Computational Methods for Mass Spectrometry-based Study of Protein-RNA or Protein-DNA Complexes and Quantitative Metaproteomics

Dissertation

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

vorgelegt von

Dipl.-Inform. Timo Sachsenberg

aus Tettnang

Tübingen
2017

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	24.04.2018
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Knut Reinert
3. Berichterstatter:	Prof. Dr. Lukas Käll

Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

Computational Methods for Mass Spectrometry-based Study of Protein-RNA or Protein-DNA Complexes and Quantitative Metaproteomics

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Abstract

In the last decade, the use of high-throughput methods has become increasingly popular in various fields of life sciences. Today, a wide range of technologies exist that allow gathering detailed quantitative insights into biological systems. With improved instrumentation and technological advances, a massive growth in data volume from these techniques has been observed. Bioinformatics copes with these heaps of data by providing computational methods that process raw data to extract biological knowledge. Computational mass spectrometry is a research field in bioinformatics that collects and analyzes data from mass-spectrometric high-throughput experiments.

In this thesis, we present two new methods as well as a new data format for computational mass spectrometry. The first method applies to a scientific problem from the field of structural biology: to determine spatial interactions between protein and nucleic acids. For this purpose, we develop experimental protocols, programs, and analysis workflows that allow identifying UV-induced cross-links in (ribo-)nucleoprotein complexes from mass spectrometry data. An outstanding feature of our method is the ability to exactly localize amino acids and (ribo-)nucleotides in contact with each other. Applied to data from yeast and human we identify new interaction partners with, to date, unmatched resolution.

The second method applies to metaproteomic studies of complex communities of microorganisms. In an unmanageable number, bacteria, simple fungi, or plants populate the most varied habitats. They are found in a high number of symbiotic or parasitic relationships which serve predominantly for the uptake of nutrients. Organisms differ in their biochemical repertoire allowing them to decompose a wide range of substrates. Remarkably, this enables functional groups of soil bacteria to even nourish themselves from environmental toxins.

We present a method from the field of metaproteomics, which allows for identification of organisms involved in substrate degradation as well as methods to group them according to their function in the degradation process. To this end, we use substrates labeled with stable isotopes, which are metabolized by the organisms. The isotope abundance in proteins serves as an indicator for the conversion of the substrate. This abundance is automatically determined by our novel computational method and assigned to the individual organisms. The automation of this process reduces the manual work from several months to a few minutes and, thus, enables large study sizes.

The third part of this work contributes to the better communication and processing of results from metabolomics and proteomics studies. We present a tabular, standardized, human-readable and machine-processable data format *mzTab* as a complement to existing data formats. We provide software components that allow processing of the format and demonstrate how the format can be integrated into complex proteomic and metabolomic workflows. The recent acceptance of *mzTab* by the largest proteomic data repositories represents a significant success. Also, we see an already widespread adoption by academic software developers and the first support by a commercial software vendor. Our novel format facilitates meta-analyses and makes research results from the field of proteomics and metabolomics available to scientists from other research areas.

Kurzfassung

In den letzten Jahren ist der Einsatz von Hochdurchsatzmethoden in den verschiedenen Feldern der Lebenswissenschaften zunehmend populärer geworden. Heutzutage existiert eine große Auswahl an Technologien die detaillierte quantitative Einblicke in biologische Systeme erlauben. Einhergehend mit technischen Fortschritten und immer besseren Instrumenten ist ein massiver Datenzuwachs durch diese Technologien zu beobachten. Der Forschungsbereich Bioinformatik hilft bei der Bewältigung dieser Datenmengen durch die Entwicklung computergestützter Methoden zur Prozessierung der Rohdaten und der Extraktion biologischer Erkenntnisse. Die computergestützte Massenspektrometrie ist ein bioinformatischer Forschungsbereich, der Daten aus massenspektrometrischen Hochdursatzexperimenten sammelt und analysiert. Im Rahmen dieser Arbeit stellen wir zwei neue Verfahren und ein neues Datenformat für die computergestützte Massenspektrometrie vor.

Das erste Verfahren dient der Strukturbiologie und hat das Ziel räumliche Interaktionen zwischen Protein und Nukleinsäuren zu bestimmen. Hierzu entwickeln wir experimentelle Protokolle, Computerprogramme und Workflows zur Datenanalyse, die es ermöglichen UV-induzierte Quervernetzungen in (Ribo-)nukleoproteinkomplexen aus massenspektrometrischen Daten zu identifizieren. Herausragendes Merkmal ist hierbei, dass unser Verfahren die exakte Lokalisation der sich in Kontakt befindenden Aminosäuren und Nukleotiden erlaubt. Wir zeigen, dass die Anwendung unseres Verfahrens auf Daten von Hefe und Mensch neue Interaktionspartner in bisher nicht erzielter Auflösung identifiziert.

Das zweite Verfahren findet Anwendung in metaproteomischen Studien komplexer Gemeinschaften von Mikroorganismen. In unüberschaubarer Anzahl besiedeln Bakterien, einfache Pilze oder Pflanzen die verschiedensten Lebensräume. Dabei stehen sie in einer Vielzahl von symbiotischen oder parasitären Verbindungen, die zum überwiegenden Teil der Nahrungsaufnahme dienen. Die Organismen unterscheiden sich teilweise erheblich in ihren biochemischen Fähigkeiten die es ihnen erlauben verschiedenste Substrate umzusetzen. Bemerkenswert ist insbesondere die Fähigkeit bestimmter Gruppen von Bodenbakterien mit Hilfe ihres biochemischen Repertoires Umwelttoxine zu verstoffwechseln. Unser Verfahren ermöglicht die Identifikation der beim Abbau beteiligter Organismen sowie deren Gruppierung bezüglich ihrer Funktion im Abbauprozess. Hierfür verwenden wir mit stabilen Isotopen markierte Substrate,

welche von den Organismen metabolisiert werden. Die relative Isotopenhäufigkeit in Proteinen dient als Indikator für die Umsetzung des Substrats. Diese wird mit Hilfe neuer computergestützter Methoden automatisch bestimmt und den einzelnen Organismen zugeordnet. Die Automatisierung dieses Prozesses reduziert die monatelange, manuelle Arbeit auf wenige Minuten und ermöglicht dadurch die Durchführung von Studien in bisher nicht erreichbarer Größe.

Der dritte Teil dieser Arbeit leistet einen Beitrag zur besseren Kommunikation und Verarbeitung von Ergebnissen aus Metabolom und Proteomstudien. Wir entwickeln hierzu ein tabellarisches, standardisiertes, menschenlesbares und maschinell prozessierbares Datenformat namens *mzTab* als Ergänzung zu existierenden Datenformaten. Zusätzlich stellen wir Softwarekomponenten zur Verfügung die es ermöglichen dieses Format zu verarbeiten. Wir demonstrieren an Beispielen wie *mzTab* sich nahtlos in komplexe Proteomik- und Metabolomikworkflows einbinden lässt. Ein großer Erfolg ist die seit kurzem eingeführte Unterstützung von *mzTab* durch die größten Datenrepositorien im Bereich der Proteomik. Erfreulicherweise beobachten wir bereits eine breite Akzeptanz von *mzTab* bei Entwicklern im akademischen Bereich sowie eine erste kommerzielle Verwendung. Unser neues Datenformat erleichtert Metaanalysen und macht Forschungsresultate der Proteomik und Metabolomik für Wissenschaftler aus anderen Forschungsbereichen zugänglich.

Acknowledgments

I would like to express my sincere gratitude to my advisor Prof. Oliver Kohlbacher for his continuous support, for his immense knowledge and guidance in my research and during the writing of this thesis.

I thank my fellow colleagues in Tübingen for the stimulating discussions, for the intense time we were working together on OpenMS releases, user meetings, or paper deadlines as well as the numerous BBQ sessions we have had in the last years. I would also like to thank all other OpenMS developers, especially the core developers Stephan Aiche, Chris Bielow, Julianus Pfeuffer, Hendrik Weisser, and Hannes Röst. They sacrificed a lot of their spare time on improving OpenMS. Special thanks to Luis de la Garza, Mathew Divine, and Oliver Alka for proofreading this thesis.

I thank my colleagues at the MPI in Göttingen, especially Katharina Kramer, Aleksandar Chernev, Saadia Qamar, Kundan Sharma, Uzma Zaman, and Henning Urlaub for making RNP^{xl} and RNP^{xl}Search possible. I want to thank Johannes Veit for integrating the tool and workflow into Proteome Discoverer. I thank my colleagues Florian-Alexander Herbst, Robert Starke, Nico Jehmlich, Vanessa Lünsmann, Martin Taubert, and Jana Seifert at the UFZ in Leipzig for introducing me to the fascinating world of microbial metaproteomics. Without them, development of MetaProSIP would not have been possible. Creation of standardized file formats for reporting of scientific results is one of the most important tasks. Unfortunately, the development process goes hand-in-hand with lengthy discussions, frustration, and little personal reward. I would like to express my deepest respect for members of the Proteomics Standardization Initiative for their continuous effort to foster reproducible science.

Last but most importantly, I would like to thank my family, Lisa, Theo, and the cat for their unlimited support and love.

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

Contents

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Structure of this Thesis . . . 3

2 Background **5**
2.1 Proteomics and Metabolomics . . . 5
2.2 Mass Spectrometry-based Proteomics and Metabolomics . . . 5
2.2.1 Sample Preparation . . . 5
2.2.2 Separation Techniques . . . 6
2.2.3 Mass Spectrometry . . . 8
2.2.4 Tandem Mass Spectrometry . . . 12
2.3 Computational Mass Spectrometry . . . 14
2.3.1 Peak Picking . . . 14
2.3.2 Quantification . . . 14
2.3.3 Identification . . . 20
2.3.4 The OpenMS Framework . . . 23

3 Single Amino Acid Assignment of Nucleotide-binding Sites in RNA- and DNA-binding Proteins **31**
3.1 Introduction . . . 31
3.1.1 Motivation . . . 32
3.1.2 Structure Elucidation . . . 33
3.2 Automated Cross-Link Identification . . . 35
3.2.1 Methods . . . 35
3.2.2 Results . . . 45
3.2.3 Discussion . . . 51
3.3 Automated Cross-Link Localization . . . 53
3.3.1 Methods . . . 53
3.3.2 Implementation . . . 62
3.3.3 Results . . . 65

3.3.4 Discussion 68

4 Dynamic Stable Isotope Probing of Metaproteomic Communities 71
4.1 Introduction 71
4.2 Methods 75
4.2.1 Experimental Setup 75
4.2.2 MetaProSIP Pipeline 76
4.2.3 MetaProSIP Tool 77
4.3 Results 86
4.3.1 Case Study 1: Performance of RIA and LR Detection 86
4.3.2 Case Study 2: Identification of Labeled Peptides 87
4.3.3 Case Study 3: Functional Grouping 88
4.3.4 False-Positive Rate Estimation of Labeled Peptides 92
4.4 Discussion 93
4.4.1 Comparison to other SIP Techniques 94
4.4.2 Comparison to other Computational Protein-SIP Methods 94
4.5 Outlook 95

5 Standardized Reporting of Experimental Results in Proteomic and Metabolomic Studies 97
5.1 Introduction 97
5.2 Methods 99
5.2.1 Design Rationales 100
5.2.2 Structure 101
5.2.3 Reporting Experimental Metadata 102
5.2.4 Reporting Peptide and Protein Identification Results 107
5.3 Results 111
5.3.1 Implementation in OpenMS 111
5.3.2 Statistical Downstream Analysis 112
5.3.3 Community Acceptance 112
5.4 Discussion 114
5.5 Outlook 115

6 Conclusion 117

Bibliography 121

Appendices 137

Abbreviations **137**

Appendix A Permissions and Contributions **141**

Appendix B Background **143**

Appendix C RNP^{xl} **147**

Appendix D MetaProSIP **159**

Appendix E MzTab **163**

Appendix F Curriculum Vitae **165**

List of Figures

2.1 High Performance Liquid Chromatography (HPLC) 7
2.2 Components of a mass spectrometer 9
2.3 Electrospray ionization 10
2.4 Orbitrap mass analyzer 11
2.5 Mass spectrum 12
2.6 Mass spectra of an MS run (peak map) 13
2.7 Precursor and isolation window 13
2.8 Label-free quantification 15
2.9 Theoretical isotope pattern (single carbon atom) 16
2.10 Theoretical isotope pattern (C_{50} molecule) 17
2.11 Theoretical isotope pattern of a peptide 18
2.12 Map alignment and linking 20
2.13 Database search 21
2.14 The OpenMS framework 25
2.15 TOPPView application 28
2.16 the KoNstanz Information MinEr (KNIME) analytics framework 29

3.1 Prokaryotic ribosome during translation initiation 32
3.2 Difference between cross-link identification and localization 34
3.3 Overview of the RNP^{xl} experimental workflow 35
3.4 UV-induced cross-linking 36
3.5 Enrichment of protein-RNA heteroconjugates 37
3.6 Overview of the computational RNP^{xl} workflow 38
3.7 XIC Filter 39
3.8 Additive cross-link masses 40
3.9 Precursor variant generation 41
3.10 Manual validation and visualization in TOPPView 44
3.11 Data reduction in yeast 45
3.12 Cross-linking sites and domains in RNA- and DNA-binding proteins (human) 46

3.13 Cross-linking sites and domains in RNA- and DNA-binding proteins (yeast) . . . 47
3.14 Precursor RNA adducts of yeast . . . 49
3.15 Overview of the main steps in peptide and cross-link identification and localization . . . 53
3.16 Main loop of the peptide identification engine . . . 55
3.17 Fragment adducts for uridine-containing RNA . . . 58
3.18 Oligonucleotide fragmentation adduct trees . . . 59
3.19 OpenMS workflow for cross-Link identification and localization . . . 64
3.20 Proteome Discoverer integration . . . 65
3.21 Identification performance . . . 66
3.22 Localization error . . . 68

4.1 Overview on MetaProSIP workflow . . . 76
4.2 Theoretical isotope pattern of a dynamically labeled peptide . . . 78
4.3 Feature-based isotope pattern extraction . . . 79
4.4 Signal decomposition and reconstruction . . . 82
4.5 Isotopic pattern of an unlabeled and labeled peptide . . . 83
4.6 Quality control items . . . 85
4.7 Performance of RIA and LR detection . . . 87
4.8 Time-course experiment with labeled substrate . . . 89
4.9 Heatmap of relative isotopic abundances (RIA) groups and annotation with phylogenetic information . . . 90
4.10 Elemental flux network reconstruction . . . 91

5.1 Standard scientific workflow. Data acquisition, data processing, analysis, publication, and data deposition . . . 98
5.2 Structure of an mzTab file . . . 102
5.3 Relation of study variable, run, sample, and assay in an experimental design . . . 104
5.4 MzTab workflow using OpenMS in KNIME (Proteomics) . . . 112
5.5 Small molecule quantification workflow in KNIME . . . 113

B.1 MS/MS spectrum . . . 143
B.2 Fragment ion nomenclature . . . 143

C.1 UV-induced cross-link reaction . . . 147
C.2 Fragment spectrum of peptide and cross-link . . . 148

C.4 RNP^{xl} OpenMS workflow 149
C.5 Structural interpretation 155

D.1 MetaProSIP OpenMS workflow for single run analysis 159
D.2 MetaProSIP OpenMS workflow for time series analysis 161

E.1 MzTab workflow using OpenMS and R in KNIME (Metabolomics) . . . 163

List of Tables

3.1 Sample output of RNP^{xl}Search 62

5.1 Draft for extended mzTab structure 115

B.1 Natural abundance of H, C, N, O, and S isotopes 144

C.1 Cross-links to mono- and dinucleotides 147
C.2 Cross-links to a uracil-containing RNA sequence 150
C.3 Cross-links with 4SU nucleotide analog substituted at a specific site. . . 150
C.4 Cross-links with isotopically labeled adenosine 151
C.5 Cross-linked proteins in human 152
C.6 Cross-linked proteins in yeast (standard uridine) 153
C.7 Cross-linked proteins in yeast (4-thiouridine) 154
C.8 RNP^{xl}Search tool parameters. 156
C.9 Supported enzymes and cutting rules 157

D.1 Basic MetaProSIP OpenMS workflow parameters 160
D.2 MetaProSIP OpenMS workflow parameters 162

Chapter 1

Introduction

1.1 Motivation

For more than two decades tremendous efforts have been made to determine the DNA sequences of organisms. Methodological advances in sequencing technologies shifted the laborious analysis from single nucleotides of single genes to high-throughput analysis of large stretches of the genome. International research efforts within "The Human Genome Project" led to the publication of a working draft[1,2] in February 2001. Followed by complete sequencing and assembling of the major part of the human genome in April 2003[3].

In many areas of life sciences[4], the unprecedented amount of genome data induced a shift from hypothesis-driven to data-driven science. Using data exploration to generate hypotheses has already resulted in answers to many unresolved scientific questions in a less biased fashion[5–7]. Analyses of genetic variations within and among populations have fundamentally expanded our understanding of gene functions. After early successes of associating, for example, diseases to specific changes in the genome, a certain disillusionment took place in the scientific community. More and more researchers came to realize the limitations of genomic studies: while the genome lays out the plan for development and functions of an organism, it is linear, static, and only indirectly involved in cellular metabolism. In contrast, its main products, RNA, proteins, and other regulatory elements act and interact in networks of astonishing complexity making sequence-based predictions of their role very difficult. Many scientists, thus, proclaimed the post-genomics era with the primary challenge of tackling the difficult task of assigning protein functions[8].

Protein functions are as diverse as protein structures and play crucial roles in all organisms. Proteins provide structural elements, transport molecules, mediate signal transduction, regulate cellular processes, and catalyze biochemical reactions[9]. Not surprisingly, proteins are particularly interesting research targets in the life sciences. The field of proteomics is the bioanalytical research branch aiming at analyzing all proteins in an organism. Compared to the genome or transcriptome, the proteome is

considered significantly closer to the phenotype of a disease or biological trait. It has recently grown in importance nurtured by the availability of large protein sequence databases derived from the genome projects and recent technological advancements[10]. High-resolution tandem mass spectrometry combined with liquid chromatography is the analytical method of choice for high-throughput protein identification and quantification today. In May 2014, more than a decade after the human genome project was completed, the first drafts of the human proteome had been published[11,12] allowing unprecedented insight into quantitative protein dynamics on the level of cells, tissues, and organs. As it turns out, this novel and detailed quantitative view is still insufficient to answer many fundamental questions, and further information needs to be integrated to achieve a comprehensive understanding of a biological system.

One important aspect typically missing in quantitative proteomics studies is the detailed characterization of interactions between components of a biological system. This set of all interactions between molecules in a biological system is often termed *interactome*. Currently, the study of interactions in a single cell or organism is dominated by the research on protein-protein interactions[13–15]. One important type of interaction, so far not accessible by high-throughput methods, is occurring between main classes of biomolecules: proteins that are in contact with RNA or DNA. More than two thousand distinct proteins[i] are expected to bind (ribo-)nucleotides[16]. It is not surprising, considering the large number of nucleotide-binding proteins, that several human diseases have been associated with RNA- and DNA-binding proteins[17,18]. Characterizing the amino acids and nucleotides in contact is likely to yield novel insights into observed phenotypes, structure, function, and dynamics of these complexes, but suitable methods have been missing. In this thesis, we present a novel method and computational tools that allow investigating interactions between RNA/DNA and proteins.

Organisms exhibit a complex interplay with other organisms and their environment. Studying an organism in isolation results in only a partial view that may not be sufficient to answer a specific question. Similar to metagenomics, which expands the study of genomes from a single organism to the study of multiple organisms in parallel, the interplay of multiple organisms is investigated in the nascent fields of metaproteomics and meta-metabolomics. The degradation of complex substrates and symbiotic processes are subjects of several recent studies. These studies help us to understand bioremediation processes or the role of the gut microbiome. In this thesis, we present a novel method and computational tools to investigate substrate metabolism.

[i] 2,234 as estimated in neXtProt, accessed 05.01.2015

We demonstrate that it is possible to determine distinct functional groups of organisms in complex microbial communities that differ in their ability to break down substrates.

Nowadays, many scientific questions cannot be answered using a single technology. For complex diseases or phenotypes, data-driven hypothesis generation is often hampered by incomplete data obtained from single high-throughput methods. Combining a variety of quantitative information from different omics levels gives a more detailed picture of the dynamics of a biological system and, ideally, lead to new causal explanations. The integration of data sources from disciplines like genomics, mass spectrometry-based metabolomics and proteomics, is currently an active research area. Reporting and exchanging results from single or combined high-throughput experiments over scientific fields requires data formats that allow for data integration and reproducible science. In this thesis, we present a novel standardized data format for reporting of mass spectrometry-based proteomic and metabolomic results.

1.2 Structure of this Thesis

Following the introduction, Chapter 2 covers the relevant technical and biological background. In Chapter 3-5 of this thesis, we describe three contributions to the field of computational mass spectrometry.

1. We develop a novel computational method, algorithms, and workflows to study RNA- and DNA-protein interactions at the level of single amino acids. We apply these methods to study whole-cell lysates of different organisms, including human and yeast. In these studies, we identify and characterize novel RNA-binding proteins. Compared to existing high-throughput methods for the analysis of Ribonucleoproteins, our method is able to pinpoint protein-nucleotide interactions with amino acid resolution and is applicable to single complexes as well as whole cell lysates. Our method, therefore, is a significant contribution to the field of protein-RNA/DNA interaction studies as well as to structural proteomics of nucleotide-binding proteins.

2. We develop a novel computational method, algorithms, and workflows to investigate the role of microorganisms that partake in substrate metabolism. Metagenomic approaches determine organisms present in complex microbial communities. In contrast, the presented computational metaproteomics approach allows answering questions about the function and biochemical repertoire of these organisms. Compared to existing, single software, or script-based solutions our approach allows analyzing high-throughput data from a variety of experimental

setups. We detect groups of organisms with similar biochemical activities and reduce the time for manual analysis from months to few minutes. Our computational method is a significant contribution to the field of functional metaproteomics with possible application to the study of nutrient flow, bioremediation, and biodegradation processes in microbiomes.

3. We develop a human-readable, computer-consumable data format for the reporting of proteomic and metabolomic results to a wider audience. We specify a tabular file format, together with members of the Proteomic Standard Initiative (PSI), and implement tools for reading and writing the file format in OpenMS. This allows to report mass spectrometry-based proteomics and metabolomics results in parallel.

This thesis finishes with a conclusion in Chapter 6.

Chapter 2

Background

2.1 Proteomics and Metabolomics

Proteomics and metabolomics are interdisciplinary research fields that study structure, function, and interaction of proteins and metabolites. They employ large-scale experimental techniques that allow acquiring data at the level of cellular systems to whole organisms. The main analytical method to identify, characterize or quantify proteins and metabolites is mass spectrometry (MS) combined with chromatographic separation.

2.2 Mass Spectrometry-based Proteomics and Metabolomics

In mass spectrometry-based proteomics and metabolomics, biological samples are extracted, prepared, and separated to reduce sample complexity. The separated analytes are ionized and measured in the mass spectrometer. Mass and abundance of ions are stored in mass spectra and used to identify and quantify the analytes in the sample using computational methods. The quantity and identity of analytes can then be used, for instance, in biomarker discovery, medical diagnostics, or basic research.

The following sections give a brief overview of the general experimental techniques employed in the context of this thesis. Metabolomics plays only a secondary role. Thus, no detailed introduction to this research field is given. Instead, the interested reader may refer to Dettmer et al.[19], Wang et al.[20], Patti et al.[21], or Kaddurah-Daouk et al.[22] for review articles. In the following sections, the term *separation techniques* exclusively refers to analytical methods employed before mass spectrometric measurement. Mass spectrometry itself is, of course, also a (mass) separation technique.

2.2.1 Sample Preparation

MS sample preparation processes biological samples to make them amenable to mass spectrometry analysis. Proper sample preparation is, therefore, a prerequisite of any mass spectrometry workflow. Not surprisingly, errors in sample preparation critically

affect reproducibility, accuracy, and sensitivity of the whole experimental setup. Because of the complexity of the proteome and the variety of different experimental setups, no standard sample preparation procedure exists. Nevertheless, exemplary steps shared by sample processing protocols may include:

- Cell lysis using reagents that break or dissolve intact cells into its components.
- Inhibition of active enzymes (e.g., endoproteases).
- Centrifugation and depletion of unwanted cell components or proteins.
- Cross-linking, the introduction of chemical bonds between molecules to fixate them and study their interaction.
- Enrichment steps that increase the relative abundance of molecules of interest. Usually by filtering of other molecules during lysate preparation.
- Desalting, the removal of ions that otherwise would interfere with mass spectrometry analysis.

The lysate can optionally be further processed in-gel or in-solution. One or two-dimensional gel electrophoresis yields bands of proteins that can be cut out and analyzed. Proteins in-gel or in-solution are typically reduced and alkylated to break disulfide bonds between cysteines irreversibly. As a result, proteins permanently lose their tertiary structure. The linearized proteins then expose enzyme cleavage sites and can be easily digested *in situ*. Most commonly, the enzyme trypsin is used, which cuts after the amino acids arginine and lysine. Because arginine and lysine have basic side chains, they may contribute a proton charge to a peptide.

Many different techniques exist which label analytes either *in vivo* or *in vitro*. Labeling reagents that introduce heavier isotopes in living organisms or cell cultures are widely used. Here the organism or cell culture consumes and metabolizes the reagent building novel proteins that incorporate the heavy isotopes. Chemical labeling is, in contrast, applied *in vitro* during sample preparation but may also involve isotopically labeled reagents.

2.2.2 Separation Techniques

Direct injection of samples into a mass spectrometer is usually not feasible for complex samples. Separation techniques are applied to reduce sample complexity to the extent necessary. In mass spectrometry-based experiments, one limiting factor is the number of analytes that can be measured in parallel. Separation of analytes in time is, thus, the

preferred mode of segregation. Most common techniques employed are fractionation and Liquid Chromatography (LC), which both exploit differences in the physicochemical properties of analytes.

Sample Fractionation

Fractionation techniques separate analytes into fractions of similar physicochemical properties, for instance, the isoelectric point. Each fraction is then separately subjected to further separation (e.g., LC) or analyzed directly.

Liquid Chromatography

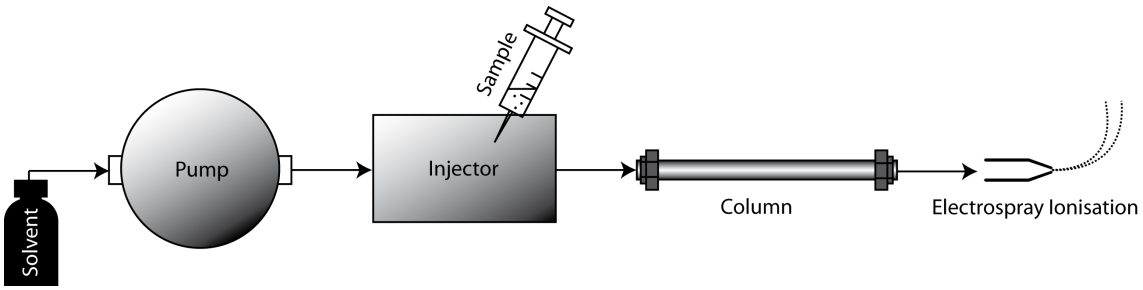


Figure 2.1: HPLC. The sample material is injected. A High-pressure pump system pumps sample analytes in a solvent (*mobile phase*) through a column with chromatographic packing material (*stationary phase*). Eluting analytes can be ionized using electrospray ionization and measured using mass spectrometry.

In mass spectrometry-based proteomics, (high-pressure) liquid chromatographic separation techniques (HPLC) are the methods of choice to achieve a high degree of separation (see Figure 2.1 for a schematic overview of an HPLC system). In HPLC, peptides are separated on a column. Solved in a pressurized liquid (*mobile phase*) they are pumped through a solid adsorbent material (*stationary phase*) packet into a capillary column. Physicochemical properties of each peptide determine how strongly it interacts with the stationary phase. The most commonly HPLC technique in proteomics uses reversed-phase chromatography (RPC) columns. RPC employs a hydrophobic stationary phase like octadecyl (C18), a nonpolar carbon chain bonded to a silica base, and a polar mobile phase. Polar molecules interact weakly with the stationary phase and elute earlier, while non-polar molecules are retained. Interaction can be further modulated by changing the gradient of solvent concentration in the mobile phase over time. Elution times in LC are inherently prone to variation, for example, due to fluctuations in the flow rate of the mobile phase or change of column. Retention time shifts between runs may be compensated using computational chromatographic

retention time alignment methods. In the LC-MS setup, the column is directly coupled to the ion source of the mass spectrometer.

2.2.3 Mass Spectrometry

MS is an analytical technique used to determine the mass of molecules. In order to achieve highly accurate and sensitive mass measurements at the atomic scale, mass spectrometers manipulate charged particles using magnetic and electrostatic fields. The fact that charged particles are much easier to manipulate was recognized several centuries ago and utilized by the early pioneers of mass spectrometry. Wilhelm Wien (1864-1928) was the first who applied magnetic and electrostatic fields to separate charged particles (1899). Sir Joseph J. Thomson (1856-1940) improved on these initial designs. In the 1950s/1960s Hans Dehmelt and Wolfgang Paul developed the ion trap, the component that balances magnetic and electrostatic forces in order to hold ions temporarily and allow further manipulation. Nowadays, mass spectrometry is used widely outside of basic research. For instance, airport security, drug control in sport, quality control of waste water, and toxicity screening of food products are just a few examples that show the wide range of practical applications.

In a typical mass spectrometer, three principal components can be identified (Figure 2.2):

- Ion Source: The component that produces ions from the analyte.
- Analyzer sorts and filters ions according to their mass and charge.
- Detector counts and records the ions.

Ion Source

A mass spectrometer only handles ions. Thus, charge needs first be transferred to uncharged particles. The component responsible for the ionization is the ion source. Different types of ion sources and ionization techniques exist with electrospray ionization being currently the most widely used ionization technique for mass spectrometry-based proteomics. In 2002, the Nobel Prize in chemistry was rewarded to John Bennet Fenn for its development. LC-separated peptides or small molecules elute from the column to pass a metal needle that is held at a high electric potential compared to the entrance of the mass spectrometer (Figure 2.3). The analytes in-solution form a cone shaped drop (*Taylor cone*) at the tip of the needle. The Taylor cone transitions into a jet of highly protonated droplets that get further dispersed into a fine aerosol. Droplets are

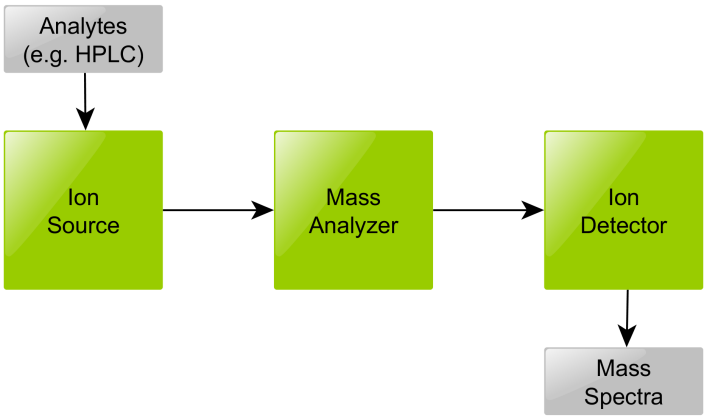


Figure 2.2: Components of a mass spectrometer. Eluting analytes are ionized in the ion source which makes them applicable to mass separation in the mass analyzer. Ion detectors record ions and store them as peaks in mass spectra.

accelerated through a heated, depressurized region resulting in solvent evaporation. With decreasing size of the droplets, the density of charged particles at the surface of the droplets is further increased. If the radius gets smaller than the *Rayleigh limit*, the electric repulsion between like charges is high enough to rapidly break the droplet into even smaller fragments (*Coulomb explosion*). At this final phase of electrospray ionization, an ion has lost all associated solvent molecules. The resulting ions potentially carry multiple charges and enter the high vacuum parts of the mass spectrometer as a continuous stream in the gas phase. Electrospray ionisation (ESI) is a so-called *soft ionization* technique as it produces charged ions without analyte degradation.

Mass Analyzer

Today, the most commonly used mass analyzer in proteomics are time-of-flight (TOF) mass analyzers, quadrupole mass filters, and orbitrap analyzers. In TOF mass analyzers, the ions are accelerated in an electric field. The flight time of an ion allows calculating the velocity which in turn is used to calculate the mass-to-charge ratio (m/z). Varying the electric field allows filtering certain mass-to-charge ratios before they enter the detector.

In quadrupole mass filters, ions pass through an oscillating electric field created by four parallel rods. For a particular voltage, only ions in a certain mass-to-charge range will reach the detector.

The orbitrap[23] is an ion trap mass analyzer (and detector) that traps ions in orbital motion between a barrel-like outer electrode and a spindle-like central electrode

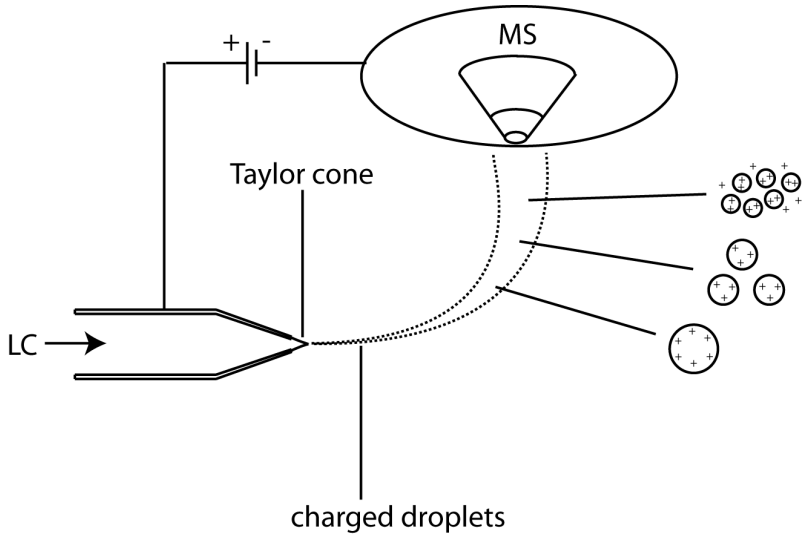


Figure 2.3: Electrospray ionization. A stream of charged droplets emerges from the cone shaped drop (*taylor cone*) at the tip of the electrospray needle. Droplets get accelerated in an electric field between the tip and the mass spectrometer. Droplets shrink due to solvent evaporation in a heated and depressurized region. The resulting, nearly solvent free ions enter the mass spectrometer.

allowing for prolonged mass measurement (Figure 2.4). As a result of the prolonged mass measurements, a high mass resolution can be achieved. Upstream to the orbitrap, a curved linear trap (C-trap) is used to decelerate and collect ions. A package of ions with reduced kinetic energies is then channeled into the orbitrap analyzer. The axial oscillation frequencies in the orbitrap are recorded via an image current. The Fourier transform of this raw signal is used to calculate the mass-to-charge ratios of ions.

Detector

The last component of the mass spectrometer is the detector. It performs the actual quantification of ions that passed through the mass analyzer. Ion intensities (a value that relates to its abundance) and the mass-to-charge ratio are recorded in a mass spectrum. The simplest type of detector is the Faraday cup. Upon collision of an ion with the detector plate, electrons get removed, and the compensating current is detected. More advanced detectors like the electron multiplier amplify the weak current of the initial collision by several orders of magnitude. An avalanche-like multiplication of emitted electrons between dynodes results in a higher current and, therefore, higher sensitivity. In both detector types, the ion loses its charge upon collision. These techniques are, thus, referred to as destructive detection techniques. In contrast, the orbitrap analyzer performs non-destructive m/z detection. Here, the

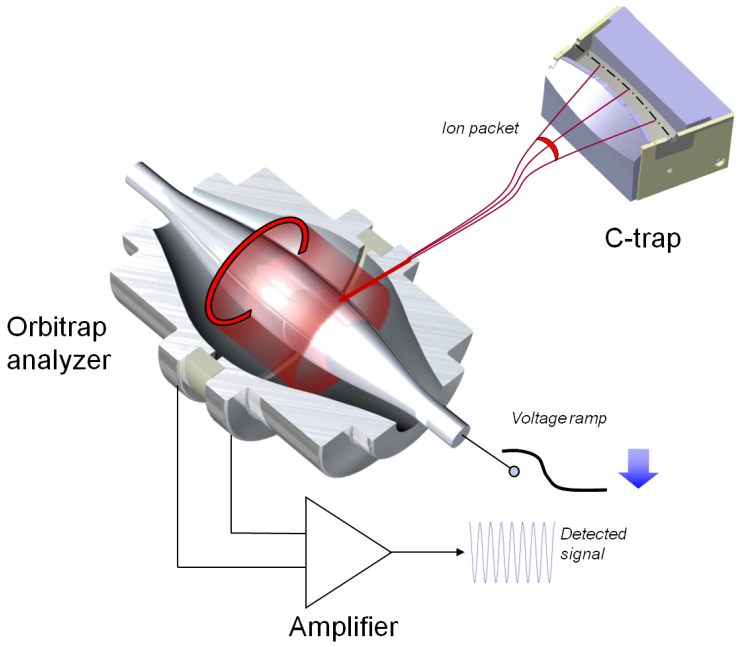


Figure 2.4: Orbitrap mass analyzer and detector. Ion packages from the C-trap are streamed into the orbitrap. The oscillating ions are modulated using a voltage ramp, and the resulting signal is read out using an amplifier circuit. Artwork by Thermo Fisher Scientific, (CC-BY-SA 3.0).

principle of electrostatic induction caused by charged particles in close proximity is used. Oscillation frequencies are then determined from this image current which in turn allows determining the mass-to-charge ratios of the ions. Because ions do not collide but oscillate in an ion trap, their mass-to-charge ratio can be measured over an extended time scale. Taken together, a higher mass accuracy is achieved at the cost of longer measurements time for each spectrum. In this thesis, we exclusively used orbitrap mass analyzers to obtain high-resolution mass spectra. A mass spectrum consists of a set of m/z, intensity pairs - so-called *mass peaks* (see Figure 2.5 for a visualization of a mass spectrum). To reconstruct the actual mass (m) from the mass-to-charge ratio (m/z), the charge number (z) must be determined. Usually, this is done by computationally detecting isotopic peaks from different isotopic composition of the molecule. m/z differences between isotopic peaks match the m/z introduced by the additional neutron. These characteristic m/z differences, thus, allow deriving mass and charge of an ion.

Complex samples typically give rise to several thousand spectra and are stored as raw data files. Figure 2.6 visualizes such set of mass spectra as a two-dimensional map of mass peaks. In computational mass spectrometry, the list of spectra is called

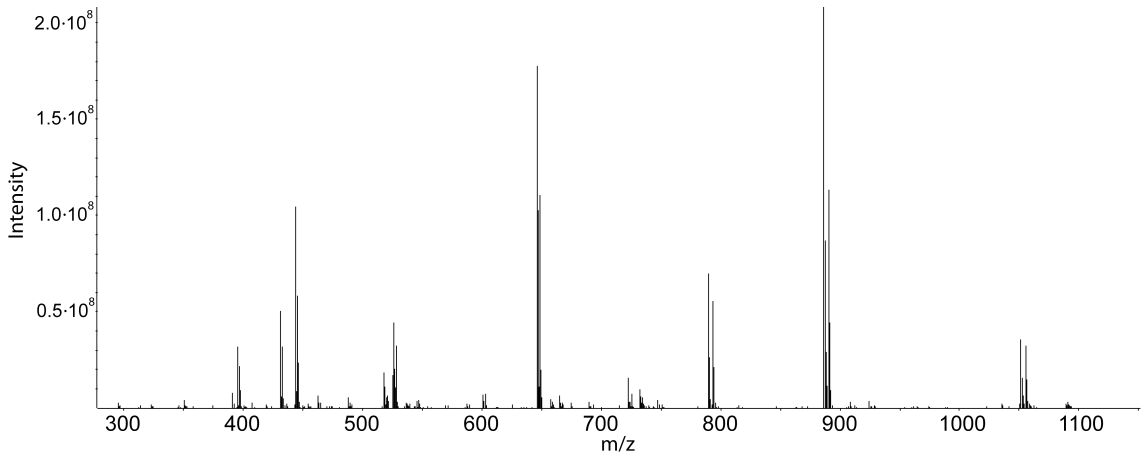


Figure 2.5: Mass spectrum as recorded by a mass spectrometer. Peak intensities are related to the ion abundances (e.g., the detected ion current).

a *peak map*. For a formal definition of basic terms relating to mass spectrometry see Appendix B.

2.2.4 Tandem Mass Spectrometry

A tandem mass spectrometer is capable of performing multiple rounds of mass measurements. In the first round, the mass spectrometer records a survey scan over the full m/z range. A subset of *precursor ions* is automatically selected from the survey scan. For each precursor, the instrument opens up a small m/z window, the so-called *precursor isolation window* (Figure 2.7), to collect ions for fragmentation in a collision cell. After fragmentation, *fragment ions* are measured in the second round of mass measurements. The resulting spectrum is called a tandem mass spectrum (MS/MS) (Appendix Figure B.1). Most widely used fragmentation techniques are collision-induced dissociation (CID), higher-energy collision dissociation (HCD)[24] as well as electron-transfer dissociation (ETD)[25]. Each method comes with different fragmentation behavior and ion types (see Appendix Figure B.2 for details).

The majority of fragmentations occur at the backbone of the peptide. Ionized fragments, corresponding to prefixes or suffixes of the parent peptide differ in length and form the *sequence ions* (also: *mass ladders* or *ion series*). In addition to the sequence ions, double backbone cleavage can give rise to internal cleavage ions. If the internal fragment contains only a single residue, it is called *immonium ion* and is referred to by the single letter code of the amino acid. Ideally, the information stored in a tandem spectrum allows identifying the peptide unambiguously[26].

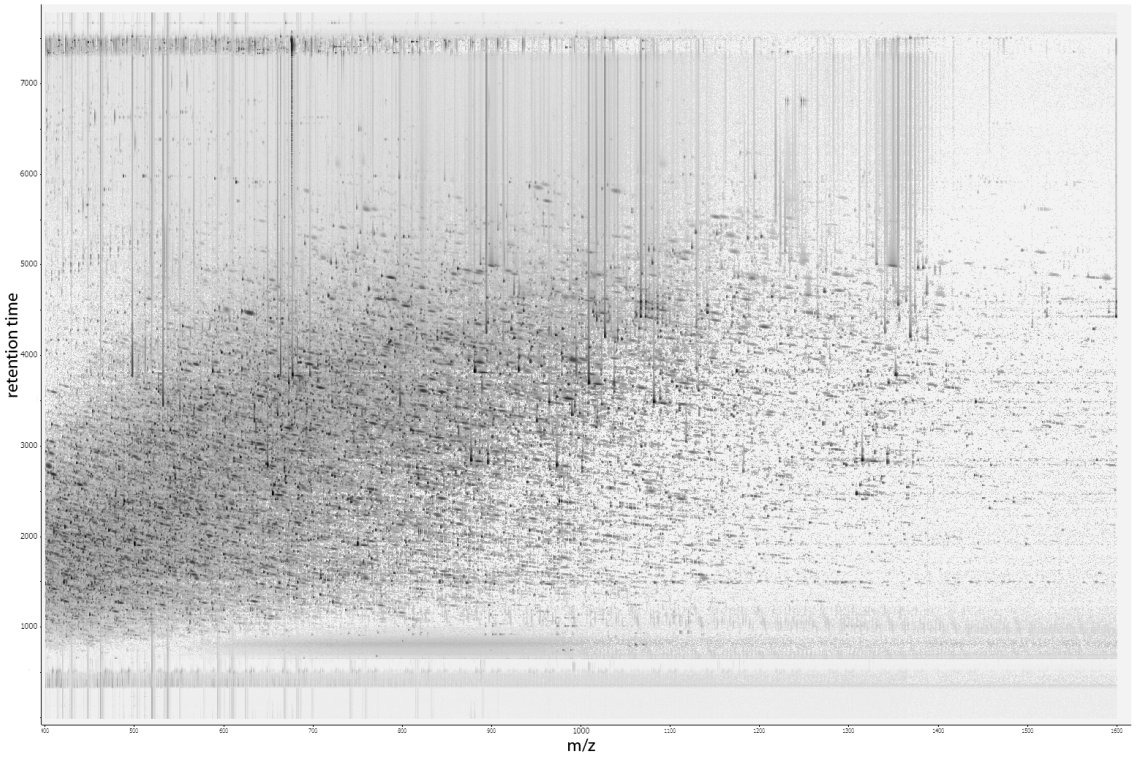


Figure 2.6: Peak Map of a complex sample displaying several thousand mass spectra, stacked in retention time. Several thousand peaks, corresponding to different analytes, can be visually spotted. Darker shades indicate higher ion counts.

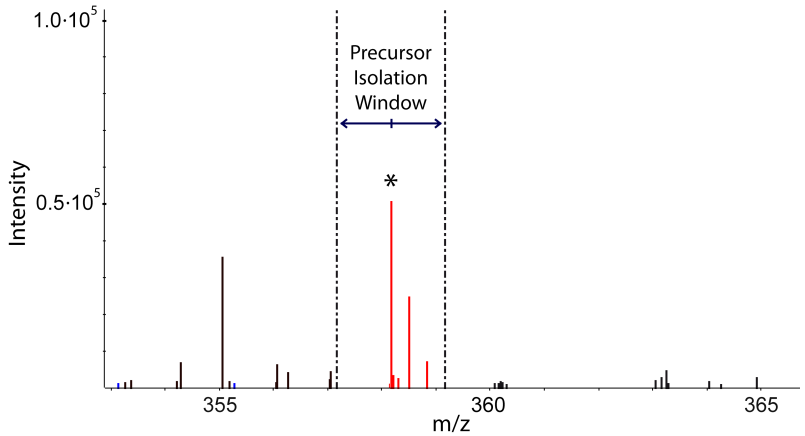


Figure 2.7: Precursor (*) and isolation window (shaded) selected for fragmentation in the collision cell.

2.3 Computational Mass Spectrometry

Computational mass spectrometry has grown to a large scientific field with a plethora of methods and tools. In this section, only methods most relevant to the work described in this thesis are introduced.

2.3.1 Peak Picking

Detectors in mass spectrometers record a continuous signal of mass-to-charge ratios, a so-called *profile spectrum*. Because of the limited resolution, the recorded profile peak is spread out in the m/z dimension and usually has a Gaussian shape. This profile peak is converted to a single mass-to-charge ratio and abundance pair by integration of the profile peak abundances. The computational methods used to convert spectra recorded in profile mode to single m/z peaks are called *peak picking* (also peak centroiding) algorithms. In the projects of this thesis, we regularly used the Open Mass Spectrometry (OpenMS) peak-picking algorithms[27].

2.3.2 Quantification

Mass spectrometry-based quantitative proteomics aims at quantifying the whole set of proteins in a biological sample. Probing protein expression between experimental conditions reveals many details on the function and dynamics of a biological system. Protein quantification is a key task regularly performed to study a wide range of scientific questions.

Existing techniques can be loosely categorized in labeled and label-free approaches. Label-free quantification is probably the most direct way of determining quantities of analytes from several biological samples.[28] Label-free quantification algorithms detect and integrate chromatographic intensities of a peptide. Quantification across several MS runs is obtained by determining and *linking* of corresponding peptide signals, so-called *features*, between runs (Figure 2.8). While label-free quantification scales to a large number of experiments, it heavily relies on correct linking of corresponding peptides. Chromatographic retention time alignment algorithms compensate for differences in chromatographic elution and reduce mislinked peptides across maps. Labeling techniques circumvent, to some extent, the problem of linking corresponding peptides as they allow measuring more than one experimental condition in a single MS run. They introduce isotopic labels that can be differentiated by the mass spectrometer. The widely used *metabolic labeling* technique stable isotope labeling by amino acids in cell culture (SILAC)[29] uses *in vivo* incorporation of ^{13}C or ^{15}N labeled amino acids. SILAC is most

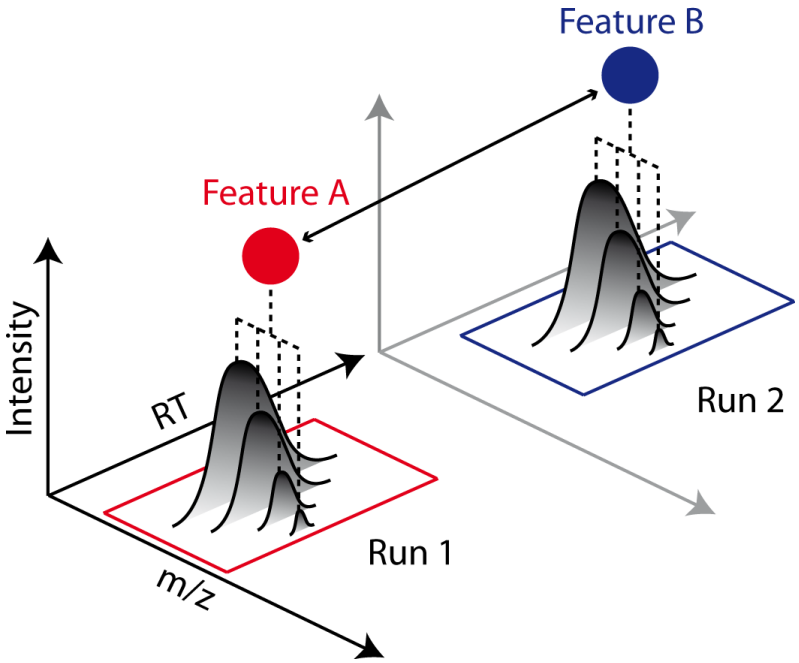


Figure 2.8: Label-free quantification. Isotopic intensities of single peptide species and charge state (*features*) are detected. Chromatographic elution profiles are integrated to yield a single feature intensity. Features A and B correspond to the same analyte. The linking of features between runs (indicated by an arrow) allows comparing feature intensities.

frequently used to measure peptides from two conditions - often referred to as the light and heavy channel - in a single MS run. So-called *dynamic metabolic labeling* techniques, like stable isotope probing of proteins (protein-SIP)[30], quantify to what degree an isotopic label has been incorporated. Isobaric tags for relative and absolute quantitation (iTRAQ)[31] and Tandem Mass Tag (TMT)[32] are *in vitro chemical labeling* techniques that attach isobaric tags to peptides from different samples. Fragment ions produced by a reporter group allows to differentiate and relatively quantify several conditions per MS run. These quantification techniques yield relative quantities for an analyte and can be used to calculate *fold changes* between conditions. Absolute quantification values (e.g., how many micromoles of a peptide is contained in a sample) are inherently more complex to obtain. One reason is that different peptides have different physiochemical properties and ionization efficiency. To obtain absolute quantities isotope-labeled standards of known concentrations can be added during sample processing. The relative quantity of a peptide can be matched to the calibration curve of its isotope-labeled version and converted into an absolute quantity[33,34].

