Novel Computational Methods for the Analysis and Interpretation of MS/MS Data in Metaproteomics

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

von	Dipl.-Bioinf. Thilo Muth
geboren am	19. Oktober 1981 in Würzburg

genehmigt durch die Fakultät für Verfahrens- und Systemtechnik
der Otto-von-Guericke-Universität Magdeburg

Promotionskommission:

Prof. Dr. Helmut Weiß
Prof. Dr.-Ing. Udo Reichl
Prof. Dr. Lennart Martens
Dr. Frédérique Lisacek

eingereicht am:	29. April 2016
Promotionskolloquium am:	26. Oktober 2016

Danksagung

Zunächst möchte ich mich herzlich bei Herrn Prof. Dr.-Ing. Udo Reichl bedanken, der es mir ermöglichte, die Promotion in der Arbeitsgruppe Bioprozesstechnik am Max-Planck-Institut für Dynamik komplexer technischer Systeme in Magdeburg in der Zeit von Juni 2011 bis Dezember 2015 durchzuführen. Besonders verdanke ich ihm die unermüdliche Unterstützung und die gewährte Freiheit bei der Bearbeitung eines hochinteressanten Themas aus dem Bereich der Bioinformatik. Sein kompetenter wissenschaftlicher Rat trug bei zahlreichen konstruktiven Diskussionen maßgeblich zum Gelingen dieser Arbeit bei.

Desweiteren gebührt mein besonderer Dank Dr. Erdmann Rapp, der mich während der Promotion mehrfach mit hervorragenden Ideen für neuartige Themenfelder in der Bioprozessanalytik begeistern konnte. Wertvolle Unterstützung erfuhr ich dabei von ihm sowohl auf der beruflichen als auch auf der menschlichen Ebene während meiner Zeit am Max-Planck-Institut.

Besonders bedanken möchte ich mich außerdem bei Prof. Dr. Lennart Martens für seine ausgezeichnete fachliche Beratung und Ideengebung zu bioinformatischen Themen während der Durchführung dieser Arbeit.

Meinen Kollegen Marcus Hoffmann und Rene Hennig danke ich für unzählige Gepräche über fachliche Themen aus der experimentellen Analytik sowie die freundschaftliche Aufnahme in die Arbeitsgruppe.

Außerdem bedanke ich mich sehr bei Dirk Benndorf, Fabian Kohrs, Robert Heyer und Carolin Kolmeder, die mir mit ihren Erfahrungen im Bereich der Metaproteomik zur Seite standen.

Ein großer Dank gilt ebenso an Alexander Behne und Sebastian Dorl, die durch ihr stetes Engagement im Rahmen ihrer studentischen Abschlußarbeiten einen wichtigen Grundstein für Teile dieser Arbeit legten.

Zusätzlich gilt größter Dank den Kollegen in der Bioprozesstechnik-Gruppe und am gesamten Max-Planck-Institut für die bereichernde Zeit über die letzten Jahre hinweg. Dabei danke ich vielen Menschen, mit denen mich eine langjährige Freundschaft verbindet, inbesondere Steffen Riethmüller, Matthias Meininger, Verena Lohr, Terry Nguyen-Khuong, Michael Pieler, Thomas Bissinger, Pavel Marichal-Gallardo und Heiko Weichelt.

An dieser Stelle gebührt besonderer Dank meiner Familie und den besten Freunden, die mich in den vergangenen Jahren in allen Phasen der Promotion unterstützen konnten.

Abschließend möchte ich meiner Verlobten Karen danken, deren unendliche Geduld und aufmunternde Worte sehr zur Vollendung dieser Arbeit beitrugen.

