DISS. ETH Nr. 19037

Design and Validation of Proteome Measurements

A dissertation submitted to the

SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of

Doctor of Sciences

presented by

Manfred Claassen
Dipl. Biochem. & Dipl. Inf.
University of Tübingen
born May 10 1977
citizen of Germany

accepted on the recommendation of

Prof. Dr. Joachim M. Buhmann examiner
Prof. Dr. Ruedi Aebersold co-examiner
Prof. Dr. Oliver Kohlbacher co-examiner

2010

Abstract

Proteomics is a branch in biology that aims to comprehensively characterize a proteome. Mass spectrometry based proteomics has proven to be the most powerful approach to achieve this goal. This thesis introduces statistical concepts to optimally design and validate shotgun proteomics experiments and thereby enables to efficiently achieve reliable and extensive proteome coverage.

The first part reports methods to estimate false discovery rates for peptide and protein identifications. These approaches enabled to reliably and comprehensively identify unusually modified protein variants. It turned out that these variants contribute to a significant fraction of the spectral evidence. This work presents a generalized target-decoy approach to estimate false discovery rates for protein identifications. This work shows evidence that the reliability of protein identifications in large studies has so far been largely overestimated and provides guidelines to compile identifications at well defined confidence. This part concludes with formulating a generic framework to compare protein inference engines based on protein identification false discovery rates. A systematic comparison of thousands of protein inference variants revealed that simple approaches yield optimal inference performance.

The second part develops a nonparametric Bayesian approach to optimally design shotgun proteomics studies. Therefore the proteome coverage prediction task is introduced. An extended infinite Markov model is presented to perform proteome coverage prediction for simple shotgun proteomics experiments is presented. To capture the intricate similarities among peptide distributions arising in integrated shotgun proteomics studies, this work developed the general concept of the fractal Dirichlet process that augments the hierarchical Dirichlet process by introducing self-referential base measures. The fractal process is successfully applied to predict proteome coverage for integrated shotgun proteomics datasets. Rational stop criteria for these studies are discussed and evaluated by means of the proteome coverage prediction approaches. Finally the proteome coverage approaches are integrated into a study design framework that enables to determine an experimental sequence that achieves maximal expected increase in proteome coverage.

Zusammenfassung

Die Proteomik ist ein Teilbereich der Biologie, der die vollständige Charakterisierung eines Proteoms zum Ziel hat. Massenspektrometrie basierte Proteomik hat sich als erfolgreichste Strategie zum Erreichen dieses Ziels herausgebildet. Diese Arbeit stellt statistische Methoden zur optimalen Planung und Validierung von Shotgun-Proteomik-Experimenten vor. Diese Methoden ermöglichen eine effiziente, zuverlässige und zugleich umfassende Proteomcharakterisierung.

Der erste Teil der Arbeit stellt Methoden zur Schätzung von False Discovery Raten für Peptid- und Proteinidentifikationen vor. Diese Methoden ermöglichen die zuverlässige und umfassende Identifikation von ungewöhnlichen chemischen Proteinmodifikationen. Die Anwendung dieser Methoden hat gezeigt, dass diese Varianten zu einem beträchtlichen Anteil der massenspektrometrischen Daten beitragen. Diese Arbeit stellt einen generalisierten Target-Decoy Ansatz zur Schätzung von False Discovery Raten für Proteinidentifikationen vor. Unsere Resultate zeigen, dass die Zuverlässigkeit von Proteinidentifikationen in grossen Studien bis dato bei weitem überschätzt wurde. Angesichts dieser Resultate schlagen wir Richtlinien für die Zusammenstellung von Proteinidentifikationen vor, die eine definierte Konfidenz gewährleisten. Dieser Teil schliesst mit der Formulierung eines generischen Systems zum Vergleich von Proteinidentifikationsmethoden, das die Zuverlässigkeit der Identifikationen berücksichtigt. Ein systematischer Vergleich von tausenden von Proteinidentifikationsvarianten hat gezeigt, dass einfache Methoden bereits optimale Performanz erzielen.

Der zweite Teil der Arbeit entwickelt einen nichtparametrischen Bayesschen Ansatz zur optimalen Planung von Shotgun-Proteomik-Studien. Hierfür wird die Aufgabe der Proteomabdeckungsvorhersage eingeführt. Ein erweitertes infinites Markovmodell wird zur Durchführung der Proteomabdeckungsvorhersage für einfache Shotgun-Proteomik-Experimente vorgestellt. Diese Arbeit stellt das neue Konzept eines fraktalen Dirichlet Prozesses vor, um die Ähnlichkeit der Peptidverteilungen in integrierten Proteomikstudien zu erfassen. Der fraktale Dirichlet Prozess erweitert den hierarchischen Dirichlet Prozess um selbstbezügliche Basismasse. Der fraktale Dirichlet Prozess wird erfolgreich zur Proteomabdeckungsvorhersage für integrierte Proteomikstudien verwendet. Diese Arbeit diskutiert rationale Stopkriterien für derartige Studien und evaluiert diese mit Hilfe der vorgestellten Methoden zur Proteomabdeckungsvorhersage. Schliesslich werden die Methoden zur Proteomabdeckungsvorhersage in einem System zur Planung von Proteomikstudien eingesetzt, dass eine Sequenz von Experimenten bestimmt, die den maximalen erwarteten Zuwachs der Proteomabdeckung erzielt.

Acknowledgments

Lots of people have accompanied me during my doctoral studies. Here, I want to specially mention some of these. For those I forgot, I hereby thank them, too.

I thank both my thesis supervisors Joachim Buhmann and Ruedi Aebersold. I am grateful to have worked in the fruitful environment that both have created in as well as across their groups. Ruedi and Joachim have inspired me throughout as scientists as well as personalities.

I thank Oliver Kohlbacher for joining my thesis committee and contributing his interdisciplinary expertise. My interaction with Oliver during my time as a student in Tübingen has influenced me a lot. I am grateful for having had such a good and inspiring start into the working as a scientist.

Special thanks go to Lukas Reiter. I am grateful for the fruitful scientific and personal interaction with Lukas. Our collaboration has benefited a lot from his creative and critical way to approach scientific problems. Working with Lukas has been a lot of fun.

I thank Alex Schmidt for our fruitful, instructive and open collaboration. Alex contributed considerably to my understanding of the secrets of mass spectrometry. I am happy to have had the possibility to work with Vinzenz Lange. His scientific rigor has impressed me a lot.

Thomas Fuchs has accompanied me throughout my doctoral studies. I am thankful for his humor and the many, not necessarily scientific, discussions that were helpful in many regards.

I thank the people in the Aebersold/Buhmann lab for the nice, productivity catalyzing atmosphere. My office mates in the center, i.e. Alberto Busetto (a.k.a constant source

of stories and wisdom), Xeno Floros and Tom Dikk and on Hoenggerberg, i.e. Lukas Müller, Abdullah Kahraman, Hendrik Weisser deserve a special mention here, as well as our part time office mates, Leila Pirhaji, Paola Picotti (and Bernd popping in every now and then). Thanks to Tom Knöpfel for his untiring and creative contributions during his master thesis. Thanks to Cheng Ong Soon for many helpful and fun discussions. And thanks to Peter Schüffler and Elias August, the few group members that are knowledgeable in Skat. The Skat sessions could not have been more fun.

I also thank the administrative staff Rita Klute, Stephanie Fischer, Rachel Schumacher and Helena Zuber for facilitating the doctoral students' lives in such a professional and kind fashion.

Special thanks go to my parents, Maria Elena and Norbert Claassen. Their constant support has been indispensable throughout my studies.

Last but foremost I want to thank my wife, Julia Claassen. I am thankful for her patience with me during the more hectic phases of my PhD. Her constant support throughout the ups and downs of my doctoral studies were an invaluable source of motivation.

Contents

1 Introduction 1

1.1 Contributions 4

1.2 Authorship 4

2 Proteomics 5

2.1 Protein primer 5

2.2 Shotgun proteomics 7

2.2.1 Fractionation techniques 8

2.2.2 Mass spectrometry 10

2.2.3 Peptide-spectrum matching 13

2.2.4 Protein inference 16

2.3 Characterization of a proteome 17

I Validation of Proteome Measurements 20

3 Iterated Target-Decoy Database Search Strategy 21

3.1 Summary 21

3.2 Introduction 22

3.3 Spectral data 24

3.4 Peptide-spectrum matching 24

3.5 Iterated target-decoy database searching 25

3.6 Peptide-spectrum match false discovery rates 26

3.7 Results 26

3.7.1 Efficient screening of hundreds of modifications 27

3.7.2 Spectral contribution of peptide variants 28

3.7.3 Discovery of novel peptide sequences 29

3.8 Discussion 30

4 Protein Identification False Discovery Rates **32**
4.1 Summary 32
4.2 Introduction 33
4.3 Methods 35
4.3.1 Spectral data and database searching 35
4.3.2 Target-decoy database generation 37
4.3.3 Estimate of protein identification FDR 37
4.3.4 Derivation of the protein identification FDR estimate 40
4.3.5 Simulation of non-uniformly distributed false positive PSM 42
4.3.6 Validation of single hit FDR using isoelectric point information . 42
4.3.7 Validation of single hit FDR using synthetic peptides 43
4.3.8 *MAYU* analysis on ProteinProphet protein identifications. 44
4.4 Results 45
4.4.1 *MAYU* - FDR for protein identifications. 45
4.4.2 Validation of protein identification FDR estimate 46
4.4.3 Comparison of decoy database types 50
4.4.4 Comparison of protein identification FDR estimates 52
4.4.5 Protein identification FDR for various data sets 54
4.4.6 Data set size dependent accumulation of false positives 56
4.5 Discussion 56

5 Generic Comparison of Protein Inference Engine Families **61**
5.1 Summary 61
5.2 Introduction 62
5.3 Dataset and data processing 65
5.4 Local false discovery rates for protein identification subsets 66
5.5 Protein identification selection schemes 68
5.6 Screening and false discovery rate evaluation of selection schemes 68
5.7 Protein inference engine benchmark 69
5.8 Results 69
5.8.1 Local false discovery rates for protein identification subsets 69
5.8.2 Pruning protein identifications does not enhance inference 72
5.8.3 Simple protein inference engines are competitive 73
5.9 Discussion 74

II Design of Proteome Measurements 77

6 Proteome Coverage Prediction with Infinite Markov Models 78
6.1 Summary . . . 78
6.2 Introduction . . . 78
6.3 Dirichlet processes priors for peptide distributions . . . 82
6.4 Infinite Markov model for LC-MS/MS experiments . . . 83
6.5 Sampling sequences of peptide identifications . . . 85
6.6 Posterior prediction of proteome coverage progression . . . 86
6.7 Proteome coverage progression with false identifications . . . 88
6.8 Results . . . 89
6.8.1 Simulation study for parameter estimation assessment . . . 89
6.8.2 Cross validation prediction accuracy . . . 90
6.8.3 Proteome coverage prediction benchmark . . . 91
6.8.4 Prediction of maximal proteome coverage . . . 91
6.9 Discussion . . . 92

7 The Fractal Dirichlet Process 95
7.1 Summary . . . 95
7.2 Introduction . . . 95
7.3 Dirichlet processes . . . 96
7.4 Hierarchical Dirichlet processes . . . 99
7.5 Fractal Dirichlet processes . . . 100
7.6 Bayesian inference . . . 104
7.7 Experiments . . . 106
7.8 Discussion . . . 107

8 Proteome Coverage Prediction for Integrated Proteomics Datasets 110
8.1 Summary . . . 110
8.2 Introduction . . . 110
8.3 Pitman-Yor processes . . . 114
8.4 Hierarchical process model for fractionation experiments . . . 115
8.5 Sampling sequences of protein Identifications . . . 117
8.6 Empirical Bayes parameter estimate . . . 118
8.7 Proteome coverage prediction with false identifications . . . 120
8.8 Results . . . 120

8.8.1 Cross validation prediction accuracy 121
8.8.2 Proteome coverage prediction benchmark 122
8.8.3 Detection of saturation coverage for *L. interrogans* 122
8.9 Discussion 123

9 Optimal Design of Integrated Proteomics Experiments 126
9.1 Summary 126
9.2 Introduction 126
9.3 Experiment Selection 128
9.4 Empirical Experiment Selection 129
9.5 Reduction to maximum k-cover 131
9.6 Computation of maximum k-cover 132
9.7 Discussion 133

10 Conclusion 135

1 Introduction

The field of proteomics is a branch of biology that aims to comprehensively characterize the protein complement of a genome, i.e. the respective proteome. Recent technological advances have enabled biologists to comprehensively monitor a proteome under systematic and specific perturbations and thereby to contribute to systems level models of biological processes at molecular resolution. Mass spectrometry based proteomics has emerged as the most powerful approach to substantially cover a proteome. Projects aiming at extensive proteome coverage typically involve extensive experimentation and thereby accumulate large amounts of noisy mass spectrometrical data [2]. The necessary statistical methods to design such studies and assess the reliability of their reported protein identifications have been poorly understood until recently.

This thesis develops novel statistical concepts to optimally design and validate shotgun proteomics experiments and demonstrates how these are applied to efficiently achieve reliable and extensive proteome coverage.

The most extensive proteome coverage has been achieved by a strategy referred to as shotgun proteomics. Briefly, proteins are extracted from their biological source, enzymatically digested and optionally fractionated. The resulting peptide mixtures are then analyzed by tandem mass spectrometry. A range of search engines is available to subsequently infer peptide and protein identities from the acquired peptide fragment ion spectra.

The stochastic relationship between the object of interest, the peptide and its indirect observation, the fragment ion spectrum, endow peptide and protein identifications with uncertainty. The ability to quantify this uncertainty constitutes an indispensable prerequisite to evaluate these identifications. The first part of this thesis presents approaches to quantify this uncertainty in terms of false discovery rates. These approaches generalize the target-decoy strategy for peptide-spectrum matches from simple database searches.

I first adapted the target-decoy strategy for iterated database searches that enables to comprehensively account for modified variants of peptides and proteins that are not reported in protein databases and found that these variants contribute to a significant proportion of the spectral evidence.

Protein identifications are defined by assemblies of peptide-spectrum matches and constitute the biologically relevant outcome of a shotgun proteomics experiments. Since the uncertainty afflicted to protein identifications has been only poorly understood until recently, there have been many debates how to properly evaluate shotgun proteomics studies aiming at large proteome coverage. This work reports a target-decoy strategy to generically estimate false discovery rates for protein identifications. This approach allowed for the first time to assess the reliability of protein identifications across data sets of arbitrary size. We particularly found that the fraction of false positive identifications has been severely underestimated in large shotgun proteomics studies. We consequently provide rational guidelines to compile reliable protein identification sets from these studies.

The lacking ability to assess the reliability of protein identifications had so far precluded a fair comparison of the many available protein inference engines. We proposed a performance measure for protein inference based on protein identification false discovery rates that enabled us to perform such a comparison. The benchmark of thousands of inference variants on the largest publicly available shotgun proteomics dataset for *C. elegans* indicated that already the most simple protein inference approaches yield optimal performance.

The second part of this thesis develops a nonparametric Bayesian framework to optimally design a shotgun proteomics study. Complementary to the efforts to optimally exploit the given data as addressed in the first part, this approach aims at deciding which experiments to carry out in order to generate the most informative data.

In a first step, I have introduced the task of proteome coverage prediction that refers to estimating the expected number of proteins to be discovered upon carrying out a specified sequence of further experiments. Proteome coverage prediction constitutes a central task in optimal design of a shotgun proteomics study. Besides its role in study design, proteome coverage prediction enables to formulate rational stop criteria for already advanced studies that already have achieved close to maximal proteome coverage.

The liquid chromatography tandem mass spectrometry marks the elementary experiment of a shotgun proteomics study. I have developed an extended infinite Markov model that enabled me to predict proteome coverage for repetitions of such an experiment. Proteome coverage prediction for a *D. melanogaster* data set revealed that maximal coverage might be constrained by the accumulation false positive peptide identifications and prematurely achieved before reaching saturation coverage.

Most large shotgun proteomics build on multidimensional fractionation experiments that study an ensemble of different and yet similar peptide or respectively protein distributions. A model for a multidimensional fractionation experiment requires to account for this similarity.

This requirement inspired the novel general concept of the fractal Dirichlet process. The fractal Dirichlet process generalizes the hierarchical Dirichlet process by introducing self referential base measures and thereby enables to explicitly capture similarities among a subset of members of a set of discrete distributions. The description of the fractal Dirichlet process is completed by providing a Gibbs sampler for fully Bayesian inference.

I present a variant of the fractal Dirichlet process to perform proteome coverage prediction for multidimensional fractionation experiments. Application of this method to a dataset acquired for the bacterium *L. interrogans* revealed that saturation coverage had already been achieved and that further experimentation is not expected to yield a significant number of protein discoveries.

In conclusion, this thesis describes the optimal design of a shotgun proteomics study. Optimal design is formulated as the experiment sequence that maximizes the expected proteome coverage. We show that this optimization task reduces to the maximum k-cover problem. We explore different routes to solve and to practically apply this optimal design task.

1.1 Contributions

(1) Target-decoy strategy for iterated database searches

(2) Comprehensive identification of modified peptide/protein variants

(3) False discovery rates for protein identifications

(4) Local false discovery rates for protein identification subsets

(5) Guidelines for statistically sound evaluation of shotgun proteomics studies

(6) Protein inference engine benchmark framework

(7) Proteome coverage prediction task

(8) Extended infinite Markov Model formulation of LC-MS/MS experiments

(9) Fractal Dirichlet processes

(10) Fractal Dirichlet process model of multidimensional fractionation experiments

(11) Rational stop criteria for shotgun proteomics studies

(12) Optimal design of shotgun proteomics studies

1.2 Authorship

Unless otherwise noted, the authorship of the following work is as follows. I developed and implemented the main ideas of each project under the supervision of Ruedi Aebersold and Joachim Buhmann. At this point I want to acknowledge Alexander Schmidt for sharing the *D. melanogaster* and *L. interrogans* dataset for several projects presented in this thesis.

2 Proteomics

The dramatic technological advances in biology have enabled researchers to monitor biological systems at ever increasing throughput, integrity and resolution. These developments gave rise to the so called omics fields, among them the field of proteomics. Proteomics focuses on the characterization of the protein complement of a genome, i.e. the respective proteome [3]. Proteomics contributes essentially to the development of systems level models for cellular dynamics since most of the processes in cellular systems are mediated by proteins [76]. The following sections will give an overview of the field of proteomics and will particularly focus on aspects of mass spectrometry based proteomics that this work elaborates on.

2.1 Protein primer

Proteins are highly structured amino acid chains that are involved in almost every process of a biological system. Twenty different basic amino acids serve as building blocks of these chains. Each of these amino acids features particular physico-chemical properties, such as e.g. volume, hydrophobicity, charge, isoelectric point, polarity (Fig. 2.1a). Considering this diversity and the large amount of possible amino acid sequences, it becomes clear that proteins constitute a very heterogeneous class of molecules in terms of physico-chemical properties and are thus suited to implement the versatile molecular machinery in biological systems.

The amino acids in a protein are coupled by means of peptide bonds (Fig. 2.1b). Peptide bonds will play an important role as “pull linkage” of mass spectrometry based proteomics approaches. A typical protein counts about 200 linearly coupled amino acids (primary structure). Short amino acid subsequences turn out to form characteristic structural patterns (secondary structure), such as alpha helices or beta sheets. The substructures of a protein fold into a defined three dimensional structure and allow the protein to exert its function in a cell (Fig. 2.1c). Some proteins further undergo a for-

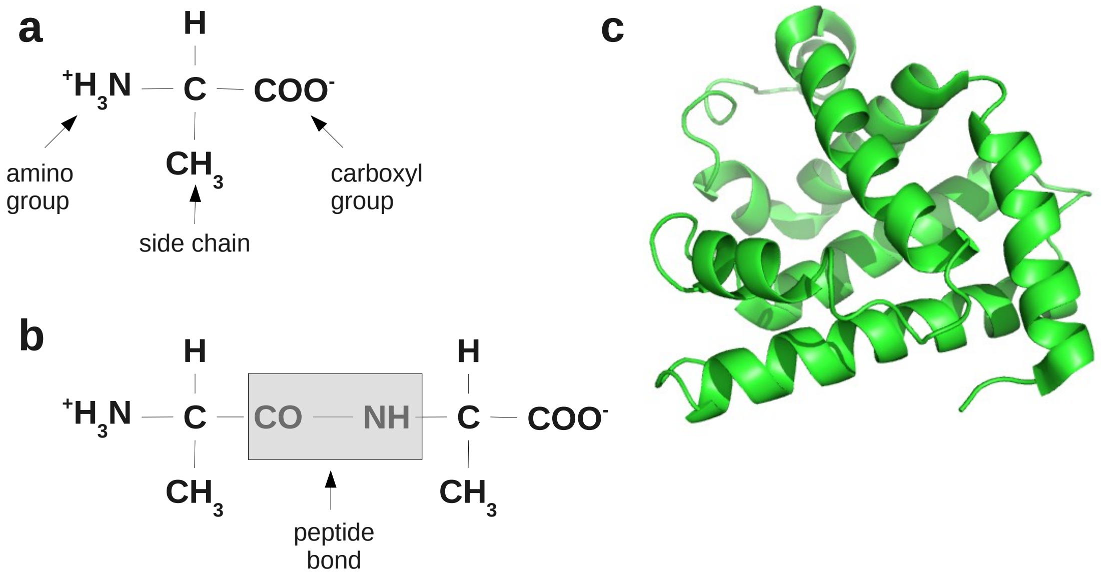


Figure 2.1: (**a**) Chemical structure of alanine, an example amino acid. Important functional groups are depicted. Amino acids other than alanine feature a different side chain. (**b**) Dimer of alanines coupled by a peptide bond. (**c**) Three dimensional structure of a myoglobin, the first protein for which a structure has been determined [73]. Secondary structures are highlighted by "cartoon" view.

mation of multimolecular complexes to be fully functional (quaternary structure).

Proteins are synthesized in the cell according to the central dogma of molecular biology. The genome constitutes the blueprint for all protein sequences of an organism. The genome is essentially a very long sequence of nucleotides, the human genome for instance counts about three billion nucleotides. Those parts of the genome that contain protein coding sequences are referred to as genes. Any cellular system implements the same steps to synthesize a protein according to the template encoded by a gene. This process is referred to as gene expression and it can be dissected into two steps, transcription and translation. Transcription refers to the process of synthesizing an mRNA transcript that corresponds to a (complementary) copy of the gene sequence. In the following translation step the transcript is processed by a complex protein machinery which catalyzes the synthesis of the amino acid chain encoded by the transcript. Optionally, proteins might be post-translationally modified by conjugation with some chemical agent, such as e.g. a phosphoryl group.

Proteins carry out a variety of functions. Enzymes are a class of proteins that serve as catalysts of chemical reactions in a cellular system. Our metabolism would not work efficiently without these protein catalysts. Another important role of proteins is to contribute to structures stabilizing a cell, such as e.g. actin filaments. Proteins play a crucial role in a cell's response to changes of its environment, by acting as molecular sensors and mediators of signal transduction. Specific chemical protein modifications, such as e.g. phosphorylations, play an important role of proteins as information carriers. These general examples are only a very small selection among the many proteins and their even more numerous functions.

A systems level understanding of biological systems requires to comprehensively monitor the entirety of proteins, i.e. the proteome. The field of proteomics aims at achieving this facet of systems biology. Due to the dramatic developments of mass spectrometry based proteomics approaches this goal is within reach now.

2.2 Shotgun proteomics

Proteins have been a long standing focus of biology research due to their important and ubiquitous role in biological systems. Until recently the intricate protein chemistry, complexity and dynamic range precluded to systematically explore a proteome.

Identification of a single protein constituted a major undertaking some decades ago. The protein of interest had to be first isolated by biochemical purification [121] and second sequenced by Edman degradation [39]. Gel based protein separation techniques alleviated the isolation step to some extent [98, 143]. Protein sequencing though remained cumbersome.

Several technological advances revolutionized the study of proteins. (1) The genomic revolution gave rise (and still does) to numerous genome sequences, with the human genome being the most prominent one [137]. This wealth of genomic information helped to compile comprehensive protein databases. It turns out that protein identification is a much easier task if expectations about possibly present proteins can be narrowed down by means of a protein database. (2) Mass spectrometry was developed up to a level to routinely measure biomolecules such as polypeptides [45, 69]. Mass spectrometrical methods turned out to be a generic high throughput alternative to conventional protein sequencing approaches [11].

These advances laid the ground for the shotgun proteomics approach, enabling biologists to identify thousands of proteins at once. This approach borrows from its namesake, the genome shotgun sequencing approach which reconstructs whole genomes from sequencing random DNA fragments [50]. The shotgun proteomics approach operates on the level of protein fragments, i.e. peptides to reconstruct the ensemble of proteins present in a biological sample [65]. Both approaches implement a divide-and-conquer strategy commonly encountered in computer science, i.e. to solve a difficult task by breaking it down to many related easy tasks [28]. The reconstruction of the difficult task's solution is typically non-trivial.

Shotgun proteomics workflows comprise three steps (Fig. 2.2). First, proteins are biochemically extracted from a biological sample and then, they are enzymatically digested to yield a complex ensemble of peptides. Protein and/or peptide ensembles are optionally further fractionated according to physical/chemical/biological properties. Second, tandem mass spectrometry is used to sample and identify individual peptide species present in the resulting ensembles and to finally recover the set of proteins initially present in the biological sample.

2.2.1 Fractionation techniques

The complexity of a proteome poses sizable challenges for its systematic exploration. The protein or respective peptide mixture resulting from a whole proteome extraction is by far too complex to be characterized directly by mass spectrometry. A variety of physico-chemical approaches have been proposed to disperse these mixtures into more tractable mixtures of lower complexity.

Two-dimensional polyacrylamide gel electrophoresis has long been the separation technique of choice. This approach operates on the level of intact proteins. Proteins are consecutively separated according to two properties. The first step typically involves isoelectric focusing according to isoelectric point. In a second electrophoresis step proteins are separated according to molecular weight. In principle, this technique is able to resolve thousands of proteins. However, several important protein classes are difficult to detect by gel based separation, such as e.g. the class of hydrophobic proteins [114]. This phenomenon particularly complicates the study of membrane proteins. In addition, low abundant proteins have shown to be underrepresented after gel separation [60]. These

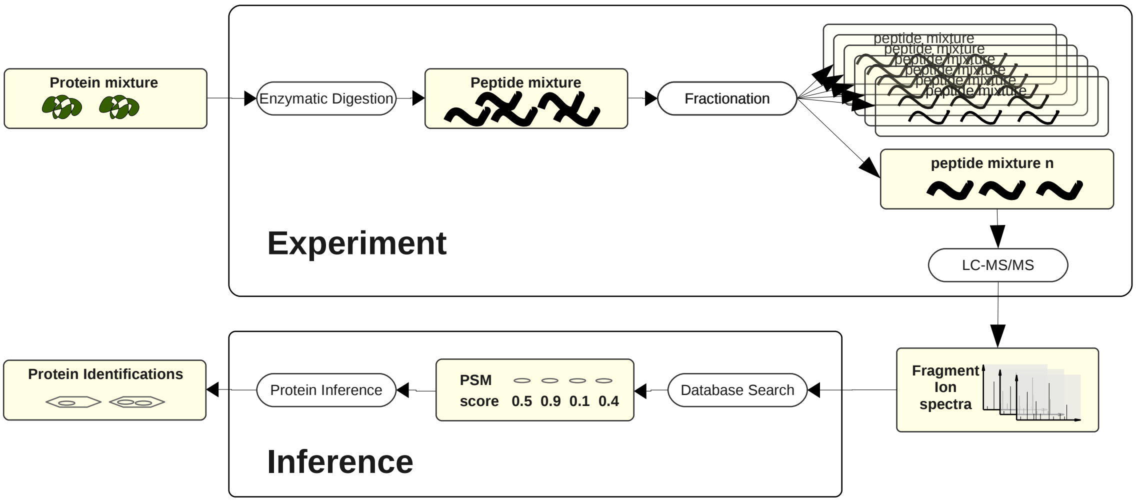


Figure 2.2: Shotgun proteomics workflow. Starting from a protein mixture, proteins are first enzymatically digested. The resulting peptide mixture is optionally fractionated according to a physical property other than hydrophobicity. The final peptide mixtures are analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS). The resulting fragment ion spectra constitute the data output of such a shotgun proteomics experiment. Peptide and protein identities are inferred in a two step procedure. Fragment ion spectra are first matched to peptide in a protein database, yielding a set of peptide-spectrum matches (PSM). Protein identifications are inferred by appropriately assembling the peptide-spectrum matches.

deficiencies disqualify gel based approaches as a separation techniques for global analysis of a proteome. Consequently, there has been a transition towards gel free separation approaches that operate on the level of peptides rather than proteins [104].

Reversed phase liquid chromatography (LC) constitutes the central fractionation technique of mass spectrometry based proteomics approaches [38]. Reversed phase liquid chromatography is typically performed on peptide mixtures. The separation device of the chromatography system consists of a column that holds a hydrophobic matrix of alkyl chains (stationary phase). A peptide mixture is injected into the column under constant buffer flow (mobile phase). The time difference between injection and elution of a peptide is referred to as retention time. The retention time of peptides of a particular peptide species is determined by the strength of their interaction with the matrix. As a rule of thumb, reversed phase liquid chromatography separates according to hydrophobicity. In the context of mass spectrometry based proteomics workflows, the chromatography system directly interfaces a mass spectrometer [142]. This setup allows to automatically analyze the eluting peptide mixtures by tandem mass spectrometry without any further intervention.

Multidimensional fractionation approaches comprise multiple different fractionation steps [55]. In order to achieve good separation performance these fractionation steps are ideally chosen to be orthogonal. Besides typically including reversed phase liquid chromatography, these approaches consider fractionation steps separating according to physicochemical properties other than hydrophobicity. Popular properties are isoelectric point, charge state, size and molecular weight. These respective fractionation steps are frequently implemented in a liquid chromatography system with an appropriate stationary phase [93, 102, 135] or an electrophoresis system [92].

Other approaches deal with the complexity of a proteome by focusing on an information rich fraction of a proteome. Previous work covers approaches enriching for e.g. cysteine-containing peptides [61], phosphorylated peptides [152, 47], or glycosylated peptides [149].

The complexity of a proteome by far exceeds the capacity of contemporary mass spectrometers in the context of shotgun proteomics workflows. Fractionation strategies effectively reduce the complexity of the peptide mixtures that are subjected to mass spectrometrical analysis and, therefore, they constitute a prerequisite for a comprehensive analysis of a proteome.

2.2.2 Mass spectrometry

Mass spectrometry has emerged as the central analytical technology to identify proteins. Compared to chemical sequencing or antibody based approaches, mass spectrometry is unsurpassed in its combination of throughput, sensitivity and information rich readout [110].

A mass spectrometer consists of three main components, an ion source, a mass analyzer that measures the mass-to-charge ratio (m/z) of the ionized analytes, and a detector that counts the number of ions at each m/z value. Soft ion sources like electrospray [45] or matrix assisted laser desorption ionization [69] intactly ionize large biomolecules. This development rendered mass spectrometry amenable to proteomics. Typical mass analyzers comprise the time-of-flight, ion trap, Fourier transform ion cyclotron or orbi trap analyzer. These analyzers differ in their characteristics, having their weaknesses and strengths. [38] provides a brief comparison in terms of mass accuracy, resolving power, sensitivity, dynamic range and throughput.

Mass spectrometry supported to easily identify substantially purified proteins by peptide mass fingerprinting [62]. This task typically arises after performing two-dimensional polyacrylamide gel electrophoresis. Each spot on the resulting gel is highly enriched for a single protein. Peptide mass fingerprinting consists of first enzymatically digesting the purified protein, measuring the masses of the resulting peptides and to reconstruct the protein's identity by matching the peptide masses to a protein database.

Mass spectrometrical analysis of complex protein or respectively peptide mixtures is more involved. Peptide mass fingerprints are not useful here since in general, they do not unambiguously identify a protein in this context. Mass spectrometers that are suited to analyze a complex mixture allow to sequence its peptide species. These instruments therefore implement a two step scanning procedure that first registers the m/z ratios of all species of a mixture (MS1 level), then selects, isolates and fragments one of these species and records the resulting fragment ion spectrum (MS2 level) [91].

Peptide species designated for fragmentation are typically selected in a data dependent fashion, i.e. randomly selected from the strongest signals at MS1 level. This selection strategy has been successfully applied in numerous studies though it tends to systematically identify highly abundant proteins and less so low abundant ones. Directed proteomics approaches circumvent this drawback by systematically registering all signals at MS1 level in preparatory experiments and to exhaustively target these in subsequent experiments [118].

Most mass spectrometers implement a peptide fragmentation mechanism referred to as collision induced dissociation, where peptide fragmentation is invoked by allowing for collision with surrounding molecules [91]. This mechanism has turned out to make peptides preferably break across the peptide bond boundaries of peptides (Fig. 2.3). This convenient phenomenon supports to read out the amino acid sequence directly from the fragment ion spectrum as long as all of the peptide bonds break. Other complementary fragmentation mechanisms have been proposed to deal with missing fragments in collision induced dissociation. Such mechanisms comprise for instance electron capture

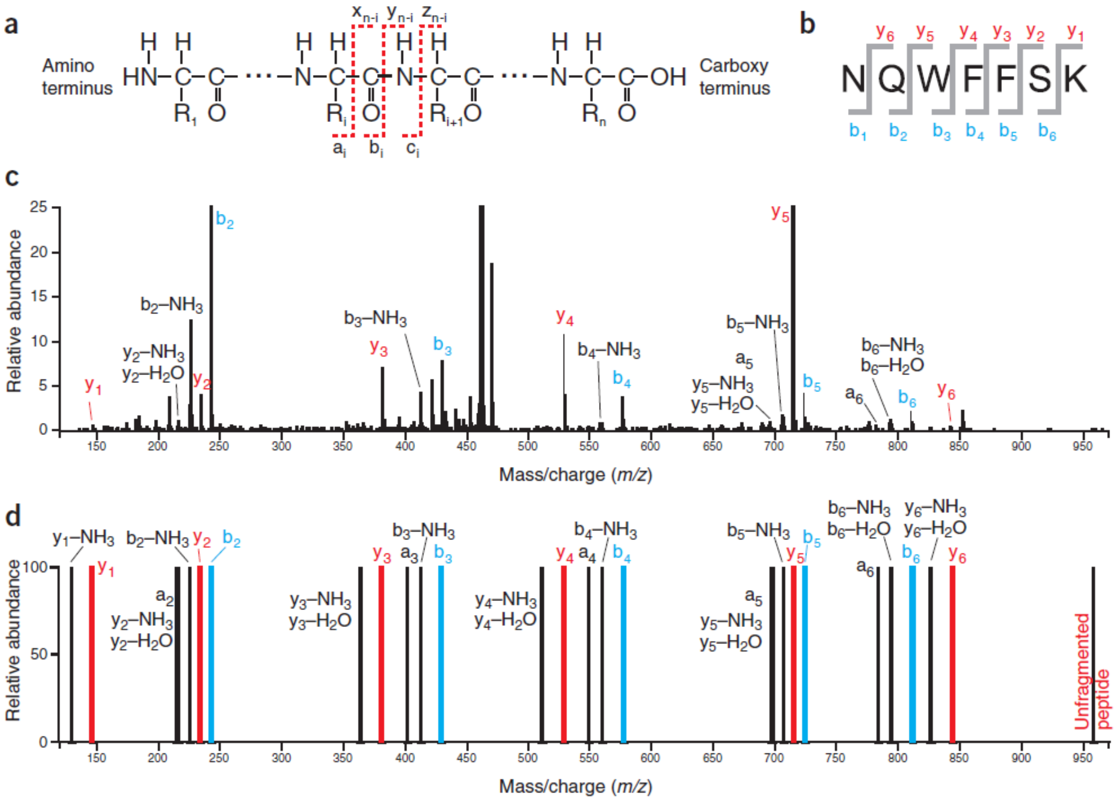


Figure 2.3: Fragmentation patterns for collision induced dissociation (adopted from [89]). (**a**) Peptides predominantly break around the peptide bond. Three different bonds can break. The respective N-terminal ions are referred to as a-, b- or c ions. The C-terminal ions are denoted x-, y- or z ions. The ions are indexed according to the position index of the next amino acid towards the N-terminus. (**b**) The predominant ions are the b- and y-ions. The respective peptide fragment ions are depicted for the exemplary peptide NQWSFFK. (**c**) Fragment ion spectrum recorded for the peptide NQWSFFK. b- and y-ions contribute most of the signal. A notable amount of signals can be attributed to b- and y-ions with neutral loss of water or ammonia. (**d**) Database search engines represent peptides with theoretical spectra with signals that are to be expected. The depicted theoretical spectrum considers a-, b- and y-ions including their neutral losses with unit signal intensity. Peptide-spectrum matches are generated by finding the peptide whose theoretical spectrum is most similar to the fragment ion spectrum.

dissociation [153] and electron transfer dissociation [130]. Nevertheless, the ideal situation where the fragment ion spectrum is complete is typically not met and calls for sophisticated approaches to infer the peptide sequence. This issue will be discussed in detail in the following section.

2.2.3 Peptide-spectrum matching

In the previous sections we have seen which experimental steps are performed to achieve the basic readout of a shotgun proteomics experiment, i.e. the peptide fragment ion spectra. This readout defines the data to infer the proteins initially present in the biological sample. Inference typically involves two steps, peptide-spectrum matching and protein inference [95]. Peptide-spectrum matching refers to assigning each fragment ion spectrum a peptide sequence that best explains its signals. Protein inference reconstructs the protein composition from the peptide-spectrum matches obtained in the first step. This section summarizes approaches to peptide-spectrum matching.

Peptide-spectrum matching is a task that admits a fragment ion spectrum as input and that consists of finding the peptide sequence best matching to the input according to a suitable objective function (score) [42]. The objective function encodes our understanding of the relation between a peptide and its fragment ion spectrum. The objective function is supposed to discriminate the peptide that gave rise to the input spectrum from all other peptides. It is non-trivial to find a good objective function since the fragmentation of peptides is only partially understood [146] and, furthermore, fragment ion spectra generated from complex peptide mixtures are exceedingly noisy [131].

Peptide-spectrum matching is usually implemented as a formalized, deterministic and fully automated process. This implementation makes peptide-spectrum matching objective and reproducible. Manual implementation of this task lacks this objectivity and is already impossible due to the large number of fragment ion spectra generated in a shotgun proteomics experiment.

Most of the peptide-spectrum matching approaches independently process each fragment ion spectrum. In a first step, a set of candidate peptides is selected according to some appropriate criteria. Each candidate is scored against the fragment ion spectrum. The top scoring candidate peptide in conjunction with the fragment ion spectrum is reported as peptide-spectrum match.

A variety of criteria are consulted to select peptide candidates for a fragment ion spectrum. These criteria can be divided into data independent and data dependent criteria. Data independent criteria comprise prior expectations regarding peptides and chemical modifications possibly present in the biological sample. Two different routes are pursued at this point. While *de novo* sequencing approaches do not restrict the set of a priori considered peptides [132, 84, 52, 49], database search approaches confine themselves to candidates from a static protein database [88, 42]. Data dependent criteria comprise constraints on the total mass of the intact peptide derived from the respective MS1 signal and possibly short sequence tags directly read from the fragment ion spectrum [88, 53].

Various scores have been proposed to match peptides against fragment ion spectra. Most of the following scores assume a theoretical spectrum of the peptide which contains unit signals for each expected fragment, e.g. all b- and y-ions. Recent scoring variants resort to spectral libraries to use consensus spectra as theoretical spectra [147, 31, 54, 80]. The fragment ion spectrum and the theoretical spectrum have been compared by means of cross correlation [42], dot product [48], hypergeometric score [113], probabilistic scores [105, 6, 27, 49] or physically motivated scores [151]. *De novo* sequencing approaches by definition consider an intractable amount of peptide candidates. It is infeasible to explicitly score each of the candidates against the respective fragment ion spectrum. The score, or more precisely the model underlying a *de novo* sequencing approach therefore has to feature an independence structure that render tricks like dynamic programming applicable to efficiently consider all candidates [34, 17].

There are other peptide-spectrum matching approaches that do not fit into the described candidate selection/scoring scheme. Some approaches do not restrict themselves to single peptide candidates and instead allow for multiple source peptides for a single fragment ion spectrum [150]. The distinction between selection and scoring is blurred in spectral alignment approaches that compute spectral networks by relating spectra that are likely to belong to overlapping peptide sequences [7].

Peptide-spectrum matches are not perfect. False positive peptide-spectrum matches arise when the top scoring candidate is not the source of the respective fragment ion spectrum. These events can mostly be attributed to flaws in the score related to the approximate encoding for the peptide fragmentation process and the lack of information in the fragment ion spectrum, e.g. in terms of lacking fragment ions. In fact, it is likely that the best fitting peptides among all imaginable peptides will not be the true peptide. This phenomenon strongly motivates the use of small protein databases to avoid the consideration of confounding peptides and demonstrates why *de novo* sequencing approaches have a hard time to be competitive with database search approaches (as long as a the genome of the studied organism is known). We note that the effect of large databases is alleviated by the trend towards mass spectrometers with increasing mass accuracies.

It is of crucial importance to control the quality of peptide-spectrum matches. Various statistical approaches have been devised to control different measures of peptide-spectrum match uncertainty, the false discovery rate being the most useful one [9]. In the context of peptide-spectrum matching, the false discovery rate corresponds to the expected fraction of false positive matches. Three routes can be pursued to estimate the false discovery rate for a set of peptide-spectrum matches.

The false discovery rate can be derived from p-values associated to each peptide-spectrum match that is considered significant [9, 129]. This approach is valid as long as p-values can be accurately computed. This requirement is though rarely met [74].

The false discovery rate can be estimated from the score distributions of true and false positive peptide-spectrum matches [72]. This mixture distribution has to be learned in an unsupervised scenario since the information whether a match is true or false positive is not known for any match. This task has been successfully implemented in e.g. PeptideProphet [72] by resorting to Expectation Maximization [36].

Recently, the target-decoy strategy became very popular to estimate the peptide-spectrum match false discovery rate [94]. A decoy database with nonsense protein sequences is searched in addition to the (target) protein database of the studied organism. The number of peptide-spectrum matches mapping to the decoy database serves as an estimate of the number of false positive matches. If the decoy database is designed similar to the target database, then we expect the false positive matches uniformly distribute across the target and decoy database. [41] have shown that reversed, pseudo-reversed as well as scrambled databases serve equally well as decoy databases, particularly ensuring uniform distribution of false positive matches. Its simplicity and generic applicability make the target-decoy strategy an appealing alternative to estimate false discovery rate of peptide-spectrum matches.