Isotope Patterns

An isotope pattern is defined as the set of peaks related to ions with the same chemical formula but containing different isotopes[35]. In Chapter 4, the detection and analysis of isotope patterns play a central role. Hence, a short introduction to theoretical isotope patterns is given. Different isotopes of carbon, hydrogen, nitrogen, oxygen, and sulfur occur in nature. Terrestrial isotope abundances (*natural abundances*) deviate only slightly and can, for our applications, be treated as a constant. Appendix Table B.1 lists mass and natural abundance of common isotopes. For the sake of clarity, we consider only single charged ions and neglect any mass introduced by the charge. We also consider only the two most abundant isotopes of an element. Consider a single carbon atom: two major isotopes occur in nature (^{12}C or the heavier ^{13}C isotope). Because the relative abundance for ^{12}C is 98.93% and for ^{13}C is 1.07% we expect it to be a carbon atom with a mass of 12 u with a probability of 98.93% and a carbon atom with a mass of 13.003355 u with a probability of 1.07%. The mass and probability pair constitute a theoretical isotope peak, and the set of all theoretical isotope peaks constitute a theoretical isotope pattern (Figure 2.9).

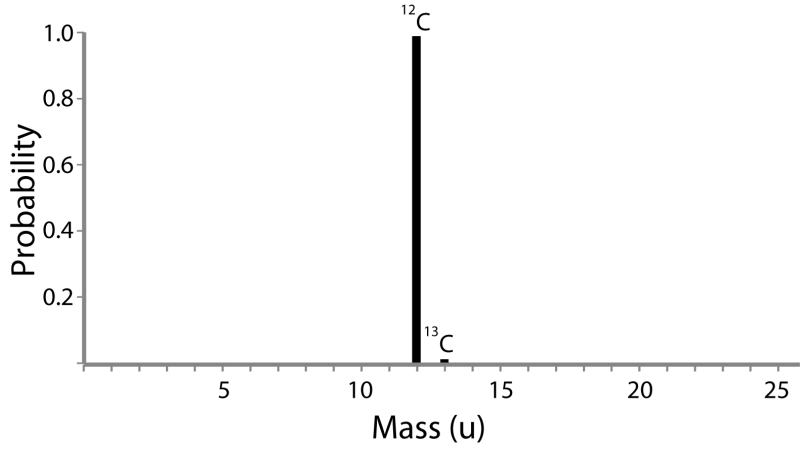


Figure 2.9: Theoretical isotope pattern of a single carbon atom. Peak intensities (probabilities) correspond to the natural abundances.

The isotope pattern consists of two peaks corresponding to the two isotopic compositions $^{12}C_1$ and $^{13}C_1$. The mass difference between neighboring isotopic peaks in an isotopic pattern is approximately[i] the neutron mass difference.

For a molecule consisting of N carbon atoms, the probabilities of the theoretical isotope pattern can be calculated using the binomial distribution[ii]. The probability of

[i] some relativistic mass defect occur due to a changed binding energy in the nucleus

[ii] For elements with more than two isotopes (e.g., sulfur) the corresponding multinomial distribution needs to be considered in Equations 2.1 - 2.2

observing the molecule with n ^{13}C and $N-n$ ^{12}C isotopes is:

$$P_p(n|N) = \binom{N}{n} p^n (1-p)^{N-n}, \tag{2.1}$$

where p corresponds to the relative abundance of ^{13}C. Note that the isotopic pattern is composed of $N+1$ peaks - one for each of the $N+1$ isotopic compositions $^{12}C_N$, $^{12}C_{N-1}{}^{13}C_1, ..., {}^{13}C_N$. Molecules that only differ in the isotopic composition are called isotopologues. The theoretical isotope pattern of a C_{50} molecule is shown in Figure 2.10.

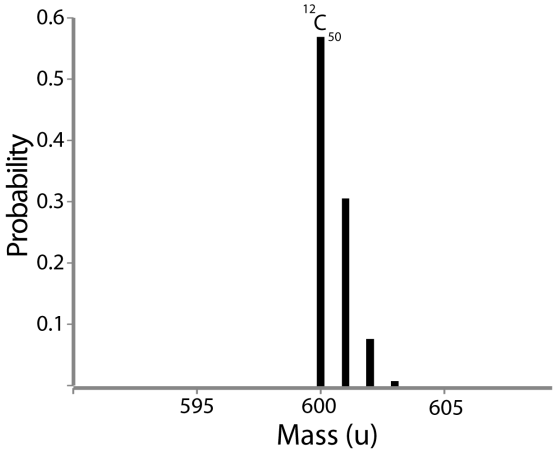


Figure 2.10: Theoretical isotope pattern of a C_{50} molecule. Note that most of the 51 isotopic peaks have a low probability of occurrence.

Larger molecules, like peptides, are composed of atoms from multiple elements. The peptide "TESTPEPTIDE" is an amino acids chain of length 11 and elemental composition $C_{50}H_{79}N_{11}O_{24}$. The number of isotopic compositions and peaks in the theoretical spectrum is thus $(50+1)\cdot(79+1)\cdot(11+1)\cdot(24+1) = 1,224,000$. The exact calculation is usually too computationally expensive for common applications in computational mass spectrometry and, therefore, rarely done in practice. In fact, most theoretical peaks have a very low probability of detecting even one ion in a mass spectrometer. Also, most peaks are so close in mass to their neighboring peak, that even high-resolution mass spectrometers are not able to resolve them into distinct peaks. Instead of calculating all theoretical peaks, approximation algorithms are commonly used to derive probabilities for nominal mass ranges. In this thesis, we exclusively used an OpenMS implementation that employs a convolution-based method to calculate probabilities for nominal masses.

In Chapter 4 we artificially increase the relative abundance of an isotope in peptides using an isotopic labeling technique. It is, thus, relevant how deviation from the natural abundance influences the shape of isotope patterns. The expected value of the binomial

distribution from Equation 2.1 is

$$E[P_p(n|N)] = np.$$

Artificially increasing the relative abundance of heavier isotopes in a peptide, thus, increases its average mass (Figure 4.2).

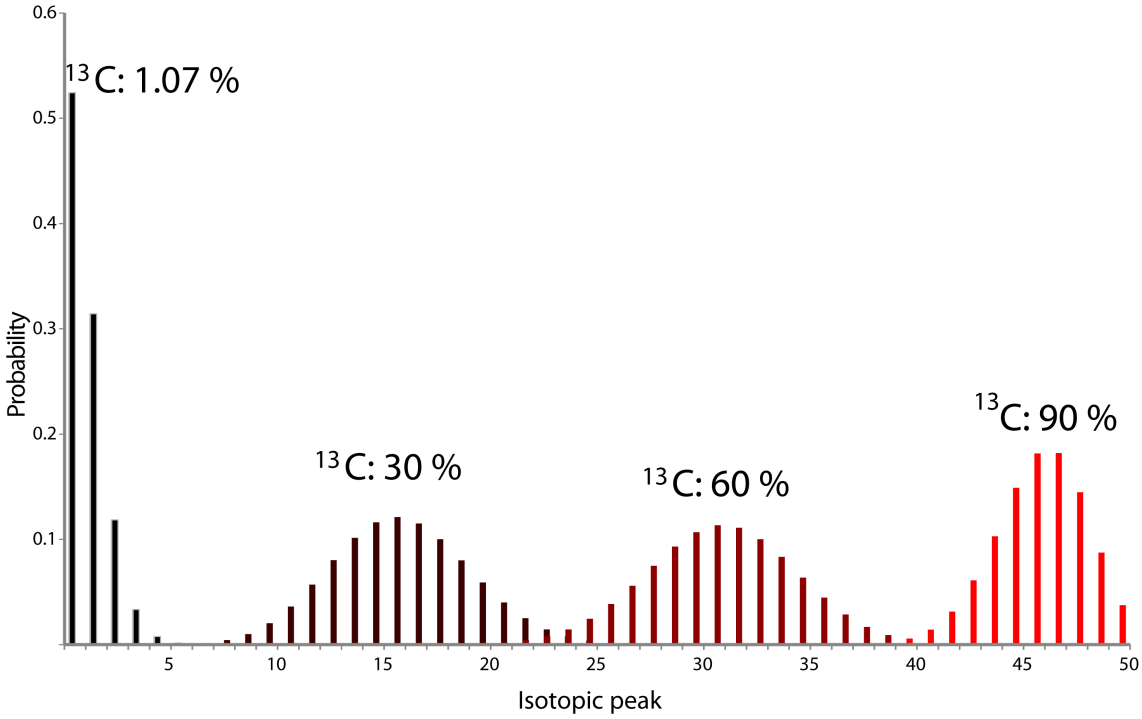


Figure 2.11: Theoretical isotope pattern calculated for a peptide with amino acid sequence *TESTPEPTIDE* and elemental composition $C_{50}H_{79}N_{11}O_{24}$ with varying RIAs.

In addition, the shape of the binomial distribution changes, as can be derived from the variance of the distribution:

$$Var[P_p(n|N)] = np(1-p) \tag{2.2}$$

If p is artificially increased, the distribution gets broader up to a maximum at $p = 0.5$, after which it decreases again (see Equation B.1 for details).

Another quantity relevant in the context of isotopic labeling techniques is the *labeling ratio*. It is defined as the proportion of labeled peptide (or protein) to total peptide (or protein) abundance.

Feature Detection

Isotopic peaks of an analyte coelute and form chromatographic mass traces. Feature detection algorithms aim to detect mass traces of a peptide charge variant and assembles them to a so-called *feature*. A feature typically stores a reduced set of data instead of all peaks and mass traces. For example, the monoisotopic m/z, the retention time of the chromatographic apex, an intensity value, the charge, and a quality score are sufficient to perform a wide range of subsequent analyses. Different ways of calculating a single intensity value from the mass trace intensities exist. Typically, the intensity is calculated by integrating chromatographic peak areas. The quality of a feature candidate is often assessed according to how well it resembles the isotope pattern of a peptide. If the peptide corresponding to a feature candidate is known, the elemental composition can be calculated from the peptide sequence. The theoretical isotopic pattern derived from the elemental composition can then be compared to the observed intensities. Depending on the feature detection approach, no sequence information might be available at that point of analysis. In that case, the feature mass can be used to calculate an approximate isotopic pattern using the *averagine*[36] model. Fundamental to this model is a hypothetical amino acid a of mass m_a that has an average elemental composition derived from large protein databases. A peptide of mass m_p is then expected to be composed of m_p/m_a average amino acids. The elemental composition of the *averagine peptide* is then obtained by multiplying the elemental composition of the average amino acid by its expected occurrence in the peptide: m_p/m_a. The approximated elemental composition is used to calculate the theoretical isotope pattern. Feature candidates that pass a quality threshold (e.g., a minimum correlation to the theoretical isotope pattern) are accepted as real features. We call the set of all features detected in a *peak map* a *feature map*.

Chromatographic Retention Time Alignment and Feature Linking

Variation in chromatographic retention times of analytes measured in different mass spectrometry runs hamper comparability and reproducibility of scientific experiments and clinical studies. Chromatographic retention time alignment algorithms typically try to find a global transformation of retention times between runs, so that same analytes share similar retention times. OpenMS implements a pose-clustering approach[37] that determines linear transformations between multiple runs (see Figure 2.12 for an illustration). Application of the transformation yields an alignment that places corresponding features in close spatial (retention time and m/z) proximity. Corresponding features are determined in a feature linking step. Linked features (also called *consensus*

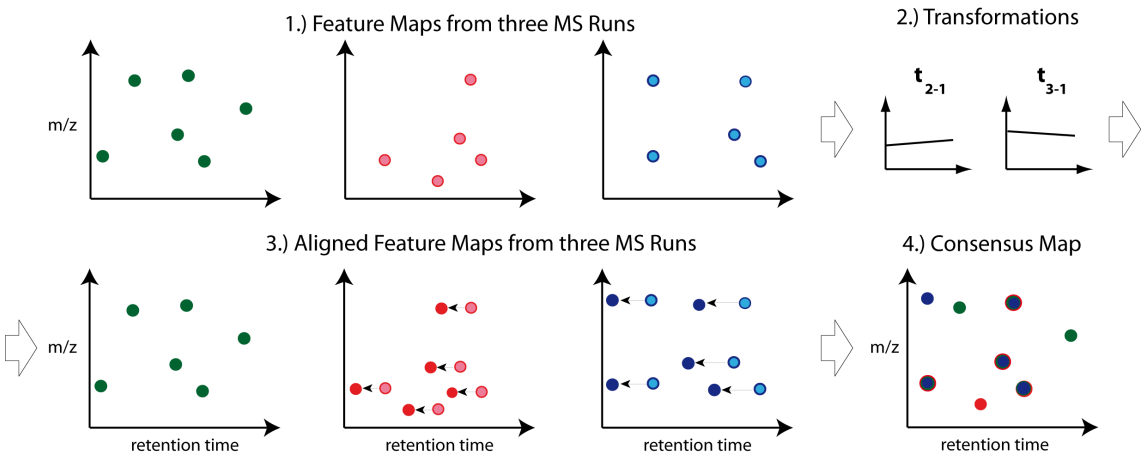


Figure 2.12: Illustration of alignment and linking. Here, the first feature map acts as a reference to which the other two maps are aligned. Transformations are calculated on corresponding features used to globally transform the feature maps. After the alignment, features are linked to form consensus features of a *consensus map*.

features) form a so-called *consensus map* (Figure 2.12). Consensus maps can be built from hundreds of runs and allow quantifying peptides in large studies.

2.3.3 Identification

Identification of proteins is an essential task in mass spectrometry-based proteomics and indispensable for interpretation of the associated quantitative values. In bottom-up proteomics, proteins are digested into shorter peptides to reduce the sample complexity and to simplify data processing. Computational methods identify these peptides and allow inferring the presence of proteins. Several computational approaches to the peptide identification problem exist.

De Novo Sequencing

De novo methods are database free approaches that directly determine the sequence of peptides from tandem mass spectra. These methods rely on high-quality spectra. In the presence of (nearly) complete mass ladders, consecutive amino acids can be reconstructed from mass differences between fragment ions. In practice, incomplete fragmentation of the parent peptide results in incomplete ion series, rendering *de novo* sequencing more difficult or ambiguous. Detector noise, peaks from co-fragmented peptides, peaks caused by neutral losses (e.g., water or ammonia loss), internal fragments and mass peaks caused by higher isotopes populate the MS/MS spectrum further impact the quality of *de novo* results.

Database Search

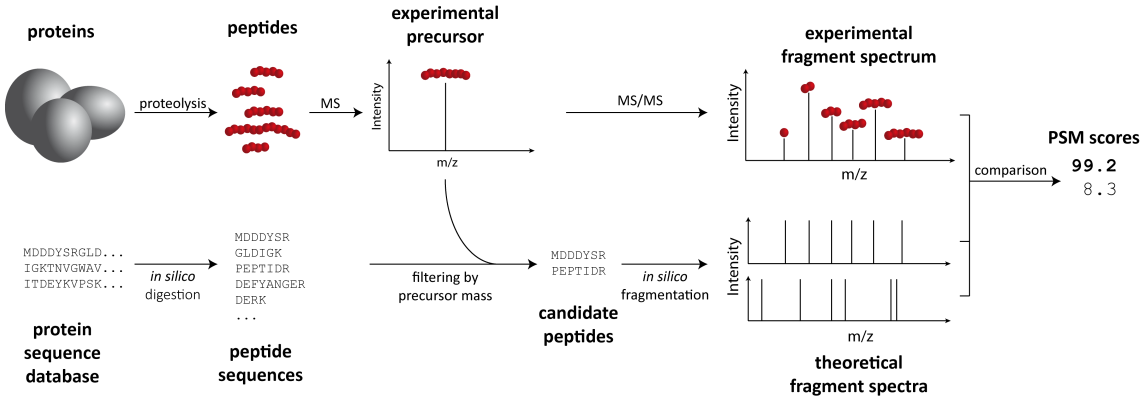


Figure 2.13: Peptide database search (schematic overview). **top:** Experimental workflow: proteins are digested into peptides and measured using tandem mass spectrometry. **bottom:** Peptide database search: *in silico* digestion of a protein sequence database generates a large set of peptide sequences. Only *in silico* peptides that match an observed precursor mass are considered as potential candidates for a peptide spectrum match (PSM). Theoretical fragment spectra of candidate peptides are calculated and compared to an observed spectrum. For each sequence assignment, a PSM score is calculated. Typically, only the top-scoring peptide (i.e., the peptide with best matching theoretical spectrum) is reported.

Peptide database search is currently the most commonly applied technique. Search engines (Figure 2.13) leverage information from genomic databases translated into protein sequences. The protein sequence database gets *in silico* digested using the specific cutting rules of the enzyme used in the experiment. Depending on the experimental setup, *in silico* peptides carrying post- or cotranslational modifications or modifications introduced in the sample processing (e.g., oxidation as a result of the exposure to air or carbamidomethylation from the cysteine blocking reagent) are created. Masses of *in silico* peptides are compared with experimentally observed precursor masses. The set of *in silico* peptides that match a precursor mass (within a specified precursor mass tolerance) are candidates for a PSM. Candidate peptides are computationally fragmented according to the mode of fragmentation (i.e., HCD, CID, or ETD). The resulting theoretical spectra are compared to the observed spectrum, and a score is assigned. Usually, only the top-scoring (e.g., best matching) PSM is reported.

Compared to *de novo* approaches, database search techniques are biased to detect only known sequences. A restriction that is more and more relativized by increasingly complete genome databases. Still, incompleteness poses a major issue for research areas where only incomplete genome information (e.g., metaproteomic studies) are available.

Spectral Library Search

Spectral library search compares a database of previously recorded spectra (with known amino acid sequence) to new spectra. Because this method further reduces the search space to proteotypic peptides, it often outperforms standard database search approaches. A major drawback of this method is that spectral libraries for special research topics (e.g., metaproteomic spectral libraries or spectral libraries of RNA-protein crosslinks) do not exist.

False Discovery Rate

One goal of many proteomics studies is to report a list of identified peptides. In an ideal world, these lists would contain only correct (true positives) but no incorrect sequence assignments (false positives). In practice, this situation is rarely the case, and one inevitably has to compromise between the number of correct and incorrect assignments that are reported. The false discovery rate (FDR) is a statistical concept developed in the context of multiple hypothesis testing that allows evaluating and assessing the trustworthiness of such a list. A brief introduction to the basic concept of target-decoy based FDR estimation is given below. Formally, the FDR is defined as the expected proportion of false positives (incorrect rejections of null hypotheses) among all positives (rejected hypotheses):

$$\text{FDR} = \text{E}[FP/P], \text{if } P > 0, \text{ FDR} = 0 \text{ otherwise}$$

where FP is the number of false positives, and P the total number of positives. In the context of peptide identification, the FDR is thus the expected ratio of incorrect identifications in the reported list of identifications. As the ratio is, a priori, not known, several methods have been developed to estimate it from the data. Today, the most widely used method for FDR estimation in peptide database searches use a target-decoy approach. The classical target-decoy approach augments the original protein database such that for every peptide sequence (target) a decoy sequence is included. Each decoy sequence needs to be constructed in a way that it is (1) not contained in the set of target sequences and (2) competes with its target sequence for the identification of a spectrum. Simple methods construct decoys by reversing or shuffling the protein sequences[i]. Thus, theoretical spectra of target and decoy peptides are scored in equal amounts against an observed spectrum. The highest scoring peptide is then assigned to a spectrum and all spectra identifications sorted by score. In the list of all identifications,

[i] Reversing and shuffling are usually only performed between enzyme cutting sites. Otherwise, number, length, or mass of decoy peptides might differ from their corresponding target peptide.

each assignment of a decoy sequence can thus be spotted as a false positive. A simple variant of the target-decoy approach assumes that the number of misassigned target sequences matches the number of assigned decoy sequences. Hence, the total number of false positives can be estimated as twice the number of reported decoys. Accordingly, the $\widehat{\text{FDR}}$ is estimated as the ratio between twice the number of observed decoys and the number of identifications. So far we only considered a list containing all identifications. In practice, we often want to report a list of top scoring identifications that do not exceed an expected FDR. Controlling the FDR by choosing a smaller threshold truncates the list of reported identifications and decreases the expected ratio of incorrect identifications. The FDR is defined on a set of identifications, while a statistically related quantity, the *q-value*[38], is defined for single identifications. The *q-value* of an identification is defined as the minimal FDR threshold at which the identification is still reported.

Protein Inference

Protein inference is the task of inferring proteins from identified peptides. The central idea of protein inference approaches is that identified peptides provide evidence for the presence of proteins that contain the peptide's sequence. In case a peptide sequence maps to a single protein sequence, the assignment is unambiguous and provides strong evidence for the presence of the protein. If multiple, unambiguous peptides, map to the same protein, then the confidence in the protein identification is increased. If a single peptide maps to multiple proteins, the assignment is ambiguous. Different approaches exist that try to resolve ambiguities, but complete removal is usually not possible. On approach is to report *ambiguity groups* that list the proteins that cannot be distinguished based on the identified peptides. Several ways to define and report protein lists and ambiguity groups have been proposed. For example, a simple maximum parsimony approach determines the smallest set of proteins and groups sufficient to explain the identified peptides.

2.3.4 The OpenMS Framework

OpenMS is an open-source framework for the analysis of mass spectrometry-based proteomic and metabolomic data. It is divided into three conceptual layers ranging from low-level programmatic access to full-featured analysis workflows (see Figure 2.14)

1. *The OpenMS core library* is a C++ library targeted to bioinformaticians and method developers with sound programming skills. It offers the richest set of functionality and is primarily intended for implementing and testing novel algorithms.

2. *The OpenMS pipeline (TOPP) tools* are command line applications that each perform a single, well-defined task. They are targeted to bioinformaticians and method developers that want to apply well-established processing steps to their data. For example, to implement novel variants of existing data processing protocols. They constitute building blocks for larger and more complex processing workflows.

3. *The OpenMS workflows* typically perform higher level analysis tasks like biomarker discovery or label-free quantification. They are constructed and executed in workflow systems (e.g., KNIME, Galaxy[39] or the OpenMS Pipeline Assistant (TOPPAS)), and allow for complex analysis tasks by chaining tools. Full-featured workflow and data integration platforms like KNIME allow combining TOPP tools with an extensive set of external tools for statistics, machine learning, or cheminformatics. This approach enables highly flexible analysis tasks by combining both computational data processing and downstream analysis in a single workflow.

OpenMS provides computational mass spectrometry functionality at the level of algorithms, tools, and workflows and can be used to quickly adapt to the ever-changing challenges in computational mass spectrometry. Novel instruments are frequently presented by manufacturers, and new experimental methods are published on a daily basis. In some cases, these changes only require a subtle change at the level of an existing workflow, like adding a novel signal processing step for spectra filtering. In other cases, a novel tool providing a functionality not contained in other tools might be needed. Conceptually different analysis approaches or experimental protocols might even require developing novel algorithms. Each of these commonly occurring tasks can be addressed with OpenMS at the appropriate level, while a maximum degree of reusability of existing code, tools or workflows is retained. As a whole, OpenMS is designed in a modular fashion to be highly flexible and customizable. This differentiates OpenMS from other existing monolithic applications that are usually specialized on one particular type of analysis.

The OpenMS Core Library

OpenMS started as pure open source C++ library for metabolomic and proteomic data analyses. Since the beginning, its aim has been to provide efficient data structures and algorithms for common data processing tasks. It builds on standardized and open formats for reading and writing raw data from mass spectrometers and analysis results. Computational biologists and bioinformaticians can build on a proven code base to

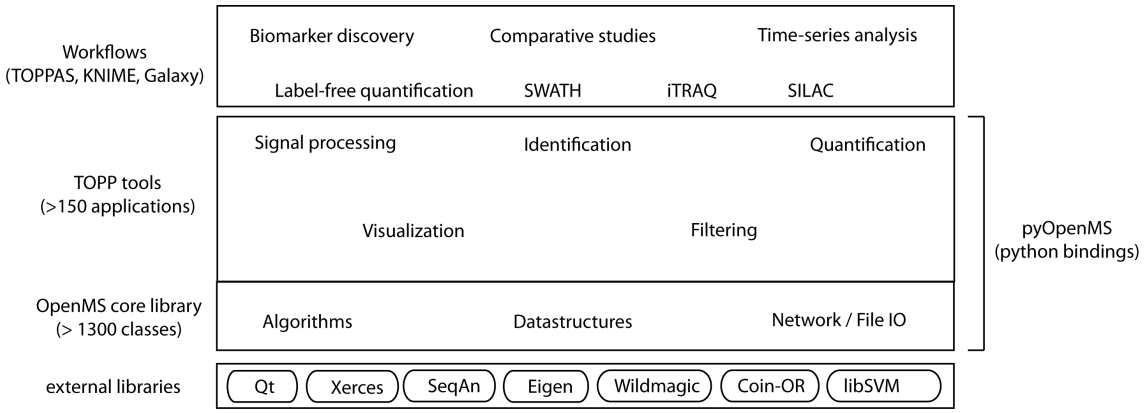


Figure 2.14: OpenMS framework: The **OpenMS core library** builds on a set of external libraries that provide database access, an abstraction layer for system specific functions as well as algorithms and data structures from other domains (e.g., machine learning). The OpenMS *TOPP tools* are small applications built using the OpenMS core library. TOPP tools act as building blocks for arbitrary complex *workflows* executed in common workflow systems like KNIME or Galaxy, as well as in the OpenMS pipeline assistant TOPPAS. *pyOpenMS* provides access to OpenMS functionality using python bindings for prototyping and scripting.

rapidly develop novel algorithms and tools. Development of the OpenMS library was started in 2003 as an academic initiative led by Prof. Knut Reinert (FU Berlin) and Prof. Oliver Kohlbacher (EKU Tübingen).

Core principles of the OpenMS library are:

- The open source model. The permissive license (3-clause BSD) allows to freely use OpenMS in both academic as well as commercial contexts. Source code and executables are available via the project homepage: `http://www.openms.de`.

- A broad platform support. OpenMS supports all major operating systems (Microsoft Windows, MacOS, and Linux-based systems).

- Extensive coverage of computational MS-related entities and tasks. OpenMS provides over 1,300 classes, including data structures for peaks, spectra, mass traces, chromatograms and features and algorithms for signal processing, spectrum generation, isotope pattern matching, and mass trace detection.

- Adhering to standards. OpenMS uses open data standards for reading and writing mass spectrometric data as well as analysis results.

OpenMS itself builds on external open-source libraries that deliver proven and well-tested functionality:

- The **Qt**[40] library acts as a platform abstraction layer providing coherent file system and web access as well as visualization components.
- The **Xerces**[41] library is used for parsing and writing the various XML-based formats supported in OpenMS.
- The **COIN-OR**[42] (Computational INfrastructure for Operations Research) library provides linear and non-linear optimization.
- The **libSVM**[43] machine learning library, which provides support vector machine based classification and regression.
- The **Eigen**[44] header-only library provides algorithms and data structures for fast linear algebra calculations.
- The **WildMagic**[45] library[iii] is used for spline interpolation and regression.
- The **Boost**[46] libraries, a collection of peer-reviewed and portable C++ libraries, is mainly used for regular expression matching.

The OpenMS library can be divided into a stable core part, consisting of basic data structures representing simple entities like amino acid sequences, peaks, spectra, or chromatograms. Building on these basic data structures, the core part also contains more complex data structures for representing all spectra of an MS experiment and its associated metadata. Higher-level functionality relies on these kernel classes and provides read and write support for several standardized data formats. File handlers allow reading the acquired spectra from files into the internal data structures. Opposed to the stable core part, a large number of data processing, data reduction, and data analysis algorithms exist that are regularly extended or adapted to novel methods and instruments. The example below demonstrates how such a multi-threaded spectrum processing algorithm is realized using the OpenMS library.

```
1    // C++ example (excerpt):
2    // Retain the 400 most intense peaks in a spectrum
3
4    // construct a spectrum filter
5    NLargest nlargest_filter = NLargest(400);
6
7    // parallelize loop for concurrent execution using OpenMP
8  #ifdef _OPENMP
9  #pragma omp parallel for
10 #endif
11   for (int i = 0; i < static_cast<int>(spectra.size()); ++i)
12   {
```

[iii] now superseded by the geometric tools library

```
13      // sort peaks by mass-to-charge position
14      spectra[i].sortByPosition();
15
16      // apply filter and keep only the 400 highest intensity peaks
17      nlargest_filter.filterPeakSpectrum(spectra[i]);
18    }
```

In addition, OpenMS provides Python bindings for most of its classes. The simple example below shows how pyOpenMS can be used for spectrum processing:

```
1  # pyOpenMS example: centroid the first spectrum in an experiment
2  import pyopenms
3
4  mse = pyopenms.MSExperiment()
5  fh = pyopenms.FileHandler()
6
7  # load spectra from mzML file
8  fh.loadExperiment(filename, mse)
9
10 # select the first spectrum
11 spec = mse[0]
12
13 # centroid spectrum using the OpenMS peak picker algorithm
14 picker = pyopenms.PeakPickerHiRes()
15 newspec_out = pyopenms.MSSpectrum()
16 picker.pick(filtered_spec, newspec_out)
```

The OpenMS Tools

In proteomic and metabolomic research it became very early evident that data analysis and data processing steps vary greatly between different experimental setups. New experimental techniques, as well as constantly evolving instruments, require flexible data processing and analysis workflows. Combining these into single software application has only been possible for a subset of techniques and experimental setups. TOPP tools are command line tools that provide a common interface for tool configuration and use open data exchange formats for passing data between tools. Each of these command-line applications acts as a building block of defined functionality. Nearly arbitrary complex analysis workflows can be obtained by chaining these tools together using scripts or as components in workflow systems. As of today, more than 158 TOPP tools for mass spectrometry analysis have been developed using the OpenMS library[47]. The functionality of TOPP tools ranges from file conversion and filtering, over MS data processing and data reduction to identification and quantification of metabolites and proteins. For several existing third-party applications, wrappers are provided in TOPP (e.g., peptide identification with OMSSA[48], X!Tandem[49], Mascot[50], MyriMatch[51], or protein inference with Fido[52]) that can be executed in combined workflows.

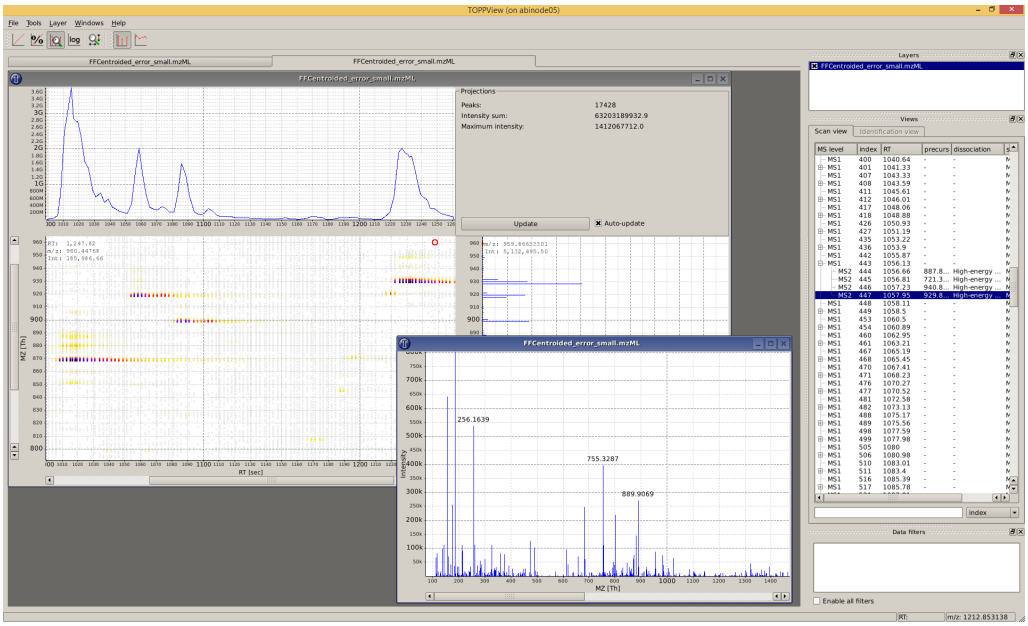


Figure 2.15: TOPPView is the main application for visualization of raw spectra, identifications and quantified analytes in OpenMS. It features, amongst others, 2D and 3D visualizations of peak maps using color-coded intensities, manual and automated annotations of mass peaks. The capability to execute TOPP tools from within TOPPView allows to quickly inspect results and optimize parameters.

In addition to command line tools, OpenMS provides the graphical application TOPPView (Figure 2.15) to examine raw spectra, the effects of data processing steps as well as results of identification and quantification tools. The TOPPView application provides graphical dialogs to configure and run TOPP tools. Inspecting the result allows optimizing tool configurations to find the best parameters for a particular type of data or instrument.

Integration into Workflow Systems

Script-based data processing using TOPP tools is a powerful way to perform complete data processing workflows. They can be run locally or on grid environments allowing the analysis of large amounts of data. Workflow systems assist the user in building complex processing pipelines by providing a graphical user interface and, hence, reduces the need to write scripts. In addition to an increase in user-friendliness, the modern workflow system KNIME acts as an integration platform for a large number of tools from different sources. For instance, plugins can be installed that bundle tools from cheminformatics, genomics, statistics, machine learning, or in the case of the OpenMS plugin: tools for the analysis of proteomics and metabolomics data. To build a workflow,

KNIME nodes are placed in the workflow editor, each node representing a processing action, input or output data. Data flow between nodes is indicated by connections between incoming and outgoing ports (Figure 2.16). Integration of TOPP tools in KNIME is performed in an automated process that:

1. invokes each tool to generate a parameter description including a short documentation stored in a common tool description file (CTD) file,

2. uses the Generic KNIME Nodes (GKN) framework to generate nodes from a CTD file. Like other KNIME nodes generated using GKN, they allow to configure and execute the underlying tool in KNIME,

3. bundles all KNIME nodes, TOPP executables and OpenMS library for integration into the OpenMS KNIME community nodes.

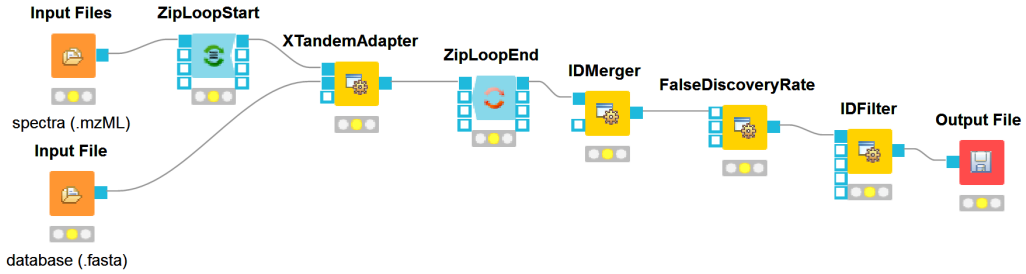


Figure 2.16: KNIME Simple OpenMS workflow. **yellow nodes:** TOPP tools, **orange/red nodes:** input and output files, **turquoise nodes:** loop logic for sequential processing of file lists. Data flow between nodes is indicated by lines connecting incoming and outgoing ports (gray squares). Traffic light symbol below nodes indicate the node's status.

Downstream processing of identification and quantification results are conveniently performed in KNIME. KNIME provides a rich set of nodes for statistical analysis and visual data exploration. The final results obtained after workflow execution may constitute potential biological relevant findings that can be published or give rise to follow-up experiments.

Upon publication, both results and computational workflow - including the raw mass spectrometry data - can be provided to other researchers to inspect and reproduce findings. Using OpenMS in KNIME, therefore, allows achieving a level of computational reproducibility rarely observed in practice. Unfortunately, custom scripts or obscure intermediate steps are still regularly employed.

Role of OpenMS in this Thesis

During the time of this Ph.D. thesis, the author has been an OpenMS core developer and responsible for maintenance, coordination of development efforts as well as release management. The MetaProSIP and RNP^{xl} TOPP tools presented in this thesis, as well as code to read and write mzTab files, have been developed by the author and were integrated into OpenMS.

Chapter 3

Single Amino Acid Assignment of Nucleotide-binding Sites in RNA- and DNA-binding Proteins

The content of this chapter is to a large extent part of the manuscript:

Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins

Katharina Kramer+, Timo Sachsenberg+, Benedikt M. Beckmann, Saadia Qamar, Kum-Loong Boon, Matthias W. Hentze, Oliver Kohlbacher, Henning Urlaub *Nature Methods* 11, 1064–1070 (2014)

+ These authors contributed equally

3.1 Introduction

Long chains of nucleic acids, RNA and DNA molecules, constitute main classes of biomolecules found in all living organisms. They play pivotal roles in a variety of cellular processes and are essential for survival and replication. Messenger RNA (mRNA), the most well-studied class of RNA, gets transcribed from endogenous DNA molecules, the primary carrier of the genetic information. Proteins are formed by the translation of the genetic information stored in the mRNA molecule into amino acid chains. Proteins carry out the main enzymatic activities and catalyze the rich set of biochemical reactions in an organism. Some recent theories suggest[53] that the predominant role of proteins is probably a more recent development in the evolution of self-replicating systems[54]. Instead, RNA-based enzymes, so-called ribozymes, have been catalyzing specific actions similar to protein enzymes in ancient cells and might have also had a role as a carrier of genetic information as DNA molecules[55] today. While

proteins, as well as DNA, have taken over many of their original tasks, many important cellular processes still rely on temporary interaction or stable complexes of nucleic acid chains with proteins. Ribonucleoproteins (RNPs) form functionally diverse complexes of RNA and protein molecules[56] and are often seen as evolutionary descendants of the old RNA processing machinery. RNPs differ significantly in function, structure, and mode of RNA-protein binding[57]. The most well-studied RNP, the ribosome[58] (Figure 3.1), is a key player in the aforementioned synthesis of proteins. Other well-known examples are heterogeneous ribonucleoprotein particles (hnRNPs), small nuclear ribonucleic proteins (snRNPs), telomerases, components of the editosom, or small RNA-associated protein complexes (e.g., the RNA-induced silencing complex (RISC)). DNA-protein complexes play an equally important role with functionally diverse classes, such as histones, helicases, or transcription factors. It should be noted that RNA- or DNA-binding protein complexes should no be considered distinct. Recently, more and more cases of DNA- and RNA-binding proteins (DRBPs)[59] have been discovered.

3.1.1 Motivation

RNPs are usually essential[17] for the survival of an organism. Often, small genetic alterations in one of its components induce severe effects on the organism scale[18] and are associated with well-known diseases like amyotrophic lateral sclerosis (ALS)[60] or premature aging syndrome[61].

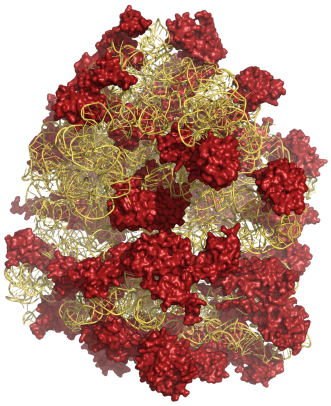


Figure 3.1: A prokaryotic ribosome during translation initiation (pdb:4v4j). Six RNA chains (4702 nucleotides) are colored in yellow, and 48 proteins chains (6163 amino acids) in red.

Recently, the use of bacterial RNPs for genome editing (e.g., CRISPR/Cas system[62]), has gained popularity as a powerful tool. Not surprisingly, there is an enormous interest in RNPs for basic as well as clinical research. Progress in these fields is currently hampered by the lack of structural information on large protein-RNA/DNA-complexes. Therefore, little information is known about the particular details of protein-RNA/DNA interactions - especially at the resolution of single amino acids and nucleic acids in contact.

Several techniques have been developed in the past and are currently used to investigate structure and interaction of protein-RNA/DNA complexes.

3.1.2 Structure Elucidation

In general, the largest amount of structural information is obtained by techniques that allow reconstructing the spatial locations of nucleotide and protein atoms directly. Classical techniques (often applied in combination) are X-ray crystallography (XRC)[63], nuclear magnetic resonance spectroscopy (NMR spectroscopy)[64], and cryo-electron microscopy (cryo-EM)[65]. In XRC, a three-dimensional image of electron densities can be calculated from diffraction patterns obtained by X-ray irradiation of crystallized molecules. Densities, in turn, allow reconstructing the atomic positions of the molecule. In NMR spectroscopy, nuclei absorb electromagnetic radiation at a specific resonance frequency that depends on the magnetic field and isotope. Atomic positions can be reconstructed from pairwise distance or angular constraints obtained by observing interactions between atoms in close proximity. Single-particle analysis using cryo-EM has shown some recent progress and allows to investigate complexes with limited conformational heterogeneity[66]. In a variant of cryo-EM, a large number of 2D projection images are generated from a frozen layer containing many copies of the protein complex of interest. Projections from different orientations are combined and allow to increase resolution as well as to generate a 3D reconstruction of the structure. While XRC, NMR, and cryo-EM have been successfully used to resolve 3D structures of protein-protein complexes, they have some restrictions when applied to RNA/DNA-protein complexes. Often these complexes do not crystallize, have too low yields, or are not applicable to *in vivo* studies. They are low-throughput methods by design and are not suitable for screening novel interactions between a large number of different molecules as obtained from complex samples. Also, they typically require highly purified complexes exhibiting a large degree of conformational homogeneity.

Recently, complementary techniques have been introduced that combine cross-linking with enrichment of the cross-linked heteroconjugates. Cross-linking immunoprecipitation (CLIP), coupled with high-throughput sequencing (HITS-CLIP or CLIP-seq) is a transcriptome-wide cross-linking method that combines UV cross-linking and immunoprecipitation[67]. Photoactivatable-Ribonucleoside-Enhanced CLIP (PAR-CLIP)[68] extends HITS-CLIP in that it introduces nucleoside analogs with increased photoreactivity to increase the yield of the cross-linking reaction. The purified and cross-linked complexes obtained by HITS- or PAR-CLIP are investigated with deep sequencing to determine the cross-linked nucleotides. Unfortunately, only partial information on the interaction is obtained, since the amino acids in contact remain concealed by these methods.

We propose a novel experimental and computational method and workflow for the identification of nucleotide-binding sites in nucleotide-binding proteins. It differs from existing approaches as it combines UV-induced cross-linking, enrichment and novel methods for computational mass spectrometry to pinpoint the cross-linking site in an automated fashion. To this end, we solve two intermediate objectives by providing computational methods for:

1. The automated identification of the cross-linked peptide-RNA/DNA pair.

2. The automated localization of the cross-linking site on the peptide.

Figure 3.2 illustrates the conceptual difference between identification and localization. Automated identification (see Section 3.2), as described in our publication[69], determines the cross-linked/oligonucleotide pairs. At that point, no automated localization of the cross-link on the peptide chain was performed. Instead, extensive manual fragment annotation of identified heteroconjugates was used to manually pinpoint the exact cross-linked amino acid in peptides. With an increased understanding of the fragmentation chemistry of cross-links, we were later able to automatize the localization of cross-links (see Section 3.3).

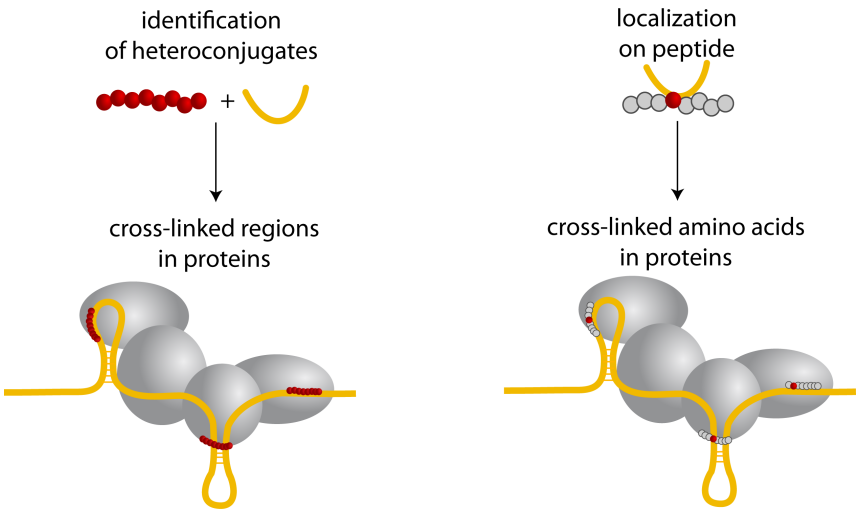


Figure 3.2: Cross-link identification and localization. Identification of peptide (red) and nucleotide (yellow) heteroconjugates provide information on the cross-linked regions in a protein complex. Localization of the cross-linked nucleotide on the peptide allows to accurately pinpoint the cross-linked amino acids.

3.2 Automated Cross-Link Identification

In order to identify cross-linked heteroconjugates, several challenges need to be addressed. First and foremost, an experimental method needs to be established that allows to induce and measure cross-links using mass spectrometry with high sensitivity. This involves experimental enrichment and sample processing steps to increase the, in general, low expected yields of cross-links. As no computational tools and algorithms for spectra processing and identification of cross-links exist, these need to be developed.

3.2.1 Methods

Similar to the PAR-CLIP and HITS-CLIP methods, we use photo-induced cross-linking to form stable protein-RNA complexes, followed by enrichment of cross-links. Instead of deep sequencing, we measure cross-links by high-resolution mass spectrometry (Figure 3.3). The resulting spectra are analyzed using an automated computational workflow that supports joint analysis of the UV-irradiated sample along with a non-irradiated control. Joint analysis of irradiated sample and control sample allows us to computationally reduce the number of false-positive detections using custom-developed tools and workflows. Additionally, we developed the tool RNP^{xl} that identifies cross-linked peptide-oligonucleotide moieties. To ease manual validation and localization of the cross-link, the OpenMS application TOPPView has been extended to visualize the identified spectra along with the identification results. The validated spectra are then compared to existing knowledge (e.g., known protein structure, annotated domains) and interpreted in its biological context.

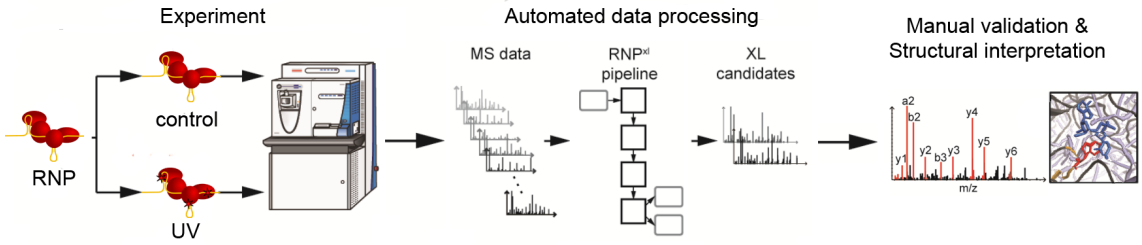


Figure 3.3: Overview of the RNP^{xl} experimental workflow UV-irradiated protein-RNA complexes and non-irradiated control are jointly prepared for LC-ESI-MS/MS on a Thermo Fisher Orbitrap mass spectrometer. Both mass spectrometry runs form the input into the automated data-processing pipeline. Potential cross-links are determined and manually validated. If available, interpretation of identified cross-links is performed based on structural, functional information or annotated domains. *Adapted from Kramer et al.* [69].

Cross-Linking and Enrichment

The majority of proteins in a cell lysate do not interact with RNA molecules. Those that do interact form either stable or transient complexes[16]. When exposed to UV light, covalent bonds between atoms of partner molecules in close proximity are formed in the complex (Figure 3.4). The chemistry of cross-link formation is not yet fully understood. According to Meisenheimer and Koch[70], Williams and Konigsberg[71] cross-links might be a formed by a free radical mechanism. Upon UV absorption, the excited base abstracts a hydrogen atom from the adjacent amino acid residue to form a pyrimidinyl radical. A zero-length cross-link is then formed by radical combination (see Appendix Figure C.1).

As the newly formed RNA-protein heteroconjugates are few (1% or lower[72]) when compared to the number of noncross-linked proteins and RNA molecules in the sample, additional enrichment steps need to be performed.

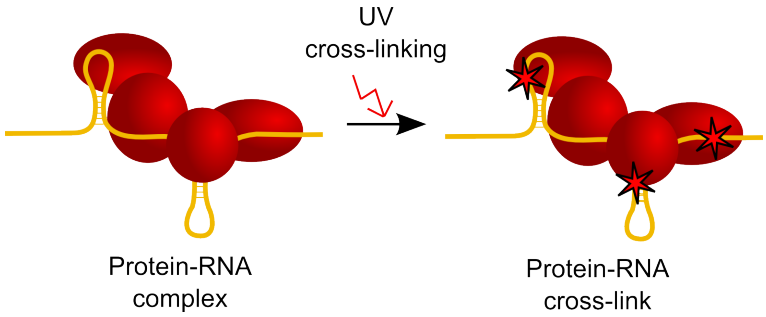


Figure 3.4: Cross-linking of amino acids and nucleotides is induced by irradiation with ultraviolet light at a wavelength of approx. 254 nm. Asterisks illustrate newly formed covalent bonds between proteins (red) and RNA (yellow).