Abstract

Microbial populations contribute strongly to the total biomass on Earth and are important key players in biochemical cycles. Microorganisms are also essential in biotechnological applications and represent the majority of cells in the human body. While the importance of microbial consortia for life and human health is increasingly recognized, this growing interest demands the holistic analysis of microbial communities. To that extent, metagenomic and metatranscriptomic approaches provide a solid blueprint, however, assessing the actual phenotypes requires the analysis of protein expression. Since the first large-scale proteome analysis of a microbial community one decade ago, metaproteomics has advanced as an indispensable tool for the detailed investigation of expression profiles in microbial samples: mass spectrometry-based proteomics provides insights into potential functions and enzymatic capabilities of microorganisms from different environmental conditions and habitats. Although many technical improvements have been made regarding the analytical tools, various severe challenges related to processing, evaluation and interpretation of high-throughput data remain unresolved in this field. These issues are mainly caused by the lack of standardized software and the limited integration of available methods from bioinformatics. Accordingly, the aim of this work was to identify the major challenges for the analysis of metaproteomic data and to provide solutions to these encountered issues.

The first part of this work presents the MetaProteomeAnalyzer, an open-source software which was developed to analyze and interpret comprehensive data sets from metaproteomic experiments. The tool includes and combines multiple search algorithms for the identification of proteins from tandem mass spectra. The server side features the automated integration of relevant taxonomic and functional meta-information for the identified proteins. The client application then allows to examine the microbial community composition and to detect key enzymes in metabolic pathways. To tackle data redundancy and protein inference issues, different rules were implemented to group protein hits to so-called meta-proteins. The software tool holds an intuitive graphical user interface with various visualization and categorization features to facilitate a detailed and unbiased exploration of the data. To handle complex questions, the included graph database back-end further extends the predefined presentation of the results by providing a user-definable query system.

The second part of this work focuses on the identification of typical bottlenecks and shortcomings which frequently arise during the data analysis in metaproteomics. Therefore, the influence of search algorithm, protein database, and enzymatic cleavage parameters on the identification outcome is evaluated by investigating metaproteomic data sets from biogas plant and human intestine samples. The results show that combining the search algorithms X!Tandem and

OMSSA as well as searching against subsets of a typical metaproteome sequence database significantly improve the identification yield. Furthermore, both metagenomic and public sequence databases result in unique peptide identifications, indicating that parallel searches against different resources lead to an information gain in terms of the proteome content of microbial samples. A benchmark experiment based on *Pyrococcus furiosus* proteome data demonstrates that identifications are lost due to an FDR overestimation in target-decoy-based searches against large protein databases. The increase of missed cleavage parameter values reduces the number of identifications for metaproteomic data sets and semi-tryptic searches fail to provide a significant gain in identifications. In addition, *de novo* sequencing is applied as alternative identification method to conventional database searching. It is shown that the outcome of *de novo* sequencing cannot justify the effort due to the low overlap with the corresponding results of database searching.

The final part of this work addresses the essential requirement of methods beyond the scope of common identification workflows in proteomics. First of all, the meta-protein generation approach is evaluated that was developed to group redundant protein hits by means of a provided set of different rules. It is shown that the grouping rule based on one shared peptide leads to the greatest redundancy reduction within a single result set and increases the comparability of results from different experiments. Next, the compliance of the taxonomic assignment process is tested using identifications derived from a mixture sample of known organism composition. In comparison to Unipept, the MPA software achieves significantly more correct assignments of taxon-specific peptides across the evaluated data sets. Regarding the analysis of human intestine samples, no significant taxon-specific abundance differences are found between two investigated groups (obese/non-obese). The results from further investigations dealing with the mapping of protein identifications into metabolic pathways suggest the combination of phylogenetic and functional annotation to increase the information content in metaproteomic data. Moreover, the analysis of the biogas samples reveals that the public databases SwissProt and TrEMBL are complementary regarding the assignment of identifications to taxonomic and functional annotations. The work is concluded by performing a supervised classification method for the results of the human intestine samples which detects 27 functional groups of Bacteria that differ significantly between the result sets of lean and obese individuals.