Peptide-spectrum matching has been studied now for more than fifteen years and several mature solutions to this task are available. Technological improvements of the mass spectrometers have largely contributed the increase in matching performance. Nowadays it is possible to confidently assign up to 80 % of the acquired fragment ion spectra on a high mass accuracy instrument (personal communication A. Schmidt). Statistical validation of peptide-spectrum matches obtained from standard database searches has also been successfully investigated from different angles. This work expands the target-decoy strategy to more intricate iterated database searches that support to efficiently consider a vast amount of possible amino acid modifications.

2.2.4 Protein inference

The previous section described the peptide-spectrum matching task, the first step of reconstructing the protein set in a biological sample from the fragment ion spectra acquired from a series of shotgun proteomics experiments. Protein inference constitutes the second step and, in simple terms, takes the peptide-spectrum matches as input and compiles a set of protein identifications.

The protein inference task is specific to the shotgun proteomics setup [110, 95]. Enzymatic digestion of the proteins into peptides facilitate sample handling and dramatically enhance throughput. These benefits come at the cost of loosing the information which proteins gave rise to which of the identified peptides. The more complex a proteome the more frequently peptide-spectrum matches turn out to ambiguously map to several protein entries, e.g. protein splice variants. Protein inference approaches aim to disambiguate these matches.

The protein inference task has been approached in various ways. Frequently, ambiguous peptide-spectrum matches are deterministically assigned to gene loci instead of resolving particular splice variants [90, 15, 5, 119]. Probabilistic approaches attempt to estimate the posterior probability of protein identifications based on confidence measures for peptide-spectrum matches [96, 112, 44]. Parsimony approaches compute the minimal set of protein identifications that are consistent with a set of peptide-spectrum matches that are considered significant [148]. After having applied some protein inference approach, it is common practice to exclude possibly unreliable protein identifications, such as e.g. single hit protein identifications. There has been considerable debate about whether such post-processing enhances protein inference [59, 58].

After having performed protein inference, it is mandatory to quantitatively assess the reliability of the resulting protein identifications. Statistical validation of protein identifications has long falsely been equated with statistical validation of peptide-spectrum matches. Therefore, this important issue is still a topic of ongoing research. This work contributes a generalized target-decoy strategy to estimate false discovery rates for protein identifications. Related work will be reviewed and discussed in this context.

2.3 Characterization of a proteome

Throughput and sensitivity of mass spectrometry based proteomics approaches allow to comprehensively characterize a proteome. In the following we will discuss approaches that have been reported to qualitatively and quantitatively describe a proteome.

Determination of a genome sequence is a fundamental goal of genomics. Shotgun sequencing technologies have evolved to a level where a genome can be routinely sequenced [10]. Due to the static, linear structure of a genome, it is in particular straightforward to tell, when the genome has been comprehensively characterized. For proteomics this task translates to identifying all proteins possibly present in a biological system. It is though not trivial to tell when a proteome can be considered to be mapped out since a proteome is not as clearly defined as a genome. This uncertainty is caused by the many variations of a gene expression product that might be introduced by alternative splicing and post translational modifications that, in addition, might only be present under particular conditions.

Shotgun proteomics has been the most successful approach to identify a large amount of proteins, i.e. to achieve substantial proteome coverage [138, 103, 100, 51, 75, 15, 5, 35, 58, 119]. All these approaches build on extensive repetition of multidimensional fractionation experiments. The respective sequencing studies seem to have reached saturation coverage with respect to the applied experimental strategy.

Various experimental strategies have been proposed to further enhance these approaches. Most contemporary shotgun proteomics studies rely on stochastic precursor selection in the mass spectrometer. This type of precursor selection results in redundantly identifying the same proteins over and over again, thereby considerably slowing down the process of discovering new proteins which would contribute to proteome coverage increase. Directed shotgun proteomics approaches circumvent this issue by systematically targeting all MS1 features for fragmentation [118]. Further improvements in sequencing speed and mass accuracy of tandem mass spectrometers will further accelerate this process [99]. Although contemporary proteome exploration studies achieve substantial coverage, it is likely that other experimental approaches will reveal that proteomes are much richer than current studies suggest. New fractionation techniques are likely to enrich for protein variants that feature particular modifications [14] or locations [144]. New modification preserving fragmentation techniques have the potential to further enhance these approaches [130]. Targeted proteomics approaches restrict themselves to measure a confined set of proteins of interest and might constitute a complementary technology to detect proteins that remain undetected with shotgun proteomics approaches [106].

Quantitative proteomics aims at augmenting the proteome characterization by quantitative information. As noted before, a proteome is not a static entity like a genome. In contrast, its members, i.e. the proteins are subject to dynamic abundancy changes. This dynamic behavior reflects responses of a biological system to its environment. It turns out that mass spectrometry based approaches lend themselves to quantify proteins. Isotope labeling [61, 101] strategies as well label free approaches [83, 16, 87] have been proposed to estimate relative or absolute protein abundancies from mass spectrometrical data. In this context targeted proteomics approaches [106] have demonstrated to be a potent alternative to shotgun proteomics approaches. Quantitative proteomics contributes to elucidate various biological processes in a local [57, 81, 141, 106, 115], as well as a proteome wide scale [35, 87].

Sizable experimental efforts have been made to achieve satisfactory proteome coverage. A variety of different biological samples or fractions arising in multidimensional fractionation strategy are typically repeatedly analyzed. It turns out that some samples contribute a lot to proteome coverage while others only redundantly cover parts of the already observed proteome [118]. This phenomenon suggests that appropriate design of a shotgun proteomics study has the potential to achieve saturation coverage at significantly reduced cost. This thesis contributes a generic nonparametric Bayesian model of shotgun proteomics experiments that estimates the coverage potential of samples and applies it to optimally design a shotgun proteomics study at an early stage to efficiently achieve saturation coverage.

The data analysis of large shotgun proteomics projects is challenging. This particularly applies to statistical validation of the database search results. While the first step from the raw fragment ion spectra to the peptide-spectrum matches is well established, there has been considerable debate how to compile a list of proteins that is as large as possible and yet reliable [138, 127, 58]. The question whether to consider single hit wonders or not, is a long-running issue in this context [59]. This thesis contributes a generic target-decoy strategy to estimate protein false discovery rates and suggests guidelines to achieve a list meeting size and quality requirements.

Part I

Validation of Proteome Measurements

3 Iterated Target-Decoy Database Search Strategy

3.1 Summary

Mass spectrometry based proteomics is suited to study post-translationally modified proteins by means of peptide fragment ion spectra. Matching the spectra to their respective peptide sequences is typically implemented as a database search, i.e. by exclusively matching candidates defined by a suitable protein database, optionally enumerating peptide variants to account for amino acid modifications. Increasing the number of considered modifications though severely affects computation and identification performance. This phenomenon limits the scope to comprehensively study the occurrence of modified peptide variants.

Here, we propose a generically applicable iterative target-decoy database search approach that enables to efficiently and reliably account for hundreds of different peptide modifications at once. This approach circumvents the combinatorial explosion coming along with exhaustively enumerating variants for all peptides in the database by considering modifications only for proteins confidently evidenced by tryptic peptides. We adapt the well established target-decoy search strategy to this iterative search approach in order to control the false discovery rates for peptide-spectrum matches. We applied our strategy to a *D. melanogaster* dataset comprising 84 LC-MS/MS runs considering more than 500 different modifications at once. We found 9.5% of all peptide-spectrum matches to map to a modified peptide variant. Due to its simplicity and generic nature, we expect that the iterated target-decoy strategy will enable us to reliably discover a diverse set of modified peptides in any other shotgun proteomic dataset.

3.2 Introduction

Comprehensively accounting for peptide modifications in mass spectrometry based proteomics poses considerable challenges to peptide identification algorithms. We present an iterated target-decoy strategy that circumvents the combinatorial explosion coming along with exhaustively enumerating all possible peptide variants and furthermore allows to control the reliability of peptide identifications in terms of false discovery rates (FDR).

Peptide identification is a task that arises in the context of shotgun proteomics experiments [3]. Briefly, protein samples are first extracted from their biological source and subjected to enzymatic digestion. These steps yield a complex peptide mixture that is analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS). Fragment ion spectra are acquired after stochastic or directed precursor ion selection [118]. More elaborate strategies adopt additional fractionation steps at the level of proteins/peptides before LC-MS/MS analysis. These steps give rise to a set of peptide fragment ion spectra that constitute the raw data to infer the peptides present in the biological source.

Interpretation of the spectral data involves peptide identification, i.e. matching the fragment ion spectra to their corresponding peptide sequences. Peptide-spectrum matches are typically generated using one of the many available search engines, e.g. [42, 105, 30]. Search engines map fragment ion spectra to the best matching peptide sequence in the protein database of the studied organism [97]. Various statistical measures, such as e.g. false discovery rates [9], have been derived to account for possibly incorrect peptide-spectrum matches [18]. In this context, the target-decoy strategy has recently gained large popularity since it is simple to implement and compatible with all currently used search engines [94, 41].

Protein databases report sequences of standard amino acids and do not capture the diversity of chemically modified peptides that happen to be present in the mass spectrometer. These modifications are either introduced by biological processes (e.g. phosphorylations) or sample preparation (e.g. carboxymethylations). Hundreds of different amino acid modifications have been reported and comprehensively documented [33]. In order to identify some of the corresponding peptide variants from fragment ion spectra, search engines typically consider a very small number of frequently observed amino acid modifications (e.g. oxidations) to exhaustively enumerate additional peptide variants from the standard sequences given in the protein database [42]. Increasing the number of considered modifications though dramatically increases the search space for peptide identification. This increase leads to a well known accumulation of false positive peptide identifications or simply becomes too demanding computationally.

Several approaches have been proposed to address the challenges involved in comprehensively accounting for modified peptides. Spectral alignment methods implicitly explore the vast number of possible modifications by dynamic programming [136]. Being compatible with any currently available search engine, iterative search strategies constitute a potent generic alternative to the specialized spectral alignment methods. Iterative search strategies explicitly constrain the search space by just enumerating peptide variants for the subset of proteins likely to be present in the biological source [29, 123]. These approaches rely on the hypothesis that present proteins are likely to give rise to at least one detectable tryptic peptide. Iterative search strategies therefore first exclusively search for tryptic peptides, giving rise to identifications pointing to a small subset entries of the complete protein database. Spectra not having been assigned to peptides in the first round are subsequently searched against a subdatabase solely comprising those protein entries that were identified by tryptic peptides before. The second search additionally enumerates peptide variants for a (possibly large) number of different modifications. Due to the small size of the subdatabase, exhaustive consideration of several hundreds of different modifications becomes feasible.

The ability to estimate sensible reliability measures, such as e.g false discovery rates [9], is essential to compile sets of peptide-spectrum matches of well defined quality [72]. Available iterative search strategies do only provide scores that are suited to enrich for correct identifications and therefore correlate with their reliability [29, 123]. These scores though do not directly lend themselves as sensible reliability measures. Therefore it remains to provide means to quantitatively assess the reliability of the peptide-spectrum matches to make iterative search approaches applicable in practice.

To close this gap, we extend the target-decoy strategy to estimate false discovery rates in the context of iterative database searches. We applied this strategy to a *D. melanogaster* dataset comprising 84 LC-MS/MS runs to study the occurrence of modified peptides. Considering hundreds of modifications at once we compiled a set of peptide identifications of well defined quality (1% FDR). We observed that peptide modifications are common. Specifically, we found almost every tenth peptide-spectrum match mapping to a modified peptide variant and 14% of all unique peptide sequences to be exclusively represented by a non-standard variant.

Methods

This section describes the spectral data that has been used in this study and the iterative target-decoy database search strategy. The presented procedures are implemented in C++ on the basis of the OpenMS [78] framework.

3.3 Spectral data

We studied the fragment ion spectrum dataset reported in [118]. Briefly, this dataset was acquired for a whole cell lysate of the *D. melanogaster* KC 167 cell line. Proteins were extracted and digested with trypsin as described in [118]. The resulting peptide mixture was repeatedly analyzed by LC-MS/MS on a high mass accuracy FT-LTQ instrument as described in [118]. This procedure gave rise to the fragment ion spectra that were further analyzed with the iterated target-decoy strategy reported in this paper.

3.4 Peptide-spectrum matching

This section summarizes how peptide-spectrum matches were generated from fragment ion spectra. We exemplarily describe the protocol starting from a single fragment ion spectrum.

In a first step the fragment ion spectrum is denoised by a common heuristic. This heuristic considers the characteristic intensity distribution of ion trap fragment ion spectra [131] and assumes that the most intense peaks correspond to the informative signals. Specifically, peak intensities are first rescaled according to their relative position to the precursor ion and second all peaks but those featuring more than 5.5% maximal intensity are removed. The remaining spectrum is binarized and discretized using a bin spacing of 1.00048 Da [105], finally constituting the denoised fragment ion spectrum.

Possible peptide explanations to the fragment ion spectrum are selected from the protein database according to the following criteria: (1) sequence length 5-40, (2) ≤ 1 tryptic miscleavages. Precursor mass tolerance is set dynamically for each LC-MS/MS run to two standard deviations of the mass deviation distribution. A preliminary search with 15 ppm precursor mass tolerance serves to estimate the mass deviation distribution. Precursor mass was typically set below 5 ppm. Mass tolerance for the fragment ion spectra was set to 0.5 Da. Each of the considered peptides is represented by its theoretical fragment ion spectrum. Theoretical spectra consider b and y ion series. Possible amino acid modifications are considered by their respective mass shifts. Each fragment is represented by a peak of intensity one.

Each of the theoretical spectra is matched to the denoised fragment ion spectrum. We use a variant of the hypergeometric score as similarity measure [113]. Specifically, the hypergeometric score is computed for each peptide. Each peptide is finally evaluated by means of the difference score that is computed as the difference of the hypergeometric score to the second rank peptide. The peptide-spectrum match corresponding to the initially supplied fragment ion spectrum is chosen as the top ranking peptide with respect to the difference score.

3.5 Iterated target-decoy database searching

Fig. 3.1 illustrates the iterated target-decoy database search strategy. As input, we assume a set of fragment ion spectra and protein database that captures our expectations regarding proteins possibly present in the biological sample. For the presented analysis of the *D. melanogaster* dataset [118] we used the Drosophila Flybase protein database (D. melanogaster, release 4.3; Mar 2006; 19645 entries) complemented with the protein sequences of bovine trypsin and human keratins.

A priori expectations about protein presence are updated by a first pass search against a concatenation of the forward and pseudo-reversed (following [41]) protein database without considering any amino acid modifications. Peptide-spectrum matches are generated as described in section 3.4. The top scoring peptide-spectrum matches (1% FDR) resulting from the first pass search are considered to be assigned. These assigned matches map to a set of proteins covering a fraction of the (target) protein database. This set of proteins is highly enriched for proteins present in the biological sample and itself constitutes a protein subdatabase.

A second pass search aims at matching so far unassigned fragment ion spectra to peptide variants mapping to entries in the subdatabase. Therefore we search the unassigned spectra against a concatenation of the forward and pseudo-reversed subdatabase also enumerating peptide variants defined by a set of amino acid modifications. For the presented analysis we considered all modifications from the Unimod database (Mar 2007) [33] excluding modifications containing elements other than H, C, N, O, P, S, Se and complemented by all possible amino acid mutations.

First and second pass searches give rise to a peptide-spectrum match for each of the initially supplied fragment ion spectra. A final subset of peptide spectrum matches is compiled according to a score threshold that achieves a user defined false discovery rate. For details see the following section 3.6.

3.6 Peptide-spectrum match false discovery rates

We want to apply the target-decoy strategy to estimate peptide-spectrum match false discovery rates. Therefore we have to ensure for each database search that target and decoy database are of equal size. In this context, database size is sensibly measured as the amount of peptides considered by the search engine, i.e. typically the number of tryptic peptides, possibly also including miscleavages. This point is important to consider in order to properly estimate false discovery rates from decoy matches acquired by the iterated database search strategy.

We ensure this requirement by choosing a concatenation of the forward and pseudo-reversed variant of database underlying the first as well as second pass search. The decoy database for the second pass search is generated dynamically after having performed and evaluated the first pass search. This database design specifically entails that every peptide in the forward database mirrors to exactly one counterpart in the decoy database and that false discovery rates for a set of peptide-spectrum matches can be estimated straightforwardly as the (normalized) amount of matches to the decoy database [41].

3.7 Results

We applied the iterative target-decoy strategy to a *D. melanogaster* whole cell lysate dataset comprising fragment ion spectra acquired over 84 LC-MS/MS runs. We screened

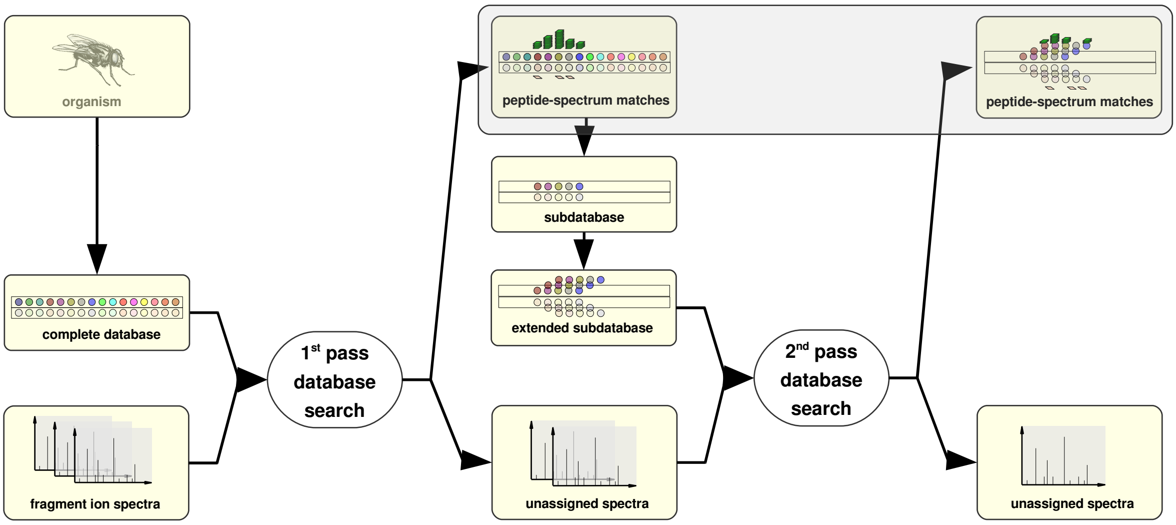


Figure 3.1: Iterated target-decoy database search pipeline. Fragment ion spectra are generated in the course of a shotgun proteomics experiment. The iterative target-decoy database search strategy comprises two steps. In the first pass search, the fragment ion spectra are matched against tryptic peptides derived from a suitable forward-reverse protein database defined by the underlying organism. The resulting high confidence peptide-spectrum matches (1% FDR) map to a subset of proteins that are likely to be present in the biological source. This subset serves to dynamically compile a forward-reverse subdatabase that subsequently is searched in the second pass search. The second pass search typically extends the subdatabse by also enumerating peptide variants derived from a set of user defined amino acid modifications. The peptide-spectrum matches from the first and second pass search constitute the results of the iterated target-decoy search. Due to the design of the decoy database in both first and second pass search, false discovery rates can be directly estimated by counting the decoy identifications.

for hundreds of different modifications at once, mapped a significant number of fragment ion spectra to modified peptide variants and furthermore found a considerable number of peptide sequences represented exclusively in a modified form.

3.7.1 Efficient screening of hundreds of modifications

The iterated database search approach enables to comprehensively consider known amino acid modifications. Specifically we considered 508 different amino acid modifications at once for the second pass search. Significantly reducing the number of entries in the protein database enables to consider such a large number of modifications. The original *D. melanogaster* protein database contains 19465 entries. The first pass search typically

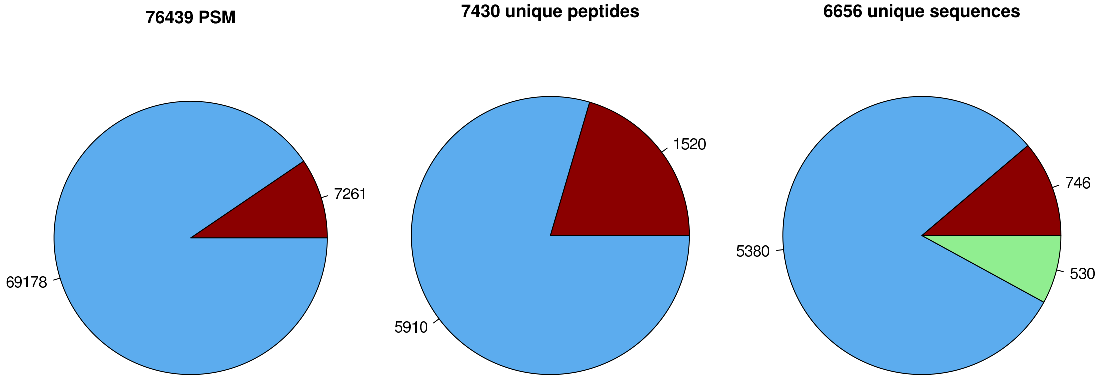


Figure 3.2: Summary of iterated target-decoy search results at 1% peptide-spectrum match false discovery rate for the *D. melanogaster* dataset. Frequency of non-modified (blue) and modified (red) instances of identification types, i.e. peptide-spectrum matches (PSM), peptides, amino acid sequences (ignoring possible modifications). Summary for amino acid sequences also covers counts of instances that are represented in both non-modified and modified variants (green).

reduced the number of considered entries in average by 40 fold. Although considering more than 500 modifications, the iterated search was computed in average in 16 minutes for a single LC-MS/MS run comprising in the order of 5000-10000 fragment ion spectra. The iterated database search approach constitutes an effective strategy to constrain the protein database to its relevant entries and therefore enables to efficiently consider a vast set of modifications.

The more modifications a database search considers the more frequent the case occurs where different peptide variants equally well match to a single fragment ion spectrum. We experienced better identification performance by consequently discarding such ambiguous peptide-spectrum matches, i.e. by scoring peptide-spectrum matches with the difference score (see section 3.4). We though note that there remains a considerable amount informative fragment ion spectra with very well matching though ambiguous peptide explanations.

3.7.2 Spectral contribution of peptide variants

We adapted the target-decoy strategy to the iterated database search scenario, enabling us to compile a set of peptide-spectrum matches at well defined false discovery rate (**Fig. 3.2**). For the *D. melanogaster* dataset, we assigned 69178 fragment ion spec-

rank	modification type	# observations
1	Isotope mass shift	2842
2	Iodoacetamide derivative	765
3	Acetylation	684
4	Pyro-glu from Q	529
5	Gln→Lys	446
6	Carbamylation	116
7	Propionaldehyde +40	115
8	Deamidation	113
9	Phosphorylation w loss	109
10	Oxidation	108

Table 3.1: Top 10 most frequent amino acid modifications discovered for the *D. melanogaster* dataset. Number of observations at 1% FDR of each modification type is reported for the complete dataset.

tra (1% FDR) to 5380 non-modified tryptic peptides (≤ 1 miscleavages). The second pass search yielded 7261 additional assignments to modified peptide variants. Comprehensively accounting for amino acid modifications in the second pass search turned out to significantly boost identification performance by contributing 9.5% of all peptide-spectrum matches.

3.7.3 Discovery of novel peptide sequences

We studied which modifications were detected by the second pass search. **Table 3.1** displays the most frequent modification types found in the *D. melanogaster* dataset. The majority of identified peptide variants were due to isotope mass shifts related to wrongly picked monoisotopic peaks. Besides other frequent modifications related to the sample preparation process, we found considerable evidence for phosphorylations. Interestingly, we frequently found the amino acid substitution from Gln to Lys. It turns out that mutating a single nucleotide at the first position of the respective codons suffices to issue this mutation. The iterative database strategy thus confirmed the occurrence of known modifications and contributed to the discovery of so far unconsidered modifications.

We investigated to what extent the iterative search strategy allowed to discover novel peptide sequences, only represented in a modified form (**Fig. 3.2**). It turns out that the second pass search contributes not less than 14% of all unique amino acid sequences. We conclude that comprehensive consideration of modifications does also significantly contribute to the discovery of novel peptide sequences predominantly present in a modified form.

3.8 Discussion

We propose a target-decoy strategy to estimate false discovery rates for iterated database searches. The iterated search strategy enables to efficiently perform database searches considering several hundred modifications at once. Our adapted target-decoy strategy makes iterative search strategies amenable to practice by providing means to control false discovery rates of peptide-spectrum matches achieved. Besides enabling to discover unexpected peptide variants per se, this approach has the potential to enhance quantitative and particularly targeted proteomics approaches by unraveling the diversity of peptide/protein variants present in the biological source [22].

The iterated target-decoy strategy is appealing due to its generic applicability. Database searching and generation are decoupled processes. Therefore, any kind of search engine flexibly allowing for modifications can be used in conjunction with this strategy.

Specification of the first pass search depends on the underlying dataset. The first pass search is designed to be computationally efficient and to give evidence for most of the proteins present in the biological source. A simple search for tryptic non-modified peptides is expected to fulfill these requirements for most shotgun proteomics datasets, including the dataset studied here. However, more care has to be taken to design the first pass search for workflows for which a considerable amount of proteins is evidenced exclusively by non-tryptic peptides [144].

Database search strategies are able to recover only peptide variants complying with the protein database and the explicitly considered modifications. Peptides including modifications not considered a priori cannot be identified. Suitable spectral alignment methods do not suffer from this limitation since they are able to discover amino acid mass shifts corresponding to novel modification discoveries [136]. As fully de novo sequencing approaches, these approaches though have to carefully deal with this additional degree of freedom in order to avoid an increased number of false positive identifications. Considering that we comprehensively accounted for all reasonable modifications in the Unimod database, we though assume that we do not suffer from the database restriction and will rarely miss to identify a peptide variant evidenced by an informative fragment ion spectrum.

We applied our approach to a dataset acquired for repeated LC-MS/MS analysis of a whole cell lysate. We found about ten percent of the peptide-spectrum matches mapping to a modified peptide variant and discovered several unexpected modifications. We expect that deep sequencing approaches involving multidimensional fractionation will reveal an even larger diversity of modifications compared to the studied dataset. It will be interesting to apply the iterated target-decoy strategy to analyze this kind of datasets and characterize their modification repertoire.

4 Protein Identification False Discovery Rates

4.1 Summary

Comprehensive characterization of a proteome is a fundamental goal in proteomics. In order to achieve saturation coverage of a proteome or specific sub-proteome via tandem mass spectrometric identification of tryptic protein sample digests, proteomic data sets are growing dramatically in size and heterogeneity. The trend towards very large integrated data sets poses so far unsolved challenges to control the uncertainty of protein identifications going beyond well established confidence measures for peptide-spectrum matches. We present *MAYU*, a novel strategy that reliably estimates false discovery rates for protein identifications in large scale data sets. We validated and applied *MAYU* using various large proteomics data sets. The data show that the size of the data set has an important and previously underestimated impact on the reliability of protein identifications. We particularly find that protein false discovery rates are significantly elevated compared to those of peptide-spectrum matches. The function provided by *MAYU* is critical to control the quality of proteome data repositories and thereby to enhance any study relying on these data sources. [1]

---

[1] Lukas Reiter and I share the first authorship on this work. Both Lukas Reiter and me have equally contributed to the main ideas and their implementation. Sabine P. Schrimpf has generated the *C. elegans* dataset that constitutes the main data source for our study. Marko Jovanovic performed most of the validation experiments with the dirty peptides. Alexander Schmidt helped with the mass spectrometrical analysis of the dirty peptides and has contributed the *L. interrogans* and *S. pombe* dataset. The project was performed under the supervision of Joachim M. Buhmann, Michael O. Hengartner and Ruedi Aebersold. We thank Vinzenz Lange, Christian Müller, Lukas Müller, Thomas Fuchs and Bernd Bodenmiller for careful reading of the manuscript. Further we thank James Eddes, Christian Panse, the Center for Model Organism Proteomes (C-MOP) and the Functional Genomics Center Zurich (FGCZ) for support.

4.2 Introduction

An explicit goal of proteomics is the complete description of a proteome and the measurement of its response to perturbations [3]. Over the last few years advances in mass spectrometry based proteomics have achieved a significant increase in proteome coverage [138, 103, 100, 51, 75, 15, 5, 35, 58, 119]. The volume and heterogeneity of proteomic data required to substantially map out a proteome pose considerable challenges to assess the confidence of peptides and proteins that are inferred from the collected fragment ion spectra [95]. While a number of statistical tools and strategies have been developed to assess the error rate of peptide-spectrum matches (PSM), estimation of the false discovery rate (FDR) of protein identifications in large datasets remains an unresolved problem. This study presents a probabilistic framework and software that addresses this issue.

The most extensive proteome coverage has generally been realized by a strategy typically referred to as shotgun proteomics. Briefly, proteins are extracted from their biological source, enzymatically digested and optionally fractionated. The resulting peptide mixtures are then analyzed by tandem mass spectrometry (MS/MS). Peptide and protein identities are inferred by computational analyzes of the acquired tandem mass spectra. The data generated by shotgun proteomics experiments are highly redundant, i.e. a subset of the peptides present is repeatedly and preferentially selected for fragmentation and identified. In contrast, other subsets of peptides, e.g. those derived from low abundance proteins are more difficult to detect and a large number of fragment ion spectra have to be acquired to increase the likelihood of their detection [15, 43, 86]. Consequently, proteomic studies aiming at extensive proteome coverage generate very large data sets consisting of up to millions of fragment ion spectra.

Shotgun proteomics experiments essentially aim at the compilation of a set of reliable protein identifications covering the proteome as extensively as possible. This goal is achieved by firstly inferring a set of protein identifications (inference) and secondly assessing the reliability of these identifications (FDR estimation) (**Fig. 4.1**). Briefly, fragment ion spectra are assigned to peptide sequences by generating peptide-spectrum matches (PSMs) using one of a range of database search engines (e.g. Mascot, Sequest, X!Tandem) [97]. Second, protein identifications are inferred from the PSMs by assembling the identified peptide sequences into proteins [110, 95]. Protein identifications are thus defined as assemblies of PSMs whose peptide sequences map to the same protein (**Fig. 4.1**).

Neither PSMs nor protein identifications are perfect. Therefore it is essential to control the reliability of PSMs and protein identifications. Various approaches have been developed to estimate the reliability of PSMs [72, 94, 41, 67]. FDR [9], i.e. the expected fraction of false positive assignments, have become a widely used measure for reliability of PSMs. FDR for PSMs can be confidently estimated by means of decoy database search strategies in which the acquired fragment ion spectra are searched against a chimeric protein database containing all (target) protein sequences possibly present in the sample analyzed and an equal number of nonsense (decoy) sequences. Target-decoy strategies are particularly appealing since they constitute a generic and independent approach to validate PSMs generated by any type of identification strategy.

Protein identifications, i.e. assemblies of PSMs, are the biologically relevant outcome of a shotgun experiment. Therefore it is highly desirable to directly control the quality of protein identifications, for example in terms of FDR. Deriving FDR for protein identifications though is not as obvious as determining FDR for PSMs. Because protein identifications are defined by assemblies of PSMs, errors determined at the PSM level propagate to the protein identification level in a non trivial manner. Therefore controlling quality on the level of PSMs does not ensure quality at the (biologically relevant) level of protein identifications. This issue has so far not been appropriately appreciated, since the distinction between PSMs and protein identifications is frequently blurred in the literature. An estimate of protein identification FDR, i.e. the expected proportion of false positive protein identifications, has to account for false positive and true positive PSMs distributing differently across the protein database. While false positive PSMs comparably distribute over all entries in the database [41], true positive PSMs map exclusively to the smaller subset of proteins being present in the biological sample. As a result, protein identification FDR in practice is larger than the PSM FDR [1].

Number, frequency and size and heterogeneity of proteomic data sets steadily increase [138, 103, 100, 51, 75, 15, 5, 35, 119]. Available approaches for protein identification focus on the protein inference task and provide reasonable to good error estimates for individual experiments (typically 10-100 LC-MS/MS runs), the complexity level at which most proteomics studies operate [85, 96, 1, 139, 108]. However, none of these approaches reliably quantifies the confidence in protein identifications in very large, integrated data sets (typically 100 or more LC-MS/MS runs), e.g. in terms of quantifying FDR for protein identifications (**Fig. 4.1**). To date, protein identifications in large proteomics data sets are compiled according to heuristic criteria for which so far no quantitative confidence measures like FDR have been derived at the protein identification level [138, 145, 19, 51, 15].

To close this gap, we developed a generic strategy enabling for the first time to quantify the confidence in protein identifications obtained from a wide range of inference methods (**Fig. 4.1**) in data sets of all sizes, especially in large to very large data sets. We refer to this approach as *MAYU* (no acronym). Our approach extends the well established target-decoy strategy designed to estimate FDR at PSM level [41, 67] to the level of protein identifications, i.e. defined assemblies of PSMs (**Fig. 4.1**). We applied *MAYU* to three different data sets varying in instrumentation and species. We found that data set size has a previously underestimated impact on protein identification FDR. The strategy developed and the tool that implements it could therefore be of critical importance for the generation and quality control of large proteome datasets and data bases. The *MAYU* software and a manual are publicly available for download.

4.3 Methods

4.3.1 Spectral data and database searching

We analyzed three different data sets, from studies varying in MS instrumentation and underlying organism. All studies were based on multi-dimensional fractionation techniques and comprised samples from *C. elegans* [119], *L. interrogans* and *S. pombe*. While the first data set was acquired on a low resolution LTQ instrument, the latter two were acquired on a high mass accuracy LTQ-FT instrument. The *C. elegans* project is part of the Center for Model Organism Proteomes (C-MOP) initiative (http://www.mop.unizh.ch/); the *C. elegans* proteome data are available on PeptideAtlas (http://www.peptideatlas.org/) [37]. We searched each data set against a composite target-decoy database using Turbo Sequest [42] and Sequest on a Sorcerer machine (Sorcerer-SEQUEST, 3.10.4 release). The search results were transformed to the pepXML format and further processed using the Trans Proteomic Pipeline [71] to

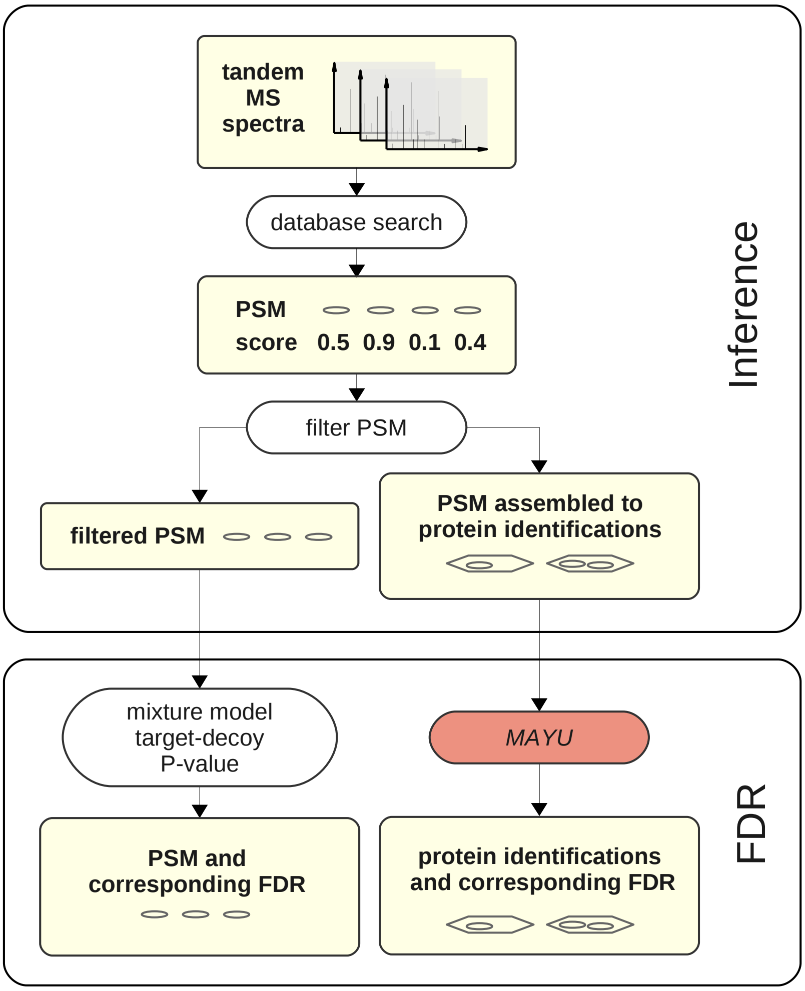

tandem MS spectra
database search
PSM
score 0.5 0.9 0.1 0.4
filter PSM
filtered PSM
PSM assembled to protein identifications
Inference
mixture model target-decoy P-value
MAYU
PSM and corresponding FDR
protein identifications and corresponding FDR
FDR

Figure 4.1: Protein inference and false discovery rate estimation. Tandem mass spectra are searched against a sequence database, where each spectrum is assigned to the best matching, i.e. highest scoring peptide sequence. These assignments are referred to as peptide-spectrum matches (PSMs). The PSM can then be filtered according to their score. The quality of the filtered PSM is usually specified in terms of PSM false discovery rates (PSM FDR). Score cutoffs for PSM are usually selected according to a user-defined maximal PSM FDR. Alternatively the filtered PSM can firstly be assembled to protein identifications. The quality of the assignments is then assessed on the level of protein identifications. *MAYU* provides a strategy to quantify this quality in terms of protein identification FDR. Compared to PSM FDR, the protein identification FDR is a more informative quality measure since it operates on biological entities of interest, i.e. proteins.

the level of PeptideProphet [72] in units of experiments. The pepXML files were then further analyzed with the *MAYU* software. If a peptide existed in more than one protein sequence the hit was associated with one protein representing the gene locus [119], see also [15, 5].

4.3.2 Target-decoy database generation

We performed all the database searches using a concatenated target-decoy database [41]. As target database for the *C. elegans* data set we chose wormpep170 . For the *L. interrogans* data set we used NC_005824 and for the *S. pombe* data set we respectively used 78.S_pombe . As decoy databases we used the reversed sequences of the target database.

For the decoy database type comparison, we further generated ten different decoy databases by sampling from a zeroth order Markov model with amino acid frequencies and protein length distribution gathered from the target database. Since randomizing of redundant sequences leads to a decoy database being effectively larger, i.e. featuring a larger amount of non-redundant sequences, than the target database [41], we corrected the target database prior to sampling of amino acids. This was done for the splice variants by removing random amino acids from the non main splice variants accordingly (with the main splice variant being the alphabetically first). If there were groups of identical protein sequences all but one of these were deleted.

4.3.3 Estimate of protein identification FDR

The set of PSMs produced in the course of a proteomics experiment give rise to protein identifications. A set of PSMs mapping to the same protein sequence defines a protein identification. A protein identification is considered to be true positive, if it contains at least one true positive PSM, and false positive if all of its PSMs are false positive. This particularly implies that a protein identification that contains false positive PSMs is not necessarily false positive. In order to estimate protein identification FDR we estimate the expected number of false positive identifications within a set of protein identifications that has been assembled from a user-defined set of PSMs, e.g. from the set of PSMs at FDR=0.01.

Based on the well established assumption that false positive PSMs equally likely map to either target or decoy database, we used the number of PSMs mapping to the decoy

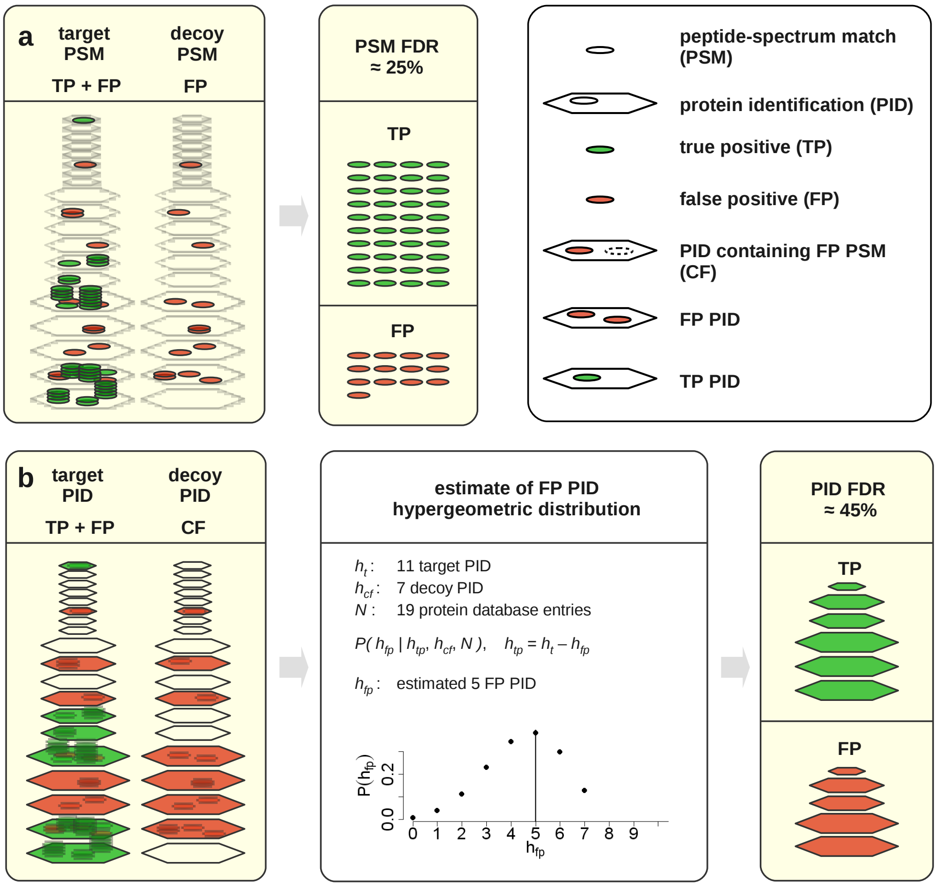


Figure 4.2: *MAYU* protein identification false discovery rate estimation. Estimation of peptide-spectrum match (PSM) false discovery rate (FDR) using a target-decoy strategy **(a)** and protein identification (PID) FDR by *MAYU* **(b)**. PSM in the target database can be false positive (FP) / true positive (TP). The PSM FDR (the expected fraction of false positive target PSM) can be estimated with the number of decoy PSM being false positive by definition. The PSM FDR is currently the major measure used for quality control of mass spectrometric data sets **(a)**. The derivation of protein identification FDR has to account for protein identifications containing false positive PSMs (CF) though not being false positive protein identifications **(b,** two proteins**)**. In order to estimate the expected number of true positive (h_{tp}) and false positive (h_{fp}) protein identifications, *MAYU* implements a hypergeometric model that takes the number of target (h_t) and decoy (h_{cf}) protein identifications and the total number of protein entries in the database (N) as input.The hypothetical example illustrates that PSM FDR (25%) and protein identification FDR (45%) can differ largely.

database as an estimate for the number of false positive PSMs mapping to the target database. The PSM FDR is then estimated as the ratio of the number of PSMs pointing to decoy- and target database, respectively. Considering that target and decoy database share the same protein length distribution, the expected number of protein identifications containing false positive PSMs can be estimated analogously using the number of protein identifications mapping to the decoy database (**Fig. 4.2b**).