Sample processing starts with the digestion of denatured proteins and RNA into manageable sizes. Proteins are first denaturated in urea buffer and digested into peptides using the endopeptidase trypsin. The single-stranded and denatured RNA is hydrolyzed by ribonucleases. As a result, noncross-linked RNA oligonucleotides are removed and short (one to four nucleotides), cross-linked RNA remains bound to peptides. If protein complexes bound to large RNAs (e.g., spliceosome or ribosome) are studied *in vitro*, an intermediate step of size exclusion chromatography (SEC) can be performed between proteolytic and nucleolytic digestion. Peptides bound to the large RNA are retained by SEC, while the smaller, noncross-linked peptides are removed. Further enrichment can be obtained by chromatographic separation (Figure 3.5).

Filtering of noncross-linked RNA is performed using reversed-phase chromatography. Small RNA oligonucleotides do not bind to the C18 material while peptides or

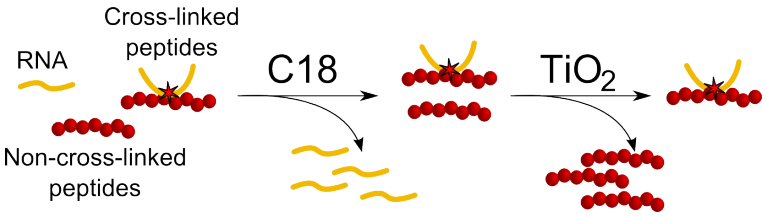


Figure 3.5: Enrichment of protein-RNA heteroconjugates. After hydrolysis, reverse phase C18 chromatography removes oligonucleotides. Titanium dioxide (TiO_2) solid phase extraction removes noncross-linked peptides.

cross-linked peptides are retained by the column. The remaining mixture of peptides with and without bound oligonucleotides can be further separated.

Filtering of noncross-linked peptides can be performed by TiO_2 chromatography. TiO_2 chromatography is a widely used technique for the separation of organic phosphates[73]. In proteomics, TiO_2 solid phase extraction is the preferred method to enrich phosphopeptides selectively[74]. Because nucleotides carry phosphate groups, we use a similar protocol to separate cross-linked peptides from noncross-linked ones.

For a detailed, protocol style description of the sample processing, the reader may consult Qamar et al.[75] or Sharma et al.[76].

Mass Spectrometry Analysis

The purified peptide-RNA cross-links are subjected to nano-liquid chromatography, electrospray ionization, and tandem mass spectrometry analysis using high-resolution instruments. We exclusively used Orbitrap instruments manufactured by Thermo Fisher Scientific. These instruments allow recording high-resolution tandem mass spectra using HCD fragmentation in data-dependent acquisition (DDA) mode. MS raw files obtained from the instrument software were subjected to the RNP^{xl} workflow.

Computational Workflow

The computational workflow of RNP^{xl} can be divided into three main steps: data preparation, data reduction, and cross-link identification with the RNP^{xl} tool (Figure 3.6).

Data Preparation

MS raw data was converted from Thermo Fisher's raw file format into the mzML format using msconvert of the ProteoWizard software package[77]. MzML files were then processed by TOPP tools (see Appendix 3.6 for a detailed visualization of the RNP^{xl}

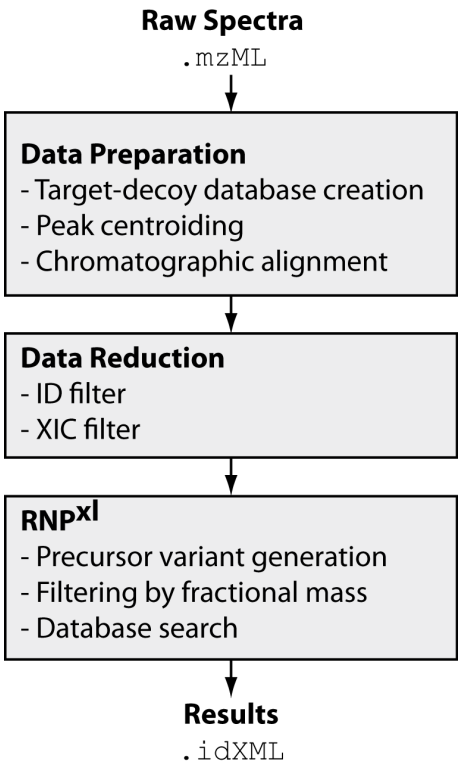


Figure 3.6: Overview of the RNP^{xl} computational workflow.

workflow topology). Mass spectra acquired in profile mode were centroided by the OpenMS tool `PeakPickerHiRes`. Both non-irradiated control and UV-irradiated MS runs were aligned to compensate for chromatographic retention time shifts. Alignment was performed using the `MapAlignerPoseClustering` tool.

Data Reduction

Several data reduction steps were applied to remove tandem spectra likely corresponding to noncross-linked peptides, RNA-derived fragments, or contaminants from the analysis.

Identification-based Filtering of Spectra Matching Peptides and Contaminants: In order to remove spectra corresponding to noncross-linked peptides as well as known contaminants, we constructed a pipeline that filters tandem spectra corresponding to confidently identified peptides and contaminants. First, a standard database search is performed using OMSSA. We use a target-decoy database created from translated gene sequences provided by UniProt[78]. In addition to the proteins of the particular organism, we add common contaminant sequences as distributed with the MaxQuant software

package[79]. We considered several common modifications of peptides induced or enriched by the sample preparation procedure in the search. In addition to oxidation of methionine, carbamylation of lysine and N-termini, we also considered phosphorylation of tyrosine, serine, and threonine as variable modifications because phosphopeptides might also be enriched by the use of TiO_2. The TOPP tool `FalseDiscoveryRate` was used to estimate *q-values* of tandem spectra. Spectra were then filtered using the OpenMS tool `IDFilter`. The list of peptides obtained from the UV-irradiated dataset was thresholded at an FDR of 1%. The remaining spectra were subjected to further analysis.

Extracted Ion Chromatogram based Filtering: Pure RNA moieties and unidentified peptides (or contaminants) may occur in both UV-irradiated sample and control. Precursors that coelute in UV-irradiated and control sample are most likely not derived from cross-links (see Figure 3.7). Their tandem spectra can be excluded from further analysis. For each precursor peak annotation in the UV-irradiated sample, we extract an extracted ion chromatogram (XIC) (extraction interval: ± 10 s in retention time and ± 10 ppm m/z around precursor peak) from both UV-irradiated and control. For both XICs, we calculate a single intensity value as the sum of individual peak intensities and compare them. If the intensity in the control is at least half the intensity in UV-irradiated, we exclude the tandem spectrum from further analyses. XIC-based filtering of spectra was implemented in our novel TOPP tool `RNPxlXICFilter`.

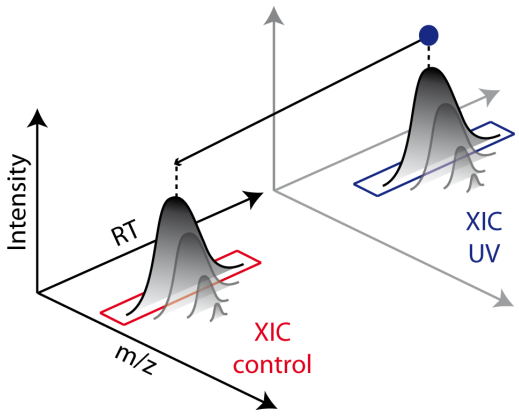


Figure 3.7: XIC Filter extracts XICs at precursor peak positions (blue circle). Signals present in both UV and control correspond to noncross-linked species and the corresponding tandem spectra are removed from the UV sample.

RNP^{xl} Tool

During cross-link formation, the oligonucleotide gets covalently bound to the peptide. After digestion with RNases, oligonucleotides of a length between one and four remain bound to the peptide. Before fragmentation, the mass of a cross-link $m_{cross-link}$ is simply the sum of peptide mass $m_{peptide}$ and oligonucleotide mass m_{RNA} minus some potential neutral losses m_{loss} that may occur on the oligonucleotide:

$$m_{cross-link} = m_{peptide} + m_{RNA} - m_{loss}$$

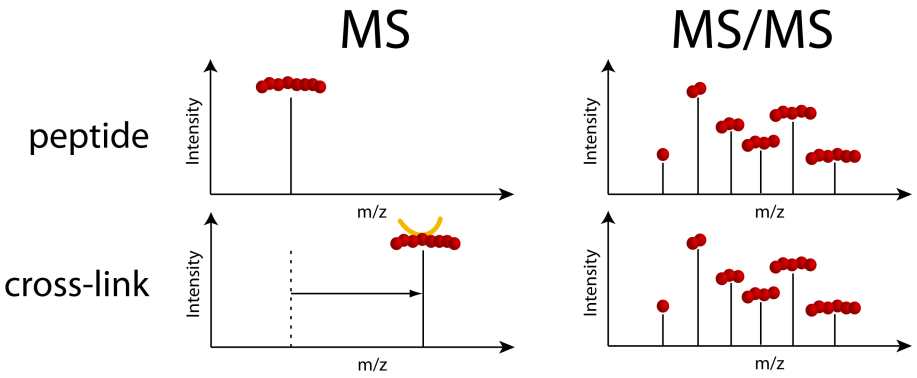


Figure 3.8: Cross-link masses are, compared to the mass of the noncross-linked peptide, shifted by the oligonucleotide mass. Fragment spectra remain virtually unchanged.

It has been previously observed that the peptide-RNA cross-link is relatively unstable. During fragmentation, the oligonucleotide easily breaks apart from the peptide. As a consequence, sequence ions in tandem spectra with oligonucleotides bound to them are less often observed. Fragmentation pattern of a cross-linked peptide therefore closely resembles the pattern of the noncross-linked peptide (see Figure 3.8 and Appendix Figure C.2). Based on these observations, cross-linkings can be modeled as a modified peptide with neutral loss of the modification upon fragmentation. The RNP^{xl} tool generates the peptide modifications that arise in the cross-linking reaction between peptides and short RNA oligonucleotides. Because the potentially thousands of different modifications surpass the number of search modifications supported by current standard database search engines we applied a simple technique we refer to as *precursor variant generation*, to make the cross-linking datasets searchable.

Precursor Variant Generation relies on the observation that tandem mass spectra of the unmodified peptide, as well as cross-links to varying oligonucleotides, share the same (unshifted) sequence ions. While in contrast, the precursor masses differ by the

modification *delta mass*

$$\Delta m_{RNA,loss} = m_{RNA} - m_{loss},$$

introduced by the oligonucleotide and associated losses. Given an unidentified precursor with mass m_p, precursor mass variants can be created that transform the precursor mass annotation of cross-links to the mass of the noncross-linked peptide.

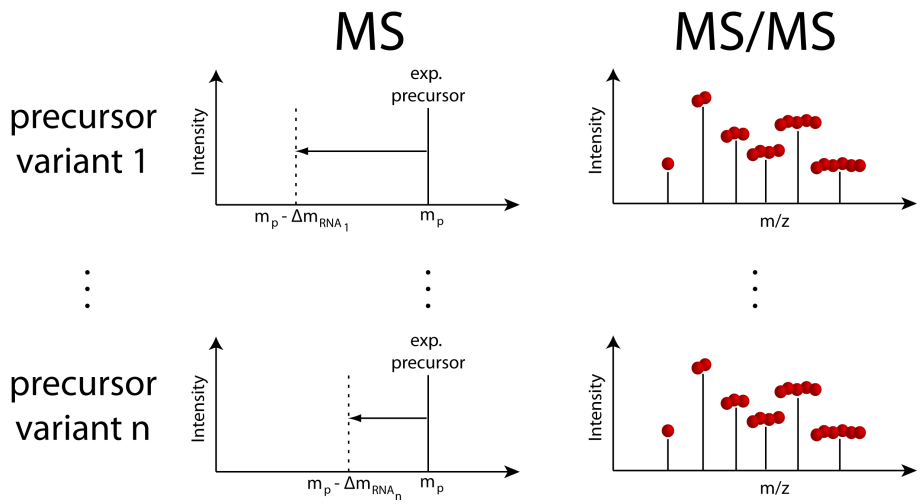


Figure 3.9: Precursor variants are generated by subtracting all oligonucleotide masses from the precursor mass. The tandem spectra recorded from the precursors are not changed. Losses are not explicitly considered in this illustration.

According to the previous equation, this transformation is simply achieved by subtracting the delta mass from the precursor mass $m_{p,unmodified} = m_p - \Delta m_{RNA,loss}$. We generate all precursor mass variants of a tandem spectrum (Figure 3.9). The synthesized spectra can be queried against the unmodified peptides using a standard peptide database search. The top scoring precursor variant, corresponding to a distinct RNA + loss (or lack thereof), is retained, annotated and reported for manual inspection.

In order to generate all possible peptide modification masses for precursor variant generation, RNP^{xl} *in silico* synthesizes all nucleotide compositions for oligonucleotides up to a specified length and applies the potential losses. Additionally, the feasibility of formed modifications and precursor variants are checked using user provided constraints and filtering by plausibility (i.e., if a chemically valid biomolecule is formed).

In vivo, RNA chains are predominantly formed by RNA polymerases. These enzymes catalyze the formation of phosphodiester bonds between the 3' carbon of one nucleotide and the 5' carbon of the other nucleotide. In the chain extension reaction, a nucleotide triphosphate is linked to a chain of nucleotide monophosphates under release of pyrophosphate. *In silico* calculation of the nucleotide chain and empirical formula

can be simplified to consider only the net change in elemental composition during the extension reaction, which corresponds to a simple condensation reaction that involves adding nucleotide monophosphates without explicitly modeling the full reaction. A further simplification arises from the fact that only all nucleotide compositions need to be considered. This implies that the order of nucleotides in the chain can be ignored.

To adapt RNP^{xl} to a variety of experimental setups, we provide several parameters that allow modifying the generation of nucleotide chains. Among others, these parameters allow employing nucleotide analogs and isotopically labeled nucleotides.

Nucleotides (or nucleotide variants) used by RNP^{xl} can be freely specified using a single letter code and associated chemical formula of the nucleotide monophosphate. In the default setup, the standard ribonucleotides are preconfigured. Nucleotide analogs have been shown to increase reactivity and yield larger amounts of cross-links compared to the standard nucleotides. In some of our experiments, we used 4-thiouridine (4SU) ($C_9H_{13}N_2O_8PS$) as uridine analog (see Appendix Figure C.3a and C.3b for chemical structures). Changing the default configuration from uridine monophosphate to 4SU (changing "U=C9H13N2O9P" to "U=C9H13N2O8PS") enables support for the 4SU analog in RNP^{xl}.

Heavy-isotope labeled nucleotide variants show similar reactivity as the unlabeled nucleotide. They are primarily used to quantitatively compare labeled and unlabeled cross-links analogous to a SILAC analysis. RNP^{xl} supports isotopically labeled nucleotide variants by specifying isotopes in the chemical formula of a nucleoside monophosphate (e.g., guanine with one heavy carbon atom is specified as: "G=(13)C1(12)C9H14N5O7P").

Specification of a sequence constrains the generation of oligonucleotides on substrings of the provided sequence. In the context of studying single protein complexes, the nucleotide sequence is often known from complementary methods like PAR-CLIP. A significant reduction of oligonucleotide candidates can be achieved if this information is provided. If no sequence is specified, all possible oligonucleotide sequences are considered.

The maximum oligonucleotide length restricts the oligonucleotide chain to a maximum of typically, one to four nucleotide oligos. We choose a maximum of four nucleotides as default for the expected oligonucleotide length after the digestion with RNases.

Mapping rules provide an easy way to specify potential locations of special nucleotides (e.g., nucleotide analogs or isotopically labeled nucleotides) in sequences. The default mapping rule is the identity which maps each letter in the sequence to its corresponding nucleotide. If, for instance, a nucleotide analog *may* be present at

the position of the letter 'X' in "AUGCC**X**AA" (e.g., corresponding to unlabeled 'U' and alternatively labeled uridine 'Y'), 'X' is mapped to 'U' and 'Y'. As a consequence, two peptides: "AUGCC**U**AA" and "AUGCC**Y**AA" are generated which in turn are then used as template sequence to produce all oligonucleotide variants (Appendix Table C.3 and Table C.4).

The minimum count of a nucleotide limits the set of oligonucleotides to those that contain a minimum number of the specified nucleotide. This feature of RNP^{xl} is particularly useful if one wants to enforce that a known cross-linked nucleotide is always part of generated sequences.

Neutral losses of small neutral molecules occur during cross-link formation. Depending on the type of nucleotide involved, different neutral losses have so far been observed. For standard nucleotides, we predominantly observe loss of water and phosphoric acid from the precursor. These losses are reflected in the default configuration of RNP^{xl}. For nucleotide analogs, like 4SU, loss of $-H_2S$ must be considered. RNP^{xl} allows specifying all neutral loss variants that should be generated for each oligonucleotide.

Dithiothreitol (DTT), as discovered in our studies, acts as highly specific, non-zero length protein-RNA cross-linker[80] that gives rises to a precursor adduct ($C_4H_8S_2O_2$, 152 Dalton) on cysteine. We added an option to include this newly characterized precursor adduct to RNP^{xl}.

Filtering by Fractional Mass. Peptides and oligonucleotides (as well as heteroconjugates) differ in their molecular composition. Particularly, in the relative amount of phosphorus. This difference in molecular composition leads to different fractional mass (= decimal fraction following the monoisotopic nominal mass) distributions characteristic for different classes of molecules. Because phosphorus exhibits a large mass defect[i], it is - in certain mass ranges and under some oligonucleotide length constraints - possible to classify molecules solely based on the accurate monoisotopic (= nominal + fractional) mass (see Pourshahian et al.[81] for details). In RNP^{xl}, we implemented fractional mass filtering of precursors that cannot correspond to a cross-link.

Automated Database Searches. All generated precursor variants passing the previous filter criteria are subjected to database search using OMSSA. To keep file sizes within reasonable bounds, the RNP^{xl} tool generates one batch of precursor variant spectra for each tandem mass spectrum.

Reporting and Annotation. Search results are read back from OMSSA by the RNP^{xl} tool. The best scoring PSM is retained and annotated with the respective cross-link (or unmodified) peptide. Additionally, mass peaks matching to known RNA marker ions are annotated and stored as metadata. Export of the final identification results

[i] fractional mass difference of approx. -0.0262385 to the atoms nominal mass

is performed as tabular separated as well as idXML[ii] file for further processing and visualization in TOPPView.

Visualization

To manually validate cross-links a component for visualization of peptide identification results was developed and integrated into TOPPView (Figure 3.10).

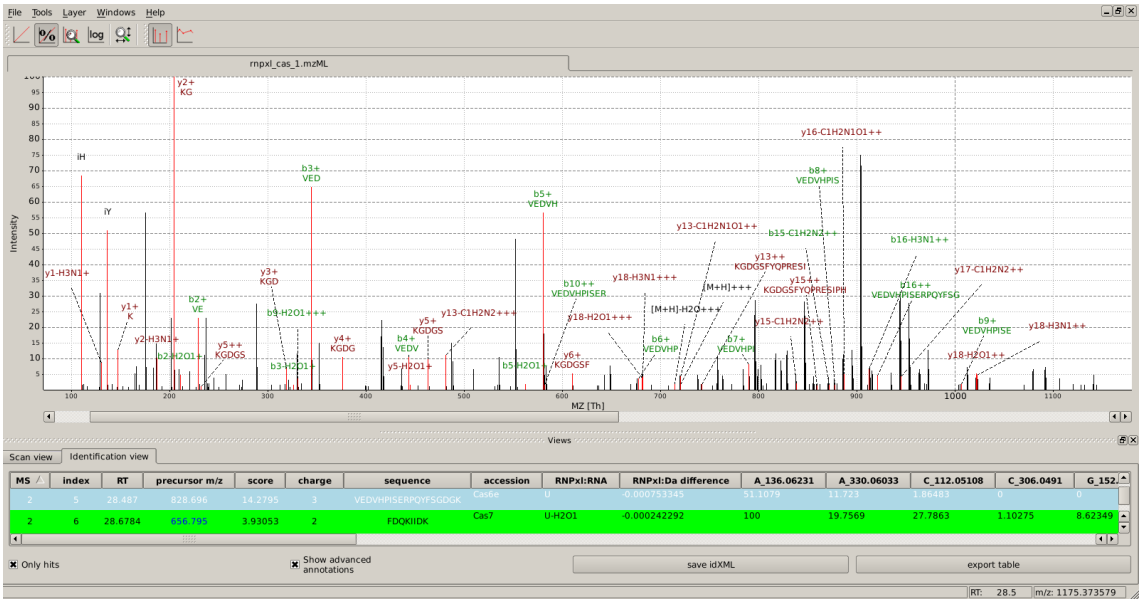


Figure 3.10: TOPPView top: annotated spectrum, bottom: table of identified sequences and associated metadata (e.g., charge, RNA sequence, marker ion intensities,...)

Manual Validation

Manual validation of cross-link candidates requires some experience in interpreting mass spectra but typically involve the following steps:

1. Candidates are sorted by search engine score and evaluated by manual inspection in TOPPView.

2. Similar to manual validation of PSMs, low-quality spectra, precursor m/z misassignments or co-fragmentation are easily spotted, and spectra can be discarded. Optionally XICs from UV-irradiated and control sample can also be compared to assess the performance of the XIC filtering step.

[ii] an OpenMS file format used to store identification results

3. Automated *a-*, *b-* and *y-ion* fragment annotations in TOPPView allow to quickly determine amino acids with adducts (e.g., U − H_3PO_4, U − H_2O or −H_2S). In addition, high-intensity RNA marker, originating from oligonucleotides can be compared to the RNA sequence.

4. If fragment spectra are dominated by low-intensity signals and incomplete mass ladders, shifted internal and immonium ions can be used to augment the information needed to obtain complete ion-ladders, which in turn allows pinpointing the cross-linked amino acid.

Prominent but unassigned mass differences of identified cross-links have been recorded as potential novel RNA adduct.

3.2.2 Results

We subjected samples from three cross-linking experiments to our analysis workflow:

1. human RNPs
2. yeast RNPs isolated with TAP purification of Cbp20
3. yeast RNPs labeled with the 4SU nucleotide analog and isolation with oligo d(T)

The data has been described in detail in our original publication.

Effect of Filtering Strategy

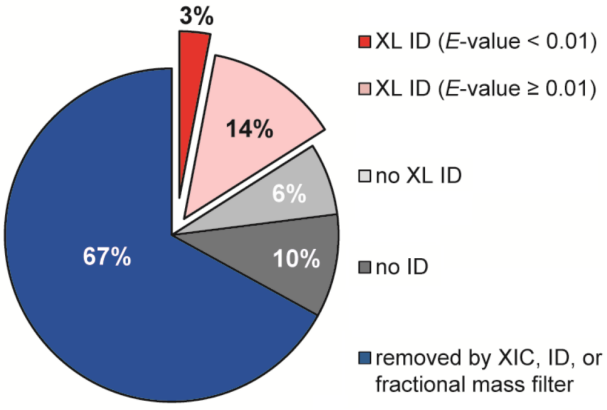


Figure 3.11: Data reduction in a yeast run. 97% of all spectra are discarded because they did not pass the XIC, ID or fractional mass filter, did not match to cross-linked heteroconjugates (no XL ID, no ID), or achieved only a low identification score (XL ID *E-value* ≥ *0.01*). *Adapted from Kramer et al.* [69].

Our filtering strategy was employed on all datasets and in all cases greatly reduced the number of candidate spectra (Figure 3.11). In a representative yeast run, two-thirds of initial 9,728 candidate spectra were removed by the ID (29%), XIC (34%) and fractional mass (3%) filters. After cross-link search, an additional 10% of spectra did not yield any identification, 6% corresponded to low-scoring peptide identifications. Of the remaining 17% potential cross-link spectra, 14% were assigned PSMs with a low score (E-value ≥ 0.01) and removed. The remaining 3% yielded the final list of cross-link candidates for manual validation. In summary, the total number of 9,728 spectra was reduced to 317 spectra.

We carefully optimized parameters of our filtering approach to remove only tandem spectra of noncross-linked analytes. Spectra excluded by the filters were routinely subjected to cross-link search and manual inspection. Except for rare cases, cross-links found by the RNP^{xl} tool were false positives and often associated with low-quality spectra. We are therefore confident that our filtering strategy is conservative and accidental removal of cross-links is expected to happen rarely.

Identified Cross-Linking Sites

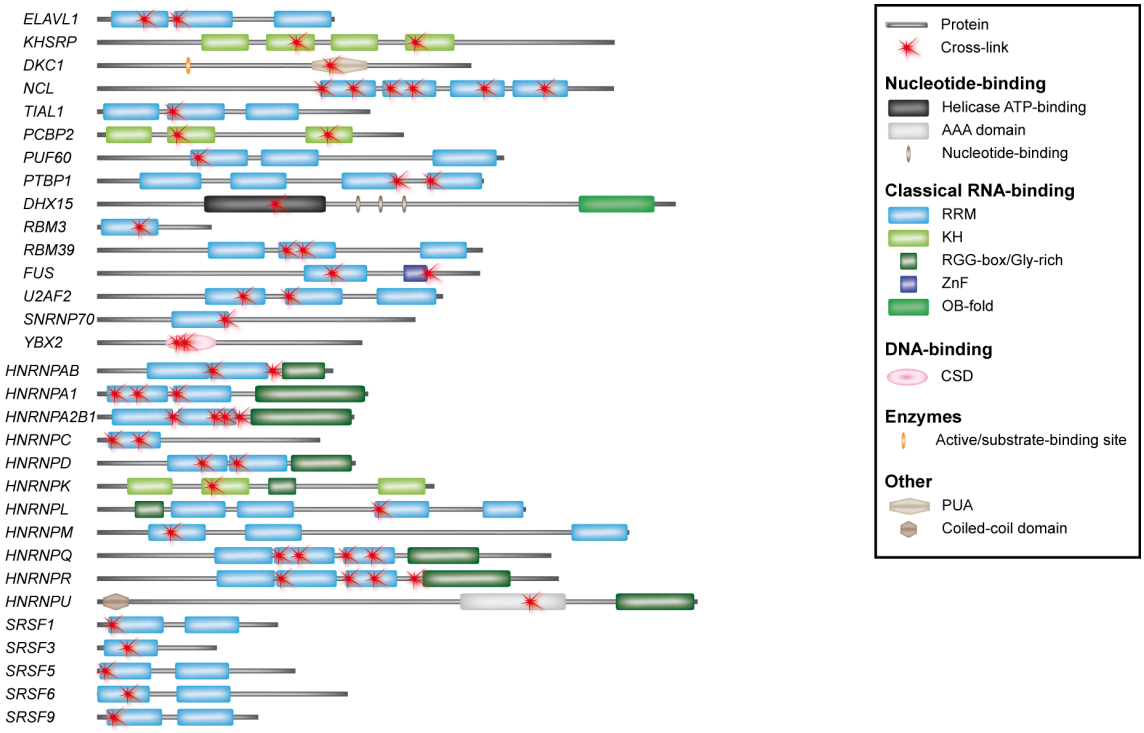


Figure 3.12: Cross-linking sites and annotated domains in human proteins (ribosomal subunits excluded). *Adapted from Kramer et al.* [69].

We identified a total of 189 cross-links on 60 tryptic peptides of 35 different proteins (Figure 3.12 and Appendix Table C.5) in human (HeLa) samples. In 79% of all cross-links, the cross-linked nucleotide and in 50% the cross-linked amino acid could be identified. Most peptides (54) lay in known RNA-binding motifs like RNA-recognition motifs (RRMs) and K Homology (KH) domain. Of the 20 annotated heterogeneous nuclear RNP proteins in the human database, our method was able to identify 13 (more than 60%) to be cross-linked. 25 distinct peptides or amino acids covered more than 55% of the 44 annotated canonical RNA-binding motifs (RRM and KH motif) in these proteins. In addition to the canonical binding motifs, we also found cross-links in the less frequently described AAA, the RanBP-type zinc finger, PUA, and coiled-coil domains.

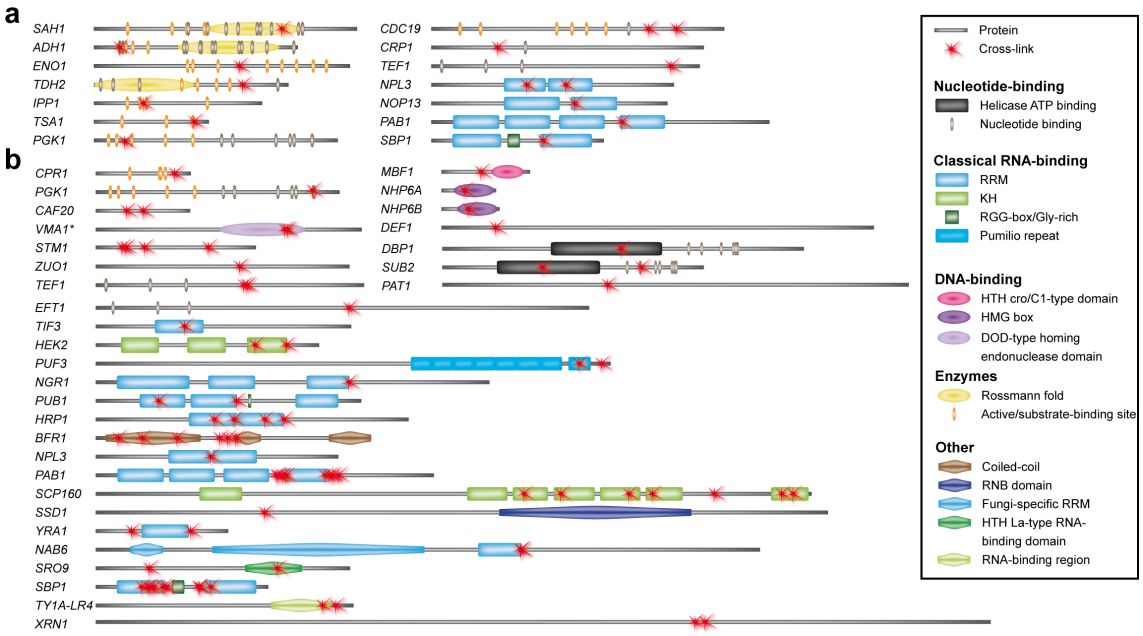


Figure 3.13: Cross-linking sites and annotated domains in yeast proteins (ribosomal subunits excluded) **(a)** yeast proteins isolated with TAP-tagged Cbp20 **(b)** yeast proteins isolated with oligo-d(T) (4SU). *Adapted from Kramer et al.* [69].

In the first yeast data set, (UV-irradiated yeast RNPs isolated by affinity purification of Cbp20), we identified 184 peptide-RNA cross-links. Cross-links on 64 tryptic peptides of 49 different proteins (Figure 3.13 and Appendix Table C.6) were identified in total. In 89% of all cross-links, the cross-linked nucleotide and in 61% the cross-linked amino acid could be identified. Most cross-links (137) mapped to ribosomal proteins (34). Other, nonribosomal proteins were well-known RBPs. Of the annotated RBPs we identified cross-links for nucleolar proteins 3 and 13 (Npl3 and Nop13), the polyadenylate-binding protein Pab1, and the single-stranded nucleic acid–binding

protein Sbp1 with cross-links located in the RRM. Interestingly, we also found cross-links in enzymes that were not annotated as RBPs: adenosylhomocysteinase Sah1, alcohol dehydrogenase Adh1, and glyceraldehyde-3-phosphate dehydrogenase Tdh2) containing a Rossmann fold[82], as well as enolase Eno1, inorganic pyrophosphatase Ipp1, peroxiredoxin Tsa1, phosphoglycerate kinase Pgk1, and pyruvate kinase Cdc19.

In the second yeast data set, (UV-irradiated yeast cells, isolation of polyadenylated mRNA) we used the 4SU nucleoside analog. Attributed to its higher reactivity, we obtained a greater number of cross-links (376) as well as a higher coverage of proteins. 161 cross-links were mapped to ribosomal and 215 to nonribosomal proteins (see Appendix Table C.7). Amongst the class of metabolic enzymes, we found peptidyl-prolyl *cis-trans* isomerase Cpr1 and phosphoglycerate kinase Pgk1. Endonuclease PI-SceI Vma1, the multiprotein bridging factor 1 Mbf1, and elongation factor 1 alpha Tef1. Interestingly, we also identified DNA-binding proteins such as nonhistone chromosomal protein NHP6A and NHP6B. Amongst the remaining RBPs, we identified putative ATP-dependent RNA helicases of the DEAD-box protein family Dbp1 and Sub2, with the latter being a homolog of the human splicing factor hUAP56. Proteins with RRMs included the Poly (A)+ RNA-binding protein Pub1. Proteins with KH (initially identified in the human hnRNP K) were the RBPs Hek2 and ligand activated Scp160. Proteins with motifs of the Pumilio-homology domain family were Puf3 (Appendix Figure C.5.e-f), a known mitochondrial surface protein that binds and promotes degradation of mRNAs encoding mitochondrial proteins. Nucleotide-binding motifs less commonly associated with RNA binding were found in the cytoplasmic RNA-binding protein Sro9 (HTA-La-type RNA-binding domain), and the polyribosomes associated RNA-binding protein Bfr1 (coiled-coil domain). Interestingly, we also identified the Transposon Ty1-LR4 Gag polyprotein Gag-p49, a retrotransposon-derived capsid protein that forms the structural component of a virus-like particle that encapsulates the retrotransposons dimeric RNA.

Precursor adducts

Figure 3.14 summarizes the identified and manually validated precursor adducts for uridine- and 4SU-containing RNA as observed in our yeast experiments. For a detailed list of precursor adducts observed from human samples, see Kramer et al.[69], Supplementary Figure 9. More than 90% of all unique cross-linking sites or regions were assigned from peptide-RNA heteroconjugates with at most two nucleotides. The preference for short oligonucleotides is a direct consequence of using endonucleases to digest RNA during sample preparation. In our experience, generation of tetranucleotides is

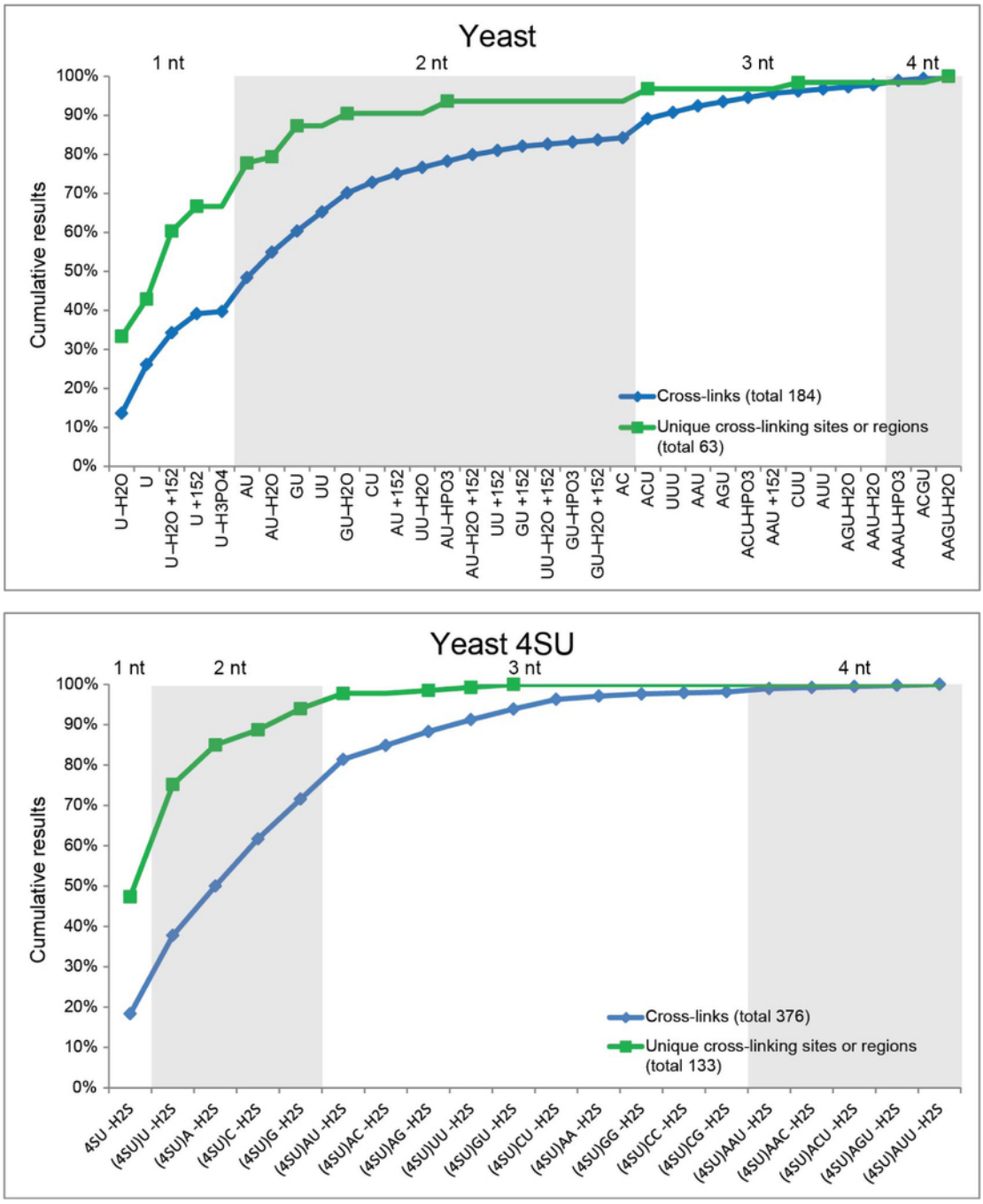


Figure 3.14: Precursor RNA adducts of yeast observed for **(top:)** RNPs isolated with TAP purification of Cbp20 (uridine) **(bottom:)** RNPs isolated with oligo d(T) (4SU). Cumulative numbers of observed precursor adducts are plotted by occurrence and ordered by number of nucleotides. *Adapted from Kramer et al.* [69].

sufficient for most types of analyses. Nevertheless, it should be noted that the localization of cross-linking sites on the RNA might be ambiguous for short, and therefore,

more likely non-unique oligonucleotide sequences. The alternative, increasing the length of oligonucleotides by omitting the digestion step with endonucleases during sample processing, has detrimental effects on the identification of cross-links. One reason is that the signal of a cross-linked peptide is distributed to an increased number of oligonucleotide variants that differ in length, composition, and losses. Another reason is that tandem spectra of longer cross-links with longer oligonucleotides are dominated by nucleotide derived fragment ions[83]. These additional ions interfere with the identification.

The vast majority of all heteroconjugates contained, at least, one uridine (>98%) or 4SU (100%). Comparison of the two cross-links obtained from both methods reveal a certain degree of complementarity: for several proteins (e.g., Npl3, Pgk1, Tef1 as well as ribosomal proteins) both methods identify different regions of the protein as cross-linked to RNA while for others (Pab1, Sbp1, and other ribosomal proteins), the same amino acids and peptides haven been identified. These findings are in accordance with previous observations[84].

Observed losses on the precursor highly depend on the cross-linked nucleotide or nucleotide analog. For uridine-containing cross-links, a wide range of losses in the precursor mass is observed. For 4SU, only loss of H_2S was observed in all cases.

Given that mainly short oligonucleotides have been found to be cross-linked, we calculated what fraction of cross-links could be possibly identified using a classic PTM search and a single nucleotide (and loss variants) as a modification. Only 14%-33% of the cross-links and 33%-82% of the cross-linking sites/regions could be identified in the three experiments.

With the exception of aspartic acid (D), asparagine (N), glutamic acid (E) and glutamine (Q) we observed every amino acid cross-linked to nucleotides (see Kramer et al.[69], Supplementary Table 1-3).

Performance

The processing time of the complete RNP^{xl} workflow typically ranges from hours to days. For a representative data set and default settings, we recorded a processing time of ≈ 83 h (single core, Intel Xeon CPU E5-2620). The vast majority of processing time, in this case, more than 99%, is spent in the OMSSA database searches invoked by the RNP^{xl} tool. Apart from compute power, disk I/O, and the number of tandem spectra, processing time mainly depends on the number of precursor variants and variable modifications considered.

3.2.3 Discussion

We presented a novel joint experimental and computational method to identify RNA-protein contact sites at the amino acid, as well as the nucleotide level. We successfully applied and validated our method on complex samples and were able to obtain additional or more detailed information on the protein-RNA interaction when compared with existing methods.

Earlier studies combined UV cross-linking and mass spectrometry to investigate moderately complex protein-RNA complexes. However, in more complex samples, especially whole-cell extracts, they only identified the entire protein but not the cross-linking site. RNA-sequencing based methods like PAR-CLIP identify cross-linked RNA and nucleotides but do not identify the cross-linked amino acid. Our method can complement these approaches by leveraging sequence information (see: sequence restriction, Section 3.2.1) obtained from these methods to more efficiently identify the cross-linked amino acid. In contrast to mutation studies, our approach is not biased towards known nucleotide-binding domains or sequence motifs. It allows discovering novel binding regions that can be interesting targets for follow-up studies - including loss-of-function mutational studies. The annotation of metabolic enzymes and transcription factors, as well as of proteins that contain multiple interaction sites, represent highly interesting research topics that merit further investigations. Recently, computational predictions of RNA-protein interactions have considerably improved. A currently open question is how strongly predictive methods are biased to known RNA-binding domains or binding motifs. In contrast, our method is, in this regards, unbiased and provides direct evidence that can potentially be used to improve predictors.

Manual validation with available 3D structures and known cross-linking sites demonstrate the specificity of the cross-linking reaction (see Appendix Figure C.5.a-f). For instance, in the 80S yeast ribosome[85], cross-linking sites were found in close proximity to the 18S or 28S rRNA. Comparison with existing annotations and 3D structures also gave rise to potential biological relevant observations. Our approach detected cross-links at Arg40 and Tyr328 of NHP6A and domain I of the homing endonuclease PI-Scel. When overlaid with the annotation and 3D structures of NHP6A in complex with dsDNA[86] or domain I of the homing endonuclease PI-Scel[87] these are located in the DNA-binding domain of these proteins with amino acids and nucleotides being in close spatial proximity. This suggests a DNA and RNA binding capability and potential role as DRBPs for these proteins that warrant further research. In the case of the receptor for activated C kinase 1 (RACK1), and in the ribosome-associated proteins Stm1[85] and Zuo1[88], we identified cross-linking sites without spatial proximity of amino

acids and nucleotides in published structures. Potential explanations are alternative conformations or involvement in mRNA binding.

The number of identified cross-links and the ability to identify so far unidentified cross-linked amino acids is a direct result of recent advancements in MS instrumentation and our RNP^{xl} workflow. Most oligonucleotide adducts and losses have not been described prior to the discovery with RNP^{xl}. Using conventional peptide database search modifications with a subset of these adducts would have identified only a fraction of cross-links. Because only a small percentage (typically about 3%) of spectra are derived from cross-links, filtering of noncross-linked spectra reduces the amount of false-positive matches which in turn improves the overall identification performance (see Kramer et. al[69], Supplementary Table 4).

Assuming a continuing improvement of mass spectrometers and sample preparation protocols we expect even more comprehensive identification of cross-linking sites in the future.

3.3 Automated Cross-Link Localization

The precursor variant approach allowed us to automatized the identification of cross-linked heteroconjugates. To further automatize the localization of the oligonucleotide on the peptide, to speed up the data processing, and to create an integrated solution independent of existing peptide search engines, we decided to develop a specialized cross-link search engine.

3.3.1 Methods

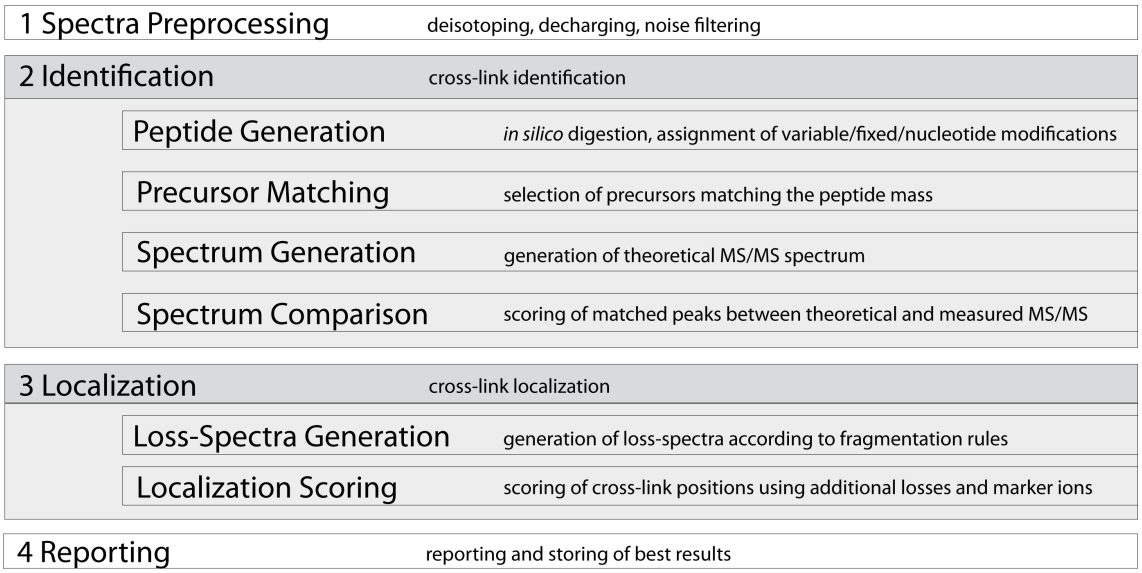


Figure 3.15: Overview: Main steps in peptide and cross-link identification and localization in RNP^{xl}Search.

Conceptually, our cross-link identification engine is an extended peptide identification engine. Many of the main processing steps in the novel engine have, thus, parallels in the majority of standard database searches. In addition to identifying noncross-linked peptides, the novel search engine provides functionality for identification and localization of cross-links. Figure 3.15 gives an overview of the main steps performed.

Step 1: Spectra Preprocessing

Comparison of theoretical and observed spectra and assignment of a match score is a central processing step in database search engines. Obtaining a PSM score that properly reflects the similarity between theoretical and observed spectrum is critical if false matches should be avoided. In practice, low-intensity detector noise, isotopic

peaks, neutral losses, and mixtures of differently charged fragments populate a tandem spectrum. Similar to existing peptide search engines[89], RNP^{xl}Search applies a spectra preprocessing step that filters tandem spectra. Ideally, only the expected fragment ion peaks (e.g., *a*-, *b*-, and *y-ions* for HCD-type fragmentation) are retained after the processing.

1. **Deisotoping:** The deisotoping filter developed for RNP^{xl}Search retains the monoisotopic peak, annotates the fragment ion charge, and removes higher-isotopic peaks from the MS/MS. To this end, m/z differences between peaks are compared whether they match the expected distance between isotopic peaks of charge three to one. If three consecutive peaks matched the expected distance within a mass tolerance window of 10 ppm, an isotope pattern is found. Because deisotoping is done in the first spectra processing step, low-intensity noise peaks are expected to be present. To reduce the impact of random matches to noise peaks, we additionally require that isotopic intensities resemble the theoretical isotope pattern of a peptide or heteroconjugate. We use an approximation of the averagine model that disallows isotopic intensities to increase after the second isotope. If an isotope pattern passed the test, fragments from higher-isotopes are removed, and the charge is annotated to the monoisotopic fragment. Fragments not part of any isotopic pattern are annotated with zero to indicate an unknown charge.

2. **Conversion to single charge:** Multiple charged fragments are converted to single charge. To this end, the mass-to-charge ratio of the single charged fragment is obtained from the multiple charged one by multiplication by the charge number z, followed by a subtraction of $(z-1)$ proton masses.

3. **Noise Filter:** The 20 highest-intensity peaks within a sliding window of m/z 100.0 are retained. This filtering step is intended to remove low-intensity signals and clusters of noise peaks. We used the `SpectraFilterWindowMower` algorithm implemented in OpenMS.

4. **Top-400 filter:** Filtering for the 400 highest-intensity peaks is merely intended to limit the maximum number of fragment peaks in pathological cases. We applied the `NLargest` algorithm implemented in OpenMS.

