To capture a variety of instruments, high-accuracy FT instruments such as the Orbitrap (Thermo Finnigan) and an FT Ultra (Thermo Finnigan) mass spectrometer, were included. Both platforms record their tandem spectra in a low resolution LTQ device. The fourth platform was an LCQ instrument, where both precursor and fragment ions are recorded in low resolution ion trap mode. Additionally, we used a complex dataset from a whole *E. coli* digest. This dataset was generated in-house. The peptides were separated on an easyLC (Proxeon) system, online coupled to an LTQ-Orbitrap. The peptide mixture was eluted from the column with a 224-min segmented gradient from 5 % to 80 % HPLC solvent B (80 % ACN in 0.5 % acetic acid) at a flow rate of 200 nL/min.

For the generation of peptide spectrum matches, Mascot, version 2.2, OMSSA, version 2.1.4 and X!Tandem, version 2008.02.01.3, were used. The modification settings were carbamidomethylation of cysteine as fixed and oxidation of methionine as variable modification. The peptide identification of our analysis was done in a two-step process. First, the identifications were performed using a precursor mass tolerance of 3.0 Da and a fragment mass tolerance of 0.5 Da. These numbers are wide enough to cover also the needs from the low-resolution instruments and appear to be the most appropriate settings to provide an instrument-independent identification pipeline. After the first identification run, the optimal tolerance values were estimated by using the peptide identifications to calculate the distribution of the errors of the precursor and fragment masses. Precursor masses were compared to the m/z values contained in the precursor information of the tandem MS spectra. Fragment mass errors were calculated using singly charged b and y-ions derived from the peptide sequence and the nearest peak within the mass tolerance (in our case 0.5 Da) in the experimental tandem mass spectra, if available. To avoid wrong error distributions, only peptide spectrum matches with an q-value (Käll et al., 2008b) of 0.01 or better were used, to estimate the optimal tolerance settings. The final tolerances were estimated manually using the error distributions. On high-resolution instruments, relative tolerances in ppm were preferred over absolute tolerances. Except for OMSSA, the search engines allow precursor settings in ppm. Fragment tolerances were always set in Da, because all tested instruments record tandem mass spectra in low-resolution mode. The error distributions observed for the Orbitrap data are shown in 4.3(a) for the MS1 data and in 4.3(b) for the MS^2 data, respectively. The tolerance settings used in the final identification runs are listed in Tbl. 4.1. If relative tolerances in ppm were not accepted by any search engines, the following absolute tolerances were set: 0.01 Da for 10 ppm and 0.03 Da for 30 ppm. The

Table 4.1: Mass tolerance settings. These tolerance values were estimated using pre-searches with large tolerance windows.

	Orbitrap	FT Ultra	LCQ
Precursor tolerance	10 ppm	30 ppm	1.5 Da
Fragment tolerance	0.5 Da	0.5 Da	0.5 Da

data for the complex *E. coli* mixture was run on an LTQ-Orbitrap instrument sequential to a four LC separation. The *E. coli* was searched against against two different databases in order to evaluate the influence of the database size to the performance of the consensus scoring. The first database contained all known *E. coli* open reading frames (Riley et al., 2006), known contaminants and reversed versions of all proteins. In total the *E. coli* database contained 8272 protein sequences. The second the database with the complete SwissProt database. The concatenated (forward/reverse) version of this database contained 712,388 protein sequences For this dataset the same search engine specific parameter were used as for the 18 protein mix LTQ-Orbitrap data. The whole ISB1 dataset contained twenty-two LC-MS/MS runs (two sets of technical replicates), resulting in 18,999 spectra. The Orbitrap data set contained ten LC-MS/MS runs and four LC-MS/MS runs were included from the FT Ultra dataset. In total there were 47,292 spectra for the Orbitrap data and 54,551 for the FT Ultra data. Search engine runs were performed against a concatenated protein database containing forward and reversed sequences of the 18 proteins, contaminants and a whole organism proteome database from the bacterium *Sorangium cellulosum* (Schneiker et al., 2007). The contaminant proteins were trace-level contaminants, as listed by Klimek et al. (2008) and additional keratin and trypsin sequences. All together the protein database contained 18,812 sequences. Peptide sequences were considered correctly identified if the sequence was found as a subsequence in one of the 18 proteins or a known contaminant. Trypsin was set as protease for all the search engines. All identifications were conducted using the respective search engine adapters available in TOPP (Kohlbacher et al., 2007).

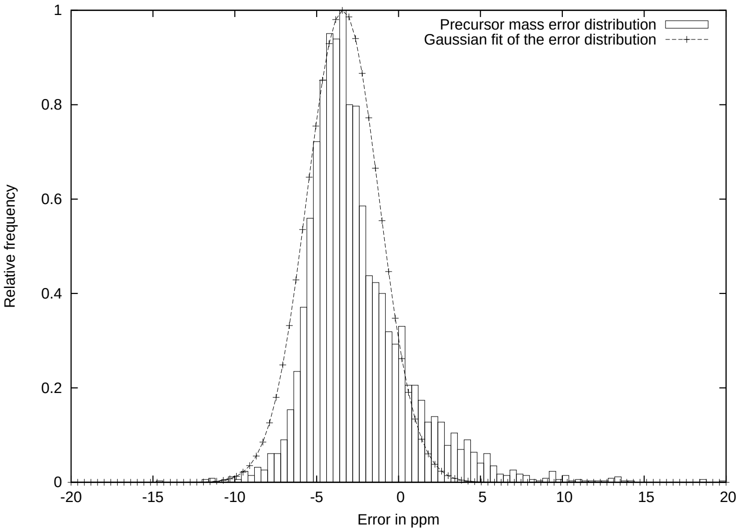
4.3 Results

In Fig. 4.1 the number of true peptides with their corresponding ranks are visualized for each individual search engine. It can be observed that if the three search engines agree on a peptide identification (all three engines suggest the same sequence on rank one), approximately 94 % of those peptides are correct sequences. However, this agreement only corresponds to approximately 8 % of all annotated spectra in the data from the low-resolution instruments and to approximately 14

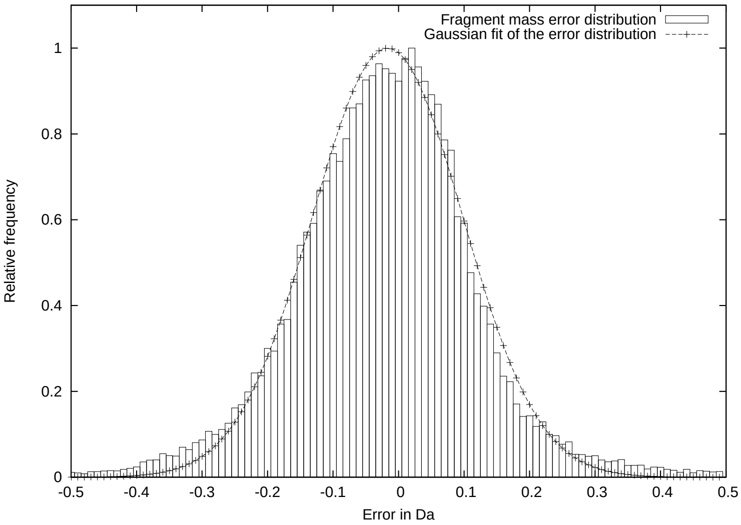
% in the data from the FT instruments. Every dataset was searched with broad initial search tolerances to assess the data quality. Fig. 4.3 shows the distributions of mass errors resulting from the initial search runs. It can be seen that error in the precursor masses were recorded with a slight shift to the left. The majority of peptide masses was measured with -4 ppm deviation. The fragment masses are distributed evenly around zero. In Fig. 4.4 the overlap of spectra that were annotated by one, two, or all three engines is shown. Here, annotated means that the search engine suggests any sequence for a given spectrum. A high number of spectra are not annotated by Mascot. Interestingly the percentage of spectra that are annotated by all engines decreases for high-accuracy FT instruments. It can also be observed that the number of spectra that are only annotated by one engine is high for OMSSA, but most peptides annotated by X!Tandem and Mascot are also annotated by other engines. Accordingly, Tbl. 4.2 shows the number of spectra that were annotated by the search engines, regardless of the significance of the scores that were assigned to the sequences. The numbers of peptides that were given a score that is high enough to accept the peptide as correctly identified within a q-value threshold of 0.01 are listed in Tbl. 4.3. For the high-accuracy datasets, the number of correctly identified peptides by X!Tandem were clearly larger than the numbers from the OMSSA and Mascot searches. For the low-accuracy LCQ instrument OMSSA performed best. On the dataset from the complex mixture X!Tandem did not perform as good as on the less complex 18 protein mix dataset.

Table 4.2: Number of spectra from the different data sets that were annotated by the different search engines. All data is based on the 18 protein mix, except the last column.

	Orbitrap	FT Ultra	LCQ	Orbitrap (*E. coli*)
Mascot	14,050	19,102	17,336	17,254
OMSSA	36,896	35,841	18,943	27,575
X!Tandem	23,035	32,024	18,011	27,218



(a)



(b)

Figure 4.3: The relative mass error of precursor masses and the absolute mass errors for fragment masses. Data based on the 18 protein mix acquired on an LTQ-Orbittap and search by OMSSA.

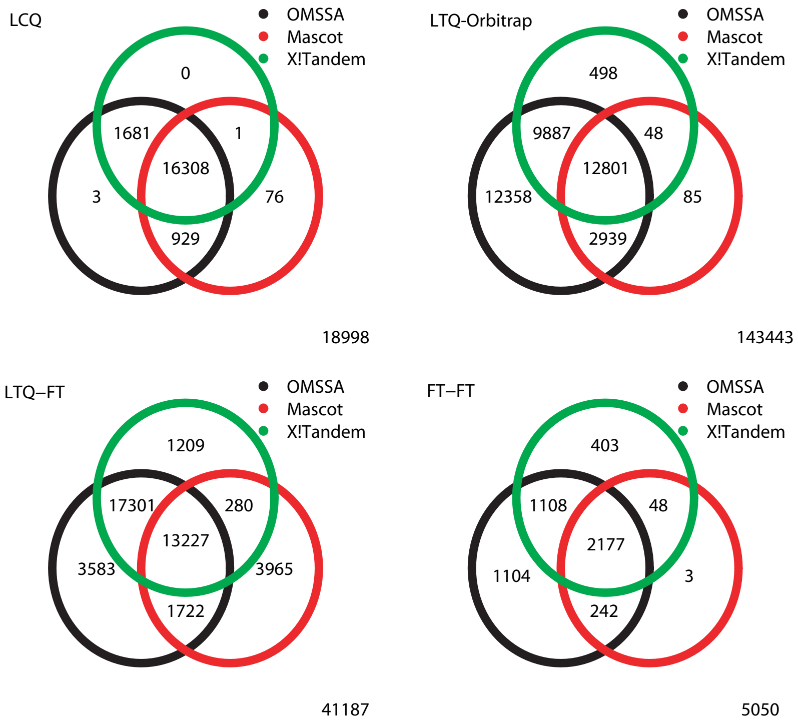


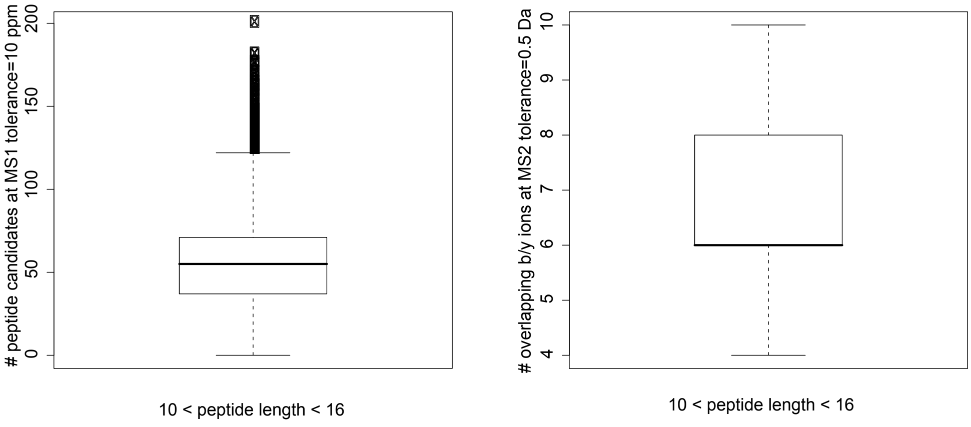
Figure 4.4: The venn diagrams show the number of spectra that were annotated by the different search engines. Annotated means that the search engine suggest a candidate sequence for the given spectrum, regardless of significance of the assignment.

Table 4.3: From the number of annotated spectra only fraction reaches the score threshold to be considered identified. All data is based on the 18 protein mix, except the last column.

	Orbitrap	FT Ultra	LCQ	Orbitrap (*E. coli*)
Mascot	5,495	3,697	1,512	10,888
OMSSA	5,405	3,809	1,955	10,611
X!Tandem	5,744	5,099	1,990	9,611

Sequence similarity

In those cases where search engines disagree on a spectrum, we frequently observed that there was significant sequence similarity. The similarity is expressed in partly overlapping sequences, which is reflected in overlapping fragment ion masses. This observation can be observed in Fig. 4.5. For both boxplots the data are averaged for peptides with length from eleven to fifteen amino acids. On average there are about 50 sequences in the search space, while searching against a human target-decoy database with a tolerance of 10 ppm. Peptide sequences in these search spaces can have a high number of overlapping fragment ion masses, as visualized in 4.5(b). Using the 18 protein mix measured on an LTQ-Orbitrap



(a) Peptide length vs. search space

(b) Overlapping fragment ions

Figure 4.5: Search space characteristics of short peptides of ten to fifteen amino acids.

Fig. 4.6 shows the similarity of top hit candidates to peptide sequences at lower ranks. It was searched with OMSSA using parameters as described in the experimental procedures. It shows the percentage of sequence similarity of peptide candidates that are suggested by OMSSA at different ranks. The sequence similarity correlates very well with the percentage of overlapping fragment ion masses. At higher ranks, the sequence similarity and the percentage of overlapping fragment ions decreases. The percentage of identical amino acids does not decrease at ranks latter than four for the OMSSA search engine results.

Consensus scoring

We evaluated different methods for the combination of the scores. The results for the comparison of different methods are shown in Fig. 4.7. Different scoring matrices were used to calculate the consensus scores. In general we observed that PAM matrices perform better compared to the Blosum matrices. We evaluated different PAM and Blosum matrices. The results in Fig. 4.7 show a selection of the best scoring matrices. In general, high penalization is more suitable for these data. However on low accuracy datasets, lower penalization shows better performance (data not shown). The spectral counter scoring (SPC) performs almost equally well as the best matrix based methods. The performance of the averaging methods is below the matrix based methods and comes even below the best single engine. However all combination methods, except for the *naive* averaging outperform the single engines.

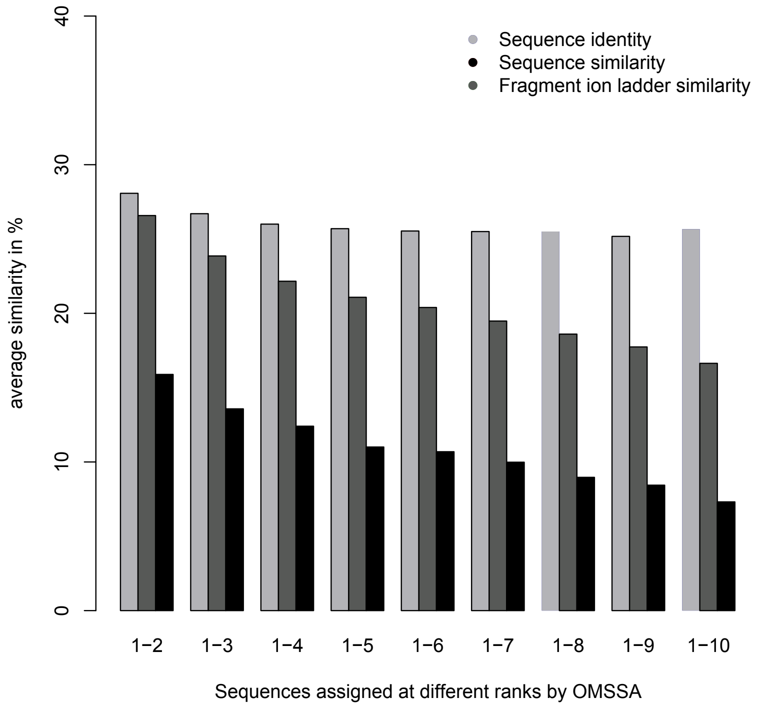


Figure 4.6: Different measures for peptide similarity.

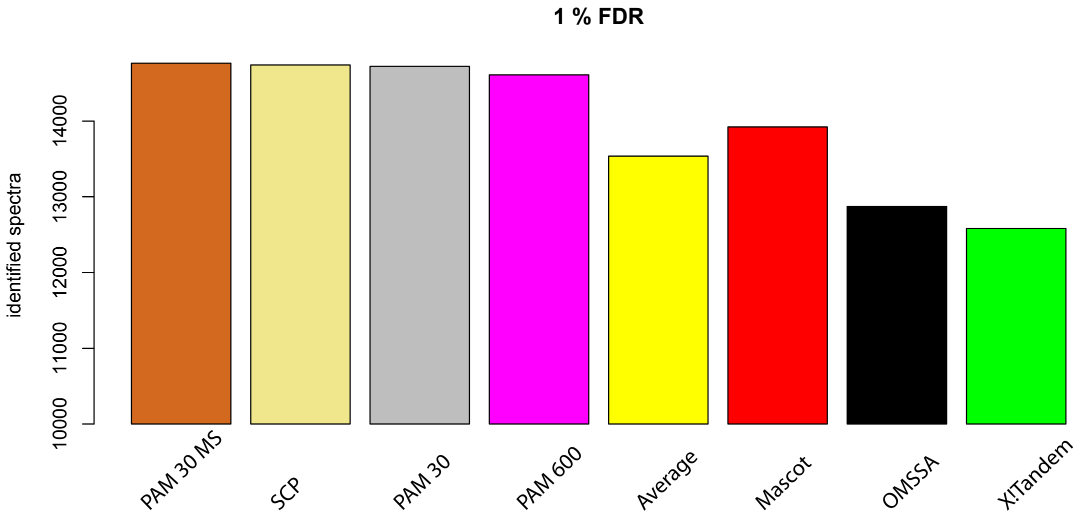


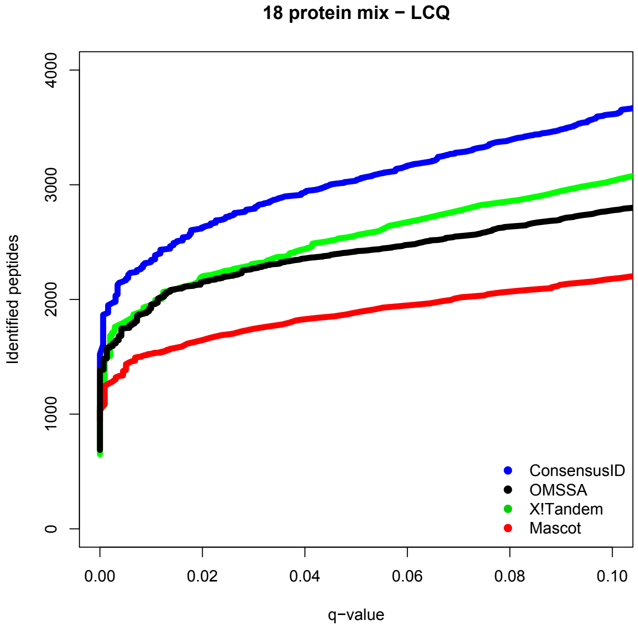
Figure 4.7: Different methods to combine single search engine results to a common consensus score. This evaluation was done on the *E.coli* dataset.

Receiver operating characteristics (ROC)

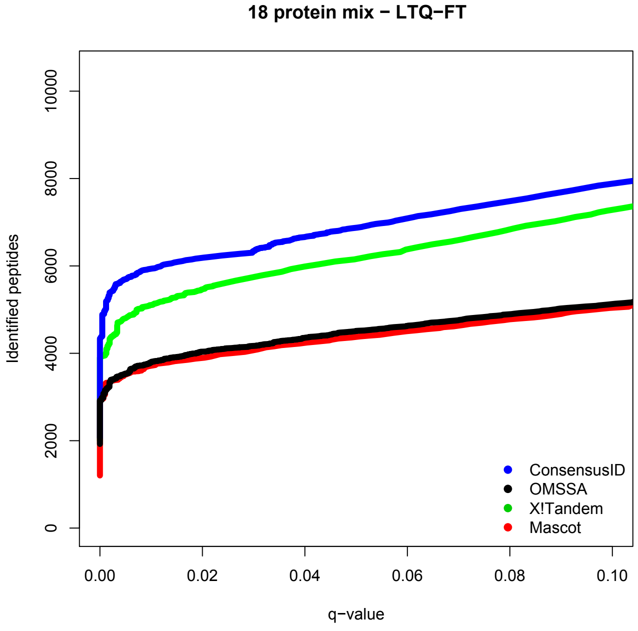
ROC analysis was performed to compare the results of the single engines and to visualize the benefit of *consensus scoring*. ROC curves help to visualize the performance of classifiers (Fawcett, 2006). In most cases, the true positive rate is plotted as a function of the false positive rate. A good classification is obtained if the slope of this function is very steep for low false positive rates and if the area under the curve comes close to one. For the purposes of this study the analysis was adapted and the number of correctly identified spectra was plotted as a function of the corresponding q-values. The results of this analysis are shown in Fig. 4.8-Fig. 4.9. The *consensus scoring* method was significantly better than the single engines for all datasets. We observed significant improvements at low error rates for all datasets. For the Orbitrap data, the consensus scoring identified 19 % more peptides compared to X!Tandem, 26 % more than OMSSA and 24 % more than Mascot at 1 % FDR. For the FT Ultra dataset the improvements were even more significant. X!Tandem was outperformed by 18%, OMSSA by 57 % and Mascot by 63 %. On the LCQ data at 1 % FDR X!Tandem was outperformed by 17 %, Mascot by 54 % and OMSSA by 19 %. The *E. coli* was the most complex among all datasets. This dataset was searched against the large SwissProt database that contains a considerable number of homologous sequences. For these data, the most significant improvement was in comparison to X!Tandem at a value of 27 %, Mascot at 13 % and OMSSA at 16 %. If more false positives are allowed the improvements gained from the consensus scoring are more significant for all datasets, compared to the 1% error rate.

Mass accuracy is known to be a crucial parameter for peptide identification. Ion traps, such as the LTQ instrument, are low mass accuracy instruments, whereas TOF instruments usually provide increased accuracy. FT instruments, such as the Orbitrap, belong to the most accurate mass spectrometers. The design of our scoring method includes the flexibility to adapt the scoring method to the most commonly used instrument types. Different values for the two gap penalties were evaluated. It was found that the performance tends to improve with the stringency of penalization. Mixture modeling allows the conversion of arbitrary search engine scores into probabilities and peptide similarity helps to assign scores to peptide sequences that were originally not assigned by a given search engine. For the combination of several single scores into a joint consensus score, various approaches can be used.

The suggested methods combines the scores from the single engines by a weighted average and that is further divided by the summed similarity (equation (4.4)). Different methods are evaluated. All other approach were found to reveal results inferior in terms of peptide identifications. Thus the similarity weighted average-like score is the default method for combining single scores. The average is proportional to the sum of the scores and can thus loosely be interpreted as the accumulation of evidence obtained from the different single sources.

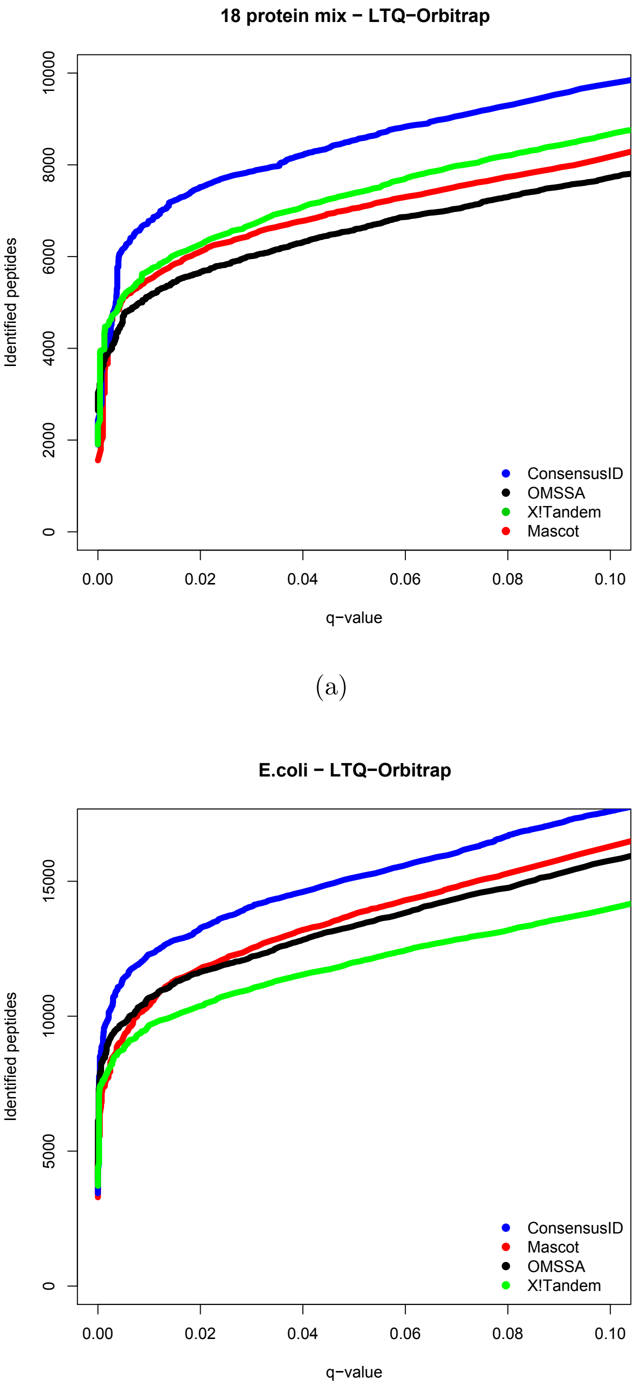


(a)



(b)

Figure 4.8: Receiver operating characteristic (ROC) curves to visualize the performance of the searches using LCQ and LTQ-FT data. Increased area under the curve indicates indicates better performance.



(a)

(b)

Figure 4.9: ROC curves to visualize the performance of the search methods using LTQ-Orbitrap data from the 18 protein mix and the complex *E. coli* sample.

Our consensus scoring method was better than any single engine. A summary of the results at 1 % FDR on all evaluated datasets is shown in Fig. 4.10.