We then estimate the expected number of false positive protein identifications given the inferred number of protein identifications containing false positive PSMs. If we assume that protein identifications containing false positive PSMs uniformly distribute over the target database, then the number of false positive protein identifications is hypergeometrically distributed (**Fig. 4.2b,** middle panel). See also section 4.3.4 for details.

This relation can be seen by regarding the protein database as an urn containing balls, each representing a protein entry. Those balls that correspond to the true positive protein identifications are green while the remaining ones are white. In the urn analogy, observing k false positive protein identifications then corresponds to hitting k white balls after drawing (without replacement) as many times from the urn as we have protein identifications containing false positive PSMs.

Having specified the probability distribution of the number of false positive protein identifications as the hypergeometric distribution, the expected number of false positive protein identifications then follows as the probability weighted average (expectation value). The estimate of protein identification FDR is computed as the ratio of expected number of false positive protein identifications and the total amount of protein identifications mapping to the target database.

We also estimated single hit FDR based on the FDR estimate for the complete set of protein identifications by applying Bayes Law. Single hit FDR is thus obtained by multiplying the FDR of the complete set of protein identifications with the fraction of single hits among the decoy protein identifications divided by the fraction of single hits among the target protein identifications.

In section 4.3.4 we also provide a formal statement of the underlying assumptions and a formal derivation of the individual estimates.

4.3.4 Derivation of the protein identification FDR estimate

The set of PSMs produced in the course of a proteomics experiment give rise to protein identifications. A set of PSMs mapping to the same protein sequence defines a protein identification. In the following we refer to the set of all protein identifications as H, the subset mapping to the target database P_t as H_t and its complement as H_d. We distinguish three types of protein identifications, i.e. (1) TP identifications, which all together we denote H_{tp}. A protein identification is considered to be TP, if it contains at least one TP PSM. While the second type (2) covers the set H_{fp} of FP protein identifications mapping to P_t, the complementing set with its identifications projecting to the decoy database P_d equals H_d. A protein identification is considered to be FP, if all of its PSM are FP. As the third type (3) we introduce the set H_{cf} that is composed of all protein identifications in P_t each containing FP PSM. Note that elements of H_{cf} can be TP as well as FP. The size of the defined sets shall be denoted by lowercase letters, as for instance $|H| = h$.

Making the reasonable assumption that FP PSM equally likely map to either target or decoy database, it is straightforward to estimate the expected value of FP PSM mapping to P_t with the number of PSM pointing to P_d . According to the definition of false discovery rates [9], we can estimate the PSM FDR as the ratio of the number of PSM pointing to P_d and P_t respectively. Considering that target and decoy database share the same protein length distribution, the expected value for h_{cf} can be estimated analogously with h_d. Note that h_{cf} does not necessarily equal h_{fp}.

In order to determine the FDR for protein identifications, we firstly calculate the conditional expectation value for $E\left[h_{fp} \mid h_t, h_d, \theta_{\text{exp}}\right]$ for the number of FP protein identifications given the proteomics experiment characterized by parameters θ_{exp} and its outcome h_t, h_{cf}. Amongst others, θ_{exp} particularly includes parameters related to the target protein database, such as the number of protein entries N. By application of Bayes formula and by assuming $P(h_{tp} \mid h_{cf}, \theta_{\text{exp}})$ and $P(h_t \mid h_{cf}, \theta_{\text{exp}})$ to be uniform and $h_d = h_{cf}$, $E\left[h_{fp} \mid h_t, h_d, \theta_{\text{exp}}\right]$ evaluates as follows.

$$
\begin{aligned}
E\left[h_{fp} \mid h_t, h_{cf}, \theta_{\text{exp}}\right] &= \sum_{h_{fp}} h_{fp} \cdot P(h_{fp} \mid h_t, h_{cf}, \theta_{\text{exp}}) && (4.1)\\
&\overset{h_{tp}=h_t-h_{fp}}{=} \sum_{h_{fp}} h_{fp} \cdot P(h_{tp} \mid h_t, h_{cf}, \theta_{\text{exp}}) && (4.2)\\
&= \sum_{h_{fp}} h_{fp} \cdot \frac{P(h_t \mid h_{tp}, h_{cf}, \theta_{\text{exp}}) P(h_{tp} \mid h_{cf}, \theta_{\text{exp}})}{P(h_t \mid h_{cf}, \theta_{\text{exp}})} && (4.3)\\
&\overset{h_{tp}=h_t-h_{fp}}{=} \sum_{h_{fp}} h_{fp} \cdot \frac{P(h_{fp} \mid h_{tp}, h_{cf}, \theta_{\text{exp}}) P(h_{tp} \mid h_{cf}, \theta_{\text{exp}})}{P(h_t \mid h_{cf}, \theta_{\text{exp}})} && (4.4)\\
&= \sum_{h_{fp}} h_{fp} \cdot P(h_{fp} \mid h_{tp}, h_{cf}, \theta_{\text{exp}}) \cdot \frac{N - h_{cf} - 1}{N+1} && (4.5)
\end{aligned}
$$

Let us assume for a moment that all protein sequences in the target and decoy database have the same size. As the probability of a FP PSM mapping to a certain protein sequence scales linearly with its size, each entry in P_t would be equally likely to be part of H_{cf}. Thus, protein identifications containing FP PSM would be uniformly distributed across P_t. Accordingly, $P(h_{fp} \mid h_{tp}, h_{cf}, \theta_{\text{exp}})$ would follow the hypergeometric distribution, where h_{fp} is modeled as a random variable representing the number of successful hits of a non-TP-identified protein in a sequence of h_{cf} draws without replacement from the N entries in P_t.

Clearly, the initial assumption about the singular size distribution does not hold for biological protein databases. So as to compile an estimate for $E\left[h_{fp} \mid h_t, h_d, \theta_{\text{exp}}\right]$ from subgroups closely meeting this assumption, we have partitioned $P = P_t \cup P_d$ into subsets P_i of protein sequences of similar size. In this context, protein sequence size is defined as number of tryptic peptides from in silico digestion (400-6000 Da, ≤ 2 missed cleavages). Variables $h_{t,i}$, $h_{cf,i}$, $h_{tp,i}$, $h_{fp,i}$, N_i are defined for each P_i in analogy to those for P. By applying the foregoing argument we approximate $E\left[h_{fp} \mid h_t, h_d, \theta_{\text{exp}}\right]$ as follows

$$
\begin{aligned}
\hat{E}\left[h_{fp} \mid h_t, h_{cf}, \theta_{\text{exp}}\right] &= \sum_i \hat{E}\left[h_{fp,i} \mid h_{t,i}, h_{cf,i}, \theta_{\text{exp}}\right] && (4.6)\\
&= \sum_i \sum_{h_{fp,i}} h_{fp,i} \cdot P(h_{fp,i} \mid h_{tp,i}, h_{cf,i}, \theta_{\text{exp}}) \cdot \frac{N_i - h_{cf,i} - 1}{N_i + 1} && (4.7)
\end{aligned}
$$

where

$$P(h_{fp,i} \mid h_{tp,i}, h_{cf,i}, \theta_{\text{exp}}) = \frac{\binom{N_i - h_{tp,i}}{h_{fp,i}} \binom{h_{fp,i}}{h_{cf,i} - h_{fp,i}}}{\binom{N_i}{h_{cf,i}}} \tag{4.8}$$

We have assessed this approximation for $E\left[h_{fp} \mid h_t, h_d, \theta_{\text{exp}}\right]$ by confirming quick convergence in experiments with various partitions featuring increasing size homogeneity within the subsets (**Fig. 4.3a**).
We obtain the final estimate for FDR by appropriately inserting $\hat{E}\left[h_{fp} \mid h_t, h_{cf}, \theta_{\text{exp}}\right]$.

$$\text{pFDR} = \frac{\hat{E}\left[h_{fp} \mid h_t, h_{cf}, \theta_{\text{exp}}\right]}{h_t} \tag{4.9}$$

4.3.5 Simulation of non-uniformly distributed false positive PSM

We performed simulation studies to assess the robustness of *MAYUs* FDR estimates. The outcome of proteomic experiments was simulated with varying types of distributions for false positive PSM. For each simulation we first distributed a fixed number of true positive protein identifications across the protein database (comprising N entries). We distributed false positive PSM according to a truncated exponential distribution ($\sim \lambda e^{-\lambda x}$). The rate parameter $\lambda = 1/(u \cdot N)$ was chosen for different degrees of uniformity u. For each simulation we determined the true protein identification FDR and its *MAYU* estimate. For each seed of distributed true positive protein identifications 50 simulations were performed and the average relative FDR deviation reported.

4.3.6 Validation of single hit FDR using isoelectric point information

To validate our model we independently derived an FDR estimate for single hits and compared this value to the estimation of *MAYU*. We used 67 LC-MS/MS runs of experiment 15 of the *C. elegans* data set where peptides were fractionated by isoelectric focusing according to their isoelectric point (pI) [119]. We used the standard deviation $\sigma_{\Delta pI}$ of isoelectric point deviations ΔpI as a quality measure for a set H of PSMs,

$$\Delta pI(i) = pI_{pr}(i) - pI_{ex}(i) \tag{4.10}$$

$$\sigma_{\Delta pI}(H) = \sqrt{\frac{1}{|H|}\sum_{i \in H} (\Delta pI(i) - m_{\Delta pI}(H))} \tag{4.11}$$

where $pI_{pr}(i)$ is the isoelectric point of a PSM i predicted by Bioperl [125]. $pI_{ex}(i)$ corresponds to the experimentally measured isoelectric point of a PSM i, determined as the mean isoelectric point of the high confident peptides of the respective LC-MS/MS run (PSM FDR 0.01). $m_{\Delta pI}(H)$ denotes the mean of $pI_{pr}(i)$ for PSM i in H.

In order to specify the correspondence of PSM FDR and $\sigma_{\Delta pI}$, we generated a calibration curve with sets $H_{c,\mathrm{x}}$ of PSMs of defined PSM FDR x. These sets were compiled from high confident target hits with zero FDR complemented with an appropriate amount of decoy hits to yield the designated PSM FDR. The corresponding decoy hits were sampled from a set of target-decoy PSMs featuring the designated PSM FDR. Standard deviations were computed using 20 bootstrap samples.

We estimated FDR for the set $H_{s,x}$ of single PSM protein identifications (single hits) with PSM FDR x by computing $\sigma_{\Delta pI}(H_{s,x})$ and reading out the corresponding FDR by linear interpolation of the calibration curve.

For very small PSM FDR x we observed a significant shift of $\sigma_{\Delta pI}(H_{s,x})$ compared to the calibration curve. Arguing that TP single hit peptides focus better (see **Fig. 4a**) in the isoelectric focusing step, we adjust $\sigma_{\Delta pI}(H_{s,x})$ to read out the FDR. The unadjusted initial FDR estimate FDR_{ini} is used to weight the adjustment according to the initially estimated TP single hits.

$$\sigma_{\Delta PI}^{adj} = \sigma_{\Delta PI}(H_{s,x}) + (\sigma_{\Delta PI}(H_{c,0}) - \sigma_{\Delta PI}(H_{s,0})) \cdot (1 - FDR(H_{s,x})) \tag{4.12}$$

4.3.7 Validation of single hit FDR using synthetic peptides

We ordered three different sets of synthetic peptides synthesized on a microscale using the SPOT-synthesis technology [140, 63]. These sets were compiled as follows:

- As positive control we randomly selected 50 peptide sequences that were identified with at least 100 PSM with a PSM FDR of zero in the search results of the complete *C. elegans* data set.
- As negative control we randomly selected 50 peptide sequences from decoy proteins with a PSM FDR of 0.01 in the search results of the complete *C. elegans* data set.
- As peptides of interest we randomly selected 150 peptide sequences whose PSM in the search results of the complete *C. elegans* data set were single hits.

The search results of the complete *C. elegans* data set were processed as follows. The PSM of the complete *C. elegans* data set were extracted. Ambiguous peptides, peptides longer than 18 amino acids and cysteine containing peptides were removed. *MAYU* was run on the remaining PSM and all PSM corresponding to PSM FDR of 0.01 were extracted. From these PSM the three sets were selected as described above.

For all the 250 synthetic peptides an inclusion list was generated [118] and measured on an LTQ-FT instrument such that the precursors corresponding to the selected PSM were targeted. The spectra were searched using SEQUEST on a Sorcerer machine (Sorcerer-SEQUEST, 3.10.4 release) and filtered for an FDR of 0.01 (protein identification FDR of 0.01 estimated by *MAYU*). The resulting tandem mass spectra were then normalized to total ion current and compared to the analogously processed tandem mass spectra of the *C. elegans* data set. Each peptide was attributed to a score comparing the corresponding *C. elegans* and inclusion list fragment ion spectrum, i.e. summed difference of normalized intensities. We trained a Gaussian mixture model for TP/FP score distributions by fitting each component to the positive and respectively negative controls and then used the mixture model to estimate the expected number of FP single hits for the peptides of interest.

4.3.8 MAYU analysis on ProteinProphet protein identifications.

ProteinProphet was run on the pepXML files using runprophet from the trans proteomic pipeline [71] and target/decoy protein identifications of ProteinProphet were used as input for *MAYUs* protein identification FDR calculation.

4.4 Results

4.4.1 MAYU - FDR for protein identifications.

MAYU implements a target-decoy strategy to estimate FDR for a set of protein identifications compiled from a selection of PSMs. Target-decoy strategies to estimate FDR of PSMs rely on the well established assumption that false positive PSMs uniformly distribute between target and decoy database. Consequently, PSM FDR is estimated as the ratio of PSMs mapping to the decoy and target database, respectively (**Fig. 4.2a**) [41]. *MAYU* extends this approach to estimate FDR for protein identification*s*, i.e. assemblies of PSMs (**Fig. 4.2b**).

Prior to *MAYU* analysis, PSMs are gathered by a target-decoy database search and processed by a protein inference engine, finally yielding a set of target and decoy protein identifications (**Fig. 4.1**). Note that *MAYU* analysis solely aims to estimate the false discovery rate of a set of already inferred protein identifications. *MAYU* analysis is applicable to the results of any search and protein inference engine (**Fig.4.8, 4.7**). The following describes the *MAYU* workflow.

MAYU processes the supplied list of protein identifications to estimate their FDR. We define a false positive protein identification as being exclusively supported by false positive PSMs and no true positive PSMs. Assuming that false positive PSMs distribute uniformly over the chimeric database, the number of the decoy protein identifications provides an estimate of target protein identifications containing false positive PSMs (seven in the example shown in **Fig. 4.2b**). However, the actual number of false positive protein identifications (five in **Fig. 4.2b**) is lower than this (naïve target-decoy) estimate, as some proteins (two in **Fig. 4.2b)** in the target database will contain both true and false positive PSMs.

MAYU uses the number of protein identifications in the target and decoy database and the total number of protein entries in the database (11, 7 and 19 respectively in **Fig. 4.2b**) to estimate the expected number of false positive protein identifications in the target database (see sections **4.3.3, 4.3.4**).

In summary, starting from a shotgun proteomic data set searched against a target-decoy database, the *MAYU* workflow provides comprehensive and quantitative error analysis

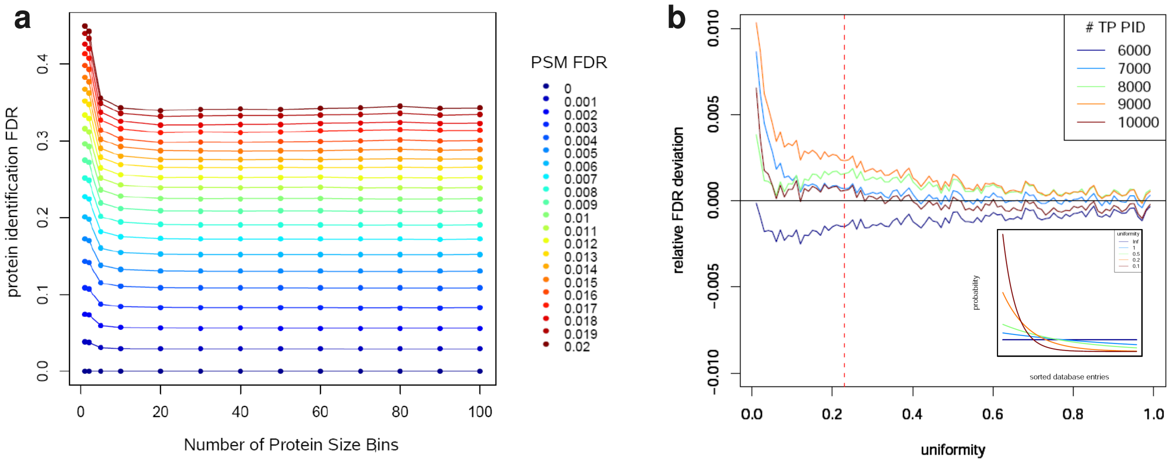


Figure 4.3: Robustness of the false discovery rate estimates of *MAYU*. *MAYU* imposes the assumption that protein identifications containing false positive PSM uniformly distribute over the protein database. To closely meet this assumption *MAYU* operates on a partition of the protein database into subsets comprising proteins of similar size. The figure depicts how the size of the partition affects the protein identification FDR estimates for different sets of PSM defined over the complete *C. elegans* data set (**a**). Partitions with more than ten size bins yield stable FDR estimates and therefore seem to yield the desired protein size homogeneity. **(b)** **S**imulation studies for the complete *C. elegans* set where we explicitly distributed false positive PSM according to distributions increasingly deviating from uniformity (see 4.3.5). We assessed the accuracy of the *MAYU* estimate in terms of relative deviation from the true FDR depending on the degree of uniformity of the false positive PSM distribution. The inserted plot exemplarily depicts four distributions of varying uniformity. We observe that the *MAYU* estimates do not deviate more than 1% from the true FDR (e.g. 0.2 ± 0.002%), even for considerable deviations from the uniformity assumption.

for protein identifications.

4.4.2 Validation of protein identification FDR estimate

We validated the *MAYU* approach in various ways. First we assessed the robustness of the FDR estimates under violations of the underlying assumptions. Second, we validated the *MAYU* FDR estimates by comparing them with an independent approach that estimates single PSM protein identifications (single hits) FDR based on isoelectric point (pI) information from an isoelectric focusing experiment (67 LC-MS/MS runs, *C. elegans* data set). Third, we validated *MAYUs* FDR estimates by confirming single hit FDR using synthesized peptides corresponding to single hits in the complete *C. elegans* data set (1,305 LC-MS/MS runs).

We studied the robustness of our FDR estimates under deviations from the assumptions underlying the hypergeometric model. *MAYUs* protein identification FDR relies on statistics gathered from a target-decoy search, most importantly the number of protein identifications mapping to the decoy database. Following [41], we assume this number to equal the number of target protein identifications containing false positive PSM. In order to estimate protein identification FDR with the hypergeometric model, we further assume that protein identifications containing false positive PSM uniformly distribute over the protein database. To closely meet this assumption *MAYU* partitions the protein database into subsets whose entries feature similar size. The protein identification FDR estimate is obtained by applying the hypergeometric model to each of these subsets (see 4.3.4). The granularity of the partition does not affect the FDR estimate as long as more than ten size bins are considered (**Fig. 4.3a**). We further conducted simulation studies to assess how deviations from the uniformity assumption influence the *MAYU* FDR estimate. For each simulation we assumed a fixed number of true positive protein identifications and distributed false positive PSM according to a truncated geometric distribution. For each simulation we determined the true protein identification FDR and compared with the *MAYU* estimate (**Fig. 4.3b**). We observe that the *MAYU* estimates are not compromised, even for considerable deviations from the uniformity assumption.

We further validated the *MAYU* FDR estimates for (non-simulated) experimental data. *MAYUs* protein identification FDR estimates are ideally validated on a test data set derived from a well-defined mix of proteins. In order to capture the relevant phenomena complicating protein identification FDR estimates, a protein reference sample of defined composition covering a significant proportion of the entire protein database (e.g. 10%) would be required. Unfortunately, such a test data set is not available and would be exceedingly difficult to construct.

We therefore validated *MAYU* on a large data set providing additional information that allows us to independently derive single hit FDR gathered from an experiment of the *C. elegans* data set where peptides were separated by isoelectric point (pI) using isoelectric focusing (experiment 15, 67 LC-MS/MS runs).

We used the standard deviation of PSM pI deviations as a quality measure for a set of PSMs. This measure grows with the fraction of false positive PSM, since their pI

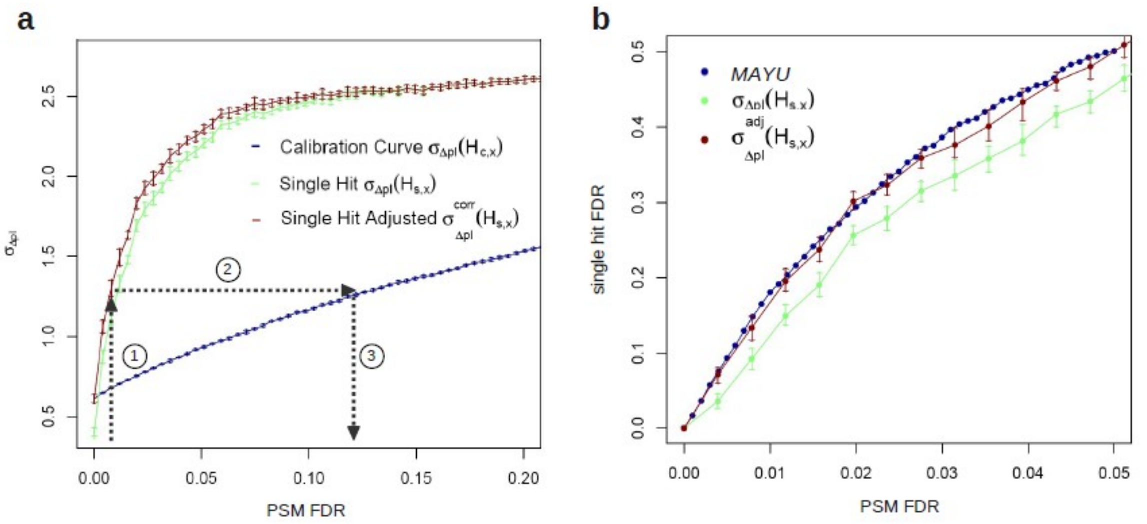


Figure 4.4: Validation of the false discovery rate estimates of *MAYU*. We validated the *MAYU* false discovery rate (FDR) using two data sets of different size and with two distinct methods. We used experiment 15 (67 LC-MS/MS runs) of the *C. elegans* data set where experimental isoelectric point (pI) information of peptides were available (**a, b**). Using experiment 15 we derived a measure of the discrepancy between the measured and the computationally predicted pIs of peptides $\sigma_{\Delta pI}$ (see 4.3.6). Sets of peptide-spectrum matches (PSMs) filtered with increasing PSM FDR up to 0.2 show an increase in $\sigma_{\Delta pI}$ (**a**, blue curve). $\sigma_{\Delta pI}$ for only the single hits is significantly higher than for all PSM over the complete range indicating that the single hit FDR is much higher compared to the PSM FDR (**a**, green and blue curve). The error bars specify standard deviations from 20 bootstraps. Using $\sigma_{\Delta pI}$ of all PSMs as a calibration curve we could estimate the single hit FDR assuming that true positive (TP) single hits are not generally different from the rest of PSMs in terms of pI (**b**). We also calculated a corrected single hit FDR (**a, b** brown curve) by making the reasonable assumption that TP single hit peptides focused better in the isoelectric focusing experiment (**a**, see offset of $\sigma_{\Delta pI}$ at zero PSM FDR between the single hits and all PSMs). We found strong consistency between the *MAYU* and independent method based on peptide pI information (**b**).

values distribute over the complete pI range, in contrast to those of true positive PSM clustering closely around the measured pI. By exploiting this phenomenon, we related pI information associated to PSM evidencing single hits to their quality in terms of FDR (4.3.6, **Fig. 4.4 a,b**). Since for single hits, PSM FDR is equivalent to the single hit FDR, we obtain a protein identification FDR estimate for the set of single hits.

MAYU analysis yielded a single hit FDR about ten fold higher than the corresponding PSM FDR of the complete set of protein identifications. We find the surprisingly high

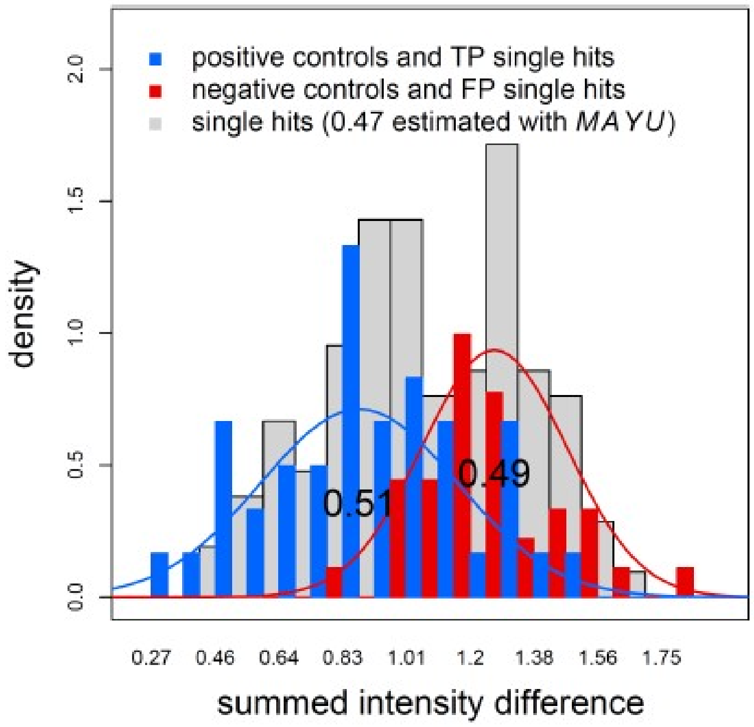


Figure 4.5: Validation of MAYU estimates with synthetic peptides. We ordered three sets of synthetic peptides corresponding to randomly picked PSMs of three different classes from the complete *C. elegans* data set (see 4.3.7). We recorded tandem mass spectra of the synthetic peptides in a targeted way using inclusion lists and compared them to the corresponding spectra of the *C. elegans* data set (**c**). 35 peptides of the negative control (**c**, red), 42 peptides of the positive control (**c**, blue) and 114 peptides of our peptides of interest (**c**, gray) were identified with a stringent cutoff. We could nicely separate the distributions of positive and negative controls using the summed intensity difference (see 4.3.7). Based on a Gaussian mixture model of the positive and negative controls we estimated the fraction of false positives of our peptides of interest as 0.49 which is very consistent with the estimated 0.47 of *MAYU*.

single hit FDRs obtained by *MAYU* analysis to be independently confirmed by the pI deviation method (**Fig. 4.4b**). We argue that the protein identification FDR estimates produced by *MAYU* are accurate in the context of typical proteomic studies in the range of 50 LC-MS/MS runs.

We also wanted to validate *MAYUs* FDR applied to the complete *C. elegans* data set, where the error propagation effects from PSM FDR to protein identification FDR are most pronounced. Since there was no pI information available for all 20 experiments we employed a different strategy. We used synthetic peptides and compared their tandem mass spectra to the tandem mass spectra from the *C. elegans* data set (see 4.3.7). We generated three sets of peptides: positive controls, negative controls and peptides of interest. The analysis was performed on the complete data set filtered with a PSM FDR of 0.01.

We recorded tandem mass spectra of the synthetic peptides in a targeted way using inclusion lists and compared them to the corresponding spectra of the *C. elegans* data set. 35 peptides of the negative control (**Fig. 4.5**, red), 42 peptides of the positive control (blue) and 114 peptides of our peptides of interest (gray) were identified.

We report the summed intensity differences distributions and observe that the peptides of interest show a bimodal distribution with the two apexes very close to the apexes of the positive and negative controls. Based on a Gaussian mixture model of for positive and negative controls we estimated the fraction of false positives of our peptides of interest as 0.49 which is very consistent with the estimated 0.47 of *MAYU.*

Other recent studies confirm this considerable error accumulation among single hits [58].

We conclude that *MAYUs* estimates are accurate in the context of a very large data set (1,305 LC-MS/MS runs). Considering the results obtained from the pI deviation method, we conclude that MAYU achieves accurate protein FDR estimates that scale well with data set size.

4.4.3 Comparison of decoy database types

There is an ongoing debate which type of decoy database to ideally choose to accurately estimate false discovery rates. [41] have convincingly shown that all types of typically used decoy database types achieve the same for PSM FDR estimates. We present results for a comparison of protein identification FDR estimates using either a reversed or zeroth order Markov model decoy database. These results confirm the situation encountered at the level of PSM (**F**ig. 4.6). Estimates based on both types of decoy database coincide across the whole range of protein identification FDR.

We also tested whether peptides present in both target- and decoy database compromise our protein identification false discovery estimates. [41] have shown that such peptides

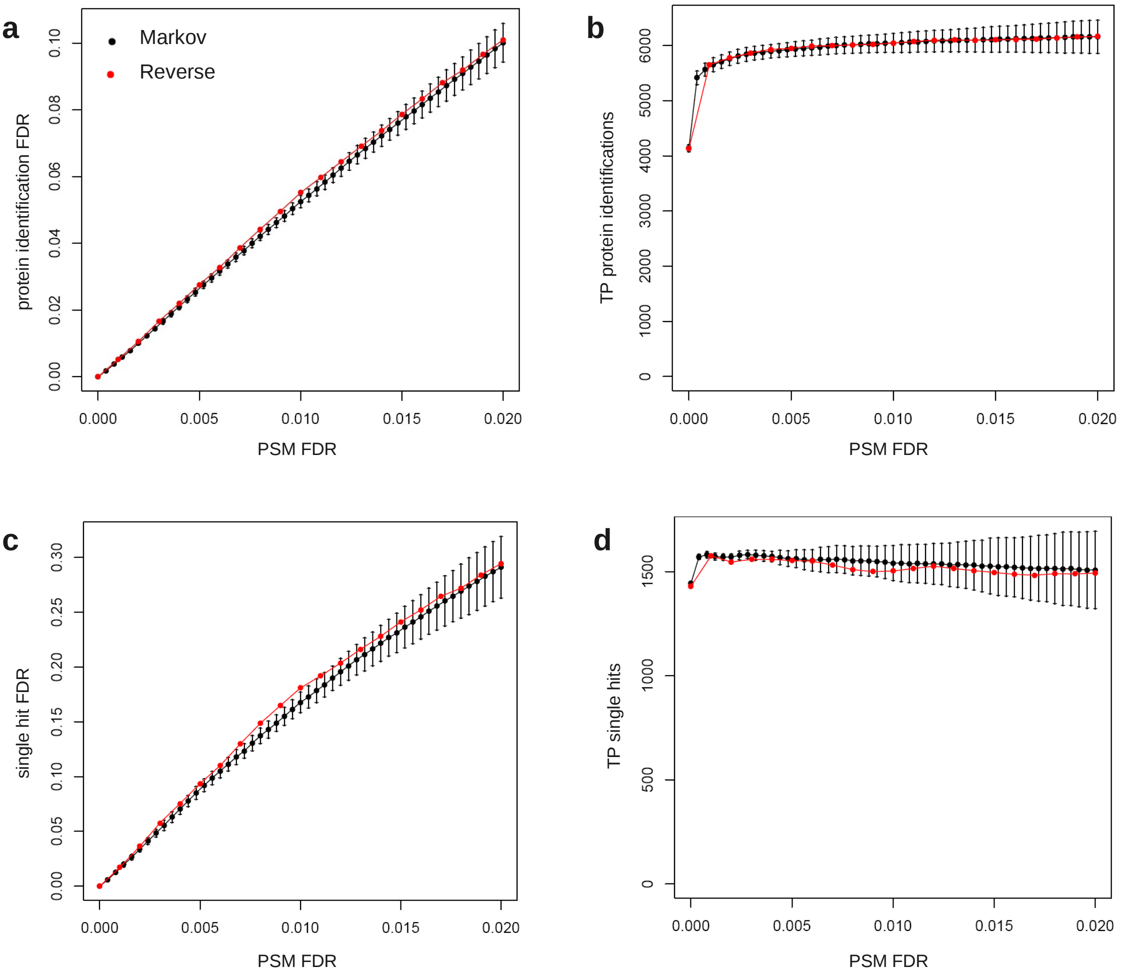


Figure 4.6: *MAYU* protein identification false discovery rates are little influenced by the choice of decoy database. Protein identification false discovery rate (FDR) estimates are stable with respect to the underlying decoy database. We show this by repeated database searches of the *C. elegans* data set, each based on a different decoy database (see section**4.3.2**). Relative standard deviation of the resulting FDR estimates in any case fell below 10% (**a**,**c**). We observe a slight trend towards larger variability of the corresponding single hit FDR estimates, revealing the limitations of the non-parametric estimates of protein identification property distributions (**c**, **d**).

occur exceedingly rare and therefore are not expected to have a major impact on our estimates. We analyzed the *C. elegans* dataset while explicitly excluding these few peptides. Our results are summarized in **Fig. 4.7** and confirm the expectation of these peptides not influencing the false discovery estimates.

We conclude that the choice of the decoy database does not have a significant impact

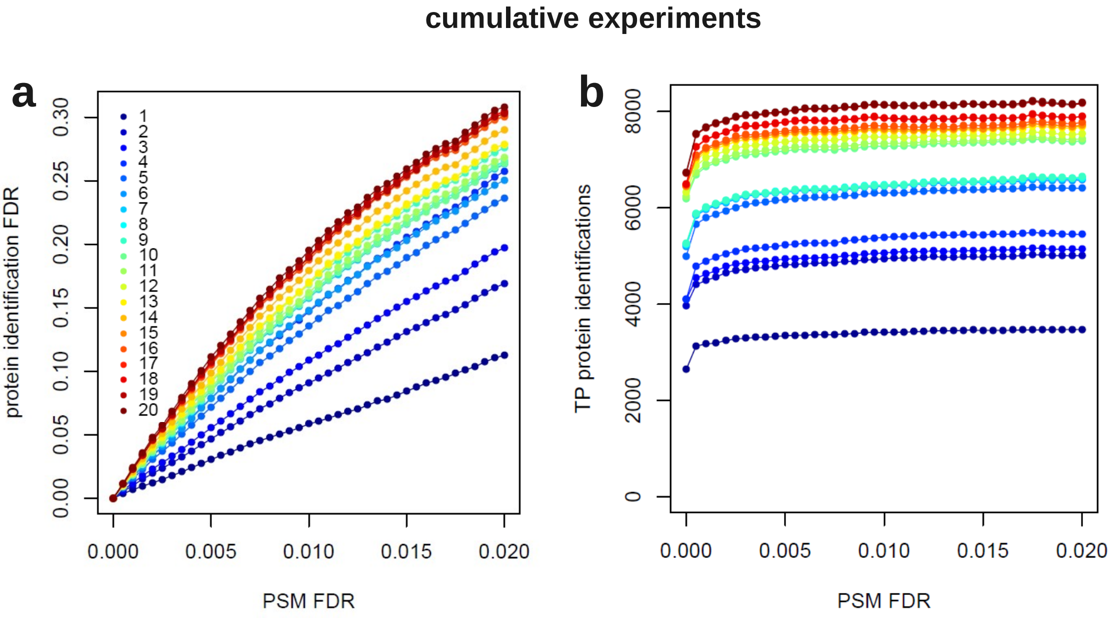


Figure 4.7: Protein identification false discovery rate for protein inference excluding ambiguous peptides. From the total data set of 20 experiments all peptide-spectrum matches (PSMs) referring to peptides pointing to more than one (target or decoy) protein, were removed. For the remaining PSMs the protein identification false discovery rate (FDR) was estimated. This protein inference method has no influence on the general behaviour of the protein identification FDR estimates as expected from the underlying model.

on estimating protein false discovery rates. In particular we note that simple reversing of the target database achieves accurate estimates while more sophisticated approaches to decoy database generation do not improve the estimates.

4.4.4 Comparison of protein identification FDR estimates

We compared protein identification FDR estimates of *MAYU*, ProteinProphet and the naïve target decoy approach. We studied four different subsets of the *C. elegans* data set varying in size (1, 5, 10 and 20 cumulative experiments). Protein identifications were inferred with ProteinProphet. Protein identification FDR for these identifications were then determined with *MAYU*, with the built-in functionality of ProteinProphet and the naïve target-decoy strategy.

The naïve target-decoy strategy estimates protein identification FDR analogously to PSM FDR, i.e. by approximating the expected number of false positive (FP) protein

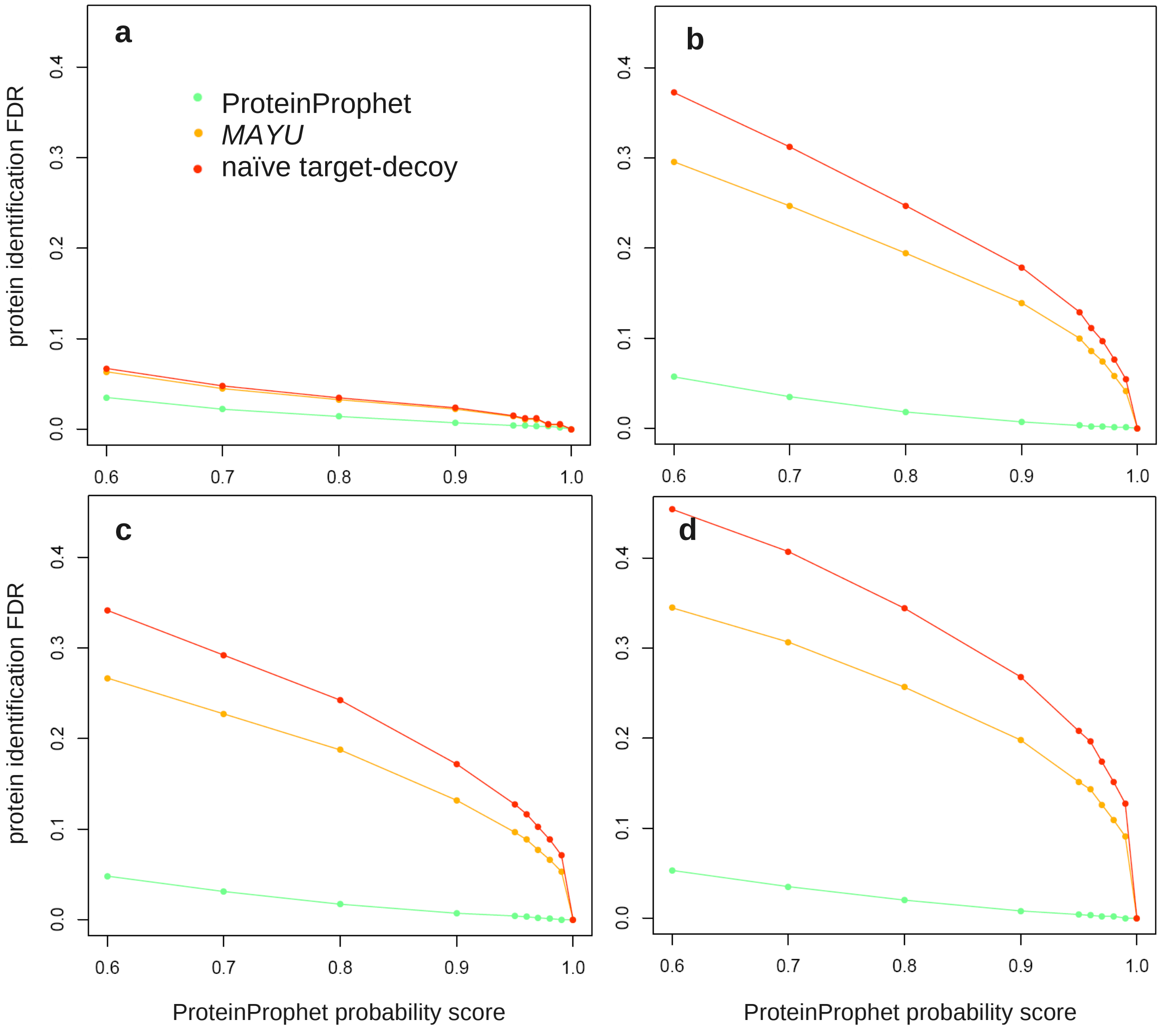


Figure 4.8: Comparison of different protein identification false discovery rate estimation strategies. We compared protein identification false discovery rate (FDR) estimates of *MAYU*, ProteinProphet and the naïve target-decoy strategy for four different data set sizes (1, 5, 10 and 20 experiments of the *C. elegans* data set, **a-d**). The discrepancy of the alternative FDR estimates and the *MAYU* estimates grow with data set size.

identification by the number of decoy protein identification (**Table 4.4.4**). We observe that the naïve target-decoy strategy estimate is overly pessimistic (**Fig. 4.8**). This is due to true positive (TP) protein identification containing FP PSMs and thus not contributing to the pool of FP protein identifications. In contrast, ProteinProphets FDR estimates are too optimistic. For typically sized data sets (**Fig. 4.8a**) ProteinProphet and naïve target-decoy still yield reasonable protein identification FDR estimates. However, the larger the data set size the more pronounced we find its discrepancy to the

MAYU estimates. Note the difference between FDR estimate and protein inference. The foregoing comparison only aims to compare different protein identification FDR estimates, it is not suitable to assess the protein inference functionality of ProteinProphet that provides an effective prioritization of protein identifications.

4.4.5 Protein identification FDR for various data sets

Proteomic studies typically report lists of protein identifications and specify confidence in terms of FDR at PSM level. We used various data sets to study how well PSM FDR reflects the relevant confidence measure for these lists, i.e. protein identification FDR. To this end, we applied *MAYU* to several shotgun proteomics data sets, varying in MS instrumentation and studied organism (**Fig. 4.9, a-c**). We analyzed isoelectric focusing experiments of a *C. elegans* [119], *L. interrogans* and *S. pombe* sample. While the first data set was acquired on a low resolution LTQ instrument, the latter two were acquired on a high mass accuracy LTQ-FT instrument. Protein identifications were compiled by lexicographical protein inference including all PSM above a score threshold (see 4.3.1). We observe that protein identification FDR behaves similarly for any of the data sets. Most importantly, we note that protein identification FDR is significantly elevated compared to the PSM FDR. We conclude that the PSM FDR is not generally an appropriate confidence measure for lists of protein identifications.