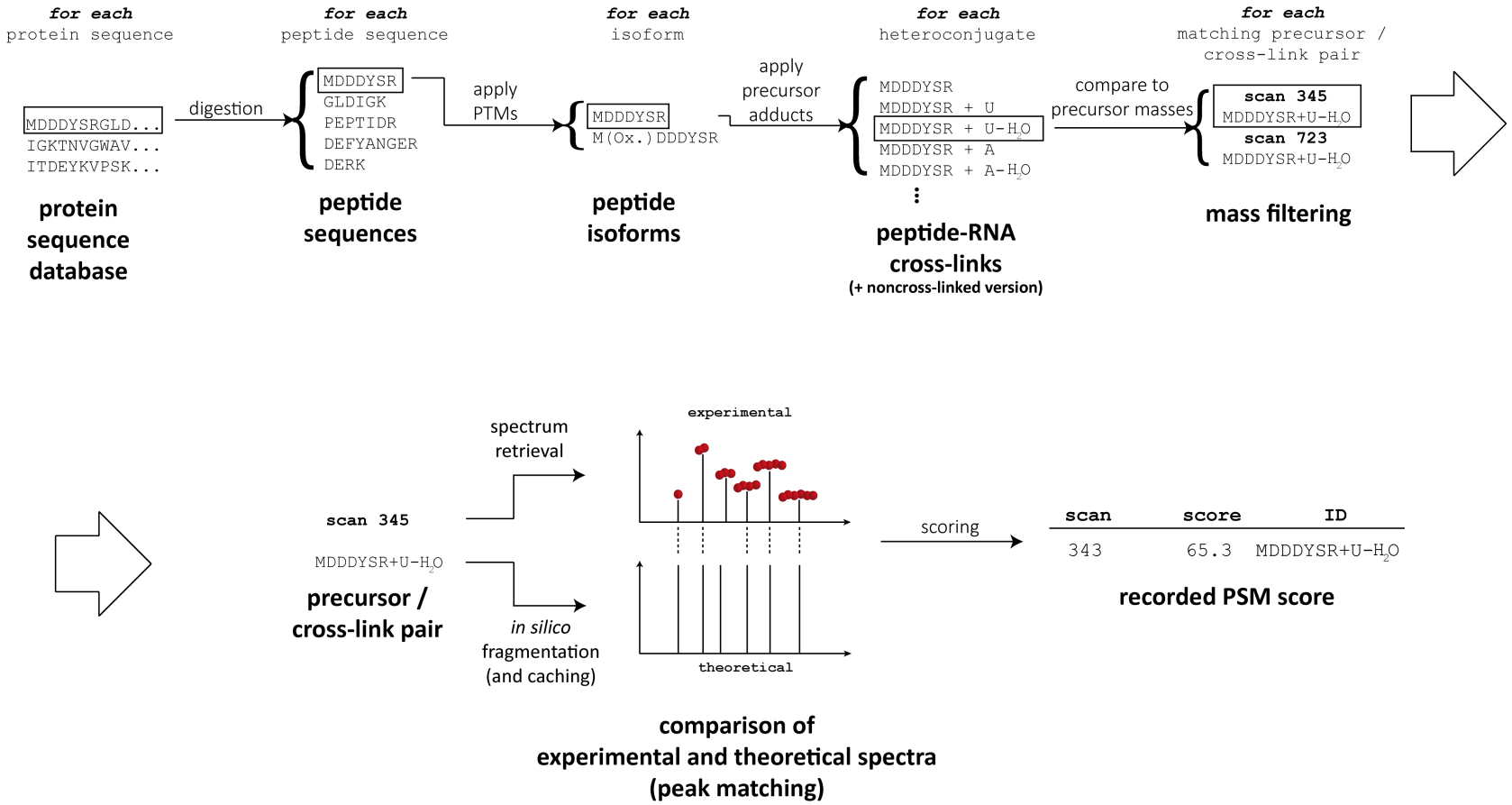


Figure 3.16: Step 2: Identification. Main loop of the peptide identification engine after spectra preprocessing. Peptide generation, mass filtering, precursor matching, spectrum generation, and spectrum comparison step.

Step 2: Identification

In the main processing loop (Figure 3.16), the protein database (in FASTA format) is digested *in silico* using user-configurable enzyme settings (see Appendix C.9 for the list of supported enzymes). The set of unique peptides is processed in parallel using OpenMP[90] as parallelization backend. For each peptide, all fixed and variable modifications are applied. In addition, and in contrast to standard database search engines, all RNA/DNA variants are created using the parameters applied in the original RNP^{xl} tool. The resulting set of masses calculated from all isoforms of a peptide (including all modifications and (oligo)nucleotide variants) is then queried against the experimental precursor masses. The precursor mass tolerance used in the matching can be configured by the user. If the mass of a peptide isoform matches an experimental precursor mass (within the specified mass tolerance window), it is a candidate peptide. In the following, we will use the term *fragment adducts* for nucleotides (with potential losses) bound to peptide fragments after fragmentation in the collision chamber.

Spectrum Generation

We generate theoretical fragment spectra of each candidate peptide using the `TheoreticalSpectrumGenerator` in OpenMS. To make the theoretical spectra compatible with the scoring function, we configured it to generate full *b-* and *y-ion* ladders but excluded unfragmented precursor ions in the MS/MS. Because a noncross-linked peptide and all of its cross-linked variants share the same fragment ions, we need to generate the shared theoretical spectrum only once. Caching and reusing it for all oligonucleotide variants effectively reduces the time spent on generating theoretical spectra.

Spectrum Comparison and Scoring

The theoretical spectrum of a candidate peptide is aligned to the measured tandem mass spectrum to determining which ions are present. For each theoretical fragment, we try to match an experimental fragment within the specified fragment mass tolerance. RNP^{xl}Search calculates the X!Tandem *HyperScore* to quantify the similarity between theoretical and experimental spectrum:

$$HyperScore = \sum_{i=1}^{n} \left(I_i \cdot P_i\right) \cdot N_b! \cdot N_y!,$$

with I_i being the measured fragment intensities and P_i is one for matched peaks (zero otherwise). N_b! and N_y! are factorials of the number of matched *b-* and *y-ions* and *n* the number of peaks in the measured spectrum. We chose the *HyperScore* in RNP^{xl}Search, because it is well established, easy to implement, and has been repeatably proven to perform well on a large range of experimental setups. Analogous to X!Tandem, log-transformed *HyperScores* are annotated (along with sequence and modification) to the matching spectrum.

Step 3: Localization

After identification, a post-scoring step is performed that aims at localizing the cross-linking site on cross-linked peptides. This step shares many similarities to modification site localization algorithms employed in standard peptide search engines. Cross-link localization relies, like post-translational modifications (PTM) localization, on the presence and absence of shifted fragment ions in the MS/MS. For example, ions of the total loss spectrum, where fragment ions never carry any fragment adduct, do not contain any information at which position the oligonucleotide was bound before fragmentation. The reason is that prefix and suffix ions corresponding to positions without nucleotides have the same mass as those that completely lost the nucleotides upon fragmentation. In the case no or partial losses occurred, some fragment ions still carry fragment adducts and a distinction is possible. Prefix and suffix ions that include the cross-linked amino acid carry an additional fragment adduct mass, while those without adduct are not shifted. In other words: while the shift in precursor mass defines which nucleotides were bound to a peptide, only the presence and absence of shifted fragment ions allow drawing conclusions on the position of the cross-link. Ideally, full mass ladders and all characteristic mass shifts can be annotated. The position of the cross-link is determined by the first shifted prefix (or suffix) ion following its unshifted version. Conceptually, PTM localization is very similar as it also leverages the information contained in the unshifted and shifted ions of mass ladders, but it differs in the following respects:

1. Typically, PTM localization does not consider mixtures of a total loss spectrum and potentially multiple partial loss spectra from different types of fragment adducts.

2. It only considers a subset of all residues (e.g., only serine, threonine and tyrosine in phosphoproteomics).

Loss-Spectra Generation

Using the RNP^{xl} tool and extensive manual annotation, we were able to gain detailed information on the fragmentation behavior of peptide-RNA heteroconjugates. A large number of additionally identified and validated spectra from follow-up experiments confirmed that fragmentation patterns of heteroconjugates are much more diverse than those of classical PTMs like phosphorylation or acetylation. Similar to other chained modifications (e.g., glycosylation) different types of fragmentation can occur. As stated before, high-intensity peaks are usually derived from fragments without nucleotide moieties. In addition, several usually lower-intensity peaks can be observed that are derived from fragments that did not completely lose the nucleotide moieties. Figure 3.17) summarizes all fragment adducts we observed for uridine-containing RNA. Fragmentation of a cross-link, thus, produces complex mixture spectra. In these mixture spectra, fragment adducts may be observed from all major fragment ion types of HCD fragmentation (*a*-, *b*-, *y*- and immonium ions).

precursor ion adducts (MS)	fragment ion adducts (observed on sequence and immonium ions) (MS/MS)						
	U	U-H_2O	U-HPO_3	U-$H_3PO_4^*$	U'	U'-H_2O	C_3O
U →	✓	✓	✓	✓	✓	✓	✓
U-H_2O →	X	✓	X	✓	✓	✓	✓
U-HPO_3 →	X	X	✓	✓	✓	✓	✓
U-$H_3PO_4^*$ →	X	X	X	✓	✓	✓	✓

Figure 3.17: Fragment adducts observed if one uridine (with potential losses) is bound to the precursor (precursor adduct). U' refers to the uracil base ($C_4H_4N_2O_2$). (*) H_3PO_4 may refer to either loss of phosphoric acid or loss of metaphosphoric acid (HPO_3) and water (H_2O) as these adducts cannot be distinguished by their elemental composition and mass.

In metabolite identification, *compositional fragmentation trees* are used to model fragments arising from compounds upon fragmentation[91]. The molecular formulas of fragments/compounds correspond to nodes in a directed acyclic graph. Labeled edges connect a parent molecule to its product. Edge labels are annotated with the losses that transform the parent's molecular formula into its product(s) formula by subtraction of the loss formula. *Compositional fragmentation trees*, thus, allow modeling which fragment ions (e.g., by specifying their composition and mass) are produced from a parent ion. We borrow from this concept to model the relation between precursor adducts and fragment adducts.

We define an *oligonucleotide adduct fragmentation tree* as a rooted tree. The root vertex is labeled with the precursor adduct and connected to fragment adducts by

directed edges. In contrast to classical fragmentation trees, *oligonucleotide adduct fragmentation trees* do not model the resulting fragment ions. Instead, each tree models which fragment adducts may arise from a particular precursor adduct (see Figure 3.18). The *oligonucleotide adduct fragmentation tree*, can, thus, be used to determines which fragment adducts (and, thus, which mass shifts) may be observed for sequence or immonium ions.

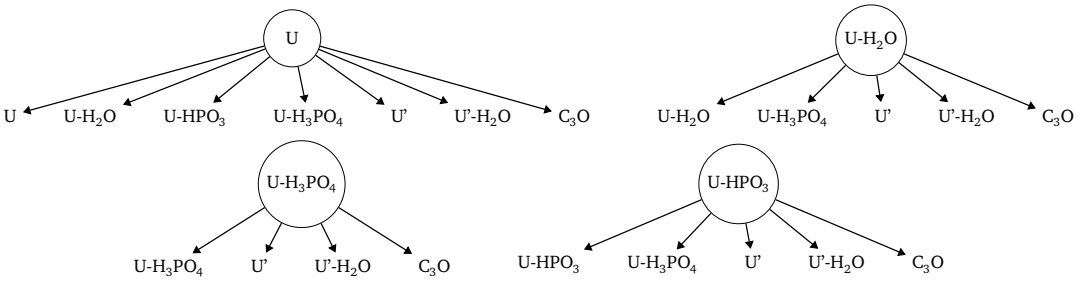


Figure 3.18: Oligonucleotide adduct fragmentation trees for uridine (with potential losses). The root of a tree (circle) is labeled with a precursor adduct and connected by edges to its fragment adducts.

So far, we mainly observed which fragment adducts are generated from precursor adducts. This partial knowledge is reflected in the topology of the tree. Our tree has no internal nodes with intermediate fragmentation products.

By encoding the oligonucleotide fragment adduct tree and using it as input for our newly developed peptide identification engine, RNP^{xl}Search, we computationally annotate fragments ions from a wide range of different precursor adducts. Based on these annotations, fragment ions without fragment adducts can be distinguished from fragment ions carrying one. The information from (potentially) multiple different fragment adducts observed or missing on sequence ions and immonium ions is then collected. This information is then used for automated localization of the cross-linked amino acid.

Localization Scoring

In UV-induced nucleotide-protein cross-linking cross-links may be formed at every amino acid. Upon fragmentation, a mixture of different fragment adducts is recorded in the tandem mass spectrum. Classical PTM localization algorithms (e.g., AScore[92]) assume that PTMs are restricted to few amino acids. In addition, classical PTM localization algorithms do not consider multiple losses and are, thus, not easily applicable to cross-link localization.

RNP^{xl}Search uses the fragmentation rules described by the empirically determined fragment adduct trees to derive feasible fragment adducts from given precursor adducts.

All fragment ions for a given precursor adduct are generated in a single theoretical spectrum and matched to the experimental spectrum. Due to the lack of a large enough training set, RNP[xl]Search employs a simple additive scoring scheme to determine a likely localization site. Conceptually, our additive scoring scheme rewards (or punishes) localization hypotheses based on the presence or absence of supporting or contradicting evidence. It makes use of information from sequence and immonium ions that carry fragment adducts.

Consider a cross-link with peptide sequence $s = (s_1, s_2, .., s_n)$, precursor adduct a and k fragment adducts $f^a = f_1, ..., f_k$. For the n possible cross-linking sites, we calculate n localization scores $\lambda = (\lambda_1, ..., \lambda_n)$. We set λ_i, the score for a specific site i, to the sum of four position-specific score components: an immonium ion score ι_i and three sequence ion scores $\alpha_i + \beta_i + \psi_i$ for *a-*, *b-* and *y-ions*, respectively:

$$\lambda_i = \alpha_i + \beta_i + \psi_i + \iota_i.$$

Immonium ions are internal fragments with a single side chain. Observation of an immonium ion with fragment adduct, thus, provides strong evidence for a cross-link at a specific amino acid. If an immonium ion is observed for a residue at position i, we increase the immonium score ι_i by the intensity of the immonium ion peak. If the peptide sequence contains multiple instances of the corresponding residue we cannot differentiate between them. In this case, we add the intensity to the immonium scores of the respective sites in question.

For sequence ion scores $\alpha_i + \beta_i + \psi_i$ the calculation differs. Consider a cross-link at the m-th amino acid. Ideally, we would observe only unshifted prefix ions up to position $m-1$ and shifted as well as unshifted (total loss) prefix ions for position m and higher.[iii] We now match the set of shifted prefix ions generated for a cross-link at position m and fragment adducts f^a to all observed fragment ions. Every fragment ion that matches to a shifted prefix ion is checked if they contradict the ideal fragmentation rule (e.g., if a shifted prefix ion was annotated before the anticipated cross-link site m). If it is in conflict with the rule, it is considered contradicting evidence for site m. If it adheres the rule, it is considered supporting evidence for site m. In practice, we cannot expect to observe all shifted ions with all possible fragment adduct shifts. We, therefore, devised ion scores for every position that consider both supporting and contradicting evidence from all fragment adducts f^a.

[iii] For suffix ions (y), the reasoning is analogous except the different indexing (first suffix ion corresponds to the position of the last amino acid in the peptide sequence) needs to be considered.

$$\alpha_i = \sum_{f \in f^a} \sum_{j=1}^{n} d(i,j)\alpha_{i,j,f}.$$

and $\alpha_{i,j,f} = w \cdot I$ the observed fragment ion intensity I.
We set:

$$w = \begin{cases} +1 & \text{, if they support the localization of the cross-link at position } i \\ -2 & \text{, otherwise.} \end{cases}$$

$\beta_i + \psi_i$ are calculated analogously.
As simple heuristic, we penalized contradicting evidence twice as much as we reward supporting evidence. The different weights are currently empirically determined and worked well on our data. In the future, we plan to perform automatic parameter selection given a larger dataset. Most information about the cross-link position is obtained from fragment ions that correspond to the cross-linking site and their direct neighbors. Fragment ions that correspond to more distant positions may still provide supporting or contradicting evidence. In the presence of noise, these peaks might interfere with the scoring. We, thus, give a linearly decreasing weight to distant evidence:

$$d(j,i) = 1 - \frac{\|(j-i)\|}{n-1} : \text{peptide} - \text{length} - \text{weighted distance.}$$

The highest-scoring localization site i^* is reported:

$$i^* = argmax_i(\lambda_i).$$

Step 4: Reporting

In the final step, identification and localization results are aggregated, and two result files similar to the output of the RNP^{xl} tool are generated. The idXML file allows for manual validation in TOPPView while the tabular file (Table 3.1) is easily opened in spreadsheet applications. In addition to the tabular output produced by RNP^{xl}, RNP^{xl}Search reports localization scores for each amino acid position in the peptide. Detailed ion annotations for the tandem mass spectrum are provided for visualization purposes (Figure 3.20):

Table 3.1: Sample output of RNP^{xl}Search (some entries shortened or omitted). Best localization is marked with lower case letter k.

RT	original m/z	proteins	RNA	peptide	charge	score	best localization(s)	...
17.788	712.7828	Cas7	U-H2O1	AEADNLDDKK	2	2.3	AEADNLDDkK	...

3.3.2 Implementation

We implemented our novel cross-link search engine as OpenMS TOPP tool (`RNP`^{xl}`Search`). Functionality common to both `RNP`^{xl} and `RNP`^{xl}`Search` tools were factored out from the `RNP`^{xl} implementation. In particular, the generation of oligonucleotides with losses and optional sequence constraints was moved into the *RNPxlModificationsGenerator* class. Both (`RNP`^{xl}) and (`RNP`^{xl}`Search`), thus, share a common code base and interface for the creation of precursor variants and precursor adducts.

Tool Parameter

Appendix Table C.8 lists the `RNP`^{xl}`Search` tool parameters. Differences in tool parameters between `RNP`^{xl} and `RNP`^{xl}`Search` are mainly related to database search options and localization functionality. Because `RNP`^{xl}`Search` is a full featured peptide identification engine, it gained additional parameters to configure the identification process (see **Precursor, Fragments, Modifications and Peptide Options** in Appendix Table C.8). Options for the creation of precursor adducts were kept mostly identical to the `RNP`^{xl} tool parameters for precursor variant generation. In addition `RNP`^{xl}`Search` gained a parameter (`fragment_adducts`) to specify fragmentation rules. Localization scoring is enabled via the parameter `localization`.

Encoding of fragment adducts in RNP^{xl}Search. RNP^{xl}Search offers two notations to encode feasible fragment adducts. The first notation is used to encode fragment adducts that may originate from every oligonucleotide adduct bound to a peptide. We only pose the restriction that fragment adducts need to part of a precursor adduct. To ensure this, the molecular formulas of fragment adducts have to be subformulas of the precursor adduct formula.

For instance, the precursor adduct $U-HPO_3$ ($C_9H_{12}N_2O_6$), may not give rise to the fragment adducts U ($C_9H_{13}N_2O_9P$) or $U-H_2O$ ($C_9H_{11}N_2O_8P$) because they would need to gain HPO_3 and O_2P_1 during fragmentation.

We encode these type of fragment adducts as a pair of molecular formula and an annotation name used in spectra visualization:

	formula		annotation
string encoding:	C4H4N2O2	,	U'

The second notation is used to encode fragment adducts that may only arise from specific precursors adducts. They allow to directly model edges between precursor and fragment adducts in complex fragmentation graphs[iv]. These fragment adducts are encoded in RNP^{xl}Search as a triplet of precursor adduct, fragment adduct molecular formula, and annotation (see Appendix Example C for an advanced use case).

	precursor adduct		formula		annotation
string encoding:	U	->	C9H13N2O9P1	,	U,
	U	->	C9H11N2O8P1	,	U-H2O,
			...		

Workflow for Cross-Link Localization

RNP^{xl}Search can, like the original RNP^{xl}, tool be flexibly combined with other OpenMS tools to build powerful analysis workflows. Figure 3.19 shows and updated version of the original RNP^{xl} workflow (see Appendix Figure C.4) with the RNP^{xl} replaced by the novel RNP^{xl}Search tool. No changes to the workflow were necessary apart from the single tool replacement and a reconfiguration of parameters.

Integration into Proteome Discoverer

Thermo Proteome Discoverer (PD) is a user-friendly and widely used commercial software for proteomics data analyses. It supports different peptide identifications engines and covers popular quantification techniques. It offers powerful visualization capabilities and a graphical GUI for workflow construction. PD is written in the C# programming language and can be extended with custom workflow nodes via a plugin mechanism. Plugins are written using PD public application programming interface (API) allowing the proteomics community to integrate own algorithms and software into PD. Today, several PD community nodes are available free of charge (e.g., the peptide identification engine MS Amanda[93] or the modification site localization tool phosphoRS[94]).
To make our method more accessible to a wider audience, we created the RNP^{xl} plugin[95]. Because PD workflows are always split into the computational expensive data-processing (*processing step*) and downstream result analysis (*consensus step*) a processing and a consensus node are provided.

[iv] Complex fragmentation graphs may, for example, arise if a mixture of standard nucleotides and nucleotide analogs (uridine (CAS 58-96-8) and 4SU (CAS 13957-31-8)) are employed that exhibit different fragmentation behavior.

Figure 3.19: OpenMS workflow for cross-Link identification and localization using the `RNP`^{xl}`Search` engine in KNIME. The workflow consists of target-decoy database creation, peak centroiding and chromatographic alignment, ID/XIC filter, and cross-link identification. Orange nodes indicate input files(s), yellow nodes TOPP tools, red nodes output files(s). Nodes are connected by edges that indicate the flow of data. Between corresponding ZipLoopStart/ZipLoopEnd nodes, a list of files is sequentially processed.

The processing node encapsulates the complete OpenMS RNP^{xl} workflow. It converts the input spectra received via the PD API into the OpenMS compatible mzML format. In addition, it is responsible for registering the workflow parameters in PD. Once registered, users can freely configure them via the graphical user interface. On workflow execution, the plugin invokes the individual OpenMS tools in their respective order, passes the corresponding workflow parameters, and handles the data flow. Once final results are written as column separated file, the RNP^{xl} node reads them back into PD internal data structures. These data structures get persisted to a SQLite-based file format via the PD internal object-relational data mapper. These results are then available for further processing in the consensus step. In the consensus step, the results are processed for visualization.

Tabular search results are displayed in the RNP^{xl} tab. Clicking on the "Show Spectrum" button opens a spectrum view with peak annotations for manual validation (Figure 3.20). The integration in PD allows executing the RNP^{xl} workflow as well

as inspecting its results in a single application. Furthermore, wrapping the complete RNP^{xl} workflow into single meta nodes hides complexity from the user.

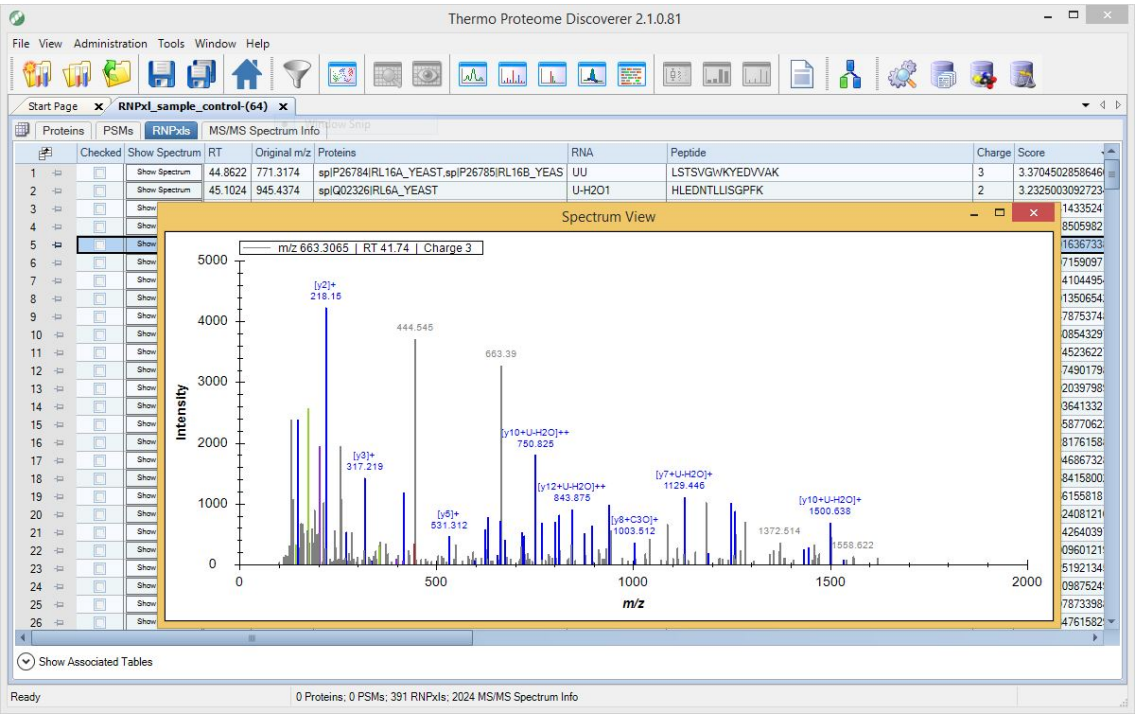


Figure 3.20: Proteome Discoverer main window showing the results and peak annotations as obtained from the RNP^{xl}Search search engine. Image courtesy of Johannes Veit who performed the integration of the RNP^{xl} workflow into PD.

3.3.3 Results

Speed Improvements

In the original RNP^{xl} tool, several hundred precursor variants that differ only in m/z annotation had to be generated for each spectrum in order to batch submit them to the OMSSA peptide database search. RNP^{xl}Search, in contrast, generates the total loss spectrum for a peptide and all of its oligonucleotide variants at most once: only if at least one mass variant matches the precursor mass of an experimental MS/MS. This leads to a drastic reduction in the number of synthesized spectra used in the initial scoring. It is, therefore, expected that RNP^{xl}Search exhibits a speedup roughly in the magnitude of precursor adducts if it is assumed that other parts of the implementation show comparable performance. We performed a search on a large orbitrap XL dataset containing approx. 30,000 spectra. RNP^{xl}Search took 0.92 h while RNP^{xl} and OMSSA

took 83.33 h resulting in a speedup factor of approx. 91. Although this is a significant speedup, it is slightly lower than expected and most likely attributed to the way OMSSA handles protein sequence databases. OMSSA performs highly optimized sequence lookups but requires additional sequence indices to be provided. These index files need to be precalculated by the NCBI makeblastdb tool[96].

Identification of Peptides

A correct implementation of all major steps in database search is crucial for a good

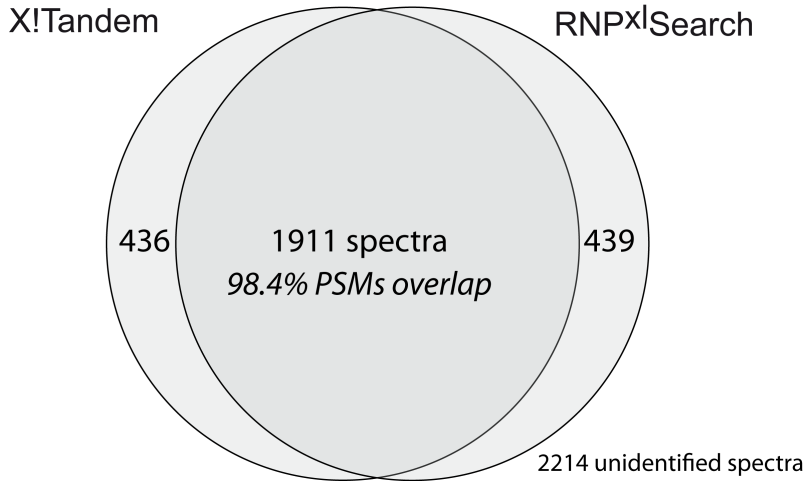


Figure 3.21: Identification performance of RNP^{xl}Search and X!Tandem on 5000 HCD spectra of a noncross-linked benchmark dataset (FDR threshold 1%). Most spectra are identified by both search engines. The same peptides were annotated by both engines in 98.4% of all PSMs (1881 of 1911).

identification performance which in turn is needed for reliable assignment of cross-linked peptides. We compared RNP^{xl}Search to X!Tandem with respect to the number of assigned PSMs (q-value < 0.01) in a human benchmark dataset[97] (5,000 HCD spectra extracted) of noncross-linked peptides. Search settings were chosen similarly for X!Tandem and RNP^{xl}Search: precursor mass tolerance set at 10 ppm, fragment mass tolerance set at 20 ppm, carbamidomethylation on cysteine (fixed modification), oxidation of methionine (variable modification), and one missed enzymatic cleavage. Both searches were performed using the same human target-decoy database containing 86,725 target proteins and their reversed version. X!Tandem (version Sledgehammer 2013.09.01.1) identified 46.94% (2347) while RNP^{xl}Search identified 47.00% (2,350) of all PSMs at an FDR of 1% (see Figure 3.21 for a graphical representation). Most spectra were identified by both RNP^{xl}Search and X!Tandem (1911). Among the spectra

identified by X!Tandem and RNP^{xl}Search, the same peptide was assigned in 98.4% of all cases. Bases on the substantial accordance in assigned peptides, we conclude that our implementation of the X!Tandem algorithm is likely without major flaws. Differences in the set of identified spectra are most likely attributed to a different peptide generation algorithm in X!Tandem. X!Tandem employs additional cutting rules (e.g., N-terminal methionine cleavage) and terminal modifications that lead to a different set of candidate peptides.

Identification of Cross-Links

The RNP^{xl} tool identifies cross-links by submitting batches of precursor mass variant to an OMSSA database search. Because the OMSSA scoring function differs from the X!Tandem scoring function used in `RNP^xl Search` differences in identification results are expected. Our main concern was, that the choice of the X!Tandem scoring function might negatively impact cross-link identification performance. In the first experiment, we compare how well `RNP^xl Search` is able to reproduce cross-link identifications determined by `RNP^xl`. In total, high-quality spectra from 61 cross-link identifications were extracted from Kramer et al.[69], Sharma et al.[76] or provided by Zaman et al., *unpublished*. Each cross-link identification was manually validated by comparing experimental and theoretical spectra in TOPPView. `RNP^xl Search` identified all 61 cross-links in the high-quality spectra. In the second experiment, we compared the identification performance of `RNP^xl Search` and `RNP^xl` on a curated dataset of 176 mixed-quality spectra (provided by A. Chernev, *unpublished*). In 75.6 % of all spectra, `RNP^xl Search` and `RNP^xl` identified the same cross-link. 9.1% were only identified by `RNP^xl Search` and 15.3% only by `RNP^xl`.

Localization Performance

To assess how accurately the cross-link position in a peptide is determined, we used the 61 manually curated high-quality cross-links described above. Manually determined localization sites were compared against those obtained from `RNP^xl Search`. 77% (47) of all cross-links (Figure 3.22) were located on the same amino acid position as determined by the expert. About 10% (6) were assigned to the neighboring amino acid. The remaining 13% (8) assignments deviated in more than one amino acid position.

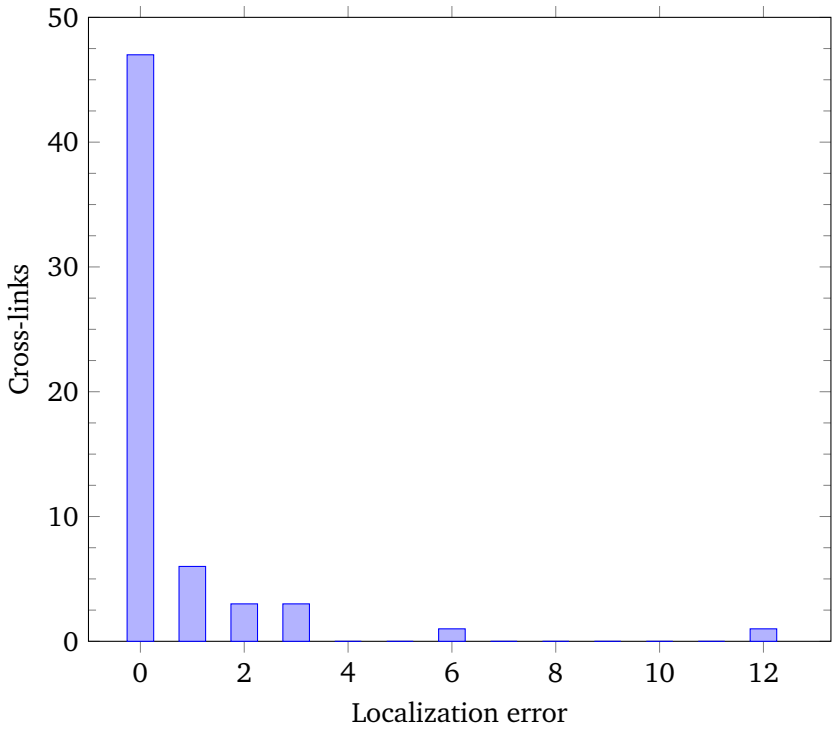


Figure 3.22: Localization error measured as the absolute difference in amino acids between automatic and manual determined cross-linking site.

3.3.4 Discussion

We presented RNP^{xl}Search, a novel tool for the automated identification and localization of UV-induced protein-nucleotide cross-links. We developed RNP^{xl}Search as a full-featured peptide identification engine. For identification of peptides, it uses a practically proven scoring function (X!Tandem *HyperScore*). It supports all OpenMS enzymes for *in silico* digestion and more than one thousand different modifications via the UniMod database. Both enzymes and modification databases can easily be extended by a user if required. We assessed the performance on a human benchmark dataset and identified approximately the same number of peptides as X!Tandem at an FDR of 1%. A different set of identified spectra between both search engines might merit further investigation. The potential differences between both implementations, though, did not result in notable differences in the number of identified spectra (2347 in X!Tandem vs. 2350 for RNP^{xl}Search). Annotations of spectra identified by X!Tandem and RNP^{xl}Search agreed in nearly all cases (98.4%).
In addition to standard database search functionality, RNP^{xl}Search includes algorithms for the identification and localization of UV-induced protein-nucleotide cross-links. Most code, originally developed for precursor generation in the RNP^{xl} tool, was reused in

RNP^{xl}Search to generate precursor adducts for cross-link identification. RNP^{xl}Search employs the X!Tandem *HyperScore* for peptide identification. On a curated dataset of high-quality spectra, RNP^{xl}Search identified all cross-links formerly identified by RNP^{xl} and manually validated by an expert. On a dataset of mixed-quality spectra identification RNP^{xl}Search performed similarly to RNP^{xl}. Regarding search speed, RNP^{xl}Search is typically about two orders of magnitude (approx. 91 times) faster. RNP^{xl}Search could also benefit from sequence indexing data structures if repeated searches are performed with the same protein database. Recently published scoring functions have been shown to perform superior to existing ones and allow deriving error probabilities for PSMs[98]. RNP^{xl}Search would also benefit from incorporation of alternative scoring functions. Novel PTM localization algorithms estimate a false-localization rate (FLR) that quantifies the confidence in a site assignment. While the most of the techniques are not easily transferable to cross-link localization, further research might allow deriving such a confidence score for RNP^{xl}Search.
RNP^{xl}Search employs a simple additive scoring scheme for localization. It generates extensive fragment annotation to score and validate the suggested cross-link position. The scoring function is able to reward (or punish) localization hypotheses based on the presence or absence of supporting or contradicting evidence. This allows making direct use of information from multiple fragment adducts. In contrast, classical PTM localization algorithms typically only account for a single loss significantly limiting their applicability to RNA/DNA cross-links. RNP^{xl}Search employs empirical weights to score supporting and contradicting evidence. In the future, we plan to automatically derive these weights from larger training datasets.
Currently, RNP^{xl} is the reference method to identify heteroconjugates, but given the various advantages of RNP^{xl}Search over our original approach, we expect a shift to the newer, specialized identification engine. To ease the transition to the novel RNP^{xl}Search tool we kept the tool interface similar to the current, state-of-the-art, implementation in RNP^{xl}. Conversion of the RNP^{xl} workflow to use the novel RNP^{xl}Search tool was achieved by a drop-in replacement of the original RNP^{xl} tool. Configuring and running workflows using generic workflow execution engine like KNIME may be considered a complex task by many lab scientists. We, thus, integrated the full workflow, including the novel RNP^{xl}Search tool, into the widely used commercial application PD. Recently, we applied RNP^{xl} to the identification of protein-DNA cross-links Flett et al.[99]. For cross-link identification, configuring the nucleotide generation in RNP^{xl} to consider deoxyribonucleoside monophosphates as building blocks was sufficient. Automated localization of DNA-peptide cross-links is an active research topic as the fragment adduct tree of DNA-peptide heteroconjugates is still to be experimentally

determined. Once completed, it can be encoded and provided as input to RNP^{xl}Search. Several approaches for quantitative analysis of cross-links are currently under investigation. Differential quantification of well-defined cross-links may employ light and heavy version of the isotopically labeled nucleotide to perform relative quantification similar to SILAC experiments. These types of experiments allow, for example, investigating the protein's affinity to different mRNAs with same nucleotide-binding motif. Taken together, RNP^{xl}Search is the first automated tool that offers a complete solution to the nucleotide cross-link identification and localization problem. We expect several new insights in the field of fragmentation chemistry, computational cross-link identification, structure biology of nucleotide-binding protein complexes, and discovery of novel protein functions that are suggested or guided by results obtained by our method.

Chapter 4

Dynamic Stable Isotope Probing of Metaproteomic Communities

Reprinted (adapted) with permission from:

MetaProSIP: automated inference of stable isotope incorporation rates in proteins for functional metaproteomics

Timo Sachsenberg[+], Florian-Alexander Herbst[+], Martin Taubert, René Kermer, Nico Jehmlich, Martin von Bergen, Jana Seifert, Oliver Kohlbacher *Journal of Proteome Research* 14(2), 619–627 (2015)

+ These authors contributed equally

4.1 Introduction

Genomic and proteomic research have been greatly expanding our understanding of the complex processes taking part in living organisms. The isolated study of a single organism, in contrast to studying multiple organisms in parallel, is often preferred and a more appropriate level of abstraction. It allows to reduce confounding factors, removes complexity from experiments and simplifies data analysis. In reality, organisms are in complex interaction with their environment. For instance, in host-pathogen interplay during infection, symbiotic processes, and the intake and metabolization of substrates in a complex microbial community. Two emerging research fields, metagenomics, and metaproteomics extend beyond single organism to multiple organisms. The more established field of metagenomics studies the genetic material of environmental or microbiome samples which are composed of a community of organisms. Heavily depending on the quality of the genetic material and existing sequence databases, metagenomics aims at identifying these organisms at the species or strain level. If

an insufficient amount of genetic information is present to unambiguously assign the identity, phylogenetic information can be leveraged to determine the most likely taxon (i.e., a higher taxonomic rank like order or family) of these organisms. To some extent, sequence-based homology studies also allow speculating on the function and biochemical repertoire of those organisms. One caveat of functional studies based on sequence homology is the lack of direct experimental evidence. Metagenomics is, therefore, best at identifying the organisms present in a biological sample, but usually unable to provide detailed functional insights. The field of metaproteomics is, at the time of this thesis, in its infancy. Its main research targets are proteomes of environmental or microbiome samples. In mass spectrometry-based metaproteomics, metagenomics-derived protein databases are typically used to identify proteins from a community of organisms. In contrast to merely identifying the organisms present in a sample, metaproteomics additionally aims at studying the interaction between microorganisms and their environment. Besides host-pathogen interaction, central topics are the degradation of substrates and nutrients - including uptake and digestion of other organisms. Simply put, metaproteomics tries to understand which organisms eat what and when.

Microorganisms vary greatly in their biochemical repertoire and pathways that allow them to decompose and metabolize different substrate molecules. Some of those most remarkable abilities include biodegradation of toxic compounds. For example, *Pseudomonas putida*, the first patented organism[100], is able to degrade organic solvents like toluene - a chemical highly toxic to a broad range of organisms. Other remarkable abilities can be found in the human microbiome. In the gut, microorganisms enable us to process many carbohydrates for which humans cannot produce the required enzymes.[101] Identifying the microorganisms and characterizing their biochemical repertoire is of great practical and commercial interest. Possible practical applications include bioremediation and waste management. Apart from environmental biology, the characterization of organisms also helps to understand clinical relevant processes in microbiomes and potentially associated diseases.

Usually, degradation of biological substrates down to simple organic compounds is a multistep process which may involve different organisms. At the top of the process, a group of organisms consumes and metabolizes an initial substrate. At some point, the organisms may not be able to further process the substrate. They dispose of molecules, which in turn may get consumed by other organisms. These organisms are able to further process the secreted molecules because they perform a different set of biochemical reactions. They form a so-called, different *functional group*. After potentially multiple steps of consumption and secretion only simplest chemical molecules are

present that are easily consumed by many organisms. At any point, an organism can also be consumed by other organisms. Elemental flux analysis techniques, track amount and flow of substrate-derived atoms between organisms and are well suited for the investigation of complex degradation processes and cross-feeding.
Tracking substrate-derived atoms in real world samples is, in general, a challenging task since a huge number of different substrate molecules are consumed by an enormous amount and variety of organisms. Most commonly, labeling of substrates with stable isotopes is employed. Because labeling with heavy isotopes changes the mass of a biomolecule, the mass change can be used as a proxy to detect incorporation of substrate-derived atoms into an organism. The methods that use this principle are called stable isotope probing (SIP) techniques and have been used to investigate microbial interaction[102,103] for more than a decade. First, organism (or a whole community of organisms) are fed with the heavy stable isotope-labeled substrates (typically ^{13}C and ^{15}N). After consumption and metabolization, biomolecules of these organisms become also isotopically labeled. These biomolecules are then extracted and analyzed using different techniques.
The different SIP techniques can be categorized by the class of biomolecule that is analyzed. RNA- or DNA-SIP investigate labeled (ribo-)nucleic acids. Fatty acid-SIP investigates labeling of phospholipid fatty acids. Nucleic acid and fatty acid-based SIP have some drawbacks: they require a high degree of labeling[104] or do not carry information that allows identifying the organisms in question[102]. This renders the application of these techniques to multiple unknown organisms difficult.
In recent years, protein-based SIP technologies have gained popularity. Central to protein-SIP is the determination of two quantities:

- The relative isotope abundance (RIA) is used to quantify to what extent isotopes from the labeled substrate were incorporated into newly synthesized proteins.

- The labeling ratio (LR) is used to characterize the speed of protein biosynthesis (protein turnover). It is the ratio of synthesized protein (labeled) to total protein abundance (labeled+unlabeled).

To accurately and sensitively determine the incorporation of stable isotopes into proteins LC-MS/MS-based quantitative analysis of peptides can be used[30,105]. In addition, identified peptide and protein sequences obtained in the analysis carry rich phylogenetic information. This information can be used to determine taxa of analyzed organisms[106].

Related Work

Previous protein-SIP approaches achieved different levels of automation in the detection and quantification of stable isotope labeled proteins. Laborious manual peak extraction combined with Excel script-assisted validation of labeled peptides and isotope patterns has been successfully applied in several studies. While this method has been shown to detect labeled peptides with high specificity, the time needed for the manual analysis (ranging up to several months) constituted the bottleneck of the metaproteomic protein-SIP study.
Published computational approaches are:

- The SIPROS software Wang et al.[107], which has been originally targeted for the analysis of protein-SIP experiments employing ^{15}N-labeled substrates. In experiments prior to the development of our approach, we were not able to detect significant numbers of ^{13}C-labeled peptides in our dataset. In ^{15}N-labeled protein-SIP experiments, the smaller (compared to ^{13}C-labeling) number of heavy isotopes considered by the algorithm still required large amounts of computational resources.

- As we later discovered Price et al.[108] independently proposed a similar decomposition algorithm for protein turnover analysis of mouse brain tissue. The calculations are performed using Matlab scripts. As a turnover-centric tool applied to single organisms, it falls short of features for metaproteomic studies.

- The SIPPER tool Slysz et al.[109] is a software for determination of RIA and labeling ratio (LR). In addition, it provides a graphical user interface for data processing and manual validation. It requires the external input of molecular formulas and feature positions of ^{13}C-labeled peptides. Labeling with other isotopes is not supported. It was published during the development of our approach.

Because metaproteomic studies typically resolve labeling states in time series experiments and with replicated measurements, a large number of MS runs (100+) need to be processed. A computational protein-SIP approach applicable to such large studies should process these runs in a fully automated fashion. Apart from detecting ^{15}N- or ^{13}C-labeled peptides our analyses require additional functionality in addition to calculation of RIAs and LRs. Organisms with different incorporation give direct evidence for different functional groups. Automated clustering according to incorporation patterns of a potentially large number of peptides suggest different functional groups and assist in data interpretation. An automatically generated quality control report is required if a large number of experiments should be manually validating. Identification of peptides

and proteins from microbiome samples is challenging. Different peptide identification engines have different strengths and weaknesses. It is, thus, beneficial to be able to choose a combination that performs best on the data. To achieve a high degree of flexibility, the tool should be compatible with all peptide identification engines supported in OpenMS and is easily integrated into complex workflows.
None of the previously described approaches fully meet our requirement for flexible and integrated metaproteomic analysis workflows. We, thus, developed a novel TOPP tool (`MetaProSIP`) and analysis workflows to fill this gap.

4.2 Methods

The following subsection gives an overview of the general Meta-Proteomics using Stable Isotope Probing (MetaProSIP) analysis workflow and the `MetaProSIP` tool. Evaluation is done in three case studies which provide details on the individual sample processing.

4.2.1 Experimental Setup

In our protein-based SIP experiments, labeled substrate is fed either *in situ* to organisms or *in vitro* to enrichment cultures (Figure 4.1**.a**). We exclusively used substrates with nitrogen or carbon atoms replaced by heavy isotopes (^{14}N by ^{15}N and ^{12}C by ^{13}C). Over time, the heavy nitrogen (or carbon) isotopes get incorporated more and more into the organisms biomolecules. As a result, the RIA increases beyond their natural isotope abundance (RIA of 1.07% for ^{13}C and 0.368% for ^{15}N) in newly synthesized proteins. These SIP-labeled proteins are heavier when compared to their unlabeled versions as they contain a higher fraction of the labeling element's heavy isotope. In the event of a time course experiment, this fraction can, for instance, increase until every protein is fully labeled. In this case, the organism only metabolizes a fully labeled substrate. Alternatively, a steady state can be reached, that may indicate a co-consumption of other, unlabeled substrates or is a direct result of only partially-labeled substrates. At different time points, proteins are extracted, digested and subjected to LC-MS/MS mass spectrometry to determine the labeling state. Optionally, a control experiment without labeling can be set up that may later be used as a reference to improve identification of labeled species. For a detailed, step-by-step protocol of the experimental procedure, the reader may consult Jehmlich and von Bergen[111].

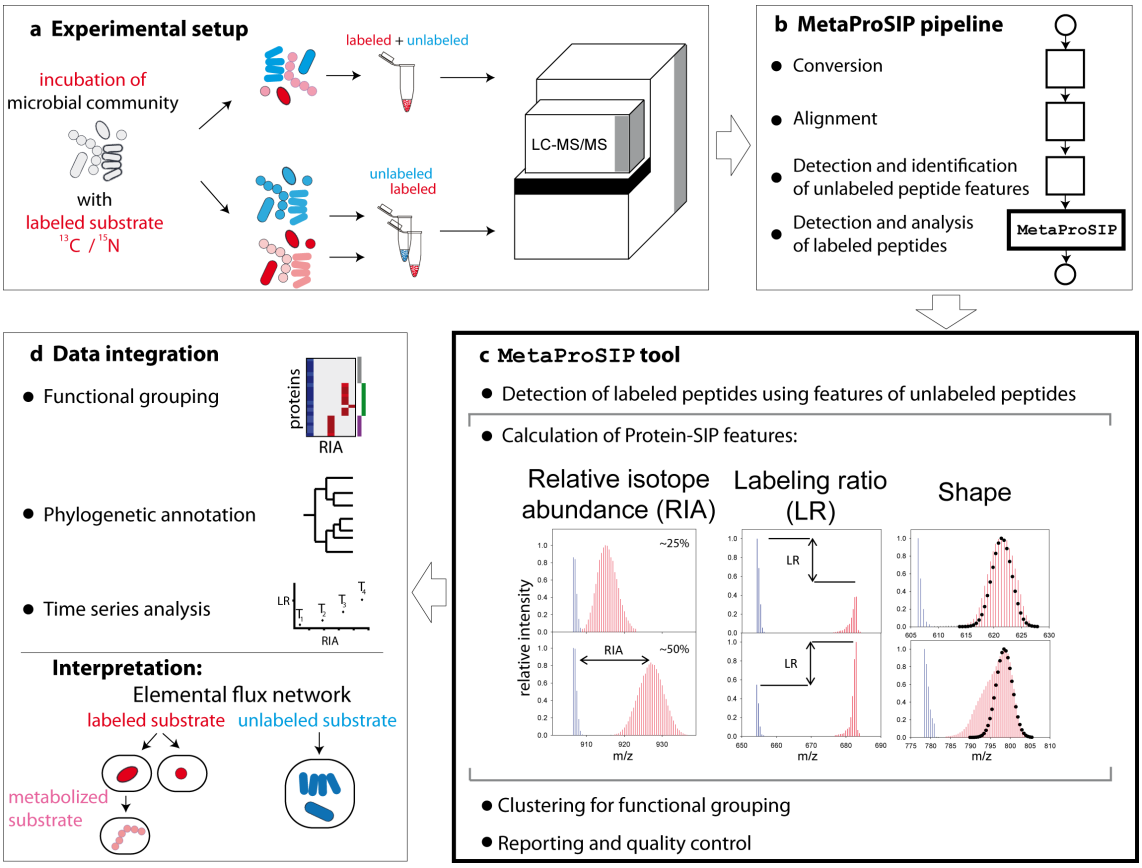


Figure 4.1: **a** Experimental setup: Stable isotope-labeled substrates are fed to microorganisms which become labeled. Optionally an unlabeled control experiment can be set up. **b** MetaProSIP pipeline: Following MS/MS data acquisition in the mass spectrometer unlabeled peptides are identified. After data conversion from the vendor format control unlabeled and labeled samples may be aligned to reduce chromatographic shifts. **c** `MetaProSIP` tool: Using retention time and mass-to-charge ratio of unlabeled peptides the `MetaProSIP` tool analyzes isotope patterns to detect incorporation of heavy isotopes. Using a decomposition algorithm, it calculates RIA and LR. Based on the number of compositions MetaProSIP can provide further information about the shape of the isotopologue distribution. **d** Data integration: Peptides and proteins may be clustered based on their incorporation behavior. Results are reported. The phylogenetic data, functional information and the calculated RIA and LR allow experts to unveil an elemental flux network. *Adapted from Sachsenberg et al.* [110].

4.2.2 MetaProSIP Pipeline

To make the acquired MS data compatible with our computational pipeline, all spectra files were converted to the open data format mzML[112] using the msconvert tool of the ProteoWizard[77] software package (version 3.0.4006). The mzML files formed the input to different workflows that were created to process and analyze the data. For

the original publication, TOPPAS[113] was used as graphical workflow editor. Workflows were later converted to KNIME to ease integration with downstream analysis tools.
In the first data processing step, MS data was subjected to signal processing to reduce the data volume. Mass spectra acquired in profile mode were centroided using the TOPP tool `PeakPickerHiRes`. When different runs (e.g., unlabeled and labeled samples) were recorded and analyzed, chromatographic shifts in retention time were corrected by aligning the experimental MS spectra using the `MapAlignerPoseClustering`[37] TOPP tool.
In the second analysis step, eluting peptides with natural isotope abundance were detected using the `FeatureFinderCentroided` tool. A database search using the freely available search engine OMSSA[48] (version 2.1.9) was used to identify peptides. Precursor mass tolerances were set to 10 ppm and to 0.5 Da for fragment masses. In the *in silico* digest of theoretical peptides used by the search engine, we allowed up to two missed cleavages. Because of the reduction agents used in sample treatment, we considered carbamidomethylation of cysteines as fixed modification. Oxidation of methionines was added as variable modification. False-discovery rates of peptide-spectrum matches were estimated using a standard target/decoy approach[114] with the database of target proteins concatenated to its sequence-reversed version (decoy proteins). The list of peptide identifications was filtered according to a q-value threshold ($q < 0.02$). Identified peptides were then mapped to features using the `IDMapper` tool in OpenMS (20 ppm m/z tolerance, 30 s retention time tolerance). The identified features formed the input to the `MetaProSIP` tool.