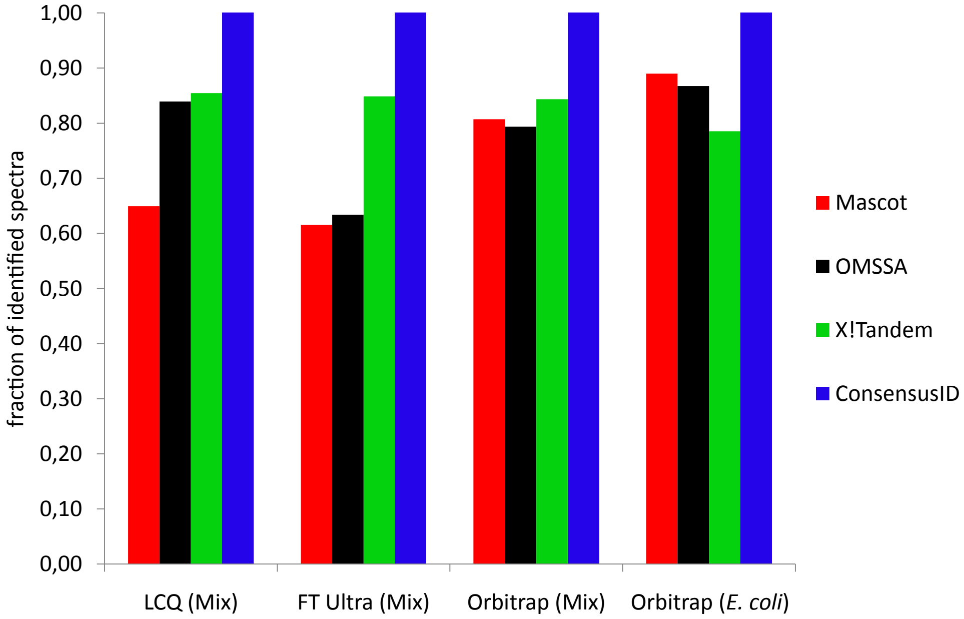


Figure 4.10: This plot summarizes the comparison of the different search methods, along with the improvements that are gained by combining the search methods. The y-axis corresponds to the number of identified spectra. The number of spectra that were identified by the best method was set to 100 %. The x-axis shows the different search datasets; Mix corresponds to the 18 protein mix datasets. Regardless of the data quality and the data complexity the ConsensusID approach always outperforms the single engines.

4.4 Discussion

Proteomics has traditionally been a dynamic area with a broad spectrum of experimental techniques and rapidly evolving instrumentation. This is accompanied by an accumulation of computational tools as an indispensable part in the analysis workflow. Our *consensus scoring* approach aims to take advantage of the large diversity of commonly used peptide identification tools. Our approach is designed to incorporate any number of different tandem MS search engines into the *consensus scoring*. For each tool, the scores are converted into probabilities in order to make them directly comparable. Missing values are estimated by our sequence similarity approach and finally the scores combined by weighted averaging.

Datasets, generated by a variety of MS instruments, were included in this study. The instrumentation in laboratories is changing rapidly and new MS instruments are continuously entering the market. More and more laboratories are equipped with several mass spectrometers to enable high-throughput and to benefit from special features available on specific instruments. This implies that software for peptide identification needs to cope with this rapid evolution. The method proposed in this work is very robust with respect to the origin of the data. Independent of which individual search engine performs best on any given dataset, the consensus approach always yields a superior performance in our tests. Previous studies aiming at combining search engine scores only focused on low mass accuracy ion traps (Searle et al., 2008). The importance of more accurate and more sensitive instruments is obvious. Furthermore, the suggested strategy allows easy interpretation of the consensus results and does not rely on complex machine learning methods that are hard to interpret. To our knowledge our *consensus scoring* is the first approach that offers significant improvements in peptide identification on data generated by a variety of different mass spectrometry platforms. By integrating the information gained by the other engines, the assigned peptide probabilities become more accurate and in cases where specific peptides were ranked best with poor scores, the information from the other engines helps to improve this score and ultimately to bring it in a range where it is accepted as correctly identified. Assuming a given mass spectrum has been recorded without any major technical bias, then the likelihood that several search engines assign the same *wrong* peptide to rank one is smaller than the likelihood that a subset of engines fails for this spectrum. This putative failure is corrected by integrating information from other engines via peptide similarity scoring. The similarity-corrected probabilities are then combined by a weighted average scoring. Another alternative for the combination of these scores would be the product of the individual probabilities. Using the product as a consensus score, one would have to assume independent scoring of the individual search engines. However, it is common sense that the ultimate prerequisite for database search algorithms is the presence of fragmentation-specific product ions. Assuming independence is thus not realistic. The average, in contrast to the product, accounts for this underlying property. If X!Tandem's refinement is used, the improvements of peptide identification rates by X!Tandem increase the number of identified peptides by Mascot and OMSSA. However due to unsolved question regarding the statistical assessment of peptide identifications that result from multi-pass searches, such as X!Tandem's refinement, we did not enable this option for routine analyses. If the precursor masses are recorded with high accuracy, significant improvements for peptide identifications are observed. The imputation procedure for peptide sequences is based on a pairwise global alignment and uses the Needlemann-Wunsch algorithm. This algorithm uses substitution matrices to score the peptide alignment. Those matrices, such as PAM matrices, are used in evolutionary biology to determine similarity of proteins, based on mutation probabilities, thus substitution matrices are *per se* not constructed to account for spectral similarity. However, we demonstrated that the sequence similarity of the different peptide candidates strongly correlates with the similarity in the ion series of the different sequences and we also showed that the usage of peptide similarity has slightly better performance than the similarity based on fragment ion similarity. This might be explained by less stringent penalization of the sequence similarity method, if one engine failed. The Needlemann-Wunsch algorithm performs a global alignment of sequences and will not adequately account for peptides that show strong local similarity. Local alignment algorithms might be useful alternatives, if the spectra are searched against very large databases, where much more similar sequences can be expected. Such strategies might be further research topics especially in the field of proteogenomics where spectra are searched against six-frame translations of the genome. If such searches are combined with the target-decoy database concept, the number of candidates in the tolerance window is very high and a lot of similar sequences will compete for spectra.

Global alignments are, in general, used to align protein or nucleotide sequences if they are expected to be similar and have roughly the same length. The lengths of the peptide candidates should not differ greatly, since precursor masses are recorded and used to restrict the search space of candidate hits. Nevertheless, global sequence alignment methods are classically implemented using two different parameters δ and ε to account for gaps that need to be introduced in the alignment. The evaluation of different values for these penalty parameters revealed that high mass accuracy instruments are not very sensitive, whereas the data that resulted from the LCQ instrument revealed better identification rates with decreased stringency in penalization. The quality of MS spectra strongly correlates with the presence of all expected fragment masses. Missing masses or imprecise precursor masses are reasons that contribute to incorrect assignments of peptide sequences by search engines and are thus more frequently observed when low mass accuracy instruments are used. With growing accuracy and sensitivity in MS instrumentation, those inadequacies are diminishing, however, they are still present to some extent. The alignment parameters are very suitable values to adjust the scoring scheme to different instruments. In lists of putative candidate peptides, there are hardly any peptides that have very large gaps, however, peptide sequences that miss amino acid masses can be found. Further development in search strategies for tandem MS spectra will also have to account for single nucleotide polymorphisms (SNPs), since SNPs are frequently observed in genomics studies. Our ConsensusID approach is very well suited to cope with search results that allow amino acid exchanges. Other improvements may include the adaption of the scoring matrix to account for isobaric di- or tripeptides. Using our suggested *consensus scoring* method, the time needed to compute the peptide identifications will be increased compared to the usage of a single engine. If three engines are used, the time for search engine calculations is three times as high and additionally the consensus scoring algorithm needs time to calculate the consensus scores, which is comparably fast; 1-2 min for moderate size datasets (0.5 - 1 GB). Using three search engines and the *consensus scoring* combination method, the peptide identification pipeline is still at least five to ten times faster as the acquisition of the tandem mass spectra and the gain in peptide identification rates is clearly in favor of the *consensus scoring*.

The proteomics community strongly relies on database search engines. Peptide identification is both, the most fundamental and the most important step in an MS-based proteomics study. In order to fulfill the high demands attributed to proteomics, combining different strengths and reducing weaknesses of individual peptide identification approaches is needed.

Chapter 5

Quantitative shotgun proteomics to analyze protein expression dynamics

5.1 Introduction

The introduction of shotgun sequencing of genomes was a milestone of modern biology. The human genome was published in two competitive papers. The human genome consortium, funded by public money used conventional methods (Lander et al., 2001), whereas Craig Venter pioneered shotgun sequencing (Venter et al., 2001). Similar to shotgun sequencing of genomes, shotgun proteomics refers to the sequencing of the entire protein complement of the genome. As in DNA sequencing, shotgun sequencing aims at assembling small parts of proteins, peptides, together to ideally build up the whole proteome. In contrast to the genome, the proteome is highly dynamic. The abundance levels of proteins change very rapidly in a biological system and the analysis of the proteome has to cope with these quantitative changes over time. In this section we will refer to shotgun proteomics, as a method to quantitatively profile the whole or large parts of the expressed proteome, using gel-free mass spectrometric techniques.

The analysis of complex biological mechanisms, such as the effects of the inhibition of signaling in cancer cells, is analytically very demanding, since only little is known about the mechanisms of these inhibitors, yet they are considered promising agents in the fight against cancer. In this section we investigate the effects of sorafenib and LY294002, two well known kinase inhibitors in human melanoma cells, using shotgun proteomics methods. In contrast to 2D-PAGE experiments, as outlined in Chapter 3, shotgun proteomics can potentially offer a much greater proteome coverage and more accurate quantitation. We chose to use *in vitro* growing cancer cells as model organisms for this study. The great advantage of cell lines is the compatibility with the SILAC labeling method. Using SILAC we can quantify the peptide ions directly after the mass spectrometric analysis. In contrast to label-free analyses, this avoids the challenging direct-comparison of multiple runs.

The adjacent section introduces all experimental and theoretical methods that were used throughout this study. These techniques cover mainly sample preparation methods, sample fractionation to reduce the sample complexity and finally LC-MS for peptide identification and quantitation. Following the experimental data acquisition, the data is processed and analyzed by statistical methods. The computational analysis allows the formulation of hypotheses that might contribute to fill gaps in the understanding of the mechanisms of action of the inhibitors. Our analysis allows to identify groups of proteins, so-called clusters that are grouped together due to their similar protein expression profile as response to inhibitor treatment. If these proteins are known to be involved in similar biological processes or pathways, such clustering can ultimately hint to the biological activities that are predominately affected by the inhibitor treatment. The time course in addition, allows to estimate the time span needed to initiate, up- or down-regulate the respective biological activity. This Chapter will present and discuss the results from the proteome-wide profiling experiments, the subsequent cluster analysis and the enrichment analysis for biological processes and pathways.

5.2 Global protein expression dynamics

Similar to the experiments that were described in Chapter 3 for the DIGE analyses, this section describes an application of quantitative shotgun proteomics.

Using SILAC labeling, the global protein expression in melanoma cells was profiled following treatment with two distinct multiple kinase inhibitors, sorafenib and LY294002. The SILAC setup allows recording of time course data, since different SILAC labels are mixed in the same experiment.

5.2.1 Material and Methods

Overview experimental setup

The experimental setup of the SILAC-based time course experiment is shown in Fig. 5.1. For each biological experiment two LC-MS analyses were performed. This setup allows profiling of five time points in parallel. In both triple SILAC experiments, one SILAC label is used as a common time point. The common time point is later used to normalize variations in intensities across the two runs.

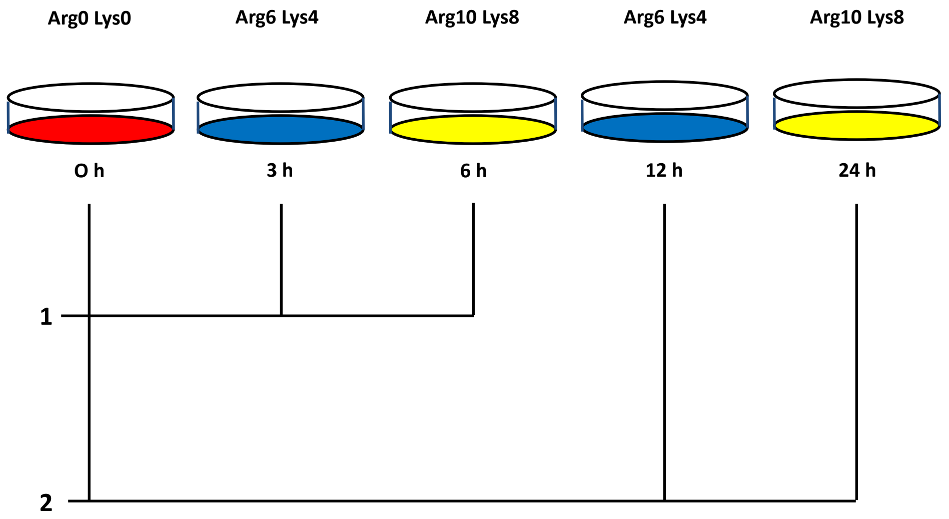


Figure 5.1: Five time points are profiled in parallel using triple SILAC LC-MS analysis.

Cell culture

Metastatically growing 451Lu melanoma cells were kindly provided by Prof. Birgit Schittek (Department for Dermatology, Tübingen). The 451Lu cell line was originally derived from lung metastases of WM164 cells that were subcutaneously injected into nude mice. The WM164 cell line is derived from a metastatic lymph node melanoma. The 451Lu cells are highly invasive and exhibit spontaneous metastasis to lung and liver (Herlyn et al., 1990). For all cell cultures, performed in this thesis, the general conditions were 37 °C and 5 % CO_2 at humidified atmosphere. The cells were cultured in RPMI medium, purchased form Invitrogen. The medium was supplemented with 10 % fetal bovine serum (Invitrogen), 1 % (10 mg/ml) streptomycin / (10,000 U/ml) penicillin (Invitrogen) and 1 % L-glutamine. Cells were usually split at 80 % confluency. The duration cycle for cell replication is approximately 24 h.

For the SILAC experiments a custom-made medium was used. The commercial custom-made medium was depleted for Arg and Lys. The SILAC medium was then supplemented with three different isotopic variants of Arg and Lys. The 'heavy' medium was supplemented with Arg10 ($\mathbf{^{13}C_6^{15}N_4}$) and Lys8 ($\mathbf{^{13}C_6^{15}N_2}$), 'medium heavy' medium with Arg6 ($\mathbf{^{13}C_6}$) and Lys4 ($\mathbf{^{2}H_4}$), the 'light' medium contained Arg0 ($\mathbf{^{12}C_6^{14}N_4}$) and Lys0 ($\mathbf{^{12}C_6^{14}N_2}$) . The final concentration of Lys in all media was set to 73 mg/l and the Arg concentration was set to 42 mg/l. The cells were grown for at least two weeks in SILAC media. This is necessary to reach sufficient incorporation rates.

Inhibitor treatment

Sorafenib was purchased as a powder from LC Laboratories (Woburn, USA). LY294002 was purchased from Cell Signaling Technology (Danvers, USA). Both inhibitors were dissolved in DMSO (dimethylsulfoxide) (Sigma, Germany) and 13 mM sorafenib and 50 mM LY294002 stock solutions were prepared. For each experiment a freshly thawed aliquot was used. According to the experimental setup cells were treated with 13 μM sorafenib, 50 μM LY294002 or the corresponding amount of DMSO.

Cell lysis

Before harvesting, the cells were washed twice with 10 ml of ice-cold PBS (PAA, Austria). The cells were lysed using 6 M urea (Sigma, Germany), 2 M thiourea

(Sigma, Germany) in 10 mM Tris at pH 8(prepared from Tris, base, Sigma, Germany and Tris, HCl,Merck, Germany). The Complete, EDTA-free protease inhibitor cocktail (Roche, Germany) was used to avoid unspecific proteolytic cleavage. This inhibitor cocktail inhibits serine and cysteine proteases, but not metalloproteases. 500 μl of lysis buffer was added to a 10 cm dish. The lysed cells were transfered into 1.5 ml reaction tube, followed by a incubation on ice for 30 min while briefly vortexing every 10 min. 1 μl benzonase (Merck, Germany) was added to every tube to destroy remaining DNA or RNA molecules. Finally the cell debris was pelleted by centrifuging at 13,000 g for 20 min at 4 °C. Protein concentration in the supernatant was determined by the Bradford method. Absorbance was measured at 590 nm and equal amounts of protein from each SILAC condition were mixed accordingly.

Proteolytic digestion of proteins

The proteins were reduced by incubating with 1 mM dithiothreitol (DTT) at room temperature (RT) for 30 min, followed by an alkylation step with 5.5 mM iodoacetamide for 30 min in the dark. The reduced and alkylated proteins were digested for 4 h with the endoproteinase Lys-C (Wako, Japan) dissolved in 20 mM ammonium bicarbonate at a concentration of 1/100 (w/w). After diluting 4 times with 20 mM ammonium bicarbonate and adjusting the pH to 8.0, sequencing-grade modified trypsin (Promega, Germany) dissolved in 20 mM ammonium bicarbonate, was added at a concentration of 1/100 (w/w). After an overnight incubation with trypsin at 37 °C the reaction was stopped by adding trifluoroacetic acid (TFA) to a final concentration of 0.1 %.

Isoelectric focusing of peptides

For the isoelectric focusing of peptides an *Agilent 3100 OFFGEL Fractionator* was used. Essentially, the experiment was performed, as described by Hubner et al. (2008). After trypsin digestion the peptides were separated based on their isoelectric points into twelve fractions. The starting material for the OFFGel separation was set to 120 μg (40 μg from each SILAC condition). The peptide sample volume was adjusted with water to 324 μl and 1.44 ml 6 % glycerol and 36 μl ampholytes (*IPG-buffer pH 3-10*, purchased from GE Healthcare, Germany) were added. The samples were loaded onto *12 well-Immobiline DryStrips pH 3-10*, purchased from GE Healthcare. Peptides were then focused for 20 kVh at a maximum current of 50 μA and maximum power of 200 mW. Each peptide fraction was mixed with 10 μl acidic solution containing 30 % ACN (Merck, Germany), 5 % acetic acid and 10 % TFA, before loading onto stage tips for desalting and storage.

Stage tipping of peptides

The procedure as published by Rappsilber et al. (2003) was followed for stage tipping of peptides resulting from OFFGel fractionation. In brief, stage tips were prepared using small discs of C18 material (Varian, Germany). The C18 material was put into a 200 μl pipette tip, using a plastic syringe prepared in-house. The columns were activated with methanol and equilibrated in 0.5 % acetic acid, 0.1 % TFA in water. Peptide samples were loaded onto the stage tips and washed once with 0.5 % acetic acid, 0.1 % TFA in water. Directly before the LC-MS measurement, peptides were eluted from the stage tips with 50 μl of 80 % ACN, 0.5 % acetic acid. The organic solvent was then evaporated by reducing in a SpeedVac (Eppendorf, Germany) to approximately 5 μl. The final volume was then adjusted to 10 μl with 0.5 % acetic acid, 0.1 % TFA and 2 % ACN in water and 5 μl were used for one LC-MS analysis.

Liquid chromatography

After the OFFGel separation, the whole-cell lysate was split into twelve fractions. For each OFFGel fraction a volume of 5μl, which approximately corresponds to 5 μg of peptides, was consecutively injected by a PAL autosampler (coupled to an Eksigent nanoLC) or by the integrated autosampler from the Proxeon EASYLC system. Samples were then injected to the a 15-cm-long 75-mm-inside-diameter column (New Objective, Woburn, USA) packed in-house with 3-mm C18 Reprosil reversed-phase beads (Dr. Maisch) with a flow rate of 500 nl/min for the first 20 min. The peptides were eluted with a linear gradient from 8 to 64 % acetonitrile in 2 h at a flowrate of 200 nl/min. Solvent A was water with 0.5 % acidic acid and solvent B was 80 % acetonitril in water with 0.5 % acidic acid. The typical gradient employed for peptide separation is shown Fig. 5.2. The increase to 80 % solvent B is used to wash the column.

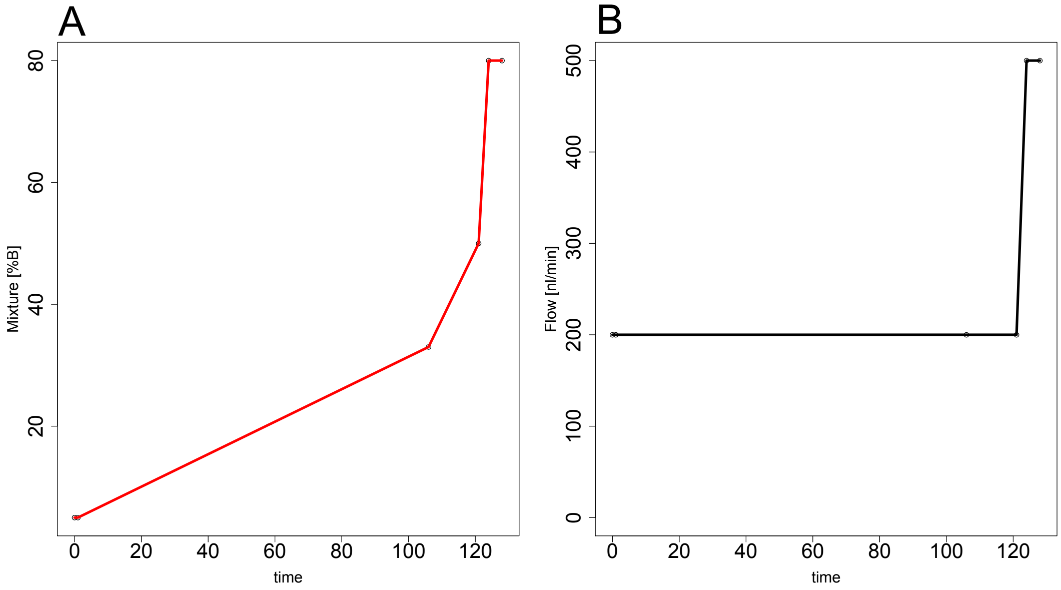


Figure 5.2: The gradient (A) and flow rates (B) used for the analysis of OFFGel fractions.

Mass spectrometry

For all analyses an LTQ-Orbitrap mass spectrometer was used, which was operated in positive ionization mode with an average ion spray voltage of 2.3 kV. Data-dependent acquisition (DDA) mode was used to sample the five most abundant ions per full scan. Full scans were recorded in the orbitrap analyzer and the fragment ion scans in the LTQ analyzer, respectively. The mass accuracy of precursor ions was routinely improved using the lock mass option. Only multiply charged ions were fragmented. 10^6 ions had to be collected for full scan analysis in the Orbitrap analyzer at a resolution 60,000 (at m/z 400). The time needed for one full scan, a lock mass cycle and five subsequent MS/MS spectra was typically below four s (Olsen et al., 2005). The activation type for all analyses was collision-induced dissociation (CID) with a normalized collision energy of 35 %, the isolation width for the precursor selection was set to two Thomson (Th). The range of masses was limited 300-2,000 m/z. Collision-induced dissociation fragmentation is induced after an accumulation of 5,000 ions. To avoid repeated fragmentation of the same ions, a dynamic exclusion list was filled with up to 500 masses that have already been fragmented. Each mass remained on the exclusion for 90 s.

Data processing

All raw MS spectra were combined and processed together using the MaxQuant software suite (version 1.0.13.9) (Cox and Mann, 2008). The Mascot (Matrix Science) search engine, version 2.2, was used to search the peak lists. The database was a concatenated IPI human forward and reversed protein database version 3.64. The database contained 84,031 forward protein sequences and 263 contaminant protein sequences, resulting in a total of 168,588 entries. Carbamidomethylation and either Lys0 and Arg0, Lys4 and Arg6 or Lys8 and Arg10 were set as fixed modifications. Methionine oxidation was set as a variable modification. Trypsin was defined as the protease and two missed cleavages were allowed. The precursor mass tolerance was set to 7 ppm and 0.5 Da was the allowed deviation at the fragment ion level. The identify module of MaxQuant was used to interpret the search results. 1 % FDR threshold was used as a acceptance threshold on the peptide and protein level. At least two quantitation events were required for protein quantitation.

5.2.2 Results

This section lists the results gained from the SILAC-based proteomics experiments. We shall start with a global overview of the results, allowing to assess the overall technical performance and to judge the data quality. This is followed by a thorough description of the results gained from clustering and enrichment analyses.

Chromatographic separation

The chromatographic separation was monitored for each LC-MS run. A typical total ion count (TIC) chromatogram is shown in Fig. 5.3. The peptides are usually widely distributed along the gradient with increasing ACN concentration. At the very end of the gradient (125 - 130 min) a peak with high intensity is observed. This peak corresponds to components originating from sample preparation procedures. However, this peak does not interfere with the peptide signals, most peptides elute at much lower ACN concentrations. Different chromatograms from the twelve OFFGel fractions are shown in Fig. 5.4. In different OFFGel fractions different peptide species are expected. The total ion count (TIC) in the different OFFGel fractions is very similar across all twelve runs. Only fraction six shows a TIC below 10^9 and fraction twelve is empty. To monitor the fractionation efficiency, all peptides that were identified in the different OFFGel fractions were compared for their occurrence in different fractions. We observed that the majority of peptides is only observed in one OFFGel fraction.

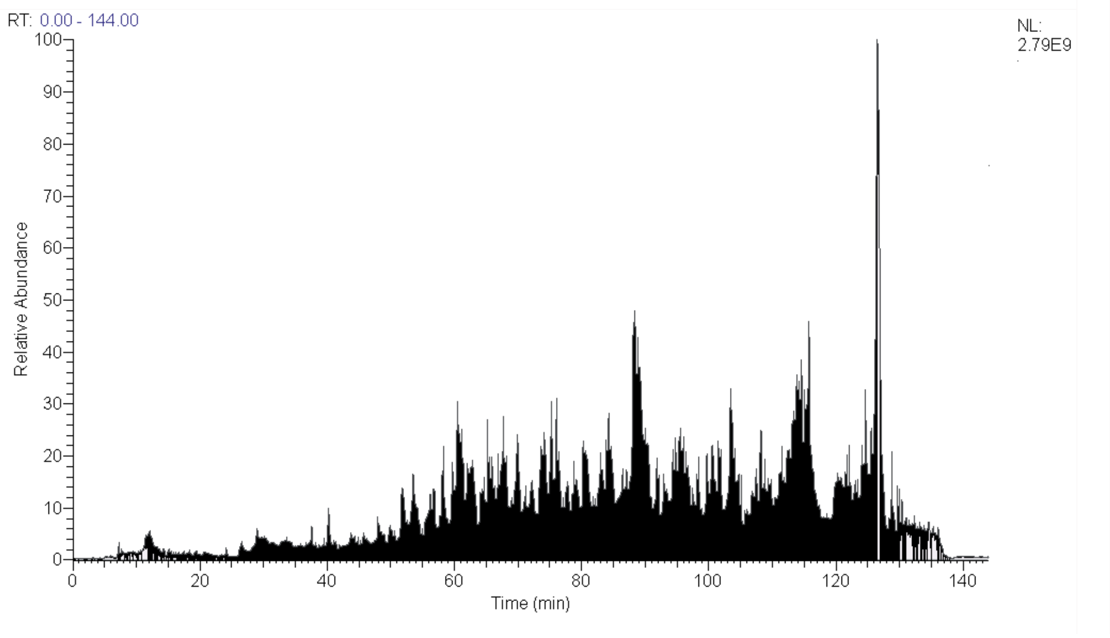


Figure 5.3: A typical TIC chromatogram for a 2 h LC gradient. All peak intensities are normalized to the highest peak, which corresponds to 100 %.