Overall, the investigations in this work highlight the importance of appropriate bioinformatic methods and protein databases to overcome limitations of the data analysis in metaproteomics. To that end, a dedicated software is presented for the processing and interpretation of metaproteomic data and recommendations regarding an optimized data analysis workflow are provided based on the knowledge gained from this work.

Zusammenfassung

Mikrobielle Gemeinschaften haben einen bedeutenden Anteil an der Gesamtbiomasse der Erde und besitzen eine Schlüsselrolle in biochemischen Kreisläufen. Mikroorganismen sind zudem wichtig für biotechnologische Anwendungen und machen einen Großteil der Zellen im menschlichen Körper aus. Während die Bedeutung von mikrobiellen Konsortien für das Leben und die menschliche Gesundheit zunehmend wahrgenommen wird, erfordert dieses gesteigerte Interesse gleichzeitig auch die Analyse von mikrobiellen Gemeinschaften als Gesamtheit. Ansätze, die auf Metagenomik und Metatranskriptomik basieren, stellen ein solides Grundgerüst bereit, jedoch können Aussagen über die tatsächlichen Phänotypen nur mithilfe der Proteinexpression getroffen werden. Seit der ersten groß angelegten Proteomanalyse einer mikrobiellen Gemeinschaft vor einem Jahrzehnt stellt die Metaproteomik ein unverzichtbares Werkzeug für die genaue Untersuchung von Expressionsprofilen in mikrobiellen Proben dar: auf Massenspektrometrie basierende Proteomik gewährt Einblicke in mögliche Funktionen und enzymatische Aktivitäten der Mikroben für unterschiedliche Umweltbedingungen und Habitate. Obwohl es viele technische Verbesserungen im Bereich der experimentellen Analyse gegeben hat, sind viele ernstzunehmende Probleme in Bezug auf Prozessierung, Auswertung und Interpretation der Hochdurchsatzdaten immer noch ungelöst. Diese Schwierigkeiten liegen hauptsächlich an einem Mangel an standardisierter Software und der unzureichenden Evaluierung von vorhandenen bioinformatischen Methoden. Das Ziel dieser Arbeit ist es, die größten Hürden der computergestützten Auswertung aufzuzeigen und Lösungen für die Analyse von Metaproteomdaten bereitzustellen.

Im ersten Teil dieser Arbeit wird die Open-Source Software MetaProteomeAnalyzer vorgestellt, welche entwickelt wurde, um umfangreiche Datensätze von metaproteomischen Experimenten auszuwerten. Das Programm enthält und vereint mehrere Suchalgorithmen zur Identifizierung von Proteinen aus Massenspektren. Die serverseitige Anwendung stellt die automatische Integration von relevanten taxonomischen und funktionellen Metainformationen für die identifizierten Proteine bereit. Die Client-Software ermöglicht es, die Zusammensetzung der mikrobiellen Gemeinschaft zu untersuchen und wichtige Enzyme in metabolischen Netzwerken zu identifizieren. Um die Problematik der Datenredundanz und des Rückschlusses auf Proteine anzugehen, wurde die auf verschiedenen Regeln basierende Gruppierung von Proteinen zu sogenannten Meta-Proteinen implementiert. Die Software besitzt eine benutzerfreundliche grafische Oberfläche, die zahlreiche Möglichkeiten zur Visualisierung und Kategorisierung erlaubt, um eine genaue und unverfälschte Untersuchung der Daten zu ermöglichen. Für komplexere Fragestellungen erweitert die zugrunde liegende Graphdatenbank die vorgegebene Darstellung der

Ergebnisse, indem sie ein vom Benutzer definierbares Abfragesystem bereitstellt.