4.2.3 MetaProSIP Tool

The `MetaProSIP` tool calculates the relevant SIP features RIA, LR, and additional shape properties of the isotope pattern (Figure 4.1.**c**). It clusters and groups peptides according to similar incorporation behavior. Cluster provide evidence for distinct functional groups among the identified organisms. In addition, `MetaProSIP` infers proteins and produces a quality report for manual validation.

Protein-SIP Mass Spectrometry

RIA and LR are the two most important parameters of a SIP peptide in a protein-SIP analysis. As discussed in the introduction, isotope patterns of peptides with increased RIA differ significantly from unlabeled ones (Figure 4.2). The ratio of signal intensities between a labeled peptide and the total peptide intensity (LR) includes

relevant information on the protein turnover. Information on both are contained in the peaks of mass spectra and need to be reconstructed from the peak intensities.

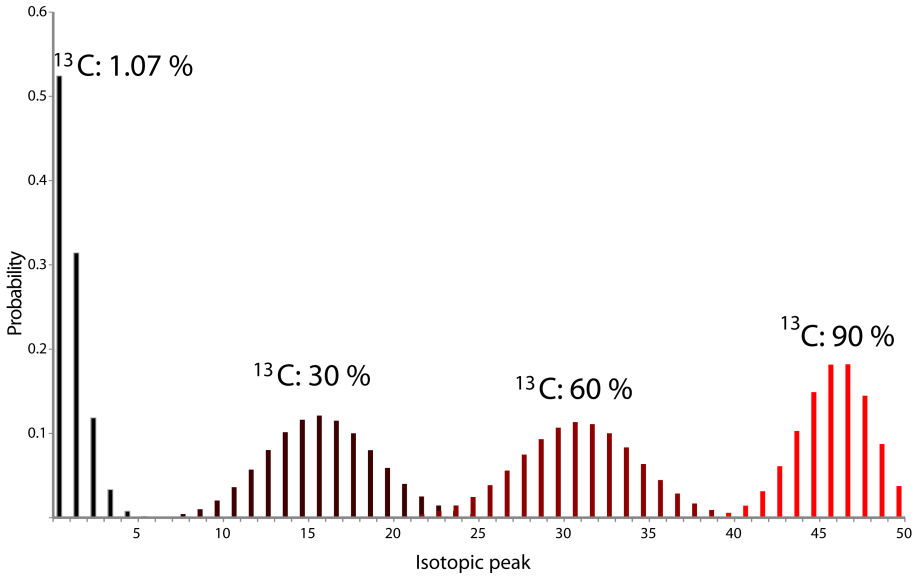


Figure 4.2: Theoretical isotope pattern calculated for a peptide (sequence TESTPEPTIDE) with varying RIAs of the heavy ^{13}C isotope. Increasing the RIA induces a shift of the isotope pattern, as well as a change in shape.

In metaproteomics, homologous peptides may be present in proteins of different organisms. Dynamic labeling of these organisms leads to a situation very different to single organism study: the same peptide may be recorded with varying degrees of labeling and abundances. This problem of mixture spectra containing signals from severally labeled species has to our knowledge not been addressed in any protein-SIP study. The central computational problem in metaproteomic SIP can, therefore, be defined as the algorithm that decomposes isotopic intensities recorded over several spectra into RIAs and associated abundances.
Given the central relevance of the problem, a more formal presentation is chosen.

Isotope Pattern Extraction

For each identified feature, the `MetaProSIP` tool calculates elemental compositions using the annotated peptide sequence. The number of atoms (n) of the labeling element (l) defines the maximum number of heavy isotopes that can be artificially introduced (i.e., yield a fully labeled peptide). Naturally occurring isotopes of other elements (e.g., heavy oxygen or sulfur) may also be incorporated in peptides. We, thus, expect that a

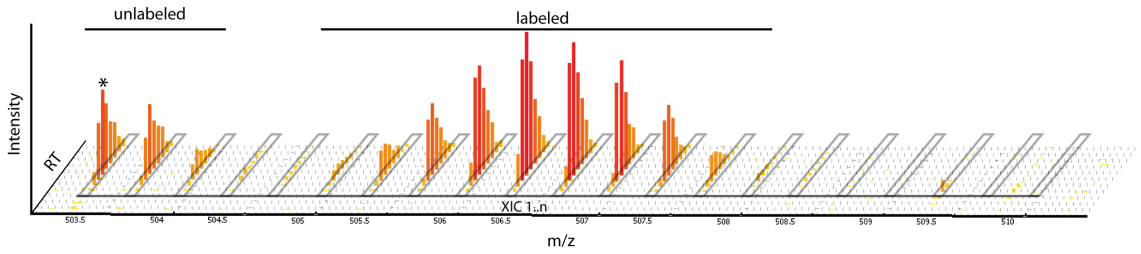


Figure 4.3: Isotope pattern extraction calculates isotopic peak intensities with extracted ion chromatograms. (*) indicates the monoisotopic peak of the unlabeled and identified feature. XICs (gray) are extracted at isotopic peak positions calculated from the elemental composition, mass, and charge of the unlabeled peptide.

small number (a) of (we consider up to five) additional isotopic peaks may be observed from fully-labeled peptides. The monoisotopic m/z of the feature (m^f), its charge (z), the maximum number of expected isotopic peaks ($n+a$) and the distance between isotopic peaks (D^l) is used to calculate the m/z positions (x_j) of the isotopic peaks:

$$x_j = m^f + j \cdot \frac{D^l}{z}, j \in \mathbb{N}_0 \wedge 0 \leq j \leq n+a,$$

where D^l corresponds to the mass difference between heavy and light isotope of the labeling element l (see Appendix Table B.1 for reference).
Centered at each isotopic peak position x_j `MetaProSIP` extracts an XIC_j (gray parallelograms in Figure 4.3). Extraction intervals in retention time and mass-to-charge dimension (Appendix Section B) were $[t^f - \Delta t, t^f + \Delta t]$ and $[x_j - \Delta m, x_j + \Delta m]$, where t^f corresponds to the retention time of the feature. Δt and Δm are user provided extraction parameters. Especially in complex samples, isotopic peak positions may overlap with mass traces of coeluting peptides or contaminants. In order to detect and remove these signals, we correlate XICs between (putative) isotopic and monoisotopic mass traces. True isotopic mass traces exhibit near perfect coelution with the monoisotopic trace resulting in Pearson correlation coefficients close to one. Signals that originate from other analytes typically differ in elution profiles leading to lower correlation coefficients. XICs with correlation above a user-specified threshold are retained while others are removed. A threshold value of 0.8 did not result in a visible loss of isotopic traces upon manual inspection and was chosen as default. `MetaProSIP` calculates for every isotopic peak a single intensity using the extracted ion chromatogram. For every XIC_i, the isotope intensity y_i is calculated as the sum of XIC intensities. For each feature, the vector of all isotope intensities $\mathbf{y} = (y_1, ..., y_{n+a})$ is used as input for the decomposition algorithm described below.

Decomposition Algorithm

Given a vector $\mathbf{y}$ of isotope intensities and a peptide sequence s. We want to approximate $\mathbf{y}$ by a linear combination of a finite set of theoretical isotope patterns $\boldsymbol{\Phi(s)}$ calculated from the elemental composition of the peptide sequence s. The coefficients of this linear combination are given by a vector $\boldsymbol{\beta}$ of non-negative weights:

$$\mathbf{y} = \boldsymbol{\Phi}(\mathbf{s})\beta$$

or

$$\begin{pmatrix} y_0 \\ \vdots \\ y_{n+a} \end{pmatrix} = \begin{bmatrix} \Phi_{0,0}(s) & \cdots & \Phi_{0,n}(s) \\ \vdots & \ddots & \vdots \\ \Phi_{n+a,0}(s) & \cdots & \Phi_{n+a,n}(s) \end{bmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix}$$

with:

n: the total number of atoms of the labeling element (e.g., carbon or nitrogen) contained in a peptide with sequence s.

a: additional isotopic traces that are collected (we chose $a = 5$).

We require the n+1 column vectors $[\boldsymbol{\Phi_0(s)}\ \boldsymbol{\Phi_1(s)}\ \cdots\ \boldsymbol{\Phi_n(s)}]$ holding the theoretical isotope patterns to be normalized to unity ($\|\boldsymbol{\Phi_0(s)}\| = 1, \cdots, \|\boldsymbol{\Phi_n(s)}\| = 1$).
As an exact solution does not usually exist because of technical variation and noise, we determine $\hat{\beta}$ to minimize the squared residual error subject to the constraint that all weights are non-negative. This leads to a standard non-negative least square formulation of our protein-SIP decomposition problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \|\boldsymbol{y} - \boldsymbol{\Phi(s)\beta}\|^2$$
$$\textit{subject to: } \beta_i \geq 0$$

To solve the optimization problem, we used a non-negative least square solver in OpenMS which is based on the FORTRAN implementation by Lawson and Hanson [115]. Figure 4.5 illustrates the decomposition of peak intensities (Figure 4.5**.a,b**) and the reconstructed signal (Figure 4.5**.c**).

Filtering of Spurious Decomposition Coefficients

Filtering of XICs by chromatographic correlation to the monoisotopic mass trace removes signals from coeluting peptides and contaminants if the elution profiles and

retention times sufficiently differ. Particularly in complex samples, it is expected that some signals from unrelated peptides, adducts, or contaminants may still be present after the XIC filtering step. Since these intensities cause spurious (non-zero) decomposition coefficients, we propose an additional filtering strategy that is based on the correlation of isotope pattern shapes. To this end, decomposition coefficients are discarded (set to zero) if less than half of the expected peaks are observed. This has been found to efficiently remove decomposition coefficients caused by isolated noise peaks. For every remaining, non-zero decomposition coefficient $\hat{\boldsymbol{\beta}}_{\boldsymbol{r}}$, we compare the observed isotope intensities to the theoretical pattern. To this end, we calculate the sample Pearson correlation coefficient $c_r(\boldsymbol{y}, \boldsymbol{\Phi_r(s)})$ between observed isotope pattern $\boldsymbol{y}$ and the theoretical isotope pattern $\boldsymbol{\Phi_r(s)}$ according to:

$$c_r(\boldsymbol{y}, \boldsymbol{\Phi_r(s)}) = \frac{\sum_{i=0}^{n+a}(y_i - \overline{y})(\Phi_{i,r}(s) - \overline{\Phi}_r(s))}{\sqrt{\sum_{i=0}^{n+a}(y_i - \overline{y})^2}\sqrt{\sum_{i=0}^{n+a}(\Phi_{i,r}(s) - \overline{\Phi}_r(s))^2}},$$

where $\overline{y}$ and $\overline{\Phi}_r(s)$ correspond to the mean of observed and theoretical isotope intensities, respectively.
Figure 4.4 shows a spectrum and the calculated decomposition and correlation coefficients.
RIAs without sufficient correlation are discarded:

$$\beta_r^* = \begin{cases} \hat{\beta}_r, & \text{if } c_r(\boldsymbol{y}, \boldsymbol{\Phi_r(s)}) > t. \\ 0, & \text{otherwise.} \end{cases}$$

Ideal isotope patterns have a c_r of one. Unrelated peaks (e.g., Figure 4.4.B) yielded consistently lower coefficients. We empirically determined a thresholding parameter t. Based on manual inspection of quality control reports, t=0.7 retained results from correct isotopic signals while it efficiently removed decomposition weights caused by coeluting peptides in our data.

Calculation of Relative Isotope Abundances and Labeling Ratios

After filtering, the solution vector $\hat{\beta}^*$ is expected to only contain the abundances of isotopologues. The i-th component $\hat{\beta}_i^*$ corresponds to a RIA of $\frac{i}{n} \cdot 100$ %. The LR is simply calculated as the ratio of labeled isotopologue abundances to the sum of all abundances.

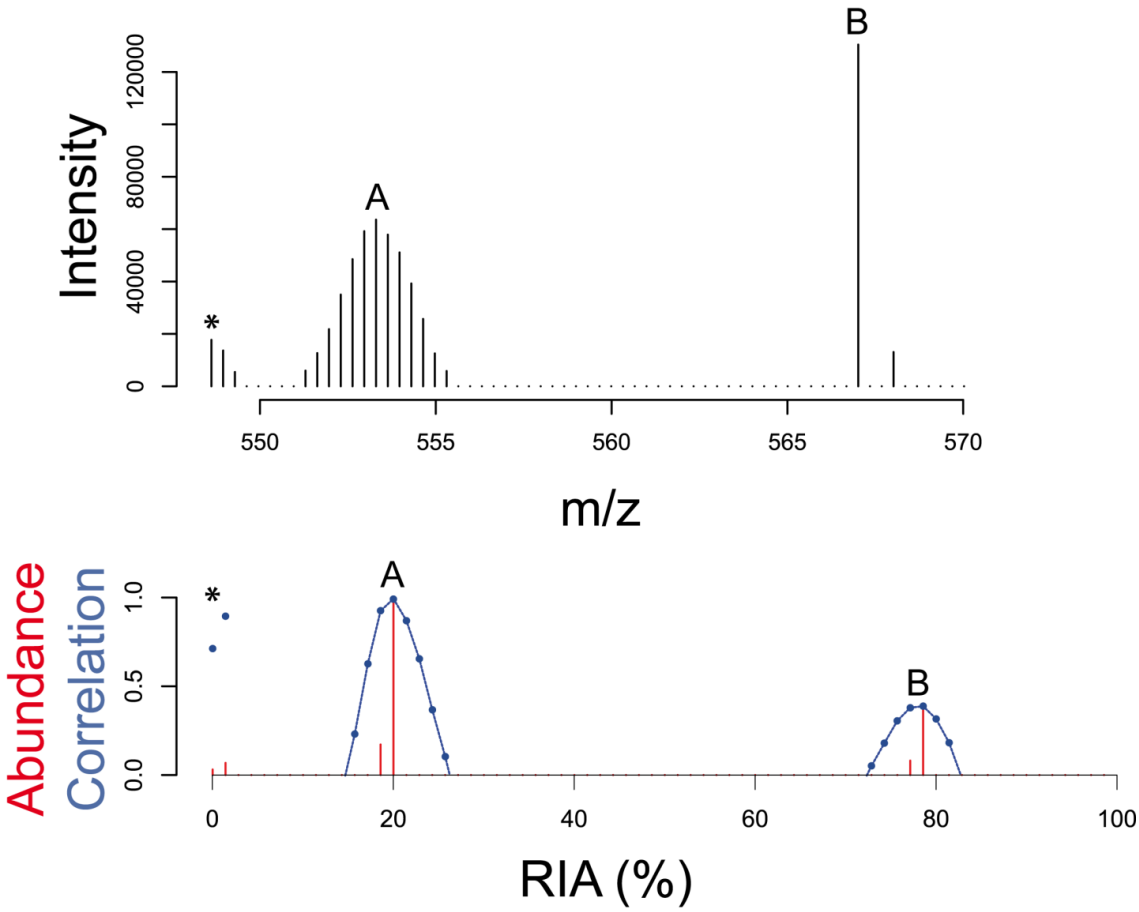


Figure 4.4: Top: Spectrum of an unlabeled peptide (*), the labeled isotopologue with RIA of approx. 20% (A), and noise peaks caused by a coeluting but otherwise unrelated peptide (B). **Bottom:** Decomposition coefficients and correlation of theoretical pattern with observed data. Noise peaks (B) have a smaller correlation compared to the labeled and unlabeled peptides because of a significant deviation from the theoretical isotope pattern. *Adapted from Sachsenberg et al.* [110].

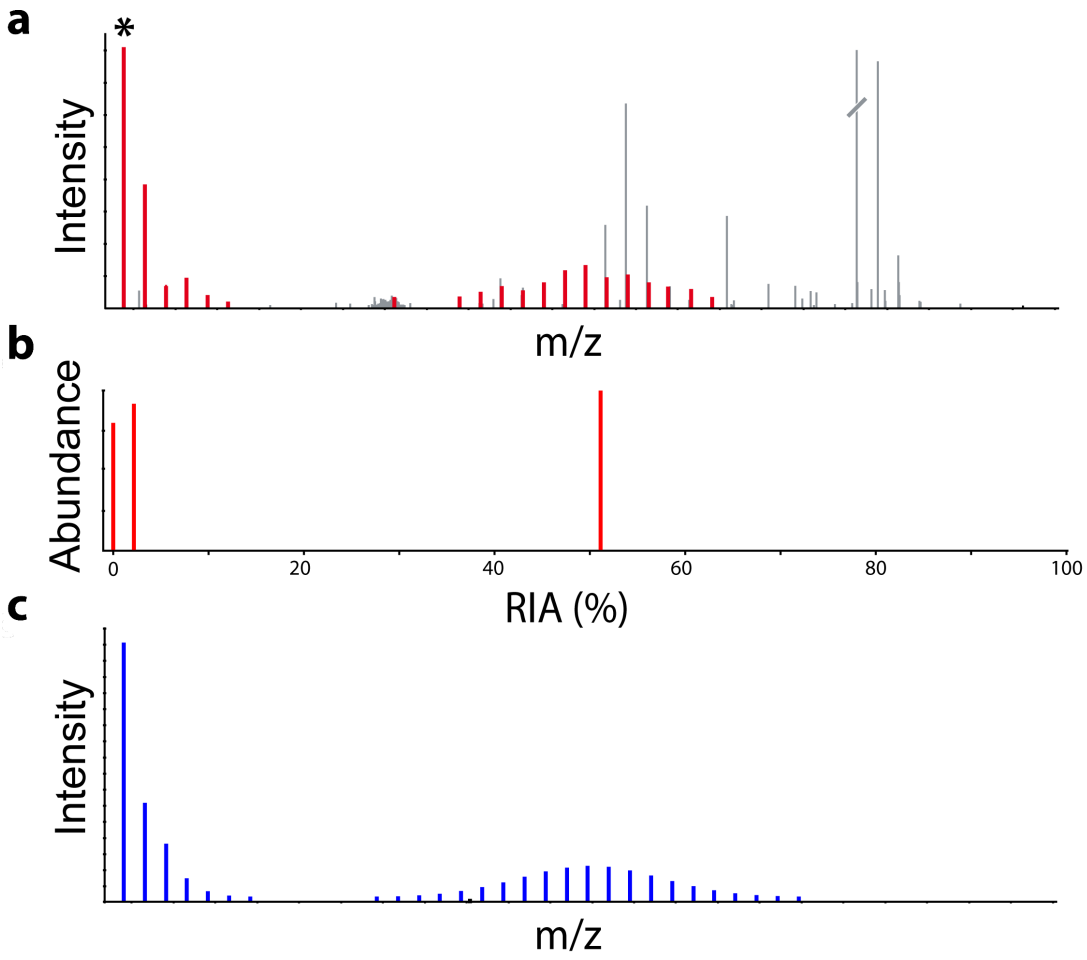


Figure 4.5: **a** MS spectrum of an unlabeled peptide and its labeled isoform. (*) indicates the monoisotopic peak. Isotopic peaks are highlighted in red. **b** Results of the decomposition algorithm. Two low RIA decomposition weights correspond to abundances of the unlabeled species. The peak at approx. 50% corresponds to the abundance of the labeled species. **c** Theoretical spectrum reconstruction as a linear combination of theoretical isotope patterns weighted by the decomposition coefficients. *Adapted from Sachsenberg et al.* [110].

Functional Grouping

To estimate the number of functional groups involved in the degradation of a supplied carbon or nitrogen source, MetaProSIP allows clustering peptides according to similar incorporation patterns. We used a density-based clustering algorithm (DBSCAN[116]) and a histogram similarity measure (FastEMD[117]) on the RIA profiles. Additionally, correlation of measured and theoretical peak shapes can also be clustered using *k*-medoids clustering (PAM[118]) and a Pearson similarity[119] measure. The number of clusters was automatically determined by the median of clusters proposed by three internal cluster validation procedures (connectivity, silhouette width, and Dunn index) as implemented in the clValid[120] R package. Once the number of clusters is selected, peptides are annotated with the cluster index for reporting.

Protein Inference

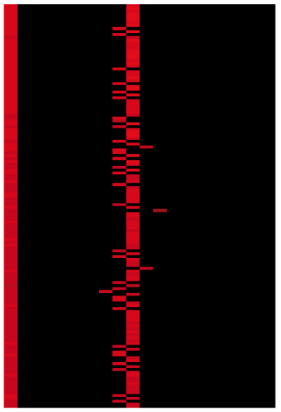
In metaproteomics, additional ambiguity in the inference of proteins by identified peptides can arise as a peptide might not only be shared by proteins of the same organism but also between organisms. We, therefore, performed a conservative protein inference and only considered proteins to be present in the sample if at least one of its unique peptides was identified. Proteins with shared peptides, for instance, peptide sequences matched to multiple proteins in the metaproteomic database are reported in a separate section in line with the shared peptide sequence.

Reporting and Quality Control

Workflow-based, high-throughput processing of mass spectrometry experiments requires concise reporting and quality control to quickly assess the outcome of an experiment. MetaProSIP generates an HTML-based report using R. Figure 4.6 shows the quality plots and tables generated by MetaProSIP:

- An overview plot and a results table for the whole experiment.
- Detailed tables and plots show extracted isotope patterns and outcome of the decomposition algorithm on the peptide and spectrum level.

a experiment level:
RIA heatmap

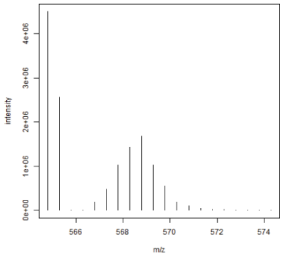


b spectrum level:
SIP peptide tables

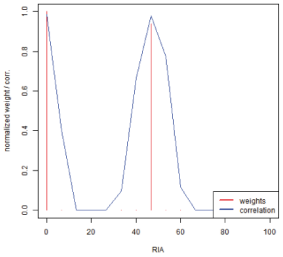
sequence	IVELNDGGVGR
rt (min.)	25.83
rt (sec.)	1549.61
mz	564.8040
theo. mz	564.8040
charge	2
accessions	tr|E2XRS7|E2XRS7_PSEFL
unique	0
search score	0
global labeling ratio	0.49
R squared	0.96

RIA1	CORR.1	INT1	RIA2	CORR.2	INT2
0.00	0.99	7796654	48.00	0.99	7329409

c spectrum level:
isotopic peaks



d spectrum level:
decomposition / correlation



e experiment level:
summary report

Group 1	# Distinct Peptides	# Unambigous Proteins	Median Global LR	median RIA 1	median RIA 2												
	279	39	0.616680413	0	46.85												
	Protein Accession	Description	# Unique Peptides	Median Global LR	median RIA 1	median RIA 2											
	SP|Q3K724|RS9_PSEPF	30S ribosomal protein S9 OS=Pseudomonas fluorescens	1	0.577454387	0	43.6											
		Peptide Sequence	RT	Exp. m/z	Theo. m/z	Charge	Score	TIC fraction	#non-natural weights		RIA 1	INT 1	Cor. 1	RIA 2	INT 2	Cor. 2	Peak intensities
		SLDTFFGR	45.6	471.7378	471.7376195	2	0.003846154	1	1		0	416796	1	43.6	734781	0.99	.....
	Protein Accession	Description	# Unique Peptides	Median Global LR	median RIA 1	median RIA 2											
	SP|Q4K3B0|ATPE_PSEF5	ATP synthase epsilon chain OS=Pseudomonas fluorescens	1	0.534933611	0	43.7											
		Peptide Sequence	RT	Exp. m/z	Theo. m/z	Charge	Score	TIC fraction	#non-natural weights		RIA 1	INT 1	Cor. 1	RIA 2	INT 2	Cor. 2	Peak intensities
		GADFDYGSAAAR	25.04	600.7678	600.7676376	2	0.012106538	1	2		0	191172	0.97	43.7	238889	0.99	.....

Figure 4.6: Main quality control items. Experiment level report summarizes results on the level of the complete MS run (**a**,**e**). Individual spectrum level assignments can be validated based on tabular information on a single SIP peptide **b**, its isotopic peaks **c** and the result of the decomposition algorithm **d**.

4.3 Results

The data has been described in detail in our original publication.

We designed three case studies aimed at evaluating MetaProSIP's performance:

- Case Study 1: Evaluates MetaProSIPs ability to quantify LR and RIA. Here we cultivated a single bacterial taxon with both natural and labeled substrates of known isotopic composition mixed in predefined ratios. We then compare RIA and LR, as determined by MetaProSIP, against ground truth values.
- Case Study 2: Feeding of organisms with a labeled substrate typically increases the LR until a steady state is reached. This might result in a partial (RIA $< 100\%$) or full labeling (RIA $= 100\%$) of the organisms. In this case study, we evaluate MetaProSIPs ability to identify peptides in a time series with increasing RIA. We demonstrate the role of unlabeled reference peptides in these type of studies.
- Case Study 3: Using a sample from a complex community of microorganisms, we demonstrate how MetaProSIP can be used to perform functional grouping and assist in reconstructing an elemental flux network.

4.3.1 Case Study 1: Performance of RIA and LR Detection

In order to reliably assess carbon and nitrogen sources as well as turnover rates, RIA and LR must be accurately determined. In the first case study, we used the `MetaProSIP` tool to automatically calculate RIA and LR in a ^{15}N-labeling experiment with *Pseudomonas fluorescens*. As the substrate, we used ammonium sulfate that either had natural ^{15}N content or 50% ^{15}N content. Cultivation was performed similarly to the protocol described in Taubert et al.[121]. Both cultures were mixed in a 3:1 and 1:3 ratio and measure in technical triplicates (in-solution digestion with trypsin, nUPLC-coupled Thermo Fisher Orbitrap XL). We processed the data using the basic MetaProSIP workflow (Appendix Figure D.1 and Table D.1).
Results are summarized in Figure 4.7. The detected RIAs showed a high degree of stability between technical replicates (1-3) and both mix ratios (3:1 and 1:3, unlabeled:labeled). The LR clearly reflect the two mix ratios but is slightly shifted to a lower-than-expected LR. A possible explanation of this deviation is an imprecision in the protein quantification using the Bradford reagent. For each mix ratio, we calculated a mean coefficient of variation for RIA and LR if a peptide was seen in at least two replicates. For the 3:1 mix we obtained 3.1% (RIA) and 7.6% (LR). In the 1:3 mix

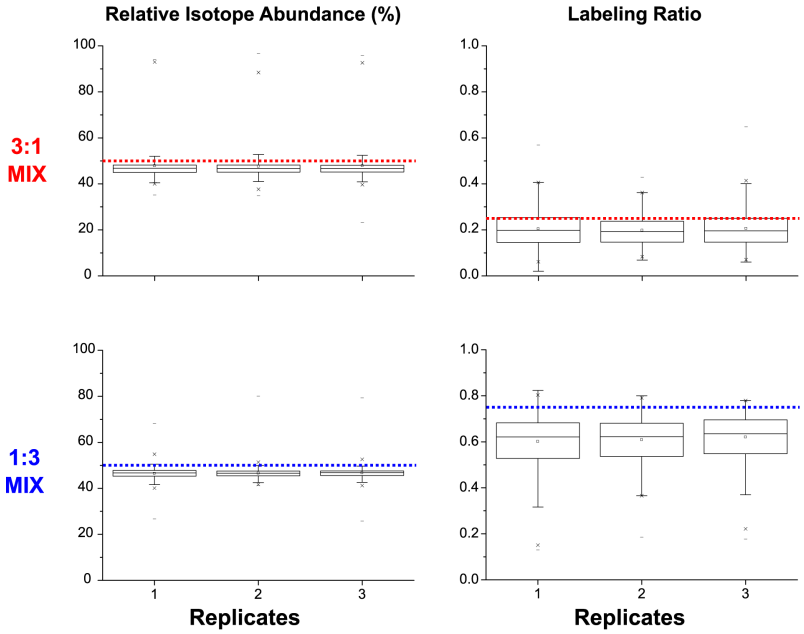


Figure 4.7: RIA and LR distributions from mixtures of 0% and 50% ^{15}N labeled *P. fluorescens* cultures. Mixing ratios were 3:1 and 1:3 unlabeled to labeled each measured in three technical replicates. The dotted line indicates the expected values. *Adapted from Sachsenberg et al.* [110].

1.4% (RIA) and 3.5% (LR). These low values for the mean coefficient of variation indicate a good repeatability between technical replicates. Variation of RIA and LR is expected to be very low for peptides of the same protein. We tested this hypothesis by calculation of median intra-protein standard deviations for the two mix ratios. In summary, we observed very low values ranging between 0.07%-1.78% for RIA and 1.23%-3.27% for LR. In total, MetaProSIP detected the labeled counterpart for 85% (3:1 mix) and 99% (1:3 mix) of all unlabeled features. While nearly all pairs of labeled and unlabeled peptides were detected in the 1:3 mix, detection rates were lower when labeled peptides were less abundant (3:1 mix).

4.3.2 Case Study 2: Identification of Labeled Peptides

In the course of a time series experiment that involves continuous feeding of organisms by a labeled substrate, more and more peptides get isotopically labeled. While the outcome of our performance study suggests that the increase of LR also positively affects the detection rate of the labeled species from an unlabeled peptide this is of course only true as long as the unlabeled peptides are abundant enough to be detected. Figure 4.8 (diamonds) clearly demonstrated that increased labeling over the time of the experiment might result in losing identifications of labeled peptides. With a decreasing

number of unlabeled reference peptides, fewer isotope patterns of labeled peptides can be extracted. We, therefore, propose an experimental setup that uses a separate dataset of unlabeled peptides as a reference. In general, several possibilities exist to generate such a dataset. The first time point of a time series, when peptides are not yet labeled, or a sample from parallel cultivation using only unlabeled substrate is generally easy to obtain.

We tested the reference dataset approach using data from an artificial mixed culture. A heterotrophic bacterium[i] (*Acidiphilium cryptum*) consumed ^{13}C-labeled galactose. The formed ^{13}C-labeled CO_2 was in turn fixed for biomass production by an autotrophic bacterium[ii] (*Acidithiobacillus ferrooxidans*) [122]. Five points in time were chosen for sample extraction during the cultivation process.

Figure 4.8 (circles) shows the result of the MetaProSIP workflow for time series analysis (see workflow in Appendix Figure D.2 and Table D.2). Here, a reference sample containing the unlabeled peptides has been measured. Instead of a decrease in labeled peptides over time (as in observed for the reference-less setup (diamonds)), a clear increase in detections are observed for the experiment with unlabeled reference samples.

Measuring reference sample and labeled sample in different MS runs imposes additional work and reagents. Therefore, it is worth discussing the advantages of this approach, compared to measuring one run containing the mixed peptides of both samples. Independent of measuring a single or separate MS runs, LR and RIA are always calculated on the signals within the MS run. Mixing of the unlabeled peptides with the labeled species prior to a single MS measurement increases the abundance of unlabeled peptides, which effectively reduces the LR of labeled peptides. In contrast, the use of reference peptide identifications from a separate sample does not alter the abundance of unlabeled peptides, which implies that the LR is not distorted. Depending on the scientific question (e.g., if only the RIA is of interest) mixing the samples might still be an adequate experimental setup.

4.3.3 Case Study 3: Functional Grouping According to Incorporation Patterns

Grouping peptides and proteins according to a phylogeny is often not sufficient to assign an ecological function to organisms. Especially, if the metagenome and annotations for the particular organisms are incomplete, additional information on an organism's

[i] A bacterium that cannot manufacture its own food from simple molecules and needs to consume organic substances from an external source.

[ii] A bacterium that produces complex organic molecules required for survival from simple molecules. Often light or inorganic reactions are used as energy source for the synthesis of these molecules.

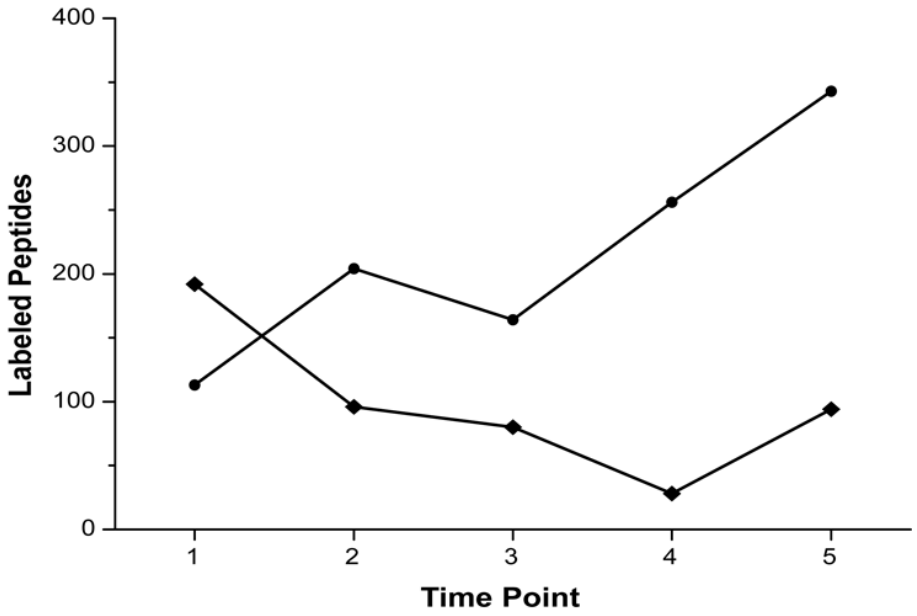


Figure 4.8: Time-course experiment with labeling substrate. If no unlabeled reference sample (diamonds) is employed, the number of detected, labeled peptides decreases as less unlabeled peptide features are detected and utilized as a reference. Adding unlabeled reference samples (circles) compensate for this effect and increase the number of labeled peptides detected by MetaProSIP. *Adapted from Sachsenberg et al.* [110].

biochemical repertoire can be deduced from its (in-)capability to metabolize certain substrates. Qualitatively grouping into active and inactive organisms (those that are able/unable to metabolize a substrate) is performed based on the presence of labeled peptides. Unlabeled peptides indicate an inactive organism. If the peptides are labeled with high RIA, the organism is likely a primary metabolizer of the substrate. Low RIA might correspond to organisms that metabolize additional substrates, including organisms that metabolize excreted molecules or scavenger organisms, that consumed biomass of labeled organisms.

Taubert et al. [123] investigated anaerobic benzene degradation using a time series with ^{13}C-labeled CO_2 and benzene substrates. Manual analysis of RIAs distinguished three active but functionally different groups of microorganisms. Benzene is first anaerobically degraded by organisms of the *Clostridium* genus. Subsequently, *Desulfobacteria* consume the formed metabolites for biomass production. The third group of organisms was suspected to be not directly involved in substrate degradation but instead are scavengers which feed on dead cell mass of other bacteria. We applied the basic MetaProSIP workflow (Appendix Figure D.1 and Table D.1) to this previously manually analyzed dataset. Figure 4.9**.a** is based on the automatically generated heatmap. The clustering performed in `MetaProSIP` clearly reveals three groups of peptides with dis-

tinct RIAs. Each of them corresponding to a different incorporation behavior. To verify that these groups indeed reproduce the three reported groups from Taubert et al. [123] we annotated the peptides using BlastP [96] and assigned phylogenetic taxa using MEGAN [124] (Figure 4.9.**b**). In concordance with the results from Taubert et al. [123] the high RIA group was annotated to predominantly originate from *Clostridiales*. *Deltaproteobacteria* were also identified for the medium RIA group. The low RIA group was, as expected and previously reported, more heterogeneous since dead cell biomass can be metabolized by a range of bacterial taxa.

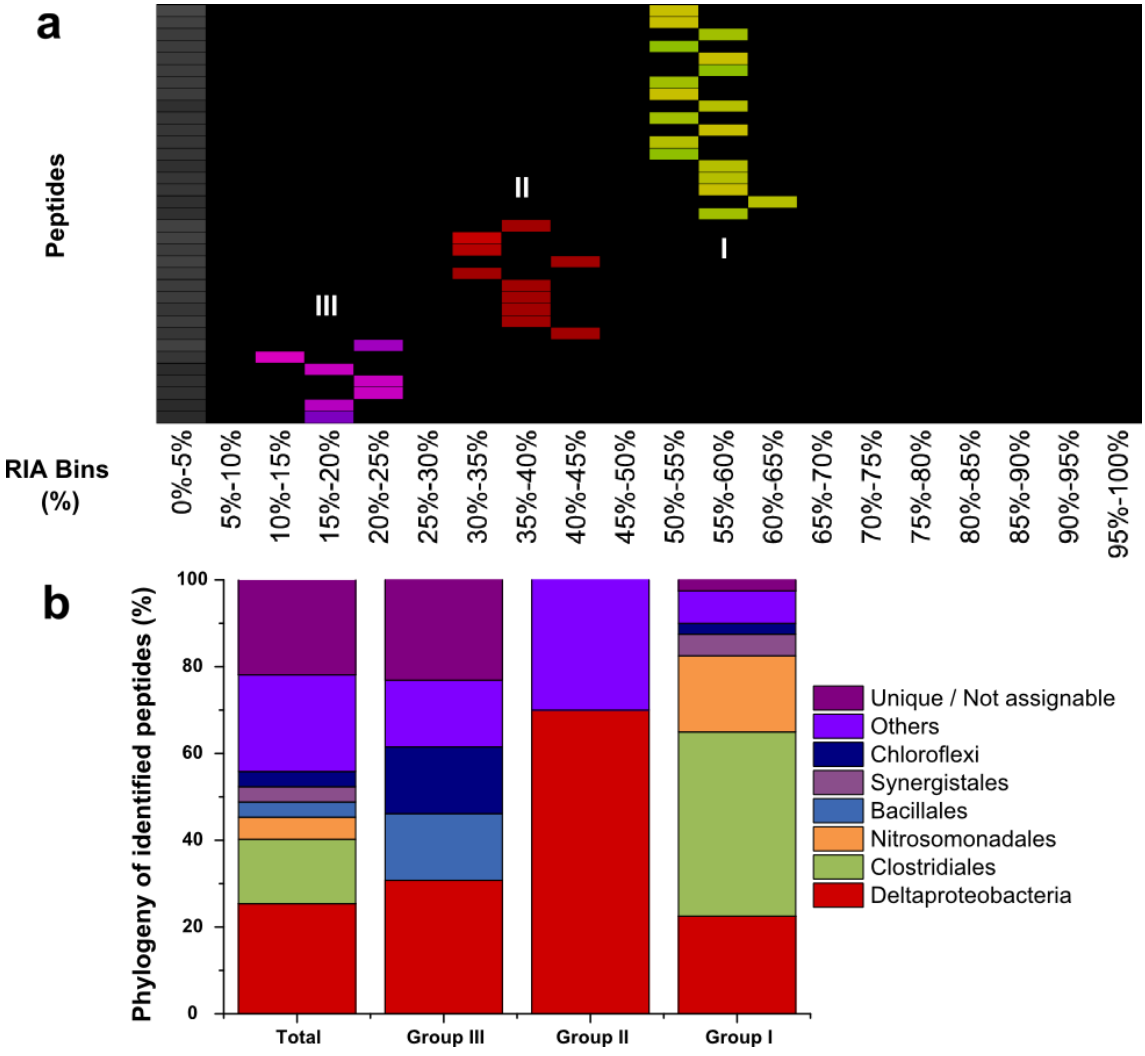


Figure 4.9: **a** color-coded heatmap showing the three RIA groups as determined by `MetaProSIP`. **b** Annotation of groups with phylogenetic information reveals a distinct composition of microorganisms. Group I is clearly dominated by *Clostridiales*, group II by *Deltaproteobacteria* while group III displays a more heterogeneous composition of phylogenetic taxa. *Adapted from Sachsenberg et al.* [110].

Taubert et al. showed that protein-SIP allows tracing of elemental fluxes between two time points.

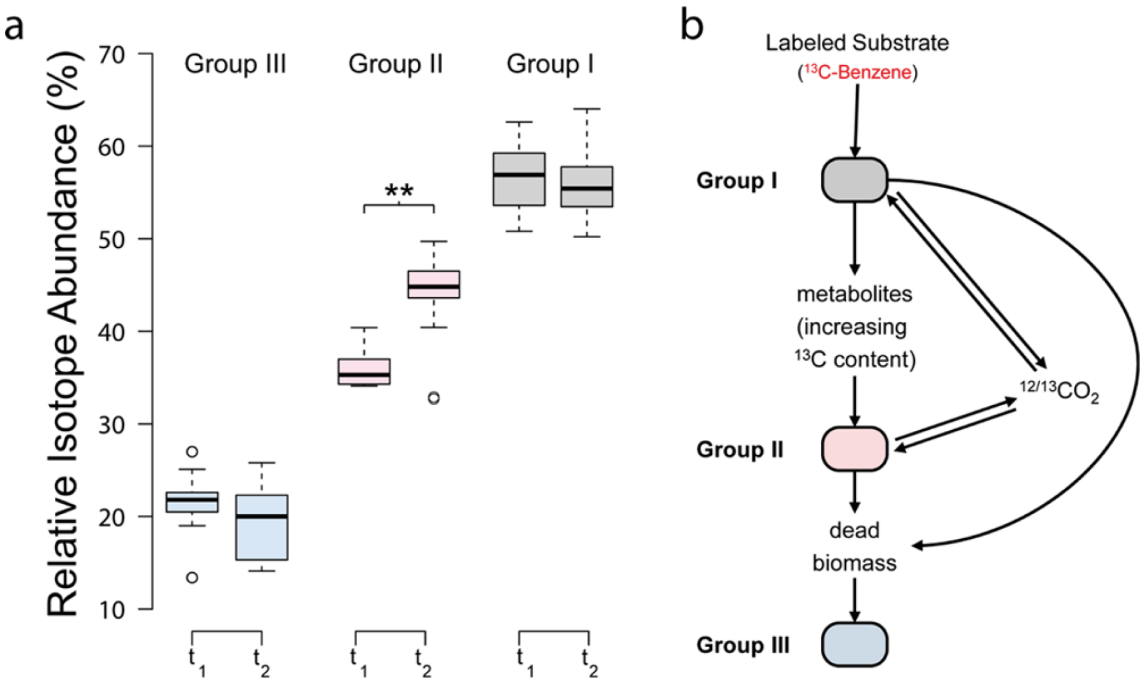


Figure 4.10: a Three groups, (*high*, *medium* and *low* RIA) have been detected and analyzed with MetaProSIP at two time points. Median RIA of group II differs significantly between t_1 and t_2. **b** Phylogenetic annotation and biological interpretation allow reconstructing the elemental flux. *Adapted from Sachsenberg et al.* [110].

Figure 4.10.**a** displays the RIA distribution obtained for the three groups at two different time points ($t_1 = 180$ d, $t_2 = 300$ d). While group I and III RIAs are relatively stable, group II shows a significant increase in RIA ($p < 0.65{\cdot}10^{-9}$) and LR ($p < 0.61{\cdot}10^{-3}$, LR not shown) confirmed by a two-tailed, heteroscedastic t-test. While an increase in LR can be explained by protein turnover and growth, the increase in RIA can be explained by the mode of substrate metabolism: An increase of RIA in group II indicates, that the ^{13}C content of its substrate pool is increased over time. As the labeling of the externally provided substrate is constant, the increase is likely a result of group II organisms consuming metabolites from group I or III. Because group I has the highest level of RIA and the phylogenetic annotation (*Clostridiales*-like) identifies them as organisms able to break the benzene ring, these are likely at the top of the degradation hierarchy. Based on their phylogenetic annotation, group III are mainly composed of scavenger organisms that consume dead biomass of group I and II organisms. They are therefore at the bottom of the degradation hierarchy with group II (*Deltaproteobacteria*-like) taking an intermediate position. Combining this information with biological knowledge

(e.g., group I and II are known to release and fixate CO_2), parts of the elemental flux network can be hypothesized (Figure 4.10.**b**). A working hypothesis could be that in the elemental flux network, ^{13}C-benzene is initially degraded by group I organisms. Release and fixation of labeled CO_2 between group I and group II yield an increase of RIA in group II. In addition, metabolites from group I (e.g., acetate) with increasing ^{13}C content, are also incorporated in organisms of group II. Group III mainly show an increase in RIA because they are composed of scavengers that feed on dead (and potentially labeled) organisms. In summary, we demonstrated that functional grouping based on incorporation behavior is indeed feasible. The biological interpretation of used substrate, LR and RIA shifts between organisms and time allows hypothesizing parts of the elemental flux network which may be confirmed using additional experiments. In our experiments, we only used two time points. More time points, different substrates (e.g., labeled acetate or CO_2) may be possible follow-up experiments to further resolve the elemental flux.

4.3.4 False-Positive Rate Estimation of Labeled Peptides

Similar to false peptide identifications in a proteomic database search, incorrect assignments of isotope-labeled peptides may occur to some extent. A high abundance of these incorrectly assigned peptides might lead to the wrong biological interpretation. It is, therefore, important to estimate the expected number of wrong assignments. The false positive rate (FPR) captures the specificity of the assignment in a statistical quantity. It is defined as the ratio of false positives to all negatives (false positives and true negatives):

$$FPR = \frac{FP}{FP + TN}$$

For protein-SIP experiments, we propose a simple procedure to estimate the FPR of labeled peptides. To clearly distinguish this quantity from commonly used terms of peptide or protein level FDR we termed the FPR of labeled peptides *incorporation FPR (iFPR)*. Contaminant peptides like trypsin or keratins are regularly present and identified in microbiological samples. These originate from external sources and are not part of the labeled organisms. Unlabeled contaminant peptides have a natural RIA. Their isotope patterns, thus, differ from labeled peptides with increased RIA and can be differentiated. The iFPR can, thus, be estimated as the fraction of contaminant peptides with (erroneously) detected stable isotope incorporation (FP) amongst all detected contaminant peptides ($FP + TN$). We estimated the iFPR for time point ($t_1 = 180$ d). In total, 287 unique contaminant peptides were identified and analyzed for incorporation analogously to the non-contaminant peptides. 266 of 287 peptides

were assigned to features, and only four were detected as labeled (RIA above 1.3% ^{13}C). This corresponds to a low iFPR of approximately 1.5%. Apart of common peptide contaminants, externally added peptides can also be employed and might be necessary if the number of contaminants is too low for confident iFPR estimation. It should be noted that the accuracy of iFPR estimation depends on the chromatographic distribution of contaminants and non-contaminant peptides. Only if these are roughly equal, the iFPR determined on the contaminants is expected to match the iFPR of non-contaminant peptides. In our experiment, chromatographic distribution of contaminants covered most parts in elution time. Comparable to FDR calculations, the quantitativeness of iFPR values should be taken with a grain of salt. Independent of this, it allows to easily spot errors in the experimental setup or data quality and is to our knowledge the first statistical measure to quantify the specificity of labeled peptide detection in protein-SIP experiments.

4.4 Discussion

Protein-SIP experiments have been successfully applied in metaproteomic studies (see von Bergen et al.[106] for a review). Being a rather novel method, it is not surprising that suitable analysis tools for protein-SIP data have only recently been developed. These novel bioinformatic tools make the data analysis part - usually the bottleneck of protein-SIP experiments - more accessible. MetaProSIP fills the important gap of providing a highly customized and automated computational pipeline for both ^{13}C and ^{15}N stable isotope labeling. We have shown that RIA can be reproducibly determined by the `MetaProSIP` tool allowing to distinguish organisms with different carbon or nitrogen sources. Additionally, the LR can be determined to measure the protein turnover.

Because metaproteomic samples can significantly differ in complexity, we developed a simple approach to estimate the false-positive rate in SIP labeling experiments. We use the identification of common contaminants or spike-in peptides. As these peptides have not been subjected to stable isotope labeling erroneously assigned incorporation events are easily detected as false positives. The estimated false-positive rate of labeled peptides assists in assessing the overall detecting problems in the experiment or workflow parameters. A low iFPR reassured that `MetaProSIP` reliably distinguishes labeled from unlabeled peptides.

Comparison of the automated data processing in MetaProSIP to a manual analysis exposes quite plainly the methodological advancement that has been achieved. Simple script-assisted, manual analysis of the data by Taubert et al. required several months.

In contrast, MetaProSIP allows processing the data and determining LR and RIA in minutes. This massive reduction in processing time shifts the overall effort required for a protein-SIP study from data generation to the biological interpretation of data analysis results. MetaProSIP, thus, enables high-throughput protein-SIP experiments in metaproteomic disciplines like environmental biology or microbiome analysis.

4.4.1 Comparison to other SIP Techniques

Studying the interaction with non-protein-based SIP approaches has several disadvantages when compared to our protein-based SIP approach. The extent of incorporation can, in most cases, only be estimated to a limited degree or no information on the biomass turnover is obtained. Protein-SIP, on the other hand, allows for accurate determination of RIA and LR in time-resolved experiments. This information allows tracing the elemental flux to a detail not obtained by other SIP approaches. Results from MetaProSIP, enriched by phylogenetic information from the peptide and protein identification, provide insights into the elemental flux and interaction within a community. Ideally, this enables determining functional groups of organisms with a distinct biochemical repertoire.