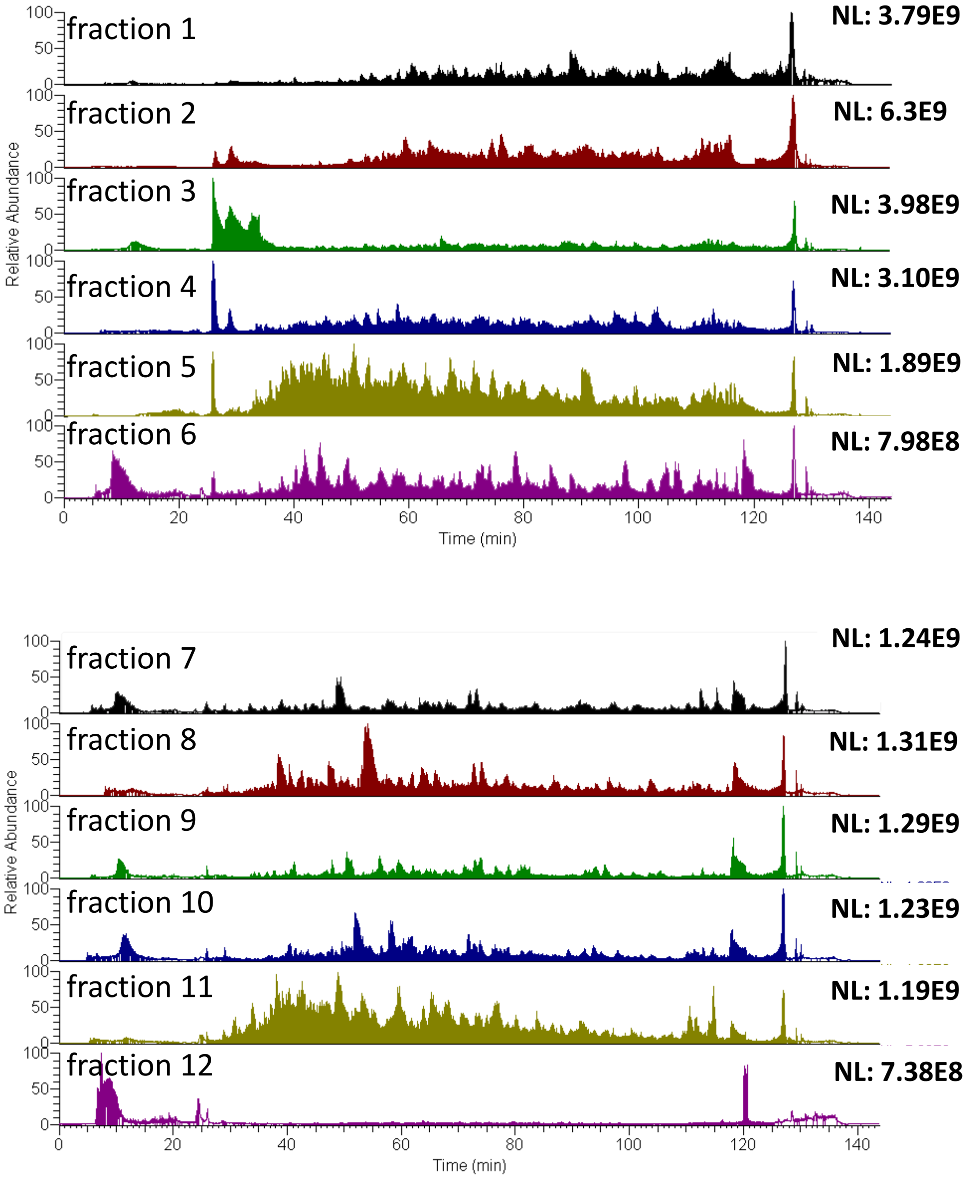


Figure 5.4: The different colors correspond to different OFFGel fractions. The NL value on the right indicates the TIC. The Y-axis are intensities normalized to 100 %.

Global assessment of data quality

All raw data from the experiments for both inhibitors including replicates were processed together. This final dataset included 240 LC-MS runs. On average this dataset was measured with a calibrated average mass error of 0.34 ppm. Tbl. 5.1 shows global results on the number of recorded spectra, identified spectra, non-redundant peptides and protein groups. The number of proteins that were found at all different time points for the sorafenib experiment is visualized in the venn diagram in Fig. 5.5. 3,465 from a total of 5,408 proteins were observed in all time points. The set of 3,465 proteins was used for the analysis of protein expression dynamics. For the global analysis of regulation the whole dataset was used. The significantly regulated proteins were determined as outlined in the appendix. Proteins were accepted as significantly regulated, if the intensity dependent significance p-value (significance B from Cox and Mann (2008)) was below 0.05. Proteins with significantly modulated expression following DMSO treatment were required to have regulation ratios at least twice as high following inhibitor treatment compared to the DMSO ratio. A typical time course profile is shown in Fig. 5.6. The five time point time course is generated from two different experiments. The number of significantly regulated proteins at different time points are shown in Fig. 5.7. Most notably the number of down-regulated proteins at the late time points is increased compared to the number of up-regulated proteins.

Table 5.1: The complete dataset set has been investigated for global result parameters.

spectra recorded	2,254,115
spectra identified	759,286
non redundant peptides	45,086
protein groups	5,408

In some cases, the protein expression was regulated selectively at only some time

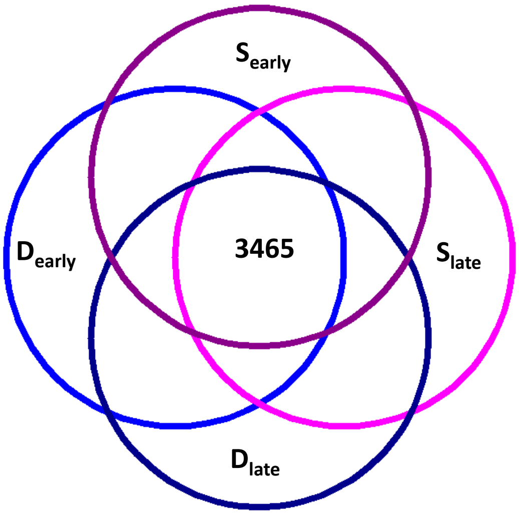


Figure 5.5: Overlap between different time points. S_{early} refers to the SILAC experiment with the early time points 3 h and 6 h of harvesting after Sorafenib treatment. S_{late} refers to the harvesting after 12 h and 24 h, respectively. The D corresponds to the control experiment with DMSO treatment.

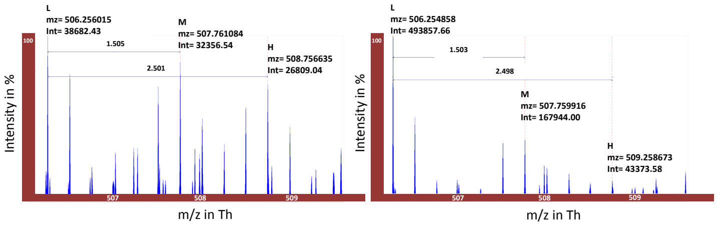


Figure 5.6: Triple SILAC based temporal profile. These spectra belong to QDFSVPQLPHSSSHWLR, a unique peptide from GP100, a known melanoma antigen.

points and in other cases the inhibitor treatment regulated the protein expression for the whole time period. The number of significantly regulated proteins at the different time points was analyzed in Fig. 5.8 for the sorafenib treatment and in Fig. 5.9 for the LY29402 treatment, respectively. Most of the proteins are significantly regulated at only one time point, however some proteins are regulated in at least two, three or even four different time points.

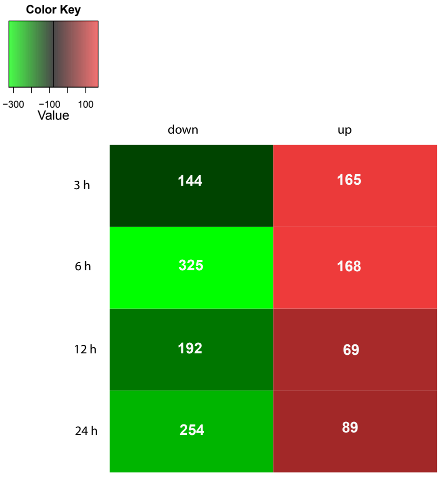
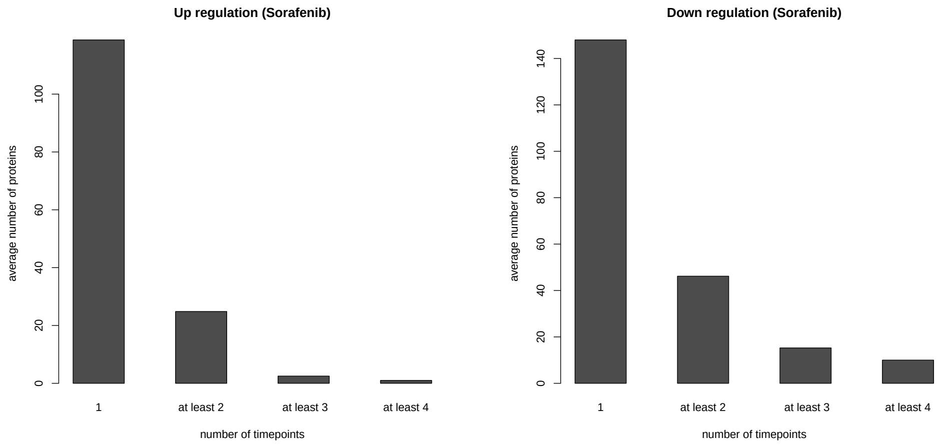


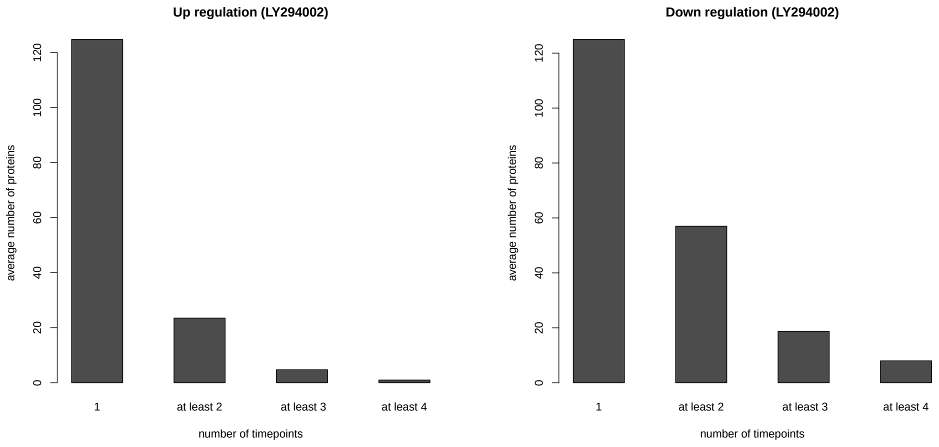
Figure 5.7: Number of significantly regulated proteins at different time points.



(a)

(b)

Figure 5.8: Number of proteins up-regulated by sorafenib (a) and number of down-regulated proteins by sorafenib (b). Proteins are found to be regulated in only one time point (1), in two (at least 2), three (at least 3) or four (at least 4)



(a) (b)

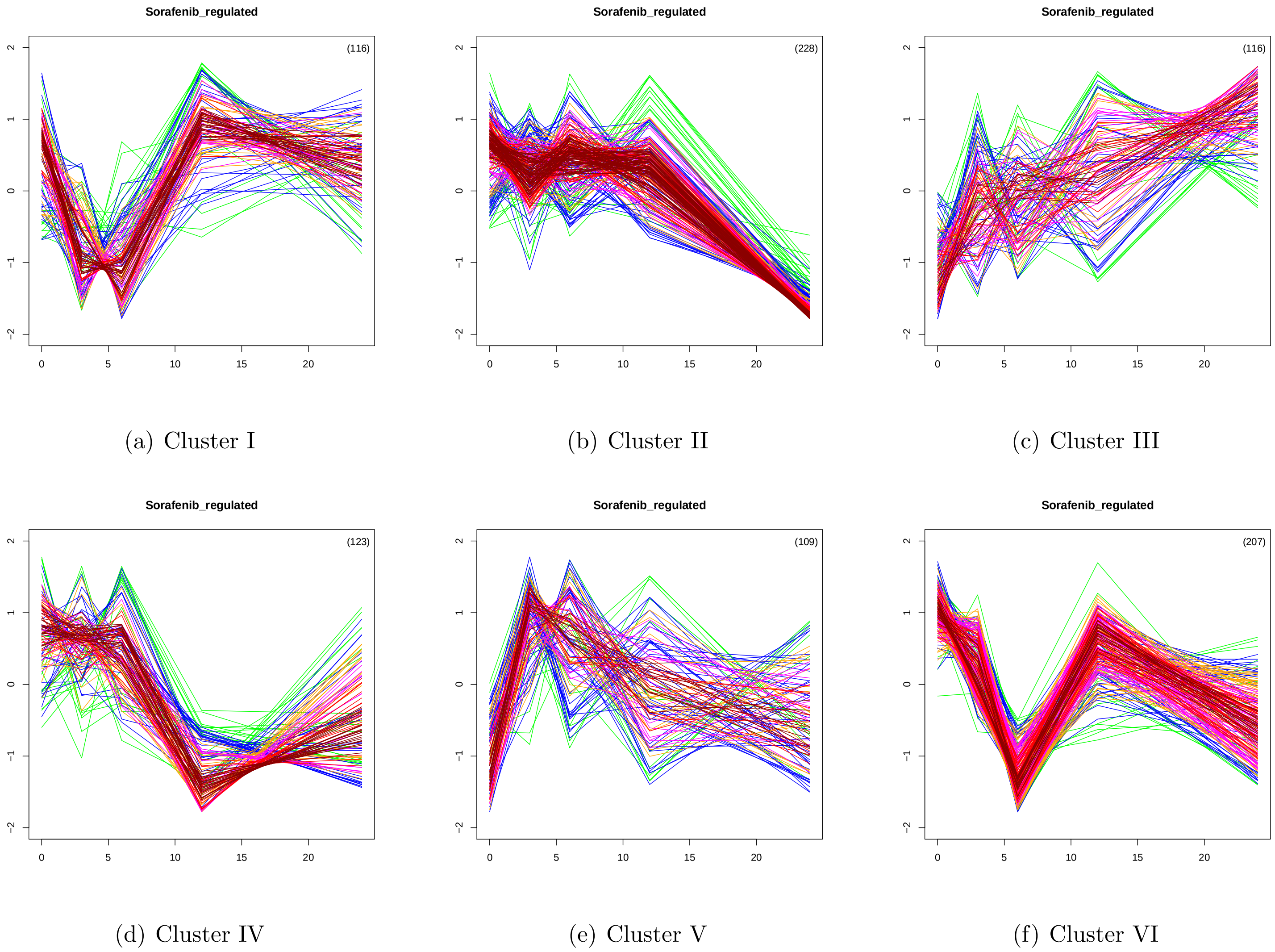
Figure 5.9: Proteins up-regulated by LY294002 (a) and down-regulated by LY294002 (b). Proteins are found to be regulated in only one time point (1), in two (at least 2), three (at least 3) or four (at least 4)

If proteins are regulated in more than one time point they are more frequently down-regulated than up-regulated for both inhibitors. The number of proteins that are up-regulated in all four time points is very low for both treatments.

Sorafenib - cluster analysis

Fuzzy c-means clustering (Bezdek, 1981) was used to identify proteins with similar time resolved protein expression patterns. The six clusters of proteins are shown in Fig. 5.10. Clusters I, II, IV and VI show very strong memberships across all associated proteins. Clusters III and V contain some proteins with lower occupancy. The number of proteins per cluster varied between 109 and 228. In general it can be observed that if a cluster contains a high number of proteins, the fraction of proteins with strong membership degrees is usually higher in in clusters with less proteins.

The number of clusters was set to six for both, the sorafenib, and the LY294002 analysis. The fuzzification parameter was set to two. In general those numbers produced the most homogeneous protein clusters. Cluster I includes 116 proteins that are slightly down-regulated at the beginning of the time course and show



(a) Cluster I

(b) Cluster II

(c) Cluster III

(d) Cluster IV

(e) Cluster V

(f) Cluster VI

Figure 5.10: Six clusters of significantly regulated proteins upon sorafenib treatment.

up-regulation at twelve h, while no difference in expression levels is observed after 24 h.

Cluster II shows strong and consistent down-regulation at the very late time point. Most of the 226 proteins included in this cluster show a very high membership degree. Strong membership degrees are assigned to proteins if the distance to the average cluster center is small.

Cluster III includes proteins that show a consistent up-regulation. The 116 proteins show more variability in the membership degrees.

Cluster IV appears to have an opposite profile compared to cluster I. Proteins are not regulated at the beginning, while they are down-regulated after twelve h and reach initial expression levels after 24 h.

The proteins in cluster V have rather weak cluster association. Cluster V is the smallest cluster with only 106 members.

A fluctuating behavior of protein expression levels is observed in cluster VI. The expression seems to be down-regulated after three to six h and after 24 h treatment, while not modulated at 12 h.

GO enrichment analysis of sorafenib regulated proteins

Cytoscape (Shannon et al., 2003) and its Plug-in ClueGo (Bindea et al., 2009) were used to analyze the large dataset, and to identify significantly over-represented biological processes of the proteins that were assigned to a common cluster. During this enrichment analysis, all Gene Ontology (GO) annotations of proteins from a cluster were compared with the annotations of the entire protein database. Hypergeometric testing and the Benjamini and Hochberg - False Discovery Rate correction (Benjamini and Hochberg, 1995) were performed to assign statistical significance to over-represented processes (Bindea et al., 2009) as described in the background Chapter.

In order to assess the main biological processes that are affected by the proteins with common time profiles, enrichment analysis for biological processes was performed for all clusters separately. Results of the enrichment analysis along with the cytoscape network visualization can be found in Fig. 5.11 - Fig. 5.13. It can be observed that proteins, clustering in cluster II, regulate most biological processes. Furthermore the network of regulated categories has a large number of hubs. Hubs are nodes in a graph that are connected to many other nodes. In this case proteins that significantly contribute to the enrichment of 'hub categories' are assigned to several groups of biological processes.

Cluster I:

Enriched biological processes for proteins from cluster I are dominated by mitochondrial respiratory chain complex assembly, β-oxidation of fatty acids, peptide cross linking, ribosome biogenesis and mRNA transport. Categories that could not form groups are related to vesicle transport, insulin response and translation.

Cluster II:

The second cluster, showing the most complex network of regulated groups with significant categories cellular respiration, mitochondria electron transport, metabolic processes of oxygen and reactive oxygen species, generation of precursor metabolites, interphase of mitotic cell cycle, cation transport, immunoglobulin transport, nucleosome assembly, establishment of organelle localization and, as already seen for cluster I, translation.

Cluster III:

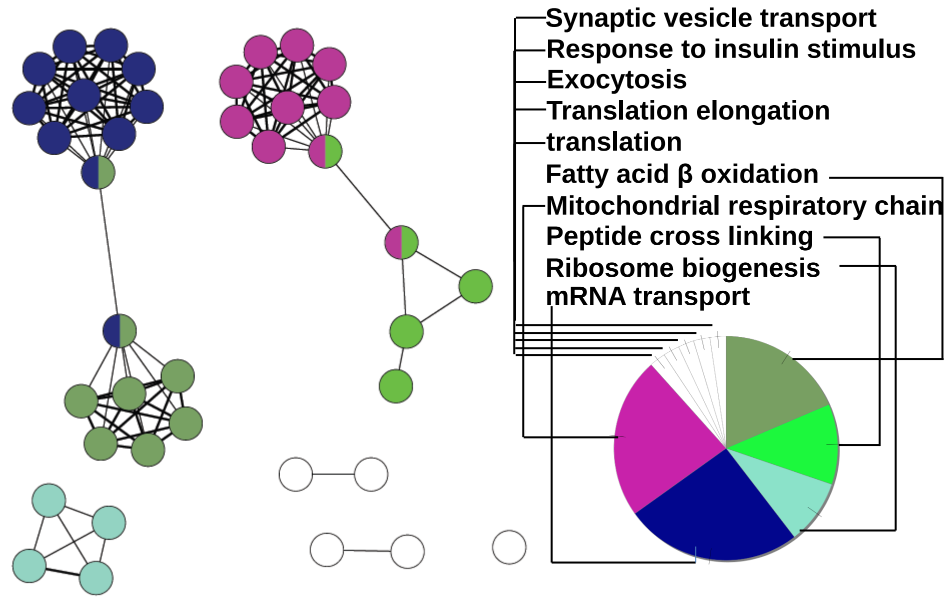
In this cluster, proteins show increasing expression levels as a function of treatment length, contains proteins, found to be enriched for RNA export from nucleus, negative regulation of ligase activity, cellular respiration, various metabolic processes, e.g. histone mRNAs and amino acids. NADPH regeneration is also significantly enriched by proteins in cluster III.

Cluster IV:

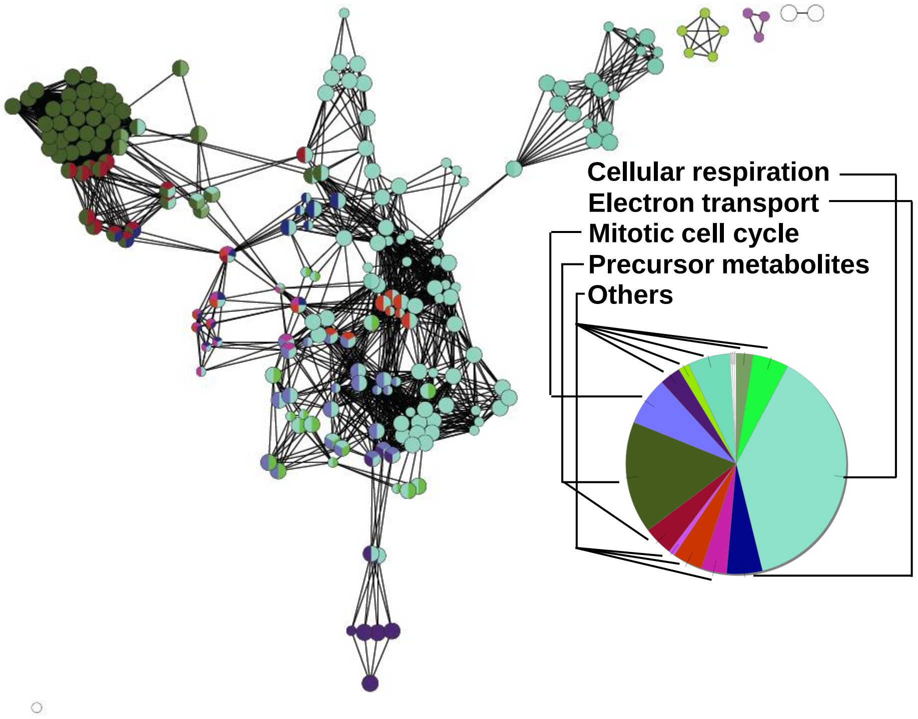
The proteins grouped here are summarized by mRNA metabolic processes, nucleus organization and maturation of SSU rRNA. GO categories that could not be assigned to groups, but show still significant enrichment, are macromolecular complex disassembly, Rho protein signaling, translation, translation elongation and autophagy.