Der zweite Teil der Arbeit konzentriert sich darauf, typische Engpässe und Unzulänglichkeiten aufzuzeigen, welche häufig während der Datenanalyse in der Metaproteomik auftreten. Dazu wird der Einfluss des Suchalgorithmus, der Proteindatenbank und der Parameter des enzymatischen Verdaus auf das Ergebnis bei der Identifizierung mittels Metaproteomdaten von Proben aus Biogasanlagen und aus dem menschlichen Darmtrakt getestet. Die Ergebnisse zeigen, dass die Kombination der Suchalgorithmen X!Tandem und OMSSA sowie die Suche gegen Teile einer typischen Metaproteom-Sequenzdatenbank signifikant die Gesamtzahl der Identifikationen erhöhen. Außerdem führen sowohl metagenomische als auch öffentlich zugängliche Sequenzdatenbanken zu spezifischen Peptididentifikationen, was darauf hindeutet, dass parallele Suchen gegen verschiedene Quellen zu einem Informationsgewinn bezüglich der proteomischen Zusammensetzung von mikrobiellen Proben führen. Ein Experiment basierend auf *Pyrococcus furiosus* Proteomdaten zeigt, dass Identifikationen wegen einer Überschätzung der *False Discovery Rate* in *Target-Decoy*-basierten Suchen gegen große Datenbanken verloren gehen. Die Erhöhung von Werten des *Missed Cleavages*-Parameters führt zu einer verringerten Anzahl an Identifikationen für metaproteomische Datensätze und semi-tryptische Suchen können keine signifikante Erhöhung der Treffer erzielen. Zusätzlich wird *de novo*-Sequenzierung als alternative Identifizierungsmethode zur gewöhnlichen Datenbanksuche verwendet. Dabei wird deutlich, dass die Ergebnisse der *de novo*-Sequenzierung den betriebenen Aufwand wegen der geringen Überschneidung mit den jeweiligen Ergebnissen der Datenbanksuchen nicht rechtfertigen können.

Der letzte Teil der Arbeit bezieht sich auf den grundlegenden Bedarf von Ansätzen, die über den Prozessschritt der Proteinidentifizierung hinausgehen. Zunächst wird dabei die Methode zum Generieren von Meta-Proteinen evaluiert, welche entwickelt wurde, um redundante Proteintreffer anhand von einem vorgegebenen Regelwerk zu gruppieren. Es wird aufgezeigt, dass die Gruppierungsmethode, welche auf einem geteilten Peptid basiert, zur höchsten Reduzierung der Redundanz innerhalb einer einzelnen Ergebnismenge führt und damit die Vergleichbarkeit von Resultaten aus verschiedenen Experimenten erhöht. Daraufhin wird die Anwendbarkeit der taxonomischen Zuordnungsmethode mit Hilfe von Identifikationen aus einer Mischprobe mit bekannter Zusammensetzung der Organismen überprüft. Die MPA Software erzielt dabei signifikant mehr richtige Zuweisungen von Peptiden zu Taxa für die getesteten Datensätze im Vergleich zu Unipept. In Bezug auf die Analyse von menschlichen Darmproben können keine signifikanten taxon-spezifischen Unterschiede in der Abundanz zwischen zwei untersuchten Gruppen (adipös/nicht-adipös) gefunden werden. Die Resultate weiterer Untersuchungen, die sich mit dem direkten Abbilden von Proteinidentifikationen in Stoffwechselwege beschäftigen, legen nahe, dass eine Kombination von phylogenetischer und funktioneller Annotierung notwendig ist, um den Informationsgehalt von Metaproteomdaten zu erhöhen. Außerdem zeigt die

Analyse der Biogasproben, dass die öffentlichen Datenbanken SwissProt und TrEMBL komplementär sind bei der Zuweisung von Identifikationen zu taxonomischen und funktionellen Annotationen. Die Arbeit wird durch die Anwendung einer überwachten Klassifizierungsmethode anhand der Ergebnisse der menschlichen Darmproben abgeschlossen, bei der 27 funktionelle Gruppen von Bakterien gefunden werden, welche sich signifikant zwischen normalgewichtigen und übergewichtigen Probanden unterscheiden.