4.4.2 Comparison to other Computational Protein-SIP Methods

In this work, we presented a novel experimental and computational approach for working with labeled and unlabeled samples in parallel. Besides the parallel analysis of control and treated samples, there are many details to MetaProSIP that provide significant added value. In contrast to Slysz et al.[109] or Wang et al.[107] we calculate an optimal decomposition into RIA and LR (similar to Price et al.[108]). Additionally, we allow for detection of the full RIA distribution of a peptide. This feature is not supported for the turnover-centric workflow by Price et al.[108]. In contrast to Price et al.[108], Slysz et al.[109] and Wang et al.[107], we are also able to detect broad RIA distributions as they may occur in yet poorly understood cross-feeding processes. On top of that, we integrate clustering by incorporation behavior to detect functional groups, a feature not integrated by the other contestants. MetaProSIP provides clustering of incorporation data in order to classify peptide identifications according to the same ^{13}C- or ^{15}N-incorporation. Furthermore, we propose the - to our knowledge first - approach for the estimation of false-positive rates in protein-SIP experiments. None of the previously published tools offers the flexibility, the degree of automation, and the methodical completeness of MetaProSIP. Its integration into user-customizable,

graphical workflows sets it clearly apart from the three previous published, related works by Wang et al. [107], Price et al. [108], Slysz et al. [109].
Especially the possibility to integrate MetaProSIP into KNIME workflows opens the way for constructing powerful downstream processing and statistical analysis pipelines required in complex metaproteomic studies. Taken together, we are confident that MetaProSIP offers a clear methodological advancement in the amount of flexibility and supported features allow for much broader experimental setups then the typical targeting of software or scripts to specific case studies. MetaProSIP offers, arguably, the to date most complete computational solution for protein-SIP experiments.

4.5 Outlook

Currently, `MetaProSIP` is being extended to support additional labeling elements (including ^{18}O and ^{2}D), which will allow to use additional substrates and investigate different biochemical pathways. While `MetaProSIP` has been primarily developed for metaproteomic studies, it might also be interesting, and of little effort, evaluating how well it performs in complex protein turnover studies of a single organism. To achieve the highest level of automation, additional tools for data pretreatment and downstream analysis need to be incorporated into a single workflow. Our workflows currently lack third-party tools for metagenome assembly and metaproteomic database generation. In many cases, additional information needs to be integrated to obtain conclusive biological interpretations of the data. Further automation of downstream analysis tasks, for example, the phylogenetic assignment and functional annotation (including mapping of proteins to known pathways) could be achieved by integrating additional tools as third-party nodes in KNIME. Another avenue of research could be opened by wrapping tools from other SIP approaches for combined SIP analyses.

Chapter 5

Standardized Reporting of Experimental Results in Proteomic and Metabolomic Studies

The content of this chapter is to the most extent part of the manuscript:

The mzTab Data Exchange Format: communicating MS-based proteomics and metabolomics experimental results to a wider audience

Johannes Griss+, Andrew R. Jones+, Timo Sachsenberg, Mathias Walzer, Laurent Gatto, Jürgen Hartler, Gerhard G. Thallinger, Reza M. Salek, Christoph Steinbeck, Nadin Neuhauser, Jürgen Cox, Steffen Neumann, Jun Fan, Florian Resinger, Qing-Wei Xu, Noemí del Toro, Yasset Pérez-Riverol, Fawaz Ghali, Nuno Bandeira, Ioannis Xenarios, Oliver Kohlbacher, Juan Antonio Vizcaíno, and Henning Hermjakob

Molecular & Cellular Proteomics 13 (10) 2765-2775 (2014)

+ These authors contributed equally

---

and supplementary material.

5.1 Introduction

Communication of experimental results is at the heart of every scientific discipline. Particularly in the context of high-throughput experiments many challenges arise in transforming the vast amount of raw data into concise and meaningful results (see Figure 5.1 for a conceptual overview). Complex data processing steps typically reduce a large amount of raw data down to simpler quantitative entities like expression values of individual genes or proteins. These intermediate results are then used in a downstream statistical analysis. Biological interpretation often complements the statistical analysis, and the final study result (e.g., the set of proteins differing between healthy patients

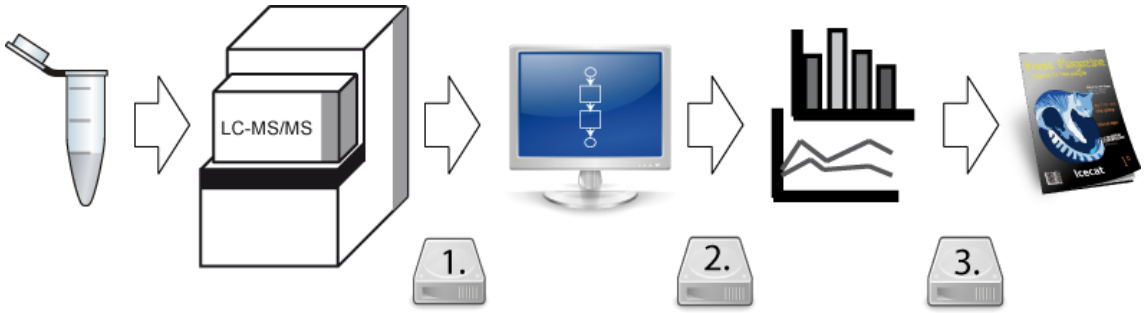


Figure 5.1: Standard scientific workflow Samples are acquired and measured using a high-throughput technology. (Semi-) automated data processing produce intermediate results which get downstream processed and statistical analyzed. Biological interpretation produces final results. During the publication process, data is deposited in public repositories.

and those with a medical condition) are reported. Today, more and more scientific journals enforce raw data deposition parallel to reporting proteomic and metabolomic study results. Unfortunately, the complete computational data processing and statistical analysis workflows are usually not reported in MS-based studies. Intermediate results, prior to statistical analysis, are not available and, consequently, valuable information for reanalysis or follow-up studies is lost.

In recent years, the Proteomics Standards Initiative (PSI) of the HUuman Proteome Organisation (HUPO) set out to solve the data sharing and archiving issues by the development of standardized formats in MS-based proteomics. Previous HUPO-PSI formats use the extensible markup language (XML) to store raw MS data (mzML[112]), identification (mzIdentML[125]) as well as quantification results (mzQuantML[126]), all accompanied by extensive metadata.

These formats allow researchers to store their data in a vendor-independent format and report their complete results in a unified, computer-readable way. From this perspective, the PSI formats have been successful in addressing the data archiving and data sharing requirements imposed by public funding agencies and enforced by new journal guidelines. However, communication of the final identification and quantification results in a human readable way is not easily possible using the existing XML-based formats. Additionally, many software employed in downstream analysis like R[127], KNIME or Microsoft Excel ideally require data in a simple tabular format. While in principle, data visualization and conversion tools can be built upon the XML data formats, this additional step often poses a significant obstacle for researchers - especially for those from other scientific fields. In other areas of research, a similar problem has been successfully addressed by the development of tabular file formats parallel to more complex, XML-based file formats (e.g., the tabular microarray file format MAGE-

TAB[128] and the XML file format MAGE-ML[129] or MITAB[130] and the XML-based format PSI-Molecular Interactions (workgroup) (PSI-MI)[131] for molecular interaction). We developed mzTab as HUPO-PSI standard file format for reporting MS-based proteomics and metabolomics results. Its primary aim is to ease communication of intermediate and final results using a human readable, computer processable, tabular file format. It is designed to complement the existing XML-based file formats by providing a comprehensive summary of both identification and quantification results similar to a result table one would expect in the supplementary material of a scientific publication. The dedicated XML formats for quantification and identification data (mzQuantML and mzIdentML) store the complete experimental evidence and information to trace all processing steps. In contrast, the mzTab format is designed to allow reporting results and data processing steps at different levels of detail but is not intended to provide the full experimental evidence and all meta information of individual processing steps. Only recently, MS-based multi-omics approaches combining metabolomic, and proteomic analysis are emerging as important analytical tools. Joint analysis of these omics levels offers a more comprehensive view of the biological system and has been shown to extend our understanding of many molecular processes. To date, no standardized way of reporting of both metabolomic and proteomic results in a single file exists, and mzTab is intended to fill this gap. Along with the experimental metadata, basic quantitative information of identified proteins, peptides, and small molecules from both omics levels can be stored. In addition to the supplementary material type summary of results, mzTab also supports a more detailed representation that includes, among others, the experimental design. This more detailed flavor of mzTab has been intended to allow for downstream processing in statistical applications.

5.2 Methods

In the following subsections, the design rationale, structure and a representative selection of main concepts and elements of the mzTab format will be described. For an exhaustive description, we refer to the standardization document (version 1.0)[132]. In addition to detailed information on the file format, several examples covering common use cases are provided.

5.2.1 Design Rationales

Common use cases and practical considerations have mainly guided the design of mzTab. MzTab is intended to:

1. be a simple text file that follows a tabular structure. That way, it can be opened and inspected in spreadsheet software like Microsoft Excel or OpenOffice Calc.
2. support two level of details: summary results as expected in a supplementary material of a publication and detailed intermediate results for downstream statistical data analysis in, for example, KNIME, R, or SPSS.
3. capture identification and quantification results from MS-based proteomic and metabolomic experiments.
4. define mandatory, optional and custom data entries. Making a defined set of the provided data mandatory ensures that a minimum amount of information is provided, optional data enriches the information and user-defined data allows to extend and adapt the format to special needs and requirements.
5. capable of reporting results for a broad range of common experimental techniques.
6. allow encoding basic experimental design.
7. be easy to export from the XML-based file formats: mzIdentML and mzQuantML.
8. retain back references to the raw data, for example, for visualization purposes.
9. rely on controlled vocabulary terms from existing ontologies to provide semantically meaningful and well-defined meta information.

Data Semantics and Controlled Vocabularies

Defining the structure of a file format controls how data is organized and, to a large extent, the meaning of the stored data. In the tabular file format we envisioned for mzTab, the structure is very simple and, thus, only allows to encode basic semantic information. To gain most from the data without increasing the complexity of the file format, we designed mzTab to employ controlled vocabularies (CVs). Controlled vocabularies provide authorized terms and definitions that help to organize knowledge and attach semantics to data. The CV terms used in PSI standard formats are manually curated and organized in ontologies[133]. Querying CV terms against an ontology allow software validating the semantics of the data stored in a file in an automated fashion. For example, consider

a data value that is annotated as CV term `spectrum probability score`. In the ontology, this term is a child of a `spectrum score`, and a parent-child relation **is a** connects the two terms. This information allows tools that process `spectrum scores` to determine that they can also process a `spectrum probability score` because `spectrum probability score` **is a** `spectrum score`. This way, CV terms help to make the data machine comprehensible and comparable by defining its semantics. In practice, ontologies usually form a directed acyclic graphs with several types of connections between terms and may be used to model complex relations. MzTab uses the same ontologies used in the other PSI file formats allowing for centralized curation of CV terms and easier conversion between XML formats and mzTab.
In mzTab, CV terms and associated values are reported using a simple format:

```
Format:  [ontology,  accession, name, value]
Example: [MS, MS:1001214, Protein level global FDR, 0.01]
```

This structure is an mzTab `CV parameter`. Similarly, `user parameters` can be defined in cases no CV term exists. These resemble the CV parameter as they should also contain a descriptive name, but the ontology and accession are omitted.

```
Format:  [, , name, value]
Example: [, , my user parameter, 24.1]
```

To ensure coherence of controlled vocabulary terms, name, label, and version of used ontologies can be stored in mzTab. Providing an URL allows tools that consume mzTab files to automatically download unknown ontologies or previous versions of an ontology and perform a basic validation of the file. The recommended ontology for mzTab is the PSI-Mass Spectrometry (PSI-MS) CV[133], which is maintained by the PSI MS and Proteomics Informatics working groups. Additional ontologies might be used, for instance, to annotate samples or sample processing (i.e., the Human Disease Ontology (DOID)[134], the Brenda Tissue Ontology (BTO)[135], Sample Processing and Separation Techniques Ontology (SEP)[136]).

```
MTD sample_processing[1]        [SEP, SEP:00210, High Performance Liquid Chromatography, ]
MTD instrument[1]-name          [MS, MS:1000448, LTQ FT, ]
MTD instrument[1]-source        [MS, MS:1000073, Electrospray Ionization, ]
MTD instrument[1]-analyzer[1]   [MS, MS:1000079, FT_ICR, ]
MTD instrument[1]-detector      [MS, MS:1000112, Faraday Cup, ]
```

5.2.2 Structure

During the early stages of mzTab development, it became apparent that a single table consisting of tab-delimited columns was insufficient to support the diverse set of requirements listed in the previous subsection. Instead, mzTab is structured in five sections:

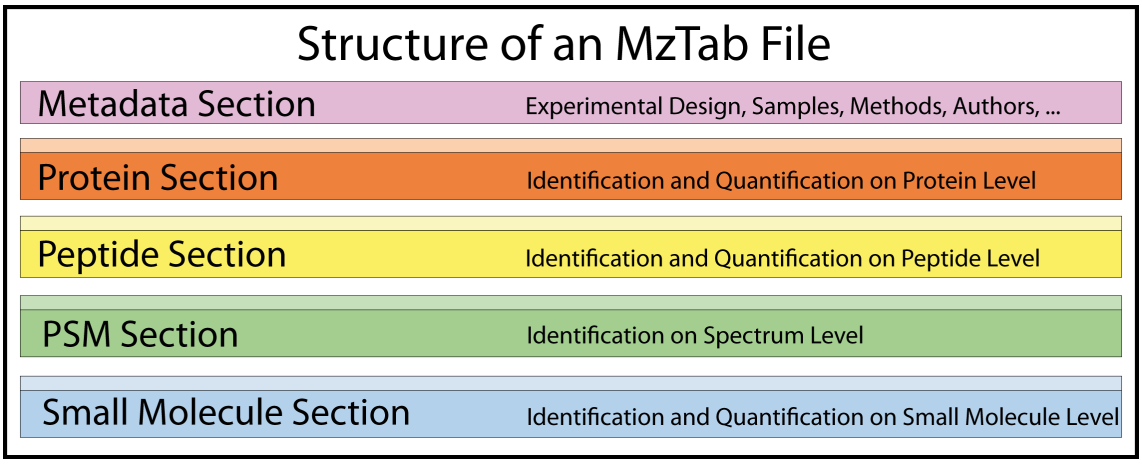


Figure 5.2: MzTab Structure Metadata section contains all metadata of reported results including the experimental design. Protein, Peptide, and Small Molecule Section may contain identification data and quantitative data. PSM contains spectrum identification data only.

Sections containing identification and quantification data (protein, peptide, PSM, and small molecule tables) are optional while the metadata section is mandatory. Each section occurs at most once and in the order indicated in Figure 5.2.
To facilitate distinction of the five sections, the first column always contains a three-letter code. This code indicates whether the line is a comment (`COM`), a metadata entry (`MTD`), protein (`PRT`), peptide (`PEP`), peptide-spectrum matching (`PSM`) or small molecules (`SMS`) row. Also , each data section must be preceded by a header line (three letter code: `PRH`, `PEH`, `PSH`, `SMH`). The use of a three letter code allows easy extraction of sections using command-line text processing tools, for instance, GNU grep.

5.2.3 Reporting Experimental Metadata

The metadata section is intended to provide basic information on the study, indicates which type of results are reported and which minimum level of detail is contained in the mzTab file. Basic study information including a human readable title and description can be provided. Information on the experimental design as well as employed software, parameters, as well as contact information and publication references allow quickly browsing relevant information that otherwise is (if at all present) buried in the method section or supplementary material of a study. The majority of metadata fields are optional but if filled, allow to annotate the complete metadata required by the Minimum Information About a Proteomics Experiment (MIAPE) [137] as well as the Core information for metabolomics reporting (CIMR) [138] guidelines.

Specifying Type and Detail of Reported Results

In the metadata section, it is mandatory to specify the *type* of mzTab file - whether identification only or quantification results are reported. As quantification data may include identification information on the quantified proteins, peptides or small molecules, a `Quantification` file is a superset of an `Identification` file. Apart from some technical reasons, like easier conversion from mzIdentML or mzQuantML, the distinction between two different mzTab types serves an additional purpose. It provides guidance to data producers on what data needs to be exported and data consumers what data can be expected in the file. We assembled tables that specify which information is mandatory to report, optional, or should be omitted (see Table 2-6 in the specification document[132]). In addition to these two types of mzTab reports, we introduced a similar concept to distinguish summary reports (e.g., a human readable digest) from the detailed report of intermediate results as generated by data converters or processing pipelines. Analogously, the `Complete` report is a more comprehensive superset of the `Summary` report. As a consequence more mandatory information and data (e.g., experimental design and quantitative values for individual replicates) must be provided to enable statistical downstream processing.

Experimental Design

We found that explicitly modeling all possible experimental designs in mzTab was out of the scope of the specification process. Instead, we use a simplified representation that handles most of the standard use cases while still allowing automated processing. In the case of more complex experimental designs, the researcher may need to consult the methods section of the original publication to interpret the reported data.
The elements used to describe the experimental design in mzTab are:

- **Sample**: The biological material that has been analyzed. Biological samples can originate from one or multiple species, cell, or tissue types or be associated with diseases.
- **MS run**: A single run on a mass spectrometer identified by a file name and format. Linking back to the original MS run is essential for tracing evidence and data visualization (e.g., tandem MS spectra of identified peptides).
- **Assay**: A measurement of a sample which produced quantitative values about peptides, proteins, or small molecules. In the case of label-free analyses, one assay maps to one MS run. In multiplexed techniques, like SILAC or iTRAQ, multiple assays are linked to a single MS run. Metadata associated with assays

are, for example, the employed quantification reagent or the biological sample that has been measured.

- **Study variable**: The study variables summarize the final quantification result, with one study variable usually associated with one investigated condition (e.g., two study variables: *cancer tissue* and *healthy tissue*). In the case of replicated measurements, the value of a study variable is usually averaged. If assays are reported, study variables reference the assays. From a conceptual point of view, a study variable may correspond to a level of an experimental factor (or, in multi-factorial experiments to a set of factor levels).

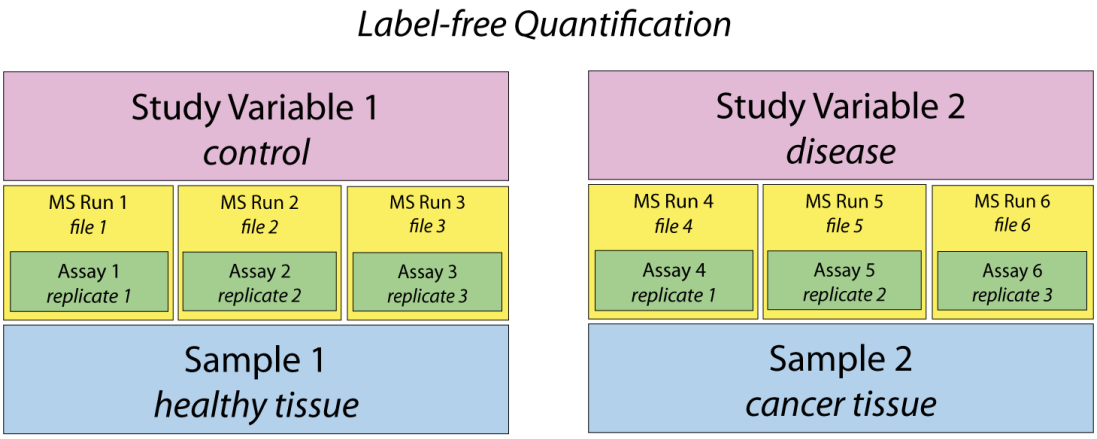


Figure 5.3: Example: Visualization of the relation of study variable, MS run, sample, and assay for a label-free experiment measuring two samples in triplicates.

MzTab represents the experimental design in Figure 5.3 as:

```
MTD   study_variable[1]-description     control
MTD   study_variable[2]-description     disease
MTD   assay[1]-ms_run_ref               ms_run[1]
MTD   assay[2]-ms_run_ref               ms_run[2]
MTD   assay[3]-ms_run_ref               ms_run[3]
MTD   assay[4]-ms_run_ref               ms_run[4]
MTD   assay[5]-ms_run_ref               ms_run[5]
MTD   assay[6]-ms_run_ref               ms_run[6]
MTD   study_variable[1]-assay_refs      assay[1],assay[2],assay[3]
MTD   study_variable[2]-assay_refs      assay[4],assay[5],assay[6]
MTD   sample[1]-description             healthy tissue
MTD   sample[2]-description             cancer tissue
MTD   sample[1]-assay_refs              assay[1],assay[2],assay[3]
MTD   sample[2]-assay_refs              assay[4],assay[5],assay[6]
```

In all cases, additional meta information related to the study can be provided (e.g., sample tissue and organisms).

Identification Metadata

Augmenting the raw identification results with metadata greatly improves the interpretability of reported results. MzTab includes metadata to specify the type of score reported by search engines at the PSM, peptide, or protein level and for small molecules. In proteomic studies, and in the context of creating summary reports, it is common practice to filter identification lists at a predefined FDR to report only the most significant ones. It is, hence, of great value to report the expected FDR if a reduced list is reported. Sample processing (e.g., cysteine blocking which reduction agents or enrichment procedure) often induces modifications of analyzed peptides and are reflected in the choice of fixed and variable search modifications. Additionally providing information on the fragmentation method is sufficient for most downstream processing use cases. Identification metadata is specified using CV parameters, and while most are optional, we highly recommend providing them. For small molecules, metadata on identification results are currently restricted to score, FDR, sample processing, and fragmentation method but might be extended in future releases of mzTab.

Quantification Metadata

Many different types of quantification methods exist and modeling each and every single one was out of the scope of the mzTab specification process. We decided to cover most widely used methods, for instance, stable isotope labeling with SILAC, isobaric labeling with iTRAQ/TMT, or label-free quantification. The `quantification_method` is specified via a CV parameter. While we do not explicitly support spectrum count-, MS^e-, or, SWATH-based methods summary results can still be reported in mzTab. To define which quantification reagents have been used in the individual channels of multiplexed experiments (or the single channel of a label-free experiment) a meta value `quantification_reagent` is provided for every mzTab assay :

```
Example iTRAQ-4plex:
MTD  quantification_method             [MS, MS:1001837, iTRAQ quantitation analysis, ]
MTD  assay[1]-quantification_reagent   [PRIDE,PRIDE:0000114,iTRAQ reagent,114]
MTD  assay[2]-quantification_reagent   [PRIDE,PRIDE:0000115,iTRAQ reagent,115]
MTD  assay[3]-quantification_reagent   [PRIDE,PRIDE:0000116,iTRAQ reagent,116]
MTD  assay[4]-quantification_reagent   [PRIDE,PRIDE:0000117,iTRAQ reagent,117]

Example SILAC:
MTD  quantification_method             [MS, MS:1001835, SILAC quantitation analysis, ]
MTD  assay[1]-quantification_reagent   [PRIDE, PRIDE:0000326, SILAC light]
MTD  assay[2]-quantification_reagent   [PRIDE, PRIDE:0000325, SILAC heavy]

Example label-free experiment:
MTD  quantification_method           [MS, MS:1001834, LC-MS label-free quantitation analysis, ]
MTD  assay[1]-quantification_reagent  [MS, MS:1002038,unlabeled sample,]
```

Additional metadata allows for the detailed specification of the unit of the reported quantification value and the modification introduced by the labeling reagent.
Small molecule quantitation shares the same metadata as peptides and proteins. Compared to ontologies in proteomics, ontologies in metabolomics currently lack behind in the number of existing CV terms. In the future, we expect more metabolomic ontologies to be defined.

Instrument and Software

Name, source, analyzer, and detector of one or multiple instruments are reported using CV parameters. Reporting of processing software using CV parameter is mandatory for `Complete` files and can be augmented by providing the software settings.

Contact and Publications

Several publications can be associated with a mzTab file and are preferably specified using PubMed IDs and DOIs. The contact information of authors may include name, affiliation, and email address.

Expressing Confidence and Reliability of Reported Results

In some cases, reliable statistical confidence measures are not easily calculated, and manual validation of identification results are required. This problem is especially prominent for metabolomic study results, which in most cases require expert inspection and annotation. In mzTab, protein, peptide, PSM, and small molecule identifications can be assigned a confidence score via the reliability column. While not strictly enforced, the reliability of a proteomic identification or quantification result should be reported as integers between one and three. These numbers correspond to a:

1 - poor reliability.

2 - medium reliability.

3 - high reliability.

In metabolomics, according to the current Metabolomics Standards Initiative (MSI) agreement[138,139], confidence values should be reported as an integer between one and four. These values correspond to:

1 - (unknown compounds).

2 - putatively characterized compound class.

3 - putatively annotated compounds.

4 - identified metabolite.

Possible applications of reliability values are "traffic light" type visualizations in graphical user interfaces.

Data Integrity

MzTab provides metadata to store checksums and hash values. These can be used for validation of referenced spectra files. In addition to these values, the hash function (e.g., SHA-1, see example below) is provided using a CV parameter.

```
MTD   ms_run[1]-hash_method   [MS, MS:1000569, SHA-1, ]
MTD   ms_run[1]-hash          ea32b3af2c7cat6e1ad3e85a0bd9b10d17087a4c
```

5.2.4 Reporting Peptide and Protein Identification Results

Most identification engines assign peptide sequences to experimental spectra. These peptide-spectrum matches are stored in the PSM section. Based on the PSMs, protein inference algorithms determine the set of identified proteins.

Peptide-Spectrum Matches

The PSM section stores detailed peptide-to-spectrum matching information. The amino acid sequence, modifications, and calculated m/z (`calc_mass_to_charge`) are stored for identified peptides. Precursor mass-to-charge ratio (`exp_mass_to_charge`), `charge`, `retention_time`, and a reference to the spectrum and its MS run (`spectra_ref`) retain spectrum specific properties.

```
PSH  sequence  ...  modifications  spectra_ref  retention_time  charge
↪ exp_mass_to_charge  ...
PSM  QTQTF...  ...      null  ms_run[1]:scan=1296  1336.62         3
↪ 600.6189        ...
```

Consensus identification approaches combine results from multiple search engines to improve the identification of peptides. Post-processing software (e.g., tools to calculate posterior error probabilities) generate derived scores from original search engine scores. MzTab supports the simple use case of a single search engine score as well as these advanced use case by allowing several score columns. Each column then corresponds to a different score type. CV parameters in the metadata section ensure that score types are properly defined. In addition to search engine scores, information on sequence database and version can be included. In the case that the peptide sequence can be unambiguously assigned to a single protein, the peptide is

annotated as unique. The protein's `accession`, the position in the protein (`start`, `end`) as well as the preceding and following amino acid (`pre`, `post`) in the protein are then annotated in the respective columns. In bottom-up proteomics, the more complex case arises that an identified peptide sequence matches to multiple proteins. The inference of the correct or most likely assignments is a non-trivial problem[140]. MzIdentML[125] features detailed protein inference information. In mzTab, we decided, for the sake of reduced complexity, to only model basic protein inference information. If the peptide is assigned to multiple proteins, the `PSM` row may be duplicated for each matching protein accession. The peptide is marked as non-unique and position as well as flanking amino acids in the respective protein are reported.

Co- or Posttranslational modifications, as well as modifications attributed to sample treatment, are represented by a list of modification objects. Each modification object is encoded as string of format:

`{position}{score (CV param.)}-{Mod. or Subst. ID}|{neutral loss (CV param.)}.`

The prefix of the string encodes the position in the peptide or protein, depending on whether the modification is reported in the PSM/peptide or protein section. Advanced use cases supported by mzTab include reporting of positional ambiguities with localization scores. Here, individual scores like the probability of a phosphorylation site assignment can be provided by a tool. Modification identifiers are either specified using identifiers from widely used modification databases (Unimod or PSI-MOD) or in the case of unknown modifications, by specification of a chemical formula or mass shift in Hill notation[141]. Neutral losses can optionally be reported using a CV term associated with a position and modification.

Reporting Protein Identification Results

Protein identifications are reported in the protein section. Columns for protein accession, description, species, GO terms, and protein database used in the identification process are expected by the mzTab standard. Statistics on the experimental evidence should be provided in columns that store the total number of PSMs and peptides that map to a protein identification, the sequence coverage by identified peptides, as well as the number of peptides that match only to the reported protein (i.e., provide strong evidence for the presence of the protein). Similar to the PSM section, the tools that created one or several scores are listed and scores from all MS runs are reported in line. This way, proteins that have been identified in only one, several, or all MS runs can be easily spotted. In addition, the best score obtained from searches in all MS runs is reported. If more than one search engine score is reported several best-score-columns

are reported (one for every individual score type). In bottom-up proteomics, so-called *shared peptides* are regularly identified. These peptides map to multiple proteins. The computational process of *protein inference* aims at determining the proteins in a sample based on the identified peptides. Because ambiguity can usually not completely be resolved in this process, protein groups are reported. How these groups are formed depends highly on the inference algorithm. In mzTab, the column `ambiguity_members` is used to list the members of a protein group. To report PSMs that map to multiple accessions the same PSM row is copied and adapted to the protein accession of the ambiguity group. That way, information on the position in proteins can be reported for each hypothesis. The example below demonstrates how a group of three proteins is reported. Here, the protein group has been identified by a single peptide that maps to three proteins:

```
COM  Protein group with three proteins P1,P2,P3
PRH  accession  ambiguity_members  ...
PRT         P1              P2,P2  ...
...
COM  PSM of peptide DEPIANGER mapping to the protein group above.
PSH  sequence  PSM_ID  accession  unique  ...  pre  post  start  end
PSM  DEPIANGER  2        P1          0     ...  N    E      34    42
PSM  DEPIANGER  2        P2          0     ...  K    I     124   132
PSM  DEPIANGER  2        P3          0     ...  M    E      23    31
```

Reporting Small Molecule Identification Results

The mzIdentML format developed by the PSI covers the majority of identification results from proteomics experiments. For MS-based metabolomics experiments, no widely used and standardized data format is available to store identification or quantification data. To fill this critical gap, mzTab offers the small molecule section that allows reporting basic identification and quantification results.

Small molecules are primarily specified using a unique identifier from metabolite or compound databases like the Human Metabolome DataBase (HMDB)[142], PubChem[143], LipidMaps[144], LipidHome[145] or Chemical Entities of Biological Interest (ChEBI)[146].

Similar to the proteomic sections, potential chemical modifications, and a human readable description can be provided. In addition, a chemical formula in Hill notation[141] can be provided. Simplified Molecular-Input Line-Entry System (SMILES) or IUPAC International Chemical Identifier (InChI) keys provide structural information on the small molecule. If SMILES are reported in their respective column, the molecule can be easily visualized by other tools. Analogous to the MS information provided in the PSM section, precursor mass-to-charge ratio (`exp_mass_to_charge`), `charge`,

retention_time, and a reference to the spectrum and its MS run (spectra_ref) are provided. Search engine and scores are recorded for every MS run. If spectral libraries are used for identification, the fields taxonomy, species, database and database_version should be filled.

SMH	identifier	chemical_formula	smiles	inchi_key	description	exp_mass_to_charge
SML	CHEBI:17562	C9H13N3O5	Nc1ccn...	UHDGC...	Cytidine	244.0928

Reporting Quantification Results

Quantification techniques measure the abundance of an analyte on the level of assays (e.g., corresponding to one channel of a multiplexed experiment or the single channel in label-free experiments). Abundance values of several assays (e.g., replicates) are summarized to form an abundance value of a study variable. The abundance value of a study variable might correspond to the average abundance of a protein in a control group or a disease state. Details on the calculation of the abundance value are currently not part of the mzTab standard. In Summary files, abundance values for study variables are sufficient. In Complete files, abundance values for every assay must be reported, too. In both cases, standard deviation and standard error of study variable abundance values should be reported. For every abundance value as well as standard deviation and error, columns are appended to the identification information in the protein, peptide, and small molecule section.

Reporting Feature Quantities

In some cases, reporting only protein abundances may not be detailed enough. MzTab offers the possibility to report feature abundances in the peptide section. This section has primarily been intended to aggregate quantitative information (e.g., linked features after map alignment) on peptides and should not be used in files containing only identification data. Quantitation columns in the peptide section are used the same way as quantitation columns in the protein section. Particularly, features that have been linked between several assays (e.g., MS runs in label-free experiments) are reported in the same format as quantified proteins (using the peptide_abundance_assay[1-n] and peptide_abundance_study_variable*[1-n] columns respectively).
Some differences to the protein section are:

- Ambiguity information is not explicitly provided for shared peptides. If protein information is required, peptide rows may be duplicated for every protein they map to.
- A column is contained to indicate if the peptide is unique.

- Chromatographic retention time start, apex, and end of one master feature are reported.

Extending MzTab with Custom Metadata and Columns

Metadata not covered by the mzTab specification can be included into a document using a `custom` meta value. It expects a CV parameter, which, if the ontology and identifier are kept empty, corresponds to a user parameter object. In many cases, it is desirable to report tool- or analysis-specific information in additional columns. These custom columns can be added to the protein, peptide, PSM, and small molecule section. Column headers of custom columns must start with the prefix `opt_`. An additional context qualifier allows to bind the column name to a specific assay, study variable or MS run (e.g., `opt_assay[1], opt_study_variable[2], or opt_ms_run[3]`). If the column is not bound the qualifier `global_` must be added (e.g., `opt_global_mycolumn`). If CV terms are available that correctly describe the content of the column, it is recommended to add the CV accession and parameter name:

```
Format:  opt_{context}_cv_{CV accession}_{parameter name}
Example: opt_global_cv_MS:1002217_decoy_peptide
```

5.3 Results

The adoption of PSI file formats by data producers and data repositories has been, in retrospect, a rather slow process. MzTab is still a rather new format but already shows a considerably quicker adoption that we mainly attribute to the simpler structure and availability of reference implementations and validators. It is hard to foresee to what extent mzTab will be accepted in the proteomic and metabolomic community and how data sharing and communication of results with mzTab will benefit research in other fields. In the following, we will take a look at existing implementations and, in our view, promising early adoptions.

5.3.1 Implementation in OpenMS

We implemented reading and writing of mzTab files, including functionality for basic semantic validation, into the core OpenMS library. MzTab data is stored in instances of the *MzTab* class and de-/serialized on disc via instances of the *MzTabFile* class. The novel TOPP tool `MzTabExporter` builds on this core functionality and exports OpenMS XML-based file formats to mzTab files. It currently supports all major OpenMS data formats for peptide and protein identification (idXML, mzIdentML), peptide quantification

(featureXML), and peptides quantified from linked features (consensusXML) to mzTab. See Figure 5.4 for an example KNIME workflow that exports peptide identification and reads them back into a KNIME table for further processing.

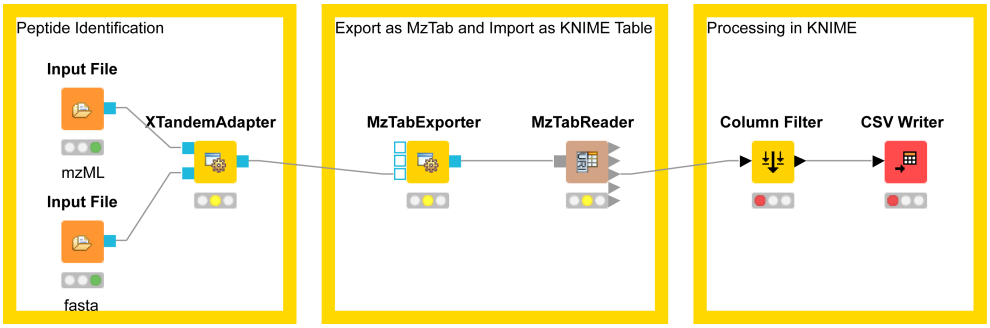


Figure 5.4: MzTab workflow using OpenMS in KNIME. Peptides are identified using `XTandemAdapter`. Results in the OpenMS idXML format are exported to MzTab using the `MzTabExporter` node. The `MzTabReader` node reads the mzTab file and converts the data to a standard KNIME table. The table can be processed using standard KNIME nodes (e.g., a simple `Column Filter`) and exported to different file formats (here a simple CSV file via the `CSV Writer` node).

The TOPP `AccurateMassSearch` tool for small molecule identification has been adapted to write mzTab files. It annotates all currently supported small molecule columns including SMILES columns with information on the molecule structure.

5.3.2 Statistical Downstream Analysis

Recently, we demonstrated how mzTab could be used to build integrated processing and statistical analysis workflows in KNIME. In a simple biomarker discovery workflow in KNIME (Figure 5.5 and Appendix Figure E.1 for a simplified workflow), we showed that mzTab captures the relevant information for downstream data analysis. Visualization of molecular structures of differentially quantified small molecules was achieved by interfacing with R and third-party KNIME extensions[147].

5.3.3 Community Acceptance

Several tools and libraries have been developed by other members of the computational mass spectrometry community that support reading and writing of mzTab files. The jmzTab[148] Java API is the current reference implementation for reading, writing, and validating mzTab files. The R Bioconductor package MSnbase[149] has been extended to support reading and writing of mzTab files (version 1.5.6). The PRIDE PRoteomics IDEntifications database (PRIDE) submission tool PRIDE Converter 2[150] converts identifications from multiple search engines to mzTab. At present, mzTab files

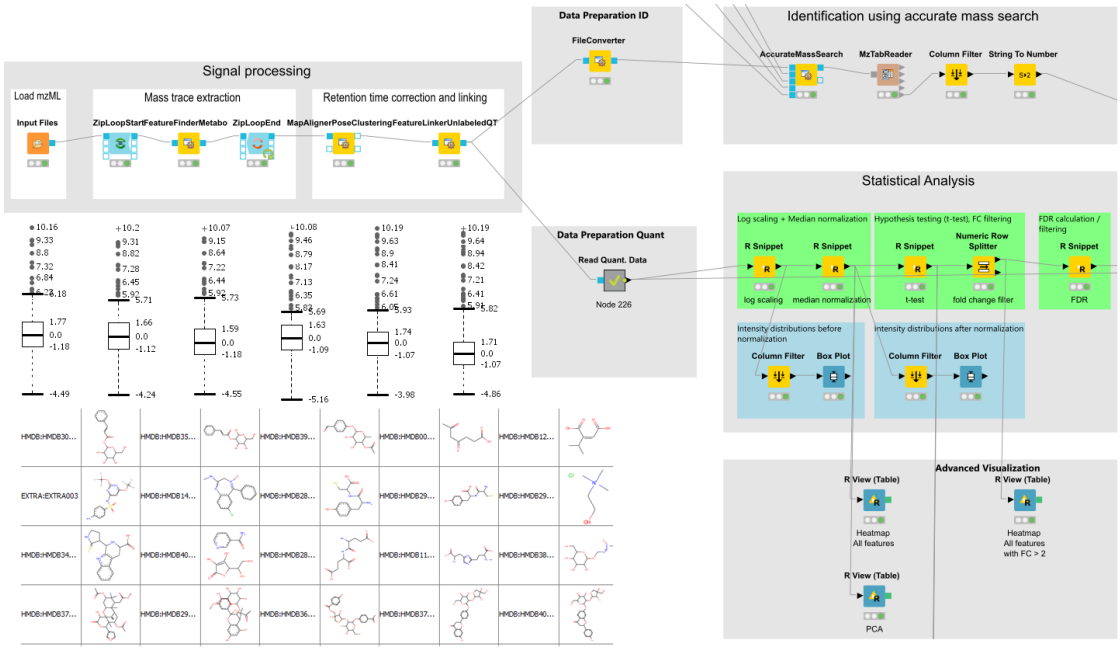


Figure 5.5: Section of a small molecule quantification workflow in KNIME. After signal processing, the table of quantified small molecules is subjected to statistical downstream analysis (green color). After log scaling and normalization, multiple *t-tests* including FDR calculation are performed to detect differentially quantified small molecules. Visualizations (blue color) are generated from the individual node outputs. Example plots on the left have been generated from the small molecule data and annotations using KNIME nodes for the visualization of boxplots and molecular structures.

can be provided by the user for conversion to the PRIDE XML submission format. That way, quantitative information can be made easily available. Alternatively, quantitative data can be provided via mzQuantML files which are converted to mzTab using the mzQuantML Java library[151]. Other tools (e.g., Mascot[50]) use mzTab as an export format for identification results. Results from other search engines (e.g., MSGFPlus[152]) can be converted to mzTab using third-party tools. A more specialized tool, the LipidDataAnalyzer[153] processes lipidomics LC-MS data and is able to export quantitative data using mzTab. PIA[154], a toolbox for MS-based protein inference and identification analysis also exports mzTab files.

MzTab was designed to ease data sharing and communication of MS-based proteomic and metabolomics results with researchers from other fields. In this regard, we are happy that it got adopted by the ProteomeXchange (PX) consortium[155], which coordinates the data submission to three major proteomics data repositories PRIDE[156], PeptideAtlas[157], and the Mass spectrometry Interactive Virtual Environment (MassIVE)[158]. Currently, mzTab files for all complete MS/MS submissions to PX via PRIDE

are automatically generated and made available to the public. In MassIVE, mzTab is primarily used for internal storage of results. It is planned that mzTab files will also be made available for download.

5.4 Discussion

We developed the mzTab file format for communication and statistical downstream analysis of experimental results in proteomic and metabolomic studies. We chose a simple structured tabular-separated format and CV terms from established ontologies to provide metadata and semantics for the stored data. Two different types of mzTab files, one for simple summary results and one for complete reports are specified, which define a set of mandatory entries and information that needs to be present. In our view, these design decisions allow mzTab to manage the balancing act of not being overly generic with loose semantics and a less versatile format with tight semantics.
In contrast to existing XML-based formats developed by the HUPO Proteomics Standards Initiative (PSI), mzTab was not intended for archiving. Therefore, it does not contain the complete trace of evidence that leads to the analysis result. Instead, the focus lied on a human- and computer-readable, tabular representation of data that can be easily interpreted by humans and processed by statistical analysis or spreadsheet software.
We, thus, expect mzTab to be more accessible for researchers in and outside the field of proteomics and metabolomics. Recently mzTab has been adapted by the ProteomXchange consortium. We, therefore, already achieved a wide accessibility of MS/MS-based proteomic study results submitted to ProteomeXchange via PRIDE.
Especially for MS-based metabolomics results there existed neither established tabular separated nor XML-based data formats, yet. MzTab fills the important gap of providing a reporting format and has already proven useful for lipidomics approaches (as implemented in the LipidDataAnalyzer[153]). Until XML-based data formats like mzQuantML catch up and support reporting and archiving of small molecule identification and quantification data, mzTab might be used as a bridge format until the data archiving issue is solved for those formats. MetaboLights[159] the first general-purpose, open-access curated repository for metabolomics studies is currently adapting mzTab.
In addition, mzTab is, to our knowledge, the first standard format for reporting proteomic and metabolomic results from multi-omics MS-based analysis.
Another possible application of mzTab is in laboratory information management systems (LIMSs). A wide range of heterogeneous data, for example, results from quality control runs or analysis results of quantitative studies, can be supported by adding optional

columns and meta information. That way, mzTab can be tailored to the requirements of different research groups.
Historically, the first motivation for mzTab was to develop a format that can often replace difficult representations of scientific results in the supplementary material of publications. At the time of this thesis, results (e.g., tables) are, in many cases, still provided as images or proprietary binary formats. Standardizing the supplementary information provided to scientific journals by the adoption of mzTab would significantly improve accessibility to the primary results. Such large an endeavor is highly depending on the willingness of journals to adapt their reporting guidelines which in turn highly depends on how well the format has been adopted by the proteomic and metabolomic research community.

5.5 Outlook

The current mzTab specification provides three sections to report proteomic data on the level of proteins, peptides, and PSMs. In contrast, reporting of metabolomic results lacks granularity and only provides a single section to report quantified small molecules. In this regard, the small molecule section currently resembles a final summary report but lacks behind the detail of proteomic data provided in complete quantification files. While it might be sufficient for reporting of general metabolomic results, a higher detail of reporting could be obtained in one of the subsequent versions of the file format. The observation that metabolomic structures in mzTab are not as expressive as the proteomic counterpart and might be missing valuable information was also fed by discussions in the Tübingen *mzTab4Metabolomics* workshop in 2014. During the workshop discussions, experts from diverse metabolomic and proteomic fields jointly formed the idea of adapting the reporting of metabolomic results in mzTab. Extensions to the current format should attribute the heterogeneity of methods and data in metabolomics. An initial draft revealed many structural similarities between the proteomic section and proposed three sections with analogous roles:

proteomics		metabolomics	
section	description	section	description
PRT	proteins and protein groups	SMS	compounds
PEP	peptide features	SMF	small molecule features
PSM	peptide spectrum matches	SSM	small molecule spectrum matches

Table 5.1: Document structure according to an initial mzTab4Metabolomics draft.

Further development of the extended mzTab version might be provided by members of the Coordination Of Standards In Metabolomics (COSMOS)[160] initiative. COSMOS aims to improve the availability of exchange formats and terms needed to describe metabolomics results and the associated metadata.

Chapter 6

Conclusion

In the past years, high-throughput methods have been developed to obtain organism-scale quantitative views of the genome, transcriptome, proteome, and metabolome. Each technique generates a vast amount of data that needs to be analyzed to obtain biologically relevant information. If examined individually, a partial view of the abundance of biomolecules in a system is obtained. Employing multiple techniques and consolidating results offers a more comprehensive view and allows uncovering biological mechanisms and causalities that otherwise remained hidden. While combining multiple omics technologies yields more quantitative information than individual ones, combining classical approaches often falls short in providing answers to fundamental biological questions. One explanation is that focusing on abundances of biomolecules provide limited information on the highly regulated and dynamic biological processes. The detailed characterization of interactions between biomolecules provides an additional level of information on involved partners which in turn facilitates resolving their function.

Proteins and RNA/DNA molecules constitute important classes of biomolecules which are involved in essential cellular processes. At the time of this thesis, high-throughput methods for the elucidation of protein-(ribo)nucleic acid interactions were missing essential features. Our first, major contribution in this thesis, RNP^{xl}, is a novel computational method and flexible workflows that allow performing comprehensive, organism-wide studies of protein-RNA or protein-DNA interactions. We successfully applied the method to whole cell lysates of different organisms and were able to pinpoint cross-linking sites down to the resolution of single amino acids. We identified novel RNA-binding proteins, including proteins without known nucleotide-binding domains. Our method, thus, contributes to the emerging notion of a much broader prevalence of protein-RNA interactions than previously anticipated. Recently, noncanonical RNA-binding enzymes have become prominent research targets. Our method already provided supporting evidence on the associated proteins, the nucleotide motif, and the location of nucleotide-binding sites (White et al.[161], White and Garcin[162]). The ability to localize cross-linked amino acids and nucleotides in contact opens the way

for a better structural elucidation of protein-RNA/DNA complexes. Notable results obtained with our method on structure, interaction, and homology in CRISPR-Cas systems that employed RNP[xl] with complementary techniques are described by Sharma et al.[76], Staals et al.[163], Gleditzsch et al.[164], Shao et al.[165]. Increasing the number of known protein-RNA/DNA binding sites will ease associating deleterious mutations with disease phenotypes. In a similar line of research, targeted mutation of binding sites can be used to study disruption of proper protein-RNA/DNA interactions. Ultimately, we expect that only integrated approaches, combining data from different sources and techniques, will uncover causal explanations for the complex biological phenomena involving protein-RNA/DNA interactions. Based on our research results and recent publications we are confident that localization of cross-links at the amino acid-level has a significant advantage compared to existing approaches.

Over the last few years, the field of microbiome research has grown substantially. Today, microbial ecology and the study of human microbiomes are attractive research areas with important practical applications and clinical relevance. Using metaproteomics approaches, the proteins of microbial communities can be investigated. Studying multiple species measured in single biological samples pose significant challenges in data analysis. Protein-SIP approaches allow analyzing substrate metabolism and elemental flux in complex samples. Our second, major contribution MetaProSIP, enables high-throughput analysis of microbial communities using protein-SIP. By developing a novel computational method and automated workflows, we were able to reduce the time-consuming manual analysis of protein-SIP experiments from several months to minutes. We successfully identified and quantified peptides and proteins of species that partake in degradation processes based on their ability to incorporate labeled substrate molecules. Automated clustering and phylogenetic annotation allowed us to identify and distinguish functional groups of organisms. Combining MetaProSIP with existing tools in complex workflows, and the use of mixed or separate unlabeled reference samples makes it the ideal tool for large (e.g., hundreds of runs) protein-SIP studies and time series analysis. Several articles have recently been published that demonstrate the broad applicability of MetaProSIP. Lünsmann et al.[166] studied toluene degradation in a rhizospheric wetland model and identified key degraders and biochemical pathways. Starke et al.[167] applied MetaProSIP to study a soil metaproteome spanning two kingdoms: fungi and bacteria. They showed that mainly bacteria were involved in the assimilation of plant-derived nitrogen, whereas fungi dominated the degradation of complex carbon compounds. Starke et al.[168] investigated the acetate utilization network within a benzene-degrading and sulfate-reducing syntrophic consortium of microorganisms. Using MetaProSIP on data from a pulsed $^{13}C_2$-acetate protein-SIP

experiment, they were able to gain detailed insight into the acetate utilization network and identified *Epsilonproteobacteria* as dominant acetate utilizers. Future developments of MetaProSIP include additional labeled substrate elements. Currently, we investigate how deuterium or ^{18}O-labeled water can be used as general activity marker in microbial communities[169]. Metaproteomic studies mainly drove the development of MetaProSIP. We currently evaluate how well MetaProSIP can be applied to single-species analysis. Protein turnover and anabolic amino acid synthesis pathways are interesting research targets MetaProSIP might find applications. Based on our research results and recent publications we are confident that MetaProSIP offers a significant methodological advancement and is widely applicable to metaproteomic studies.

The third contribution of this thesis mzTab is a data format that eases the communication and automated processing of experimental results from proteomics and metabolomics studies. MzTab was developed with members of the Proteomics Standard Initiative. MzTab is a standardized, tabular file format that is both human-readable and computer-consumable. In contrast to existing XML file formats, it enables convenient downstream statistical analysis of proteomic and metabolomic results. We implemented tools for reading and writing of mzTab files in OpenMS. Development of a KNIME community node by other OpenMS developers significantly lowered the bar for non-experts as it allows performing workflow-based data analytics on proteomic and metabolomic results. MzTab provides a concise presentation of study results for easier communication of results to a wider community. Several tools and libraries have been developed for reading and writing mzTab by other members of the mass spectrometry community. While most of these originate from academic research, commercial applications are starting to support mzTab. MzTab got adopted by the ProteomeXchange consortium, which coordinates the data submission to three major proteomics data repositories. In two of the main proteomic data repositories, mzTab is used for storing proteomics and metabolomics results. Thus, mzTab already plays a major role in carrying proteomic results over to other areas of research. In the future, improved and more detailed representation of metabolomics results will be a major format extension. We hope that these format extensions ease further adoption of mzTab in the field of metabolomics.