Cluster V:

The enrichment analysis of cluster V revealed no hubs, but seven distinct groups, summarized as nucleosome assembly, RNA splicing, B cell differentiation, Vitamin B6 and sphingolipid metabolic processes, synaptic vesicle endocytosis and

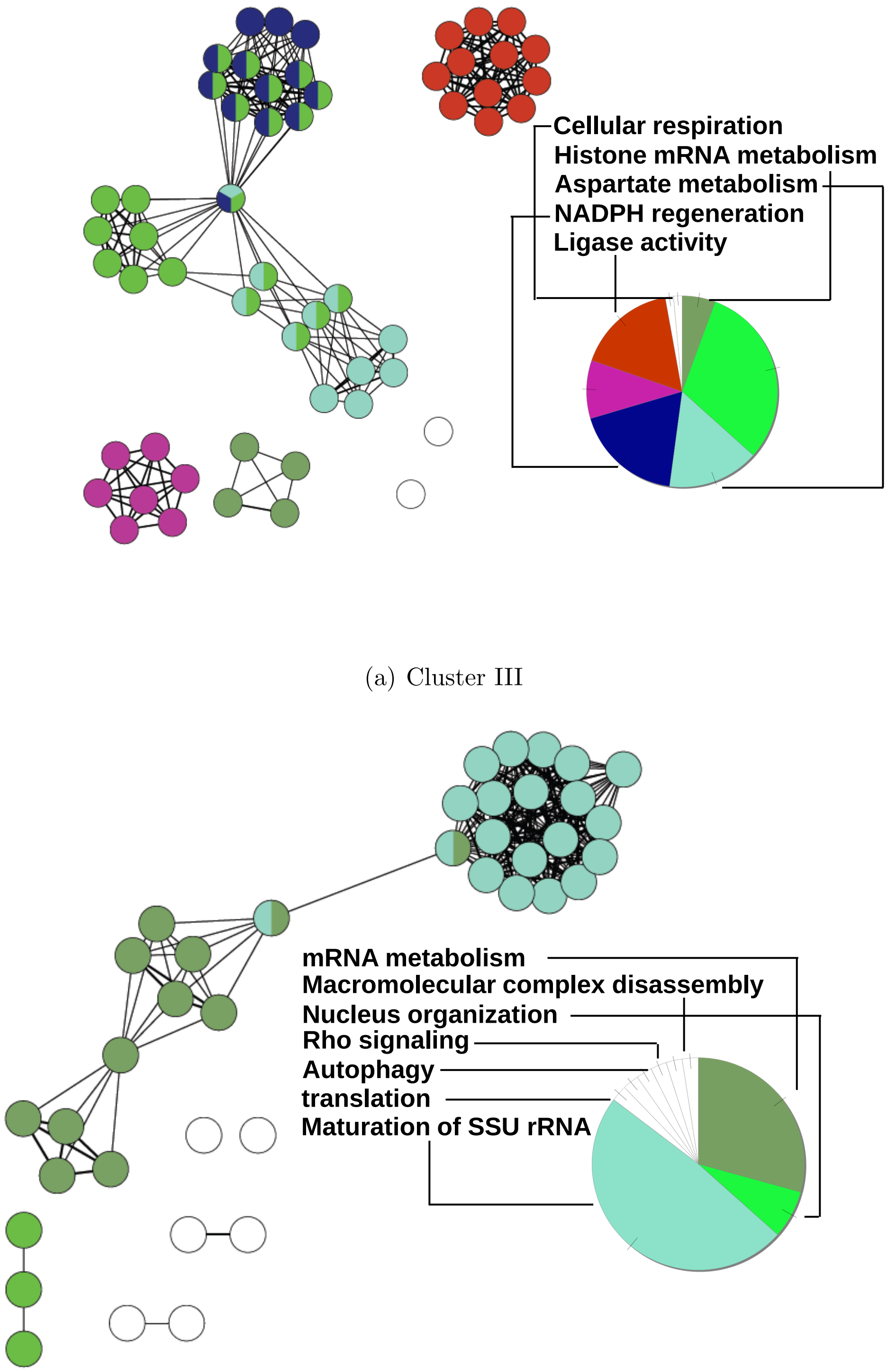


(a) Cluster I



(b) Cluster II

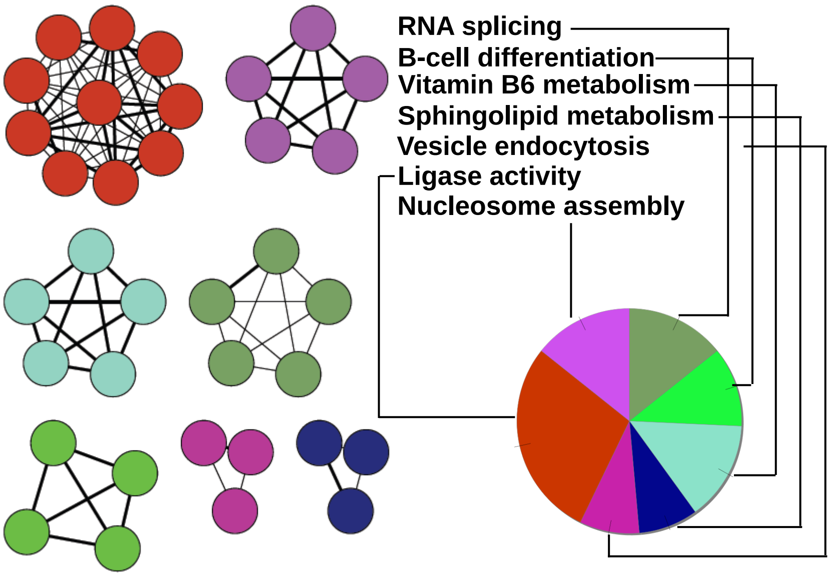
Figure 5.11: Network analysis of groups of significantly enriched biological processes upon sorafenib treatment in cluster I and II.



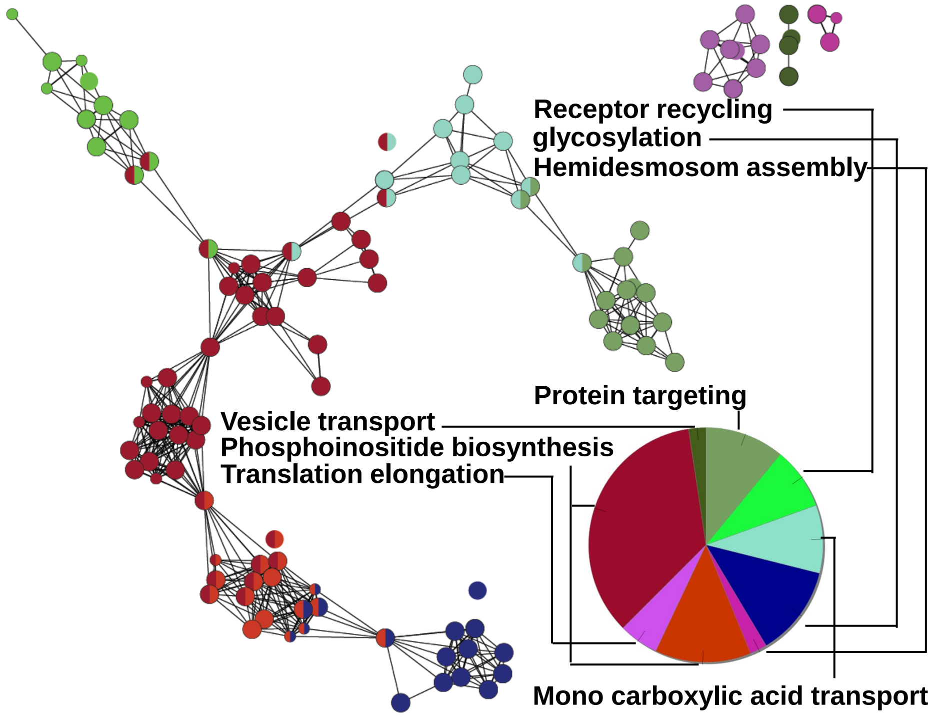
(a) Cluster III

(b) Cluster IV

Figure 5.12: Network analysis of groups of significantly enriched biological processes upon sorafenib treatment in cluster III and IV.



(a) Cluster V



(b) Cluster VI

Figure 5.13: Network analysis of groups of significantly enriched biological processes upon sorafenib treatment in cluster V and VI.

positive regulation of ligase activity.

Cluster VI:

Cluster VI shows a more interconnected network of different enriched groups, that can mainly by summarized as protein targeting, receptor recycling, mono carboxylic acid transport, glycosylation, hemidesmosome assembly, phosphoinositide biosynthetic processes, ER to Golgi vesicle-mediated transport and translation elongation.

KEGG enrichment analysis of sorafenib regulated proteins

Cytoscape (Shannon et al., 2003), version 2.6.3 along with its plugin ClueGo (Bindea et al., 2009) was used to check whether specific pathways are enriched in clusters. Similar to the GO (Ashburner, 2000) analysis, a hypergeometric test followed by a Benjamini-Hochberg (Benjamini and Hochberg, 1995) correction was performed. KEGG (Kanehisa and Goto, 2000) categories were accepted if the corrected p-value was smaller than 0.05. The results of the KEGG enrichment analysis of all proteins in cluster II are shown in Fig. 5.14. These proteins are significantly down-regulated at the late time points.

Oxidative phosphorylation was found as an enriched category in the results from the KEGG enrichment of proteins from cluster II. Pathways from neurodegenerative diseases show very significant enrichment. The proteins that are associated with the KEGG category oxidative phosphorylation are shown in Fig. 5.15, where they are mapped to the original KEGG pathway oxidative phosphorylation. Proteins from this pathway are also the main proteins in the pathways from the neurodegenerative diseases. Results from the KEGG enrichment of the other clusters can be found in the appendix. The main proteins that lead to the enrichment of pathways from neurodegenerative diseases are ATPase related proteins, proteins from the mitochondrial respiratory chain, such cytochrome c oxidase or the ubiquinol-cytochrome c reductase. The proto-oncogene tyrosine-protein kinase Fyn is down-regulated and grouped to cluster II. We also found the cycline dependent kinase 4 (CDK4), the Ras-related C3 botulinum toxin substrate 1 (RAC1) and the small GTPase, Ras homolog gene family, member A (RhoA),

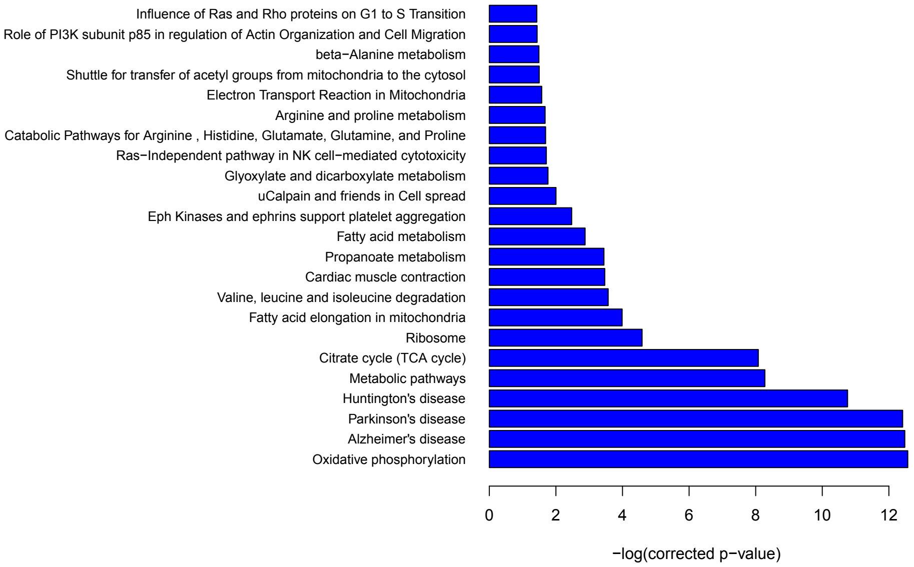
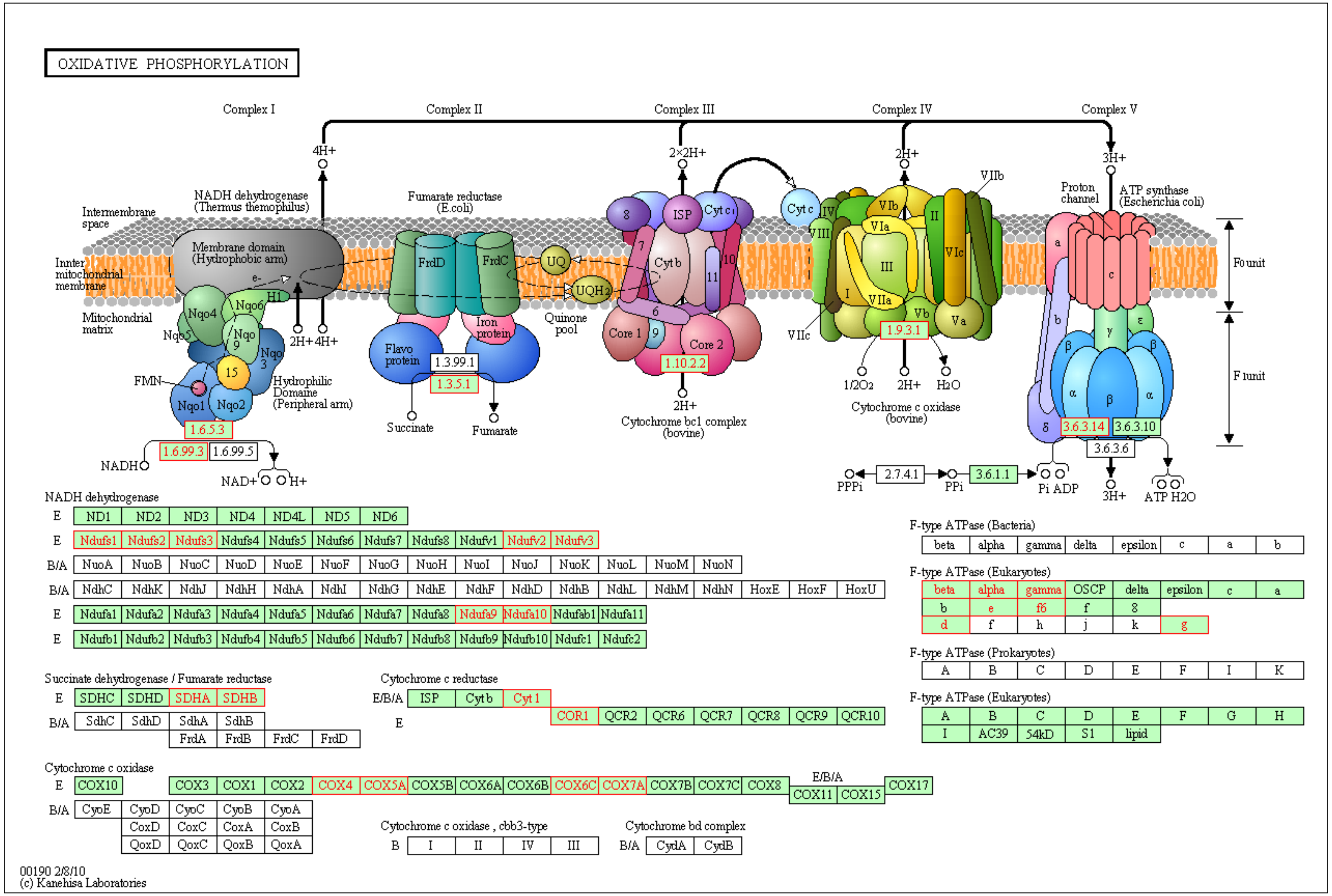


Figure 5.14: KEGG enrichment analysis of sorafenib-regulated proteins from cluster II.

as members of cluster II. These three proteins contribute to the significant enrichment of the KEGG category 'Influence of Ras and Rho proteins on G1 to S transition'.

Summary: GO and KEGG enrichment analysis - sorafenib treatment:

Interestingly translation or translation elongation seems to be affected at various times of the treatment. Proteins involved in translation have been found in all clusters and with very low p-values in clusters I, IV and VI. The mitochondrial respiratory chain complex was found as a significantly regulated category in cluster I and VI, implying significant down-regulation in early time points and also that this complex is influenced at different time points by the treatment. The enrichment in cluster I relies on the regulation of the NADH dehydrogenase, iron-sulfur proteins 4 and 5. Finally the mitochondrial electron transport and oxidative phosphorylation were found as down-regulated at the very last time point in cluster II. The enrichment of mitochondrial electron transport and oxidative phosphorylation in cluster II is due to seven down-regulated NADH



NADH dehydrogenase

E	ND1	ND2	ND3	ND4	ND4L	ND5	ND6										
E	Ndufs1	Ndufs2	Ndufs3	Ndufs4	Ndufs5	Ndufs6	Ndufs7	Ndufs8	Ndufv1	Ndufv2	Ndufv3						
B/A	NuoA	NuoB	NuoC	NuoD	NuoE	NuoF	NuoG	NuoH	NuoI	NuoJ	NuoK	NuoL	NuoM	NuoN			
B/A	NdhC	NdhK	NdhJ	NdhH	NdhA	NdhI	NdhG	NdhE	NdhF	NdhD	NdhB	NdhL	NdhM	NdhN	HoxE	HoxF	HoxU
E	Ndufa1	Ndufa2	Ndufa3	Ndufa4	Ndufa5	Ndufa6	Ndufa7	Ndufa8	Ndufa9	Ndufa10	Ndufab1	Ndufa11					
E	Ndufb1	Ndufb2	Ndufb3	Ndufb4	Ndufb5	Ndufb6	Ndufb7	Ndufb8	Ndufb9	Ndufb10	Ndufc1	Ndufc2					

Succinate dehydrogenase / Fumarate reductase

E	SDHC	SDHD	SDHA	SDHB		
B/A	SdhC	SdhD	SdhA	SdhB		
			FrdA	FrdB	FrdC	FrdD

Cytochrome c reductase

E/B/A	ISP	Cytb	Cyt1						
E				COR1	QCR2	QCR6	QCR7	QCR8	QCR9	QCR10

Cytochrome c oxidase

															E/B/A		
E	COX10		COX3	COX1	COX2	COX4	COX5A	COX5B	COX6A	COX6B	COX6C	COX7A	COX7B	COX7C	COX8	COX11	COX15	COX17
B/A	CyoE	CyoD	CyoC	CyoB	CyoA													
		CoxD	CoxC	CoxA	CoxB													
		QoxD	QoxC	QoxB	QoxA													

Cytochrome c oxidase , cbb3-type

B	I	II	IV	III

Cytochrome bd complex

B/A	CydA	CydB

F-type ATPase (Bacteria)

beta	alpha	gamma	delta	epsilon	c	a	b

F-type ATPase (Eukaryotes)

beta	alpha	gamma	OSCP	delta	epsilon	c	a
b	e	f6	f	8			
d	f	h	j	k	g		

F-type ATPase (Prokaryotes)

A	B	C	D	E	F	I	K

F-type ATPase (Eukaryotes)

A	B	C	D	E	F	G	H
I	AC39	54kD	S1	lipid			

Figure 5.15: Oxidative phosphorylation. In red are the significantly regulated proteins. (Image taken from the KEGG database)

dehydrogenases, seven proteins related to the ATP synthase and the mitochondrial cytochrome b-c1 complex subunit 1. Rho protein signaling in Cluster IV was enriched due to regulation of Rho GTPase-activating protein 5 and the LIM kinase 1. Autophagy, significantly enriched in cluster IV, is based on the regulation of Sequestosome-1 and the WD repeat domain phosphoinositide-interacting protein 1 (WIPI-1). Another autophagy protein, the autophagy marker LC-3 was found to be down-regulated after 12 h of sorafenib treatment. Although down-regulated in both replicates, LC-3 was only detected as significantly ($p < 0.05$) down-regulated in one experiment. Another interesting observation is the down-regulation of proteins involved in mitochondrial respiratory chain complex assembly at early time points and proteins involved in cellular respiration were found cluster II with a down-regulation at the very late time point. The KEGG analysis mainly revealed oxidative phosphorylation and Ras and Rho signaling as affected categories.

Ly294002 - cluster analysis

Fuzzy c-means clustering was used exactly as for the sorafenib dataset. We evaluated different parameters for the clustering and found that, as for the sorafenib experiment, six clusters and fuzzification parameter of two led to most consistent clusters (Fig. 5.16).

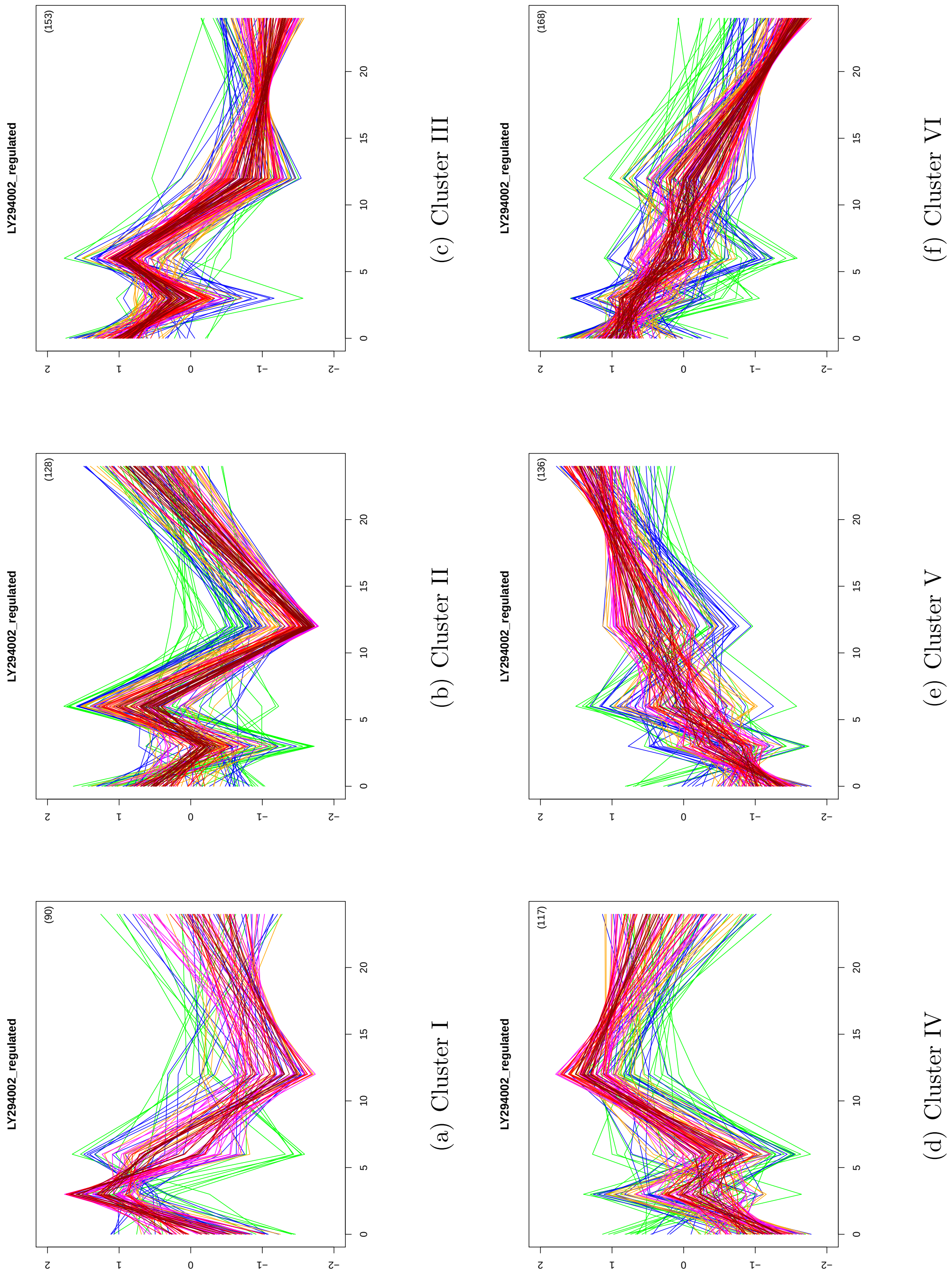
In average there were 130 proteins per cluster. Cluster I had the lowest number of members (90 proteins). The protein expression levels in cluster I reach a maximum in the very early time point and no difference in expression in observed thereafter. In general cluster I is rather sparse, meaning the average membership degree is low compared to cluster II or III.

The 128 proteins in cluster II show a fluctuating expression level, going down at three h, up at six h, down at twelve and finally up at the 24 h time point. Cluster III shows down-regulation at time points twelve and 24 h.

Cluster III contains 163 and shows strong membership degrees.

Cluster IV is very sparse, including proteins with large distances to the cluster center. It contains 117 proteins that peak around the twelve h time point.

Monotonous up-regulation is shown by proteins in cluster V. 136 proteins show increasing expression levels as a function of treatment time.



(a) Cluster I

(b) Cluster II

(c) Cluster III

(d) Cluster IV

(e) Cluster V

(f) Cluster VI

Figure 5.16: Six clusters of significantly regulated proteins upon LY294002 treatment.

Proteins in cluster VI behave inversely to the previous cluster. A monotonous down-regulation is observed for the 168 proteins in cluster VI.

GO enrichment analysis of LY204002 regulated proteins

All proteins that were quantified in all time points following the treatment with LY294002 were subjected to cluster analysis and the biological process of the proteins that were grouped in the same cluster were analyzed using Cytoscape. This analysis was performed in the same way as for the sorafenib data. The enriched groups of GO categories and the resulting cytoscape networks are shown in Fig. 5.17 - Fig. 5.19.

Cluster I:

Dominating groups of biological processes in cluster I are base pair excision repair, DNA topology change, DNA packing, rRNA processing and various categories related to translation. The network analysis shows four disjoint groups. No hubs have been found. Members of cluster I include proteins from the histone macro-H2A family and transcription factors, such as the eukaryotic translation initiation factor 5A-1.

Cluster II:

This cluster in contrast to the previous one, shows a more interconnected network. Several hubs, proteins associated with several groups of biological processes, are observed. The main groups in cluster II are grouped into oxidative phosphorylation, cell maturation, iron ion transport and various metabolic processes.

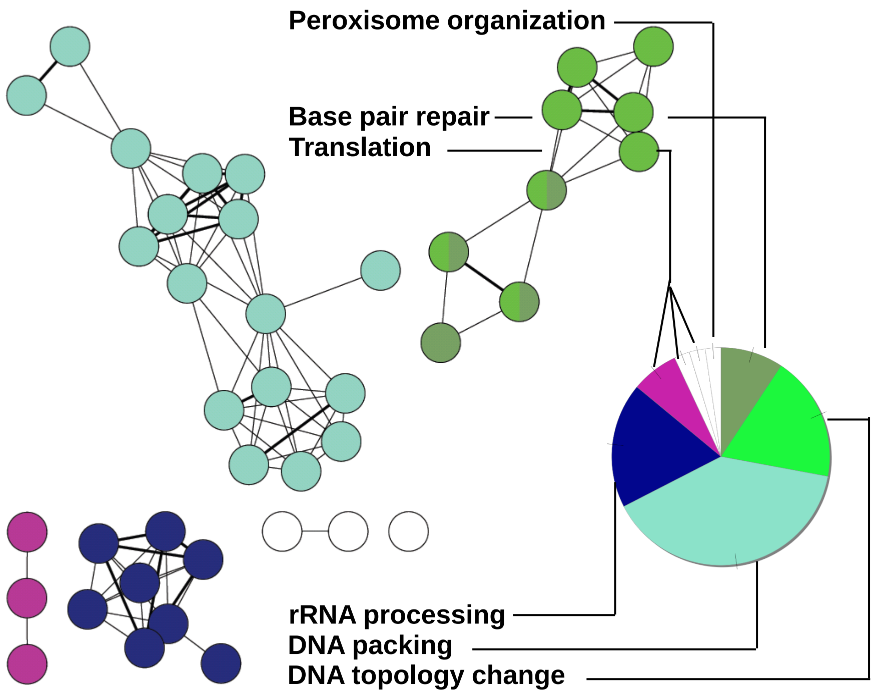
Cluster III:

The enrichment of biological processes that are affected by proteins in cluster III are mainly summarized as the organization of mitochondria, substrate adhesion and also different metabolic processes, which build most of the nodes in the network.