	PSMs			peptide identifications			protein identifications		
PSM FDR	target	decoy	ratio	target	decoy	ratio	target	decoy	ratio
0.05	954,661	47,725	0.05	117,293	36,419	0.310	16,459	14,354	0.872
0.01	795,502	7,947	0.01	82,628	6,394	0.077	11,089	4,974	0.449
0.001	614,486	614	0.001	65,779	519	0.008	8,477	506	0.060

Table 4.1: Results of a target-decoy database search of the complete *C. elegans* data set. Number of target and decoy peptide-spectrum matches, peptide identifications and protein identifications for three different PSM FDRs are shown. For peptides mapping to several protein sequences only the alphabetically first protein id was considered. For any PSM FDR, the ratio of decoy to target hits is higher for peptides and again higher for proteins. Unlike for the PSMs, this ratio is not to be mistaken for FDR for peptide or protein identifications.

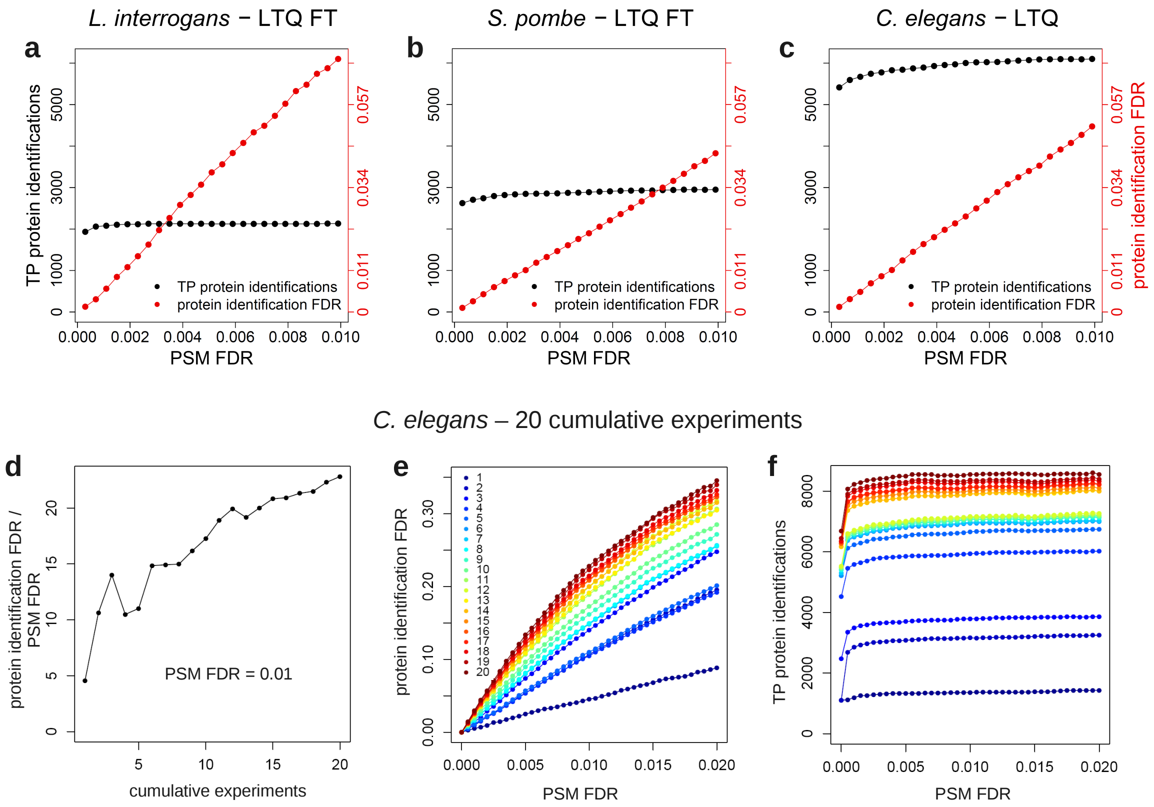


Figure 4.9: Protein identification false discovery rates behave similarly for data sets of different species and instruments and largely depend on the size of the data set. We applied *MAYU* to three different data sets of similar size but from different organisms and instruments (59,918 **a**, 40,008 **b**, 65,553 **c** target PSMs for a PSM FDR of 0.01). In all three data sets the protein identification false discovery rate (FDR) is roughly 5 times higher than the peptide-spectrum match (PSM) FDR. The number of estimated true positive (TP) protein identifications saturates for very low PSM FDR (**a-c, f**). We investigated the influence of data set size using 20 compilations from the *C. elegans* data set representing 1 to 20 cumulative experiments. The ratio of protein identification FDR to PSM FDR (protein identification FDR / PSM FDR) shows clear dependence on data set size (**d**). In the complete data set (1,305 LC-MS/MS runs) the protein identification FDR is more than 20 fold higher than the PSM FDR. For all data set sizes the protein identification FDR is elevated compared to the PSM FDR over the whole range of PSM FDR (**e**) and the apparent maximal number of TP protein identifications is reached for very stringent PSM FDR of roughly 0.005 (**f**). This data suggests that increasing the PSM FDR beyond 0.005 mainly entails an accumulation of FP protein identifications.

4.4.6 Data set size dependent accumulation of false positives

Using *MAYU* we assessed the impact of data set size on protein identification FDR. For this purpose, we analyzed the currently largest shotgun proteomic data set for *C. elegans* [119] generated at the Center for Model Organism Proteomes (C-MOP). We sub sampled this data set (5,897,279 tandem mass spectra, 1,305 LC-MS/MS runs) into 20 data units of increasing size (**Fig. 4.9, d-f**). For each of these units we estimated the FDR of the protein identifications defined for varying PSM FDR cutoffs.

Our analysis revealed that protein identification FDR is strongly influenced by the chosen FDR of PSMs and the size of the respective data set (**Fig. 4.9, d,e**). For the 20 data units, protein identification FDR increases dramatically with growing PSM FDR (**Fig. 4.9d**). In the largest data unit, protein identification FDR is more than 20 times the corresponding PSM FDR (**Fig. 4.9e**).

For all data sets shown, the apparent maximal number of true positive protein identifications achievable by the respective data unit is approached already at very low PSM FDR, in the range of 0.005 (**Fig. 4.9, a-c,f**). This quick convergence of the expected number of TP protein identifications suggests that including less reliable PSMs mainly entails accumulation of FP protein identifications without gaining new TP protein identifications. We conclude that in order to achieve acceptable protein identification FDR, PSMs have to be selected exceedingly stringently with increasing data set size.

4.5 Discussion

MAYU is a generic strategy to estimate false discovery rates for protein identifications inferred from shotgun proteomics data sets. An implementation of *MAYU* is publicly available.

Unlike other well established strategies, which quantify the uncertainty of PSMs (frequently also referred to as peptide identifications), *MAYU* evaluates quality at the level of protein identifications. *MAYU* implements a novel and generic strategy that generalizes the established target-decoy database search approach for PSMs in order to estimate FDR for protein identifications. This approach constitutes a shift from assessing confidence of proteomic data sets at PSM level by providing instead a confidence measure at protein level. It should be noted that *MAYU* is not designed for protein inference, i.e. for the assembly of protein identifications. Instead *MAYU* generically assesses the reliability of protein identifications already inferred by any sequence database driven identification strategy (e.g. search engines such as Sequest, Mascot or protein inference strategies such as ProteinProphet). Besides exemplarily showing *MAYU*s compatibility to applications such as lexicographical and ProteinProphet protein inference, we also applied *MAYU* to non-ambiguous protein inference (**Fig. 4.7**). With regards to conceptual as well as computational issues, *MAYU* scales well with data set size and is particularly suited for the analysis of very large integrated data sets comprising millions of tandem mass spectra. This concept is also expected to be applicable to other high throughput experiments in biology and medicine which are characterized by indirect observations.

In this study, we assessed *MAYU* on three heterogeneous data sets including the largest shotgun proteomics data set for *C. elegans* available to date [119]. FDR estimation for protein identifications on data sets of this size has not been solved satisfactorily prior to *MAYU*. Widely used protein inference tools like ProteinProphet [96] have proven to yield reliable error estimates on data sets at the experiment level (typically 10-50 LC-MS/MS runs) but fail to estimate accurate protein identification FDR for large data sets (**Fig. 4.8**). Current approaches to assemble protein identification from such large data sets rely on common sense criteria for which no quantitative confidence measure at protein identification level has been reported yet. *MAYU* overcomes this limitation by providing FDR for protein identifications in arbitrarily large data sets.

We found that data set size critically influences protein identification FDR. For the integrated data set (1,305 LC-MS/MS runs), the discrepancy in FDR rises to a more than 20-fold difference, even when stringent PSM FDR thresholds are used. Besides these results obtained for protein inference as described in 4.3.1, we found the same trend towards larger protein identification FDR for various other protein inference strategies.

This study aims to quantify the uncertainty of protein identifications in the context of a large-scale data set. To the best of our knowledge, this is the first study that independently confirms the scale of FDR estimates. More precisely, we showed that the scale of FDR estimates for a subset of single hit are in very good agreement with an independent method relying on experimentally acquired isoelectric points of peptides (**Fig. 4.4a**).

We also showed that *MAYU*s protein identification FDRs are reproducible regardless of the underlying decoy database (**Fig. 4.6**).

Other approaches like the protein inference engine ProteinProphet have been successfully applied to estimate confidence measures for protein identifications in the context of smaller data sets. ProteinProphet relies on probability estimates of given PSMs to be false, to compute the probability of the respective protein identification to be false. Our results show that in large data sets, certain classes of PSMs are enriched in false positive PSMs. This particularly applies to PSMs defining single hits: Their actual proportion of false positive instances was nearly two orders of magnitude larger than the average FDR for the complete set of PSMs (data not shown). This discrepancy is not a contradiction: Because false positive PSM randomly map to a very large target-decoy database, they are prone to map to previously unoccupied protein entry and therefore give rise to a single hit. Phenomena like these complicate a reasonable estimate for false positive probabilities for single PSM and thus challenge approaches like ProteinProphet to estimate FDRs at protein level in the context of large-scale data sets (**Fig. 4.8**). In contrast, *MAYU* estimates protein identification FDR without relying on false positive probabilities for single hit PSM, since FDR estimates are derived solely from statistics gathered at the protein identification level.

In a similar spirit, a Poisson model has been proposed to estimate the proportion of false positive protein identifications given the number of supporting PSMs [1]. The parametric model requires the Poisson distribution parameter to be estimated. This estimate is obtained in a heuristic way by assuming different scenarios for the validity of single hits. This model implicitly assumes statistical independence of all PSMs. Our results indicate that this assumption does not hold in general (data not shown), which confirms the coarse approximate nature of the Poisson model.

MAYU circumvents the shortcomings of such parametric assumptions. *MAYU* exploits the underlying target-decoy database search strategy and particularly addresses the phenomenon of true positive protein identifications containing false positive PSMs. This component clearly distinguishes *MAYU* from naïve target-decoy strategies that approximate the number of false positive protein identifications with the number of decoy protein identifications [139]. These strategies overestimate protein identification FDR since they implicitly assume that all protein identifications containing false positive PSMs are false positive (**Table 4.4.4**). In particular, the degree of protein identification FDR overestimation grows with data set size (**Fig. 4.8**) [139].

Consider the following example where all proteins of a proteome (e.g. *E. coli*) have been truly identified. The correct protein identification FDR would thus be zero. Due to the accumulation of false positive, i.e. decoy PSM (not invalidating the true evidence for the protein identifications) the naïve target-decoy strategy will falsely estimate an FDR differing significantly from zero. Furthermore, the naïve target-decoy estimate has the undesired property of diverging stronger the more experiments will be carried out.

*MAYU*s FDR builds on an estimate of the number of protein identifications containing false positive PSMs. In this study we estimate this quantity by the number of decoy protein identifications. While in principle there are other means to estimate the number of protein identifications containing false positive PSMs, *MAYU* uses target-decoy database searched data sets to estimate protein identification FDRs since this represents a well understood and well accepted strategy.

In addition, we find the assumptions underlying the target-decoy search strategy to be well met. The central assumption comprises that false positive PSMs uniformly distribute between target and decoy database. Foregoing studies have discussed and shown the general validity of the target-decoy search strategy [41]. Recurrently occurring chemical entities (e.g. unusually modified peptides), which are not represented by the protein database, could potentially challenge the validity of target-decoy strategies since each of these give rise to false positive PSM preferably mapping to the same false peptide sequence. However, the overall balanced distribution of all false positive PSMs as well as protein identifications containing false positive PSMs is not compromised, due to the large number of such entities.

We have seen that protein length has a small and controllable effect on *MAYU*'s FDR estimates (**Fig. 4.3a**). We observed that deviations from the uniformity assumption regarding the distribution of protein identifications containing false positive PSM do not compromise the FDR estimates (**Fig. 4.3b**). We furthermore observed that *MAYU*'s FDR estimates are not dependent on the underlying type of decoy database, i.e. reversed or Markov model type (**Fig. 4.6**). Most importantly, we were able to independently reproduce single hit FDR (**Fig. 4.4 & 4.5**), altogether providing a strong indication that the assumptions underlying *MAYU* analysis are reasonable and provide reliable estimates of protein identification FDR.

Throughput and sensitivity of mass spectrometers applied to proteomics are steadily increasing. Data repositories have been created to store the vast amount of mass spectrometric data [32, 37, 90, 75]. These repositories constitute a cornerstone for proteomics contributing to a wide range of genome-wide studies. Well curated data repositories are a prerequisite of the success of applications like spectrum library searching [128, 32, 80], protein expression estimates by spectral counting [112] and targeted proteomics approaches based on the selection of proteotypic peptides [79]. *MAYU* enables to more efficiently utilize existing and upcoming data sets in this context by allowing a quantitative quality control of the of protein identifications. *MAYU* is the first approach to quantify the uncertainty of protein identifications in the context of large scale data sets, thereby allowing to automatically curate proteomics repositories of steadily increasing size. We conclude that approaches like *MAYU* will significantly enhance genome-wide studies based on shotgun proteomics strategies.

5 Generic Comparison of Protein Inference Engine Families

5.1 Summary

Protein inference defines a key step in mass spectrometry based proteomics and refers to the reconstruction of protein identities from the fragment ion spectra generated by shotgun proteomics experiments. There has been an ongoing debate about how to optimally infer protein identities. The inability to estimate false discovery rates of protein identifications for large integrated datasets has so far hindered to generally assess protein inference approaches in the context of contemporary datasets featuring ever increasing size and heterogeneity.

We present a simple generic strategy to benchmark a wide range of protein inference engine. This strategy essentially builds on a performance measure for protein inference that evaluates the number of correct protein identifications while accounting for false discovery rates at the level of protein identifications. Specifically, a family of several thousand protein inference approaches is benchmarked to systematically explore the benefit of excluding possibly unreliable protein identifications, such as e.g. single hit wonders. In a preliminary study we identified particularly unreliable protein identification subsets, in terms of local false discovery rates. On the basis of this study, a family of protein inference engines is defined by extending a simple inference engine by thousands of pruning variants, each excluding a different set of unreliable identifications. None of the pruning strategies improves protein inference performance when applied to the currently largest reported shotgun proteomics dataset for *C. elegans.* We conclude that the maximal number of reliable protein identifications can be effectively inferred by considering all spectral evidence of high quality, including single hit wonders. [1]

---

[1] Lukas Reiter and I share the first authorship on this work. Both Lukas Reiter and me have equally contributed to the main ideas and their implementation. The project was performed under the supervision of Joachim M.Buhmann, Michael O. Hengartner and Ruedi Aebersold. We thank Sabine Schrimpf for providing the *C. elegans* dataset.

5.2 Introduction

A fundamental goal of mass spectrometry based proteomics is to determine the true protein composition of biological samples. Protein inference denotes the task of recovering the protein identities from the fragment ion spectra acquired in the course of shotgun proteomics experiments. Assessment of protein inference methods so far suffered from the lack of a generally applicable performance criterion that takes protein identification reliability into account. We extend the statistical validation framework Mayu [26] to define such a criterion and apply it to benchmark a family of prototypical protein inference approaches.

Protein inference is a task that arises in the context of shotgun proteomics experiments [3]. In their simplest implementation, protein samples are first extracted from their biological source, subjected to enzymatic digestion, yielding a complex peptide mixture that is analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS). Fragment ion spectra are acquired after stochastic or directed precursor ion selection [118]. More elaborate strategies augment this workflow by additional fractionation steps at the level of proteins/peptides before LC-MS/MS analysis. These steps give rise to a set of peptide fragment ion spectra that constitute the raw data to infer the proteins present in the biological source.

Interpretation of the spectral data consists of first matching the fragment ion spectra to their corresponding peptide sequences (peptide spectrum matching) and second to integrate these results to infer the set of proteins initially present in the biological sample (protein inference) [95]. See also **5.1**. These steps are typically automated due to the vast amount of spectra generated in the course of contemporary shotgun proteomics approaches.

Peptide-spectrum matches are typically generated using one of the many available search engines, e.g. [42, 105, 30]. Search engines map fragment ion spectra to the best matching peptide sequence in the protein database of the studied organism [97]. Various statistical measures, such as e.g. false discovery rates [9], have been derived to account for possibly incorrect peptide-spectrum matches [18]. In this context the target-decoy strategy has recently grown very popular since it is simple to implement and compatible with all currently used search engines [94, 41].

Protein inference uses peptide-spectrum matches to infer the identities of proteins initially present in the biological source [95]. A protein identification comprises an assembly of supporting peptide-spectrum matches. Protein inference engines integrate possibly redundant spectral evidence to compile a set of protein identifications that are expected to be correct, i.e. contain at least one correct peptide-spectrum match. The more complex a proteome the more frequently peptide-spectrum matches turn out to ambiguously map to several protein entries, e.g. protein splice variants. It is common practice to circumvent this issue by effectively reducing a protein identification to a gene locus identification in case of ambiguity ("gene locus inference")[90, 15, 5, 119]. More sophisticated protein inference engines though implement statistical or algorithmic approaches to disambiguate peptide-spectrum matches where required [96, 112, 44, 108, 148]. After having applied an inference engine, it is common practice to exclude possibly unreliable protein identifications, such as e.g. single hit protein identifications. There has been considerable debate about whether this kind of post-processing enhances protein inference [59, 58].

Protein identifications are not perfect since peptide-spectrum matches can be spurious. Errors at the level of peptide-spectrum matches though propagate non-trivially to the level of protein identifications. While error rate estimation for peptide-spectrum matches is well established [18], several attempts have been made to control protein identification error rates throughout. Besides their inference functionality, most protein inferences engines also estimate error rates based on probabilities of individual peptide-spectrum matches being wrong [96, 112, 44, 108]. It turns out, however, that this kind of approach does not scale well with dataset size [26]. Another approach estimates the number of incorrect protein identifications assuming that false positive peptide-spectrum matches distribute according to a Poisson distribution across the protein database. [90, 126]. The estimates from such models though give ambiguous estimates depending on assumptions regarding single hit protein identifications. Simple target-decoy approaches, estimating the number of false positive protein identifications by the number of decoy identifications [97, 108, 148, 59], have shown to give too pessimistic estimates [26]. Considering the limitations of the latter approaches, none of these qualifies as a general purpose method to control protein identification error rates for datasets of different quality and size. To close this gap, we recently proposed the Mayu approach that appropriately adapted the target-decoy strategy to the protein inference task and achieved accurate, independently validated protein identification false discovery rates (i.e. the expected proportion of incorrect among all accepted identifications) for a range of diverse datasets differing in size, underlying proteome and experimental setting [26].

The literature does not provide a starting point to decide which protein inference engine to choose for a particular application scenario (in contrast to the rich literature about search engine comparisons, see e.g. [68]). Specifically, none of the studies presenting a novel protein inference engine [96, 112, 44, 108] reports identification performance by means of a thorough benchmark against at least one baseline method. Instead each study shows that its approach is to some extent able to recover protein identities from different kinds of datasets, i.e. artificial, well characterized protein mixtures comprising at most dozens of sufficiently abundant proteins and real world complex whole proteome mixtures. A single study ([108]) shows results of a competing approach ([96]). This study lacks, however, a systematic benchmark with a sensible performance measure.

Typically (though not always), performance of a protein inference engine is positively correlated with the number of protein identifications attributed to the respective engine. In this context, protein inference performance is sensibly measured by specifying the number of correct and incorrect protein identifications, i.e. by not only counting the total number of protein identifications but by also considering identification specificity. In the case of an artificial protein mixture, identification performance is easily measured since the protein composition is known and identifications are therefore trivially recognized as true or false positive. In the real world case, identification performance is not straightforward to measure since the true protein composition of the test sample is not known. As delineated before, it is has been only partially understood how to count the number of correct and incorrect protein identifications in this case until recently [26] and therefore protein inference engine performance has only been reasonably approximated and reported for scenarios that do not reflect the heterogeneity and size of contemporary shotgun proteomics datasets [96, 112, 44, 108].

The present study contributes a sensible and generic performance measure that enables to easily benchmark protein inference engines (**5.1**). This measure evaluates the number of true and the proportion of false positives in a particular set of protein identifications. We show that these numbers can be easily and generically estimated on the basis of protein false discovery rates [26]. We apply this performance measure to compare a family of a widely used protein inference engines. This family is based on the popular "gene locus inference" approach [90, 15, 5, 119] . For these base protein inference engines we additionally study the impact of post-processing schemes related to the exclusion of protein identifications subsets featuring low spectral counts. We report a target-decoy strategy for local false discovery rates [40] to quantify the reliability of various protein identification subsets. In order to systematically study the exclusion of protein identifications after applying one of the base inference engines, we introduce the concept of a selection scheme that formally characterizes properties of a subset of protein identifications (**5.2**). By systematically varying selection schemes we effectively benchmark thousands of different variants of the base protein inference engine (**5.1**). Finally, we apply the benchmark strategy to compare "gene locus inference" with ProteinProphet. For the largest reported shotgun proteomics dataset for *C. elegans* [119] we find that "gene locus inference" without any further pruning achieves the highest performance.

Material and Methods

5.3 Dataset and data processing

This work builds on a heterogeneous dataset acquired for *Caenorhabditis elegans* in a study in which varying sample preparation and MS instrumentation were applied [119]. The spectral data was searched against a composite target-decoy database using Turbo Sequest [42] and Sequest on a Sorcerer machine (Sorcerer$^{\text{TM}}$-SEQUEST$^{®}$, 3.10.4 release). The search results were transformed to the pepXML format and further processed using the Trans-Proteomic Pipeline [71] to the level of PeptideProphet [72] in units of experiments. The pepXML files were then further analyzed with the Mayu software [26]. If a peptide existed in more than one protein sequence the hit was associated with one protein representing the gene locus ("gene locus identification") [119]. All database searches were performed using a concatenated target-decoy database [41]. As target database we chose wormpep170 (WormBase). ProteinProphet was run on the pepXML files using runprophet from the Trans-Proteomic Pipeline, and target/decoy protein identifications of ProteinProphet were used as input for the Mayu protein identification false discovery rate calculation.

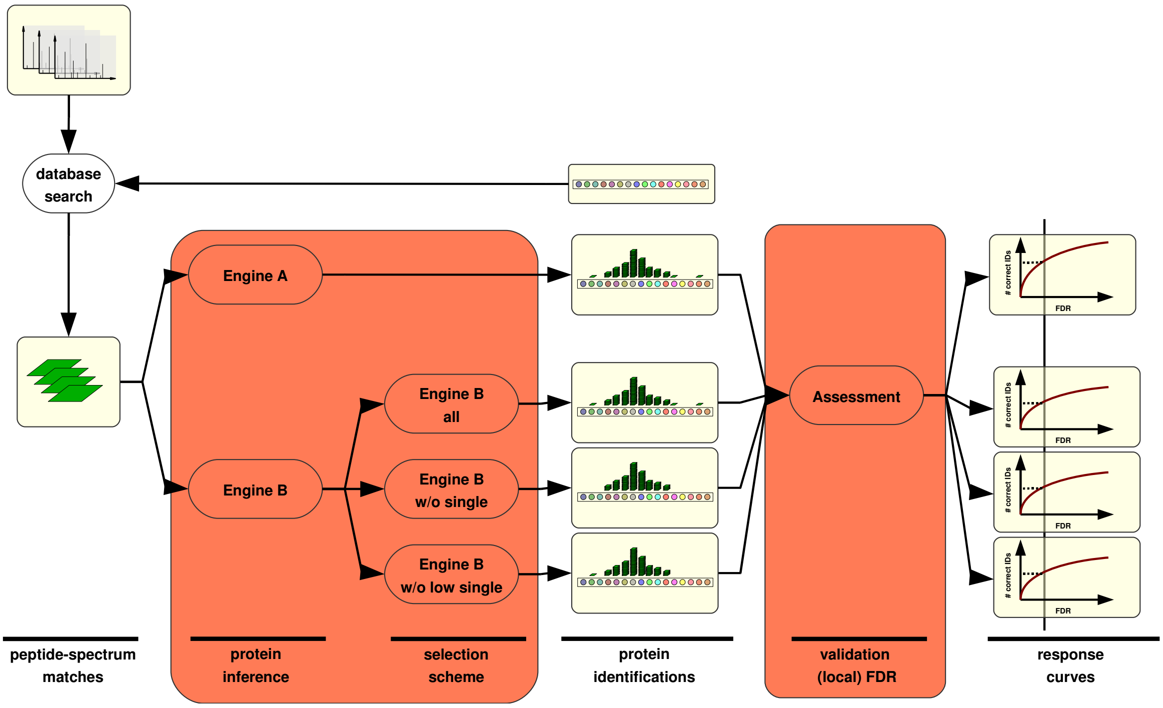


Figure 5.1: Schema to benchmark protein inference engines. Tandem mass spectra are generated in the course of a shotgun proteomics experiment. Protein identities are recovered in two distinct steps, i.e. (1) peptide identification yielding peptide-spectrum matches and (2) protein inference assembling peptide-spectrum matches to protein identifications. Optionally, protein inference is followed by additionally pruning particular protein identifications sets, e.g. single hit identifications. We formally characterize these sets by means of selection schemes to systematically study different pruning strategies. Protein identification reliability is assessed in terms of (possibly local) protein identification false discovery rates. Protein inference performance is measured by estimating the number of correct identifications over a range of different protein identification false discovery rates, thereby giving rise to inference engine characteristic response curves. Comparison and ranking of protein inference engines is usually performed for a user defined protein identification false discovery rate. Processes studied in this work are highlighted in red. Specifically, these are (1) selection scheme variants of available protein inference engines and (2) assessment and comparison of protein inference performance.

5.4 Local false discovery rates for protein identification subsets

Local false discovery rates can be used to quantify the reliability of protein identification subsets. We use simple properties, such as e.g. number of supporting peptide-spectrum matches, to characterize protein identification subsets. More generally, an individual

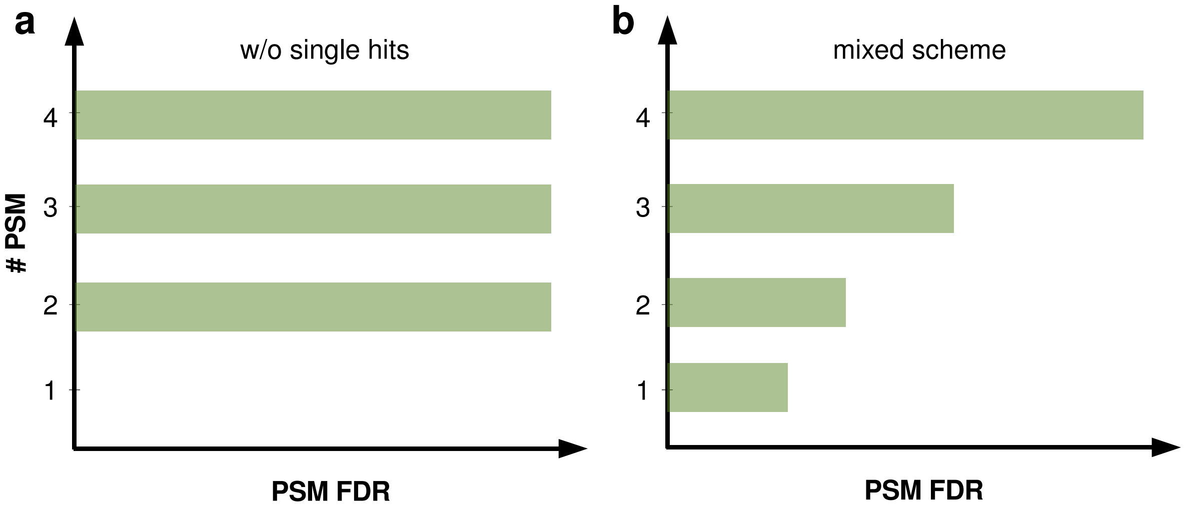


Figure 5.2: Selection scheme illustration. Selection schemes aim to formalize the notion of selecting the spectra more stringently for protein identifications evidenced by few spectra than for those featuring more redundant evidence. Selection schemes characterize protein identification subsets according to the reliability of peptide spectrum matches (PSM FDR) and some property of a protein identification, e.g. the number of supporting peptide spectrum matches (# PSM). Formally, a selection scheme specifies a series of peptide-spectrum match false discoveries $m_1, m_2, \ldots$ and accordingly considers protein identifications that for some $i = 1, 2, 3, \ldots$ are supported by at least i peptide-spectrum matches afflicted with false discovery rate of less than m_i. **(a)** depicts the selection scheme for excluding all single hit protein identifications and considering all other protein identifications supported by at least two peptide-spectrum matches at false discovery rate lower than some threshold. **(b)** depicts a more intricate selection scheme that allows to consider single hit protein identifications as long as the respective peptide-spectrum matches feature a low false discovery rate. With increasing support the spectral quality requirements decrease.

property Y is used to split the complete set of protein identifications into subsets, each featuring the same property value (e.g. single hits) and to measure their quality by local false discovery rates $\mathrm{FDR}(y)$ [40]. By definition of local false discovery rates we can write

$$\mathrm{FDR}(y) := P(\mathrm{fp} \mid y) = \frac{P(y \mid \mathrm{fp}) \cdot P(\mathrm{fp})}{P(y)} \tag{5.1}$$

While y corresponds to the property value of a single identification, fp denotes the identification to be false positive. $\mathrm{FDR}(y)$ thus corresponds to $P(\mathrm{fp})$ scaled by the ratio of $P(y \mid \mathrm{fp})$ to $P(y)$. Calculation of $\mathrm{FDR}(y)$ requires to specify the distributions $P(y \mid \mathrm{fp})$, $P(\mathrm{fp})$ and $P(y)$. $P(y)$ can be estimated with its empirical distribution defined by all protein identifications mapping to the target database. Recalling that protein identifications mapping to the decoy database are false positive by definition, $P(y \mid \text{fp})$ can be approximated analogously by its empirical distribution defined by all decoy protein identifications. The protein identification false-discovery rate for the complete identification set is straightforwardly estimated with Mayu (for details please see [26]) and provides an estimate for the prior $P(\text{fp})$, finally allowing to estimate the local false discovery rate by plugging in the latter estimates.

5.5 Protein identification selection schemes

We use selection schemes to characterize (presumably high quality) protein identification subsets that we wish to report in the final identification list. A very simple selection scheme could for instance characterize the subset of all non-single hit protein identifications and the single-hit identifications whose peptide-spectrum matches score higher than any decoy match.

Generally, protein identification sets were generated by various selection schemes considering the “number of supporting peptide-spectrum matches” property. Selection schemes are characterized by a sequence of peptide-spectrum match false discovery rates $m_1, m_2, ...$ The selection scheme considers those protein identifications which are supported by at least i peptide-spectrum matches that map to the peptide-spectrum match set with false discovery rate less than m_i. Selection schemes thus allow us to define protein identification sets where protein identifications evidenced by very few high confidence peptide-spectrum matches and protein identifications supported by a large number of lower confidence peptide-spectrum matches. For an illustration see also **5.2**.

5.6 Screening and false discovery rate evaluation of selection schemes

We exhaustively enumerated all selection schemes ($m = 0$, 0.0001, 0.0002, 0.0003, 0.0004, 0.0005, 0.001, 0.0015, 0.002, 0.0025, 0.003, 0.0035 for m_1, ..., m_6, $m_{>6}$) to identify optimal protein identification sets for some desired false discovery rate. We considered a protein identification set to be optimal if it maximized the expected number of true positive protein identifications for a chosen protein identification false discovery rate. The protein identification false discovery rate of each selection scheme was determined as the local false discovery rate estimate for the protein identification property "selection scheme predicate". This predicate is true for a protein identification being considered by the underlying selection scheme and false otherwise.

5.7 Protein inference engine benchmark

For the protein inference engine benchmark we assume that for each competitor the (target-decoy) identification results are available in terms of an identification list. We measure performance of each competing approach by evaluating the absolute number of true positive identifications and the respective proportion of false positives, i.e. the false discovery rate. Since the total number of identifications is trivially given by length of protein identification list it is sufficient to estimate the false discovery rate in order to complete the performance measure. We estimate the false discovery rate directly with Mayu [26]. In case the competitors inference strategy involves a pruning step according to a selection scheme, the proportion of false positives is estimated as local false discovery rate as described in the preceding section.

Most inference engines assign a score to each protein identification and therefore produce a series of protein identifications with increasing size and false discovery rate. Assessing this series of identification sets yields a response curve that characterizes the performance of the inference engine across the whole spectrum of false discovery rates. In summary, competitors can now be sensibly ranked according to the amount of true positive identifications at a user defined false discovery rate. See also **5.1**.

5.8 Results

The following sections report the quantitative impact of certain properties on protein identifications reliability, such as sequence length and spectral support of protein identificationsm, a systematic benchmark of "gene locus inference" in conjunction with a multitude of selection schemes based on spectral support and the more sophisticated protein inference engine ProteinProphet.

5.8.1 Local false discovery rates for protein identification subsets

To date, deciding on the final set of protein identifications in a given proteomics dataset is frequently based on heuristic criteria supposed to enrich for valid identifications. A

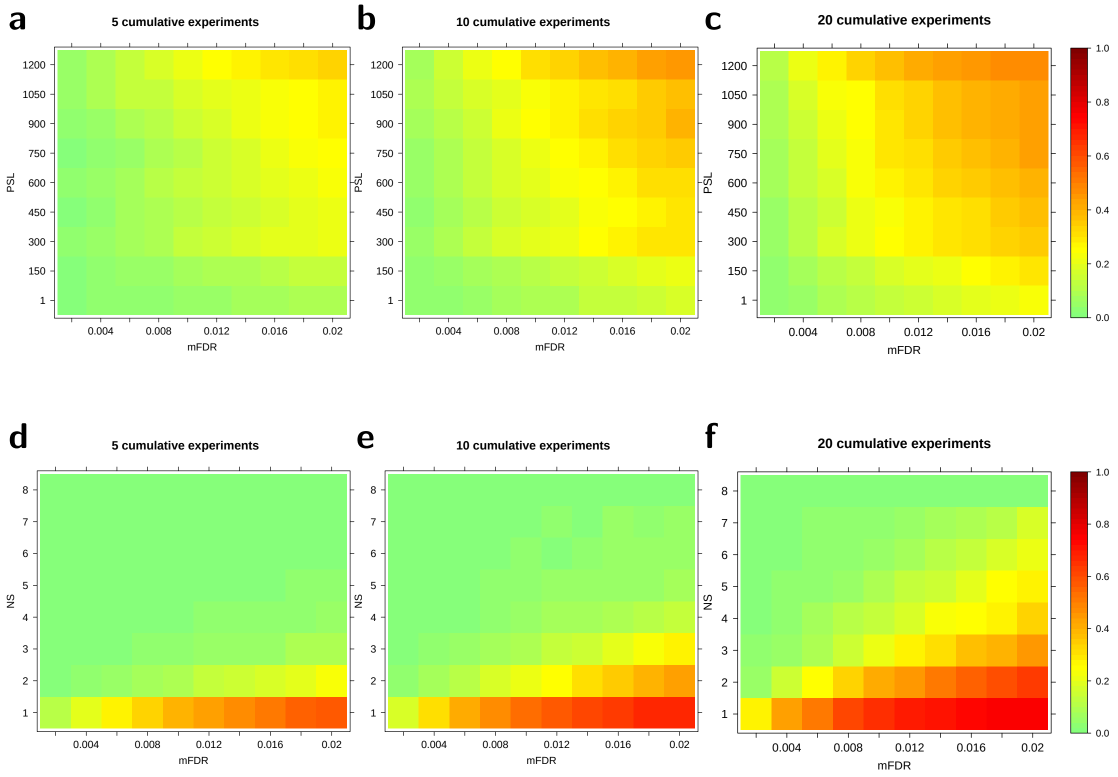


Figure 5.3: Local protein identification false discovery rate (FDR) for protein identification groups characterized by either protein sequence length in amino acids (specified by lower bin boundaries) (a-c) or number of peptide-spectrum matches defining the identification (d-f) Each heat map depicts results for dataset partitions of varying size (amount of cumulative experiments). Magnitude of local protein identification false discovery rate is color coded as indicated. Certain protein identification subsets feature more than 60 fold higher protein identification false discovery rate than the underlying peptide-spectrum match false discovery rate.

widely used strategy filters for protein identifications whose peptide-spectrum matches score above a certain threshold. More stringent strategies furthermore require a valid protein identification to be composed of a minimal number of supporting peptide-spectrum matches (e.g., neglect all single hits) . There have been substantial debates about the validity of such criteria to compile protein identification sets from large datasets.

To determine whether removal of particular subsets of protein identifications improves the quality of the remaining identifications, we estimated local false discovery rates for identification subsets. These subsets were characterized by a certain property (e.g. supported by single hits) expected to have an impact on the quality of protein identifications (**5.3**). We studied the effect of two properties that were expected to have an impact on the quality of protein identifications. Specifically, the effects of protein sequence length in amino acids and number of peptide-spectrum matches supporting the protein identification (number of supporting peptide-spectrum matches) were explored.

In a first step we generated local false discovery rate estimates for protein identification sets characterized by protein sequence length (**5.3a-c**). Increasing protein sequence length is expected to amplify false discovery rate since false positive peptide-spectrum matches map to the database by chance and therefore are more likely to map to larger proteins. Our local false discovery rate estimates clearly confirm this expectation. The local false discovery rate for proteins of for instance sequence length 400 is two fold higher than for proteins with sequence length 100 using a peptide-spectrum matches false discovery rate cutoff of 0.01. We find a similar though not so pronounced trend for smaller datasets, i.e. subsets of the complete *C. elegans* dataset.

We went on to study protein identification sets characterized by a varying number of supporting peptide-spectrum matches (**5.3d-f**). Note that a protein identification with one supporting peptide can be supported by several peptide-spectrum matches. As expected, the confidence in a protein identification scales with the number of peptide-spectrum matches supporting it. Surprisingly high false discovery rate were observed for protein identifications only supported by a single peptide-spectrum match (single hits) in the complete dataset, exceeding 0.65 for a presumably stringent peptide-spectrum match false discovery rate of 0.01. Even protein identifications being apparently approved by two peptide-spectrum matches complying with a cutoff of 0.01 featured false discovery rate of 0.4. While not being so pronounced, we encounter a similar situation for smaller subsets of the *C. elegans* datasets. The discrepancy between false discovery rate of peptide-spectrum matches and corresponding protein identifications was most pronounced for the subset of single hits. These results confirm a similar discrepancy estimated by other (not so generically applicable) methods, such as validation with synthetic peptides [26], considering deviations in measured and predicted isoelectric point [26] or manual curation [58].

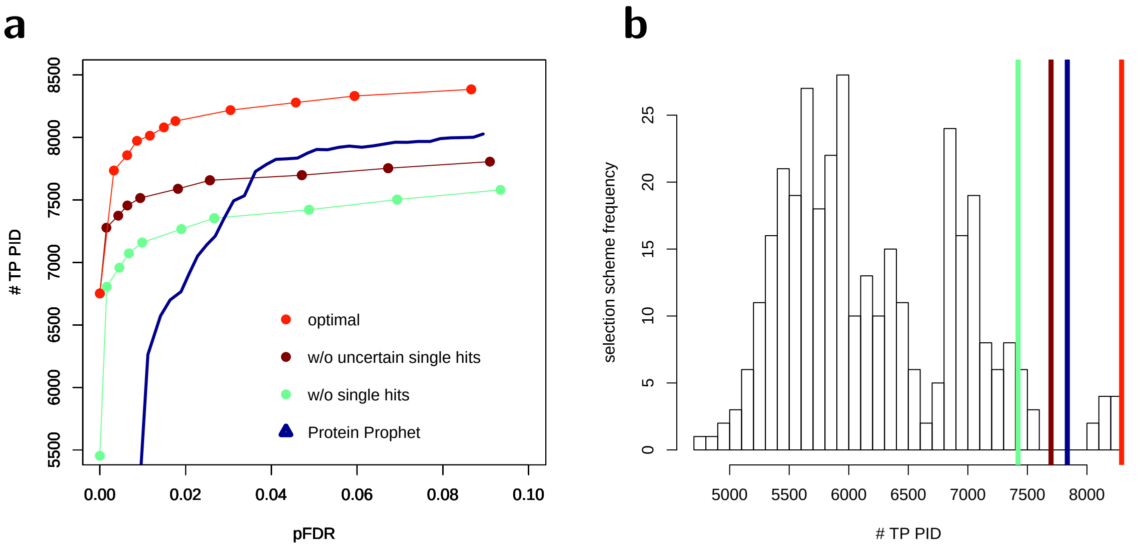


Figure 5.4: Optimal pruning strategies with respect to expected number of true positive protein identifications. **(a)** Comparison of protein identification selection schemes. All optimal selection schemes (see text) consider all peptide-spectrum matches (PSMs) of sufficiently low peptide-spectrum match false discovery rate (FDR) (red). Two alternative selection schemes are shown exemplarily. Selection schemes consequently neglecting single hits (green) or solely neglecting single hits with nonzero uncertainty (brown). The response curve of ProteinProphet is shown in blue.**(b)** Histogram of expected number of correct protein identifications for all selection schemes at protein identification false discovery rate < 0.05. The performance of the exemplary schemes and ProteinProphet are plotted according to their color code in (a). While three groups can be discerned, the clearly detached top group only considers selection schemes retaining single hit identifications.