Computational mass spectrometry-based proteomics is an ever-changing field with broad applications. In this thesis, we demonstrated that joint development of biochemical and computational methods complement each other. We expanded the methodological repertoire in the fields of protein-DNA/RNA biology and microbial metaproteomics. Designed with high-throughput analysis in mind, our methods allow approaching biological questions in a data-driven way. Data from successful applications of our methods have resulted in novel biological insights, and have led to

prospective follow-up studies. Additionally, we contributed to the development of a data format for communication of proteomics and metabolomics research results. In conclusion, we are certain that the results of our efforts contribute to the advancement of computational mass spectrometry, and is applicable to various fields of life sciences.

Bibliography

[1] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. 1

[2] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. 1

[3] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004. 1

[4] Douglas B Kell and Stephen G Oliver. Here is the evidence, now what is the hypothesis? the complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, 26(1):99–105, 2004. 1

[5] Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nature reviews. Genetics*, 7(2):85–97, 2006. 1

[6] Siân Jones, Xiaosong Zhang, D Williams Parsons, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, Hirohiko Kamiyama, Antonio Jimeno, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, 321(5897):1801–1806, 2008.

[7] Benjamin J Blencowe. Alternative splicing: new insights from global analyses. *Cell*, 126(1): 37–47, 2006. 1

[8] David Eisenberg, Edward M Marcotte, Ioannis Xenarios, and Todd O Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, 2000. 1

[9] David Whitford. *Proteins: structure and function*. John Wiley & Sons, 2013. 1

[10] Mike Tyers and Matthias Mann. From genomics to proteomics. *Nature*, 422(6928):193–197, 2003. 2

[11] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, et al. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, 2014. 2

[12] Min-Sik Kim, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S Manda, Raghothama Chaerkady, Anil K Madugundu, Dhanashree S Kelkar, Ruth Isserlin, Shobhit Jain, et al. A draft map of the human proteome. *Nature*, 509(7502):575–581, 2014. 2

[13] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437 (7062):1173–1178, 2005. 2

[14] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, 21(6): 697–700, 2003.

[15] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005. 2

[16] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human RNA-binding proteins. *Nature reviews. Genetics*, 2014. 2, 36

[17] Emily F Freed, Franziska Bleichert, Laura M Dutca, and Susan J Baserga. When ribosomes go bad: diseases of ribosome biogenesis. *Molecular BioSystems*, 6(3):481–493, 2010. 2, 32

[18] Marie Morimoto and Cornelius F Boerkoel. The role of nuclear bodies in gene expression and disease. *Biology*, 2(3):976–1033, 2013. 2, 32

[19] Katja Dettmer, Pavel A Aronov, and Bruce D Hammock. Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, 26(1):51–78, 2007. 5

[20] Yang Wang, Shuying Liu, Yuanjia Hu, Peng Li, and Jian-Bo Wan. Current state of the art of mass spectrometry-based metabolomics studies–a review focusing on wide coverage, high throughput and easy identification. *RSC Advances*, 5(96):78728–78737, 2015. 5

[21] Gary J Patti, Oscar Yanes, and Gary Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature reviews. Molecular cell biology*, 13(4):263–269, 2012. 5

[22] Rima Kaddurah-Daouk, Bruce S Kristal, and Richard M Weinshilboum. Metabolomics: a global biochemical approach to drug response and disease. *Annual review of pharmacology and toxicology*, 48:653–683, 2008. 5

[23] Richard H Perry, R Graham Cooks, and Robert J Noll. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass spectrometry reviews*, 27(6):661–699, 2008. 9

[24] Jesper V Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods*, 4 (9):709–712, 2007. 12

[25] John EP Syka, Joshua J Coon, Melanie J Schroeder, Jeffrey Shabanowitz, and Donald F Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9528–9533, 2004. 12

[26] Hanno Steen and Matthias Mann. The ABC's (and XYZ's) of peptide sequencing. *Nature reviews. Molecular cell biology*, 5(9):699–711, 2004. 12

[27] Hendrik Weisser, Sven Nahnsen, Jonas Grossmann, Lars Nilse, Andreas Quandt, Hendrik Brauer, Marc Sturm, Erhan Kenar, Oliver Kohlbacher, Ruedi Aebersold, et al. An automated pipeline for high-throughput label-free quantitative proteomics. *Journal of proteome research*, 12(4): 1628–1644, 2013. 14

[28] Ole Schulz-Trieglaff, Rene Hussong, Clemens Gröpl, Andreas Hildebrandt, and Knut Reinert. A fast and accurate algorithm for the quantification of peptides from mass spectrometry data. In *Annual International Conference on Research in Computational Molecular Biology*, pages 473–487. Springer, 2007. 14

[29] Shao-En Ong, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, 1(5): 376–386, 2002. 14

[30] Jana Seifert, Martin Taubert, Nico Jehmlich, Frank Schmidt, Uwe Völker, Carsten Vogt, Hans-Hermann Richnow, and Martin von Bergen. Protein-based stable isotope probing (protein-SIP) in functional metaproteomics. *Mass spectrometry reviews*, 31(6):683–697, 2012. 15, 73

[31] Philip L Ross, Yulin N Huang, Jason N Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, Sasi Pillai, Subhakar Dey, Scott Daniels, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, 3(12):1154–1169, 2004. 15

[32] Andrew Thompson, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms. *Analytical chemistry*, 75(8): 1895–1904, 2003. 15

[33] Mark Brönstrup. Absolute quantification strategies in proteomics based on mass spectrometry. *Expert Review of Proteomics*, 1(4):503–512, 2004. 15

[34] Scott A Gerber, John Rush, Olaf Stemman, Marc W Kirschner, and Steven P Gygi. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem ms. *Proceedings of the National Academy of Sciences*, 100(12):6940–6945, 2003. 15

[35] A. D. Compiled by McNaught and A. Wilkinson. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. Blackwell Scientific Publications, 1997. 16

[36] Michael W Senko, Steven C Beu, and Fred W McLaffertycor. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229–233, 1995. 19

[37] Eva Lange, Clemens Gröpl, Ole Schulz-Trieglaff, Andreas Leinenbach, Christian Huber, and Knut Reinert. A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics*, 23(13):i273–i281, 2007. 19, 77

[38] John D Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of statistics*, pages 2013–2035, 2003. 23

[39] Enis Afgan, Dannon Baker, Marius Van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Carl Eberhard, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research*, page gkw343, 2016. 24

[40] The Qt library. `https://www.qt.io/`. Accessed: 2016-09-06. 26

[41] The Xerces library. `http://xerces.apache.org/`. Accessed: 2016-09-06. 26

[42] Robin Lougee-Heimer. The Common Optimization INterface for Operations Research: Promoting open-source software in the operations research community. *IBM Journal of Research and Development*, 47(1):57–66, 2003. 26

[43] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 26

[44] The Eigen C++ Library for linear algebra: matrices, vectors, numerical solvers, and related algorithms. `http://eigen.tuxfamily.org/`. Accessed: 2016-09-06. 26

[45] The Geometric Tools Library. `http://www.geometrictools.com/`. Accessed: 2016-09-06. 26

[46] Björn Karlsson. *Beyond the C++ standard library: an introduction to boost*. Pearson Education, 2005. 26

[47] Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature methods*, 13 (9):741–748, 2016. 27

[48] Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open mass spectrometry search algorithm. *Journal of proteome research*, 3(5):958–964, 2004. 27, 77

[49] David Fenyö and Ronald C Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry*, 75(4):768–774, 2003. 27

[50] John S Cottrell and U London. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999. 27, 113

[51] David L Tabb, Christopher G Fernando, and Matthew C Chambers. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of proteome research*, 6(2):654–661, 2007. 27

[52] Oliver Serang, Michael J MacCoss, and William Stafford Noble. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of proteome research*, 9(10):5346–5357, 2010. 27

[53] Thomas A Steitz and Peter B Moore. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends in biochemical sciences*, 28(8):411–418, 2003. 31

[54] Leslie E Orgel. Evolution of the genetic apparatus. *Journal of molecular biology*, 38(3):381–393, 1968. 31

[55] Francis HC Crick et al. The origin of the genetic code. *Journal of molecular biology*, 38(3): 367–379, 1968. 31

[56] Franziska Bleichert and Susan J Baserga. Ribonucleoprotein multimers and their functions. *Critical reviews in biochemistry and molecular biology*, 45(5):331–350, 2010. 32

[57] Charles G Hoogstraten and Minako Sumita. Structure–function relationships in RNA and RNP enzymes: recent advances. *Biopolymers*, 87(5-6):317–328, 2007. 32

[58] Thomas A Steitz. A structural understanding of the dynamic ribosome machine. *Nature reviews. Molecular Cell Biology*, 9(3):242–253, 2008. 32

[59] William H Hudson and Eric A Ortlund. The structure, function and evolution of proteins that bind DNA and RNA. *Nature reviews. Molecular Cell Biology*, 15(11):749–760, 2014. 32

[60] Stefan Maas, Yukio Kawahara, Kristen M Tamburro, and Kazuko Nishikura. A-to-I RNA editing and human disease. *RNA biology*, 3(1):1–9, 2006. 32

[61] Tom Vulliamy, Richard Beswick, Michael Kirwan, Anna Marrone, Martin Digweed, Amanda Walne, and Inderjeet Dokal. Mutations in the telomerase component NHP2 cause the premature ageing syndrome dyskeratosis congenita. *Proceedings of the National Academy of Sciences*, 105(23): 8073–8078, 2008. 32

[62] Jeffry D Sander and J Keith Joung. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature biotechnology*, 32(4):347–355, 2014. 32

[63] John C Kendrew, G Bodo, Howard M Dintzis, RG Parrish, Harold Wyckoff, and David C Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181 (4610):662–666, 1958. 33

[64] Kurt Wüthrich. The way to NMR structures of proteins. *Nature structural & molecular biology*, 8 (11):923–925, 2001. 33

[65] Werner Kühlbrandt. Cryo-EM enters a new era. *Elife*, 3(e03665):e03665, 2014. 33

[66] Jacqueline LS Milne, Mario J Borgnia, Alberto Bartesaghi, Erin EH Tran, Lesley A Earl, David M Schauder, Jeffrey Lengyel, Jason Pierson, Ardan Patwardhan, and Sriram Subramaniam. Cryo-electron microscopy–a primer for the non-microscopist. *FEBS Journal*, 280(1):28–45, 2013. 33

[67] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, 2008. 33

[68] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010. 33

[69] Katharina Kramer, Timo Sachsenberg, Benedikt M Beckmann, Saadia Qamar, Kum-Loong Boon, Matthias W Hentze, Oliver Kohlbacher, and Henning Urlaub. Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nature methods*, 11(10):1064–1070, 2014. 34, 35, 45, 46, 47, 48, 49, 50, 52, 67, 141, 147, 150, 151, 152, 153, 154, 155

[70] Kristen M Meisenheimer and Tad H Koch. Photocross-linking of nucleic acids to associated proteins. *Critical Reviews in Biochemistry and Molecular Biology*, 32(2):101–140, 1997. 36

[71] Kenneth R Williams and William H Konigsberg. Identification of amino acid residues at interface of protein-nucleic acid complexes by photochemical cross-linking. *Methods in Enzymology*, 208: 516, 1991. 36

[72] Henning Urlaub, Volker Kruft, O Bischof, EC Müller, and B Wittmann-Liebold. Protein-rRNA binding features and their structural and functional implications in ribosomes as determined by cross-linking studies. *The EMBO journal*, 14(18):4578, 1995. 36

[73] Yoshihiko Ikeguchi and Hiroshi Nakamura. Determination of organic phosphates by column-switching high performance anion-exchange chromatography using on-line preconcentration on titania. *Analytical Sciences*, 13(3):479–483, 1997. 37

[74] Boris Macek, Matthias Mann, and Jesper V Olsen. Global and site-specific quantitative phosphoproteomics: principles and applications. *Annual Review of Pharmacology and Toxicology*, 49: 199–221, 2009. 37

[75] Saadia Qamar, Katharina Kramer, and Henning Urlaub. Studying RNA–protein interactions of pre-mRNA complexes by mass spectrometry. *Methods in Enzymology*, 2015. 37

[76] Kundan Sharma, Ajla Hrle, Katharina Kramer, Timo Sachsenberg, Raymond HJ Staals, Lennart Randau, Anita Marchfelder, John van der Oost, Oliver Kohlbacher, Elena Conti, et al. Analysis of protein–RNA interactions in CRISPR proteins and effector complexes by UV-induced cross-linking and mass spectrometry. *Methods*, 89:138–148, 2015. 37, 67, 118

[77] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, 2008. 37, 76

[78] UniProt Consortium et al. The universal protein resource (UniProt). *Nucleic acids research*, 36 (suppl 1):D190–D195, 2008. 38

[79] Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–1372, 2008. 39

[80] Uzma Zaman, Florian M Richter, Romina Hofele, Katharina Kramer, Timo Sachsenberg, Oliver Kohlbacher, Christof Lenz, and Henning Urlaub. Dithiothreitol (DTT) acts as a specific, UV-inducible cross-linker in elucidation of protein–RNA interactions. *Molecular & Cellular Proteomics*, 14(12):3196–3210, 2015. 43

[81] Soheil Pourshahian and Patrick A Limbach. Application of fractional mass for the identification of peptide–oligonucleotide cross-links by mass spectrometry. *Journal of mass spectrometry*, 43(8): 1081–1088, 2008. 43

[82] Matthias W Hentze. Enzymes as RNA-binding proteins: a role for (di)nucleotide-binding domains? *Trends in biochemical sciences*, 19(3):101–103, 1994. 48

[83] Carla Schmidt, Katharina Kramer, and Henning Urlaub. Investigation of protein–RNA interactions by mass spectrometry - Techniques and applications. *Journal of proteomics*, 75(12):3478–3494, 2012. 50

[84] Alfredo Castello, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M Beckmann, Claudia Strein, Norman E Davey, David T Humphreys, Thomas Preiss, Lars M Steinmetz, et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 149(6): 1393–1406, 2012. 50

[85] Adam Ben-Shem, Nicolas Garreau de Loubresse, Sergey Melnikov, Lasse Jenner, Gulnara Yusupova, and Marat Yusupov. The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science*, 334 (6062):1524–1529, 2011. 51, 155

[86] Frédéric H-T Allain, Yi-Meng Yen, James E Masse, Peter Schultze, Thorsten Dieckmann, Reid C Johnson, and Juli Feigon. Solution structure of the HMG protein NHP6A and its interaction with DNA reveals the structural determinants for non-sequence-specific binding. *The EMBO journal*, 18(9):2563–2579, 1999. 51

[87] Erik Werner, Wolfgang Wende, Alfred Pingoud, and Udo Heinemann. High resolution crystal structure of domain I of the Saccharomyces cerevisiae homing endonuclease PI-SceI. *Nucleic acids research*, 30(18):3962–3971, 2002. 51

[88] Christoph Leidig, Gert Bange, Jürgen Kopp, Stefan Amlacher, Ajay Aravind, Stephan Wickles, Gregor Witte, Ed Hurt, Roland Beckmann, and Irmgard Sinning. Structural characterization of a

eukaryotic chaperone - the ribosome-associated complex. *Nature structural & molecular biology*, 20(1):23–28, 2013. 51

[89] Jurgen Cox, Nadin Neuhauser, Annette Michalski, Richard A Scheltema, Jesper V Olsen, and Matthias Mann. Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research*, 10(4):1794–1805, 2011. 54

[90] The OpenMP API specification for parallel programming. `http://www.openmp.org/`. Accessed: 2017-02-06. 56

[91] Franziska Hufsky, Kai Dührkop, Florian Rasche, Markus Chimani, and Sebastian Böcker. Fast alignment of fragmentation trees. *Bioinformatics*, 28(12):i265–i273, 2012. 58

[92] Sean A Beausoleil, Judit Villén, Scott A Gerber, John Rush, and Steven P Gygi. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature biotechnology*, 24(10):1285–1292, 2006. 59

[93] Viktoria Dorfer, Peter Pichler, Thomas Stranzl, Johannes Stadlmann, Thomas Taus, Stephan Winkler, and Karl Mechtler. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of proteome research*, 13(8):3679–3684, 2014. 63

[94] Thomas Taus, Thomas Köcher, Peter Pichler, Carmen Paschke, Andreas Schmidt, Christoph Henrich, and Karl Mechtler. Universal and confident phosphorylation site localization using phosphoRS. *Journal of proteome research*, 10(12):5354–5362, 2011. 63

[95] Johannes Veit, Timo Sachsenberg, Aleksandar Chernev, Fabian Aicheler, Henning Urlaub, and Oliver Kohlbacher. LFQProfiler and RNPxl: Open-source tools for label-free quantification and protein–RNA cross-linking integrated into Proteome Discoverer. *Journal of proteome research*, 15 (9):3441–3448, 2016. 63, 141

[96] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC bioinformatics*, 10 (1):421, 2009. 66, 90

[97] Craig D Wenger and Joshua J Coon. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *Journal of proteome research*, 12(3):1377–1386, 2013. 66

[98] Sangtae Kim and Pavel A Pevzner. Universal database search tool for proteomics. *Nature communications*, 5:5277, 2014. 69

[99] Fiona J. Flett, Timo Sachsenberg, Oliver Kohlbacher, C. Logan Mackay, and Heidrun Interthal. Differential enzymatic 16O/18O labelling for the detection of cross-linked nucleic acid-protein heteroconjugates. *(submitted)*, 2017. 69

[100] J.A. Schofield, P.R. Betteridge, G. Ryback, and P.J. Geary. New strains of pseudomonas putida and their use, May 22 1990. URL `https://www.google.de/patents/US4927759`. US Patent 4,927,759. 72

[101] Patricia Lepage, Marion C Leclerc, Marie Joossens, Stanislas Mondot, Hervé M Blottière, Jeroen Raes, Dusko Ehrlich, and Joel Doré. A metagenomic insight into our gut's microbiome. *Gut*, 62 (1):146–158, 2013. 72

[102] Marc G Dumont and J Colin Murrell. Stable isotope probing - linking microbial identity to function. *Nature reviews. Microbiology*, 3(6):499–504, 2005. 73

[103] Josh D Neufeld, Michael Wagner, and J Colin Murrell. Who eats what, where and when? Isotope-labelling experiments are coming of age. *The ISME journal*, 1(2):103–110, 2007. 73

[104] Stefan Radajewski, Philip Ineson, Nisha R Parekh, and J Colin Murrell. Stable-isotope probing as a tool in microbial ecology. *Nature*, 403(6770):646–649, 2000. 73

[105] Nico Jehmlich, Frank Schmidt, Martin Taubert, Jana Seifert, Felipe Bastida, Martin von Bergen, Hans-Hermann Richnow, and Carsten Vogt. Protein-based stable isotope probing. *Nature protocols*, 5(12):1957–1966, 2010. 73

[106] Martin von Bergen, Nico Jehmlich, Martin Taubert, Carsten Vogt, Felipe Bastida, Florian-Alexander Herbst, Frank Schmidt, Hans-Hermann Richnow, and Jana Seifert. Insights from quantitative metaproteomics and protein-stable isotope probing into microbial ecology. *The ISME journal*, 7 (10):1877–1885, 2013. 73, 93

[107] Yingfeng Wang, Tae-Hyuk Ahn, Zhou Li, and Chongle Pan. Sipros/ProRata: a versatile informatics system for quantitative community proteomics. *Bioinformatics*, 29(16):2064–2065, 2013. 74, 94, 95

[108] John C Price, Shenheng Guan, Alma Burlingame, Stanley B Prusiner, and Sina Ghaemmaghami. Analysis of proteome dynamics in the mouse brain. *Proceedings of the National Academy of Sciences*, 107(32):14508–14513, 2010. 74, 94, 95

[109] Gordon W Slysz, Laurey Steinke, David M Ward, Christian G Klatt, Therese RW Clauss, Samuel O Purvine, Samuel H Payne, Gordon A Anderson, Richard D Smith, and Mary S Lipton. Automated data extraction from in situ protein-stable isotope probing studies. *Journal of proteome research*, 13(3):1200–1210, 2014. 74, 94, 95

[110] Timo Sachsenberg, Florian-Alexander Herbst, Martin Taubert, Rene Kermer, Nico Jehmlich, Martin von Bergen, Jana Seifert, and Oliver Kohlbacher. MetaProSIP: automated inference of stable isotope incorporation rates in proteins for functional metaproteomics. *Journal of proteome research*, 14(2):619–627, 2014. 76, 82, 83, 87, 89, 90, 91, 141, 159, 161

[111] Nico Jehmlich and Martin von Bergen. Protocol for performing protein stable isotope probing (protein-SIP) experiments. *Springer Protocols Handbooks*, 2016. 75

[112] Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Römpp, Steffen Neumann, Angel D Pizarro, et al. mzML - a community standard for mass spectrometry data. *Molecular & Cellular Proteomics*, 10(1): R110–000133, 2011. 76, 98

[113] Johannes Junker, Chris Bielow, Andreas Bertsch, Marc Sturm, Knut Reinert, and Oliver Kohlbacher. TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data. *Journal of proteome research*, 11(7):3914–3920, 2012. 77

[114] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207–214, 2007. 77

[115] Charles L Lawson and Richard J Hanson. *Solving least squares problems*. SIAM, 1995. 80

[116] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996. 84

[117] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *12^{th} international conference on Computer Vision*, pages 460–467. IEEE, 2009. 84

[118] Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, pages 68–125, 1990. 84

[119] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, pages 240–242, 1895. 84

[120] Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta. clValid, an R package for cluster validation. *Journal of Statistical Software (Brock et al., March 2008)*, 2011. 84

[121] Martin Taubert, Martin von Bergen, and Jana Seifert. Limitations in detection of 15N incorporation by mass spectrometry in protein-based stable isotope probing (protein-SIP). *Analytical and bioanalytical chemistry*, 405(12):3989–3996, 2013. 86

[122] René Kermer, Sabrina Hedrich, Martin Taubert, Sven Baumann, Michael Schlömann, D Barrie Johnson, and Jana Seifert. Elucidation of carbon transfer in a mixed culture of Acidiphilium cryptum and Acidithiobacillus ferrooxidans using protein-based stable isotope probing. *Journal of Integrated OMICS*, 2(1):37–45, 2012. 88

[123] Martin Taubert, Carsten Vogt, Tesfaye Wubet, Sabine Kleinsteuber, Mika T Tarkka, Hauke Harms, François Buscot, Hans-Hermann Richnow, Martin von Bergen, and Jana Seifert. Protein-SIP enables time-resolved analysis of the carbon flux in a sulfate-reducing, benzene-degrading microbial consortium. *The ISME journal*, 6(12):2291–2301, 2012. 89, 90

[124] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. MEGAN analysis of metagenomic data. *Genome research*, 17(3):377–386, 2007. 90

[125] Andrew R Jones, Martin Eisenacher, Gerhard Mayer, Oliver Kohlbacher, Jennifer Siepen, Simon J Hubbard, Julian N Selley, Brian C Searle, James Shofstahl, Sean L Seymour, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular & Cellular Proteomics*, 11(7):M111–014381, 2012. 98, 108

[126] Mathias Walzer, Da Qi, Gerhard Mayer, Julian Uszkoreit, Martin Eisenacher, Timo Sachsenberg, Faviel F Gonzalez-Galarza, Jun Fan, Conrad Bessant, Eric W Deutsch, et al. The mzQuantML data standard for mass spectrometry–based quantitative studies in proteomics. *Molecular & Cellular Proteomics*, 12(8):2332–2340, 2013. 98

[127] RDevelopment Core Team et al. R: A language and environment for statistical computing. *R foundation for Statistical Computing*, 2005. 98

[128] Tim F Rayner, Philippe Rocca-Serra, Paul T Spellman, Helen C Causton, Anna Farne, Ele Holloway, Rafael A Irizarry, Junmin Liu, Donald S Maier, Michael Miller, et al. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *Bmc Bioinformatics*, 7(1):489, 2006. 99

[129] Paul T Spellman, Michael Miller, Jason Stewart, Charles Troup, Ugis Sarkans, Steve Chervitz, Derek Bernhart, Gavin Sherlock, Catherine Ball, Marc Lepage, et al. Design and implementation of microarray gene expression markup language (mage-ML). *Genome biology*, 3(9):research0046, 2002. 99

[130] Samuel Kerrien, Sandra Orchard, Luisa Montecchi-Palazzi, Bruno Aranda, Antony F Quinn, Nisha Vinod, Gary D Bader, Ioannis Xenarios, Jérôme Wojcik, David Sherman, et al. Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC biology*, 5(1):44, 2007. 99

[131] Henning Hermjakob, Luisa Montecchi-Palazzi, Gary Bader, Jerome Wojcik, Lukasz Salwinski, Arnaud Ceol, Susan Moore, Sandra Orchard, Ugis Sarkans, Christian Von Mering, et al. The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nature biotechnology*, 22(2):177–183, 2004. 99

[132] MzTab specification. `https://github.com/HUPO-PSI/mzTab/`. Accessed: 2017-01-25. 99, 103

[133] Gerhard Mayer, Luisa Montecchi-Palazzi, David Ovelleiro, Andrew R Jones, Pierre-Alain Binz, Eric W Deutsch, Matthew Chambers, Marius Kallhardt, Fredrik Levander, James Shofstahl, et al. The HUPO proteomics standards initiative-mass spectrometry controlled vocabulary. *Database: the journal of biological databases and curation*, 2013, 2013. 100, 101

[134] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012. 101

[135] Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research*, 39(suppl 1):D507–D513, 2011. 101

[136] The Sample Processing and Separation Techniques Ontology. `https://bioportal.bioontology.org/ontologies/SEP`. Accessed: 2017-03-16. 101

[137] Chris F Taylor, Norman W Paton, Kathryn S Lilley, Pierre-Alain Binz, Randall K Julian, Andrew R Jones, Weimin Zhu, Rolf Apweiler, Ruedi Aebersold, Eric W Deutsch, et al. The minimum information about a proteomics experiment (MIAPE). *Nature biotechnology*, 25(8):887–893, 2007. 102

[138] Lloyd W Sumner, Alexander Amberg, Dave Barrett, Michael H Beale, Richard Beger, Clare A Daykin, Teresa W-M Fan, Oliver Fiehn, Royston Goodacre, Julian L Griffin, et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3(3):211–221, 2007. 102, 106

[139] Reza M Salek, Christoph Steinbeck, Mark R Viant, Royston Goodacre, and Warwick B Dunn. The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience*, 2(1):13, 2013. 106

[140] Alexey I Nesvizhskii and Ruedi Aebersold. Interpretation of shotgun proteomic data the protein inference problem. *Molecular & Cellular Proteomics*, 4(10):1419–1440, 2005. 108

[141] Edwin A Hill. On a system of indexing chemical literature; adopted by the classification division of the us patent office. 1. *Journal of the American Chemical Society*, 22(8):478–494, 1900. 108, 109

[142] David S Wishart, Craig Knox, An Chi Guo, Roman Eisner, Nelson Young, Bijaya Gautam, David D Hau, Nick Psychogios, Edison Dong, Souhaila Bouatra, et al. HMDB: a knowledgebase for the human metabolome. *Nucleic acids research*, 37(suppl 1):D603–D610, 2009. 109

[143] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 40(D1): D13–D25, 2012. 109

[144] Manish Sud, Eoin Fahy, Dawn Cotter, Edward A Dennis, and Shankar Subramaniam. LIPID MAPS-Nature Lipidomics Gateway: an online resource for students and educators interested in lipids. *Journal of Chemical Education*, 89(2):291–292, 2011. 109

[145] Joseph M Foster, Pablo Moreno, Antonio Fabregat, Henning Hermjakob, Christoph Steinbeck, Rolf Apweiler, Michael JO Wakelam, and Juan Antonio Vizcaíno. LipidHome: a database of theoretical lipids optimized for high throughput mass spectrometry lipidomics. *PloS one*, 8(5): e61951, 2013. 109

[146] Paula de Matos, Nico Adams, Janna Hastings, Pablo Moreno, and Christoph Steinbeck. A database for chemical proteomics: ChEBI. In *Chemical Proteomics*, pages 273–296. Springer, 2012. 109

[147] Stephan Aiche, Timo Sachsenberg, Erhan Kenar, Mathias Walzer, Bernd Wiswedel, Theresa Kristl, Matthew Boyles, Albert Duschl, Christian G Huber, Michael R Berthold, et al. Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry. *Proteomics*, 15(8):1443–1447, 2015. 112

[148] Qing-Wei Xu, Johannes Griss, Rui Wang, Andrew R Jones, Henning Hermjakob, and Juan Antonio Vizcaíno. jmzTab: A Java interface to the mzTab data standard. *Proteomics*, 14(11):1328–1332, 2014. 112

[149] Laurent Gatto and Kathryn S Lilley. MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28(2):288–289, 2012. 112

[150] Richard G Côté, Johannes Griss, José A Dianes, Rui Wang, James C Wright, Henk WP van den Toorn, Bas van Breukelen, Albert JR Heck, Niels Hulstaert, Lennart Martens, et al. The PRoteomics IDEntification (PRIDE) Converter 2 framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium. *Molecular & Cellular Proteomics*, 11(12):1682–1689, 2012. 112

[151] Da Qi, Huaizhong Zhang, Jun Fan, Simon Perkins, Addolorata Pisconti, Deborah M Simpson, Conrad Bessant, Simon Hubbard, and Andrew R Jones. The mzqlibrary–an open source java library supporting the hupo-psi quantitative proteomics standard. *Proteomics*, 15(18):3152–3162, 2015. 113

[152] Sangtae Kim, Nikolai Mischerikow, Nuno Bandeira, J Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert JR Heck, and Pavel A Pevzner. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Molecular & Cellular Proteomics*, 9(12):2840–2852, 2010. 113

[153] Jürgen Hartler, Martin Trötzmüller, Chandramohan Chitraju, Friedrich Spener, Harald C Köfeler, and Gerhard G Thallinger. Lipid Data Analyzer: unattended identification and quantitation of lipids in LC-MS data. *Bioinformatics*, 27(4):572–577, 2011. 113, 114

[154] Julian Uszkoreit, Alexandra Maerkens, Yasset Perez-Riverol, Helmut E Meyer, Katrin Marcus, Christian Stephan, Oliver Kohlbacher, and Martin Eisenacher. PIA–an intuitive protein inference engine with a web-based user interface. *Journal of proteome research*, 2015. 113

[155] Juan A Vizcaíno, Eric W Deutsch, Rui Wang, Attila Csordas, Florian Reisinger, Daniel Ríos, José A Dianes, Zhi Sun, Terry Farrah, Nuno Bandeira, et al. Proteomexchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology*, 32(3):223–226, 2014. 113

[156] Juan Antonio Vizcaíno, Richard G Côté, Attila Csordas, José A Dianes, Antonio Fabregat, Joseph M Foster, Johannes Griss, Emanuele Alpi, Melih Birim, Javier Contell, et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic acids research*, 41 (D1):D1063–D1069, 2013. 113

[157] Frank Desiere, Eric W Deutsch, Nichole L King, Alexey I Nesvizhskii, Parag Mallick, Jimmy Eng, Sharon Chen, James Eddes, Sandra N Loevenich, and Ruedi Aebersold. The peptideatlas project. *Nucleic acids research*, 34(suppl 1):D655–D658, 2006. 113

[158] The Qt library. `https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp`. Accessed: 2017-03-16. 113

[159] Kenneth Haug, Reza M Salek, Pablo Conesa, Janna Hastings, Paula de Matos, Mark Rijnbeek, Tejasvi Mahendraker, Mark Williams, Steffen Neumann, Philippe Rocca-Serra, et al. MetaboLights - an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research*, 41(D1):D781–D786, 2013. 114

[160] Christoph Steinbeck, Pablo Conesa, Kenneth Haug, Tejasvi Mahendraker, Mark Williams, Eamonn Maguire, Philippe Rocca-Serra, Susanna-Assunta Sansone, Reza M Salek, and Julian L Griffin. MetaboLights: towards a new COSMOS of metabolomics data management. *Metabolomics*, 8(5): 757–760, 2012. 116

[161] Michael R White, Mohd M Khan, Daniel Deredge, Christina R Ross, Royston Quintyn, Beth E Zucconi, Vicki H Wysocki, Patrick L Wintrode, Gerald M Wilson, and Elsa D Garcin. A dimer interface mutation in glyceraldehyde-3-phosphate dehydrogenase regulates its binding to AU-rich RNA. *Journal of Biological Chemistry*, 290(3):1770–1785, 2015. 117

[162] Michael R White and Elsa D Garcin. The sweet side of RNA regulation: glyceraldehyde-3-phosphate dehydrogenase as a noncanonical RNA-binding protein. *Wiley Interdisciplinary Reviews: RNA*, 7(1):53–70, 2016. 117

[163] Raymond HJ Staals, Yifan Zhu, David W Taylor, Jack E Kornfeld, Kundan Sharma, Arjan Barendregt, Jasper J Koehorst, Marnix Vlot, Nirajan Neupane, Koen Varossieau, et al. RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. *Molecular cell*, 56(4):518–530, 2014. 118

[164] Daniel Gleditzsch, Hanna Müller-Esparza, Patrick Pausch, Kundan Sharma, Srivatsa Dwarakanath, Henning Urlaub, Gert Bange, and Lennart Randau. Modulating the Cascade architecture of a minimal Type IF CRISPR-Cas system. *Nucleic acids research*, page gkw469, 2016. 118

[165] Yaming Shao, Hagen Richter, Shengfang Sun, Kundan Sharma, Henning Urlaub, Lennart Randau, and Hong Li. A non-stem-loop CRISPR RNA is processed by dual binding Cas6. *Structure*, 24(4): 547–554, 2016. 118

[166] Vanessa Lünsmann, Uwe Kappelmeyer, René Benndorf, Paula M Martinez-Lavanchy, Anja Taubert, Lorenz Adrian, Marcia Duarte, Dietmar H Pieper, Martin Bergen, Jochen A Müller, et al. In situ protein-SIP highlights Burkholderiaceae as key players degrading toluene by para ring hydroxylation in a constructed wetland model. *Environmental microbiology*, 2016. 118

[167] Robert Starke, René Kermer, Lynn Ullmann-Zeunert, Ian T Baldwin, Jana Seifert, Felipe Bastida, Martin von Bergen, and Nico Jehmlich. Bacteria dominate the short-term assimilation of plant-derived N in soil. *Soil Biology and Biochemistry*, 96:30–38, 2016. 118

[168] Robert Starke, Andreas Keller, Nico Jehmlich, Carsten Vogt, Hans H Richnow, Sabine Kleinsteuber, Martin von Bergen, and Jana Seifert. Pulsed 13C2-acetate protein-SIP unveils Epsilonproteobacteria as dominant acetate utilizers in a sulfate-reducing microbial community mineralizing benzene. *Microbial ecology*, 71(4):901–911, 2016. 118

[169] Timo Sachsenberg, Robert Starke, Nico Jehmlich, Martin von Bergen, and Oliver Kohlbacher. Stable isotope probing of proteins (protein-SIP) with heavy water (d2o and h2 18o): a general marker for microbial activity. In *Proteomic Forum*, 2017. 119

[170] Johannes Griss, Joseph M Foster, Henning Hermjakob, and Juan Antonio Vizcaíno. PRIDE Cluster: building a consensus of proteomics data. *Nature methods*, 10(2):95–96, 2013. 142

[171] Lingdong Quan and Miao Liu. CID, ETD and HCD fragmentation to study protein post-translational modifications. *Modern Chemistry and Applications*, 1(1), 2013. 143

[172] William M Haynes. *CRC handbook of chemistry and physics*. CRC press, 2016. ISBN 9781498754286. 144

[173] Katharina Kramer. Investigation of protein-RNA interactions by UV cross-linking and mass spectrometry: methodological improvements toward in vivo applications. *GGNB - Göttinger Graduiertenschule für Neurowissenschaften, Biophysik und molekulare Biowissenschaften*, (495), 2014. 147

[174] Cameron D Mackereth, Tobias Madl, Sophie Bonnal, Bernd Simon, Katia Zanier, Alexander Gasch, Vladimir Rybin, Juan Valcárcel, and Michael Sattler. Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature*, 475(7356):408–411, 2011. 155

[175] Deyu Zhu, Craig R Stumpf, Joseph M Krahn, Marvin Wickens, and Traci M Tanaka Hall. A 5' cytosine binding pocket in Puf3p specifies regulation of mitochondrial mRNAs. *Proceedings of the National Academy of Sciences*, 106(48):20192–20197, 2009. 155

Abbreviations

4SU 4-thiouridine. 42, 43, 48–50, 63

ALS amyotrophic lateral sclerosis. 32

API application programming interface. 63, 64, 112

BTO the Brenda Tissue Ontology. 101

C18 octadecyl. 7

ChEBI Chemical Entities of Biological Interest. 109

CID collision-induced dissociation. 12, 21

CIMR Core information for metabolomics reporting. 102

CLIP cross-linking immunoprecipitation. 33

COSMOS Coordination Of Standards In Metabolomics. 116

cryo-EM cryo-electron microscopy. 33

CTD common tool description file. 29

CV controlled vocabularies. 100

DDA data-dependent acquisition. 37

DOID the Human Disease Ontology. 101

DRBP DNA- and RNA-binding protein. 32

DTT dithiothreitol. 43

ESI electrospray ionisation. 9

ETD electron-transfer dissociation. 12, 21, 125

FDR false discovery rate. 22, 23, 39, 92, 93, 105, 113

FLR false-localization rate. 69

FPR false positive rate. 92

GKN Generic KNIME Nodes. 29

HCD higher-energy collision dissociation. 12, 21, 37, 58, 66, 125

HMDB Human Metabolome DataBase. 109

hnRNP heterogeneous ribonucleoprotein particle. 32

HPLC High Performance Liquid Chromatography. 7

HUPO HUuman Proteome Organisation. 98

iFPR incorporation FPR. 92, 93

InChI IUPAC International Chemical Identifier. 109

iTRAQ isobaric tags for relative and absolute quantitation. 15, 105

KH K Homology. 47, 48

KNIME the KoNstanz Information MinEr. 24, 25, 29, 30, 95, 98, 100, 112, 113

LC Liquid Chromatography. 7

LIMS laboratory information management system. 114

LR labeling ratio. 74, 76, 77, 81, 86, 87, 89, 91–94

m/z mass-to-charge ratio. 9, 10, 12, 14, 54

MassIVE Mass spectrometry Interactive Virtual Environment. 113, 114

MetaProSIP Meta-Proteomics using Stable Isotope Probing. 75

MIAPE Minimum Information About a Proteomics Experiment. 102

mRNA messenger RNA. 31

MS mass spectrometry. 5, 8, 76, 107–111

MS/MS tandem mass spectrum. 12, 21, 76, 113

MSI Metabolomics Standards Initiative. 106

NMR spectroscopy nuclear magnetic resonance spectroscopy. 33

OpenMS Open Mass Spectrometry. 14, 17, 19, 24–30

PAR-CLIP Photoactivatable-Ribonucleoside-Enhanced CLIP. 33

PD Thermo Proteome Discoverer. 63–65, 69

PRIDE PRIDE PRoteomics IDEntifications database. 112, 113

protein-SIP stable isotope probing of proteins. 15

PSI Proteomics Standards Initiative. 98

PSI-MI PSI-Molecular Interactions (workgroup). 99

PSM peptide spectrum match. 21, 53

PTM post-translational modifications. 57, 69

PX ProteomeXchange. 113, 119

RIA relative isotopic abundances. 18, 74–78, 81–84, 86, 87, 89–94

RISC RNA-induced silencing complex. 32

RNP ribonucleoprotein. 32

RPC reversed-phase chromatography. 7

RRM RNA-recognition motif. 46–48

SEC size exclusion chromatography. 36

SEP Sample Processing and Separation Techniques Ontology. 101

SILAC stable isotope labeling by amino acids in cell culture. 14, 105

SIP stable isotope probing. 73, 95

SMILES Simplified Molecular-Input Line-Entry System. 109, 112

snRNP small nuclear ribonucleic protein. 32

TMT Tandem Mass Tag. 15, 105

TOF time-of-flight. 9

TOPP OpenMS pipeline. 24, 27

TOPPAS the OpenMS Pipeline Assistant. 24

XIC extracted ion chromatogram. 39, 79–81

XML extensible markup language. 98

XRC X-ray crystallography. 33

Appendix A: Permissions and Contributions

Single amino acid assignment of nucleotide-binding sites in RNA- and DNA-binding proteins

Cross-link identification[69]: Permission to reuse of text, figures, and charts was granted by the Nature Publishing Group for the article: "Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins, Kramer K, Sachsenberg T, Beckmann B, Qamar S, Boon K, Hentze M, Kohlbacher O, Urlaub H. Nature Methods 11, 1064–1070 (2014). KK, BB, SQ, KB, MH, and HU designed biochemical experiments. KB and KK designed and transformed the yeast strain. KK and BB. carried out experiments for the yeast systems; KK analyzed the resulting data. SQ performed experiments in the human system; KK and SQ analyzed the resulting data. KK, **TS**, OK and HU designed data analysis strategy; **TS** developed algorithms and implemented RNP^{xl} tools. **TS** implemented TOPPView visualization. KK and **TS** tested the data analysis and visualization tools. KK, **TS**, BB, MH, OK and HU wrote the paper. KK, **TS** and SQ compiled the supplementary materials. **TS** deposited data to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD000513.
Cross-link localization (Application note on integration of RNPxlSearch into Proteome Discoverer[95]): KK, UZ, SQ, KS, and AC designed, performed and validated experiments. **TS** developed algorithms and implemented `RNP`^{xl}`Search` tools and workflows. AC and **TS** and performed data analyses. JV integrated RNP^{xl}Search into Proteome Discoverer. JV, **TS**, AC, FA, HU, and OK wrote the paper.

Dynamic Stable Isotope Probing of Metaproteomic Communities[110]

Permission to reuse of text, figures, and charts was granted by the American Chemical Society. In compliance with the permission and copyright policy, we state that our publication was: "Reprinted and adapted with permission from MetaProSIP: automated inference of stable isotope incorporation rates in proteins for functional metaproteomics. Sachsenberg T, Herbst, FA, Taubert M, Kermer R, Jehmlich N, von Bergen M, Seifert J, Kohlbacher O. Journal of Proteome Research 14(2), 619–627 (2015). Copyright 2014 American Chemical Society." FH, MT, RK, NJ, MB and JS designed biochemical experiments. FH performed and validated experiments. FH, MT, RK, NJ, MB, JS, **TS**, OK designed data analysis strategy; **TS** formalized decomposition, developed algorithms, and implemented the `MetaProSIP` tool and workflows. **TS** and FH performed data

analyses. **TS** deposited data to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD000382.

Standardized Reporting of Experimental Results in Proteomic and Metabolomic Studies[170]

Permission to reuse of text, figures, and charts was granted by the American Society for Biochemistry and Molecular Biology. In compliance with the permission and copyright policy, we state that: "This research was originally published in Mol Cell Proteomics. Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, Hartler J, Thallinger GG, Salek RM, Steinbeck C, Neuhauser N, Cox J, Neumann S, Fan J, Reisinger F, Xu QW, Del Toro N, Pérez-Riverol Y, Ghali F, Bandeira N, Xenarios I, Kohlbacher O, Vizcaíno JA, Hermjakob H. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol Cell Proteomics. 2014; 13:2765-2775. © the American Society for Biochemistry and Molecular Biology." AJ, OK, JAV, and HH designed research. JG, AJ, **TS**, MW, LG, JH, GT, RMS, CS, NN, JC, SN, JF, FR, QX, Nd, YP, FG, NB, IX, and JAV performed research. JG, JAV, and HH wrote the paper. **TS** has been actively involved in the standardization process and creation of examples. **TS** implemented mzTab I/O file handler and data structures in OpenMS. EK and **TS** adapted OpenMS tools (`AccurateMassSearch` & `SpectralLibrarySearch`) to support writing of MzTab. JP, EP, and SA developed a KNIME node (`MzTabReader`) for importing mzTab files into KNIME tables. Workflows were developed and refined by several members of OpenMS.

SA: Stephan Aiche, NB: Nuno Bandeira, BB: Benedikt Beckmann, KB: Kum-Loong Boon, MB: Martin von Bergen, AC: Aleksandar Chernev, JC: Jürgen Cox, JF:Jun Fan, LG: Laurent Gatto, FG :Fawaz Ghali, JG: Johannes Griss, JH: Jürgen Hartler, MH: Matthias W Hentze, FA: Florian-Alexander Herbst, HH: Henning Hermjakob, NJ: Nico Jehmlich, AJ: Andrew Robert Jones, EK: Erhan Kenar, RK: René Kermer, OK: Oliver Kohlbacher, KK: Katharina Kramer, NN: Nadin Neuhauser, SN: Steffen Neumann, YP: Yasset Pérez-Riverol, JP: Julianus Pfeuffer, EP: Enes Poyraz, SQ: Saadia Qamar, FR: Florian Reisinger, **TS: Timo Sachsenberg**, RS: Reza Salek, JS: Jana Seifert, KS: Kundan Sharma, CS: Christoph Steinbeck, MT: Martin Taubert, GT: Gerhard Thallinger, NT: Noemi del Toro, HU: Henning Urlaub, JV: Johannes Veit, JAV: Juan Antonio Vizcaíno, MW: Mathias Walzer, IX: Ioannis Xenarios, QX: Qing-Wei Xu, UZ: Uzma Zaman

Appendix B: Background

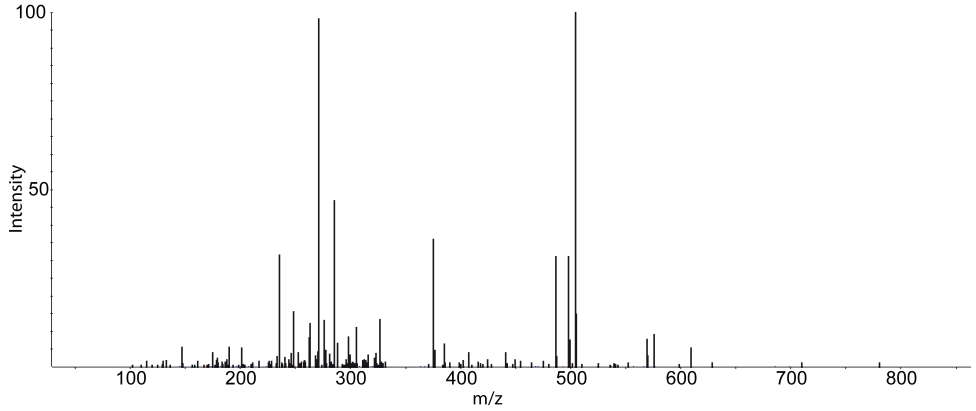


Figure B.1: MS/MS spectrum. Fragmentation of ions in the precursor isolation window. Fragment ions were recorded in the second round of mass measurement. The resulting spectrum is called *tandem spectrum*.

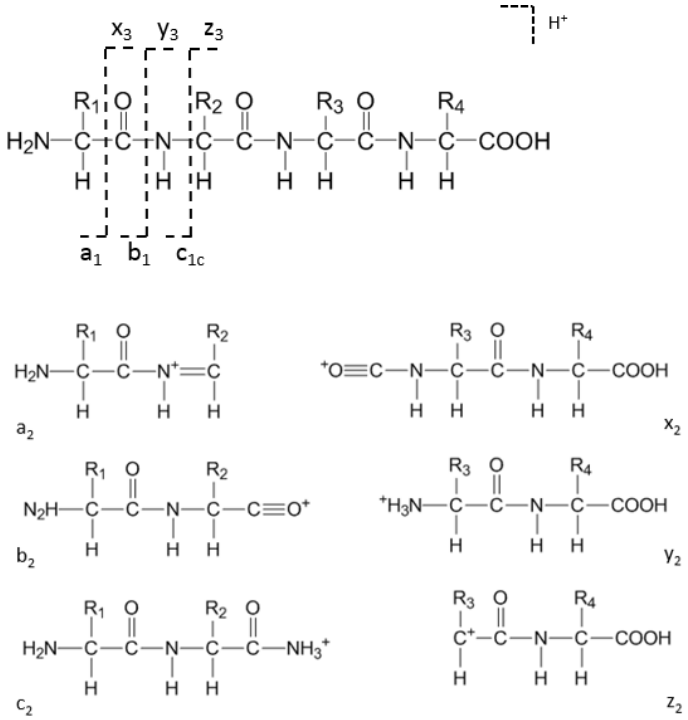


Figure B.2: Fragment ion nomenclature. Fragmentation at different backbone positions gives rise to different fragment ions[171]. CID spectra are dominated by *b*- and *y-ions*. HCD mainly generates *y-ions*. *b-ions* occur to a lesser extent and may be further fragmented to *a-ions*. ETD yields mainly *c*- and *z-ions*. *Created by Huimin Zhong, (CC-BY-SA 4.0)*

Table B.1: Natural abundance of hydrogen, carbon, nitrogen, oxygen, and sulfur isotopes[172]. Isotopes marked with (*) indicate that it is not present in nature or no meaningful abundance can be given.