Insgesamt stellen die Untersuchungen dieser Arbeit die Wichtigkeit von geeigneten bioinformatischen Methoden und Proteindatenbanken heraus, um Engpässe bei der metaproteomischen Datenanalyse zu überwinden. Dazu wird eine spezielle Software für die Prozessierung und Interpretation von Metaproteomdaten vorgestellt, und es werden anhand der gewonnenen Erkenntnisse aus dieser Arbeit Empfehlungen bezüglich eines optimierten Workflows zur Datenanalyse von metaproteomischen Proben gegeben.

Contents

Abstract **VI**

Zusammenfassung **IX**

List of Abbreviations **XVI**

1 Introduction **1**

2 Theoretical Background **5**

2.1 Analysis of Microbial Communities 5

2.1.1 Role of Microbial Communities in Humans 6

2.1.2 Microbial Analysis Techniques 7

2.1.3 Beyond the Genome to the Proteome 8

2.1.4 Microbial Community Proteomics 10

2.1.5 Experimental Bottom-Up Workflow 13

2.2 Data Analysis Workflow in Metaproteomics 17

2.2.1 Filtering and Clustering of MS/MS Spectra 18

2.2.2 Tailor-Made Database Construction 18

2.2.3 Protein Identification by Database Searching 19

2.2.4 *De Novo* Sequencing and Homology Search 21

2.2.5 Protein Inference and Taxonomic Assignment 22

2.2.6 Functional and Metabolic Pathway Analysis 23

2.2.7 Protein Quantification Methods 24

2.2.8 Data Storage and Online Data Repositories 25

3 Material and Methods **27**
3.1 MetaProteomeAnalyzer . . . 28
3.1.1 Software Workflow . . . 28
3.1.2 Meta-Protein Generation . . . 31
3.1.3 Graph Database System . . . 34
3.2 Experimental Data . . . 36
3.2.1 Biogas Plant Samples . . . 37
3.2.2 Human Intestine Metaproteomes . . . 37
3.2.3 Pyrococcus Furiosus . . . 38
3.2.4 Mixture of Nine Organisms . . . 38
3.3 Protein Sequence Databases . . . 39
3.3.1 UniProtKB (SwissProt/TrEMBL) . . . 39
3.3.2 Biogas Plant Metagenome (BGPMG) . . . 39
3.3.3 Human Intestinal Metaproteome Database (HIMPdb) . . . 40
3.3.4 Pyroccoccus Furiosus Database (Pyrodb) . . . 40
3.4 Employed Software . . . 41
3.4.1 X!Tandem . . . 41
3.4.2 OMSSA . . . 41
3.4.3 MASCOT . . . 41
3.4.4 DeNovoGUI . . . 42
3.4.5 Unipept . . . 43
3.4.6 EggNOG . . . 43
3.4.7 LEfSe . . . 44
3.5 Applied Methods . . . 44
3.5.1 Target-Decoy Approach . . . 44
3.5.2 Quality Control and Results Combination . . . 45
3.5.3 Identification Rescoring . . . 46
3.5.4 Two-Step Searching . . . 46
3.5.5 Jaccard Index . . . 46

4 Results **47**
4.1 Search Algorithm Comparison . . . 47
4.1.1 Preliminary Analysis . . . 48
4.1.2 Performance of X!Tandem and OMSSA . . . 49
4.2 Database Searching . . . 50
4.2.1 Influence of Protein Database . . . 50