In summary, the investigated protein identification properties are powerful indicators of protein identification quality and therefore represent a promising starting point to selectively prune protein identification sets in large to very large datasets to enrich for valid protein identifications.

5.8.2 Pruning protein identifications does not enhance inference

The foregoing analysis suggests that a selection scheme that consequently excludes single hits might enrich for correct protein identifications (**5.4a**, w/o single hits). However, it might be more beneficial to opt for a selection scheme that selectively retains high quality (very low false discovery rate) single hits (**5.4a**, w/o uncertain single hits). A variety of similar and more complex selection schemes that include or exclude protein identifications according to the “number of supporting peptide-spectrum matches” property can be considered to optimally explore the underlying dataset.

We systematically searched for optimal selection schemes, i.e. selection schemes that maximized the number of expected true positive protein identifications for a desired false discovery rate. Regardless of the desired protein identification false discovery rate, optimal selection schemes turned out to be the ones that consider all high quality peptide-spectrum matches, irrespective of the “number of supporting peptide-spectrum matches” property (**5.4a**, optimal). For instance, the optimal protein identification set featuring false discovery rate of 0.015 is compiled from the complete set of peptide-spectrum matches with false discovery rate of 0.0005. All other selection schemes turned out to be inferior. In particular our results clearly ruled out selection schemes that consequently neglected single hits, as well as selection schemes that selectively included protein identifications supported by a large number of low confidence peptide-spectrum matches (**5.4b**).

In summary, selection schemes considering all peptide-spectrum matches, including those giving rise to the less reliable single hits, turned out to be optimal for this dataset. However, peptide-spectrum matches have to be selected much more carefully than appreciated so far, in order to achieve reasonable protein identification false discovery rate for datasets of large size.

5.8.3 Simple protein inference engines are competitive

We compared “gene locus inference” and its selection scheme variants to ProteinProphet on the complete *C. elegans* dataset (**5.4a**). “Gene locus inference” without any pruning performs clearly better over the complete range of reasonable protein identification false discovery rates. ProteinProphet is also inferior to the single hit exclusion schemes for small false discovery rates, though outperforms these for less stringent identifcation quality requirements.

At a first glance it might be surprising that a simple protein inference strategy like “gene locus inference” outperforms a sophisticated inference engine like ProteinProphet. Considering that ProteinProphet effectively implements a probabilistically motivated selection scheme this result is though consistent with the forgoing analysis. To see this, consider ProteinProphet’s probabilistic model that is supposed to also recover proteins that are redundantly evidenced by less reliable fragment ion spectra. In the previous analysis, we systematically assessed all reasonable selection schemes for the complete *C. elegans* dataset and demonstrated them all to be inferior to the simple “gene locus inference” without any further pruning. These results suggest that, at least for large datasets like the *C. elegans* dataset, low quality spectra do not contribute novel protein identifications and they potentially mislead approaches that aim to exploit them as an additional information source.

5.9 Discussion

This work systematically assesses how pruning unreliable protein identification subsets affects protein inference performance [25]. An exploratory study investigated possibly unreliable protein identifications subsets by means of local false discovery rates. We further studied whether pruning such unreliable protein identifications is beneficial for protein inference performance. Therefore the concept the concept of selection schemes is introduced to enable a systematic enumeration of thousands of conceivable pruning strategies. In a second step a performance measure is introduced that allowed to generically compare the protein inference results obtained by each of the many pruning strategies. This measure evaluates the number of correct identifications and the involving false discovery rate. This measure is applied to benchmark pruning strategies defined by selection schemes computed for protein identifications obtained by the “gene locus inference” approach on the largest reported shotgun proteomics dataset for *C. elegans*.

This work reports the influence of various protein identification properties on identification false discovery rates. The number of peptide-spectrum matches supporting a protein identification has a severe impact on the identifications reliability. The following benchmark of protein inference engines therefore focuses on pruning strategies based on this property. The generic concept of selection schemes though also lends itself to define pruning strategies based on other protein identification properties. Future studies might benefit from incorporating these, too.

Our results confirm a recently published study that advocates to retain single hit wonders instead of discarding them [59] since these typically comprise many correct identifications. Here we studied these two pruning strategies among thousands of other strategies defined by the selection schemes. We could consistently rule out all conceivable pruning strategies to improve protein inference performance. While our heterogeneous dataset is presumably representative of most large scale shotgun proteomics datasets, it is still conceivable that for other datasets these conclusions do not hold. Consider for instance repetitive measurements of a large number of very similar samples. Such a scenario might result in a situation where true single hits become exceedingly rare, redeeming selection schemes that exclude single hits.

In this study we benchmark ProteinProphet and "gene locus inference" including its selection scheme variants. The approach is, however, equally applicable to a wide range of inference engines since the evaluation of the performance criterion is performed on the list of (target-decoy) protein identifications. This criterion can be easily evaluated since it simply involves the estimation of protein identification false discovery rates [26]. The result of a benchmark involving a larger number of competitors might vary across different application scenarios. Our approach enables the experimentalist to perform a benchmark and choice that is tailored to his application.

In order to ensure a fair comparison, competing protein inference engines should be comparable with respect to our performance measure. Since our performance measure rewards large numbers of correct protein identifications, the competitors should base their inference on a similar sized repertoire of possible protein identities. This requirement is mainly ensured by providing the same protein database to all competitors. However, it is conceivable that cases arise, where a protein inference engine would be intrinsically disadvantaged if it were to infer less resolved entities, such as e.g. exclusively gene loci, compared to competitors that could possibly report a larger number of higher resolved entities, such as e.g. splice variants[2].

For the protein inference engine benchmark, we optimize a trade-off between identification sensitivity and specificity. Despite being appealing, this objective is not necessarily always suitable. This is particularly the case where shotgun proteomics studies focus on a small set of proteins and aim to make explicit or even resolve possible ambiguities, such as e.g. gene products of a single gene locus. These cases necessitate protein inference engines like e.g. ProteinProphet [96] or IDPicker [148] that provide a protein grouping functionality.

---

[2] Note that this example does not apply to "gene locus inference" as introduced here. Given unambiguous peptide-spectrum matches, "gene locus inference" is well able to identify splice variants.

In the context of large shotgun proteomics projects aiming at extensive proteome coverage it is desirable to (1) decide upon the experiments that are expected to produce the most informative data, i.e. to most effectively explore a proteome [43, 15, 23, 24] and (2) to optimally evaluate the finally acquired data, i.e. to optimally perform protein inference. The presented benchmark approach contributes to the second step by generically enabling to choose the best protein inference engine. In our case we observe that processing the data with simple protein inference approaches and keeping all the spectral evidence achieves competitive proteome coverage.

Part II

Design of Proteome Measurements

6 Proteome Coverage Prediction with Infinite Markov Models

6.1 Summary

Liquid chromatography tandem mass spectrometry (LC-MS/MS) is the predominant method to comprehensively characterize complex protein mixtures such as samples from pre-fractionated or complete proteomes. In order to maximize proteome coverage for the studied sample, i.e. identify as many traceable proteins as possible, LC-MS/MS experiments are typically repeated extensively and the results combined. Proteome coverage prediction is the task of estimating the number of peptide discoveries of future LC-MS/MS experiments. Proteome coverage prediction is important to enhance the design of efficient proteomics studies. To date, there does not exist any method to reliably estimate the increase of proteome coverage at an early stage.

We propose an extended infinite Markov model DiriSim to extrapolate the progression of proteome coverage based on a small number of already performed LC-MS/MS experiments. The method explicitly accounts for the uncertainty of peptide identifications. We tested DiriSim on a set of 37 LC-MS/MS experiments of a complete proteome sample and demonstrated that DiriSim correctly predicts the coverage progression already from a small subset of experiments. The predicted progression enabled us to specify maximal coverage for the test sample. We demonstrated that quality requirements on the final proteome map impose an upper bound on the number of useful experiment repetitions and limit the achievable proteome coverage.

6.2 Introduction

Over the last few years mass spectrometry based proteomics has emerged as the most powerful approach to comprehensively characterize a proteome. The experimental workflows for mass spectrometry based proteomics have sufficiently advanced to enable ex-

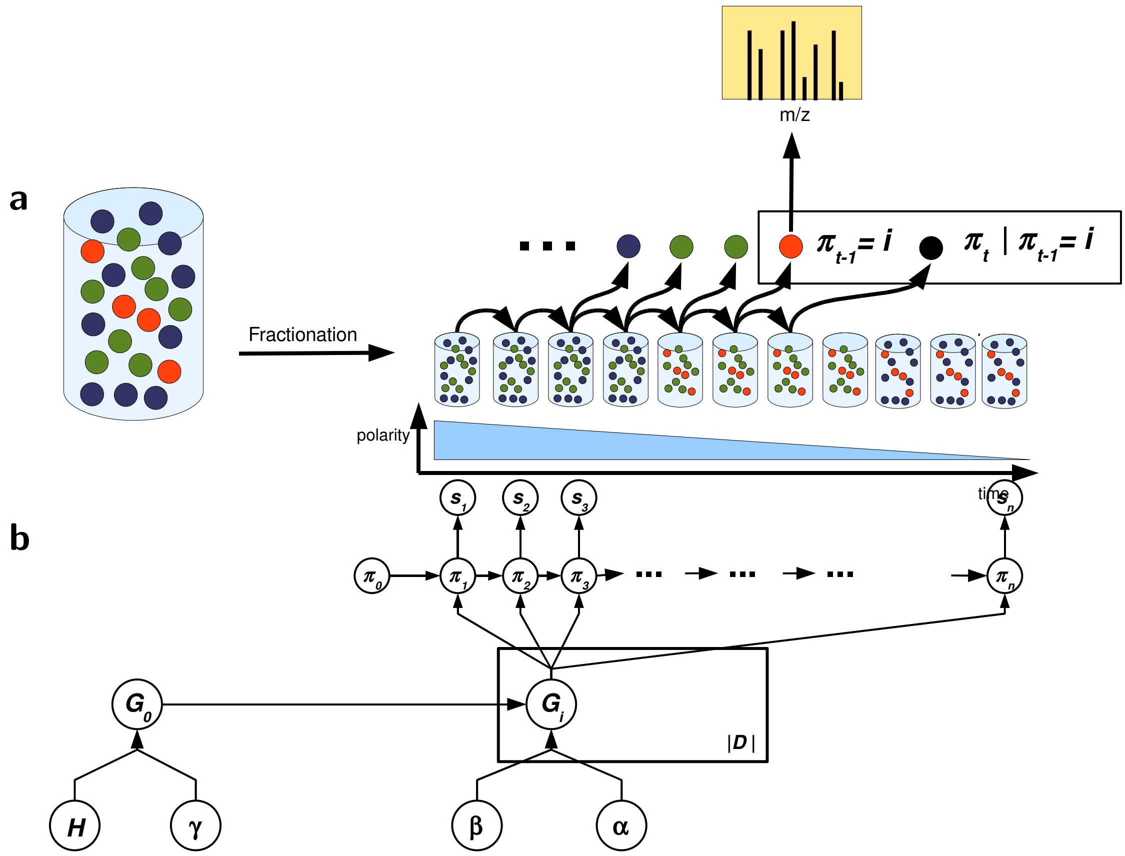


Figure 6.1: Illustration of an LC-MS/MS experiment. (**a**) Liquid chromatography fractionation generates a sequence of local peptide ensembles from the initial ensemble. Each of these ensembles is derived from the initial ensemble by pooling peptides of similar polarity. The sequence of ensembles features descending overall polarity in the course of the experiment. During the experiment peptides π_t are drawn from the sequence of ensembles and analyzed by the mass spectrometer coupled to the liquid chromatography system. (**b**) Graphical representation of the infinite Markov model. The initial ensemble is represented by its peptide distribution G_0. G_0 is assumed to have a Dirichlet process prior with concentration parameter γ and uniform distribution H over the protein database $\mathcal{D}$ as base probability measure. Local ensembles for which representative peptides have been detected are represented explicitly. Each of these ensembles is indexed by its representative peptide i and characterized by its peptide distribution G_i. G_i is assumed to be sampled from a biased Dirichlet process with G_0 as base probability measure. The peptide π_t following the series $\pi_1, ..., \pi_{t-1} = i$ of detected peptides is sampled from G_i. Each peptide π_t gives rise to an observable fragment ion spectrum s_t, defining the peptide-spectrum match (s_t, π_t).

tensive exploration of complex biological samples ([38]). While conceptional studies provided rough a priori insights about the scope of these workflows ([43]), there are still no means to dynamically infer the a posteriori potential, i.e. to predict the increase in proteome coverage for their real-world implementations. This work contributes the extended infinite Markov model DiriSim to predict proteome coverage (in terms of peptide discoveries) upon repetition of liquid chromatography tandem mass spectrometry (LC-MS/MS) experiments. By explicitly modeling false and true positive peptide identifications, DiriSim enables us to specify the maximally achievable proteome coverage for a specified quality constraint on the final set of peptide discoveries.

The most successful strategy to achieve extensive proteome coverage is referred to as shotgun proteomics. In its simplest implementation, protein samples are extracted from their biological source, subjected to enzymatic digestion and the resulting peptide mixtures are finally analyzed by LC-MS/MS. More elaborate strategies essentially adopt the same workflow, additionally augmented by fractionation steps for proteins/peptides before LC-MS/MS analysis. Finally, peptide identities are inferred from the acquired fragment ion spectra and they are used to recover the protein composition of the initial biological sample.

The complexity of the protein, and hence peptide mixtures, poses a formidable challenge to mass spectrometrical analysis. The reversed phase liquid chromatography step effectively reduces the complexity of the peptide mixture by selecting peptides for tandem mass spectrometry analysis according to their polarity. For the duration of the LC-MS/MS experiment the mass spectrometer coupled to the liquid chromatography system constantly acquires tandem mass spectra from eluting peptides. The elution time of a particular peptide is defined by its polarity. Any time during the LC-MS/MS experiment, the mass spectrometer is thus exposed to a local peptide mixture that is less complex than the initial mixture (**Fig. 6.1a**). Nevertheless, these mixtures are typically still far too complex to allow the mass spectrometer to acquire tandem mass spectra for all peptides in a single LC-MS/MS experiment. Consequently, LC-MS/MS experiments are usually repeated extensively, in order to increase the number of peptides for which tandem mass spectra are acquired.

Using one of a range of database search engines, tandem mass spectra are then assigned to peptide giving rise to a series of peptide-spectrum matches ([97]). Note that peptide-spectrum matches are typically highly redundant, i.e. the number of peptide discoveries covered by the peptide-spectrum matches is typically much smaller than the total number of peptide-spectrum matches. Not all peptide-spectrum matches are correct. Various approaches are available to estimate the reliability of peptide-spectrum matches ([72, 41]). Target-decoy strategies have shown to be a generic and reliable strategy to estimate false discovery rates for peptide-spectrum matches, i.e. the expected fraction of false positive peptide assignments ([41]). At this point, the preliminary result of a series of

LC-MS/MS experiments reduces to a series of peptide-spectrum matches that is additionally characterized by some false discovery rate.

Shotgun proteomics studies should ideally be designed such that proteome coverage, i.e. discovered peptides increases efficiently with consecutive measurements. For a given series of already performed LC-MS/MS experiments this requirement translates into the task of estimating the required number of additional experiments that have to be performed to achieve a reasonable increase in proteome coverage. If the estimated effort turns out to be too large, it might be more convenient to consider other experimental setups to analyze the underlying sample. Besides simply giving existing workflows a try, there have been approaches to rationally design promising setups according to statistical analysis of the already acquired peptide-spectrum matches ([15]). To the best of our knowledge, no method specifies the remaining potential of the currently performed experiments by predicting their proteome coverage progression.

To close this gap, we present DiriSim, an extended infinite Markov model for LC-MS/MS experiments that yields a posterior prediction of the proteome coverage progression. DiriSim explicitly accounts for true and false positive peptide-spectrum matches by modeling a set of LC-MS/MS experiments as a mixture of an infinite Markov model ([8]) and an error model distribution. The expected proteome coverage progression for additional experiments is estimated by sampling from the posterior predictive distribution. We have assessed this approach by cross validation on a set of 37 LC-MS/MS measurements of a complete proteome sample. We show that the extended infinite Markov model outperforms simple extrapolation methods and correctly predicts proteome coverage progression. Extrapolation of the proteome coverage progression further enabled us to specify the maximal coverage of the test set.

Methods

The data utilized by DiriSim consists of a list of LC-MS/MS experiments where peptide-spectrum matches have been generated by searching against a protein database $\mathcal{D}$. Each peptide-spectrum match (s, π) corresponds to a tandem mass spectrum s and its peptide assignment $\pi \in \mathcal{D}$. Each LC-MS/MS experiment R_l defines a series of n_l peptide assignments $\boldsymbol{\pi}^{(l)} = \pi_1^{(l)}, ..., \pi_{n_l}^{(l)}$. A fraction q of all peptide-spectrum matches is assumed to be erroneously assigned.

The following sections describe how to predict the progression of proteome coverage conditioned on the given data. In summary, this estimate is achieved by sampling from the posterior predictive distribution given a series of LC-MS/MS experiments and counting the amount of newly discovered peptides.

Section 6.3 briefly introduces Dirichlet processes and how these can be used to formally characterize peptide distributions arising in shotgun proteomics experiments. Section 6.4 characterizes the distribution from which peptides are sampled during an ideal LC-MS/MS experiment without false positive peptide-spectrum matches. Section 6.5 describes how to sample a series of peptides from such a distribution. Section 6.6 first describes how to sample from this distribution conditioned on the given data and second how to predict the progression of proteome coverage from the a posteriori sampled trajectories. Section 6.7 completes the framework description by introducing a component accounting for false positive peptide-spectrum matches.

Unless otherwise noted, $\boldsymbol{\pi}$ will in the following denote a series of sampled peptides π_t. Capital italic Latin letters like G, H will denote distributions.

6.3 Dirichlet processes priors for peptide distributions

In the course of a shotgun proteomics experiment peptides are sampled from an unknown distribution and then identified by mass spectrometrical analysis. This distribution is defined by the biological sample contributing a characteristic set of proteins/peptides and by the experimental setup enriching/depleting particular types of proteins/peptides. The more samples we draw from this distribution, i.e. the more experiments we perform, the better we are able to characterize the distribution and thereby predict the future progression of peptide discoveries.

The incremental estimation procedure is captured by a non-parametric Bayesian technique, denoted as *Chinese restaurant processes* ([12]). The Chinese restaurant process can be envisioned as a schematic task where n customers are to be seated in a restaurant with an infinite number of tables. At each table a particular dish is served that is denoted by its number in the menu. The first customer is seated at the first table and offered the corresponding dish π_1. The t-th subsequent customer is offered his dish π_t after having been seated either at an already populated table or at a new unpopulated table according to the following probabilities:

$$P(\pi_t = i \mid \pi_1, ..., \pi_{t-1}, \gamma) \quad = \quad \begin{cases} \frac{n_i}{t-1+\gamma} & \text{populated table} \\ \frac{\gamma}{t-1+\gamma} & \text{next unpopulated table} \end{cases} \tag{6.1}$$

where n_i corresponds to the number of customers already sitting at the table serving dish i. In case a customer happens to be seated at a new table, the dish served at this table is drawn from the base probability measure H. γ is referred to as the concentration parameter of the process. The larger γ, the higher the chances that a new customer is seated at a new table. The more customers have already been seated, the less likely it will open up a new table.

Let us now assume that we do not know γ and have seated n customers. We want to estimate how many tables will be occupied, or equivalently how many different dishes will be served after m additional customers have been seated. In a first step we characterize the seating distribution by fitting γ according to the observed seating arrangement, i.e. the more tables we find populated the larger we choose γ. We can now simulate m additional seating events using the γ estimate and thereby estimate the number of tables occupied afterwards.

By identifying dishes with peptides and respectively customers with mass spectra, we obtain a simple model to sample peptide assignments, i.e. simulate experiments and in particular estimate the expected number of new peptide discoveries. Although being overly simple, this model captures an essential property of shotgun proteomics experiments. While always allowing to discover a novel peptide with non-zero probability, the overall progression of new discoveries slows down for a growing number of experiments.

It turns out that a Chinese restaurant process with concentration parameter γ implements draws π_t from a discrete distribution G that itself is drawn from a prior distribution referred to as Dirichlet process DP with concentration parameter γ and base probability measure H ([46, 4]).

$$\begin{aligned} G \mid \gamma, H &\sim \mathrm{DP}(\gamma, H) \\ \pi_t \mid G &\sim G \end{aligned} \tag{6.2}$$

Dirichlet processes have proven to be useful to formally express and deal with the uncertainty of an unknown discrete distribution, e.g. mixing distributions of mixture models. In this work we assume Dirichlet process priors for distributions over peptides and sample from them by using the Chinese restaurant process construction.

6.4 Infinite Markov model for LC-MS/MS experiments

During an LC-MS/MS experiment, peptides designated for tandem mass spectrometry are sampled from a multitude of unknown distributions (**Fig. 6.1**). This section describes how to model these distributions with an infinite Markov model.

The peptides in the initial ensemble are distributed according to an unknown discrete distribution G_0. We assume a Dirichlet process prior $\mathrm{DP}(\gamma, H)$ for G_0 with base probability measure H and concentration parameter γ. H is assumed to be the uniform distribution over the peptides defined by the protein database $\mathcal{D}$. Note that the prior $\mathrm{DP}(\gamma, H)$ does not necessarily identify G_0 with H, i.e. the uniform distribution over the protein database $\mathcal{D}$.

Peptides are not directly sampled from G_0 in an LC-MS/MS experiment (**Fig. 6.1**). In the course of liquid chromatography, the mass spectrometer is exposed to a subpopulation of the initial ensemble, confined to members within a time dependent polarity range. Depending on the time point t, peptides are thus sampled from a characteristic peptide distribution G_t that is "related" to G_0. The prior for G_t has to capture the dependency on G_0. We particularly require the support of G_t to be contained in the support of G_0. While retaining flexibility, this requirement is met by choosing the prior for G_t to be a Dirichlet process with base probability measure G_0 and concentration parameter β ([134]).

Due to technical difficulties to reproduce absolute time courses for a series of LC-MS/MS experiments, we abstain from explicitly modeling polarity and, thereby, G_t. Instead we represent time or respectively ensemble polarity by peptide identities. We denote G_i as the local peptide distribution at the time points where peptide i has been identified. Assume that we have sampled $\pi_{t-1} = i$ in the course of an experiment. Since $\pi_{t-1} = i$ is indicative for the current polarity, we assume the subsequent peptide π_t to be sampled from the local distribution G_i (**Fig. 6.1**).

This representation induces a Markov chain whose states correspond to the identified peptides. We assume each state sequence $\boldsymbol{\pi}$ to begin at a distinguished start state π^*, i.e. we assume $\pi_0 \sim \delta_{\pi^*}$. Following ([8]), we define the prior of G_i to be a biased Dirichlet Process DP_i with base probability measure G_0, concentration parameter β and additional prior weight α on state i. Thereby, α explicitly controls the rate of sampling self-transitions $\pi_t = \pi_{t-1} = i$. Having a Dirichlet process prior on G_0, the number of sampled states is not fixed a priori and steadily grows with the number of sampled transitions. Due to the Dirichlet process prior on the local probability distributions G_i, the occurrence of transitions evolves in a similar fashion. We obtain the full characterization

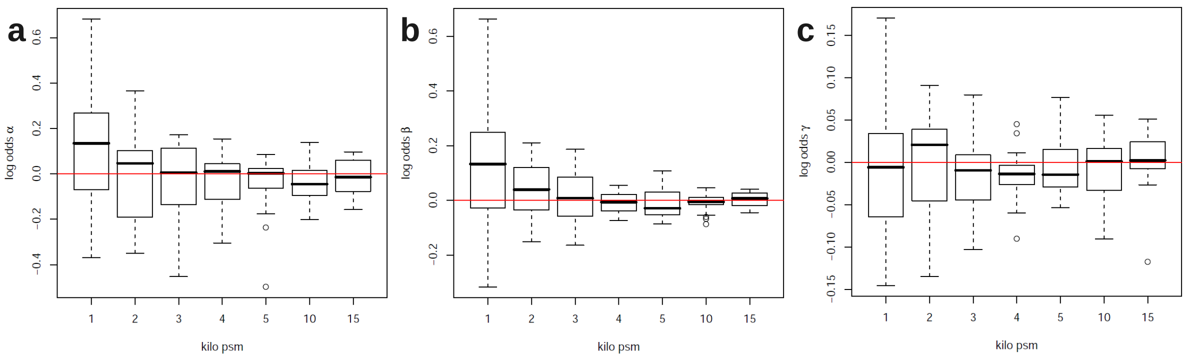


Figure 6.2: $\boldsymbol{\theta_{ML}}$ estimate on simulated data. Performance is evaluated for different training set sizes, i.e. series of peptide assignments (psm) of length ranging from 1000 to 15000. Performance is reported as log odds of predicted and true parameter value. Results are shown for parameters α, β, γ respectively governing the events of self-transitions (**a**), new transitions (**b**) and globally new discoveries (**c**). It can be seen that the parameters can be confidently estimated considering a training series of 10000 peptide assignments.

of the distribution that is sampled in the course of an LC-MS/MS experiment:

$$\begin{array}{lll} G_0 \mid \gamma, H & \sim & \mathrm{DP}(\gamma, H) \\ G_i \mid G_0 & \sim & \mathrm{DP}_i(\alpha, \beta, G_0) \\ \pi_t \mid \pi_{t-1} = i & \sim & G_i \\ \pi_0 & \sim & \delta_{\pi^*} \end{array} \tag{6.3}$$

6.5 Sampling sequences of peptide identifications

In the following we describe how to sample series of peptides from the distribution defined in the preceding section. Assume that $\alpha, \beta, \gamma, H, q$ are given and m series $\boldsymbol{\pi} = \boldsymbol{\pi}^{(1)}, ..., \boldsymbol{\pi}^{(m)}$ are to be sampled sequentially.

We assume each series $\boldsymbol{\pi}^{(k)}$ to begin at a distinguished start state π^*. $\boldsymbol{\pi}$ can be sampled in ascending order. To see this, assume that we already sampled the trajectory $\pi_0, \pi_1, ..., \pi_{t-1}$. In order to sample the subsequent peptide we have to specify the distribution for $\pi_t \mid \pi_0, \pi_1, ..., \pi_{t-1}, \alpha, \beta, \gamma, H$. Starting from the hierarchy of Dirichlet processes (6.3) and after integrating out $G_{\pi_{t-1}}$ and G_0 we obtain a nested variant of the

Chinese restaurant process construction (6.1) for the infinite Markov model:

$$P(\pi_t = j \mid \pi_0, \pi_1, ..., \pi_{t-1} = i, \alpha, \beta, \gamma, H) = \\ = \begin{cases} [\, n_{ii}(t) + \alpha \,] \cdot T_i(t) & \text{self} \\ [\, n_{ij}(t) \,] \cdot T_i(t) & \text{non-self} \\ [\, \beta \cdot [\, n^o_j(t) \,] \cdot T^o(t) \,] \cdot T_i(t) & \text{new target} \\ [\, \beta \cdot [\, \gamma \,] \cdot T^o(t) \,] \cdot T_i(t) & \text{new state} \end{cases} \tag{6.4}$$

$n_{ij}(t)$ corresponds to the number of occurrences of observing the transition from peptide i to peptide j in the series $\pi_0, ..., \pi_{t-2}$. $n^o_j(t)$ denotes how many times peptide j has been observed as a new transition target in the series $\pi_0, ..., \pi_{t-1}$. $T_i(t)$ is shorthand for $(\sum_j n_{ij}(t) + \alpha + \beta)^{-1}$ and $T^o(t)$ for $(\sum_j n^o_j + \gamma)^{-1}$.

The outcome “self” denotes to a self-transitions $\pi_t = \pi_{t-1}$. Accordingly, “non-self” corresponds to already observed transitions $\pi_t \neq \pi_{t-1}$. Note the distinguished role of self-transitions by the prior weight α. While the event “new target” refers to the discovery of a new transition to a peptide already observed in another context, “new state” denotes the discovery of a yet unobserved peptide. It is straight forward to sample the random variable $\pi_t \mid \pi_0, \pi_1, ..., \pi_{t-1} = i, \alpha, \beta, \gamma, H$ since its distribution has a closed form and only depends on the given parameters and quantities defined by the series of preceding peptide assignments.

6.6 Posterior prediction of proteome coverage progression

This section describes how to sample peptide series conditioned on already observed series. This task translates to sampling the posterior predictive distribution for $\boldsymbol{\pi}_{new}$ given the observed peptides $\boldsymbol{\pi}$. Proteome coverage progression for future experiments is estimated by approximating the expected number $E\left[|\mathcal{U}(\boldsymbol{\pi}_{new})| \mid \boldsymbol{\pi}, H\right]$ of new peptide discoveries $\mathcal{U}(\boldsymbol{\pi}_{new})$ upon posterior predictive sampling.

The posterior predictive distribution for $\boldsymbol{\pi}_{new} \mid \boldsymbol{\pi}, H$ has no closed form. For sufficiently large series $\boldsymbol{\pi}$, the posterior predictive distribution can be reasonably approximated by $\boldsymbol{\pi}_{new} \mid \boldsymbol{\pi}, \boldsymbol{\theta}_{ML}, H$ where $\boldsymbol{\theta}_{ML}$ corresponds to the maximum likelihood estimate for

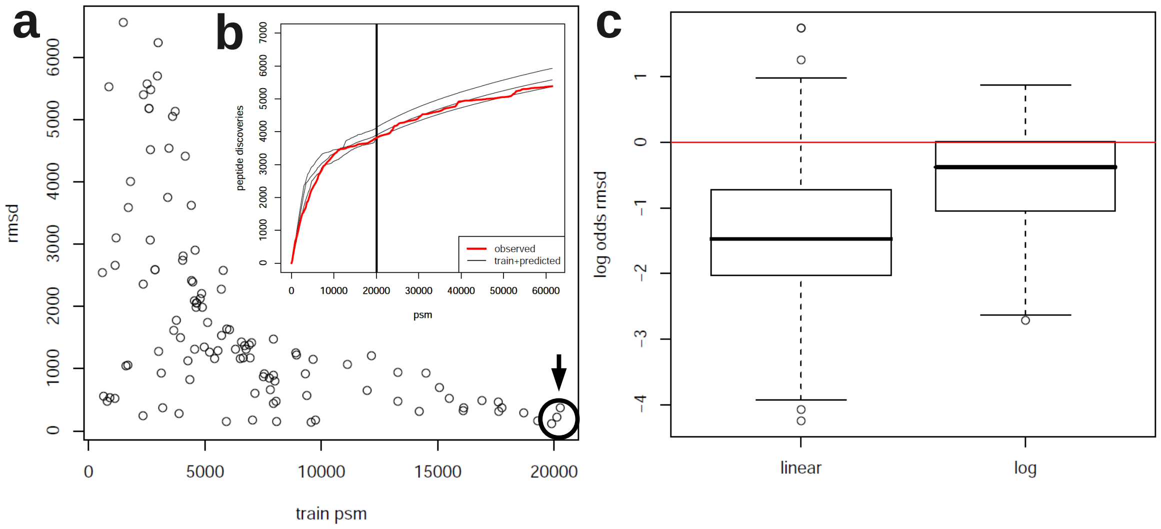


Figure 6.3: Prediction of proteome coverage progression for a data set comprising 37 LC-MS/MS experiments each giving rise to a series of peptide assignments (psm). We generated 120 training series of varying size (train psm) by subsampling complete LC-MS/MS experiments. We predicted the progression of proteome coverage (peptide discoveries) for each training series and compared to the progression observed for the series of the complete data set. (**a**) Prediction accuracy for the 120 training series. Prediction accuracy is given as root mean square deviation (rmsd) from the observed progression of peptide discoveries. (**b**) Concatenated training and respective predicted progressions (black) from the largest three training series (corresponding items in (a) are encircled) compared to observed progression (red). Vertical lines denote the size of the training series. Vertical lines overlap due to similar sizes around 20000. (**c**) Comparison of DiriSim with linear extrapolation of proteome coverage progression of last LC-MS/MS experiment in training series (linear) or respectively extrapolation of logarithmic regression of training series (log). Box plot of log odds of rmsd ($\log(\mathrm{rmsd}_{\mathrm{DiriSim}}/\mathrm{rmsd}_{\mathrm{compare}})$) for DiriSim and compared method (linear, log) on the 120 training series. Median log odds for comparison with the extrapolation methods linear and log are lower than zero, indicating weaker performance than DiriSim.

$\boldsymbol{\theta} := (\alpha, \beta, \gamma)$ based on the seating event probabilities in equation (6.4).

$$\boldsymbol{\theta}_{ML} \quad = \quad \arg\max_{\boldsymbol{\theta}} \prod_{t=1}^{n} P(\pi_t \mid \pi_0, ..., \pi_{t-1} = i, \boldsymbol{\theta}, H) \tag{6.5}$$

We predict the proteome coverage progression by approximating $E\left[|\mathcal{U}(\boldsymbol{\pi}_{new})| \mid \boldsymbol{\pi}, H\right]$ by averaging over a set of trajectories $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, ...$ sampled from $\boldsymbol{\pi}_{new} \mid \boldsymbol{\pi}, \boldsymbol{\theta}_{ML}, H$ as described in section 6.5.

6.7 Proteome coverage progression with false identifications

Sequences $\boldsymbol{\pi}$ of peptide assignments were assumed to be perfect in the preceding sections. Obviously this assumption does not hold in practice. This section describes an extension of the infinite Markov model by an error model that is able to deal with series of peptide assignments that are afflicted with a nonzero false discovery rate q.

We observe that false positive peptide assignments map to the decoy database in a non-redundant fashion, i.e. 83% of all decoy peptide discoveries of the test data set (see Results) are supported only by a single peptide assignment. Assuming that false positive peptide assignments distribute like decoy peptide assignments ([41]), we approximate the distribution of false positive peptide assignments with H, i.e. the uniform distribution over the protein database. In order to model the fraction q of false positive peptide assignments, we assume that peptide assignments are sampled from a mixture model with two components. The first component accounting for the true positive peptide assignments is given by the infinite Markov model as described in section 6.4. The second component is given by the distribution of false positive peptide assignments, i.e. H. Component weights are chosen according to the false discovery rate q. Consequently, the first and second component are weighted $1-q$ or q respectively.

Series of peptide assignments are generated by sampling each peptide assignment π_t either from the infinite Markov model as described in section 6.5 or directly from H, according to the components weights. Posterior sampling requires the estimate $\boldsymbol{\theta}_{ML}$ from an already observed series $\boldsymbol{\pi}$. Exact computation of $\boldsymbol{\theta}_{ML}$ though involves an intractable sum over configurations of false positive peptide assignments. We approximate $\boldsymbol{\theta}_{ML}$ by assuming that the number of false positive peptide assignments equals the expected value $n(1-q)$ and that these distribute uniformly over $\boldsymbol{\pi}$. This assumption allows us to approximate $P(\boldsymbol{\pi} \mid \boldsymbol{\theta}, H, q)$ with adjusted transition counts, e.g. $\hat{n}_{ij} := (1-q)n_{ij}$.

$$\boldsymbol{\theta}_{ML} \approx \arg\max_{\boldsymbol{\theta}} \prod_{t=1}^{n} P(\pi_t \mid \hat{n}_{ij}(t), \hat{n}_i(t), \hat{n}^o_j(t), \hat{n}^o(t), \boldsymbol{\theta}, H) \tag{6.6}$$

Proteome coverage progression is then predicted as described in section 6.6.

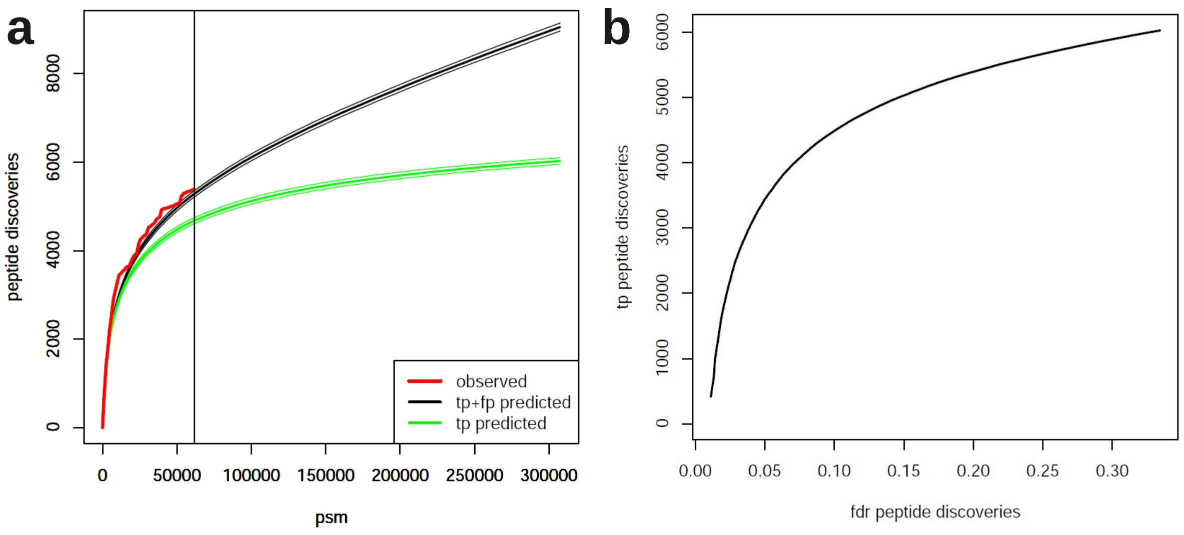


Figure 6.4: Five fold extrapolation beyond the range of the test data set (61582 peptide-spectrum matches). (**a**) Observed progression of the test data set in red, predicted progression with standard deviations of all (black) and only true positive (green) peptide discoveries. The progression of true positive discoveries stagnates considerably. (**b**) relates the absolute number of true positive (tp) peptide discoveries to the fraction of false positive discoveries (fdr peptide discoveries). The fraction of false positive peptide discoveries grows steadily with the total amount of peptide discoveries. Quality requirements on the final set of peptide discoveries limit the maximally achievable proteome coverage as well as the sensible number of LC-MS/MS experiments.

6.8 Results

In the following we show results that first, demonstrate that prediction of proteome coverage progression is a non-trivial task that is not solved satisfactory by simple extrapolation methods and second, that the extended infinite Markov model can confidently predict proteome coverage progression from a small number of already performed experiments and third, that we can identify the putative number of LC-MS/MS experiments to be carried out until reaching maximal coverage.

6.8.1 Simulation study for parameter estimation assessment

We conducted simulation studies to ensure that we can confidently estimate α, β, γ. Therefore we generated a data set by simulating peptide series with false discovery rate of 1% as described in section 6.7. Parameters α, β, γ were chosen in a range also observed in the real-world test data set that is introduced later. We assessed the estimates on 20 simulated series, each corresponding to multiple LC-MS/MS experiments. Each set of 20 series was chosen to be of length ranging from 1000 to 15000 peptide assignments. For each of these series we estimated α, β, γ as described in 6.6 and 6.7 (**Fig. 6.2**). It can be seen that α, β, γ can be reasonably recovered even from the smallest training series. The larger the series grows the more precise the estimates become. The approximations introduced in section 6.7 to account for false positive peptide assignments do not compromise the parameter estimates. Considering the equivalent of six or more LC-MS/MS experiments already yielded satisfactory estimates.

6.8.2 Cross validation prediction accuracy

We assessed DiriSim's ability to predict proteome coverage progression for real LC-MS/MS experiments. We consider proteome coverage to be the number of peptide discoveries, i.e. the number of different peptides represented in the series of peptide assignments. We were particularly interested to see how many LC-MS/MS experiments are needed to confidently extrapolate the progression of peptide discoveries. We expected that confident extrapolation is feasible after training DiriSim on a small training series of peptide assignments corresponding to a small number of LC-MS/MS experiments.

To this end, we applied DiriSim to a test data set covering 37 LC-MS/MS experiments of the complete *D. melanogaster* proteome ([118]). Peptide-spectrum matches were generated by searching against a target-decoy protein database (tryptic, ≤ 1 missed cleavage), for details see ([118]). For our study, we selected top-scoring peptide-spectrum matches mapping to the target database at a false discovery rate of 1% as described in [41]. By this means, we finally considered 61582 peptide-spectrum matches. We generated training series of varying size by subsampling the data set, extrapolated the progression of peptide discoveries for each training series and compared to the observed progression of the complete data set.

In total, we subsampled 120 training series of peptide assignments. Note that the subsampling procedure has to preserve the peptide assignments order within the individual LC-MS/MS experiments. Therefore we generated the training series by subsampling complete LC-MS/MS experiments. We subsampled 1, 2, 3, 4, 5 and 10 LC-MS/MS experiments, giving rise to 6 training series of peptide assignments. By repeating this step 20 times we generated a total of 120 training series. For instance, one of the training series comprised the series of 1139 peptide assignments defined by the 2 LC-MS/MS experiments with index 14 and 18 (out of all 37 experiments). The 120 training series varied in size, ranging from 596 to 20277 peptide assignments, i.e. covering up to one third of the complete data set's peptide assignments. Note that two training series that were generated by subsampling the same number of LC-MS/MS experiments do not necessarily comprise the same number of peptide assignments. This is due to the heterogeneous number of peptide assignments contributed by the individual LC-MS/MS experiments.

We extrapolated the progression of peptide discoveries for each training sequence and compared to the observed progression of the complete data set. Therefore, we estimated α, β, γ and estimated the expected proteome coverage progression by averaging over 50 series sampled from the posterior predictive distribution of the extended infinite Markov model (see sections 6.6 and 6.7). Goodness of the prediction was evaluated as root mean square deviation from the observed progression of the complete data set. Training series in corresponding to six or more average LC-MS/MS experiments (~ 1600 peptide assignments) yield good matches (**Fig. 6.3 a,b**). These results demonstrate that first, the principles governing the yield of LC-MS/MS experiments seem to be well captured by the extended infinite Markov model and second, proteome coverage progression can be confidently predicted from a considerably small set of experiments.

6.8.3 Proteome coverage prediction benchmark

We compared DiriSim with other extrapolation methods. We chose two simple general purpose extrapolation methods since there do not exist specific methods for proteome coverage prediction. We first considered an extrapolation scheme that linearly extrapolated proteome coverage progression of the last LC-MS/MS experiment of a training series. Second, we considered the extrapolation of a logarithmic regression ($y = a \log x + b$). We assessed prediction performance on the 120 training series as described above and observed that DiriSim clearly outperforms both extrapolation methods (**Fig. 6.3c**). These results indicate that proteome coverage prediction is a non-trivial task that is not solved satisfactory by ad hoc extrapolation methods.