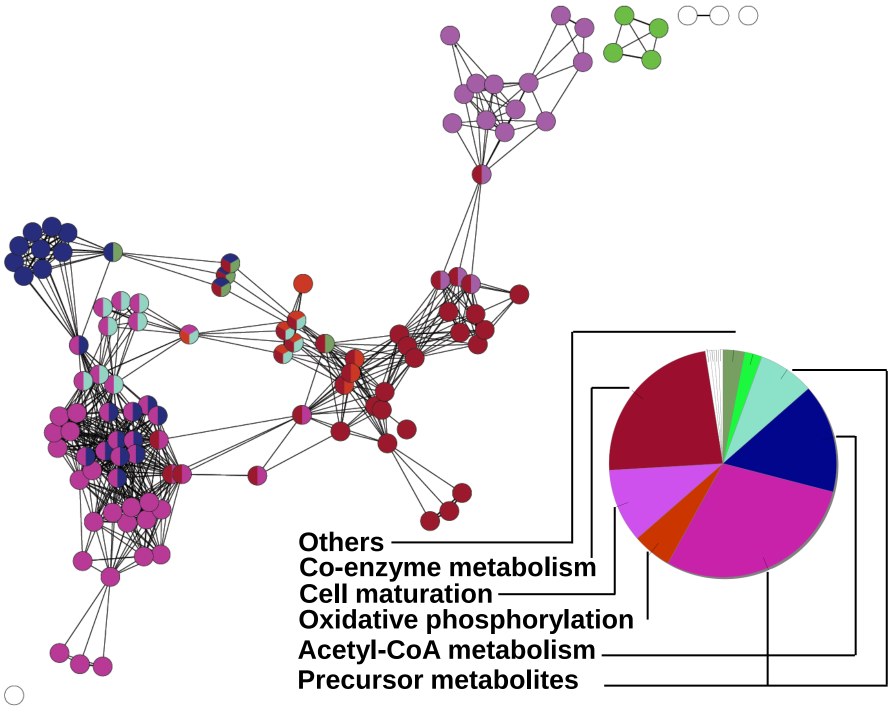
Cluster IV:

The biggest group of regulated proteins in cluster IV is named regulation of transporter activity that is connected to the nuclear division group. Further groups include nucleotide metabolic processes, as well as glycolipid transport.

Cluster V:

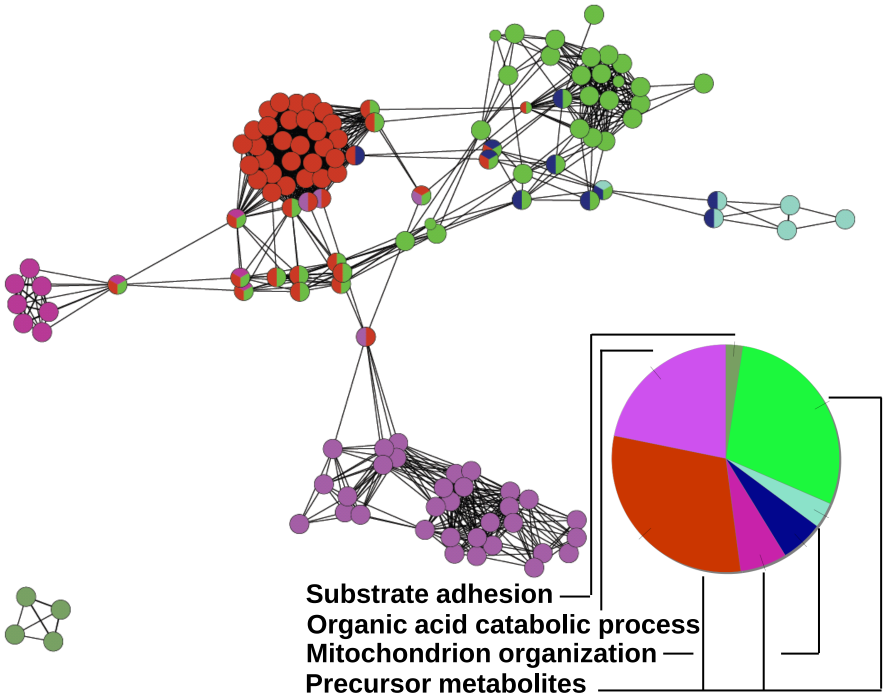


(a) Cluster I

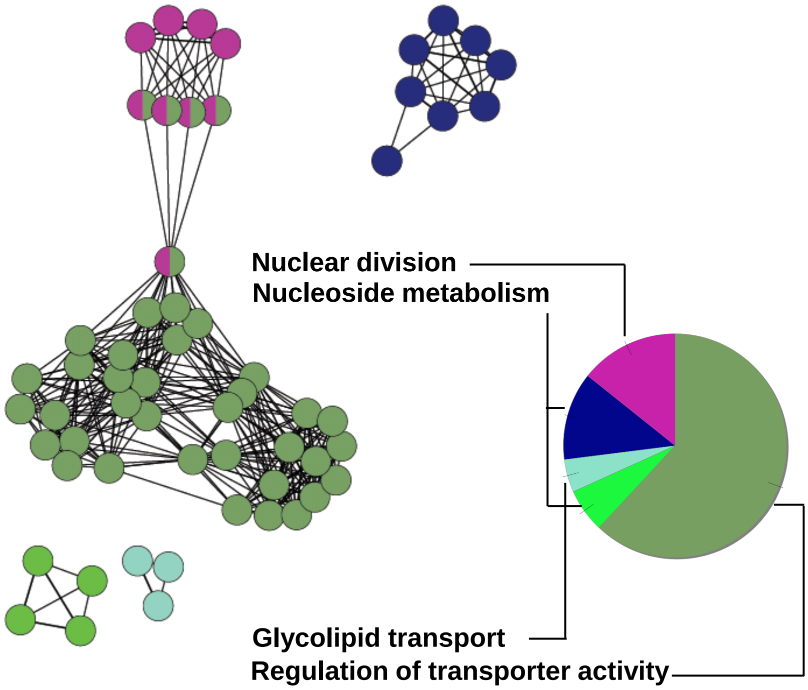


(b) Cluster II

Figure 5.17: Network analysis of groups of significantly enriched biological processes upon LY294002 treatment in cluster I and II.

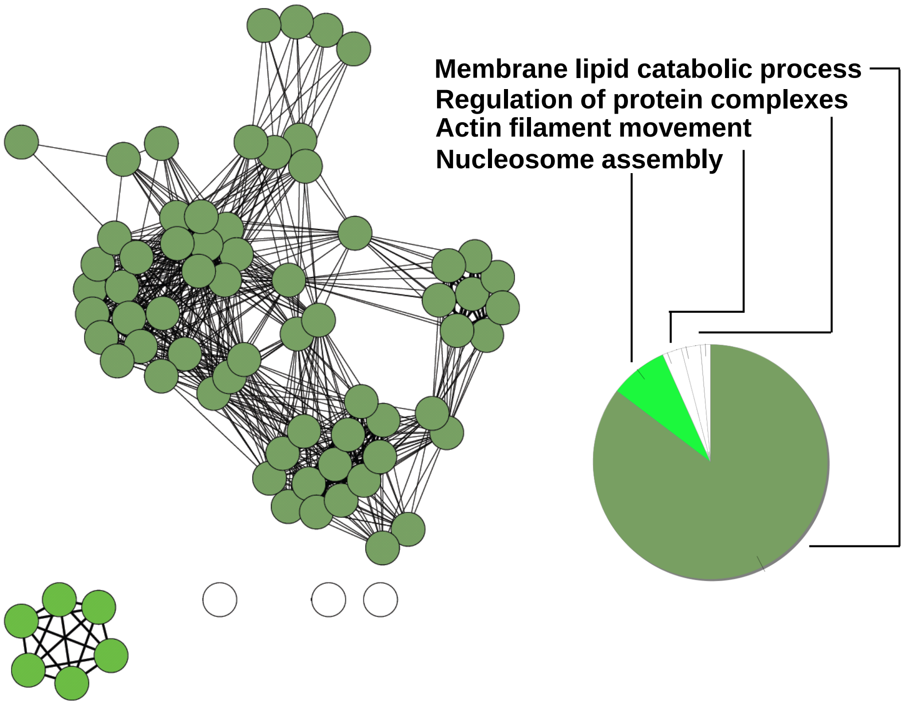


(a) Cluster III

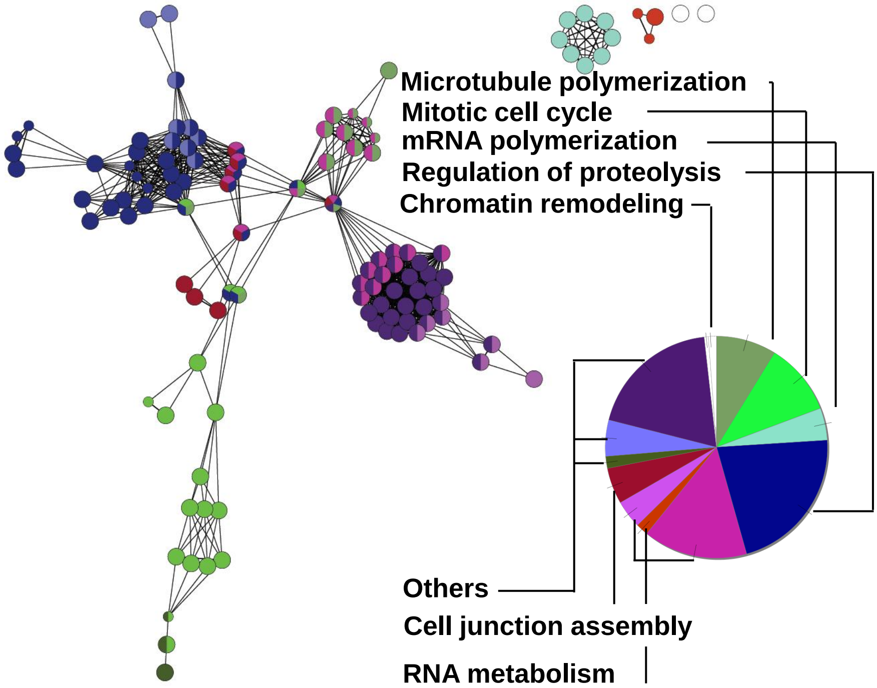


(b) Cluster IV

Figure 5.18: Network analysis of groups of significantly enriched biological processes upon LY294002 treatment in cluster III and IV.


Membrane lipid catabolic process
Regulation of protein complexes
Actin filament movement
Nucleosome assembly

(a) Cluster V



(b) Cluster VI

Figure 5.19: Network analysis of groups of significantly enriched biological processes upon LY294002 treatment in cluster V and VI.

The enriched categories in cluster V are dominated by a group labeled as membrane lipid catabolic process and a much smaller group, containing six nodes, labeled as nucleosome assembly. Categories that have not been grouped in this cluster are actin filament based movement, regulation of protein polymerization and regulation of protein complexes.

Cluster VI:

This cluster contains proteins showing monotonous down-regulation. As observed for the similar cluster of the sorafenib treatment (cluster II from the sorafenib experiment), this cluster revealed most enriched categories with a highly interconnected network. The most prominent categories include regulation of proteolysis, mitotic cell cycle, negative regulation of microtubule polymerization, nucleoside and bioribunucleoside biosynthetic processes. Interestingly, the group labeled as regulation of proteolysis, contains the category 'Ras protein signal transduction' and the hub that connects 'Ras signal transduction' to 'Rho signal transduction'. Apolipoprotein E and LIM kinase I are shared between the two categories.

KEGG enrichment analysis of LY204002-regulated proteins

The KEGG enrichment was done separately for the proteins grouped in different clusters. The KEGG enrichment results for cluster III and VI are shown in Fig. 5.20 - Fig. 5.21. *Cluster III:*

Interestingly, the KEGG categories found for cluster III contain different tumor pathways, Ras signaling, 'Erk and PI3 kinases in collagen binding'. Proteins related to the influence of PI3 kinase subunit p85 in actin polymerization. As seen for other clusters, this cluster reveals numerous categories related to metabolism. As for the sorafenib treatment, pathways related to oxidative phosphorylation are also enriched.

Cluster VI:

This cluster contains proteins that are continuously down-regulated and the KEGG analysis revealed among others the mTor pathway and DNA replication as significantly enriched categories. DNA replication was found with the most significant p-value.

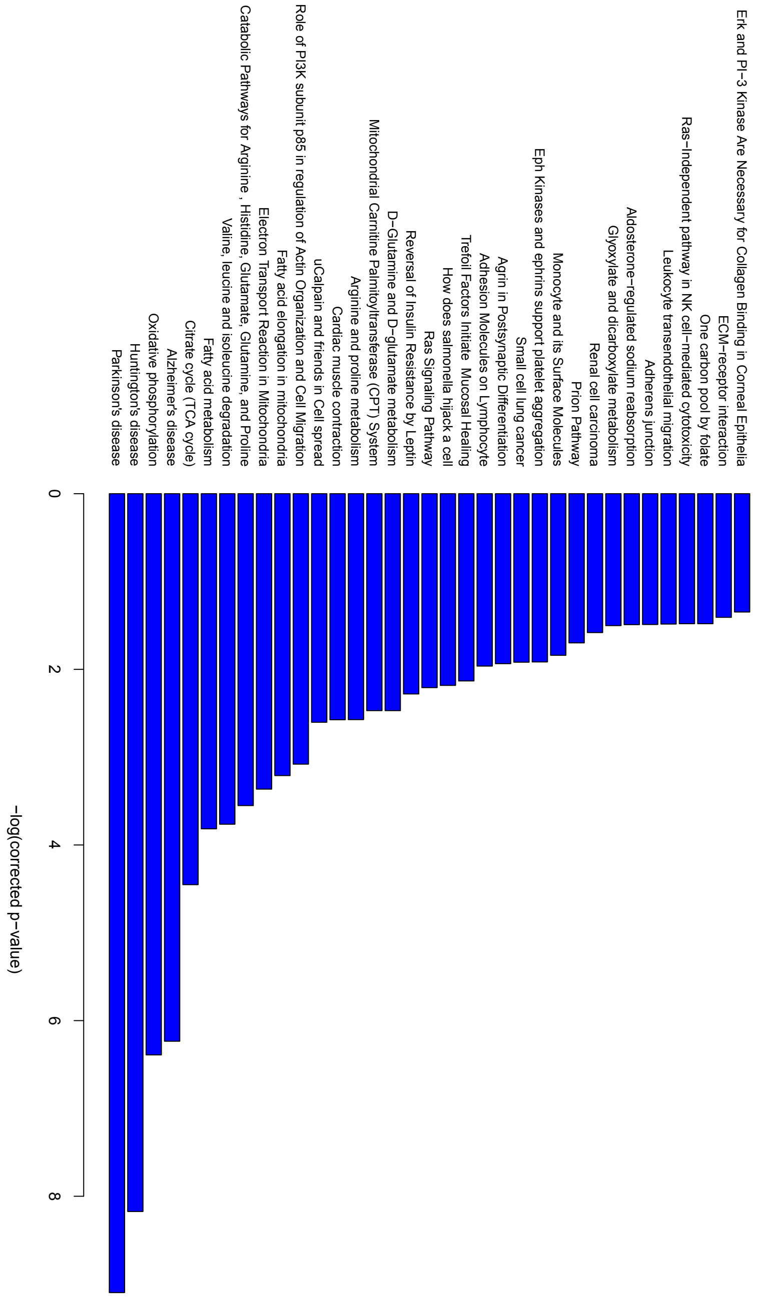


Figure 5.20: KEGG enrichment for LY294002-regulated proteins in cluster III.

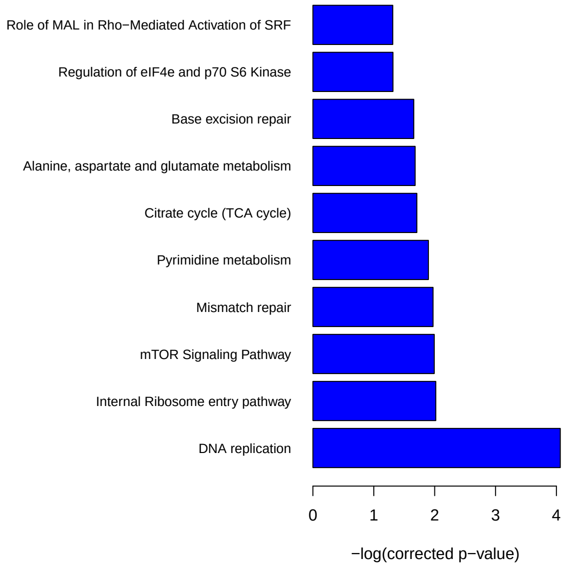


Figure 5.21: KEGG enrichment for LY294002-regulated proteins in cluster VI.

The KEGG enrichment of clusters I, IV and V can be found in the appendix.

Summary: Enrichment analysis - LY294002 treatment:
According to the GO enrichment analysis LY294002 the strongest influence of the inhibitor in observed in mechanisms related to DNA replication and translation. The cluster analysis revealed a set of clusters that contain strong members (proteins with high membership degrees), but have fluctuating protein expression profile. The expression levels of proteins that are involved in polymerization increase over time, whereas Ras and Rho signaling was observed to be down-regulated. Interestingly, the KEGG analysis revealed down-regulated pathway that involve PI3K, and mTOR, the known targets of LY294002.

Regulation of autophagy related proteins

The cluster analysis of the sorafenib experiment revealed that sorafenib treatment has an impact on the expression levels of Sequestosome-1 and the WD repeat domain phosphoinositide-interacting protein 1 (WIPI-1). Proteins related to autophagy were not found to be regulated upon LY294002 treatment.

5.2.3 Discussion

This study presents the first unbiased system wide study on the dynamic effects of the signal cascade inhibitors sorafenib and LY294002 in malignant melanoma cells. 5,408 protein groups were identified and 3,465 (approximately two thirds) could be identified and quantified in all experiments. OFFGel peptide fractionation with twelve fractions per experiment was used and at least two biological replicates were performed for each time point and treatment. The TIC chromatograms from the different OFFGel fraction show very different patterns. This suggests that different fractions contain different peptide species and allows concluding that the fractionation worked well. The analysis of peptide identifications, gained from different OFFGel fractions, revealed that peptides did not spread in more than two additional fractions. We decided to use twelve fractions in the OFFGel step. The peptide fractionation methodology theoretically allows separation of up to 24 fractions, which should theoretically lead to even further reduction of complexity, however, since each fraction was analyzed with a two h LC gradient, followed by tandem MS. Additional fractions would increase the analysis time by two h per fraction. It has been shown that using 24 fractions, the gain in protein identification rate is only at 20 % compared to twelve fractions (Hubner et al., 2008).

Fuzzy c-means clustering

Protein expression is a very dynamic process, as the cell needs to be able to quickly adapt its protein content to external stimuli or stress situations. Kinase inhibitors are strong stress factors for cells, because main functional units (signal cascades) are inhibited. The complex influence of kinase inhibitors to essential cellular processes can not be accurately described by picking just one time point. In this study the global and dynamic influence of kinase inhibitors was analyzed using the fuzzy c-means clustering methods. A variety of different clustering methods for high-throughput data have been suggested in recent years. In contrast to hierarchical clustering (Meunier et al., 2007) or probabilistic and Bayesian methods (Bensmail et al., 2005), the fuzzy c-means approach offers some advantages compared to other methods; proteins can be assigned to several clusters by calculating probability-like membership values. If a cluster consistently contains proteins with high membership degrees, the likelihood that the proteins in this cluster are truly co-regulated is increased compared to a cluster that contains only proteins with low membership degrees. However a drawback of the method is that the optimal selection of the cluster numbers is not trivial. Six clusters were chosen, as this choice results in comparable numbers of proteins in all clusters with strong membership degrees. Recently, Schwämmle and Jensen (2010) introduced a method for the estimation of parameters for fuzzy c-means clustering. Integrating those concepts for the analysis of large-scale, time-resolved proteomics data, the application of fuzzy c-means clustering will be facilitated, since the parameters optimization can be done automatically.

Sorafenib treatment

Cluster IV from the sorafenib cluster analysis revealed biological processes such as the organization and maturation of rRNAs involved in ribosome activity. It also contains the categories autophagy and Rho signaling. This enrichment is due to proteins that are strongly down-regulated at the late time points, especially after twelve h. These findings suggest that translational activity is reduced at late time points, Rho GTPases and main players in the autophagy pathway are reduced in their expression. Autophagy, the process that degrades the cell's own components has crucial roles in cellular homeostasis. Autophagy might significantly contribute to cell death introduced by sorafenib. It has been found previously that WIPI expression is dys-regulated in cancer (Proikas-Cezanne et al., 2004). Furthermore, the expression of LC-3 is down-regulated at late time points. LC-3 expression ratios were consistently below one, however significance could only be assigned in one experiment. This is due to the way significance is calculated. Peptides of very low abundance need higher ratios to be accepted as significantly regulated. The role of sorafenib in the regulation of autophagy is still unclear.

These results suggest functional follow-up experiments, investigating the different LC-3 isoforms that are used as an indicator of autophagy induction (Tanida et al., 2005). It has been shown in prostate cancer cell lines that sorafenib can induce autophagy. Ullen et al. (2010) used sorafenib at concentrations up to 25 μM and treated the cells for 72 h. These conditions are more intense compared to the sorafenib concentrations of 13 μM and the maximum of 24 h treatment used in this study. Furthermore the influence of sorafenib to the induction of autophagy can not be directly compared, as different cell lines will have different sensitivities to sorafenib treatment.

The GO term enrichment analysis of proteins showing common time-resolved expression patterns revealed translation as an affected category in several clusters. Translation was significantly regulated in clusters IV and VI. As reduced translation is directly linked to reduced proliferation this observation can be associated with the cell-death-inducing effects of sorafenib. Another main biological process is the assembly of the mitochondrial respiratory chain, which is enriched in cluster I. Cluster II is dominated by proteins that are involved in oxidative phoshorylation and electron transport chain. The regulation of these proteins might hint to the mitochondrial apoptotic pathway. In fact, especially the down-regulation of proteins like cytochrome-c oxidase can suggest that the cells reduce their respiratory activity and start to die. Recently sorafenib has been shown to induce mitochondria-dependent oxidative stress that results in cell death. Chiou et al. (2009) showed in their study for the first time in a hepatocellular carcinoma cell line that sorafenib can provoke an alternative pathway for apoptosis induction through a mitochondria-dependent oxidative stress mechanism. The involvement of sorafenib in the mitochondrial activity was also suggested by the enrichment analysis of KEGG categories. KEGG analysis was done for all clustered proteins separately. Oxidative phosphorylation was the most significant KEGG pathway among the significantly enriched pathways. Interestingly the KEGG analysis also revealed enriched pathways for neurodegenerative diseases, such as Huntington's, Alzheimer's and Parkinson's disease. Fyn kinase was found to be down-regulated and grouped to cluster II. This kinase is known to be involved in Parkinson's disease (Dunah et al., 2004). The apolipoprotein E is another candidate in cluster II and this protein is known to be involved in Alzheimer's disease (Wolk et al., 2010).

The KEGG analysis of cluster II also suggests the 'influence of Ras and Rho proteins on G1 to S transition' as an enriched category. CDK4, RAC1 and RHOA are members of cluster II. In hematopoietic cells it has been shown that activated Raf can increase CK4 expression (Chang and McCubrey, 2001). This observation can be brought forward by these data, since the decreased activity of the Raf kinase by sorafenib treatment is known and has been confirmed by others in the same cell line (Lasithiotakis et al., 2008). Furthermore, this findings can easily be integrated with the hypotheses, generated from the DIGE experiments. We have shown that the treatment with sorafenib revealed several differentially regulated isoforms of actin and vimentin. Although not yet confirmed, the DIGE analysis suggests that these isoforms might correspond to different phospho-isoforms of those proteins. Rho kinases have previously been suggested to phosphorylate filamentous proteins, such as vimentin. In our dataset we found Rho kinases to be down-regulated following sorafenib treatment. These findings support the hypotheses formulated in Chapter 3.

LY294002 treatment

The cluster analysis of the proteins regulated by LY294002 revealed six significant clusters of co-regulated proteins. Proteins in cluster II, III, V and VI show strong membership degrees. Proteins in clusters I and especially V show more variations around the center of the cluster. Both clusters I and V include comparable low numbers of proteins and there is no clear trend of regulation. The expression of their members appears to be rather fluctuating. This might be explained by unspecific regulation. Although all proteins included in the clustering analysis show corrected significance ratios below the threshold of 0.05, there is still the probability that this regulation might be independent of the treatment. Cluster VI contains proteins involved in relevant signaling pathways, such as the Ras or Rho signaling pathway. Crosstalk of the MAPK and PI3k-Akt pathway has previously been investigated in the same cellular model. Lasithiotakis et al. (2008) showed that LY294002 significantly inhibits Akt phosphorylation, but does not have any influence on Erk phosphorylation, suggesting no interference of LY294002 in the Ras-Raf-Mek-Erk pathway. These data show a significant enrichment of the Ras signaling pathway within proteins that show monotonous reduction in expression as a function of treatment length. These findings further suggest an indirect influence of LY294002 to the Ras pathway by down-regulating the expression of apolipoprotein E and LIM kinase I. Apolipoprotein E has been shown to be involved in the proliferation and survival in apoE-expressing ovarian cancer cells (Chen et al., 2005). Although the clear mechanism is still missing, non affected Erk activity in combination with down-regulated expression of proteins in the Ras pathway allows hypothesizing for crosstalk between the two pathways. Intriguingly proteins that are up-regulated at early time points followed by down-regulation at late time points, three or six h are frequently associated with metabolism and translation (especially proteins grouped in cluster I). Whereas transcription factors are known to be immediate early genes, genes involved in metabolism are *per se* not associated with early responses to perturbation. However it is known that the induction of genes for the generation of precursor metabolites, such as phosphofructokinase-1 (PFK-1), glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and L-pyruvate kinase (L-PK) can be induced very rapidly following stimulation with glucose (Roche et al., 1997). It was found that mRNA of glycolytic enzymes can be increased by up to seven fold as early as 30 min after stimulation. Also not commonly known as immediate early genes, these results suggest that the induction of metabolic activity as an early response to the inhibition of major signaling pathways might be of biological relevance for the cancer cell. This mechanism might be associated with first unspecific gene expression as a reaction to perturbation. Proteins involved in metabolic processes are not enriched in clusters showing no or moderate regulation at early time point (clusters IV to VI). Although less pronounced, similar observations are made for the sorafenib treatment. In cluster V, the only cluster where sorafenib regulated proteins peak at early time points, metabolic process are enriched. In contrast to the LY29402 experiment, metabolic processes were found to be enriched in the cluster with monotonous up-regulation (cluster II). Some proteins were found to be equally regulated in both treatments. Prominent examples are the apolipoprotein E and the LIM 1 kinase. Furthermore categories related the oxidative phosphorylation and translation were also found to be affected by both inhibitors. Although it can not be pinpointed whether the regulation of the expression of these proteins is a direct influence of the multiple inhibition of major signaling cascades, the disruption of the cellular homeostasis may impact on a variety of mechanisms. Mao et al. (2007) introduced the notion of 'balancer' proteins. Balancer proteins are defined as proteins that buffer or cushion a system and can therefore oppose multiple system disturbances. The proteomic response to stimuli or perturbations results in many cases in large numbers of regulated proteins. Mao et al. (2007) hypothesized that the observed protein changes might partially be explained by a proteomic network response and that mainly the modulated expression of 'balancer' proteins can be used to explain this response. In most quantitative proteomics experiments the protein samples are mixed equally according the protein content. Yet if proteins appear to be regulated by any stimulus, other proteins have to be regulated in the opposite direction to balance the 'weight-loss'. Interesting future research directions might be to investigate, if specific sets of balancing effects might be correlated to specific perturbations of a biological system. Mao et al. (2007) based their results on work using 2D-PAGE; using shotgun proteomics datasets, insights towards the effects of balancer proteins can be founded on more reliable statistics and on a much greater proteome coverage.