4.2.2 Evaluation of Search Strategies . . . 53
4.2.3 Missed Cleavage Parameter Testing . . . 58
4.2.4 Non-Tryptic Enzyme Settings . . . 60
4.2.5 Benchmark Evaluation of Proteomic Sample . . . 61
4.3 *De Novo* Sequencing . . . 66
4.3.1 Method Evaluation and Identification Recall . . . 66
4.3.2 Comparison of Classic and Two-step Searching . . . 67
4.4 Protein Grouping . . . 70
4.4.1 Testing Meta-Protein Generation Rules . . . 70
4.4.2 Evaluating Reproducibility between Replicates . . . 73
4.4.3 Comparing Data Sets from Different Samples . . . 77
4.5 Taxonomic Assignment . . . 80
4.5.1 Influence of Protein Database . . . 80
4.5.2 Assignment Performance Evaluation . . . 81
4.5.3 Phylogenetic Overview on Human Intestine Microbiota . . . 86
4.6 Functional Analysis . . . 90
4.6.1 Methods of Functional Annotation . . . 90
4.6.2 Quantifying the Functional Profile . . . 95
4.6.3 Postprocessing Unannotated Data . . . 99

5 Discussion 105
5.1 Combining Multiple Search Algorithms . . . 105
5.2 Evaluating Parameters of Database Searching . . . 107
5.2.1 Influence of Protein Database . . . 107
5.2.2 Evaluation of Search Strategies . . . 109
5.2.3 Missed Cleavages and Enzyme Specificity . . . 112
5.3 Testing Performance of *De Novo* Sequencing . . . 114
5.4 Generating Meta-Proteins by Protein Grouping . . . 117
5.5 Investigating Techniques of Taxonomic Assignment . . . 120
5.5.1 Influence of Protein Database . . . 120
5.5.2 Assignment Performance Evaluation . . . 121
5.5.3 Phylogenetic Overview on Human Intestine Microbiota . . . 123
5.6 Assessing Methods of Functional Analysis . . . 125
5.6.1 Elucidating Functional Annotation Methods . . . 125
5.6.2 Quantifying the Functional Profile . . . 128
5.6.3 Postprocessing Unannotated Data . . . 129

6 Conclusion and Outlook **133**

List of Figures **141**

List of Tables **145**

List of Contributions **147**

Bibliography **175**

Appendices **177**

List of Abbreviations

BGP	Biogas plant
BGPMG	Biogas plant metagenome
BLAST	Basic local alignment search tool
BMI	Body mass index
DNA	Deoxyribonucleic acid
EC	Enzyme commission
ED	Edit distance
EggNOG	Evolutionary genealogy of genes: non-supervised orthologous groups
ESI	Electrospray ionization
FASP	Filter-aided sample preparation
FDR	False discovery rate
FP	False positives
HIMP	Human intestine metaproteome
KEGG	Kyoto encyclopedia of genes and genomes
KO	KEGG orthology
LC	Liquid chromatography
LCA	Lowest common ancestor
LDA	Linear discriminant analysis
LIMS	Laboratory information management system
MALDI	Matrix-assisted laser desorption/ionization
MC	Missed cleavages
MPA	MetaProteomeAnalyzer
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MST	Most specific taxonomy
NCBI	National center for biotechnology information

NOG	Non-supervised orthologous group
NSAF	Normalized spectral abundance factor
PFU	Pyrococcus furiosus
PPID	Protein precipitation followed by in-solution digestion
PSM	Peptide-spectrum match
PTM	Post-translational modification
RDA	Redundancy analysis
RMIC	Relative matched ion count
RNA	Ribonucleic acid
SQL	Structured query language
TDA	Target-decoy approach
TIC	Total ion current
TP	True positives

1

Introduction

Microorganisms account for the major proportion of biomass on Earth and are omnipresent in any environment. Microbes exhibit a remarkable degree of diversity and commonly live as complex communities in natural habitats [1]. These consortia are essential in geochemical cycles, renewable energy production, waste water treatment, agricultural and biotechnological applications [2, 3, 4, 5]. Moreover, the analysis of microbial communities is highly relevant for human and animal health where microbes have a beneficial or—in rare cases—harmful role to their hosts [6]. In contrast to pure culture studies, the holistic approach of studying complex microbial communities increases the chance to decipher the interactions between hundreds or thousands of different species and the environment with the ultimate goal to gain comprehensive knowledge about their functions in diverse ecosystems.