6.8.4 Prediction of maximal proteome coverage

We further extrapolated the coverage progression five fold beyond the range covered by the test data set (**Fig. 6.4a**). The progression of peptide discoveries for all peptide assignments shows a linear increase. Since DiriSim explicitly models true and false positive samples, we could exclusively monitor the series of true positive peptide assignments. We observe a pronounced divergence of the progression for all assignments and the exclusively true positive ones. We particularly see, that the progression of true positive discoveries stagnates considerably. While the fraction of false positive peptide assignments is constantly held at 1%, the fraction of false positive peptide discoveries at the end of the predicted progression amounts to more than 30%. The fraction of false positives among the novel discoveries beyond the range of the test set even surmounts 60%. Tolerating a limited amount of false positive peptide discoveries, bounds the maximal number of possible peptide discoveries as well as the number of experiments having to be performed (**Fig. 6.4b**). For instance, assume that we require that at most 15% of all peptide discoveries are false positive. This constraint restricts the maximally achievable coverage since we can discover at most 5000 distinct true positive peptides. According to **Fig. 6.4a** we will have reached this point after having acquired 90000 peptide assignments.

6.9 Discussion

To date, it is not clear beforehand how often to repeat an LC-MS/MS experiment on a single biological sample in order to efficiently achieve satisfactory proteome coverage. Furthermore, the maximally achievable proteome coverage with a particular method is not known. We address these issues by presenting DiriSim, a framework to predict the progression of proteome coverage for LC-MS/MS experiments.

DiriSim models a series of LC-MS/MS experiments as an infinite Markov model, whose states correspond to peptides. We apply DiriSim to extrapolate the proteome coverage progression of a small number of already performed LC-MS/MS experiments. Note that this task is different to the a posteriori inference of the state sequence of these experiments. In contrast to previous applications ([8, 124]), a posteriori inference of the state sequence is furthermore not necessary, since the states (peptides) are already assigned to the observable variables (tandem mass spectra) by means of the corresponding peptide-spectrum matches. Besides its application in proteome coverage prediction, the infinite Markov model could though serve as a prior in a Bayesian peptide identification setting and, in particular, prevent the accumulation of false positive peptide discoveries coming along with increasing data set size.

LC-MS/MS experiments are typically analyzed by database searching. The underlying protein databases are large but still of finite size and therefore define a finitely large set of possibly identified peptides. De novo sequencing approaches infer peptide identities without relying on protein databases and thereby implicitly support an infinite number of possible peptide identities. Using an appropriate base probability measure H, the proposed infinite Markov model for LC-MS/MS experiments naturally lends itself to predict the proteome coverage in this context.

We have shown that DiriSim correctly extrapolates proteome coverage progression from at most 10 LC-MS/MS experiments and outperforms ad hoc extrapolation methods. Proteome coverage prediction appears to be a non-trivial task due to the intricate dependency structure of an LC-MS/MS experiment. DiriSim provides a comprehensive non-parametric Bayesian characterization of an LC-MS/MS experiment that enabled us to confidently predict proteome coverage. Although capturing the dependencies of LC-MS/MS experiments, DiriSim remains a robust, non-complex model since it only needs three parameters that are to be learned from data.

By explicitly modeling false and true positive peptide assignments, DiriSim enables us to specify the maximally achievable proteome coverage with regards to true positive peptide discoveries. We have seen in the simulations that new peptide discoveries from extensive repetition of LC-MS/MS experiments mostly accumulate false positive discoveries. This observation reflects the difference between the distributions for true and false positive peptide assignments. While true positive peptide assignments concentrate over a small subset of the protein database, false positive peptide assignments distribute broadly over the protein database and therefore mostly contribute false positive peptide discoveries. Due to the exceedingly broad distribution of decoy matches, we do not expect that errors possibly introduced by the uniformity approximation compromise the observed accumulation of false positive peptide discoveries. We conclude that performing more and more experiments seeking for maximal coverage mainly deteriorates the overall quality of the complete peptide discovery set. Depending on the false discovery rate of the peptide assignments, a quality requirement on the set of peptide discoveries imposes an upper bound to the total number of experiments which therefore, potentially limits the maximally achievable proteome coverage before the progression of true positive peptide discoveries is fully saturated. This limitation accrues from the occurrence of erroneous peptide-spectrum matches and their broad distribution over the protein database. As long as peptide-spectrum matches are afflicted with uncertainty, this reasoning holds for any proteome being studied. It will though be interesting to apply DiriSim to other data sets in order to study the quantitative impact of factors like proteome size and experimental setup on the maximally achievable proteome coverage. In summary, our results suggest that the design of large shotgun proteomics studies should focus on efficiency not only to save resources but most importantly to yield reliable peptide discoveries.

7 The Fractal Dirichlet Process

7.1 Summary

Hierarchical Dirichlet processes are rich priors for ensembles of discrete distributions. These processes though do not explicitly account for the similarities among subsets of such ensembles. To this end, we propose a novel concept, the fractal Dirichlet process. Fractal Dirichlet processes generalize hierarchical Dirichlet processes by introducing self-referential base measures. We present an efficient Gibbs sampler for Bayesian parameter and hidden variable inference. We expect that explicitly accounting for similarity among distributions by means of fractal Dirichlet processes will add to various statistical learning tasks that benefit from hierarchical Dirichlet processes.

7.2 Introduction

Characterization of a set of related mixture distributions defines an essential task in diverse statistical learning scenarios, such as e.g. image segmentation, language modeling or proteome coverage prediction. Hierarchical Dirichlet processes address this task by providing a non-parametric Bayesian formalism that supports to globally share mixture components across the set of mixtures [134]. Here we present the fractal Dirichlet process that generalizes this formalism to explicitly capture similarity among subsets of mixture distributions.

Consider a situation where the data is partitioned into a set of J groups. Each group constitutes an exchangeable sequence of observations. Each observation within a group can thus be considered a conditionally independent draw from a latent variable mixture model. We optionally assume that latent variable assignments for each observation have already been inferred. We want to characterize the posterior distribution of the mixture model parameters and thereby make shared structure across the groups explicit.

Dirichlet processes are non-parametric Bayesian priors to characterize discrete distributions [46]. A Dirichlet process $\mathrm{DP}(\gamma, G)$ defines a probability measure on probability measures and it is characterized by two parameters, a concentration parameter γ and a base probability measure G. Briefly, draws from $\mathrm{DP}(\gamma, G)$ give rise to distributions that are similar to the base measure G to an extent defined by the concentration parameter γ. Since measures drawn from a Dirichlet process are discrete (with probability one), these processes serve as priors for mixing distributions of mixture models by associating mixture components to atoms of a Dirichlet process draw [122, 4].

Hierarchical Dirichlet processes in the sense of [134] have been proposed as versatile priors over a set of related discrete distributions. A Dirichlet process prior is assumed for each distribution. By making these processes share the same discrete base measure (also drawn from a Dirichlet process) the set of distributions may share atoms or, in the context of a mixture model, mixture components. Hierarchical Dirichlet processes have been widely used to characterize sets of related mixing distributions in applications like e.g. language modeling [133], topic modeling [13] and proteome coverage prediction [23].

Hierarchical Dirichlet processes are able to reveal nested relationships among subsets of distributions though they are not able to make non-nested relationships explicit. We propose the fractal Dirichlet process, a generalization of the hierarchical Dirichlet process that explicitly captures non-nested relationships among the random distributions by choosing self-referential base measures for the respective Dirichlet process priors. For inference we provide a Gibbs sampler based on the Chinese Restaurant construction.

The remaining manuscript is organized as follows. Section 7.3 briefly reviews Dirichlet processes and section 7.4 summarizes hierarchical Dirichlet processes in the sense of [134]. While section 7.5 introduces the fractal Dirichlet process, section 7.6 describes the Gibbs sampler based on the Chinese Restaurant construction. Section 7.7 presents the experimental results on synthetic data before we summarize and conclude this study.

7.3 Dirichlet processes

To render the chapter self-contained we will define Dirichlet processes in this section and discuss one of the constructive views on the Dirichlet process, i.e. the Chinese Restaurant construction.

A Dirichlet process DP(γ, H) is defined to be the distribution of a random probability measure G over a measurable space (Θ, Σ) with probability measure H such that, for any finite measurable partition $(B_1,B_2,...,B_r)$ of Θ, the random vector $(G(B_1),...,G(B_r))$ is distributed as a finite-dimensional Dirichlet distribution with parameters $(\gamma H(B_1),...,\gamma H(B_r))$. We write $G \mid \gamma, H \sim$ DP(γ, H) if G is a random probability measure with distribution given by the Dirichlet process [46].
The explicit construction of a measure drawn from a Dirichlet process provides another view on the Dirichlet process. Two popular construction schemes have been reported, the stick breaking construction [122] and the Pólya urn scheme, also known as the Chinese Restaurant construction [12]. Since we will not make use of the stick breaking construction, we focus on Chinese Restaurant construction here.

We want to construct a particular draw G from a Dirichlet process DP(γ, H). We achieve this goal incrementally by realizing an (infinite) sequence of i.i.d. random variables $\pi_1, \pi_2, ...$ that each follow G. Blackwell has shown that, given the parameters of the Dirichlet process and the realizations of the variables $\pi_1, ..., \pi_t$, marginalizing out G yields the following measure according to which $\pi := \pi_{t+1}$ is distributed [12].

$$\pi \mid \pi_1, ..., \pi_t, \gamma, H \sim (t+\gamma)^{-1} \left(\sum_{l=1}^{t} \delta_{\pi_l} + \gamma H \right) \tag{7.1}$$

π is though distributed according to a mixture. The larger the concentration parameter γ the more likely π is sampled from the component comprising the base measure H. Thus, the larger γ the more the finally constructed measure G will resemble the base measure H. This construction scheme exhibits a clustering property, i.e. allows for a newly sampled instance π to assume an already observed value with positive probability. Furthermore note the "rich-get-richer" type of behavior where the probability of sampling an already seen value ω scales linearly with the number of events of realizing ω. These observations can be made explicit by reformulating equation (7.1). Let $\boldsymbol{\omega} := \omega_1, ..., \omega_K$ be the set of values and $\boldsymbol{n} := n_1, ..., n_K$ denote the respective frequencies realized by the sequence $\boldsymbol{\pi} := \pi_1, ..., \pi_t$. We can now summarize equation (7.1) by the following mixture

$$\pi \mid \boldsymbol{n}, \boldsymbol{\omega}, \gamma, H \sim T(\boldsymbol{n}, \gamma) \left(\sum_{k=1}^{K} n_k \delta_{\omega_k} + \gamma H \right) \tag{7.2}$$

with $T(\boldsymbol{n}, \gamma)$ being an abbreviation for $\left(\sum_{k=1}^{K} n_k + \gamma \right)^{-1}$.
This construction scheme is also known as the Chinese restaurant construction. This naming convention becomes apparent when casting this scheme into a scenario where customers are seated according to (7.2) into a restaurant with an infinite number of tables. The sequence $\boldsymbol{n} := n_1, ..., n_K$ translates into the number of customers seated at tables $\boldsymbol{t} := t_1, ..., t_K$. Each table $t_k \in \boldsymbol{t}$ serves a single dish $\omega_k \in \boldsymbol{\omega}$ to its respective customers. A new customer entering the restaurant is assigned a table and dish according to (7.2). If this assignment involves sampling the dish from the component comprising the base measure H then a new table t_{K+1} with dish ω_{K+1} is opened and populated with the new customer. Note that for a discrete base measure it is possible to have several tables serving the same dish.

A finite number of seating events partially constructs a Dirichlet process sample, i.e. G. We represent a partial construction by its (Chinese) restaurant process $C := (\boldsymbol{n}, \boldsymbol{\omega}, \gamma, H)$. In the following we concisely refer to iterating the restaurant process if we sample an additional instance π according to (7.2) and thereby change the restaurant process accordingly.

Besides the Chinese restaurant construction, we furthermore studied construction scheme variants whose mixture weights do not depend linearly on the number of customers seated at the respective table.

$$\pi \mid \boldsymbol{n}, \boldsymbol{\omega}, \gamma, H \sim T_f(\boldsymbol{n}, \gamma) \left(\sum_{k=1}^{K} f(n_k) \delta_{\omega_k} + \gamma H \right) \tag{7.3}$$

f can now be an arbitrary function for which we only require homogeneity, i.e. $f(0) = 0$. f can for instance be chosen as $f(n) = \sqrt{n}$ or $f(n) = n^2$. The normalization constant is adapted accordingly: $T_f(\boldsymbol{n}, \gamma) := \left(\sum_{k=1}^{K} f(n_k) + \gamma \right)^{-1}$. It turns out that construction schemes with non-linear f generate sequences $\boldsymbol{\pi} := \pi_1, ..., \pi_t$ that are not exchangeable, i.e. for such schemes it does not hold for any permutation $\boldsymbol{\pi}'$ that the probabilities according to (7.3) coincide [77]. Since the exchangeability property is a necessary condition to generate a sequence of i.i.d. samples from an implicit multinomial, no construction scheme of the form (7.3) and non-linear f is suited as a suitable variant to construct a process.

7.4 Hierarchical Dirichlet processes

Hierarchical Dirichlet processes are sets of random measures. The measures in such a set are related in the sense that they share atoms with positive probability.

Specifically, a hierarchical Dirichlet process constitutes a distribution over a set of discrete measures $\boldsymbol{G}^c = (G^c_j)_{1\leq j\leq M}$. Each G_j is distributed according to a Dirichlet process with concentration parameter γ_c and base measure G^r, which itself is distributed according to a Dirichlet process with concentration parameter γ^r and base measure H.

$$\begin{aligned} G^r \mid \gamma^r, H \ &\sim \mathrm{DP}(\gamma^r, H) \\ G^c_j \mid \gamma^c, G^r &\sim \mathrm{DP}(\gamma^c, G^r) \end{aligned} \tag{7.4}$$

G^r is discrete since it is drawn from a Dirichlet process. Consequently, the G^c_j share atoms with positive probability since their respective Dirichlet process is endowed with a discrete base measure, i.e. G^r.

In the following we refer to the members of $\boldsymbol{G}^c$ as the child distributions and to G^r as the root distribution. Accordingly, we indicate parameters and variables related to the child/root distributions with the superscript naming convention. For instance, γ^c refers to the concentration parameter of the Dirichlet process that defines the probability measure for the child distributions G^c_j.

The Chinese restaurant franchise constitutes an explicit scheme to construct a set of measures $\boldsymbol{G}$ that is drawn from a hierarchical Dirichlet process with given concentration parameters γ^c, γ^r and base measure H. The Chinese restaurant franchise straightforwardly implements the Chinese Restaurant construction for the Dirichlet processes involved in (7.4). Each G^c_j is a draw from a Dirichlet process and can thus be constructed by iterating its restaurant process C^c_j. Following (7.2) and considering the conditional independence statement in (7.4) we obtain the distribution of π given the selected child j, its corresponding child process $C^c_j := (\boldsymbol{n}^c_j, \boldsymbol{\omega}^c_j, \gamma^c, G^r)$, the other child processes $C^c_{i\neq j}$

and the root process C^r. To keep the notation uncluttered, we write $\boldsymbol{n}^c := \boldsymbol{n}^c_1, \boldsymbol{n}^c_2, \ldots$ and $\boldsymbol{n} := (\boldsymbol{n}^c, \boldsymbol{n}^r)$ and $\boldsymbol{\omega} := (\boldsymbol{\omega}^c, \boldsymbol{\omega}^r)$ accordingly.

$$\pi \mid j, \boldsymbol{n}, \boldsymbol{\omega}, \gamma^c, G^r, (\gamma^r, H) \;\sim\; T(\boldsymbol{n}^c_j, \gamma^c) \left(\sum_{k=1}^{K^c_j} n^c_{jk} \delta_{\omega^c_{jk}} + \gamma^c G^r \right) \tag{7.5}$$

It is not possible to directly sample from the mixture component comprising G^r. If π turns out to be sampled from this component (event flagged by b) we can though proceed by integrating out G^r and iterating the restaurant process $C^r := (\boldsymbol{n}^r, \boldsymbol{\omega}^r, \gamma^r, H)$ corresponding to G^r on the basis of the following distribution.

$$\pi \mid b, \boldsymbol{n}, \boldsymbol{\omega}, \gamma^r, H \;\sim\; T(\boldsymbol{n}^r, \gamma^r) \left(\sum_{k=1}^{K^r} n^r_k \delta_{\omega^r_k} + \gamma^r H \right) \tag{7.6}$$

In this case we subsequently update both restaurant processes C^r and C^c_j.

Samples from hierarchical Dirichlet processes produce sets of similar (child) distributions by providing a mechanism that allows for sharing of atoms. Specifically, the child distributions $\boldsymbol{G}^c$ can share atoms globally defined by the root distribution G^r. The Chinese restaurant franchise construction makes this mechanism apparent by explicitly representing and relating the child and root restaurant processes.

7.5 Fractal Dirichlet processes

In this section we introduce the concept of the fractal Dirichlet process. It generalizes the hierarchical Dirichlet process to explicitly capture the similarity among subgroups of a set of distributions. As for the hierarchical Dirichlet process, we put Dirichlet priors on the distributions and extend this process by appropriately choosing their base measures.

Consider a set of (child) distributions $\boldsymbol{G}^c$ for which we assume Dirichlet process priors. In the previous section we have seen for the hierarchical Dirichlet process that the choice of the base measure establishes correspondences between members of these sets. Specifically, choosing the base measure discrete permits the child distribution to share atoms from the base measure with nonzero probability. For each G_j we now want to choose a base measure A_j that explicitly captures possible similarity to distribution subsets $\boldsymbol{G}' \subseteq \boldsymbol{G}$.

$$G_j^c \mid \gamma^c, A_j \sim \mathrm{DP}(\gamma^c, A_j) \tag{7.7}$$

To this end, we choose A_j as a self-referential mixture of all child distributions G_i^c with $i \neq j$ and the discrete root distribution G^r. This base measure enables G_j^c to share atoms with any other child distribution and the root distribution. The mixture weights $\boldsymbol{a}_j$ explicitly express to what extent G_j^c inherits atoms from the respective component.

$$A_j = a_{jj} G^r + \sum_{i \neq j} a_{ji} G_i^c \tag{7.8}$$

The mixture weights $\boldsymbol{a}_j$ constitute a discrete measure G_j^a. We treat G_j^a as a random measure with a (biased) Dirichlet process prior $\mathrm{DP}(\gamma^a, \alpha^a, F)$ that additionally puts prior weight α^a on the distinguished atom j [8]. The base measure F is defined over possible indices i of child distributions. Identifying F with G^r renders the fractal Dirichlet process amenable to define an non-parametric Markov chain [8, 134]. Furthermore, we assume a Dirichlet process prior for the root distribution G^r. This completes the specification of the fractal Dirichlet process. Note that the hierarchical Dirichlet process is a special case of the fractal Dirichlet process with $\alpha^a > 0$ and $\gamma^a = 0$. See also **Fig. 7.1** for a graphical model representation of the fractal Dirichlet process.

$$\begin{aligned} G^r \mid \gamma^r, H &\sim \mathrm{DP}(\gamma^r, H) \\ G_j^a \mid \gamma^a, \alpha^a, G^r &\sim \mathrm{DP}(\gamma^a, \alpha^a, F) \\ G_j^c \mid \gamma^c, A_j &\sim \mathrm{DP}(\gamma^c, A_j) \end{aligned} \tag{7.9}$$

Besides adopting the naming and notation introduced for hierarchical Dirichlet processes in section 7.4, we refer to the members of $\boldsymbol{G}^a := G_1^a, G_2^a, ...$ as the adapter distributions. Accordingly, we indicate parameters and variables related to the adapter distributions with the superscript ^{a}. Furthermore, we denote $\boldsymbol{A} := A_1, A_2, ...$

We provide a Chinese restaurant franchise scheme to construct a sample from a fractal Dirichlet process given the parameters $\boldsymbol{\gamma} := \gamma^r, \gamma^a, \alpha^a, \gamma^c$ and the base measures H, F. As for the hierarchical Dirichlet process, each G_j^c is a draw from a Dirichlet process and can thus be constructed by iterating a restaurant process C_j^c. Considering that we introduced the mixture A_j as base measure of the respective Dirichlet process prior, we iterate C_j^c analogously to (7.2). After augmenting the variable for the process frequencies

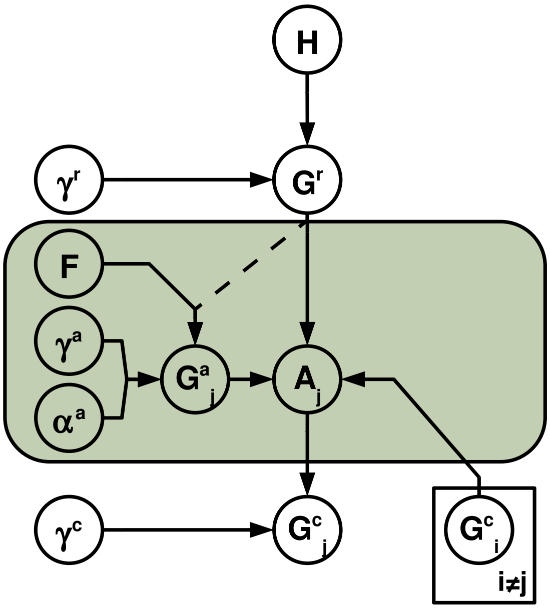


Figure 7.1: Graphical model representation of the fractal Dirichlet Process. For clarity, dependencies are depicted only for a single child distribution G^c_j and those for $G^c_{i \neq j}$ are omitted. Removing the components in the box and establishing the dependency of G^c_j from G^r yields the well established hierarchical Dirichlet process. Identifying the measure F with G^r results in an infinite fractal Markov chain (dotted arrow).

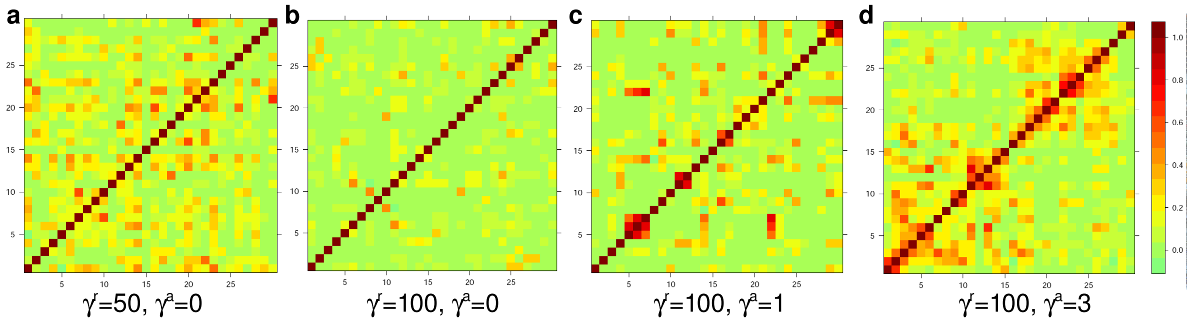
and labels by writing $\boldsymbol{n} := (\boldsymbol{n}^c, \boldsymbol{n}^a, \boldsymbol{n}^r)$ and $\boldsymbol{\omega} := (\boldsymbol{\omega}^c, \boldsymbol{\omega}^a, \boldsymbol{\omega}^r)$ we obtain.

$$\pi \mid j, \boldsymbol{n}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \boldsymbol{A}, \boldsymbol{G}^a, G^r, H, F \; \sim \; T(\boldsymbol{n}^c_j, \gamma^c) \left(\sum_{k=1}^{K^c_j} n^c_{jk} \delta_{\omega^c_{jk}} + \gamma^c A_j \right) \tag{7.10}$$

If π turns out to be sampled from this mixture component comprising A^j (event denoted by b) we first have to determine which of A^j's mixture components i is to be sampled. Therefore we have to sample i from adapter distribution G^a_j by iterating the respective restaurant process $C^a_j := (\boldsymbol{n}^a_j, \boldsymbol{\omega}^a_j, \gamma^a, \alpha^a, F)$.

$$i \mid j, \boldsymbol{n}, \boldsymbol{\omega}, \boldsymbol{\gamma}, F \; \sim \; T(\boldsymbol{n}^a_j, \gamma^a) \left(\sum_{k=1}^{K^a_j} n^a_{jk} \delta_{\omega^a_{jk}} + \gamma^a F \right) \tag{7.11}$$

If $i \neq j$ then π is recursively sampled as described in (7.10) with updated $j := i$. Otherwise, $\pi \mid b, (i = j) \boldsymbol{n}, \boldsymbol{\omega}, \boldsymbol{\gamma}, H$ is sampled directly from G^r by iterating the respective restaurant process C^r, analogously to (7.6). For each of the latter sampling events the



γ^r=50, γ^a=0 γ^r=100, γ^a=0 γ^r=100, γ^a=1 γ^r=100, γ^a=3

Figure 7.2: Pairwise similarity among the empirical child distributions obtained after running the Chinese Restaurant Franchise construction for various parameter settings. Similarity is reported in terms of Pearson correlation. **(a,b)** Parameter settings corresponding to the special case of the hierarchical Dirichlet process, i.e. $\gamma^a = 0$. **(c,d)** representative parameter settings for the fractal Dirichlet process. Pairwise similarity is more dispersed for the hierarchical Dirichlet process as for the fractal Dirichlet process sample.

respective restaurant processes are updated accordingly.

A single step of Chinese restaurant franchise construction for the fractal Dirichlet process is concisely summarized by the FRACTALFRANCHISE procedure. This representation is particularly suited to make the fractal structure of the process apparent. We write $\boldsymbol{C}$ for the set of all restaurant processes C^c_1, C^c_2, ..., C^a_1, C^a_2, ..., C^r.

```
FRACTALFRANCHISE($j, \boldsymbol{C}$)
$\pi \leftarrow$ ITERATE $\left(C^c_j\right)$                    # $\pi \neq$ null $\Rightarrow$ return
if $\pi =$ null
   then $i \leftarrow$ ITERATE $\left(C^a_j\right)$
        if $i = j$ then $\pi \leftarrow$ ITERATE $(C^r)$          # return
        else $\pi \leftarrow$ FRACTALFRANCHISE($i, \boldsymbol{C}$)   # recursion
        OPENTABLE $\left(C^c_j, \pi\right)$
return $\pi$
```

Samples from fractal Dirichlet processes produce sets of similar (child) distributions by generalizing the sharing mechanism implemented by hierarchical Dirichlet processes. By means of self-referential base measures of the Dirichlet process priors, the child distributions $\boldsymbol{G}^c$ can share atoms globally defined by the root distribution G^r, as well as directly among themselves. The structure of the fractal Chinese restaurant franchise construction makes this mechanism apparent by explicitly specifying the recursive relationship of the self-referential base measures.

7.6 Bayesian inference

We present a Gibbs sampler for posterior inference of the parameters of the fractal Dirichlet process and the hidden variables governing the Chinese restaurant franchise construction.

We start off with a sequence $\boldsymbol{s} := (s_t)_{1..n}$ whose elements $s_t := (j_t, \pi_t)$ correspond to variables π_t realized from the child distribution G_{j_t}. We want to sample from the posterior distribution $\boldsymbol{n}, \boldsymbol{\omega}, \boldsymbol{\gamma} \mid \boldsymbol{s}, H$ by first, specifying and second, cyclically sampling from conditional posterior distributions for the individual parameters or hidden variables.

The conditional posteriors for the parameters $\boldsymbol{\gamma}$ follow easily from the posterior of the parameters γ, α of a single (biased) Dirichlet process given a partial construction in terms of the respective restaurant process frequencies $\boldsymbol{n}$. Additionally placing Gamma priors on γ, α we obtain:

$$P(\gamma,\alpha \mid \boldsymbol{n}) \propto \mathcal{G}(a_\gamma, b_\gamma)\mathcal{G}(a_\alpha, b_\alpha) \cdot \frac{\Gamma\left(\alpha+\gamma\right) \cdot \gamma^K \cdot \Gamma(n_0+\alpha) \cdot \prod_{k=1}^{K-1} \Gamma(n_k)}{\Gamma\left(n+\alpha+\gamma\right)} \tag{7.12}$$

The respective posterior for a standard Dirichlet process follows by assuming $\alpha := 0$. This setting applies to the priors for the child and the root distributions. The posterior $P(\gamma, \alpha \mid \boldsymbol{n}_1, .., \boldsymbol{n}_M)$ for the parameters given $M > 1$ independent restaurant processes realizations simply evaluates as $\prod_{i=1}^{M} P(\gamma, \alpha \mid \boldsymbol{n}_i)$. This situation particularly applies to the parameter posterior given the adapter and child processes. Given the frequencies n in the restaurant processes, we can independently sample the parameters for each the child, adapter and root Dirichlet process. We apply adaptive rejection metropolis sampling [56] to sample from the respective conditional posterior distributions [111].

The conditional posteriors for the hidden variables $\boldsymbol{n}, \boldsymbol{\omega}$ are more involved to specify. For clarity we from now on consider the label vectors $\boldsymbol{\omega}$ to be absorbed into the frequency vectors $\boldsymbol{n}$ and omit explicit conditioning on H, F. We thus want to sample from $\boldsymbol{n} \mid \boldsymbol{s}, \boldsymbol{\gamma}$. To this end we construct $\boldsymbol{n}_t := \boldsymbol{n}$ iteratively by sampling $\boldsymbol{n}_1, \boldsymbol{n}_2, ...$ from $\boldsymbol{n}_{t'} \mid \boldsymbol{s}_{t'}, \boldsymbol{n}_{t'-1}, \boldsymbol{\gamma}$ where $\boldsymbol{n}_0 = \emptyset$.

In the following we will specify the events covered by the distribution of type $\boldsymbol{n}' \mid \pi, j, \boldsymbol{n}, \boldsymbol{\gamma}$. Sampling an event from this conditional distribution translates in sampling a sequence of restaurant process iterations that finally yields π. This sequence starts with iterating the restaurant process C_j^c. The configurations of all restaurant processes are determined by $\boldsymbol{n}$. Note that to realize $\boldsymbol{n}'$ we cannot simply run the Chinese restaurant franchise (effectively sampling $\boldsymbol{n}', \pi \mid j, \boldsymbol{n}, \boldsymbol{\gamma}$) since we are conditioning on the outcome π.

$$P(\boldsymbol{n}' \mid \pi, j, \boldsymbol{n}, \boldsymbol{\gamma}) = \frac{P(\boldsymbol{n}'\pi \mid j, \boldsymbol{n}, \boldsymbol{\gamma})}{P(\pi \mid j, \boldsymbol{n}, \boldsymbol{\gamma})} \tag{7.13}$$

To sample the conditional posterior for the hidden variables, the marginal $P(\pi \mid j, \boldsymbol{n}, \boldsymbol{\gamma})$ has to be evaluated by integrate over all possible sampling events yielding π after iterating the child restaurant process j. We distinguish two classes of events, recursion and return events. A recursion event happens when iterating the child process results in further iterating its adapter restaurant process, in turn resulting into recursing to iterate a child restaurant process $i \neq j$. We denote the probability of each of these events by $P^{\mathrm{rec}}_{j,\boldsymbol{n},\boldsymbol{\gamma}}(i)$. Return events are the complementary events that do not lead to further recursion. We denote their probability by $P^{\mathrm{emi}}_{j,\boldsymbol{n},\boldsymbol{\gamma}}(\pi)$. For further illustration of the event classes see also the FractalFranchise procedure. Having defined these quantities the marginal adopts the following form.

$$\begin{aligned} P(\pi \mid j, \boldsymbol{n}, \boldsymbol{\gamma}) &= \sum_{\boldsymbol{n}'} P(\pi, \boldsymbol{n}' \mid j, \boldsymbol{n}, \boldsymbol{\gamma}) \\ &= P^{\mathrm{emi}}_{j,\boldsymbol{n},\boldsymbol{\gamma}}(\pi) + \sum_{i \neq j} P^{\mathrm{rec}}_{j,\boldsymbol{n},\boldsymbol{\gamma}}(i) P(\pi \mid i, \boldsymbol{n}^{i+}, \boldsymbol{\gamma}) \end{aligned} \tag{7.14}$$

where $\boldsymbol{n}^{i+}$ corresponds to the updated restaurant counts induced by iterating the adapter process j. Note that the summation over recursion targets also comprises a child process i^{new} that represents so far undiscovered children, though might be sampled by iterating the root restaurant process. Given $j, \boldsymbol{n}, \boldsymbol{\gamma}$, various configurations can be enumerated to achieve a return event. The same is true for the recursion events. These configurations and their probabilities are evident from the Chinese restaurant franchise construction and are not reported explicitly here. It turns out to be useful to investigate the stationary approximation $\hat{P}(\pi \mid j, \boldsymbol{n}, \boldsymbol{\gamma})$ that assumes the restaurant counts constant along the marginalization.

$$\hat{P}(\pi \mid j, \boldsymbol{n}, \boldsymbol{\gamma}) = P^{\text{emi}}_{j,\boldsymbol{n},\boldsymbol{\gamma}}(\pi) + \sum_{i \neq j} P^{\text{rec}}_{j,\boldsymbol{n},\boldsymbol{\gamma}}(i) P(\pi \mid i, \boldsymbol{n}, \boldsymbol{\gamma}) \tag{7.15}$$

The marginal approximation $\hat{P}(\pi \mid j, \boldsymbol{n}, \boldsymbol{\gamma})$ can be efficiently computed. The recursion (7.14) gives rise to a linear equation system that can be solved for the marginal. Let M be the number of indexed child distributions (including i^{new}). Introducing $\boldsymbol{P} := (P(\pi \mid i, \boldsymbol{n}, \boldsymbol{\gamma}))_{1 \leq i \leq M}$, $\boldsymbol{P}^{\text{emi}} := \left(P^{\text{emi}}_{j,\boldsymbol{n},\boldsymbol{\gamma}}(\pi)\right)_{1 \leq i \leq M}$ and $\boldsymbol{P}^{\text{rec}} := (p_{ji})_{1 \leq j,i \leq M}$, where $p_{ji} := P^{\text{rec}}_{j,\boldsymbol{n},\boldsymbol{\gamma}}(i)$, we can rewrite (7.14).

$$\boldsymbol{P}^{\text{emi}} = (\mathbb{1} - \boldsymbol{P}^{\text{rec}})\, \boldsymbol{P} \tag{7.16}$$

We can easily solve for $\boldsymbol{P}$ and obtain the marginals for all indexed children, including j.

The structure of the conditional posterior $\boldsymbol{n}' \mid \pi, j, \boldsymbol{n}, \boldsymbol{\gamma}$ allows to straightforwardly sample a sequence of recursion and return events yielding $\pi_{t'}$. Given $\pi, j, \boldsymbol{n}, \boldsymbol{\gamma}$, we can specify $\boldsymbol{P}^{\text{emi}}$ and $\boldsymbol{P}^{\text{rec}}$ and compute the marginals $\boldsymbol{P}$. Starting from the child restaurant process j we can specify the posterior probability whether to directly emit π or to recurse into another child restaurant process.

7.7 Experiments

We qualitatively studied the properties of fractal Dirichlet process samples and particularly compared them to hierarchical Dirichlet process samples (**Fig. 7.2**). We sampled four trajectories $\boldsymbol{s}$ ($|\boldsymbol{s}| = 2000$) in the infinite fractal Markov chain model where we identify the adapter base measure F with G^r (see also section 7.5). The process parameters $\gamma^c = 10, \alpha^a = 1$ were the same for all simulations. We varied the concentration parameter γ^a of the adapter process to investigate the transition from the well known hierarchical Dirichlet process ($\gamma^a = 0$) to the (truly) fractal Dirichlet process ($\gamma^a > 0$). We determined the Pearson correlation among the empirical child distributions as pairwise similarity measure. We observed that pairwise similarity is more dispersed for the hierarchical Dirichlet process than for its fractal extension. The larger the ratio γ^a/α^a, the more the fractal Dirichlet process gives rise to clusters of strongly related child distributions.

We report the posterior inference with the Gibbs sampler (**Fig. 7.3**). We sampled two trajectories each comprising 200 data points (infinite fractal Markov chain mode). Parameters were set to $\gamma^r = 100, \gamma^c = 1, \alpha^a = 1$. For the first trajectory $\gamma^a = 0$ and for the second $\gamma^a = 3$. We assumed vague Gamma priors $\mathcal{G}(1, 10)$ for all parameters. We estimated the posterior distributions by considering every 100^{th} realization from in total 10^6 Gibbs sampler iterations. We generally observed that the posterior parameter distributions deviated considerably from the initially supplied prior distributions and accumulated their probability mass predominantly closely around the true value. The first case assessed to what extent the special case of the hierarchical Dirichlet process could be recovered while allowing for freedom in the adapter concentration parameter. We observed that this is actually the case since the posterior distribution puts most of its weight close to zero. Regarding the detailed Gibbs sampler trajectories, we assume that convergence is not an issue.

7.8 Discussion

We present the fractal Dirichlet process, a generalization of the hierarchical Dirichlet process. The fractal Dirichlet process incorporates self-referential base measures, thereby providing a mechanism to explicitly capture pairwise similarity in a set of discrete measures [20].

This chapter presents the fractal process framework on the basis of Dirichlet processes. It is though straightforward to formulate this concept on the basis of the more general Pitman-Yor processes [107].

We propose an efficient Gibbs sampler for Bayesian parameter inference. The Gibbs sampler involves a marginalization step with respect to all possible sampling events to yield some specified outcome. We show how to reduce the marginalization to a tractable solution of a linear equation system. We note that this solution assumes that the transition probabilities are stationary in the course of further recursion. This is not exactly the case. However, diagnostic experiments have shown that our stationary approximation effectively holds in all encountered situations. For all experiments carried out, the average recursion depth for each fractal franchise iteration was only around three. For some cases we explicitly computed marginal probabilities by accounting for expected count changes up to a recursion depth of ten and compared to the stationary approximation. We did not observe significant deviations for both estimates of the marginal probabilities. This result is not surprising since the contribution of count changes decrease exponentially with recursion depth. We conclude that the Gibbs sampler efficiently achieves accurate posterior parameter inference.

It will be interesting to investigate how fractal Dirichlet processes behave in statistical learning scenarios, such as e.g. language or topic modeling. Those application scenarios where prototypes have to be inferred from indirect observations will necessitate to extend the inference scheme to this end. Constituting a rich and practical prior for ensembles of intricately related distributions, we expect the fractal Dirichlet process to enhance various machine learning applications. The following chapter of this thesis will investigate the application of the fractal Dirichlet process to characterize peptide distributions arising in heterogeneous shotgun proteomics studies and to thereby allow for proteome coverage prediction in this context.

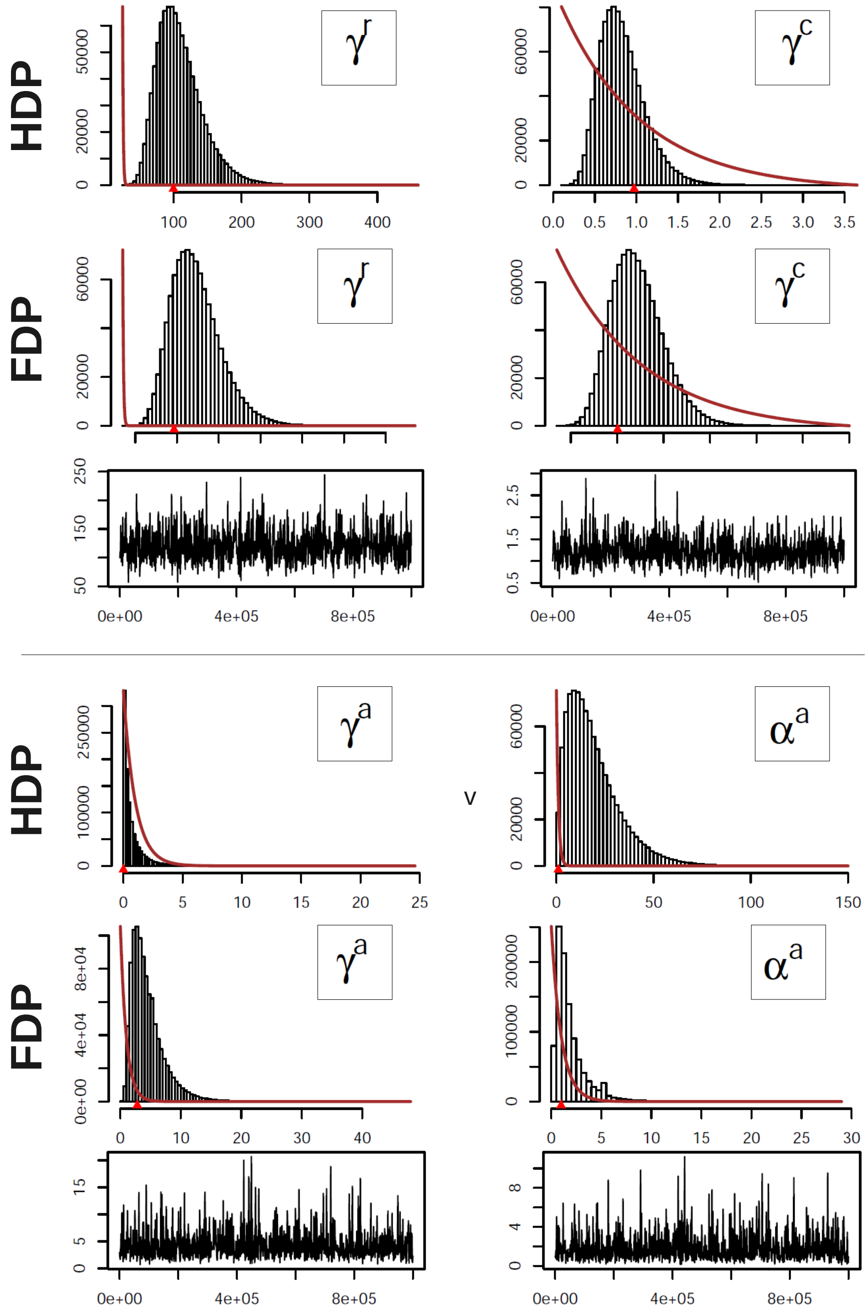


Figure 7.3: Posterior inference of the parameters of the fractal Dirichlet process. We consider a setting corresponding to the special case of the hierarchical Dirichlet process where $\gamma^a_{\mathrm{true}} = 0$ (HDP, first row) or where $\gamma^a_{\mathrm{true}} \neq 0$ (FDP, second row). The Gamma prior (brown) and the estimated posterior distribution (histogram) of the parameters for simulated data ($|\boldsymbol{s}| = 200$). The red triangles indicate the true parameter values. Posterior expectations deviate considerably from prior beliefs and well match the true parameter value. The bottom row exemplarily illustrates the sampling trajectories (10^6 iterations) for the individual parameters for the $\gamma^a_{\mathrm{true}} \neq 0$ setting. The trajectories demonstrate good convergence behavior of the Gibbs sampler.