Chapter 6

Conclusions and perspectives

In this thesis we developed novel methods for the analysis of proteomics experiments, a map alignment method for DIGE-based proteomics and a probabilistic method for the integration of different tandem MS search engine results. Proteomics methods were applied to an important topic in cancer research, the profiling of modulated protein expression following treatment with small molecular weight kinase inhibitors. Proteomics has classically been done by using 2D-PAGE to separate proteins before the mass spectrometric identification. In recent years shotgun methods have gained popularity as methods for whole proteome analysis. In this last chapter the results gained from gel-based approaches will be compared to those gained from SILAC-based shotgun analyses. Future research directions will be indicated accordingly. Furthermore we will speculate about the future impact of proteomics and high-throughput biology to cancer research.

The innovation of difference gel electrophoresis (DIGE) was a great asset for 2D-PAGE based proteomics. The technology allows separating up to three proteomes on one gel. Before the invention of the CyDye labeling every sample had to be separated on different 2D gels. The use of CyDye labels for 2D-PAGE allows separating several samples on one gel. This results in a perfect matching of samples separated on the same gel. However, in most experiments it is not enough to run one DIGE gel, as more samples are necessary for statistical analysis. Running several DIGE gels, the matching problem is essentially the same as in classical 2D-PAGE without CyDye labeling. Protein spots need to be matched across gels.

Although the graph-based method that was developed in this thesis facilitates this task, the technology has other drawbacks. The whole analysis workflow is very labor-intensive, because all steps need manual intervention. The main advantage of 2D gel analysis is the analysis of protein isoforms. Proteins are separated according to their isoelectric point and their molecular mass. These properties allow separating most isoforms, such as phosphorylated isoforms or cleaved proteins. In Chapter 3 we have shown that the cytoskeleton proteins β-actin and vimentin occur in different isoforms that are differentially regulated. Shotgun proteomics experiments, in contrast, investigate peptides, the smaller fragments of proteins, and do not analyze intact proteins. The detection of isoforms is by far not as easy as on a 2D PAGE. Identified peptides can originate from various isoforms of the same gene product. To account for proteins that are indistinguishable based their identified peptides the term protein group is frequently used in shotgun proteomics. A protein group refers to all proteins that originate from the same gene. Approaches to overcome this drawback of shotgun proteomics are specialized workflows, designed for the enrichment of specific isoforms of proteins, such as phosphorylated (Olsen et al., 2006), acetylated (Basu et al., 2009) or ubiquitinated (Peng et al., 2003) forms of proteins. Although such dedicated protocols increase the workload, the gain of information is high. One important advantage of gel-free proteomics is the potential for high automation. Using state of the art auto sampling injectors in combination with liquid chromatography systems that are online coupled to a mass spectrometer, the whole experimental workflow from the tryptic peptide to the raw LC-MS data can be automated. Running the experiment with calibrated instruments and defined methods, complete automation of the data analysis process is possible. In contrast to the 2D-PAGE approach, the automation in shotgun workflows reduces the degree of subjectivity tremendously. In conclusion, the main advantage of gel-based proteomics, the resolution of isoforms, might be important to specialized questions; however, the depth of the proteome, phosphoproteome or the acetylome that is be covered by shotgun proteomics, is very suitable for global system-wide analyses. It is foreseeable that the importance of gel-free approaches in proteomics will increase further. Future research directions in proteomics should focus on proteome-wide profiling of multiple modifications in parallel. This demanding task, if accompanied by protein expression analysis, will allow to assess the stoichiometry of modification site occupancy. Such data might yield unprecedented insights to the biological system.

The second major result of this thesis is the development of a computational framework for the integration of multiple search engines. We showed that this new approach significantly increases the peptide identification rates in shotgun proteomics. Although methods for the combination of search engines have been suggested before, the consensus scoring method, presented here, is the first method that considers peptide similarity if search engines do not agree on their suggested peptide-spectrum match. The comparison of different algorithms, that are widely used for peptide identification also showed that the performance of single engines can vary in different datasets. Taken together, this suggests that multiple search engines should be applied in high-throughput studies. It has been shown that this technique is highly beneficial, even for high-accuracy datasets, where high precision mass measurements facilitate the annotation of tandem MS spectra. Future research directions in the post-processing of database search results can be the integration of multi-pass methodologies and especially the statistical assessment of results gained from multi-pass strategies. Multi-pass strategies perform several runs of the search engine by integrating information from the previous run. A common multi-pass search strategy selects a subset of proteins, which have been reliably identified by an initial search, and constructs a new database with only those proteins that have already been found. This strategy relies on the assumption that if peptides from a protein have already been found, there is a high change that the unidentified spectra correspond to the other peptides from the same proteins. Additional search runs are then performed with different modification settings, allowing more missed cleavages or polymorphisms. Current multi-pass solutions (e.g., X!Tandem's refinement function) do not properly provide solutions to assess the significance of results gained by repeated searches against a smaller database. Besides the development of strategies for scoring with multiple engines, the construction of spectral libraries is a strategy that will gain importance in the future. It has become evident that a relatively large fraction of peptides is rediscovered in every proteomics experiment. This might be avoided by intelligent design of libraries based on retention time and accurate mass of the precursor ions combined with algorithms for *on the fly* decision if a precursor needs to be subjected to tandem mass spectrometry analysis or not. The development of novel computational and mass spectrometric methods that enable such workflows might become a further direction of proteomics research.

Throughout the thesis, proteomics technology has been applied to open questions in cancer research. All experiments, investigating the effects of kinase inhibitors to global protein expression, were done using cancer cell lines. Such cancer cell lines are very appropriate models to study the disease in the laboratory. For quantitative proteomics, cell line models can easily be SILAC-labeled, allowing accurate proteome-wide quantitation of protein expression. However, important aspects for clinical applications are neglected by looking only at cell culture models. Genetic variability, tissue environment and different life conditions are playing important roles in the development of cancers in eukaryotic organisms and especially in humans. These factors can only be addressed by looking at clinical samples. The analysis of the latter will become more important in future directions of proteomics-based cancer research. The decreased sample amount and the inability for stable isotope labeling may be reasons why proteomics was so far not successful with clinical samples. The ultimate goal for high-throughput platforms in clinical applications is the discovery of drug targets or biomarker molecules. Biomarkers are disease-related molecules that function as monitors for disease progression, status or therapy effectiveness. Drug targets are key molecules involved in a particular metabolic or signaling pathway that is specific to a disease condition or pathology, or to the infectivity or survival of a microbial pathogen (Kaplan, 2007). The future of proteomics applications in clinical research, however, is promising. Mass spectrometric technology has reached a level where very sensitive and reproducible assays can be brought to clinical application. Accompanied by further development in computational proteomics, as discussed above, high-throughput analyses of different tissues from patients will become possible.

The general advantage of proteomics methods over more classical, hypothesis-driven methods for protein identification and quantitation, such as western blotting, is the ability to profile large parts of the expressed proteome in parallel. In this way high-throughput proteomics experiments will create large datasets that are very valuable for molecular biology and especially for the emerging field of systems biology. Proteomics being just one example of a high-throughput platform, future investigations on complex questions should employ global analyses on different biological levels, such as the genome, transcriptome and the metabolome. All these data need to be integrated in the right way to mine this information. The community will need better methods for data processing and data mining to allow rapid queries of such system-wide comprehensive datasets. Only properly measured and analyzed high-throughput data will allow mathematical modeling. Theoretical models of cellular mechanisms can significantly contribute to discover aspects in this mechanisms that can not be seen, by a purely qualitative assessment of the data.

Appendix A

Appendix

A.1 PAM30MS substitution matrix

PAM (point accepted mutation) matrices belong to the first amino acid substitution matrices. In 1970, The concept of PAM matrices was developed by Dayhoff et al. (1978). The entries in the PAM matrices are calculated on the basis of closely related proteins. The PAM1 matrix indicates the rate of expected amino acid substitutions, if 1 % of the amino acids would be changed. In this case the proteins would have 99 % sequence identity. The PAM1 matrix is used as the basis for calculating other matrices by assuming that repeated mutations would follow the same pattern as those in the PAM1 matrix, and multiple substitutions can occur at the same site (Fulekar, 2009). The PAM30 matrix is one of the most commonly used matrix. In the PAM30MS, a similarity based on tandem MS peptide spectra is added integrated into the PAM similarity concept. The substitution probabilities in a PAM matrices are multiplied by 1,000.

Table A.1: The PAM30MS substitution matrix.

test	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	6	-7	-4	-3	-6	-4	-2	-2	-7	-5	-6	-7	-5	-8	-2	0	-1	-13	-8	-2	-7	-6	0	-17
R	-7	8	-6	-10	-8	-2	-9	-9	-2	-5	-7	0	-4	-9	-4	-3	-6	-2	-10	-8	5	-1	0	-17
N	-4	-6	8	2	-11	-3	-2	-3	0	-5	-6	-1	-9	-9	-6	0	-2	-8	-4	-8	-4	-2	0	-17
D	-3	-10	2	8	-14	-2	2	-3	-4	-7	-10	-4	-11	-15	-8	-4	-5	-15	-11	-8	-7	-3	0	-17
C	-6	-8	-11	-14	10	-14	-14	-9	-7	-6	-11	-14	-13	-13	-8	-3	-8	-15	-4	-6	-11	-14	0	-17
Q	-4	-2	-3	-2	-14	8	1	-7	1	-8	-7	-3	-4	-13	-3	-5	-5	-13	-12	-7	-3	4	0	-17
E	-2	-9	-2	2	-14	1	8	-4	-5	-5	-7	-4	-7	-14	-5	-4	-6	-17	-8	-6	-7	-2	0	-17
G	-2	-9	-3	-3	-9	-7	-4	6	-9	-11	-11	-7	-8	-9	-6	-2	-6	-15	-14	-5	-8	-7	0	-17
H	-7	-2	0	-4	-7	1	-5	-9	9	-9	-8	-6	-10	-6	-4	-6	-7	-7	-3	-6	-4	-3	0	-17
I	-5	-5	-5	-7	-6	-8	-5	-11	-9	8	5	-6	-1	-2	-8	-7	-2	-14	-6	2	-6	-7	0	-17
L	-6	-7	-6	-10	-11	-7	-7	-11	-8	5	5	-7	0	-3	-8	-8	-5	-10	-7	0	-7	-7	0	-17
K	-7	0	-1	-4	-14	-3	-4	-7	-6	-6	-7	7	-2	-14	-6	-4	-3	-12	-9	-9	5	4	0	-17
M	-5	-4	-9	-11	-13	-4	-7	-8	-10	-1	0	-2	11	-4	-8	-5	-4	-13	-11	-1	-3	-3	0	-17
F	-8	-9	-9	-15	-13	-13	-14	-9	-6	-2	-3	-14	-4	9	-10	-6	-9	-4	2	-8	-12	-14	0	-17
P	-2	-4	-6	-8	-8	-3	-5	-6	-4	-8	-8	-6	-8	-10	8	-2	-4	-14	-13	-6	-5	-5	0	-17
S	0	-3	0	-4	-3	-5	-4	-2	-6	-7	-8	-4	-5	-6	-2	6	0	-5	-7	-6	-4	-5	0	-17
T	-1	-6	-2	-5	-8	-5	-6	-6	-7	-2	-5	-3	-4	-9	-4	0	7	-13	-6	-3	-5	-4	0	-17
W	-13	-2	-8	-15	-15	-13	-17	-15	-7	-14	-10	-12	-13	-4	-14	-5	-13	13	-5	-15	-7	-13	0	-17
Y	-8	-10	-4	-11	-4	-12	-8	-14	-3	-6	-7	-9	-11	2	-13	-7	-6	-5	10	-7	-10	-11	0	-17
V	-2	-8	-8	-8	-6	-7	-6	-5	-6	2	0	-9	-1	-8	-6	-6	-3	-15	-7	7	-9	-8	0	-17
B	-7	5	-4	-7	-11	-3	-7	-8	-4	-6	-7	5	-3	-12	-5	-4	-5	-7	-10	-9	5	1	0	-17
Z	-6	-1	-2	-3	-14	4	-2	-7	-3	-7	-7	4	-3	-14	-5	-5	-4	-13	-11	-8	1	4	0	-17
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-17
*	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	1

A.2 Marginal distribution

Lemma 1. *The marginal distribution of* x *is a mixture of Gaussian distributions*

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sigma_k)$$

Proof. In Chapter 2 the general mixture model and the marginal distribution were introduced. Let us introduce $\mathbf{z} = \{z_1, ..., z_k\}$, a k-dimensional, binary random variable with

$$\exists\ i \text{ with } z_i = 1 \text{ and } \forall\ j\ \neq\ i:\ z_j = 0.$$

This is formulated as: $z_j \in \{0; 1\}_{\forall j}$ and $\sum_{j=1}^{k} z_j = 1$. For the joint distribution

$$p(\mathbf{x}, \mathbf{z}) := p(\mathbf{z})\ p(\mathbf{x}|\mathbf{z})$$

we can denote

$$p(z_k = 1) = \pi_k$$

with the π_i satisfying the criteria introduced above

$$0 \leq \pi_i \leq 1; \sum_i \pi_i = 1$$

In that sense π_i are valid probabilities and $p(z)$ can thus be written as

$$p(z) = \prod_{j=1}^{k} \pi_j^{z_j}$$

In a similar way, we can write the conditional distribution of x for a given value of $\mathbf{z}$ as a Gaussian

$$p(\mathbf{x}|z_k = 1) = N(\mathbf{x}|\mu_k, \sigma_k)$$

which equals

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x}|\mu_k, \sigma_k)^{z_k}$$

For the joint distribution of $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ we get

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sigma_k)$$

□

A.3 Determination of significance of SILAC ratios

The significance of a differentially regulated protein was determined as follows (these calculations are based on the significance A and significance B formulas described by (Cox and Mann, 2008)). We assume a normal distribution (μ=0 and σ=1) of the natural logarithms of all ratios. r_{-1}, r_0 and r_1 respectively correspond to the 15.87, 50, and 84.13 percentiles of all ratios. z is defined as the significance measure for a ratio r.

$$z = \begin{cases} \frac{r-r_0}{r_1-r_0}, \text{ for } r > r_0 \\ \frac{r_0-r}{r_0-r_{-1}}, \text{ for } r < r_0 \end{cases}$$

Given a ratio r $\in$ $\mathbb{R}$, the probability that any ratio (random variable) $X \in N(\mu, \sigma^2)$ takes a value less or equal to r is given by

$$P\{X \leq r\} = \int_{-\infty}^{r} f(x)dx = \int_{-\infty}^{r} \frac{1}{\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

The Gaussian error function is defined as,

$$\text{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{\frac{-y^2}{2}}$$

with $x = \frac{r-\mu}{\sigma}$, which corresponds the z-transformation of the ratio r, we get

$$P\{X \leq r\} = \begin{cases} 0.5 - \text{erf}(\frac{\mu-r}{\sigma}), \text{ for } r \leq \mu \\ 0.5 + \text{erf}(\frac{r-\mu}{\sigma}), \text{ for } r \geq \mu \end{cases}$$

The probability that any measurement X is in $[-r, r]$ is

$$P\{X \leq r\} - P\{X \leq -r\} = 0.5 + \text{erf}(\frac{\mu - r}{\sigma}) - (0.5 - \text{erf}(\frac{r - \mu}{\sigma}))$$

$$\Leftrightarrow \mathrm{erf}(\frac{r-\mu}{\sigma}) + \mathrm{erf}(\frac{\mu-r}{\sigma})$$
$$\Leftrightarrow 2\mathrm{erf}(\frac{r-\mu}{\sigma})$$
$$\Leftrightarrow \mathrm{erf}(\frac{r-\mu}{\sigma\sqrt{2}})$$

The probability that any measurement X is not in $[-r, r]$ is thus,

$$P\{X \geq r\} = 1 - \mathrm{erf}(\frac{r-\mu}{\sigma\sqrt{2}}) = \mathrm{erfc}(\frac{r-\mu}{\sigma\sqrt{2}})$$

with the percentiles as defined above, we have

$$\mathrm{erfc}(\frac{r-\mu}{\sigma\sqrt{2}}) = \mathrm{erfc}(\frac{r-r_0}{(r_1-r_0)\sqrt{2}})$$
$$\Leftrightarrow \mathrm{erfc}(\frac{z}{\sqrt{2}})$$

with $z \geq 0 \; \forall$ r $\in \Re$

$$P\{X \geq r\} = \frac{1}{2}\mathrm{erfc}(\frac{z}{\sqrt{2}})$$

To account for the intensity dependence of the ratio, all proteins are binned in intensity bins. One intensity bin contains at least 300 proteins (Cox and Mann, 2008). All calculations above are done in each intensity bin separately. A protein with ratio r was set to be differentially regulated if the significance was 5 % or smaller. This 5 % significance is the likelihood to observe proteins with a ratio significance measure z that high or higher. This can loosely by interpreted as a 5 % probability that the null hypothesis, *differentially expressed protein*, is wrong. If proteins were found to be differentially regulated following DMSO and inhibitor treatment, the regulation in the inhibitor experiment had to be twice as high as in the DMSO treatment to accept this differential regulation.

A.4 KEGG categories

Sorafenib-regulated proteins grouped in cluster I and III did not reveal any significant KEGG enrichment results. The results from the KEGG enrichment analysis of the proteins, grouped in clusters IV, V and VI are shown in Fig. A.1. KEGG enrichment results for the LY294002-regulated proteins grouped in clusters I, II, IV and V are shown in Fig. A.2.

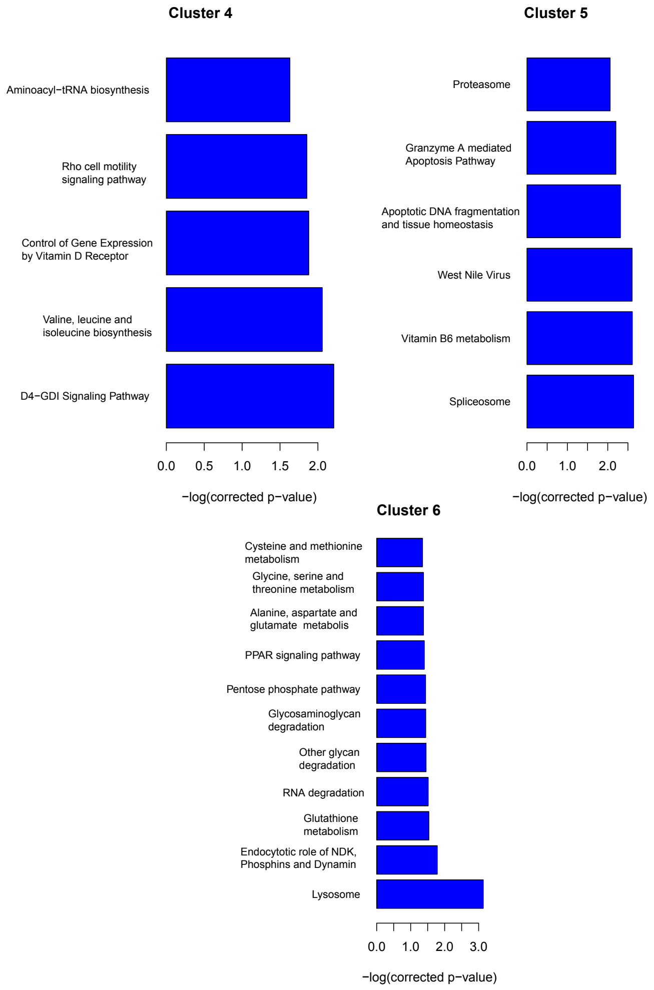


Figure A.1: KEGG enrichment for sorafenib-regulated proteins in clusters IV, V and VI.

A.5 Reproducibility in shotgun proteomics

A.5.1 Reproducibility of SILAC ratios

In the scientific literature there has always been a discussion on the reproducibility in shotgun proteomics. Issues on the reproducibility in shotgun proteomics have been systematically investigated by Tabb et al. (2010). Cultivation of mammalian cells produces the protein material that is used in many proteomics re-

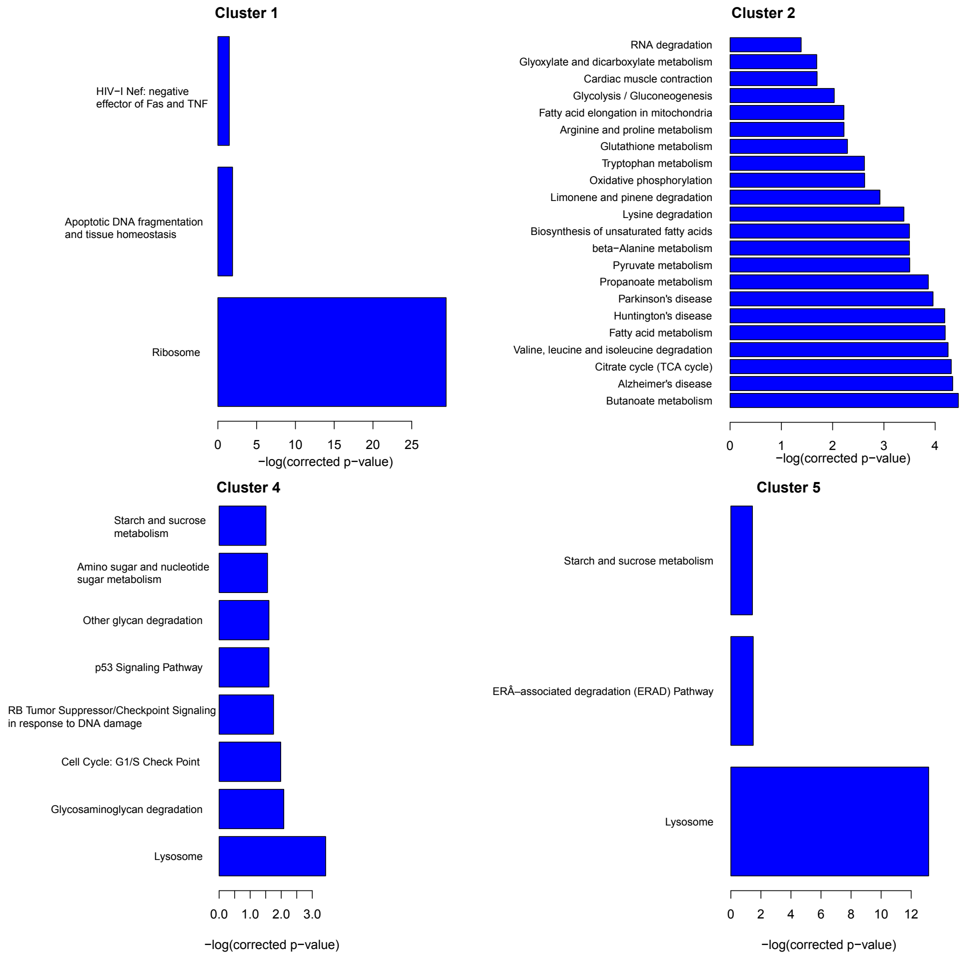


Figure A.2: KEGG enrichment for LY294002-regulated proteins in clusters I, II, IV and V.

search projects. The assess the biological reproducibility of the results obtained by quantitative shotgun proteomics, different analyses have to be performed. For each analysis new biological material is used. These independent assays can either be based on parallel or consecutive cell culture experiments.

Parallel cell culture

In a *parallel cultivation* all cells originate from the same split population. The cells that are subsequently used to evaluate the biological reproducibility of the experiment are grown at the same time in the same incubator. The cells are split at the same time.

Consecutive cell culture

In contrast to parallel cultivation, cells can be grown in *consecutive cultivation*. This means that the experiment is repeated at a different times and the initial cells are not split populations from the some culture dish. In the studies for this thesis, for both inhibitors two consecutive cell culture experiments were performed. The time gap between the two experiments is in the range of six months. Biological and technical reproducibility were investigated.

The following plots show the reproducibility of SILAC ratios. The SILAC ratios correspond to the quotients 12 hours DMSO / t_0 for the parallel cultivation, and 12 hours sorafenib / t_0 and 12 hours Ly294002 / t_0, respectively for the consecutive cultivation. t_0 denotes the time point zero without any treatment. Reproducibility of consecutive cultivation was investigated with 451Lu cells that were grown in April 2009 (run1) and in November 2009 (run2). Reproducibility of parallel cultivation is based on the cultivation from November 2009. The comparison of cell populations that were grown in parallel are shown in Fig. A.3. In run1 and run2 two different dishes of cells grown on M medium were treated with DMSO for 12 hours. Protein extracts from both cell populations were compared to cells grown on L medium in separate LC-MS runs. The red bars correspond to ratios from experiment one and the blue bars are ratios from experiment two. Green bars show proteins with common regulation in both experiments. Dark green shows significant regulation. Most of the proteins had common regulation in both experiments. Approximately 20 % of all proteins show different directions of expression. These proteins are shown in the middle of the block with red (run1) and blue (run2) colors. The rest of the proteins show the same direction of expression. In dark green the plot highlights proteins whose expression has significant p-values in either of the experiments and the same direction in both experiments. A contaminant protein was found as the only identified protein with significant regulation in different directions during the parallel experiments

(Tbl. A.2). All non contaminant proteins with significant regulation are regulated in the same direction.

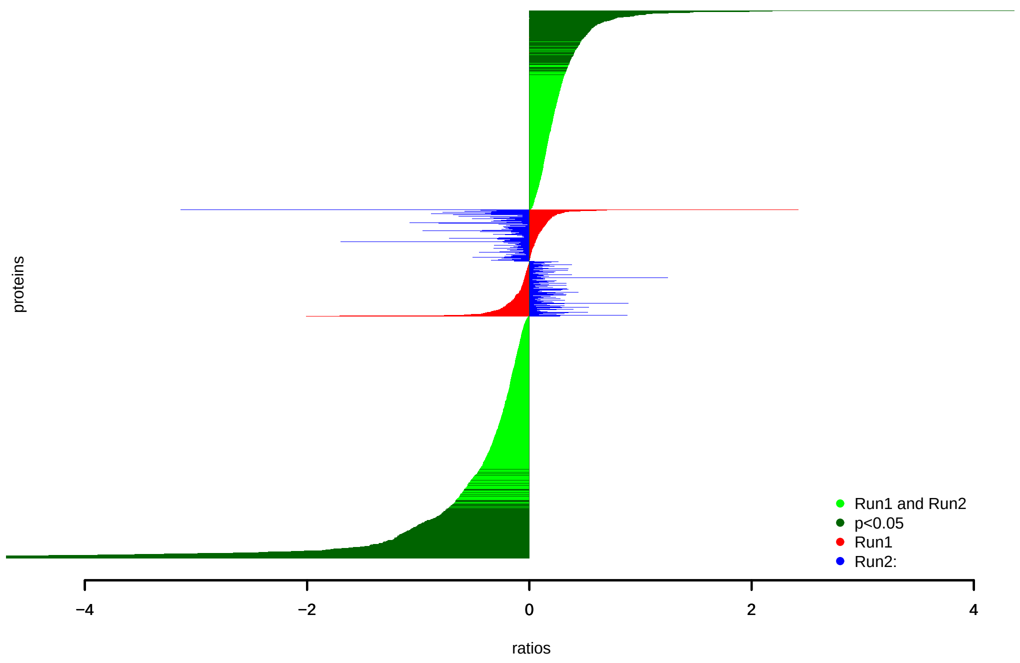


Figure A.3: Parallel culture and DMSO treatment. The x-axis corresponds to log_2 ratios of the two parallel experiments. The bars correspond to different proteins.