Name	Symbol	Mass of Atom (u)	abundance %
Hydrogen	^{1}H	1.007825	99.9885
Deuterium	^{2}H	2.014102	0.0115
Tritium	^{3}H	3.016049	(*)
Carbon	^{12}C	12.000000	98.93
	^{13}C	13.003355	1.07
	^{14}C	14.003242	(*)
Nitrogen	^{14}N	14.003074	99.636
	^{15}N	15.000109	0.364
Oxygen	^{16}O	15.994915	99.757
	^{17}O	16.999132	0.038
	^{18}O	17.999160	0.205
Sulfur	^{32}S	31.972071	94.99
	^{33}S	32.971458	0.75
	^{34}S	33.967867	4.25
	^{36}S	35.967081	0.01

Notation and Definition of Mass Spectrometry Related Terms

In computational mass spectrometry, basic mass spectrometry terms are used in a slightly different context and deviate from the official IUPAC definitions. For example, IUPAC defines a mass spectrum as "a plot of relative abundances (...) as function of their m/z values." To ease formal description of methods and algorithms, we use following definitions and mathematical notations:

Mass Peak: A mass peak p is defined by a 3-tuple of retention time t, mass m and intensity i:

$$p_k = (t_k, m_k, i_k).$$

If the retention time of the mass peak can be deduced from the context, we might simply write:

$$p_k = (m_k, i_k).$$

Mass Spectrum: A mass spectrum s is a set of mass peaks $s = \{p_k\}$ with same retention time.

Peak Map: A peak map P is a set of mass spectra $P = \{s_j\}$.

Extracted Ion Chromatogram: We define an extracted ion chromatogram XIC as the set of peaks in P that fall in a retention time $[t_a, t_b]$ and mass-to-charge interval $[m_a, m_b]$:

$$XIC = \{p_i \in P : \ t_i \in [t_a, t_b] \wedge m_i \in [m_a, m_b]\}.$$

Variance of the Binomial Distribution

The variance of the binomial distribution is:

$$Var[P_p(n|N)] = np(1-p) = -np^2 + np, \tag{B.1}$$

a quadratic function of p ($n \in \mathbb{N}_{>0}$).
From basic analysis follows a maximum at: $p = -n/(-2n) = 0.5$.

Appendix C: RNP^{xl}

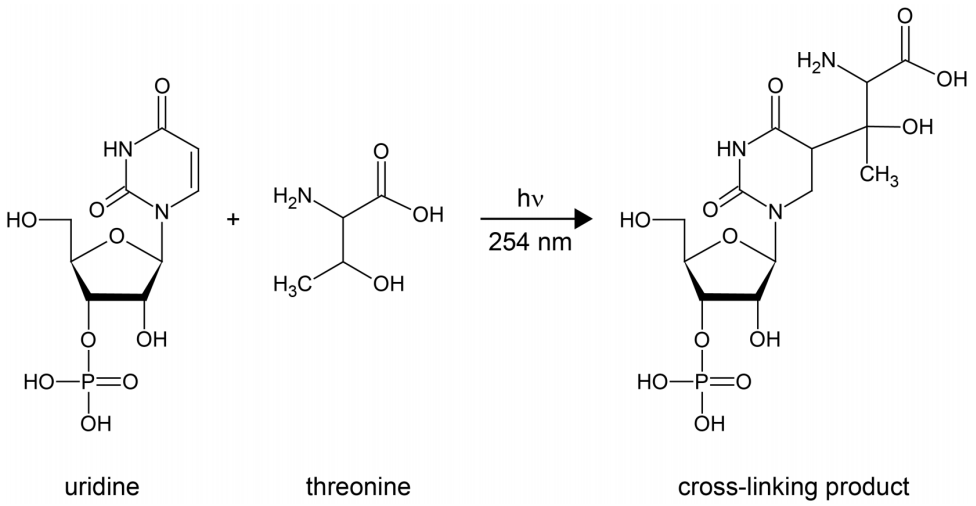


Figure C.1: Possible UV-induced cross-linking reaction between uridine and threonine. Note that the exact mechanism of cross-link formation is not fully understood and might differ. *Image kindly provided by Kramer* [173]

Table C.1: Cross-links to mono- and dinucleotides.
In order generate mono- and dinucleotides, the oligo length is restricted to two. No sequence is provided to limit oligonucleotide generation. To allow for all RNA combinations, no restriction on the minimum number of oligonucleotide occurrence is set. No DTT cysteine adduct is generated and no losses are specified in the modifications row. *Adapted from Kramer et al.* [69].

Parameter	Value
length	2
sequence	
target_nucleotides	A=C10H14N5O7P, C=C9H14N3O8P, G=C10H14N5O8P, U=C9H13N2O9P
mapping	A->A, C->C, G->G, U->U
restrictions	A=0, C=0, G=0, U=0
modifications	
CysteineAdduct	false

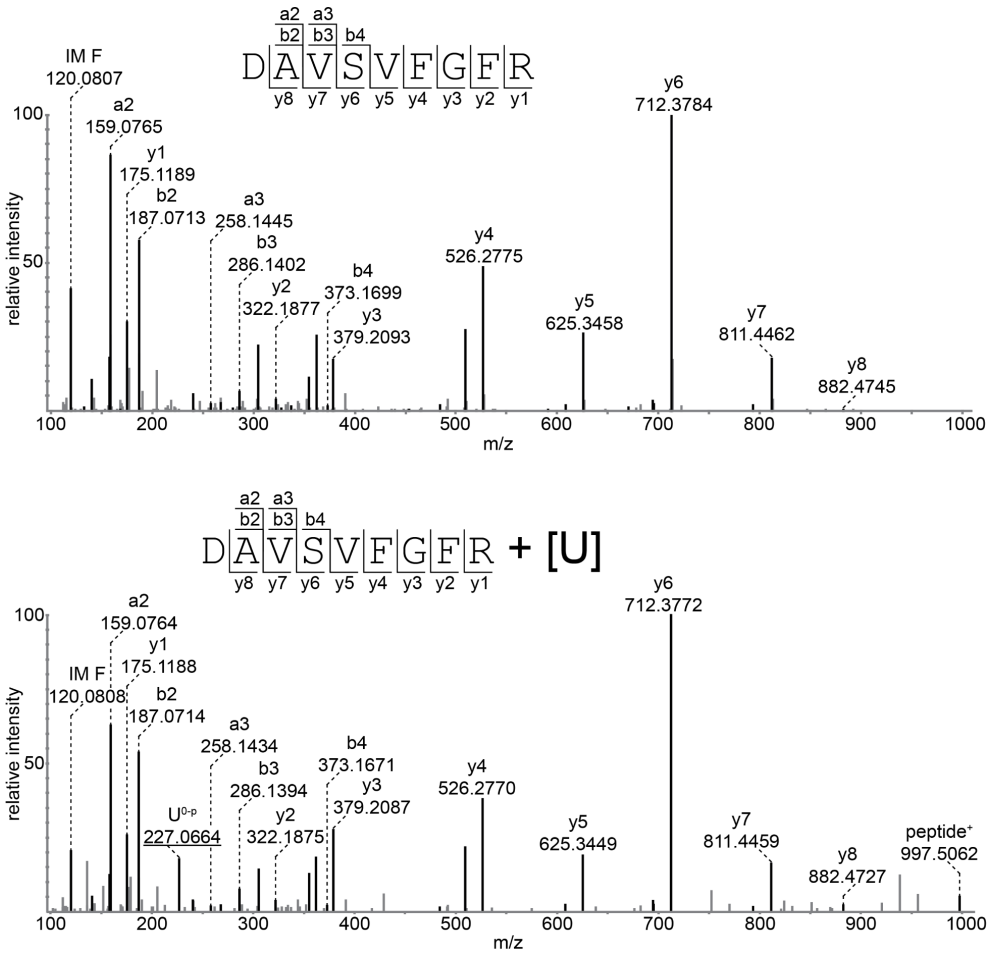
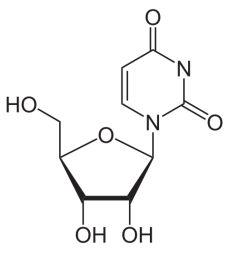
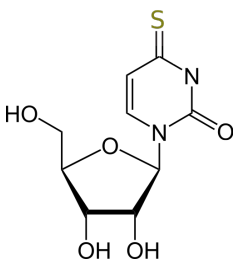


Figure C.2: Fragment spectrum of peptide (**top**) and cross-link (**bottom**) may be very similar because of a predominant total loss of the RNA upon fragmentation. *Image kindly provided by K. Kramer.*



(a) uridine



(b) 4-thiouridine

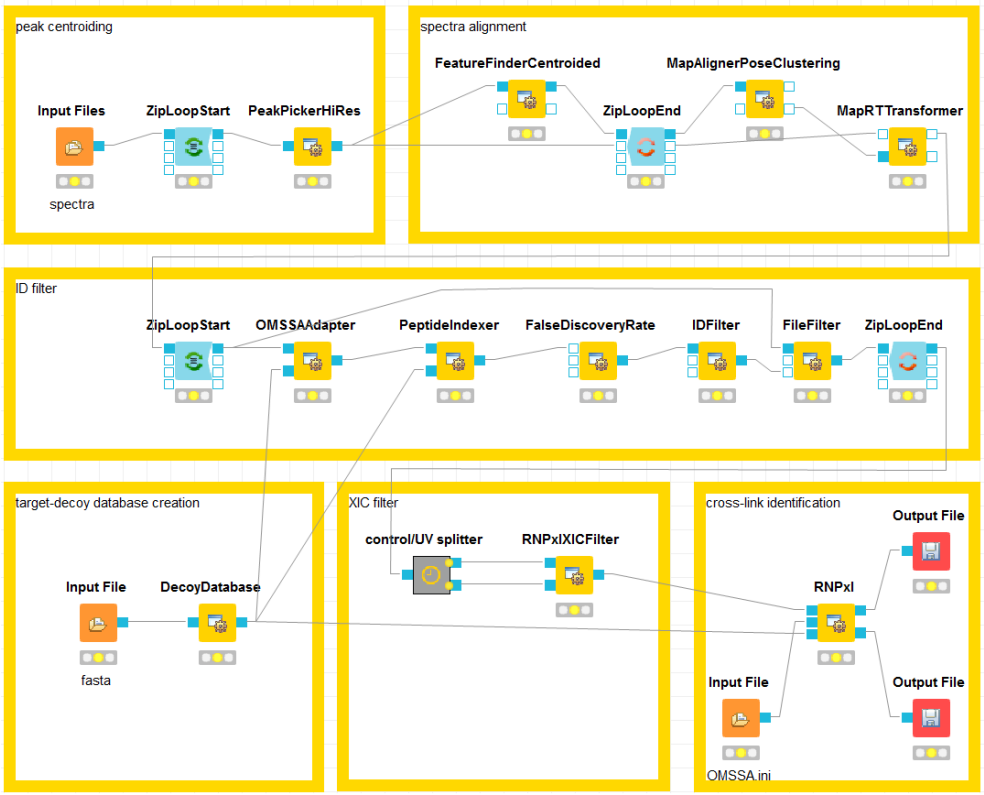


Figure C.4: OpenMS workflow for cross-Link identification using the RNP^{xl} tool in KNIME. Workflow consists of target-decoy database creation, peak centroiding, chromatographic alignment, ID/XIC filter and cross-link identification. Orange nodes indicate input files(s), yellow nodes TOPP tools, red nodes output files(s). Nodes are connected by edges that indicate the flow of data. Between corresponding ZipLoopStart/ZipLoopEnd nodes, a list of files is sequentially processed.

Table C.2: Cross-links to a uracil-containing RNA sequence.
The maximum length of generated oligonucleotides is limited to three. The composition of generated oligonucleotides is restricted by the provided sequence: only mono-, di- and trinucleotides are generated that form a valid substring. All standard nucleotides are defined via their sum formula and no further mapping is performed. Oligonucleotides without loss, loss of water, metaphosphoric acid as well as loss of both metaphosphoric acid and water are generated. "U=1" ensures that at least one uridine is present in the RNA combinations. In addition, the presence of the DTT cysteine adducts is considered for each RNA. *Adapted from Kramer et al.* [69].

Parameter	Value
length	3
sequence	ACUGCAUGAG
target_nucleotides	A=C10H14N5O7P, C=C9H14N3O8P, G=C10H14N5O8P, U=C9H13N2O9P
mapping	A->A, C->C, G->G, U->U
restrictions	A=0, C=0, G=0, U=1
modifications	-H2O, , -H2O-HPO3, -HPO3
CysteineAdduct	true

Table C.3: Cross-links with 4SU nucleotide analog substituted at a specific site.
4-thiouridine (defined as nucleotide "Y") is used to site-specifically label position 5 of the nucleotide sequence. All trinucleotides containing at least one 4SU are generated. Modifications are chosen according to the typical 4SU specific losses. *Adapted from Kramer et al.* [69].

Parameter	Value
length	2
sequence	ACUGYCAUGAG
target_nucleotides	A=C10H14N5O7P, C=C9H14N3O8P, G=C10H14N5O8P, U=C9H13N2O9P, Y=C9H13N2O8PS
mapping	A->A, C->C, G->G, U->U, Y->Y
restrictions	A=0, C=0, G=0, U=0, Y=1
modifications	-H2S, , -H2S-HPO3
CysteineAdduct	false

Table C.4: Cross-links with isotopically labeled adenosine.
Heavy adenosine (defined as nucleotide "Y") is used to site-specifically label position 5. All trinucleotides are generated without further restrictions. Standard losses are assumed. *Adapted from Kramer et al.* [69].

Parameter	Value
length	2
sequence	ACUGYUCAUGAG
target_nucleotides	A=C10H14N5O7P, C=C9H14N3O8P, G=C10H14N5O8P, U=C9H13N2O9P, Y=(13)C1(12)C9H14N5O7P
mapping	A->A, C->C, G->G, U->U, Y->Y
restrictions	A=0, C=0, G=0, U=0, Y=0
modifications	-H2O, , -H2O-HPO3, -HPO3
CysteineAdduct	false

Table C.5: Cross-linked proteins in human. *Adapted from Kramer et al.* [69].

type	protein	accession
RBPs	ELAV-like protein 1	Q15717
	Far upstream element-binding protein 2	Q92945
	H/ACA ribonucleoprotein complex subunit 4	O60832
	Heterogeneous nuclear ribonucleoprotein A/B	Q99729
	Heterogeneous nuclear ribonucleoprotein A1/A1-like 2	P09651/Q32P51
	Heterogeneous nuclear ribonucleoproteins A2/B1	P22626
	Heterogeneous nuclear ribonucleoproteins C1/C2	P07910
	Heterogeneous nuclear ribonucleoprotein D0	Q14103
	Heterogeneous nuclear ribonucleoprotein K	P61978
	Heterogeneous nuclear ribonucleoprotein L	P14866
	Heterogeneous nuclear ribonucleoprotein M	P52272
	Heterogeneous nuclear ribonucleoprotein Q	O60506
	Heterogeneous nuclear ribonucleoprotein R	O43390
	Heterogeneous nuclear ribonucleoprotein U	Q00839
	Nucleolin	P19338
	Nucleolysin TIAR	Q01085
	Poly(rC)-binding protein 1/2/3	Q15365/Q15366/P57721
	Poly(U)-binding-splicing factor PUF60	Q9UHX1
	Polypyrimidine tract-binding protein 1	P26599
	Putative pre-mRNA-splicing factor ATP-dependent RNA helicase DHX15	O43143
	Putative RNA-binding protein 3	P98179
	RNA-binding protein 39	Q14498
	RNA-binding protein FUS	P35637
	Serine/arginine-rich splicing factor 1	Q07955
	Serine/arginine-rich splicing factor 3	P84103
	Serine/arginine-rich splicing factor 5	Q13243
	Serine/arginine-rich splicing factor 6	Q13247
	Serine/arginine-rich splicing factor 9	Q13242
	Splicing factor U2AF 65 kDa subunit	P26368
	U1 small nuclear ribonucleoprotein 70 kDa	P08621
	Y-box-binding protein 1/2/3	P67809/Q9Y2T7/P16989
ribosomal subunits	40S ribosomal protein S2	P15880
	60S ribosomal protein L5	P46777
	60S ribosomal protein L6	Q02878
	60S ribosomal protein L34	P49207

Table C.6: Cross-linked proteins in yeast (standard uridine). *Adapted from Kramer et al.* [69].

type	protein	accession
metabolic enzymes	Adenosylhomocysteinase	P39954
	Alcohol dehydrogenase 1/3	P00330/P07246
	Enolase 1/2	P00924/P00925
	Glyceraldehyde-3-phosphate dehydrogenase 2/3	P00358/P00359
	Inorganic pyrophosphatase	P00817
	Peroxiredoxin TSA1	P34760
	Phosphoglycerate kinase	P00560
	Pyruvate kinase 1	P00549
DNA binding	Cruciform DNA-recognizing protein 1	P38845
nucleotide binding	Elongation factor 1-alpha	P02994
RNA binding	Nucleolar protein 3	Q01560
	Nucleolar protein 13	P53883
	Polyadenylate-binding protein	P04147
	Single-stranded nucleic-acid binding protein	P10080
40S small ribosomal subunit	40S ribosomal protein S1-A/-B	P33442/P23248
	40S ribosomal protein S3	P05750
	40S ribosomal protein S5	P26783
	40S ribosomal protein S11-A/-B	P0CX47/P0CX48
	40S ribosomal protein S14-A/-B	P06367/P39516
	40S ribosomal protein S16-A/-B	P0CX51/P0CX52
	40S ribosomal protein S17-A/-B	P02407/P14127
	40S ribosomal protein S24-A/-B	P0CX31/P0CX32
	40S ribosomal protein S29-A	P41057
	40S ribosomal protein S29-B	P41058
	Guanine nucleotide-binding protein subunit beta-like protein (RACK1)	P38011
60S large ribosomal subunit	60S ribosomal protein L1-A/-B	P0CX43/P0CX44
	60S ribosomal protein L2-A /-B	P0CX45/P0CX46
	60S ribosomal protein L3	P14126
	60S ribosomal protein L4-A	P10664
	60S ribosomal protein L4-B	P49626
	60S ribosomal protein L5	P26321
	60S ribosomal protein L6-A	Q02326
	60S ribosomal protein L6-B	P05739
	60S ribosomal protein L8-A	P17076
	60S ribosomal protein L8-B	P29453
	60S ribosomal protein L16-A	P26784
	60S ribosomal protein L16-B	P26785
	60S ribosomal protein L18-A/-B	P0CX49/P0CX50
	60S ribosomal protein L23-A/-B	P0CX41/P0CX42
	60S ribosomal protein L26-B	P53221
	60S ribosomal protein L28	P02406
	60S ribosomal protein L31-A/-B	P0C2H8/P0C2H9
	60S ribosomal protein L33-A/-B	P05744/P41056
	60S ribosomal protein L35-A/-B	P0CX84/P0CX85
	60S ribosomal protein L37-A	P49166
	60S ribosomal protein L37-B	P51402
	Ubiquitin-60S ribosomal protein L40	P0CH08/P0CH09
	60S ribosomal protein L42-A/-B	P0CX27/P0CX28
rRNA binding	Ribosome biogenesis protein RLP7	P40693

Table C.7: Cross-linked proteins in yeast (4-thiouridine). *Adapted from Kramer et al.* [69].

type	protein	accession
metabolic enzymes	Peptidyl-prolyl cis-trans isomerase	P14832
	Phosphoglycerate kinase	P00560
translation regulator	Cap-associated protein CAF20	P12962
DNA binding	Endonuclease PI-Scel1	P17255
	Multiprotein-bridging factor 1	O14467
	Non-histone chromosomal protein 6A	P11632
	Non-histone chromosomal protein 6B	P11633
	RNA polymerase II degradation factor 1	P35732
	Suppressor protein STM1	P39015
	Zuotin	P32527
nucleotide binding	Elongation factor 1-alpha	P02994
RNA binding	5'-3' exoribonuclease 1	P22147
	ATP-dependent RNA helicase DBP1/ ATP-dependent RNA helicase DED1	P24784/ P06634
	ATP-dependent RNA helicase SUB2	Q07478
	DNA topoisomerase 2-associated protein PAT1	P25644
	Elongation factor 2	P32324
	Eukaryotic translation initiation factor 4B	P34167
	Heterogeneous nuclear rnp K-like protein 2	P38199
	mRNA-binding protein PUF3	Q07807
	Negative growth regulatory protein NGR1	P32831
	Nuclear and cytoplasmic polyadenylated RNA-binding protein PUB1	P32588
	Nuclear polyadenylated RNA-binding protein 4	Q99383
	Nuclear segregation protein BFR1	P38934
	Nucleolar protein 3	Q01560
	Polyadenylate-binding protein, cytoplasmic and nuclear	P04147
	Protein SCP160	P06105
	Protein SSD1	P24276
	RNA annealing protein YRA1	Q12159
	RNA-binding protein NAB6	Q03735
	RNA-binding protein SRO9	P25567
	Single-stranded nucleic acid-binding protein	P10080
	Transposon Ty1-LR4 Gag polyprotein2	P0C2I8
ribosomal subunits	40S ribosomal protein S0-A/-B	P32905/P46654
	40S ribosomal protein S2	P25443
	40S ribosomal protein S3	P05750
	40S ribosomal protein S4-A/-B	P0CX35/P0CX36
	40S ribosomal protein S5	P26783
	40S ribosomal protein S7-A/-B	P26786/P48164
	40S ribosomal protein S8-A/-B	P0CX39/P0CX40
	40S ribosomal protein S14-A/-B	P06367/P39516
	40S ribosomal protein S15	Q01855
	40S ribosomal protein S17-A/-B	P02407/P14127
	40S ribosomal protein S22-A/-B	P0C0W1/Q3E7Y3
	40S ribosomal protein S24-A/-B	P0CX31/P0CX32
	40S ribosomal protein S26-A/-B	P39938/P39939
	40S ribosomal protein S30-A/-B	P0CX33/P0CX34
	60S ribosomal protein L4-A	P10664
	60S ribosomal protein L4-B	P49626
	60S ribosomal protein L6-A	Q02326
	60S ribosomal protein L6-B	P05739
	60S ribosomal protein L8-A	P17076
	60S ribosomal protein L8-B	P29453
	60S ribosomal protein L14-A/-B	P36105/P38754
	60S ribosomal protein L24-A/-B	P04449/P24000
	60S ribosomal protein L26-A/-B	P05743/P53221
	60S ribosomal protein L27-A/-B	P0C2H6/P0C2H7
	60S ribosomal protein L33-A/-B	P05744/P41056

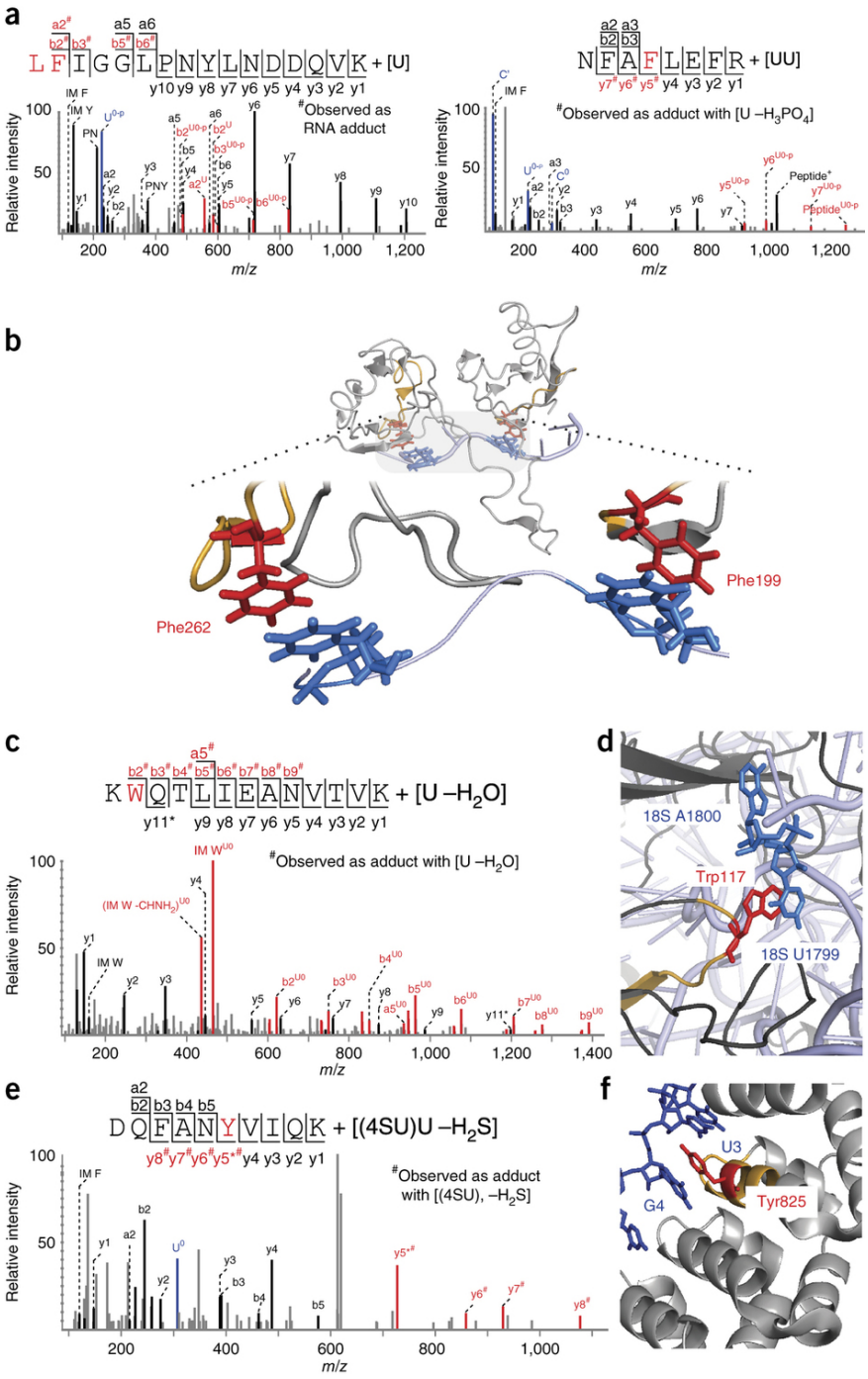


Figure C.5: Structural interpretation. MS/MS spectra with manual peak annotations: RNA fragments (blue) and shifted peptide fragment ions (red) in (a,c and e) have been used to validate cross-links. Cross-links are investigated in their structural context in (b,d and f) with cross-linked amino acids highlighted in red and nucleotides in blue. **a left:** Leu261 or Phe262 is cross-linked to U, **right:** Phe199 is cross-linked to UU. **b** Compared to an existing structure of U2AF with a poly(U)-nucleotide [174] , both Phe262 and Phe199 are found in close spatial proximity to uracil in the RRM1/2. **c** According to the MS/MS, Trp117 in 40S ribosomal protein S1 is cross-linked to U-H_2O. **d** In the 3D structure Trp117 is in close spatial proximity to U1799 of the 18S ribosomal RNA [85]. **e** The MS/MS identified Tyr825 in mRNA-binding protein Puf3 as cross-linked amino acid. **f** In the structure of Puf3 and a cocrystallized recognition sequence [175], Tyr825 is placed between U3 and G4. *Adapted from Kramer et al.* [69].

Table C.8: RNP^{xl}Search tool parameters.

Parameter	Description
in	Spectra (mzML)
database	Protein sequence database (fasta)
out	Output file (idXML)
Precursor (Parent Ion) Options:	
mass_tolerance	Precursor mass tolerance (+/- around precursor m/z)
mass_tolerance_unit	Unit of precursor mass tolerance. (valid: 'ppm', 'Da')
min_charge	Minimum precursor charge to be considered.
max_charge	Maximum precursor charge to be considered.
Fragments (Product Ion) Options:	
mass_tolerance	Fragment mass tolerance (+/- around fragment m/z)
mass_tolerance_unit	Unit of fragment m (valid: 'ppm', 'Da')
Modifications Options:	
fixed	Fixed modifications, e.g., 'Carbamidomethyl (C)'
variable	Variable modifications, e.g., 'Oxidation (M)'
variable_max_per_peptide	Maximum number of variable modifications per peptide
Peptide Options:	
min_size	Minimum size a peptide must after digestion.
missed_cleavages	Number of missed cleavages.
enzyme	The enzyme used for digestion.
Reporting Options:	
top_hits	Maximum number of hits per spectrum that are reported.
RNPxl Options:	
length	Oligonucleotide maximum length.
sequence	Sequence to restrict the generation of oligonucleotides.
target_nucleotides	Target nucleotides
mapping	Mapping rules.
restrictions	Restrictions.
fragment_adducts	Fragmentation adducts.
modifications	Format: empirical formula e.g -H2O, ..., H2O+PO3
CysteineAdduct	Use this flag if the +152 adduct from DTT is expected.
filter_fractional_mass	Filter non-crosslinks by fractional mass.
localization	Perform cross-link localization.
carbon_labeled_fragments	Generate fragment shifts assuming full labeling of carbon.
filter_small_peptide_mass	Filter non-crosslinks.
marker_ions_tolerance	Tolerance used to determine marker ions (Da).

Table C.9: Supported enzymes and cutting rules

Name (OpenMS)	Cutting Rule (Regular Expression)
Trypsin	(?<=[KR])(?!P)
Trypsin/P	(?<=[KR])
Lys-C	(?<=K)(?!P)
Lys-C/P	(?<=K)
Formic_acid	((?<=D))|((?=D))
TrypChymo	(?<=[KRFLWY])(?!P)
Chymotrypsin	(?<=[FLWY])(?!P)
Asp-N	(?=[BD])
PepsinA	(?<=[FL])
2-iodobenzoate	(?<=W)
Asp-N_ambic	(?=[DE])
Arg-C	(?<=R)(?!P)
glutamyl endopeptidase	(?>=[DE])
proline endopeptidase	(?>=[HKR]P)(?!P)
CNBr	(?<=M)
V8-DE	(?<=[BDEZ])(?!P)
V8-E	(?<=[EZ])(?!P)
leukocyte elastase	(?>=[AILV])(?!P)
unspecific cleavage	(?<=[A-Z])
no cleavage	()

Example: Explicitly specifying edges in the fragment adduct graph

Consider an experiment with uridine and the 4-thiouridine analog. Both give rise to different fragmentation adducts. Therefore, no common fragmentation behavior can be defined and we need to explicitly model the fragmentation adducts of uridine and 4-thiouridine.
Assume the character 'Y' was used as a placeholder for the 4SU nucleotide analog. 4SU only produces fragment with 4SU and loss of H_2S:

	precursor adduct		formula		annotation
string encoding:	Y	->	C9H11N2O8P	,	4SU-H2S

Fragment adducts for U are analogously specified using the explicit notation:

	precursor adduct		formula		annotation
string encoding:	U	->	C9H13N2O9P1	,	U,
	U	->	C9H11N2O8P1	,	U-H2O,
	U	->	C9H12N2O6	,	U-HPO3,
			...		

Appendix D: MetaProSIP

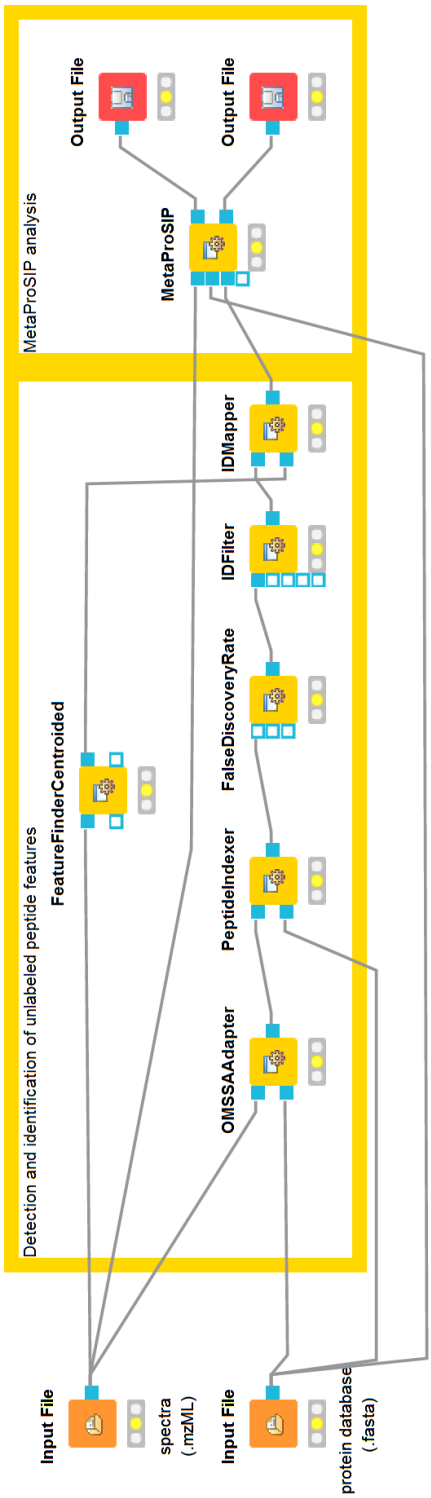


Figure D.1: Basic MetaProSIP OpenMS workflow for single file analysis without reference identifications. Peptide identifications are performed with the OMSSA search engine, followed by FDR calculation and filtering. Identifications are mapped to features and input into `MetaProSIP`. Result files are written as comma separated files. *TOPPAS workflow from Sachsenberg et al.*[110] *converted to KNIME.*

Table D.1: Basic MetaProSIP OpenMS workflow parameters compatible with a High-Resolution Orbitrap-type mass spectrometer.

Parameter	Value	Description
OMSSAAdapter (Peptide identification engine)		
precursor_mass_tolerance	10	precursor mass tolerance window for candidate peptides
precursor_mass_tolerance_unit_ppm	true	use relative tolerances (parts per million)
fragment_mass_tolerance	0.5	fragment mass tolerance window for candidate peptides (m/z)
fixed_modifications	Carbamidomethyl (C)	expect carbamidomethylation of cysteines from sample preparation
variable_modifications	Oxidation (M)	expect some methionine to be oxidized
IDFilter (Filtering to retain peptides at a given FDR)		
score:pep	0.02	filter results at a q-value threshold of 0.01 to obtain a FDR of 1%
FeatureFinderCentroided (Detection of eluting features)		
mass_trace:mz_tolerance	0.004	expected mass trace fluctuations
IDMapper (Mapping of identifications to features)		
rt_tolerance	30	tolerance (in seconds) for the matching of peptide identifications and features
mz_tolerance	20	m/z tolerance for the matching of peptide identifications and features.
mz_measure	ppm	use relative tolerances (parts per million)
MetaProSIP		
labeling_element	N	element used to introduce heavy isotopes
mz_tolerance_ppm	10	relative m/z tolerance for isotopic trace collection
rt_tolerance_s	30	tolerance (in seconds) for isotopic trace collection
correlation_threshold	0.7	minimum required pearson similarity with theoretical isotope pattern
use_unassigned_ids	false	whether precursor positions of unassigned peptides should be included
use_averagine_ids	false	whether averagine peptides should be included for unidentified masses
cluster	true	cluster by incorporation patterns

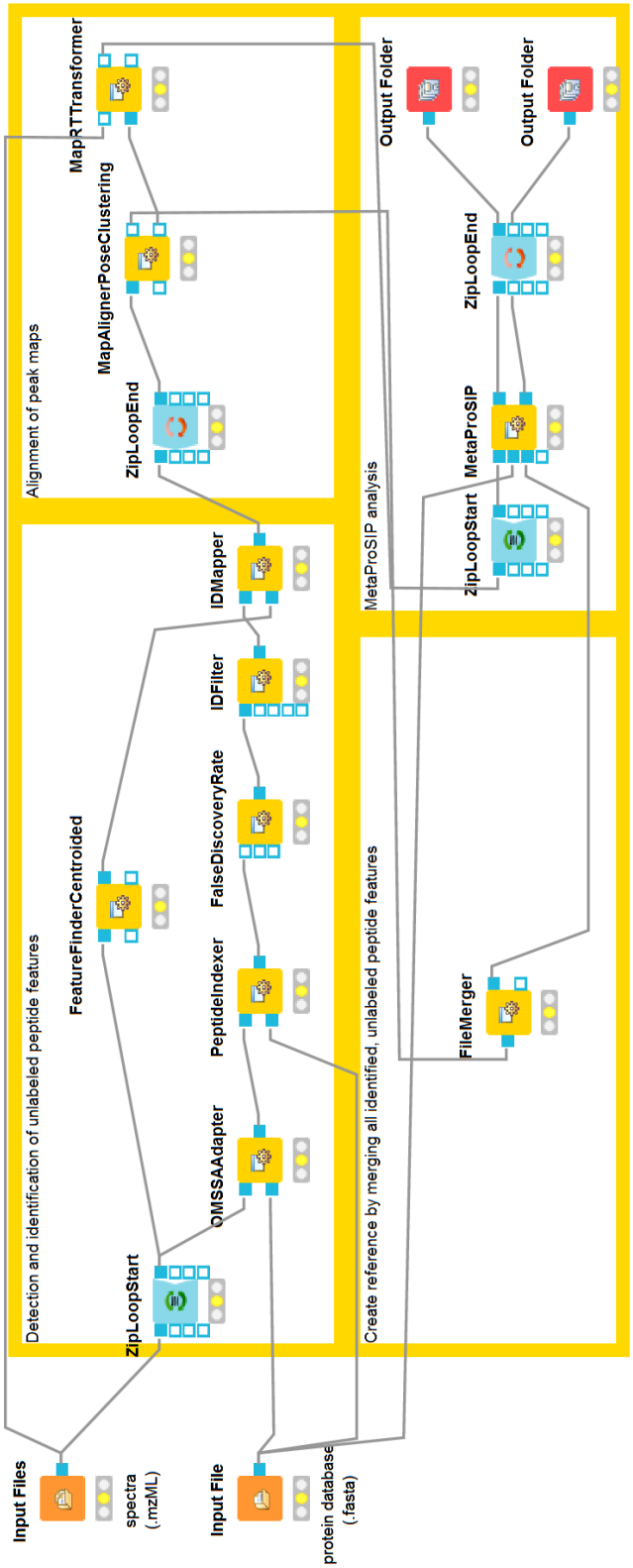


Figure D.2: MetaProSIP OpenMS workflow for time series analysis, including pooling of reference identifications. Peptide identifications (including FDR calculation and filtering) are mapped to features. After retention time alignment of features and peak maps, identified features are merged (pooled) and input into MetaProSIP. Result files are written as comma separated files. *Adapted from Sachsenberg et al.*[110] *and converted to KNIME.*

Table D.2: MetaProSIP OpenMS workflow configuration for a time series experiment with pooled reference features. Parameters are compatible with a High-Resolution Orbitrap-type mass spectrometer.

Parameter	Value	Description
OMSSAAdapter (Peptide identification engine)		
precursor_mass_tolerance	10	precursor mass tolerance window for candidate peptides
precursor_error_units	ppm	use relative tolerances (parts per million)
fragment_mass_tolerance	0.5	fragment mass tolerance window for candidate peptides (Dalton)
fixed_modifications	Carbamidomethyl (C)	expect carbamidomethylation of cysteines from sample preparation
variable_modifications	Oxidation (M)	expect some methionine to be oxidized
IDFilter (Filtering to retain peptides at a given FDR)		
score:pep	0.02	filter results at a q-value threshold of 0.01 to obtain a FDR of 1%
FeatureFinderCentroided (Detection of eluting features)		
mass_trace:mz_tolerance	0.004	expected mass trace fluctuations
IDMapper (Map identifications to detected features)		
rt_tolerance	30	tolerance (in seconds) for the matching of peptide identifications and features
mz_tolerance	20	m/z tolerance for the matching of peptide identifications and features.
mz_measure	ppm	use relative tolerances (parts per million)
MapAlignerPoseClustering (Chromatographic alignment)		
reference:index	1	select first file as reference in map alignment
MetaProSIP		
labeling_element	N or C	element used to introduce heavy isotopes
mz_tolerance_ppm	10	relative m/z tolerance for isotopic trace collection
rt_tolerance_s	30	tolerance (in seconds) for isotopic trace collection
correlation_threshold	0.7	minimum required pearson similarity with theoretical isotope pattern
use_unassigned_ids	false	whether precursor positions of unassigned peptides should be included
use_averagine_ids	false	whether averagine peptides should be included for unidentified masses
cluster	true	cluster by incorporation patterns
xic_threshold	-1	disable XIC filtering (no monoisotopic peak are expected at later time points)

Appendix E: MzTab

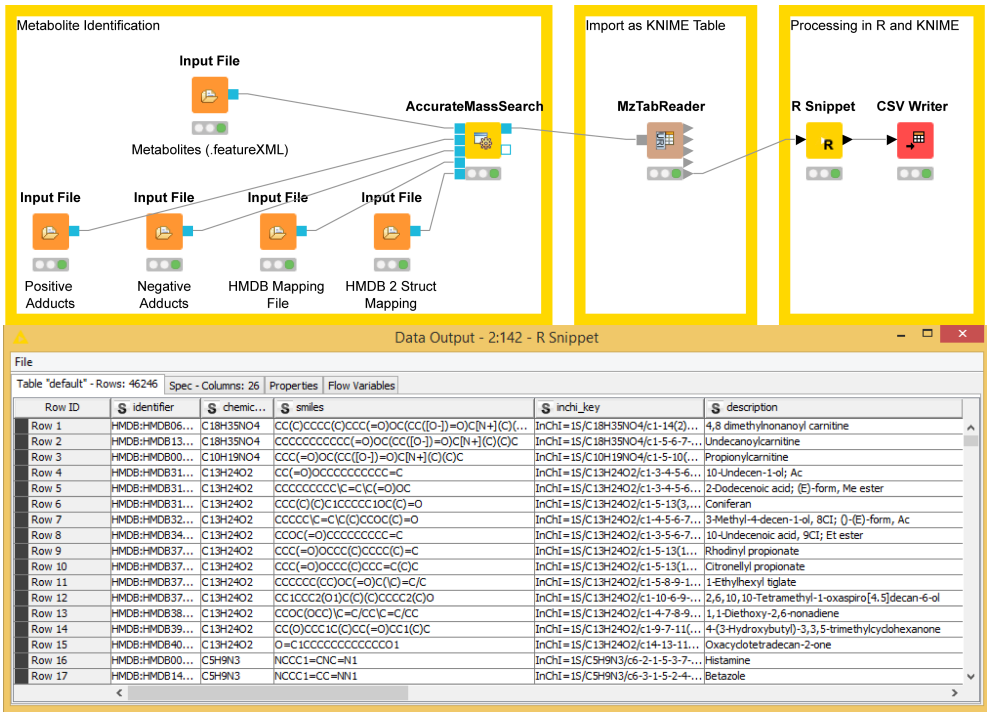


Figure E.1: MzTab workflow using OpenMS and R in KNIME. Identification of metabolite features using the OpenMS Accurate mass search results are directly exported to an mzTab file. An `MzTabReader` node reads the file and converts it to a standard KNIME table. In this example workflow, the KNIME `R snipped` node is used to further process the data. The final analysis results are written to a text file.

Appendix F:

Curriculum Vitae

Education

08/2010 – present

PhD candidate — Applied Bioinformatics (Prof. Oliver Kohlbacher), Center for Bioinformatics, **Eberhard Karls Universität Tübingen**, Germany

10/2007 – 09/2008

Diploma thesis — *Analysis of small RNA pathway mutants using whole genome tiling arrays*, supervised by Dr. Kay Nieselt (bioinformatics), Wilhelm-Schickard-Institute Eberhard Karls Universität Tübingen, Germany and Prof. Dr. Detlef Weigel (biology), Max Planck Institute for Developmental Biology, Germany

10/2000 – 04/2010

Bioinformatic studies — **Eberhard Karls Universität Tübingen**, Germany

Professional Work Experience

2006 – 2010

Software development (self-employed) — C++ und C# Application development in the field of computer vision (video-based 3D-Reconstruction)

Scientific Work Experience

2010 – present	
Software development (OpenMS core developer / release manager)	Applied Bioinformatics (Prof. Oliver Kohlbacher), Center for Bioinformatics, **Eberhard Karls Universität Tübingen**, Germany
2013/WS	
Teaching assistant	Tutoring of lecture *computational mass spectrometry*, Applied Bioinformatics (Prof. Oliver Kohlbacher), Center for Bioinformatics, **Eberhard Karls Universität Tübingen**, Germany
2011 – present	
Teaching assistant	Tutoring of practical courses accompanying the lecture *software engineering*, Applied Bioinformatics (Prof. Oliver Kohlbacher), Center for Bioinformatics, **Eberhard Karls Universität Tübingen**, Germany

Accepted manuscripts

2017

J Pfeuffer, **T Sachsenberg**, O Alka, M Walzer, A Fillbrunn, L Nilse, O Schilling, K Reinert, O Kohlbacher OpenMS - A platform for reproducible analysis of mass spectrometry data. *Journal of Biotechnology* 2017 May 27, PMID: 28559010, DOI: 10.1016/j.jbiotec.2017.05.016

FD Leprevost*, BA Grüning*, S Alves Aflitos, HL Röst, J Uszkoreit, H Barsnes, M Vaudel, P Moreno, L Gatto, J Weber, M Bai, RC Jimenez, **T Sachsenberg**, J Pfeuffer, R Vera Alvarez, J Griss, AI Nesvizhskii, Y Perez-Riverol BioContainers: An open-source and community-driven framework for software standardization. *Bioinformatics* 2017 Mar 30, PMID: 28379341, DOI: 10.1093/bioinformatics/btx192 (* *contributed equally*)

E Audain*, J Uszkoreit*, **T Sachsenberg**, J Pfeuffer, X Liang, H Hermjakob et al. In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *Journal of Proteomics* 150, 170-182, PMID: 27498275, DOI: 10.1016/j.jprot.2016.08.002 (* *contributed equally*)

2016

HL Röst*, **T Sachsenberg***, S Aiche*, C Bielow*, H Weisser* et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods* 13 (9), 741-748, PMID: 27575624, DOI: 10.1038/nmeth.3959 (* *contributed equally*)

J Veit, **T Sachsenberg**, A Chernev, F Aicheler, H Urlaub, O Kohlbacher LFQProfiler and RNPxl: Open-Source Tools for Label-Free Quantification and Protein–RNA Cross-Linking Integrated into Proteome Discoverer. *Journal of Proteome Research* 15 (9), 3441-3448, PMID: 27476824, DOI: 10.1021/acs.jproteome.6b00407

Y Perez-Riverol, L Gatto, R Wang, **T Sachsenberg**, J Uszkoreit et al. Ten Simple Rules for Taking Advantage of git and GitHub. *PLOS Computational Biology*, PMID: 27415786, DOI: 10.1371/journal.pcbi.1004947

2015

U Zaman, FM Richter, R Hofele, K Kramer, **T Sachsenberg**, O Kohlbacher et al. Dithiothreitol (DTT) Acts as a Specific, UV-inducible Cross-linker in Elucidation of Protein–RNA Interactions. *Molecular & Cellular Proteomics* 14 (12), 3196-3210, PMID: 26450613, DOI: 10.1074/mcp.M115.052795

K Sharma, A Hrle, K Kramer, **T Sachsenberg**, RHJ Staals, L Randau et al. Analysis of protein–RNA interactions in CRISPR proteins and effector complexes by UV-induced cross-linking and mass spectrometry. *Methods* 89, 138-148, PMID: 26071038, DOI: 10.1016/j.ymeth.2015.06.005

S Aiche, **T Sachsenberg**, E Kenar, M Walzer, B Wiswedel, T Kristl et al. Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry. *Proteomics* 15 (8), 1443-1447, PMID: 25604327, DOI: 10.1002/pmic.201400391

2014

T Sachsenberg*, FA Herbst*, M Taubert, R Kermer, N Jehmlich et al. MetaProSIP: automated inference of stable isotope incorporation rates in proteins for functional metaproteomics. *Journal of Proteome Research* 14 (2), 619-627, PMID: 25412983, DOI: 10.1021/pr500245w (* *contributed equally*)

K Kramer*, **T Sachsenberg***, BM Beckmann, S Qamar, KL Boon et al. Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nature Methods* 11 (10), 1064-1070, PMID: 25173706, DOI: 10.1038/nmeth.3092 (* *contributed equally*)

J Griss*, AR Jones*, **T Sachsenberg**, M Walzer, L Gatto, J Hartler et al. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & Cellular Proteomics* 13 (10), 2765-2775, PMID: 24980485, DOI: 10.1074/mcp.O113.036681 (* *contributed equally*)

2013

J Novák, **T Sachsenberg**, D Hoksza, T Skopal, O Kohlbacher On Comparison of SimTandem with State-of-the-Art Peptide Identification Tools, Efficiency of Precursor Mass Filter and Dealing with Variable Modifications. *Journal of Integrative Bioinformatics* 10 (3), 228, PMID: 24231142, DOI: 10.2390/biecoll-jib-2013-228

M Walzer, D Qi, G Mayer, J Uszkoreit, M Eisenacher, **T Sachsenberg** et al. The mzQuantML data standard for mass spectrometry–based quantitative studies in proteomics. *Molecular & Cellular Proteomics* 12 (8), 2332-2340, PMID: 23599424, DOI: 10.1074/mcp.O113.028506

S Nahnsen*, **T Sachsenberg***, O Kohlbacher PTMeta: Increasing identification rates of modified peptides using modification prescanning and meta-analysis. *Proteomics* 13 (6), 1042-1051, PMID: 23335442, DOI: 10.1002/pmic.201200315 (* *contributed equally*)

2010

S Laubinger*, G Zeller*, SR Henz, S Buechel, **T Sachsenberg**, JW Wang et al. Global effects of the small RNA biogenesis machinery on the Arabidopsis thaliana transcriptome. *Proceedings of the National Academy of Sciences* 107 (41), 17466-17473, PMID: 20870966, DOI: 10.1073/pnas.1012891107 (* *contributed equally*)

2009

G Zeller, SR Henz, CK Widmer, **T Sachsenberg**, G Rätsch, D Weigel et al. Stress-induced changes in the Arabidopsis thaliana transcriptome analyzed using whole-genome tiling arrays. *The Plant Journal* 58 (6), 1068-1082, PMID: 19222804, DOI: 10.1111/j.1365-313X.2009.03835.x

2008

S Laubinger, G Zeller, SR Henz, **T Sachsenberg**, CK Widmer, N Naouar et al. At-TAX: a whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. *Genome Biology* 9 (7), 1, PMID: 18613972, DOI: 10.1186/gb-2008-9-7-r112

S Laubinger, **T Sachsenberg**, G Zeller, W Busch, JU Lohmann, G Rätsch et al. Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* 105 (25), 8795-8800, PMID: 18550839, DOI: 10.1073/pnas.0802493105