8 Proteome Coverage Prediction for Integrated Proteomics Datasets

8.1 Summary

In order to maximize proteome coverage for a complex protein mixture, i.e. to identify as many proteins as possible, various different fractionation experiments are typically performed and the individual fractions are subjected to mass spectrometric analysis. The resulting data are integrated into large and heterogeneous datasets. Proteome coverage prediction refers to the task of extrapolating the number of protein discoveries by future measurements conditioned on a sequence of already performed measurements. Proteome coverage prediction at an early stage enables experimentalists to design and plan maximally informative proteomics studies. To date, there does not exist any method that reliably predicts proteome coverage from integrated datasets. We present a generalized hierarchical Pitman-Yor process model that explicitly captures the redundancy within integrated datasets by means of self-referential base measures. Proteome coverage prediction accuracy of our approach is assessed by applying it to an integrated proteomics dataset for the bacterium *L. interrogans* and by demonstrating that it outperforms ad hoc extrapolation methods and prediction methods designed for non-integrated datasets. Furthermore, we estimate the maximally achievable proteome coverage for the experimental setup underlying the *L. interrogans* dataset. We discuss the implications of our results to determine rational stop criteria and their influence on the design of efficient and reliable proteomics studies.

8.2 Introduction

Recent developments in mass spectrometry based proteomics have enabled biologists to comprehensively characterize proteomes, the protein inventories of biological samples [38]. To achieve extensive proteome coverage, a range of different experiments have to be carefully planned and extensively repeated. Proteome coverage prediction denotes the task of estimating the expected yield of protein discoveries upon experiment repetitions. This task is essential to guide experimental planning and to infer maximal coverage for a particular series of experiments. Here we present a generalized hierarchical Pitman-Yor process to reliably predict proteome coverage for multidimensional fractionation experiments.

The most successful strategy to achieve extensive proteome coverage is referred to as shotgun proteomics. Briefly, proteins are biochemically extracted from a biological sample and are enzymatically digested to yield a complex ensemble of peptides. Protein and/or peptide ensembles are optionally further fractionated according to physical/chemical/biological properties (multidimensional fractionation). Tandem mass spectrometry is then used to sample and identify individual peptide species present in the resulting ensembles and to finally recover the set of proteins initially present in the biological sample [97] (**Fig. 8.1**).

The capacity of mass spectrometers limits the number of peptides possibly identified at a time. Due to this constraint it is by far too difficult to identify the entirety of species in a peptide ensemble arising after enzymatic digestion of a typical complex biological sample such as a complete proteome. Two experimental routes are pursued to circumvent this limitation and to enable comprehensive characterization of a complex peptide ensemble. First, peptide ensembles are fractionated into a multitude of less complex and, therefore, more tractable ensembles before being analyzed by tandem mass spectrometry and second, experiments are extensively repeated. Popular fractionation schemes separate peptides with respect to properties such as e.g. size or isoelectric point. Reversed phase liquid chromatography (LC) is the most common fractionation technique and separates peptide ensembles according to hydrophobicity and is typically directly coupled to a tandem mass spectrometry system (LC-MS/MS). Multidimensional fractionation strategies comprise multiple steps of fractionation, typically fractionation according to some physico-chemical property other than hydrophobicity followed by LC-MS/MS analysis. (**Fig. 8.1**). Shotgun proteomics studies that achieved significant proteome coverage for a variety of organisms have shown to build on extensive repetition of multidimensional fractionation experiments (see e.g. [15]).

Methods for proteome coverage prediction estimate the expected number of peptide/protein discoveries when experiments are repeated. Proteome coverage prediction is essential for rational experimental planning of shotgun proteomics studies. Projects aiming at extensive proteome coverage require a considerable amount of experimentation. Proteome coverage should ideally increase efficiently with consecutive experiments. The choice between competing experimental setups should thus be guided by their potential to increase proteome coverage. Methods for proteome coverage prediction enable to rationally determine the optimal setup. Proteome coverage prediction furthermore enables to estimate the maximal coverage as well the volume of experiments required to achieve this coverage.

Proteome coverage prediction and related tasks have not been addressed until recently. Fenyo *et al.* conducted simulation studies to generally investigate how fractionation of peptide or protein ensembles might affect the efficiency of shotgun proteomics experiments [43]. Brunner *et al.* roughly estimated upper and lower bounds for proteome coverage from a real data set by assuming worst/best case scenarios [15]. Recently, an infinite Markov model based on Dirichlet processes [8] has been proposed to characterize LC-MS/MS experiments and for the first time to predict proteome coverage for one dimensional fractionation experiments [23].

In practice, it is highly desirable to predict proteome coverage of multidimensional fractionation experiments since these strategies have shown to have the largest potential to map out a proteome. However, there does not exist any method for proteome coverage prediction of these experiments. This task is particularly challenging since the proteomes represented by each fraction overlap to an unknown extent. Proteome coverage prediction methods for multidimensional fractionation experiments have to account for this phenomenon.

In this chapter we generalize the non-parametric approach to characterize peptide distributions arising in LC-MS/MS experiments [23] to further enable proteome coverage prediction from integrated datasets compiled from multidimensional fractionation experiments. Specifically, we propose a novel generalized hierarchical Pitman-Yor process [134, 133] with self-referential base measures that addresses the issue of distribution overlap which is introduced by the fractionation preceding the LC-MS/MS analysis. Besides the possibility to characterize peptide distributions arising in the course of multidimensional fractionation experiments, this approach also lends itself to characterize the biologically more relevant protein distributions. We assess our method on a set of 24 experiments from multidimensional fractionation of a *L. interrogans* whole proteome sample and report better performance than ad hoc extrapolation schemes and other approaches designed for one dimensional fractionation experiments. We discuss our re-

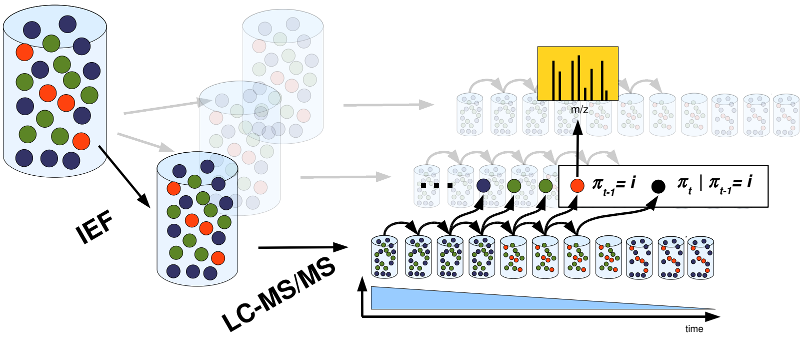


Figure 8.1: Illustration of a typical multidimensional fractionation experiment. The initial root peptide ensemble obtained from the biological source is separated by some fractionation method (e.g. isoelectric focussing (IEF)), giving rise to a set of related peptide ensembles. LC-MS/MS analysis is performed for each of these fractions. Liquid chromatography fractionation generates a sequence of child peptide ensembles from the root ensemble. Each of these ensembles is derived from the root ensemble by pooling peptides of similar polarity. The sequence of ensembles features descending overall polarity in the course of the experiment. During the experiment peptides π_t are drawn from the sequence of ensembles and analyzed by the mass spectrometer coupled to the liquid chromatography system and subsequently identified computationally. We propose a non-parametric Bayesian approach to characterize the distributions governing the peptide ensembles. We simulate further experiments and thereby predict proteome coverage by sampling from these peptide distributions.

sults with respect to maximally achievable proteome coverage from a peptide- as well as protein-centric perspective.

Methods

The following sections give technical background and details on the hierarchical Pitman-Yor process framework for proteome coverage prediction based on integrated datasets. Briefly, our approach characterizes the peptide/protein distributions arising in a multidimensional fractionation experiment and simulates further experiments by sampling from these distributions. Proteome coverage is predicted by counting the number of novel peptide/protein discoveries in the simulations. In the following sections we will assume a peptide-centric view for clarity, i.e. consider peptide distributions instead of its protein counterparts. Note that peptides, by virtue of being protein fragments, also refer to protein identities. Therefore, the following sections can also be read by consequently substituting peptides with proteins. Complications arising from peptides ambiguously referring to several protein identities are discussed in section 8.9.

8.3 Pitman-Yor processes

We apply Pitman-Yor processes to characterize peptide distributions arising in the course of a series of proteomics experiments. In the following we briefly review the concept of Pitman-Yor processes in the context of this work.

Like the Gaussian distribution is an appropriate distribution for a real valued random variable in numerous applications, the Pitman-Yor process frequently is an appropriate distribution for complex objects such as discrete distributions [66]. Loosely spoken, Pitman-Yor processes are suited as priors over discrete distributions that are expected to have most of their probability mass on a small number of atoms and only little probability mass on the vast majority of atoms [133]. As various proteomics studies have shown that protein/peptide frequencies exhibit such a property (see e.g. [26]), we use Pitman-Yor processes as priors for distributions G over a set Π of peptides defined by a protein database of the studied organism.

$$G \mid \gamma, d, H \quad \sim \quad \mathrm{PY}(\gamma, d, H) \tag{8.1}$$

where $\mathrm{PY}(\gamma, d, H)$ is a Pitman-Yor process with a concentration parameter γ, a discount parameter d and a base probability measure H. The base measure is defined over Π (sample space). H is frequently chosen as the uniform measure, assigning $1/|\Pi|$ probability mass to each $\pi \in \Pi$.

The so called *Chinese Restaurant* construction [12, 107] provides an intuitive way to see which kind of distributions are likely to be drawn from a Pitman-Yor process $\mathrm{PY}(\gamma, d, H)$. Imagine a restaurant with an infinite number of tables. At each table a specific dish is served. We construct a distribution G over dishes after having seated an infinite number of customers. Customers are seated according to a probabilistic rule. Specifically, the probability of the t-th customer being seated at the table serving dish $\pi_t = k$ assumes the values

$$P(\pi_t = k \mid \pi_1, ..., \pi_{t-1}, \gamma, d, H) \quad = \quad \begin{cases} \frac{n_k - d}{t-1+\gamma} & \text{populated table} \\ \frac{\gamma + kd}{t-1+\gamma} & \text{next unpopulated table} \end{cases} \tag{8.2}$$

where n_k corresponds to the number of customers already sitting at the table serving dish i. In case a customer happens to be seated at a new table, the dish served at this table is drawn from the base probability measure H. A procedural description of serving a new customer in a restaurant with seating arrangement $R = n_1, n_2...$ is as follows:

```
Seat($R, \gamma, d, H$)
t ← SampleTable($R, \gamma, d$)
if $t \neq new$
   then return Dish($R, t$)
   else return Sample($H$)
```

The larger the concentration parameter γ, the higher the chances that a new customer is seated at a new table. The more customers have already been seated, the less likely a new dish will be served. The larger the discount parameter d the less likely a customer is seated at an already populated table. Note that $d < 1$. In summary, the parameters γ and d control, though in different ways, the deviation of G from the base measure H. The *Chinese Restaurant* construction specifies the posterior to iteratively sample from $\pi_t \mid \pi_1, ..., \pi_{t-1}, \gamma, d, H$ after marginalizing out G.

Pitman-Yor Processes are generalizations of the more commonly known Dirichlet processes [12, 4]. More precisely, a Dirichlet Process DP(γ, H) is equivalent to a Pitman-Yor process PY(γ, d, H) with $d = 0$. Both Dirichlet and Pitman-Yor processes will be used as priors for peptide distributions that arise in the course of a multidimensional fractionation experiment. After having estimated the process parameters we will simulate further experiments by sampling according to the *Chinese Restaurant* construction.

8.4 Hierarchical process model for fractionation experiments

In the following we characterize the distributions which arise in a multidimensional fractionation experiment. We specifically describe a typical setup that comprises two consecutive fractionation steps, where the first step splits the initial peptide ensemble into a set of I fractions that are each analyzed by LC-MS/MS (**Fig. 8.1**). Besides enforcing consistency along subsequent fractionation steps using hierarchical processes, we further want our model to explicitly capture the similarity of corresponding peptide distributions across different fractions.

The initial peptide ensemble follows the root distribution G. We assume a Pitman-Yor process prior $\mathrm{PY}(\gamma_r, d_r, H)$ for G. The base measure H is chosen to be the uniform distribution over the peptides defined by the protein database of the studied organism.

Peptides are not directly sampled from the root distribution G. Consider some time point t during the LC-MS/MS analysis of fraction i. The peptide π_t^i is sampled from the child peptide distribution G_t^i of the peptide ensemble currently eluting from the liquid chromatography column. Following [23] we assume that the preceding peptide $\pi_{t-1}^i := j$ is indicative for the current polarity of the chromatography and thereby the current peptide distribution, i.e. with a slight abuse of notation we assume $G_t^i = G_j^i$. Further we assume a Dirichlet process prior for G_j^i, resulting in an infinite Markov model for LC-MS/MS experiments similar to [23].

$$\begin{aligned} G_j^i \mid \gamma_c^i, A_j^i &\quad\sim\quad \mathrm{DP}(\gamma_c^i, A_j^i) \\ \pi_t \mid \pi_{t-1}^i = j &\quad\sim\quad G_j^i \end{aligned} \tag{8.3}$$

We want the child distributions G_j^i to be consistent with the root distribution G, i.e. we want to ensure that peptides having zero probability mass in the initial peptide ensemble still have zero probability mass during an LC-MS/MS experiment. This notion is captured by choosing G as base measure A_j^i in (8.3), yielding a hierarchical process [134]. This choice ensures (1) that G_j^i is consistent with G, i.e. the support of G_j^i is enclosed by the support of G and (2) that G_j^i will have similarity to G to an extent defined by the concentration parameter γ_c^i. Furthermore, we want to capture the similarity between G_j^i and its corresponding distributions $G_j^{i'}$ in all other fractions $i' \neq i$. Therefore we extend the base measure A_j^i in (8.3) to a (self-referential) linear combination of the distributions $(G_j^{i'})_{i'=1}^I$ and G.

$$A_j^i \quad = \quad a_i^i G + \sum_{l \neq i} a_l^i G_j^l \tag{8.4}$$

Since the values $a^i := (a_l^i)_{l=1}^I$ are not known beforehand, it is natural to treat them as a random discrete distribution with a Dirichlet process prior. The a_i^i reflect the dissimilarity of fraction i from the other fractions by controlling the rate of sampling peptides directly from the root distribution G. We account for their distinguished role by putting prior weight α_a^i on a_i^i and incorporating this parameter by assuming for the a^i a biased (in the sense of [23]) Dirichlet process prior $\mathrm{DP}_i(\gamma_a^i, \alpha_a^i, M)$ with uniform base measure $M := (1/I)_{1..I}$. In the following, we will refer to the a^i as the adapter distributions.

The self-referential base measures A^i_j are a crucial component of this process since they capture the important overlap of peptide distributions across the fractions j arising in a multidimensional fractionation experiment. The step from the simple base measure G as described in [23] to the self-referential base measure enables to appropriately characterize the peptide distributions describing such an experiment.

Putting together the precedent considerations we fully characterize the stochastic source of a multidimensional fractionation experiment by

$$\begin{aligned}
G \mid \gamma_r, d_r, H &\sim \mathrm{PY}(\gamma_r, d_r, H) \\
a^i \mid \gamma^i_a, \alpha^i_a, M &\sim \mathrm{DP}_i(\gamma^i_a, \alpha^i_a, M) \\
G^i_j \mid \gamma^i_c, A^i_j &\sim \mathrm{DP}(\gamma^i_c, A^i_j) \\
\pi_t \mid \pi^i_{t-1} = j &\sim G^i_j
\end{aligned} \tag{8.5}$$

Note that it is straightforward to assume Pitman-Yor process priors for all distributions. This choice though comes at the cost of additional parameters that have to be learned from data. In this work we wanted to focus on robustness and therefore we decided to keep the priors of the child distributions as simple as possible.

8.5 Sampling sequences of protein Identifications

This section describes a nested, recursive *Chinese Restaurant* construction to sample peptides from the hierarchical process model with self-referential base measures given an already observed series $\boldsymbol{\pi}$ of already observed peptides, i.e. how to simulate further experiments.

For each distribution in the hierarchical process model we have a restaurant representation, i.e. a seating arrangement. Specifically, we denote the restaurants corresponding to the G^i_j as $R^c_{ij} = (n^c_{ijk})^K_{k=1}$, those to the a^j as $R^a_i = (n^a_{ii'})^I_{i'=1}$ and the root restaurant as $R^r = (n_k)^K_{k=1}$. To keep the notation uncluttered we incorporate the prior weights α^i_a into the counts n^a_{ii} and respectively R^a_i. $\boldsymbol{R}$ denotes the set of all restaurants. Note that a set of seating arrangements $\boldsymbol{R}$ implies a series $\boldsymbol{\pi}$ of observed identifications. We further summarize the set of parameters by $\boldsymbol{\theta} := (\gamma_r, d_r, \gamma^1_a, ..., \gamma^I_a, \gamma^1_c, ..., \gamma^I_c)$.

For a given set of seating arrangements $\boldsymbol{R}$ we now want to sample the identification π_t for fraction i and preceding identification $\pi_{t-1} = j$. Verbally, we first have to iterate the

Chinese Restaurant construction for the corresponding child distribution. In case this iteration triggers a sampling event of its base measure, we have to determine which of its mixture components is to be sampled. Therefore we iterate the *Chinese Restaurant* construction of the corresponding adapter distribution. Subsequently, either the root restaurant or, recursively, some of the sibling child restaurants of another fraction is iterated. This procedure can summarized as shown below.

```
SampleIdentification($i, j, \boldsymbol{R}, \boldsymbol{\theta}, H, M$)
$\pi \leftarrow$ Seat($R^c_{ij}, \gamma^i_c, 0, 0$)   // sample child
if $\pi = 0$
   then $i' \leftarrow$ Seat($R^a_i, \gamma^i_a, 0, M$)   // sample adapter
        if $i' \neq i$
           then $\pi \leftarrow$ SampleIdentification($i', j, \boldsymbol{R}, \boldsymbol{\theta}, H, M$)
           else $\pi \leftarrow$ Seat($R^r, \gamma_r, d_r, H$)   // sample root
return $\pi$
```

The nested, recursive *Chinese Restaurant* construction serves to simulate further experiments, i.e. to sample more peptides given an already observed series $\boldsymbol{\pi}$ of peptides and will be useful in the following section to derive a likelihood function for paramater estimation.

8.6 Empirical Bayes parameter estimate

Parameters of the hierarchical process model from section 8.4 can be estimated from a series π of identifications by empirical Bayes inference, i.e. by choosing the parameters to maximize a likelihood function $\mathcal{L}_{\widehat{\boldsymbol{R}}}$.

$$\widehat{\boldsymbol{\theta}} \quad := \quad \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{\widehat{\boldsymbol{R}}}(\boldsymbol{\theta}) \tag{8.6}$$

In the following we will specify $\mathcal{L}_{\widehat{\boldsymbol{R}}}$. Sampling a series $\boldsymbol{\pi}$ of identifications reduces to iterate various *Chinese Restaurant* constructions according to the probabilities in (8.2). We can define a likelihood function $\mathcal{L}_{\boldsymbol{R}}(\boldsymbol{\theta})$ for a set of seating arrangements $\boldsymbol{R}$, or the corresponding series $\boldsymbol{\pi}$ of identifications.

$$\mathcal{L}_{\boldsymbol{R}}(\boldsymbol{\theta}) \quad = \quad \mathcal{L}_{\mathrm{cr}}(R^r, \gamma_r, d_r) \cdot \prod_{i=1}^{I} \mathcal{L}_{\mathrm{cr}}(R^a_i, \gamma^i_a) \cdot \prod_{j=1}^{J} \mathcal{L}_{\mathrm{cr}}(R^c_{ij}, \gamma^i_c) \tag{8.7}$$

where $\mathcal{L}_{\text{cr}}(R,\gamma,d)/\mathcal{L}_{\text{cr}}(R,\gamma)$ corresponds to the likelihood of achieving a seating arrangement R in a single restaurant representation of a Pitman-Yor/Dirichlet process sample with parameters $\gamma, d/\gamma$. Note that prior weights α_a^i of the adapter processes are appropriately incorporated into R_i^a and they are therefore not explicitly listed.

$$\mathcal{L}_{\text{cr}}(R,\gamma,d) \quad = \quad \frac{\prod_{k=1}^{K}(\gamma+kd)\cdot\prod_{n=1}^{n_k}(n-d)}{\prod_{n=1}^{N}(n+\gamma)} \tag{8.8}$$

with $N=\sum_{k=1}^{K} n_k$ and K corresponding to the number of populated tables.

We do observe the series $\boldsymbol{\pi}$ of identifications. Though we only have incomplete knowledge about $\boldsymbol{R}$. We observe the seating arrangements R_{ij}^c of the child processes.

$$n_{ijk}^c \quad = \quad \left|\pi_t^i \ : \ (\pi_{t-1}^i = j)\wedge(\pi_t^i = k)\right| \tag{8.9}$$

where the $\pi_t^i \in \boldsymbol{\pi}^i$ denote identifications observed exclusively in fraction i. We do not directly observe R^r and the R_i^a. We present a sparse estimate for $\boldsymbol{R}$ that is consistent with $\boldsymbol{\pi}$ and complies with a minimal number of seating events in the root restaurant representation R^r of the root distribution G. Consider the representation matrix M with entries m_{ik} equaling one if a peptide k has been observed in fraction i or zero otherwise. We want each peptide discovery k to be represented by some fraction f_k. We further want to choose the number of representing fractions to be as small as possible. This problem is more commonly known as the NP-hard set cover problem [70]. We compute the f_k with the greedy heuristic, choosing at each step the fraction which covers the largest number of remaining different peptides. Every time the peptide k is discovered, i.e. sampled for the first time in a child process, we choose the corresponding adapter process to trigger a sampling event in f_k. Accordingly, we estimate the hidden seating arrangements of the adapter and root restaurant representations.

$$\begin{aligned} n_{ii'}^a \quad &= \quad \left|i,j,k \ : \ (f_k = i')\wedge(\exists\, t : (\pi_{t-1}^i = j)\wedge(\pi_t^i = k))\right| \\ n_k^r \quad &= \quad \left|i,j,k \ : \ (f_k = i\ )\wedge(\exists\, t : (\pi_{t-1}^i = j)\wedge(\pi_t^i = k))\right| \end{aligned} \tag{8.10}$$

We finally determine the parameters $\widehat{\boldsymbol{\theta}}$ by optimizing $\mathcal{L}_{\widehat{\boldsymbol{R}}}$ with a quasi-Newton method [109]. In summary, we obtain an empirical Bayes parameter estimate from an observed series $\boldsymbol{\pi}$ of identifications.

8.7 Proteome coverage prediction with false identifications

At this point we can specify how to predict the number of new peptide discoveries for future experiments from a series $\boldsymbol{\pi}$ of already observed identifications. In a first step, we estimate the parameters and hidden variables of the hierarchical process model (8.4) as described in the preceding section 8.6. Second, we sample m peptide series $(\boldsymbol{\pi}_{new,i})_{i=1}^{m}$ by means of the nested *Chinese Restaurant* construction (8.5). For each $\boldsymbol{\pi}_{new,i}$ we count the number of new discoveries. The expected proteome coverage we estimate as the mean of discovery counts across all $\boldsymbol{\pi}_{new,i}$.

In practice, the series $\boldsymbol{\pi}$ of observed peptides corresponds to a series of peptide-spectrum matches that have been inferred computationally. Obviously peptide-spectrum matches are not perfect. Fortunately, the fraction of false positive peptide-spectrum matches is typically known [72, 41]. Furthermore it has been observed that false positive peptide-spectrum matches distribute in a uniform-like manner across the protein database [23, 26]. To account for false positive peptide-spectrum matches we adaptively estimate parameters and we adaptively sample novel peptide identifications as described in [23].

8.8 Results

We present results that demonstrate the proteome coverage prediction performance of our hierarchical process model. To this end we studied a large multidimensional fractionation experiment of a *L. interrogans* sample. We compared to a recent approach designed for (one dimensional) LC-MS/MS experiments [23] and to ad hoc extrapolation methods. We further extrapolated proteome coverage for the *L. interrogans* sample to make statements about maximal coverage.

We studied an integrated dataset acquired from multidimensional fractionation experiments for the bacterium *L. interrogans*. After protein extraction and tryptic digestion, the resulting peptide mixture was fractionated according to the isoelectric point of the peptides by off gel electrophoresis and each of the 24 fractions analyzed by LC-MS/MS coupled to a FT-LTQ high mass accuracy instrument. Target-decoy database search with Sequest/PeptideProphet [72] resulted in 59918 peptide-spectrum matches at a false discovery rate of 1% [116].

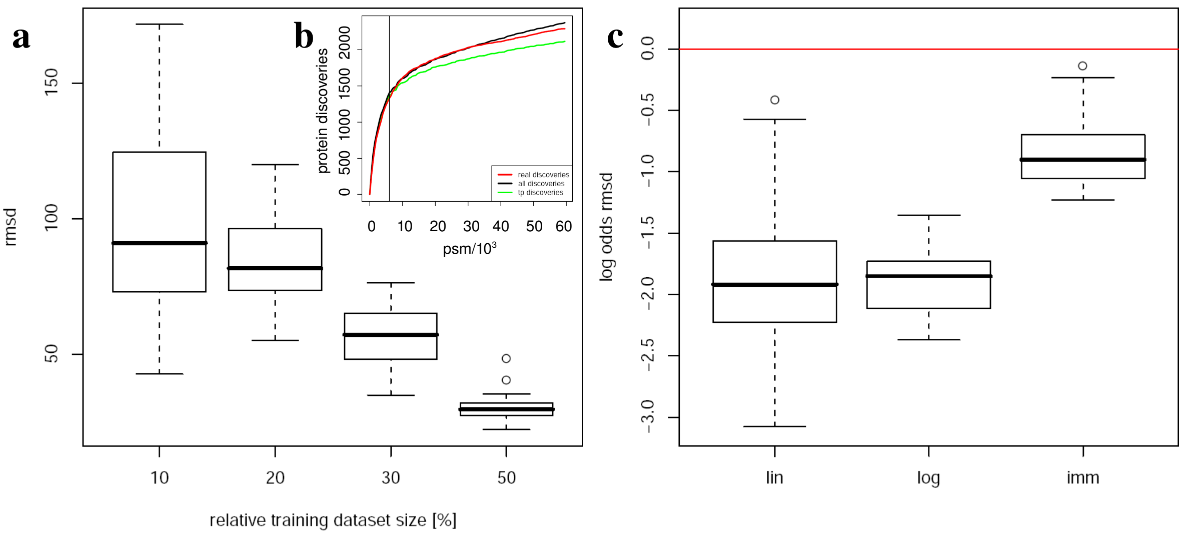


Figure 8.2: Proteome coverage prediction performance by cross validation. Training datasets generated by subsampling the complete set of peptide-spectrum matches. Test of prediction performance on complete dataset. (**a**) Hierarchical process model accuracy in terms of root mean square deviation (rmsd) from the true progression of proteome coverage. Columns correspond to relative training dataset size compared to the complete *L. interrogans*. (**b**) Example trajectory for prediction from dataset instance with 10% relative size. Plot shows trajectory of observed (real), predicted true positive (tp) and including false positive protein discoveries (all). (**c**) Performance comparison of hierarchical process model with infinite Markov model (imm), extrapolation of logarithmic regression (log) and linear extrapolation of last experiment (lin). Box plot of log odds of rmsd ($\log(\mathrm{rmsd}_{\mathrm{ref}}/\mathrm{rmsd}_{\mathrm{comp}})$) for reference and compared method (lin, log, imm). Median log odds for comparison with the other methods are significantly lower than zero, indicating weaker performance than our approach. The hierarchical process model is capable to reliably predict proteome coverage from a small amount of identifications and clearly outperforms other applicable methods.

8.8.1 Cross validation prediction accuracy

We assessed proteome coverage prediction performance in a cross validation scenario. Briefly, we generated various training datasets of decreasing size by subsampling the complete set of peptide-spectrum matches. We performed proteome coverage prediction for each training dataset and assessed accuracy by comparing to the real proteome coverage progression of the complete dataset. Precisely, we generated 20 training datasets by 20 times sampling 10% of all peptide-spectrum matches in the dataset while preserving their fraction association. We repeated this procedure by also sampling 20, 30 or 50% of all peptide-spectrum matches, finally obtaining 80 training datasets of varying size.

We assessed the prediction accuracy of the hierarchical process model (**Fig. 8.2a**). Prediction accuracy is measured as root mean square deviation of predicted and actually observed progression of proteome coverage. Proteome coverage corresponds to number of protein discoveries. Prediction accuracy is reasonable already for the smallest training dataset sizes, i.e. 10% of the complete *L. interrogans* dataset. **Fig. 8.2b** depicts an example prediction for the set of smallest training datasets. As expected, prediction accuracy improves further for training datasets of larger size. Similar results are obtained for prediction of proteome coverage in terms of peptide discoveries (data not shown). We conclude that our approach is able to reliably predict proteome coverage already from a small amount of data.

8.8.2 Proteome coverage prediction benchmark

We compared the hierarchical process model to other methods. We chose two simple general purpose extrapolation methods and a method designed for proteome coverage prediction of non-integrated datasets. We first considered an extrapolation scheme that linearly extrapolated proteome coverage progression of the last LC-MS/MS experiment of a training series. Second, we considered the extrapolation of a logarithmic regression ($y = a \log x + b$). We assessed prediction performance on the 80 training series as described above and observed that the hierarchical process model clearly outperforms the other methods (**Fig. 8.2c**). These results indicate that proteome coverage prediction for integrated datasets is a non-trivial task that is not solved satisfactory by ad hoc extrapolation methods and is different from the related task of proteome coverage prediction for non-integrated datasets.

8.8.3 Detection of saturation coverage for L. interrogans

We estimated saturation proteome coverage for *L. interrogans* given the experimental workflow described above. Therefore we performed proteome coverage prediction for in silico repetition of all experiments. Proteome coverage in terms of peptide discoveries appears to steadily increase (**Fig. 8.3a**). Proteome coverage in terms of protein discoveries also seems to increase (**Fig. 8.3b**). This observation is however only true for all protein discoveries including the false positive ones. Since our approach separately accounts for the contribution of false and true positive protein discoveries (see section 8.7), we could exclusively monitor the progression of true protein discoveries. We ob-

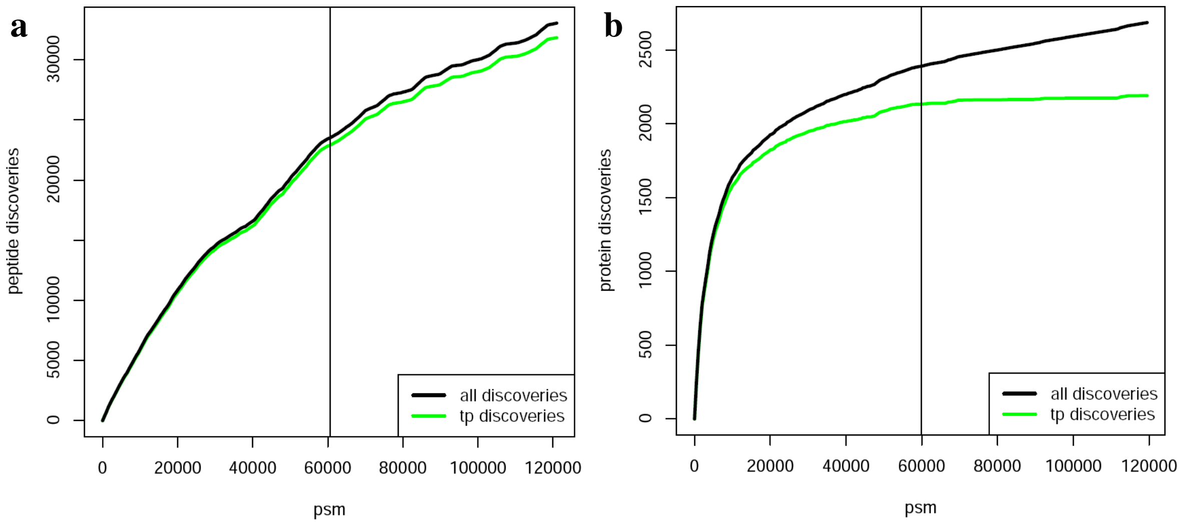


Figure 8.3: Proteome coverage prediction beyond the *L. interrogans* dataset. Vertical lines denote the extent of the dataset in terms of acquired peptide-spectrum matches (psm). Trajectories correspond to predicted true positive (tp) and including false positive discoveries (all) (**a**) Progression of peptide discoveries. (**b**) Progression of protein discoveries. Protein discovery rate stagnates compared to the steadily increasing number of peptide discoveries. The *L. interrogans* dataset achieves saturation coverage at the level of protein discoveries.

serve that the number of true positive protein discoveries does not change significantly. Considering the rate of new true positive discoveries, we effectively have reached saturation coverage for *L. interrogans.*

8.9 Discussion

For the first time, we propose a method to predict proteome coverage for multidimensional fractionation experiments. This achievement is an important enabling step for experimentalists since multidimensional fractionation experiments so far have the largest potential to comprehensively characterize a proteome. We present a novel hierarchical process to characterize distributions arising in the course of these experiments. This approach conceptionally extends methods exclusively suited for single fraction experiments [23], by introducing self-referential base measures that accommodate similarities among different experiment fractions. Our approach is generic since it operates on the level of peptide or protein distributions and, therefore, it conceptually accommodates any kind of heterogeneous set of fractions being analyzed by LC-MS/MS. Fractions do not necessarily have to originate from a single fractionation experiment. The considered fractions might also be derived from different tissues or cell cultures as long as their analysis is based on the same sequence database. Although we explicitly describe an approach that accounts for two fractionation steps, it is conceptually straightforward to extend it from a two level to a higher level hierarchy. However, the corresponding experimental setups are rarely encountered in practice. We show that our model reliably predicts proteome coverage of future experiments from a small amount of already performed experiments and clearly outperforms other methods.

Besides providing predictions at the level of peptide discoveries, we demonstrate that our approach yields reliable predictions of proteome coverage in terms of protein discoveries. Specifically, we require the set of considered fragment ion spectra to be unambiguously assigned to a protein identity to estimate future proteome coverage. This requirement is usually met, since possible ambiguities introduced by peptide-spectrum matches whose sequence maps to several protein identities are typically resolved by protein inference engines, e.g. by reporting a minimal consistent set of protein identifications [96]. It will though be interesting to extend our approach to allow for ambiguity in the protein identity assignments.

There has been considerable discussion in the past about when to consider a proteome to be mapped out. Our approach to proteome coverage prediction enables us to detect saturation coverage for any kind of shotgun proteomics dataset. In this study the *L. interrogans* dataset reaches saturation coverage at the level of protein discoveries. Out of 3740 proteins reported in the sequence database, roughly 2000 proteins can be faithfully observed — not less but also not a lot more. This analysis is a remarkable result considering the manageable amount of experimentation (24 LC-MS/MS runs). It should be noted that this result is valid for the given experimental setup, such as type of protein extraction, enzymatic digestion, fractionation method, type of mass spectrometer. Despite the sensitive state-of-the-art approach reported here, it remains conceivable that other experimental approaches turn out to be able to explore other parts of the *L. interrogans* proteome. Their potential could though be evaluated with the hierarchical process model presented here. Therefore the presented method is suited to assist method development since it objectively assesses the potential of a particular method to explore a proteome.

Characterizing more complex proteomes (e.g. human) necessitates a considerably larger amount of experimentation. In this context it will be promising to perform proteome coverage prediction for different experimental strategies at an early stage of the project to design future experiments such that maximal proteome coverage is achieved efficiently [120]. Our approach enables for the first time to accommodate any multidimensional fractionation strategy to perform this task. Efficient study design will help to save costly experiments, contribute to the reliability of the final set of protein discoveries [23, 26] and furthermore enhance subsequent directed/targeted proteomics studies [118, 82].

9 Optimal Design of Integrated Proteomics Experiments

9.1 Summary

Large shotgun proteomics studies typically aim at substantial proteome coverage. The complexity of any organism's proteome necessitates extensive repetition of multidimensional fractionation experiments. Such studies turn out to be obstructed by spending a considerable amount of resources on performing non-informative experiments that do not contribute novel protein discoveries. Shotgun proteomics studies would benefit from a rational design approach that prioritizes future experiments according to their expected impact on proteome coverage.

We present a non-parametric Bayesian approach to optimally design a shotgun proteomics study, i.e. to select a fixed length experimental sequence that maximizes the expected proteome coverage. Starting from a small amount of different experiments, we efficiently estimate expected proteome coverage for all possible experiment sequences of specified length. We build on our approach for proteome coverage prediction for single experiment sequences in the context of multidimensional fractionation studies. We formulate optimal experiment design as an optimization problem and show how to reduce it to the commonly known maximum k-coverage problem. Our approach to optimal design of shotgun proteomics studies enables researchers to accelerate the progression of proteome coverage by focusing resources on truly informative experiments.

9.2 Introduction

Careful design of large shotgun proteomics studies considerably accelerates the progression of proteome coverage. We present a rational approach to optimally design a shotgun proteomics study after having performed a small amount of experiments.

The complexity of a proteome constitutes the main challenge for experimental approaches aiming at characterization of a proteome. Protein databases count thousands of different protein sequences that each in average give rise to hundreds of different peptide fragments. Also accounting for post translational modifications, it is reasonable to assume that a biological sample gives rise to more than a million different peptide species in a shotgun proteomics experiment. Contemporary mass spectrometers are far from capable to characterize such a complex peptide mixture.

Fractionation strategies aim at reducing the complexity of peptide mixtures derived from whole proteome digests by splitting them according to some physical property into more tractable, less complex peptide mixtures. Reversed phase liquid chromatography constitutes the central fractionation approach in the context of mass spectrometer based proteomics (LC-MS/MS) [38]. It separates with respect to peptide polarity. Multidimensional fractionation strategies add additional steps of fractionation according to different physical properties other than polarity [55]. Peptide mixtures arising in the course of multidimensional fractionation strategies are still too complex to be characterized by a single experiment. Consequently, each fraction is repeatedly analyzed by LC-MS/MS. This setup inevitably entails redundant and therefore nonessential identification of peptides/proteins that have been seen in a previous experiment. Despite their redundancy, the multidimensional fractionation strategy still constitutes the most successful strategy to achieve extensive proteome coverage [138, 15, 5, 35, 119].

Several approaches have been presented to increase the efficiency of shotgun proteomics studies. Directed mass spectrometer approaches aim at reducing the sequencing redundancy by recording all peptide signals measurable in the mass spectrometer in a first step and to then systematically issue sequencing events for each signal [118, 117]. Other approaches aim at appropriately designing shotgun proteomics studies. Eriksson *et. al* theoretically studied whether fractionation at the level of peptides or proteins is expected to better promote progression of proteome coverage. Brunner *et. al* proposed an “analysis-driven experimentation“ strategy that consists of designing experiments that are supposed to close the gap between the set of expected proteins and those actually discovered by previous experiments.

We propose a novel complementary approach to optimally design a shotgun proteomics study. Starting from a small amount of LC-MS/MS experiments performed for a set of biological samples and/or fractions, we estimate a sequence of LC-MS/MS experiments (for this set of samples) that maximizes the expected proteome coverage. This approach aims at reducing redundancy by identifying and focusing the experimental efforts on the most information rich samples/fractions. Our approach builds on proteome coverage prediction for a single experiment sequence [23, 24]. Finding the optimal experiment sequence implicitly involves proteome coverage prediction of all possible sequences. We show that this task reduces to the maximum k-coverage problem, a variant of the well known set cover problem. We exemplify the impact of optimal design for a proteomics study of the human proteome.

Notation

We use boldfaced variables to represent vectorial variables. We use the following index notation. Normal font indices refer to single elements of a vector at position specified by the index (e.g. $\boldsymbol{e}_l$). Boldface (i.e. vectorial) indices refer to possibly several vector elements at positions specified by the vectorial index (e.g. $\boldsymbol{e}_{\boldsymbol{l}} := (\boldsymbol{e}_l)_{l \in \boldsymbol{l}}$).

We consider a scenario where we dispose of J different biological samples that are to be analyzed by LC-MS/MS experiments. In the following we refer to an LC-MS/MS experiment of a biological sample simply as an experiment.

We introduce the notion of an experiment sequence $\boldsymbol{e} := (e_l)_{1 \leq i \leq m}$. Each entry e_l denotes an LC-MS/MS experiment for the biological sample e_l. The realization of experiment e_l gives rise to a protein identification trajectory $\boldsymbol{\pi}_l$ which comprises the proteins $\boldsymbol{\Pi}_l := \text{unique}(\boldsymbol{\pi}_l)$. The complete experiment sequence gives rise to the trajectory $\boldsymbol{\pi} := (\boldsymbol{\pi}_l)_{1 \leq l \leq m}$ and protein discoveries $\boldsymbol{\Pi} := \bigcup_{1 \leq l \leq m} \boldsymbol{\Pi}_l$. We refer to proteome coverage as the cardinality of a protein discovery set.

9.3 Experiment Selection

This section describes how to select an optimal sequence of follow up experiments given a set of already performed experiments. *Experiment Selection* can be formulated as an optimization task.

We consider a situation where we have already performed l LC-MS/MS experiments for each of the J different samples. These experiments we denote by the experiment sequence $\boldsymbol{e}^0$. We denote the trajectory of protein identifications as $\boldsymbol{\pi}^0$. We discussed the task of proteome coverage prediction, i.e. the task of estimating the expected proteome coverage $E\left[c \mid \boldsymbol{e}, \boldsymbol{\pi}^0\right]$ in terms of protein discoveries for the sequence of experiments $\boldsymbol{e}$ after already having performed the experiments $\boldsymbol{e}^0$. For optimal experiment selection we now want to perform m additional LC-MS/MS experiments and select them such that they maximize the expected proteome coverage.

Experiment Selection Let $\boldsymbol{\pi}^0$ be a set of identification trajectories and $\boldsymbol{\Pi}^0$ be the respective set of protein discoveries obtained for J biological samples. Select the optimal sequence $\boldsymbol{e}^*$ comprising m additional experiments. $\boldsymbol{e}^*$ is considered optimal if its expected proteome coverage c is maximal.

$$\boldsymbol{e}^* = \arg\max_{|\boldsymbol{e}|=m} E\left[c \mid \boldsymbol{e}, \boldsymbol{\pi}^0\right] \tag{9.1}$$

9.4 Empirical Experiment Selection

In the following, we will describe a tractable estimate for the optimal sequence $\boldsymbol{e}^*$. Briefly, we first describe how to evaluate the optimization objective, i.e. how to estimate the expectation coverage for a given experiment sequence. In a second step we reduce the optimization task to the well known NP-hard maximum k-coverage problem for which good approximations are known.