The comparison of consecutive LY294002 treatments is shown in Fig. A.4. For this comparison the cells were grown on M medium and treated with LY for 12 hours. The expression was compared to untreated cells grown on L medium. The number of proteins with different directions in their log scaled ratios is significantly increased compared to the DMSO treatment in parallel culture. The proteins that are significantly regulated in different directions can be found in Tbl. A.3. Interestingly 67 % of the proteins whose regulation is significant in dif-

Table A.2: Biological reproducibility of DMSO treatment in Parallel cultivation. One contaminant protein with significant regulation in opposite directions in two consecutive experiments.

	Protein names	log_2 ratio run1	log_2 ratio run2
1	Keratin, type II cytoskeletal 6A	2.42	-3.14

ferent direction, are mitochondrial proteins. The other 35 % cannot be grouped by common properties.

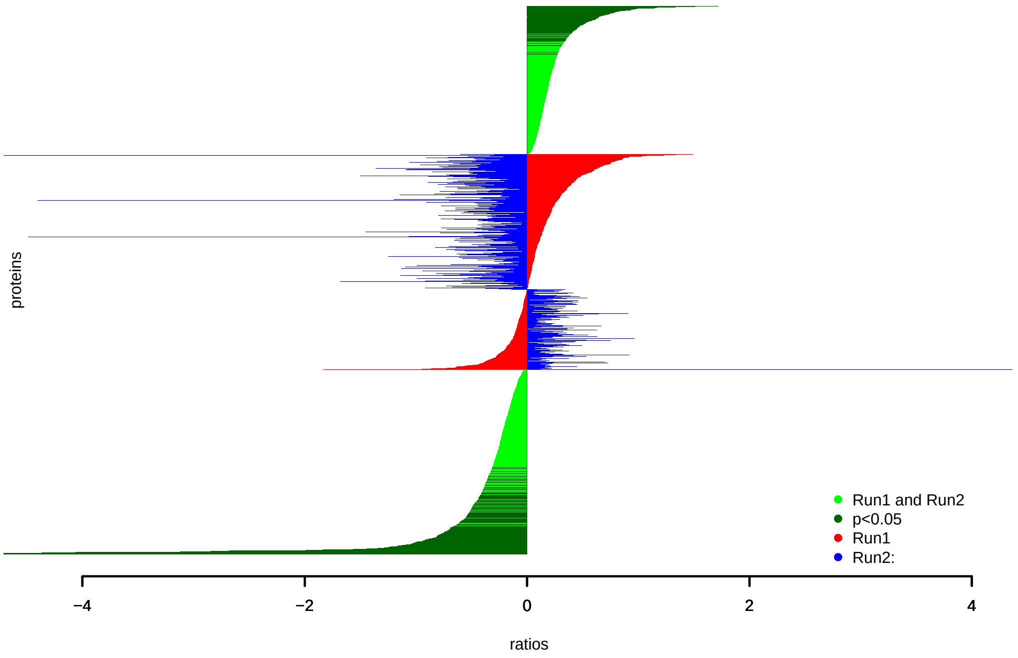


Figure A.4: Consecutive culture and Ly294002 treatment. Both runs from 12 h treatment. The red bars correspond to ratios from experiment one/ treatment of 12 h and the blue bars are ratios from experiment two/ treatment 12 h. Green bars show proteins with common regulation in both experiments. Dark green shows significant regulation.

Table A.3: Biological reproducibility of sorafenib treatment in consecutive cultivation. Proteins with significant regulation in opposite directions in two consecutive experiments.

IDs	Protein names	run 1	run 2
1	Serine hydroxymethyltransferase, mitochondrial	0.55	-0.49
2	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 9, mitochondrial	0.61	-1.08
3	Electron transfer flavoprotein subunit beta	0.57	-0.52
4	Lon protease homolog, mitochondrial	0.75	-0.67
5	Stress-70 protein, mitochondrial	0.47	-0.67
6	Sodium/potassium-transporting ATPase subunit beta-3	0.81	-0.7
7	Electron transfer flavoprotein subunit alpha, mitochondrial	0.72	-0.42
8	NAD-dependent malic enzyme, mitochondrial	0.68	-0.71
9	Cytochrome b-c1 complex subunit 1, mitochondrial	0.53	-0.52
10	Aconitase 2, mitochondrial	0.47	-0.58
11	28S ribosomal protein S29, mitochondrial	0.87	-0.91
12	Transmembrane protein 111	0.48	-0.89
13	Transmembrane emp24 domain-containing protein 9	0.38	-0.52
14	Putative RNA-binding protein 3	-0.38	0.7
15	NADH dehydrogenase [ubiquinone] iron-sulfur protein 2, mitochondrial	0.6	-1.1
16	Citrate synthase, mitochondrial	0.67	-0.79
17	Ceroid-lipofuscinosis neuronal protein 5	-0.6	0.45
18	Succinyl-CoA:3-ketoacid-coenzyme A transferase 1, mitochondrial	0.59	-0.85
19	Elongation factor Tu, mitochondrial	0.64	-0.61
20	cDNA FLJ56425, mitochondrial (EC 1.3.99.-)	0.77	-0.5
21	GrpE protein homolog 1, mitochondrial	0.75	-0.64
22	Heat shock protein 75 kDa, mitochondrial	0.61	-0.51
23	Trifunctional enzyme subunit alpha, mitochondrial	0.4	-0.72

Continued on next page

Table A.3 – continued from previous page

IDs	Protein names	run 1	run 2
24	28S ribosomal protein S35, mitochondrial	0.64	-0.76
25	Probable asparaginyl-tRNA synthetase, mitochondrial	0.75	-0.82
26	Aldehyde dehydrogenase X, mitochondrial	0.73	-0.6
27	Uncharacterized protein KIAA2013	0.48	-1.5
28	SDHA protein	0.44	-0.72
29	Delta-1-pyrroline-5-carboxylate dehydrogenase, mitochondrial	0.63	-1.36
30	Annexin A1	0.58	-0.57
31	Glutaminase kidney isoform, mitochondrial	0.36	-0.63
32	ATP synthase subunit beta, mitochondrial	0.65	-0.44
33	Mitochondrial import inner membrane translocase subunit TIM44	0.71	-0.96
34	cDNA FLJ56153, transcript variant 1	0.57	-0.79
35	Stomatin-like protein 2	0.89	-0.71
36	Mitochondrial carrier homolog 1	0.76	-1.06
37	Pyrroline-5-carboxylate reductase 2	0.61	-0.8
38	Transforming protein RhoA	0.4	-0.55
39	Pyrroline-5-carboxylate reductase	0.78	-0.81
40	Coiled-coil and C2 domain-containing protein 1B	-1.83	5.9
41	Leucine-rich PPR motif-containing protein, mitochondrial	0.48	-0.74
42	60 kDa heat shock protein, mitochondrial	0.49	-0.73

The results from the same analysis with sorafenib treated cells can be found in Fig. A.5. Two consecutive cultivation experiments were performed. 12 hours sorafenib treatment was compared to t_0. The percentage of proteins that show different signs in the log_2 ratios is comparable to the observations from the LY experiment (Fig. A.4), however the number of proteins with opposite significant regulation is below to one observed above. 20 % of the proteins with different regulation are again mitochondrial proteins and another 20 % are found to be ribosomal proteins. Another significant part of those proteins (25 %) are associated with the cytoskeleton.

Table A.4: Biological reproducibility of LY294002 treatment in consecutive cultivation. Proteins with significant regulation in opposite directions in two consecutive experiments.

	Protein names	log_2 ratio run1	log_2 ratio run2
1	Keratin, type I cytoskeletal 10	0.39	-5.41
2	Protein transport protein Sec61 subunit gamma	0.35	-0.86
3	PRA1 family protein 3	0.52	-0.45
4	28S ribosomal protein S21, mitochondrial	0.45	-0.82
5	Putative RNA-binding protein 3	-0.35	0.66
6	Lamina-associated polypeptide 2, isoforms beta/gamma	0.52	-1.11
7	Long-chain-fatty-acid–CoA ligase 3	0.27	-0.32
8	Growth arrest-specific protein 7	-0.55	0.31
9	Elongation factor G 2, mitochondrial	1.25	-3.72
10	Sarcoplasmic/endoplasmic reticulum calcium ATPase 2	0.35	-0.36
11	60S ribosomal protein L3-like	6.06	-4.99
12	Adipophilin	0.53	-1.36
13	28S ribosomal protein S31, mitochondrial	0.54	-1.44
14	DNA-dependent protein kinase catalytic subunit	0.26	-0.7
15	Heat shock 70 kDa protein 1	0.25	-0.43
16	Fibronectin type-III domain-containing protein 3a	-0.38	0.46
17	NADH-ubiquinone oxidoreductase 75 kDa subunit	0.59	-0.43
18	Coiled-coil and C2 domain-containing protein 1B	-1.79	5.8
19	Sterol O-acyltransferase 1	0.54	-0.57
20	60 kDa heat shock protein, mitochondrial	0.78	-0.29

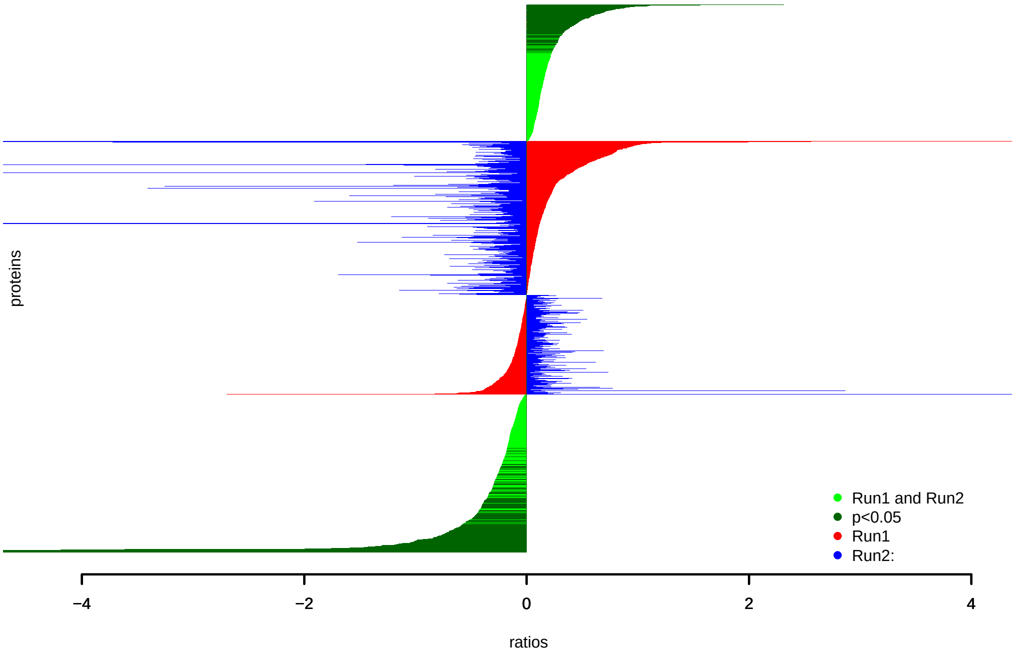


Figure A.5: Consecutive culture and sorafenib treatment. Both runs from 12 h treatment. The red bars correspond to ratios from experiment one/ treatment of 12 h and the blue bars are ratios from experiment two/ treatment 12 h. Green bars show proteins with common regulation in both experiments. Dark green shows significant regulation.

Appendix B

Curriculum vitae

Education

12/2006 - 12/2010
PhD

Eberhard-Karls University Tübingen, Joint PhD thesis, Proteome Center, Tübingen (Prof. Boris Macek/ Prof. Alfred Nordheim) and Center for Bioinformatics (Prof. Oliver Kohlbacher), Germany.

01/2006 - 11/2006
Master thesis

University of Cambridge, UK.

09/2003 - 09/2006
Diplôme d'Ingnieur en Biotechnologie

Ecole Supérieure de Biotechnologie de Strasbourg, France.

09/2001 - 09/2003
Vordiplom Biomathematics

University of Greifswald, Germany.

Scientific work experience

11/2009 to present **Team leader Orbitrap mass spectrometry**	**Eberhard-Karls University Tübingen**, Proteome Center, Tübingen, Germany.
01/2007 to 01/2010 **Teaching assistant** in course work in systems biology, systems immunology and bioinformatics II	**Eberhard-Karls University Tübingen**, Center for Bioinformatics, Germany.
01/2007 to 01/2010 **Teaching assistant** in course and lab work in proteomics	**Eberhard-Karls University Tübingen**, Proteome Center Tübingen, Germany.
10/2007 - present **Software development** as part of the OpenMS team	**Eberhard-Karls University Tübingen**, Center for Bioinformatics, Germany.
10/2005 - 11/2005 **Internship** student Bioinformatics	**IGBMC Strasbourg**, Bioinformatics Core Facility, France.
6/2005 - 7/2005 **Internship** student Molecular biology	**Tumorbiology Research Center**, Freiburg, Germany.
6/2004 - 7/2004 **Internship** student Biomathematics	**Politecnico di Torino**, Turin, Italy.

Publications

Peer reviewed papers

- K. Krug, S. Nahnsen and B. Macek. Mass spectrometry at the interface of proteomics and genomics. 2010. *Mol Biosyst. 2010 Oct 21. Epub ahead of print.*

- E. Schwarz, P. Whitfield, S. Nahnsen, L. Wang, H. Major, F. M. Leweke, D. Koethe, P. Lio, S. Bahn. Alterations of primary fatty acid amides in serum of patients with severe mental illness. 2010. *In press: Frontiers in Bioscience.*

- S. Nahnsen, A. Bertsch, J. Rahnenfürer, A. Nordheim and O. Kohlbacher. Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. 2010. *Submitted to J Proteome Res.: in revision.*

- T. Sinnberg, M. Menzel, S. Kaesler, T. Biedermann, B. Sauer, S. Nahnsen, M. Schwarz, C. Garbe and B. Schittek . Suppression of Casein Kinase 1α in Tumor Cells Induces a Switch in β-Catenin Signaling to Promote Metastasis. 2010. *Cancer Res. 2010 Sep 1;70(17):6999-7009. Epub 2010 Aug 10.*

- A. Bertsch, S. Jung, A. Zerck, N. Pfeifer, S. Nahnsen, C. Henneges, A. Nordheim and O. Kohlbacher. Optimal selection and scheduling of MRM transitions for rapid quantitation assay development. 2010. *J Proteome Res. 2010 May 7;9(5):2696-704.*

- S. Nahnsen, A. Nordheim and O. Kohlbacher. A geometric matching approach improves throughput and accurary in DIGE based proteomics. 2009. *In Proceedings of the 6th international workshop for computational systems biology.*

Manuscripts in preparation

- S. Nahnsen, S. Freiberger, T. Proikas-Cezanne, O. Kohlbacher, B. Macek, A. Nordheim. *Global protein expression dynamics in tumor cells following pharmacological intervention.*

Diploma thesis

- S. Nahnsen *High Throughput Metabolomics Profiling of Disease and Medication Effects in Schizophrenia.* Diploma thesis, Ecole Suprieure de biotechnology de Strasbourg, Strasbourg, France, 2006.

Poster Publications

- S. Nahnsen, O. Kohlbacher, A. Nordheim and B. Macek. Global protein expression dynamics in human tumor cells upon treatment with the *Raf* inhibitor sorafenib. In: *Proceedings of the 58th Conference of the ASMS, Salt Lake City 2010*, Poster abstract.

- S. Nahnsen, A. Bertsch, J. Rahnenfürer, A. Nordheim and O. Kohlbacher. Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. In: *Proceedings of the 57th Conference of the ASMS, Philadelphia 2009*, Poster abstract.

- S. Nahnsen, S.H. Rho, A. Bertsch, O. Kohlbacher, A. Nordheim. Integrative transcriptomic and proteomic profiling of tumor cells upon pharmacological intervention In: *Proceedings of the 9th International Conference of Systems Biology, Gothenborg 2008*, Poster abstract.

- E. Schwarz, S. Nahnsen, M. Leweke, D. Koethe, S. Gross, H. Major, S. Bahn. Metabolic profiling of serum from patients suffering from schizophrenia and affective disorder In: *Proceedings of the 55th Conference of the ASMS, Indianapolis 2007*, Poster abstract.

- S. Nahnsen, E. Schwarz, S. Bahn. Multivariate profiling of psychotropic drug action In: *Perspectives of Metabolomics and Proteomics Investigations in Clinical Science, Rome 2006*, Poster abstract.

- H. Major, T. McKenna, C. Hughes, J. Vissers, J. Huang, S. Nahnsen, S. Bahn, E. Schwarz. Metabolic and proteomic profiling of cerebrospinal fluid and serum for schizophrenia In: *Proceedings of the 54th Conference of the ASMS, Washington 2006*, Poster abstract.

Talks

- S. Nahnsen. A geometric matching approach improves throughput and accuracy in DIGE based proteomics. Presented: *6th international workshop for computational systems biology, Aarhus, Denmark, June 10-12, 2009.*

- S. Nahnsen. Proteomic Profiling of tumor cells upon pharmacological intervention. Presented at: *3rd graduate school network meeting, Rothenburg ob der Tauber,Germany, July 6 - 8, 2008.*

- S. Nahnsen. Quantitative Profiling of tumor cell HLA ligandome, proteome and phospho-proteome after pharmacological intervention. Presented at: *2nd Baden-Württemberg-Shanghai Workshop on Systems Biology and Biosystems Engineering, Lake Constance, Germany, April 22 - 24, 2007.*

References

M. Ashburner. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000. 45, 121

Amrita Basu, Kristie L Rose, Junmei Zhang, Ronald C Beavis, Beatrix Ueberheide, Benjamin A Garcia, Brian Chait, Yingming Zhao, Donald F Hunt, Eran Segal, C. David Allis, and Sandra B Hake. Proteome-wide prediction of acetylation substrates. *Proc Natl Acad Sci U S A*, 106(33):13785–13790, Aug 2009. 140

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57:289–300, 1995. 46, 116, 121

Halima Bensmail, Jennifer Golek, Michelle M Moody, John O Semmes, and Abdelali Haoudi. A novel approach for clustering proteomics data using bayesian fast fourier transform. *Bioinformatics*, 21(10):2210–2224, May 2005. 134

James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Kluwer Academic Publishers, Norwell, MA, USA, 1981. ISBN 0306406713. 44, 114

Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pags, Zlatko Trajanoski, and Jrme Galon. Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–1093, Apr 2009. 47, 116, 121

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition, October 2007. ISBN 0387310738. 77

M. M. Bradford. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem*, 72:248–254, May 1976. 60

Malorye A Branca. Multi-kinase inhibitors create buzz at asco. *Nat Biotechnol*, 23(6):639, Jun 2005. 2

G. J. Brunn, J. Williams, C. Sabers, G. Wiederrecht, J. C. Lawrence, and R. T. Abraham. Direct inhibition of the signaling functions of the mammalian target of rapamycin by the phosphoinositide 3-kinase inhibitors, wortmannin and ly294002. *EMBO J*, 15(19):5256–5267, Oct 1996. 15

F. Chang and J. A. McCubrey. P21(cip1) induced by raf is associated with increased cdk4 activity in hematopoietic cells. *Oncogene*, 20(32):4354–4364, Jul 2001. 136

Chris Cheadle, Marquis P Vawter, William J Freed, and Kevin G Becker. Analysis of microarray data using z score transformation. *J Mol Diagn*, 5(2):73–81, May 2003. 44

Yu-Chi Chen, Gudrun Pohl, Tian-Li Wang, Patrice J Morin, Bjrn Risberg, Gunnar B Kristensen, Albert Yu, Ben Davidson, and Ie-Ming Shih. Apolipoprotein e is required for cell proliferation and survival in ovarian cancer. *Cancer Res*, 65(1):331–337, Jan 2005. 137

Jeng-Fong Chiou, Cheng-Jeng Tai, Yu-Huei Wang, Tsan-Zon Liu, Yee-Min Jen, and Chia-Yang Shiau. Sorafenib induces preferential apoptotic killing of a drug- and radio-resistant hep g2 cells through a mitochondria-dependent oxidative stress mechanism. *Cancer Biol Ther*, 8(20):1904–1913, Oct 2009. 135

Hyungwon Choi and Alexey I Nesvizhskii. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J Proteome Res*, 7(1):254–265, Jan 2008. 72, 79

Jacques Colinge and Keiryn L Bennett. Introduction to computational proteomics. *PLoS Comput Biol*, 3(7):e114, Jul 2007. 35

Jacques Colinge, Alexandre Masselot, Marc Giron, Thierry Dessingy, and Jrme Magnin. Olav: towards high-throughput tandem mass spectrometry data identification. *Proteomics*, 3(8):1454–1463, Aug 2003. 30

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 2001. 51

Jürgen Cox and Matthias Mann. Is proteomics the new genomics? *Cell*, 130(3): 395–398, Aug 2007. 3

Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, 26(12):1367–1372, Dec 2008. 42, 43, 108, 111, 148, 149

Robertson Craig and Ronald C Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, Jun 2004. 30, 34

M. B. Datto, P. P. Hu, T. F. Kowalik, J. Yingling, and X. F. Wang. The viral oncoprotein e1a blocks transforming growth factor beta-mediated induction of p21/waf1/cip1 and p15/ink4b. *Mol Cell Biol*, 17(4):2030–2037, Apr 1997. 11

S. P. Davies, H. Reddy, M. Caivano, and P. Cohen. Specificity and mechanism of action of some commonly used protein kinase inhibitors. *Biochem J*, 351(Pt 1):95–105, Oct 2000. 15

Margaret O Dayhoff, Robert M Schwartz, and Bruce C Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5 Suppl. 3:345352, 1978. 145

Lyris M F de Godoy, Jesper V Olsen, Jürgen Cox, Michael L Nielsen, Nina C Hubner, Florian Fröhlich, Tobias C Walther, and Matthias Mann. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217):1251–1254, Oct 2008. 3

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977. 72, 77

Peter A DiMaggio and Christodoulos A Floudas. De novo peptide identification via tandem mass spectrometry and integer linear optimization. *Anal Chem*, 79 (4):1433–1446, Feb 2007. 29

Anthone W Dunah, Ana C Sirianni, Allen A Fienberg, Elena Bastia, Michael A Schwarzschild, and David G Standaert. Dopamine d1-dependent trafficking of striatal n-methyl-d-aspartate glutamate receptors requires fyn protein tyrosine kinase but not darpp-32. *Mol Pharmacol*, 65(1):121–129, Jan 2004. 135

Nathan J Edwards, , Xue Wu, and Chau-Wen Tseng. An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra. *Clinical Proteomics*, 5:23–36, 2009. 72

Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–214, Mar 2007. 37

Jimmy K. Eng, Ashley L. McCormack, and John R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am SW Mass Spectrometry*, 5:976–989, 1994. 30

Ellen Mosleth Faergestad, Morten Rye, Beata Walczak, Lars Gidskehaug, Jens Petter Wold, Harald Grove, Xiaohong Jia, Kristin Hollung, Ulf G Indahl, Frank Westad, Frans van den Berg, and Harald Martens. Pixel-based analysis of multiple images for the identification of changes: a novel approach applied to unravel proteome patterns [corrected] of 2-d electrophoresis gel images. *Proteomics*, 7(19):3450–3461, Oct 2007. 38, 50

Tom Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8): 861–874, 2006. 92

David Fenyo and Ronald C Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem*, 75(4):768–774, Feb 2003. 31, 79

Hunter B Fraser and Joshua B Plotkin. Using protein complexes to predict phenotypic effects of gene mutation. *Genome Biol*, 8(11):R252, 2007. 3

M.H. Fulekar. *Bioinformatics: Applications in Life and Environmental Sciences*. Springer-Verlag, 2009. 145

C. Gauss, M. Kalkum, M. Lwe, H. Lehrach, and J. Klose. Analysis of the mouse proteome. (i) brain proteins: separation by two-dimensional electrophoresis and identification by mass spectrometry and genetic variation. *Electrophoresis*, 20 (3):575–600, Mar 1999. 19

Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open mass spectrometry search algorithm. *J Proteome Res*, 3(5):958–964, 2004. 30, 35, 36

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004. 50

Scott A Gerber, John Rush, Olaf Stemman, Marc W Kirschner, and Steven P Gygi. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem ms. *Proc Natl Acad Sci U S A*, 100(12):6940–6945, Jun 2003. 41

Severine I Gharbi, Marketa J Zvelebil, Stephen J Shuttleworth, Tim Hancox, Nahid Saghir, John F Timms, and Michael D Waterfield. Exploring the specificity of the pi3k family inhibitor ly294002. *Biochem J*, 404(1):15–21, May 2007. 15

Kamran Ghoreschi, Arian Laurence, and John J O'Shea. Selectivity and therapeutic inhibition of kinases: to be or not to be? *Nat Immunol*, 10(4):356–360, Apr 2009. 12

Hidemasa Goto, Kazushi Tanabe, Edward Manser, Louis Lim, Yoshihiro Yasui, and Masaki Inagaki. Phosphorylation and reorganization of vimentin by p21-activated kinase (pak). *Genes Cells*, 7(2):91–97, Feb 2002. 69

Jonas Grossmann, Bernd Roschitzki, Christian Panse, Claudia Fortes, Simon Barkow-Oesterreicher, Dorothea Rutishauser, and Ralph Schlapbach. Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. *J Proteomics*, 73(9):1740–1746, Aug 2010. 42

S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, 17(10):994–999, Oct 1999. 41

D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000. 9, 10, 11

D. Herlyn, D. Iliopoulos, P. J. Jensen, A. Parmiter, J. Baird, H. Hotta, K. Adachi, A. H. Ross, J. Jambrosic, and H. Koprowski. In vitro properties of human melanoma cells metastatic in nude mice. *Cancer Res*, 50(8):2296–2302, Apr 1990. 104

Qizhi Hu, Robert J Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R. Graham Cooks. The orbitrap: a new mass spectrometer. *J Mass Spectrom*, 40(4):430–443, Apr 2005. 26

L. Huang, R. J. Jacob, S. C. Pegg, M. A. Baldwin, C. C. Wang, A. L. Burlingame, and P. C. Babbitt. Functional assignment of the 20 s proteasome from trypanosoma brucei using mass spectrometry and new bioinformatics approaches. *J Biol Chem*, 276(30):28327–28339, Jul 2001. 83

Nina C Hubner, Shubin Ren, and Matthias Mann. Peptide separation with immobilized pi strips is an attractive alternative to in-gel protein digestion for proteome analysis. *Proteomics*, 8(23-24):4862–4872, Dec 2008. 105, 133

Daehee Hwang, Inyoul Y Lee, Hyuntae Yoo, Nils Gehlenborg, Ji-Hoon Cho, Brianne Petritis, David Baxter, Rose Pitstick, Rebecca Young, Doug Spicer, Nathan D Price, John G Hohmann, Stephen J Dearmond, George A Carlson, and Leroy E Hood. A systems approach to prion disease. *Mol Syst Biol*, 5:252, 2009. 2

P. James. Protein identification in the post-genome era: the rapid rise of proteomics. *Q Rev Biophys*, 30(4):279–331, Nov 1997. 4

Ole N Jensen. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol*, 7(6):391–403, Jun 2006. 3

Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res*, 7(1):40–44, Jan 2008a. 37, 72

Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*, 7(1):29–34, Jan 2008b. 85

M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000. 45, 121

Stanley P. Kaplan. *Drug Design Research Perspectives*. Nova Science Publishers, Inc., 2007. 142

Eugene A Kapp, Frédéric Schütz, Lisa M Connolly, John A Chakel, Jose E Meza, Christine A Miller, David Fenyo, Jimmy K Eng, Joshua N Adkins, Gilbert S Omenn, and Richard J Simpson. An evaluation, comparison, and accurate benchmarking of several publicly available ms/ms search algorithms: sensitivity and specificity analysis. *Proteomics*, 5(13):3475–3490, Aug 2005. 71, 72

Mazen W Karaman, Sanna Herrgard, Daniel K Treiber, Paul Gallant, Corey E Atteridge, Brian T Campbell, Katrina W Chan, Pietro Ciceri, Mindy I Davis, Philip T Edeen, Raffaella Faraoni, Mark Floyd, Jeremy P Hunt, Daniel J Lockhart, Zdravko V Milanov, Michael J Morrison, Gabriel Pallares, Hitesh K Patel, Stephanie Pritchard, Lisa M Wodicka, and Patrick P Zarrinkar. A quantitative analysis of kinase inhibitor selectivity. *Nat Biotechnol*, 26(1):127–132, Jan 2008. 12

Michael Karas and Franz Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60(20): 2299–2301, 1988. 22

Andrew Keller, Alexey I Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal Chem*, 74(20):5383–5392, Oct 2002a. 71, 73

Andrew Keller, Samuel Purvine, Alexey I Nesvizhskii, Sergey Stolyar, David R Goodlett, and Eugene Kolker. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS*, 6(2):207–212, 2002b. 84

Paul J Kersey, Jorge Duarte, Allyson Williams, Youla Karavidopoulou, Ewan Birney, and Rolf Apweiler. The international protein index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985–1988, Jul 2004. 21

Young Ho Kim, Kwang-Hae Choi, Jong-Wook Park, and Taeg Kyu Kwon. Ly294002 inhibits lps-induced no production through a inhibition of nf-kappab activation: independent mechanism of phosphatidylinositol 3-kinase. *Immunol Lett*, 99(1):45–50, Jun 2005. 15

John Klimek, James S. Eddes, Laura Hohmann, Jennifer Jackson, Amelia Peterson, Simon Letarte, Philip R. Gafken, Jonathan E Katz, Parag Mallick, Hookeun Lee, Alexander Schmidt, Reto Ossola, Jimmy K. Eng, Ruedi Aebersold, and Daniel B Martin. The standard protein mix database: A diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.*, ,7:96 – 103, 2008. 84, 86

J. Klose. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. a novel approach to testing for induced point mutations in mammals. *Humangenetik*, 26(3):231–243, 1975. 4, 17, 19

R D Knight. Storage of ions from laser-produced plasmas. *Appl Phys Lett*, 38: 221–222, 1981. 26

Oliver Kohlbacher, Knut Reinert, Clemens Gröpl, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, and Marc Sturm. Topp–the openms proteomics pipeline. *Bioinformatics*, 23(2):e191–e197, Jan 2007. 43, 86

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics*, 52:7–21, 2005. 53

Chanchal Kumar and Matthias Mann. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett*, 583(11):1703–1712, Jun 2009. 42

U. K. Laemmli. Cleavage of structural proteins during the assembly of the head of bacteriophage t4. *Nature*, 227(5259):680–685, Aug 1970. 19

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis,

L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001. 2, 101

Konstantinos G Lasithiotakis, Tobias W Sinnberg, Birgit Schittek, Keith T Flaherty, Dagmar Kulms, Evelyn Maczey, Claus Garbe, and Friedegund E Meier. Combined inhibition of mapk and mtor signaling inhibits growth, induces cell death, and abrogates invasive growth of melanoma cells. *J Invest Dermatol*, 128(8):2013–2023, Aug 2008. 136

Friedrich Leisch. Flexmix: A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software*, 11:1–18, 2004. 77, 80

LTQ XL Hardware LTQ-Manual. *LTQ XL Hardware Manual*, 2006. 26

Makarov. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem*, 72(6):1156–1162, Mar 2000. 27

G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, Dec 2002. 12

Lei Mao, Claus Zabel, Marion Herrmann, Tobias Nolden, Florian Mertes, Laetitia Magnol, Caroline Chabert, Daniela Hartl, Yann Herault, Jean Maurice Delabar, Thomas Manke, Heinz Himmelbauer, and Joachim Klose. Proteomic shifts in embryonic stem cells with gene dose modifications suggest the presence of balancer proteins in protein regulatory networks. *PLoS One*, 2(11):e1218, 2007. 138

Edward M Marcotte. How do shotgun proteomics algorithms identify proteins? *Nat Biotechnol*, 25(7):755–757, Jul 2007. 17

Fred W McLafferty, Kathrin Breuker, Mi Jin, Xuemei Han, Giuseppe Infusini, Honghai Jiang, Xianglei Kong, and Tadhg P Begley. Top-down ms, a powerful complement to the high capabilities of proteolysis proteomics. *FEBS J*, 274 (24):6256–6268, Dec 2007. 17

Bruno Meunier, Emilie Dumas, Isabelle Piec, Daniel Bchet, Michel Hbraud, and Jean-Franois Hocquette. Assessment of hierarchical clustering methodologies for proteomic data mining. *J Proteome Res*, 6(1):358–366, Jan 2007. 134

Henrik Molina, Yi Yang, Travis Ruch, Jae-Woo Kim, Peter Mortensen, Tamara Otto, Anuradha Nalli, Qi-Qun Tang, M. Daniel Lane, Raghothama Chaerkady, and Akhilesh Pandey. Temporal profiling of the adipocyte proteome during differentiation using a five-plex silac based strategy. *J Proteome Res*, 8(1): 48–58, Jan 2009. 40

Lucia Monteoliva and Juan Pablo Albar. Differential proteomics: an overview of gel and non-gel based approaches. *Brief Funct Genomic Proteomic*, 3(3): 220–239, Nov 2004. 4

K. Moriyama, K. Iida, and I. Yahara. Phosphorylation of ser-3 of cofilin regulates its essential function on actin. *Genes Cells*, 1(1):73–86, Jan 1996. 69

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5:32–38, 1957. 53

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3): 443–453, Mar 1970. 81

Alexey I Nesvizhskii, Olga Vitek, and Ruedi Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods*, 4 (10):787–797, Oct 2007. 71, 75

M. Nistér, C. H. Heldin, and B. Westermark. Clonal variation in the production of a platelet-derived growth factor-like protein and expression of corresponding receptors in a human malignant glioma. *Cancer Res*, 46(1):332–340, Jan 1986. 10

P. H. O'Farrell. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*, 250(10):4007–4021, May 1975. 4, 17, 19

Jesper V Olsen, Lyris M F de Godoy, Guoqing Li, Boris Macek, Peter Mortensen, Reinhold Pesch, Alexander Makarov, Oliver Lange, Stevan Horning, and

Matthias Mann. Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics*, 4(12): 2010–2021, Dec 2005. 107

Jesper V Olsen, Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3):635–648, Nov 2006. 44, 140

Shao-En Ong, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1(5):376–386, May 2002. 39

D. J. Pappin, P. Hojrup, and A. J. Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol*, 3(6):327–332, Jun 1993. 34

Junmin Peng, Daniel Schwartz, Joshua E Elias, Carson C Thoreen, Dongmei Cheng, Gerald Marsischky, Jeroen Roelofs, Daniel Finley, and Steven P Gygi. A proteomics approach to understanding protein ubiquitination. *Nat Biotechnol*, 21(8):921–926, Aug 2003. 140

Sabine Peres, Laurence Molina, Nicolas Salvetat, Claude Granier, and Franck Molina. A new method for 2d gel spot alignment: application to the analysis of large sample sets in clinical proteomics. *BMC Bioinformatics*, 9:460, 2008. 51

David N. Perkins, Darryl J. C. Pappin, David M. Creasy, and John S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551 – 3567, 1999. 30, 34

Richard H Perry, R. Graham Cooks, and Robert J Noll. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom Rev*, 27(6):661–699, 2008. 27

Julia Poland, Thierry Rabilloud, and Pranav Sinha. Silver staining of 2-d gels. In John M. Walker, editor, *The Proteomics Protocols Handbook*, pages 177–184. Humana Press, 2005. 10.1385/1-59259-890-0:177. 62

Tassula Proikas-Cezanne, Scott Waddell, Anja Gaugel, Tancred Frickey, Andrei Lupas, and Alfred Nordheim. Wipi-1alpha (wipi49), a member of the novel 7-bladed wipi protein family, is aberrantly expressed in human cancer and is linked to starvation-induced autophagy. *Oncogene*, 23(58):9314–9325, Dec 2004. 134

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0. 77

Juri Rappsilber, Yasushi Ishihama, and Matthias Mann. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and lc/ms sample pretreatment in proteomics. *Anal Chem*, 75(3):663–670, Feb 2003. 106

Alexandra Resch, Stefan Leicht, Marc Saric, Linda Psztor, Andreas Jakob, Friedrich Götz, and Alfred Nordheim. Comparative proteome analysis of staphylococcus aureus biofilm and planktonic cells and correlation with transcriptome profiling. *Proteomics*, 6(6):1867–1877, Mar 2006. 63

Monica Riley, Takashi Abe, Martha B Arnaud, Mary K B Berlyn, Frederick R Blattner, Roy R Chaudhuri, Jeremy D Glasner, Takashi Horiuchi, Ingrid M Keseler, Takehide Kosuge, Hirotada Mori, Nicole T Perna, Guy Plunkett, Kenneth E Rudd, Margrethe H Serres, Gavin H Thomas, Nicholas R Thomson, David Wishart, and Barry L Wanner. Escherichia coli k-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Res*, 34(1):1–9, 2006. 86

Isabelle Rivals, Lon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, Feb 2007. 46

E. Roche, F. Assimacopoulos-Jeannet, L. A. Witters, B. Perruchoud, G. Yaney, B. Corkey, M. Asfari, and M. Prentki. Induction by glucose of genes coding for glycolytic enzymes in a pancreatic beta-cell line (ins-1). *J Biol Chem*, 272(5): 3091–3098, Jan 1997. 137

Philip L Ross, Yulin N Huang, Jason N Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, Sasi Pillai, Subhakar Dey, Scott Daniels, Subhasish Purkayastha, Peter Juhasz, Stephen Martin, Michael Bartlet-Jones, Feng He, Allan Jacobson, and Darryl J Pappin. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*, 3(12):1154–1169, Dec 2004. 41

Rovshan G Sadygov and John R Yates. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem*, 75(15):3792–3798, Aug 2003. 35

Susanne Schneiker, Olena Perlova, Olaf Kaiser, Klaus Gerth, Aysel Alici, Matthias O Altmeyer, Daniela Bartels, Thomas Bekel, Stefan Beyer, Edna Bode, Helge B Bode, Christoph J Bolten, Jomuna V Choudhuri, Sabrina Doss, Yasser A Elnakady, Bettina Frank, Lars Gaigalat, Alexander Goesmann, Carolin Groeger, Frank Gross, Lars Jelsbak, Lotte Jelsbak, Jrn Kalinowski, Carsten Kegler, Tina Knauber, Sebastian Konietzny, Maren Kopp, Lutz Krause, Daniel Krug, Bukhard Linke, Taifo Mahmud, Rosa Martinez-Arias, Alice C McHardy, Michelle Merai, Folker Meyer, Sascha Mormann, Jose Munoz-Dorado, Juana Perez, Silke Pradella, Shwan Rachid, Günter Raddatz, Frank Rosenau, Christian Rückert, Florenz Sasse, Maren Scharfe, Stephan C Schuster, Garret Suen, Anke Treuner-Lange, Gregory J Velicer, Frank-Jörg Vorhölter, Kira J Weissman, Roy D Welch, Silke C Wenzel, David E Whitworth, Susanne Wilhelm, Christoph Wittmann, Helmut Blöcker, Alfred Pühler, and Rolf Müller. Complete genome sequence of the myxobacterium sorangium cellulosum. *Nat Biotechnol*, 25(11):1281–1289, Nov 2007. 86

Ole Schulz-Trieglaff, Rene Hussong, Clemens Gröpl, Andreas Hildebrandt, and Knut Reinert. A fast and accurate algorithm for the quantification of peptides from mass spectrometry data. In *Proceedings of the 11th annual in-*

ternational conference on Research in computational molecular biology, RECOMB'07, pages 473–487, Berlin, Heidelberg, 2007. Springer-Verlag. 42

Veit Schwämmle and Ole Nørregaard Jensen. A simple and fast method to determine the parameters for fuzzy c-means cluster analysis. *Bioinformatics*, 26 (22):2841–2848, Nov 2010. 134

Brian C Searle, Mark Turner, and Alexey I Nesvizhskii. Improving sensitivity by probabilistically combining results from multiple ms/ms search methodologies. *J Proteome Res*, 7(1):245–253, Jan 2008. 72, 79, 97

Ian P Shadforth, Tom P J Dunkley, Kathryn S Lilley, and Conrad Bessant. i-tracker: for quantitative proteomics using itraq. *BMC Genomics*, 6:145, 2005. 41

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003. 47, 50, 116, 121

D. Shteynberg, E. Deutsch, H. Lam, R. Aebersold, and A. Nesvizhskii. iprophet: Improved validation of peptide identification in shotgun proteomics. In *HUPO World Congress, Amsterdam, The Netherlands.*, 2008. 73

M. Sirava, T. Schfer, M. Eiglsperger, M. Kaufmann, O. Kohlbacher, E. Bornberg-Bauer, and H. P. Lenhof. Biominer–modeling, analyzing, and visualizing biochemical pathways and networks. *Bioinformatics*, 18 Suppl 2:S219–S230, 2002. 50

D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire. Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene. *Science*, 235(4785):177–182, Jan 1987. 10

B. Snel, G. Lehmann, P. Bork, and M. A. Huynen. String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res*, 28(18):3442–3444, Sep 2000. 63, 65

Hanno Steen and Matthias Mann. The abc's (and xyz's) of peptide sequencing. *Nat Rev Mol Cell Biol*, 5(9):699–711, Sep 2004. 81

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, Aug 2003. 37

Marc Sturm, Andreas Bertsch, Clemens Grpl, Andreas Hildebrandt, Rene Hussong, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, Alexandra Zerck, Knut Reinert, and Oliver Kohlbacher. Openms - an open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9:163, 2008. 43

David L Tabb, Lorenzo Vega-Montoto, Paul A Rudnick, Asokan Mulayath Variyath, Amy-Joan L Ham, David M Bunk, Lisa E Kilpatrick, Dean D Billheimer, Ronald K Blackman, Helene L Cardasis, Steven A Carr, Karl R Clauser, Jacob D Jaffe, Kevin A Kowalski, Thomas A Neubert, Fred E Regnier, Birgit Schilling, Tony J Tegeler, Mu Wang, Pei Wang, Jeffrey R Whiteaker, Lisa J Zimmerman, Susan J Fisher, Bradford W Gibson, Christopher R Kinsinger, Mehdi Mesri, Henry Rodriguez, Stephen E Stein, Paul Tempst, Amanda G Paulovich, Daniel C Liebler, and Cliff Spiegelman. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res*, 9(2):761–776, Feb 2010. 150

Koichi Tanaka, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, and Tamio Yoshida. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 2:151–153, 1988. 22

Isei Tanida, Naoko Minematsu-Ikeguchi, Takashi Ueno, and Eiki Kominami. Lysosomal turnover, but not a cellular level, of endogenous lc3 is a marker for autophagy. *Autophagy*, 1(2):84–91, Jul 2005. 135

Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X. Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329 (5991):533–538, Jul 2010. 3

Stephen Tanner, Hongjun Shu, Ari Frank, Ling-Chi Wang, Ebrahim Zandi, Marc Mumby, Pavel A Pevzner, and Vineet Bafna. Inspect: identification of post-translationally modified peptides from tandem mass spectra. *Anal Chem*, 77 (14):4626–4639, Jul 2005. 30

Ravi Tharakan, Nathan Edwards, and David R M Graham. Data maximization by multipass analysis of protein mass spectra. *Proteomics*, 10(6):1160–1171, Mar 2010. 73

Barbara Tolloczko, Petra Turkewitsch, Mustafa Al-Chalabi, and James G Martin. Ly-294002 [2-(4-morpholinyl)-8-phenyl-4h-1-benzopyran-4-one] affects calcium signaling in airway smooth muscle cells independently of phosphoinositide 3-kinase inhibition. *J Pharmacol Exp Ther*, 311(2):787–793, Nov 2004. 15

Mike Tyers and Matthias Mann. From genomics to proteomics. *Nature*, 422 (6928):193–197, Mar 2003. 4

E. Tzima, P. J. Trotter, M. A. Orchard, and J. H. Walker. Annexin v relocates to the platelet cytoskeleton upon activation and binds to a specific isoform of actin. *Eur J Biochem*, 267(15):4720–4730, Aug 2000. 65

Anders Ullen, Marianne Farnebo, Lena Thyrell, Salah Mahmoudi, Pedram Kharaziha, Lena Lennartsson, Dan Grandr, Theoharis Panaretakis, and Sten Nilsson. Sorafenib induces apoptosis and autophagy in prostate cancer cells in vitro. *Int J Oncol*, 37(1):15–20, Jul 2010. 135

M. Unlü, M. E. Morgan, and J. S. Minden. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, 18(11): 2071–2077, Oct 1997. 38

J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman,

M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen,

M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001. 2, 101

C. J. Vlahos, W. F. Matter, K. Y. Hui, and R. F. Brown. A specific inhibitor of phosphatidylinositol 3-kinase, 2-(4-morpholinyl)-8-phenyl-4h-1-benzopyran-4-one (ly294002). *J Biol Chem*, 269(7):5241–5248, Feb 1994. 15

R. A. Weinberg. The retinoblastoma protein and cell cycle control. *Cell*, 81(3): 323–330, May 1995. 11

Robert A. Weinberg. *The Biology of Cancer*. Garland Science, 2007. 12

Elaine Welsh, Marina Jirotka, and David Gavaghan. Post-genomic science: cross-disciplinary and large-scale collaborative research and its organizational and technological challenges for the scientific research process. *Philos Transact A Math Phys Eng Sci*, 364(1843):1533–1549, Jun 2006. 2

C. M. Whitehouse, R. N. Dreyer, M. Yamashita, and J. B. Fenn. Electrospray interface for liquid chromatographs and mass spectrometers. *Anal Chem*, 57 (3):675–679, Mar 1985. 22

Scott M Wilhelm, Christopher Carter, Liya Tang, Dean Wilkie, Angela McNabola, Hong Rong, Charles Chen, Xiaomei Zhang, Patrick Vincent, Mark McHugh, Yichen Cao, Jaleel Shujath, Susan Gawlak, Deepa Eveleigh, Bruce Rowley, Li Liu, Lila Adnane, Mark Lynch, Daniel Auclair, Ian Taylor, Rich Gedrich, Andrei Voznesensky, Bernd Riedl, Leonard E Post, Gideon Bollag, and Pamela A Trail. Bay 43-9006 exhibits broad spectrum oral antitumor activity and targets the raf/mek/erk pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis. *Cancer Res*, 64(19):7099–7109, Oct 2004. 14

David A Wolk, Bradford C Dickerson, Alzheimer's Disease Neuroimaging Initiative, Michael Weiner, Marilyn Aiello, Paul Aisen, Marilyn S Albert, Gene

Alexander, Heather S Anderson, Karen Anderson, Liana Apostolova, Steve Arnold, Wes Ashford, Michele Assaly, Sanjay Asthana, Dan Bandy, Rob Bartha, Vernice Bates, Laurel Beckett, Karen L Bell, Amanda L Benincasa, Howard Bergman, Charles Bernick, Matthew Bernstein, Sandra Black, Karen Blank, Michael Borrie, Connie Brand, James Brewer, Alice D Brown, Jeffrey M Burns, Nigel J Cairns, Curtis Caldwell, Horacio Capote, Cynthia M Carlsson, Owen Carmichael, Janet S Cellar, Dzintra Celmins, Kewei Chen, Howard Chertkow, Munir Chowdhury, David Clark, Donald Connor, Stephen Correia, Karen Crawford, Anders Dale, Mony J de Leon, Susan M De Santi, Charles Decarli, Leyla Detoledo-Morrell, Michael Devous, Ramon Diaz-Arrastia, Sara Dolen, Michael Donohue, Rachelle S Doody, P. Murali Doraiswamy, Ranjan Duara, Jessica Englert, Martin Farlow, Howard Feldman, Joel Felmlee, Adam Fleisher, Evan Fletcher, Tatiana M Foroud, Norm Foster, Nick Fox, Richard Frank, Anthony Gamst, Curtis A Given, Neill R Graff-Radford, Robert C Green, Randall Griffith, Hillel Grossman, Ann M Hake, Peter Hardy, Danielle Harvey, Judith L Heidebrink, Barry A Hendin, Scott Herring, Lawrence S Honig, Chris Hosein, Ging-Yuek Robin Hsiung, Leon Hudson, M. Saleem Ismail, Clifford R Jack, Sandra Jacobson, William Jagust, Annapurni Jayam-Trouth, Kris Johnson, Heather Johnson, Nancy Johnson, Kathleen Johnson, Keith A Johnson, Sterling Johnson, Zaven Kachaturian, Jason H Karlawish, Maria Kataki, Jeffrey Kaye, Andrew Kertesz, Ronald Killiany, Smita Kittur, Robert A Koeppe, Magdalena Korecka, John Kornak, Nicholas Kozauer, James J Lah, Mary M Laubinger, Virginia M-Y Lee, T-Y. Lee, Alan Lerner, Allan I Levey, Crystal Flynn Longmire, Oscar L Lopez, Joanne L Lord, Po H Lu, Martha G Macavoy, Paul Malloy, Daniel Marson, Kristen Martin-Cook, Walter Martinez, George Marzloff, Chet Mathis, Catherine Mc-Adams-Ortiz, Marsel Mesulam, Bruce L Miller, Mark A Mintun, Jacobo Mintzer, Susan Molchan, Tom Montine, John Morris, Ruth A Mulnard, Donna Munic, Anil Nair, Scott Neu, Dana Nguyen, Alexander Norbash, Maryann Oakley, Thomas O Obisesan, Paula Ogrocki, Brian R Ott, Francine Parfitt, Sonia Pawluczyk, Godfrey Pearlson, Ronald Petersen, Jeffrey R Petrella, Steven Potkin, William Z Potter, Adrian Preda, Joseph Quinn, Michelle Rainka, Stephanie Reeder, Eric M Reiman, Dorene M Rentz, Brigid Reynolds, Jennifer Richard, Peggy Roberts, John Rogers, Allyson Rosen, Howard J Rosen, Henry Rusinek, Marwan Sabbagh, Carl Sadowsky, Stephen Salloway, Robert B Santulli, Andrew J Saykin, Douglas W Scharre, Lon Schneider, Stacy Schneider, Norbert Schuff, Raj C Shah, Les Shaw, Li Shen, Daniel H S Silverman, Donna M Simpson, Kaycee M Sink, Charles D Smith, Peter J Snyder, Bryan M Spann, Reisa A Sperling, Kenneth Spicer, Bojana Stefanovic, Yaakov Stern, Edward Stopa, Cheuk Tang, Pierre Tariot, Lisa Taylor-Reinwald, Gaby Thai, Ronald G Thomas, Paul Thompson, Jared Tinklenberg, Arthur W Toga, Geoffrey Tremont, J. Q. Trojanowki, Dick Trost, Raymond Scott Turner, Christopher H van Dyck, Helen Vanderswag, Daniel Varon, Javier Villanueva-Meyer, Teresa Villena, Sarah Walter, Paul Wang, Franklin Watkins, Michael Weiner, Jeff D Williamson, David Wolk, Chuang-Kuo Wu, Maria Zerrate, and Earl A Zimmerman. Apolipoprotein e (apoe) genotype has dissociable effects on memory and attentional-executive network function in alzheimer's disease. *Proc Natl Acad Sci U S A*, 107(22):10256–10261, Jun 2010. 136

A. H. Wyllie, J. F. Kerr, and A. R. Currie. Cell death: the significance of apoptosis. *Int Rev Cytol*, 68:251–306, 1980. 11

Y. Yamaguchi and S. E. Pfeiffer. Highly basic myelin and oligodendrocyte proteins analyzed by nephge-two-dimensional gel electrophoresis: recognition of novel developmentally regulated proteins. *J Neurosci Res*, 56(2):199–205, Apr 1999. 68

Xinwen Yang, Jianhua Wang, Cunren Liu, William E Grizzle, Shaohua Yu, Shuangqin Zhang, Stephen Barnes, William J Koopman, John D Mountz, Robert P Kimberly, and Huang-Ge Zhang. Cleavage of p53-vimentin complex enhances tumor necrosis factor-related apoptosis-inducing ligand-mediated apoptosis of rheumatoid arthritis synovial fibroblasts. *Am J Pathol*, 167(3): 705–719, Sep 2005. 70

Roman A Zubarev. Electron-capture dissociation tandem mass spectrometry. *Curr Opin Biotechnol*, 15(1):12–16, Feb 2004. 28