Experiment Selection requires to evaluate the expectation value $E\left[c \mid \boldsymbol{e}, \boldsymbol{\pi}^0\right]$. We are not aware of a possibility to do this analytically. We can though estimate $E\left[c \mid \boldsymbol{e}, \boldsymbol{\pi}^0\right]$ for some sequence $\boldsymbol{e}$ of experiments as the empirical mean of coverage from a sufficiently large set of trajectories $\left(\boldsymbol{\pi}^k\right)_{1\leq k\leq r}$ sampled from $\boldsymbol{\pi} \mid \boldsymbol{e}, \boldsymbol{\pi}^0$ and the respective sets of protein identifications $\boldsymbol{\Pi}^k := \text{unique}(\boldsymbol{\pi}^k) \setminus \boldsymbol{\Pi}^0$. Each of these trajectories is obtained by running the proteome coverage prediction as described in section 8.

$$\hat{E}\left[c \mid \boldsymbol{\pi}^0\right] = \left|\boldsymbol{\Pi}^0\right| + \frac{1}{r}\sum_k \left|\boldsymbol{\Pi}^k\right| \tag{9.2}$$

The expected coverage for the experiment subsequence $\boldsymbol{e}' := \boldsymbol{e}_{\boldsymbol{l}}$ with $\boldsymbol{l} \subseteq (1,..,m)$ can be estimated from trajectories sampled from $\boldsymbol{\pi} \mid \boldsymbol{e}, \boldsymbol{\pi}^0$. Specifically, we assume that we already sampled the protein discovery sets $\left(\boldsymbol{\Pi}^k\right)_{1\leq k\leq r}$. We then obtain $\hat{E}\left[c \mid \boldsymbol{e}', \boldsymbol{\pi}^0\right]$ by simply considering those $\boldsymbol{\Pi}_l^k$ that correspond to experiments specified in $\boldsymbol{e}'$.

$$\hat{E}\left[c \mid \boldsymbol{e}', \boldsymbol{\pi}^0\right] = \left|\boldsymbol{\Pi}^0\right| + \frac{1}{r}\sum_k \left|\boldsymbol{\Pi}_{\boldsymbol{l}}^k\right| \tag{9.3}$$

The expected coverage for any sequence of m experiments can be estimated by sampling a tractable amount of protein trajectories. Therefore, we assume an experiment sequence $\boldsymbol{e} = \left(\{1\}^m, ..., \{J\}^m\right)$ and sample r trajectories $\left(\boldsymbol{\pi}^k\right)_{1\leq k\leq r}$ from $\boldsymbol{\pi} \mid \boldsymbol{e}, \boldsymbol{\pi}^0$, thereby obtaining protein discovery sets $\left(\boldsymbol{\Pi}^k\right)_{1\leq k\leq r}$. By definition $\boldsymbol{e}$ comprises all experiments sequences $\boldsymbol{e}'$ of length m. The expected coverage for any experiment sequence $\boldsymbol{e}'$ of length m can now be estimated on the basis the protein discovery sets for $\boldsymbol{e}$ as summarized in equation (9.3).

Experiment Selection can now be performed on the basis of the trajectories sampled for the $J \times m$ sized experiment sequence $\boldsymbol{e} = \left(\{1\}^m, ..., \{J\}^m\right)$. We therefore introduce the *Empirical Experiment Selection* task that substitutes the exact coverage expectation values with the respective empirical estimates obtained from a finite sample of protein trajectories.

Empirical Experiment Selection Let $\boldsymbol{\pi}^0$ be a set of identification trajectories obtained for J biological samples. $\left(\boldsymbol{\pi}^k\right)_{1\leq k\leq r}$ and $\left(\boldsymbol{\Pi}^k\right)_{1\leq k\leq r}$ are r trajectories/discovery sets sampled from $\boldsymbol{\pi} \mid \boldsymbol{e} = \left(\{1\}^m, ..., \{J\}^m\right), \boldsymbol{\pi}^0$. Estimate the optimal sequence $\hat{\boldsymbol{e}}^*$ comprising m additional experiments as the sequence maximizing the empirical estimate for coverage expectation.

$$\hat{\boldsymbol{e}}^* = \arg\max_{|\boldsymbol{l}|=m} \sum_k \left|\boldsymbol{\Pi}_{\boldsymbol{l}}^k\right| \tag{9.4}$$

9.5 Reduction to maximum k-cover

We reduce *Empirical Experiment Selection* to maximum k-cover [70]. Therefore we introduce the expanded sets and subsequently show that *Empirical Experiment Selection* is a maximum k-cover instance over a set of expanded sets.

Maximum K-Cover Let $(\boldsymbol{U}_l)_{1 \leq l \leq m}$ be a collection of sets. The maximum k-cover C is defined as

$$C = \arg\max_{|C'|=k} \left| \bigcup_{l \in C'} \boldsymbol{U}_l \right| \tag{9.5}$$

Expanded Sets Let $\boldsymbol{\pi}^0$ be a set of identification trajectories obtained for J biological samples, $\boldsymbol{e}$ an experiment sequence and $\left(\boldsymbol{\Pi}^k\right)_{1 \leq k \leq r}$ r sets of protein discoveries obtained from $\boldsymbol{\pi} \mid \boldsymbol{e}, \boldsymbol{\pi}^0$. Expanded sets are defined as follows.

$$\breve{\boldsymbol{\Pi}}_l := \bigcup_k \breve{\boldsymbol{\Pi}}_l^k := \bigcup_k \left\{ (k, \pi) \mid \pi \in \boldsymbol{\Pi}_l^k \right\} \tag{9.6}$$

It is easy to see the following properties of expanded sets:

(1) $\left|\bigcup_l \boldsymbol{\Pi}_l^k\right| = \left|\bigcup_l \breve{\boldsymbol{\Pi}}_l^k\right|$

(2) $\left|\breve{\boldsymbol{\Pi}}_l^{k_1}\right| + \left|\breve{\boldsymbol{\Pi}}_{l'}^{k_2}\right| = \left|\breve{\boldsymbol{\Pi}}_l^{k_1} \cup \breve{\boldsymbol{\Pi}}_{l'}^{k_2}\right| \quad \forall k_1 \neq k_2$

(3) $\sum_k \left|\boldsymbol{\Pi}_{\boldsymbol{l}}^k\right| = \left|\bigcup_{l \in \boldsymbol{l}} \breve{\boldsymbol{\Pi}}_l\right|$

Proof. $\sum_k \left|\boldsymbol{\Pi}_{\boldsymbol{l}}^k\right| \overset{(\text{def})}{=} \sum_k \left|\bigcup_l \boldsymbol{\Pi}_l^k\right| \overset{(1)}{=} \sum_k \left|\bigcup_l \breve{\boldsymbol{\Pi}}_l^k\right| \overset{(2)}{=} \left|\bigcup_l \bigcup_k \breve{\boldsymbol{\Pi}}_l^k\right| \overset{(\text{def})}{=} \left|\bigcup_l \breve{\boldsymbol{\Pi}}_l\right|$

Empirical Experiment Selection can now be reduced to maximum k-cover by resorting to expanded sets.

Proposition. 9.5.1 *Empirical Sample Selection reduces to a maximum set coverage instance that consists of finding m sets from* $\left(\breve{\boldsymbol{\Pi}}_l\right)_{l \in \boldsymbol{l} \subseteq \{1,..,J \times m\}}$ *such that their union has maximal cardinality.*

Proof. $\hat{\boldsymbol{e}}^* \overset{(\text{def})}{=} \arg\max_{|\boldsymbol{l}|=m} \sum_k \left|\boldsymbol{\Pi}_{\boldsymbol{l}}^k\right| \overset{(3)}{=} \arg\max_{|\boldsymbol{l}|=m} \left|\bigcup_{l \in \boldsymbol{l}} \breve{\boldsymbol{\Pi}}_l\right|$

9.6 Computation of maximum k-cover

Although finding a maximum k-cover is NP-hard [70], there are several approaches to this algorithmic problem that show good performance in practice. The most popular approach is an efficient greedy heuristic.

```
Greedy K-Cover($k, \boldsymbol{U}_1, ..., \boldsymbol{U}_m$)
$\boldsymbol{U} \leftarrow \bigcup_{l=1}^{m} \boldsymbol{U}_l$
$C \leftarrow \emptyset$
repeat
    $l \leftarrow \arg\max_{l'} |\boldsymbol{U}_{l'} \cap \boldsymbol{U}|$
    $C \leftarrow C \cup \{l\}$
    $\boldsymbol{U} \leftarrow \boldsymbol{U} \setminus \boldsymbol{U}_l$
until $|C| = k$
return $C$
```

Greedy k-cover achieves approximates the size of the globally optimal solution at least up to a factor of $1 - 1/e$ [64].

Maximum k-cover can also be formulated as an integer linear program [28]. For each $u_i \in \bigcup_l \boldsymbol{U}_l$ and for each set $\boldsymbol{U}_j$ we introduce variables y_i and respectively x_j. The assignment of the variables x_j indicates whether the respective set $\boldsymbol{U}_j$ is included in the cover ($x_j = 0$) or not ($x_j = 1$). The constraints and the objective of the program ensure that $y_i = 1$ iff the variable assignments for the x_j correspond to a cover including u_i and that $y_i = 0$ iff the respective cover does not include u_i.

$$\begin{array}{llll}
\text{max} & \sum_i y_i & & \\
\text{s.t.} & & & \\
 & \sum_j x_j & \leq & k \\
 & \sum_{j:u_i \in \boldsymbol{U}_j} x_j & \geq & y_i \quad \forall i \\
 & y_i & \geq & 0 \quad \forall i \\
 & y_i & \leq & 1 \quad \forall i \\
 & x_j & \in & \{0,1\} \quad \forall j
\end{array}$$

This formulation makes maximum k-cover amenable to standard solvers for integer linear programs. A lot of instances occurring in practice can be solved efficiently and exactly with this formulation.

9.7 Discussion

This chapter describes how the delineated proteome coverage prediction approaches integrate into the task of optimally designing a shotgun proteomics study, i.e. to maximize the expected proteome coverage for a user defined amount of experimentation [21].

We assume that an initial number of LC-MS/MS experiments has been performed for a set of protein mixture samples that were obtained from possibly various biological sources and/or multiple fractionation steps. Our approach estimates a sequence of LC-MS/MS experiments that yields maximal expected proteome coverage. We formulate this task as an optimization problem and show how it reduces to maximum k-coverage, a variant of the well known set cover problem.

This approach enables researches, at an early stage of a large proteomics study, to quantify the potential of the individual samples/fractions to map out a proteome in the context of all others and to which extent each deserves to be further studied by LC-MS/MS experiments. By these means only the most informative data is acquired. This strategy therefore contributes to the efficiency of a shotgun proteomics study aiming at extensive proteome coverage. We discussed the deteriorating effect of dataset size in the context of the error propagation from peptide-spectrum matches to the level of protein identifications in chapter 4. Therefore, it is conceivable that, beyond efficiency considerations, avoiding redundant experiments leads to higher sensitivity than strategies affording exhaustive experimentation.

Our approach motivates a new strategy towards more comprehensive coverage of a proteome of interest. Knowing about the dynamic nature of a proteome, we would like to account for a possibly very large number of different biological samples that represent the proteome under a variety of conditions/perturbations. While it might be feasible to probe each of these many samples by a couple of whole cell lysate LC-MS/MS experiments in such a scenario, it is not possible to follow up with extensive multidimensional fractionation experiments on each sample. Our approach enables us to identify the most informative subset of samples to be analyzed in depth. It is likely that the biological samples feature highly redundant proteomes. In this case the informative subset could turn out to be small enough to allow comprehensive in depth analysis and to reveal more of the true richness of the underlying proteome.

We expect that our approaches to proteome coverage prediction and optimal design of shotgun proteomics studies provide unique strategies to contribute to a more comprehensive view on proteomes.

10 Conclusion

The mechanisms underlying complex diseases like cancer or diabetes are not well understood until today. Systems biology aims to elucidate these mechanisms by means of expressive models of biological systems that take into account the entirety of their components. Mass spectrometry based proteomics significantly contributes to systems biology approaches by providing technologies to comprehensively characterize a proteome, i.e. the protein components of a biological system.

Mass spectrometry based proteomics constitutes an efficient and generic approach to measure almost any member of a proteome. Throughput and sensitivity of contemporary shotgun proteomics approaches allow to measure hundreds of proteins with a manageable amount of experimentation and to further map out large parts of a proteome for more extensive and elaborate experimental setups.

Shotgun proteomics experiments generate data that is very informative about the studied proteome. However, the inference steps from the mass spectrometrical data up to the level of protein identifications are not trivial. The typically generated amount of data and the need for objective and reproducible inference routines presuppose automation of the inference steps. This challenging requirement translates into a diversity of statistical and algorithmic problems that have attracted a lot of interest in the statistics and machine learning community.

Shotgun proteomics data is inherently noisy. The interpretation of mass spectrometrical data is therefore inevitably afflicted with uncertainty. The ability to term quantitative confidence measures for interpretations like peptide-spectrum matches or protein identifications is an indispensable prerequisite to appropriately evaluate the outcome of a proteomics study.

This thesis contributes methods to estimate confidence measures for proteome measurements. First, a target-decoy strategy to estimate false discovery rates for peptide-spectrum matches from iterated database searches is presented. This approach allowed to compile a set of reliable identifications from an iterated database search allowing for hundreds of amino acid modifications. This target-decoy approach thereby enabled to reliably reveal various novel and yet frequent variations of a proteome. Second, this thesis introduces a generalized target-decoy strategy to estimate false discovery rates for protein identifications. Although protein identifications constitute the biologically relevant outcome of a proteomics study, confidence measures have been typically reported at the level peptide-spectrum matches, implicitly assuming them to be a reasonable approximation for those of protein identifications. We show how errors for peptide-spectrum matches propagate non-trivially to the level of peptide-spectrum match assemblies, i.e. protein identifications. We discovered that false discovery rates for proteins identifications are significantly larger than for peptide-spectrum matches. We found this discrepancy to be more pronounced the larger the underlying volume of mass spectrometrical data. This finding has implications for the interpretation of data acquired in large shotgun proteomics studies aiming at extensive proteome coverage. This thesis presents a formal approach to derive guidelines how to optimally interpret the mass spectrometrical data for a given set of interpretation tools, such as search engines or protein inference methods. For the datasets studied in this thesis we found that the best strategy consists of accounting for all spectral data of sufficiently high quality. We particularly found that for large studies the spectral data has to be much more carefully selected than appreciated before. Beyond individual studies, this finding also applies to proteomics data repositories. Our approach to estimate false discovery rates for protein identifications can be used to automatically curate such repositories and thereby enhance systems biology projects building on these valuable resources.

It turns out that shotgun proteomics studies aiming at extensive proteome coverage acquire a lot of redundant, non-informative data by measuring the same peptides over and over again. It is beneficial to avoid this redundancy for efficiency and sensitivity considerations. Avoiding non-informative experiments obviously saves resources without loosing in terms of proteome coverage. Avoiding these experiments yields smaller datasets and therefore, as we have seen in our study on protein false discovery rates, also avoids accumulation of false positive protein identifications, potentially allowing to confidently identify more weakly evidenced proteins. In conclusion, it is desirable to design a shotgun proteomics study such that it focuses on informative experiments.

This thesis contributes a framework to design shotgun proteomics studies in order to maximize their expected proteome coverage. This framework lends itself to predict the optimal sequence of experiment repetitions from a small amount of already performed (LC-MS/MS) experiments on a set of protein mixtures. This approach builds on the ability to predict proteome coverage for an individual sequence of experiments. This thesis develops a non-parametric Bayesian approach to this task. The peptide distributions arising in the course of LC-MS/MS experiments are characterized by means of hierarchical Dirichlet processes and variants thereof. The intricate relationship among similar peptide distributions over multidimensional fractionation experiments inspired the formulation of a novel class of hierarchical processes, the fractal Dirichlet process. We showed how these processes can be applied to accurately predict proteome coverage from a small amount of experiments. Proteome coverage prediction can be used to define quantitative stop criteria that take into account the accumulation of false positive protein identifications as well as the rate of novel protein discoveries for some unit of experimentation. We showed cases where the maximally achievable coverage at a user defined false discovery rate did not coincide with maximal number of true positive identifications. Proteome coverage prediction can furthermore assist experimental method development by providing an additional quantitative measure for coverage potential. We finally described how proteome coverage prediction formally integrates into estimating an optimal sequence of experiments that maximizes the expected proteome coverage. This optimization task can be reduced to maximum k-cover, a variant of the well known set cover problem. It will be interesting to evaluate the impact of optimal design on efficiency as well as sensitivity of the respective shotgun proteomics study.

The statistical concepts developed in this thesis are not confined to application scenarios in mass spectrometry based proteomics. The first part of this thesis has generalized the target-decoy strategy to estimate false discovery rates for protein identifications, i.e. assemblies of peptide-spectrum matches. The target-decoy strategy is more generally applicable to assess the confidence of inference results obtained from assigning hypotheses (e.g. proteins) from a collection of hypotheses (e.g. protein database) to observations (e.g. fragment ion spectra). After having compiled a suitable collection of decoy hypotheses, the target-decoy strategy can be straightforwardly applied, e.g. to a retrieval task that consists of assigning song snippets to songs from a music database. The second part of this thesis introduces the fractal Dirichlet process, i.e. a novel measure over a set of discrete measures generalizing the hierarchical Dirichlet process. We could show that such a process is better suited to capture the intricate relationships among the peptide distributions arising in integrated shotgun proteomics experiments. It will be interesting to see whether other application scenarios, like e.g. language modeling, also exhibit structures best captured by the fractal Dirichlet process.

In conclusion, this thesis contributes novel statistical methods that enable the experimentalist to rationally decide which data to acquire and which of the many available data analysis strategy to choose in order to efficiently achieve the most extensive and yet reliable proteome coverage. The resulting curated data will constitute an important resource for targeted quantitative proteomics approaches such as selected reaction monitoring and thereby strengthen the role of proteomics data in the context of systems biology projects building on heterogeneous data sources.

This thesis addresses questions and develops methods that lie at the interface of biology and machine learning. This work exemplifies the power of machine learning concepts to tackle biologically relevant problems as well as how biology can inspire a novel kind of general tasks that lead to novel concepts in machine learning. I am convinced that both fields will keep on benefiting from this synergy in the future.

Bibliography

[1] M. Adamski, T. Blackwell, R. Menon, L. Martens, H. Hermjakob, C. Taylor, G. S. Omenn, and D. J. States. Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project. *Proteomics*, 5(13):3246–61, 2005.

[2] R. Aebersold. A stress test for mass spectrometry-based proteomics. *Nat Methods*, 6(6):411–2, 2009.

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

[4] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[5] K. Baerenfaller, J. Grossmann, M. A. Grobei, R. Hull, M. Hirsch-Hoffmann, S. Yalovsky, P. Zimmermann, U. Grossniklaus, W. Gruissem, and S. Baginsky. Genome-Scale Proteomics Reveals Arabidopsis thaliana Gene Models and Proteome Dynamics. *Science*, 320(5878):938–941, 2008.

[6] V. Bafna and N. Edwards. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 17 Suppl 1:S13–21, 2001.

[7] N. Bandeira, D. Tsur, A. Frank, and P. A. Pevzner. Protein identification by spectral networks analysis. *Proc Natl Acad Sci U S A*, 104(15):6140–5, 2007.

[8] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

[9] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[10] D. R. Bentley. Whole-genome re-sequencing. *Curr Opin Genet Dev*, 16(6):545–52, 2006.

[11] K. Biemann. Mass spectrometry of peptides and proteins. *Annu Rev Biochem*, 61:977–1010, 1992.

[12] D. Blackwell and J. B. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[13] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16:106, 2004.

[14] B. Bodenmiller, L. N. Mueller, M. Mueller, B. Domon, and R. Aebersold. Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat Methods*, 4(3):231–7, 2007.

[15] E. Brunner, C. H. Ahrens, S. Mohanty, H. Baetschmann, S. Loevenich, F. Potthast, E. W. Deutsch, C. Panse, U. de Lichtenberg, O. Rinner, H. Lee, P. G. Pedrioli, J. Malmstrom, K. Koehler, S. Schrimpf, J. Krijgsveld, F. Kregenow, A. J. Heck, E. Hafen, R. Schlapbach, and R. Aebersold. A high-quality catalog of the Drosophila melanogaster proteome. *Nat Biotechnol*, 25(5):576–83, 2007.

[16] D. Chelius and P. V. Bondarenko. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res*, 1(4):317–23, 2002.

[17] T. Chen, M. yang Kao, M. Tepel, J. Rush, and G. M. Church. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 8:325–337, 2001.

[18] H. Choi and A. Nesvizhskii. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *Journal of Proteome Research*, 7(01):47–50, 2007.

[19] D. S. Chu, H. Liu, P. Nix, T. F. Wu, E. J. Ralston, J. R. Yates, 3rd, and B. J. Meyer. Sperm chromatin proteomics identifies evolutionarily conserved fertility factors. *Nature*, 443(7107):101–5, 2006.

[20] M. Claassen, R. Aebersold, and J. M. Buhmann. The fractal dirichlet process. *submitted.*

[21] M. Claassen, R. Aebersold, and J. M. Buhmann. Optimal design of integrated proteomics experiments. *in prep.*

[22] M. Claassen, R. Aebersold, and J. M. Buhmann. Reliable, efficient and comprehensive identification of modified peptides with an iterated target-decoy database search strategy. *in prep.*

[23] M. Claassen, R. Aebersold, and J. M. Buhmann. Proteome coverage prediction with infinite Markov models. *Bioinformatics*, 25(12):i154–60, 2009.

[24] M. Claassen, R. Aebersold, and J. M. Buhmann. Proteome Coverage Prediction for Integrated Proteomics Datasets. *RECOMB*, 2010.

[25] M. Claassen*, L. Reiter*, M. O. Hengartner, J. M. Buhmann, and R. Aebersold. Generic comparison of protein inference engine families. *RECOMB Satellite for Computational Proteomics*, 2010.

[26] M. Claassen*, L. Reiter*, S. P. Schrimpf, M. Jovanovic, A. Schmidt, J. M. Buhmann, M. O. Hengartner, and R. Aebersold. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Mol Cell Proteomics*, 8(11):2405–2417, 2009.

[27] J. Colinge, A. Masselot, I. Cusin, E. Mahe, A. Niknejad, G. Argoud-Puy, S. Reffas, N. Bederr, A. Gleizes, P. A. Rey, and L. Bougueleret. High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. *Proteomics*, 4(7):1977–84, 2004.

[28] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms.* MIT Press.

[29] R. Craig and R. Beavis. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid communications in mass spectrometry*, 17(20):2310–2316, 2003.

[30] R. Craig and R. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.

[31] R. Craig, J. C. Cortens, D. Fenyo, and R. C. Beavis. Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res*, 5(8):1843–9, 2006.

[32] R. Craig, J. P. Cortens, and R. C. Beavis. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res*, 3(6):1234–42, 2004.

[33] D. Creasy and J. Cottrell. Unimod: Protein modifications for mass spectrometry. *PROTEOMICS-Clinical Applications*, 4(6):1534–1536.

[34] V. Dank, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6(3-4):327–342, 1999.

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

[36] 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):1–38, 1977.

[37] F. Desiere, E. W. Deutsch, A. I. Nesvizhskii, P. Mallick, N. L. King, J. K. Eng, A. Aderem, R. Boyle, E. Brunner, S. Donohoe, N. Fausto, E. Hafen, L. Hood, M. G. Katze, K. A. Kennedy, F. Kregenow, H. Lee, B. Lin, D. Martin, J. A. Ranish, D. J. Rawlings, L. E. Samelson, Y. Shiio, J. D. Watts, B. Wollscheid, M. E. Wright, W. Yan, L. Yang, E. C. Yi, H. Zhang, and R. Aebersold. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol*, 6(1):R9, 2005.

[38] B. Domon and R. Aebersold. Mass spectrometry and protein analysis. *Science*, 312(5771):212–7, 2006.

[39] P. Edman. A method for determination of the amino acid sequence in peptides. *Acta Chem. Scand*, 4:283–293, 1950.

[40] B. Efron and R. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol*, 23(1):70–86, 2002.

[41] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–14, 2007.

[42] J. Eng, A. McCormack, J. Yates, et al. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

[43] J. Eriksson and D. Fenyo. Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nat Biotechnol*, 25(6):651–5, 2007.

[44] J. Eriksson, D. Fenyo, et al. Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results. *Journal of Proteome Research*, 3(1):32–36, 2004.

[45] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.

[46] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[47] S. B. Ficarro, M. L. McCleland, P. T. Stukenberg, D. J. Burke, M. M. Ross, J. Shabanowitz, D. F. Hunt, and F. M. White. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat Biotechnol*, 20(3):301–5, 2002.

[48] H. I. Field, D. Fenyo, and R. C. Beavis. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics*, 2(1):36–47, 2002.

[49] B. Fischer, V. Roth, F. Roos, J. Grossmann, S. Baginsky, P. Widmayer, W. Gruissem, and J. M. Buhmann. NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal Chem*, 77(22):7265–73, 2005.

[50] R. Fleischmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, and e. al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, 1995.

[51] L. J. Foster, C. L. de Hoog, Y. Zhang, Y. Zhang, X. Xie, V. K. Mootha, and M. Mann. A mammalian organelle map by protein correlation profiling. *Cell*, 125(1):187–99, 2006.

[52] A. Frank and P. Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem*, 77(4):964–73, 2005.

[53] A. Frank, S. Tanner, V. Bafna, and P. Pevzner. Peptide sequence tags for fast database search in mass-spectrometry. *J Proteome Res*, 4(4):1287–95, 2005.

[54] B. E. Frewen, G. E. Merrihew, C. C. Wu, W. S. Noble, and M. J. MacCoss. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal Chem*, 78(16):5678–84, 2006.

[55] J. Giddings. Concepts and comparisons in multidimensional separation. *Journal of High Resolution Chromatography*, 10(5):319–323, 1987.

[56] W. Gilks, N. Best, and K. Tan. Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):455–472, 1995.

[57] T. Glatter, A. Wepf, R. Aebersold, and M. Gstaiger. An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol Syst Biol*, 5:237, 2009.

[58] M. A. Grobei, E. Qeli, E. Brunner, H. Rehrauer, R. Zhang, B. Roschitzki, K. Basler, C. H. Ahrens, and U. Grossniklaus. Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. *Genome Research*, 19(10):1786–1800, 2009.

[59] N. Gupta and P. Pevzner. False discovery rates of protein identifications: a strike against the two-peptide rule. *Journal of Proteome Research*, 8(9):4173–4181, 2009.

[60] S. P. Gygi, G. L. Corthals, Y. Zhang, Y. Rochon, and R. Aebersold. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A*, 97(17):9390–5, 2000.

[61] 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):994–9, 1999.

[62] W. J. Henzel, T. M. Billeci, J. T. Stults, S. C. Wong, C. Grimley, and C. Watanabe. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A*, 90(11):5011–5, 1993.

[63] K. Hilpert, D. F. Winkler, and R. E. Hancock. Peptide arrays on cellulose support: SPOT synthesis, a time and cost efficient method for synthesis of large numbers of peptides in a parallel and addressable fashion. *Nat Protoc*, 2(6):1333–49, 2007.

[64] D. Hochbaum and A. Pathria. Analysis of the greedy approach in problems of maximum k-coverage. *Naval Research Logistics*, 45(6):615–627, 1998.

[65] D. F. Hunt, J. R. Yates, J. Shabanowitz, S. Winston, and C. R. Hauer. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 83(17):6233–6237, 1986.

[66] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

[67] L. Kall, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*, 7(1):29–34, 2008.

[68] E. Kapp, F. Schütz, L. Connolly, J. Chakel, J. Meza, C. Miller, D. Fenyo, J. Eng, J. Adkins, G. Omenn, et al. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics*, 5(13):3475, 2005.

[69] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem*, 60(20):2299–301, 1988.

[70] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[71] A. Keller, J. Eng, N. Zhang, X. Li, and R. Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular systems biology*, 1, 2005.

[72] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*, 74(20):5383–92, 2002.

[73] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–6, 1958.

[74] S. Kim, N. Gupta, and P. A. Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res*, 7(8):3354–63, 2008.

[75] N. L. King, E. W. Deutsch, J. A. Ranish, A. I. Nesvizhskii, J. S. Eddes, P. Mallick, J. Eng, F. Desiere, M. Flory, D. B. Martin, B. Kim, H. Lee, B. Raught, and R. Aebersold. Analysis of the Saccharomyces cerevisiae proteome with PeptideAtlas. *Genome Biol*, 7(11):R106, 2006.

[76] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–4, 2002.

[77] T. Knöpfel. Proteome Coverage Prediction with Variants of the Chinese Restaurant Process Construction . Master's thesis, ETH Zurich, 2010.

[78] O. Kohlbacher, K. Reinert, C. Gropl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, and M. Sturm. TOPP–the OpenMS proteomics pipeline. *Bioinformatics*, 23(2):e191, 2007.

[79] B. Kuster, M. Schirle, P. Mallick, and R. Aebersold. Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol*, 6(7):577–83, 2005.

[80] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein, and R. Aebersold. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, 7(5):655–67, 2007.

[81] V. Lange, J. A. Malmstrom, J. Didion, N. L. King, B. P. Johansson, J. Schafer, J. Rameseder, C. H. Wong, E. W. Deutsch, M. Y. Brusniak, P. Buhlmann, L. Bjorck, B. Domon, and R. Aebersold. Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. *Mol Cell Proteomics*, 7(8):1489–500, 2008.

[82] V. Lange, J. A. Malmstrom, J. Didion, N. L. King, B. P. Johansson, J. Schafer, J. Rameseder, C.-H. o. Wong, E. W. Deutsch, M.-Y. Brusniak, P. Buhlmann, L. Bjorck, B. Domon, and R. Aebersold. Targeted Quantitative Analysis of Streptococcus pyogenes Virulence Factors by Multiple Reaction Monitoring. *Mol Cell Proteomics*, 7(8):1489–1500, 2008.

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

[84] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom*, 17(20):2337–42, 2003.

[85] M. J. MacCoss, C. C. Wu, and J. R. Yates, 3rd. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal Chem*, 74(21):5593–9, 2002.

[86] P. Mallick, M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, and R. Aebersold. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol*, 25(1):125–31, 2007.

[87] J. Malmstrom, M. Beck, A. Schmidt, V. Lange, E. W. Deutsch, and R. Aebersold. Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature*, 460(7256):762–5, 2009.

[88] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*, 66(24):4390–9, 1994.

[89] E. M. Marcotte. How do shotgun proteomics algorithms identify proteins? *Nat Biotechnol*, 25(7):755–7, 2007.

[90] L. Martens, H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove, and R. Apweiler. PRIDE: the proteomics identifications database. *Proteomics*, 5(13):3537–45, 2005.

[91] F. W. McLafferty. Tandem mass spectrometry. *Science*, 214(4518):280–287, 1981.

[92] P. E. Michel, F. Reymond, I. L. Arnaud, J. Josserand, H. H. Girault, and J. S. Rossier. Protein fractionation in a multicompartment device using Off-Gel isoelectric focusing. *Electrophoresis*, 24(1-2):3–11, 2003.

[93] A. W. Moore, Jr, J. P. Larmann, Jr, A. V. Lemmo, and J. W. Jorgenson. Two-dimensional liquid chromatography-capillary electrophoresis techniques for analysis of proteins and peptides. *Methods Enzymol*, 270:401–19, 1996.

[94] R. Moore, M. Young, and T. Lee. Qscore: an algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, 2002.

[95] A. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & Cellular Proteomics*, 4(10):1419, 2005.

[96] A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75(17):4646–58, 2003.

[97] A. I. Nesvizhskii, O. Vitek, and R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods*, 4(10):787–97, 2007.

[98] P. H. O'Farrell. High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, 250(10):4007–4021, 1975.

[99] J. V. Olsen, J. C. Schwartz, J. Griep-Raming, M. L. Nielsen, E. Damoc, E. Denisov, O. Lange, P. Remes, D. Taylor, M. Splendore, E. R. Wouters, M. Senko, A. Makarov, M. Mann, and S. Horning. A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics*, 8(12):2759–69, 2009.

[100] G. S. Omenn, D. J. States, M. Adamski, T. W. Blackwell, R. Menon, H. Hermjakob, R. Apweiler, B. B. Haab, R. J. Simpson, J. S. Eddes, E. A. Kapp, R. L. Moritz, D. W. Chan, A. J. Rai, A. Admon, R. Aebersold, J. Eng, W. S. Hancock, S. A. Hefta, H. Meyer, Y. K. Paik, J. S. Yoo, P. Ping, J. Pounds, J. Adkins, X. Qian, R. Wang, V. Wasinger, C. Y. Wu, X. Zhao, R. Zeng, A. Archakov, A. Tsugita, I. Beer, A. Pandey, M. Pisano, P. Andrews, H. Tammen, D. W. Speicher, and S. M. Hanash. Overview of the HUPO Plasma Proteome Project:

results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics*, 5(13):3226–45, 2005.

[101] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1(5):376–86, 2002.

[102] G. J. Opiteck and J. W. Jorgenson. Two-dimensional SEC/RPLC coupled to mass spectrometry for the analysis of peptides. *Anal Chem*, 69(13):2283–91, 1997.

[103] J. Peng, J. E. Elias, C. C. Thoreen, L. J. Licklider, and S. P. Gygi. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res*, 2(1):43–50, 2003.

[104] J. Peng and S. P. Gygi. Proteomics: the move to mixtures. *J Mass Spectrom*, 36(10):1083–91, 2001.

[105] D. Perkins, D. Pappin, D. Creasy, J. Cottrell, et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[106] P. Picotti, B. Bodenmiller, L. N. Mueller, B. Domon, and R. Aebersold. Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell*, 138(4):795–806, 2009.

[107] J. Pitman. Combinatorial stochastic processes. *Technical Report 621, Dept.Statistics, U.C. Berkeley*, 2002.

[108] T. S. Price, M. B. Lucitt, W. Wu, D. J. Austin, A. Pizarro, A. K. Yocum, I. A. Blair, G. A. FitzGerald, and T. Grosser. EBP, a Program for Protein Identification Using Multiple Tandem Mass Spectrometry Datasets. *Mol Cell Proteomics*, 6(3):527–536, 2007.

[109] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[110] J. Rappsilber and M. Mann. What does it mean to identify a protein in proteomics? *Trends Biochem Sci*, 27(2):74–8, 2002.

[111] C. Rasmussen. The infinite Gaussian mixture model. *Advances in neural information processing systems*, 12:554–560, 2000.

[112] R. Sadygov, H. Liu, and J. Yates. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem*, 76(6):1664–1671, 2004.

[113] R. Sadygov and J. Yates 3rd. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Analytical chemistry*, 75(15):3792, 2003.

[114] V. Santoni, M. Molloy, and T. Rabilloud. Membrane proteins and proteomics: un amour impossible? *Electrophoresis*, 21(6):1054–70, 2000.

[115] R. Schiess, B. Wollscheid, and R. Aebersold. Targeted proteomic strategy for clinical biomarker discovery. *Mol Oncol*, 3(1):33–44, 2009.

[116] A. Schmidt, M. Beck, J. Malmstroem, H. H. N. Lam, M. Claassen, D. Campell, and R. Aebersold. Proteome-wide high-throughput screening using directed mass spectrometry: Application to the human pathogen l. interrogans. *in prep.*

[117] A. Schmidt, M. Claassen, and R. Aebersold. Directed mass spectrometry: towards hypothesis-driven proteomics. *Curr Opin Chem Biol*, 13(5-6):510–7, 2009.

[118] A. Schmidt, N. Gehlenborg, B. Bodenmiller, L. N. Mueller, D. Campbell, M. Mueller, R. Aebersold, and B. Domon. An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Mol Cell Proteomics*, 7(11):2138, 2008.

[119] S. Schrimpf, M. Weiss, L. Reiter, C. Ahrens, M. Jovanovic, J. Malmström, E. Brunner, S. Mohanty, M. Lercher, P. Hunziker, et al. Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol*, 7(3):e48, 2009.

[120] S. P. Schrimpf, M. Jovanovic, L. Reiter, M. Claassen, J. Malmstrm, A. Sendoel, E. Brunner, B. Roschitzki, C. Panse, R. Schlapbach, P. E. Hunziker, R. Aebersold, and M. O. Hengartner. Complementary separation techniques to identify complex proteomes. *in prep.*

[121] R. Scopes. *Protein purification: principles and practice.* Springer.

[122] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[123] I. Shilov, S. Seymour, A. Patel, A. Loboda, W. Tang, S. Keating, C. Hunter, L. Nuwaysir, and D. Schaeffer. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Molecular & Cellular Proteomics*, 6(9):1638, 2007.

[124] K.-A. Sohn and E. P. Xing. Hidden Markov Dirichlet process: modeling genetic recombination in open ancestral space. In B. Schölkopf, J. Platt, and T. Hofman, editors, *Advances in neural information processing systems 19*, pages 1305–1312. MIT Press, Cambridge, MA, 2007.

[125] J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–8, 2002.

[126] D. States, G. Omenn, T. Blackwell, D. Fermin, J. Eng, D. Speicher, and S. Hanash. Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nature Biotechnology*, 24(3):333, 2006.

[127] D. J. States, G. S. Omenn, T. W. Blackwell, D. Fermin, J. Eng, D. W. Speicher, and S. M. Hanash. Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat Biotechnol*, 24(3):333–8, 2006.

[128] S. E. Stein. Chemical substructure identification by mass spectral library searching. *Journal of the American Society for Mass Spectrometry*, 6(8):644 – 655, 1995.

[129] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–5, 2003.

[130] J. E. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A*, 101(26):9528–33, 2004.

[131] D. Tabb, L. Smith, L. Breci, V. Wysocki, D. Lin, and J. Yates III. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem*, 75(5):1155–1163, 2003.

[132] J. A. Taylor and R. S. Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem*, 73(11):2594–604, 2001.

[133] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 985–992, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[134] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, December 2006.

[135] W. Tong, A. Link, J. K. Eng, and J. R. Yates, 3rd. Identification of proteins in complexes by solid-phase microextraction/multistep elution/capillary electrophoresis/tandem mass spectrometry. *Anal Chem*, 71(13):2270–8, 1999.

[136] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P. Pevzner. Identification of post-translational modifications via blind search of mass-spectra. *Nature biotechnology*, 23:1562–1567, 2005.

[137] J. C. Venter, M. D. Adams, E. W. Myers, et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

[138] M. P. Washburn, D. Wolters, and J. R. Yates, 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, 19(3):242–7, 2001.

[139] D. B. Weatherly, J. A. Atwood, 3rd, T. A. Minning, C. Cavola, R. L. Tarleton, and R. Orlando. A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol Cell Proteomics*, 4(6):762–72, 2005.

[140] H. Wenschuh, R. Volkmer-Engert, M. Schmidt, M. Schulz, J. Schneider-Mergener, and U. Reineke. Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides. *Biopolymers*, 55(3):188–206, 2000.

[141] A. Wepf, T. Glatter, A. Schmidt, R. Aebersold, and M. Gstaiger. Quantitative interaction proteomics using mass spectrometry. *Nat Methods*, 6(3):203–5, 2009.

[142] C. M. Whitehouse, R. N. Dreyer, M. Yamashita, and J. B. Fenn. Electrospray interface for liquid chromatographs and mass spectrometers. *Anal Chem*, 57(3):675–9, 1985.

[143] M. R. Wilkins, C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J. C. Sanchez, J. X. Yan, A. A. Gooley, G. Hughes, I. Humphery-Smith, K. L. Williams, and D. F. Hochstrasser. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)*, 14(1):61–5, 1996.

[144] B. Wollscheid, D. Bausch-Fluck, C. Henderson, R. O'Brien, M. Bibel, R. Schiess, R. Aebersold, and J. Watts. Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nature Biotechnology*, 27(4):378–386, 2009.

[145] C. C. Wu, M. J. MacCoss, K. E. Howell, and J. R. Yates, 3rd. A method for the comprehensive proteomic analysis of membrane proteins. *Nat Biotechnol*, 21(5):532–8, 2003.

[146] V. H. Wysocki, G. Tsaprailis, L. L. Smith, and L. A. Breci. Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom*, 35(12):1399–406, 2000.

[147] J. R. Yates, 3rd, S. F. Morgan, C. L. Gatlin, P. R. Griffin, and J. K. Eng. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal Chem*, 70(17):3557–65, 1998.

[148] B. Zhang, M. Chambers, D. Tabb, et al. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res*, 6(9):3549–3557, 2007.

[149] H. Zhang, X. J. Li, D. B. Martin, and R. Aebersold. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol*, 21(6):660–6, 2003.

[150] N. Zhang, X. J. Li, M. Ye, S. Pan, B. Schwikowski, and R. Aebersold. ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics*, 5(16):4096–106, 2005.

[151] Z. Zhang. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal Chem*, 76(14):3908–22, 2004.

[152] H. Zhou, J. D. Watts, and R. Aebersold. A systematic approach to the analysis of protein phosphorylation. *Nat Biotechnol*, 19(4):375–8, 2001.

[153] R. A. Zubarev, D. M. Horn, E. K. Fridriksson, N. L. Kelleher, N. A. Kruger, M. A. Lewis, B. K. Carpenter, and F. W. McLafferty. Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal Chem*, 72(3):563–73, 2000.

Curriculum Vitae

Name	Manfred Claassen
Date of birth	10.5.1977, Maracay, Venezuela

06/1996	Abitur at Theodor-Heuss-Gymnasium in Göttingen
10/1996 - 11/1997	Alternative civilian service at the nursing home *Blindenheim*, Freiburg
01/1998 - 06/1998	Volunteer in the children's home *Siqem*, Montevideo
10/1998 - 10/2000	Basic studies (~ B.Sc.) in biochemistry University of Regensburg
10/2000 - 12/2004	Advanced studies in biochemistry University of Tübingen
08/2002 - 03/2003	Advanced studies in biochemistry University Claude-Bernard Lyon, France
12/2004	**Diplom (M.Sc.) in Biochemistry** University of Tübingen
04/2001 - 04/2002	Basic studies (~ B.Sc.) in computer science University of Tübingen
04/2002 - 02/2006	Advanced studies in computer science University of Tübingen
08/2006	**Diplom (M.Sc.) in Computer Science** University of Tübingen
09/2006 - now	Doctoral studies at ETH Zürich