Latest advances in high throughput DNA sequencing have provided exciting opportunities to study a microbial population in its ecological habitat by means of metagenomic techniques [7]. While investigations at the genomic and transcriptomic level provide valuable insights into the genetic diversity and taxonomic composition of microbial consortia, the protein expression profile cannot be readily assessed by these approaches. By characterizing the entire set of expressed proteins of environmental microbiota at a given time point, metaproteomics—also referred to as whole community or environmental proteomics [8, 9]—aims to examine the functional components of a microbial ecosystem. Thereby, the application of proteomic techniques for analyzing samples of microbial communities allows to investigate potential metabolic activities carried out by these consortia [10].

In recent years, microbial community proteomics has been driven forth by enormous advances concerning analytical tools, in particular, by the rapidly evolving technological platform of mass spectrometry. Compared to pure-culture proteomics, however, metaproteomic research poses several unique challenges. In particular, samples of microbial communities are complex and heterogeneous exhibiting highly dynamic protein expression levels. Although various improvements regarding computational methods have facilitated the analysis of single-organism proteomic data over the last decade, metaproteomics is still an untapped field which lacks the detailed evaluation of database search algorithms and parameter selection. For instance, protein identification algorithms are designed to process single-organism samples and are therefore challenged by size and redundant composition of microbial sequence databases. Moreover, a critical obstacle presents the protein inference problem [11] which is more difficult to resolve in the context of metaproteomics due to the high amount of shared peptides found in homologous proteins from different organisms. Eventually, only a low proportion of microbial genomes has been sequenced, which negatively impacts the computational analysis of metaproteomic data: the lack of appropriate sequence databases is a serious bottleneck for the identification of proteins from microbial samples and results in a large proportion of unidentified tandem mass spectra. Beyond protein identification, the taxonomic assignment and functional annotation of the results in microbial data sets gains relevance: the integration of meta-information, referring to phylogenetic origin and involvement in metabolic pathways, is essential for conducting in-depth analyses, such as understanding the biological processes in biogas and wastewater treatment plants [12, 13] or examining the enzymatic interactions of complex microbial communities in the human gastrointestinal tract [14, 15].

Due to the lack of appropriate computational methods, the analysis and interpretation of data constitute the major bottleneck for metaproteomic research. The main goal of this work is to tackle the aforementioned challenges by developing a data analysis pipeline that is tailored towards samples from microbial communities. The objectives for such a software workflow are to integrate multiple protein identification algorithms, to provide a user-friendly, yet powerful processing and storage framework for high volumes of upcoming data, and to enable a comprehensive analysis of microbial community samples at the taxonomic and functional level. Moreover, the aforementioned inference issue should be addressed by the integration of meaningful protein grouping strategies. Mainly, the dedicated software aims at analyzing and interpreting metaproteomic data sets originating from microbial community samples: for this purpose, metaproteomic data sets derived from biogas plant and human feces samples are investigated in this work. Using this kind of data, conventional and alternative computational methods are evaluated to identify shortcomings and provide recommendations for optimized parameter selection and analysis strategies in metaproteomics. The last objective focuses on the detailed anal-

ysis of results from processed metaproteomic data and regards essential steps that come after the identification of proteins, namely, approaches for protein grouping, taxonomic assignment and functional annotation. Therefore, the performance of computational methods developed in this work as well as of external software tools is tested using aforementioned data sets derived from metaproteomic samples.

This work is structured in six chapters. After explaining the motivation and aim of the work in this chapter, relevant background information on the analysis of microbial communities and available computational methods is given in Chapter 2. Chapter 3 describes the developed data analysis pipeline and provides further details on employed software tools, applied methods, protein sequence databases and experimental data sets. In Chapter 4, the results are presented; Section 4.1 starts with findings on the performance of different database search algorithms used for the identification of metaproteomic data sets. In Section 4.2, search algorithm parameters, such as chosen protein database and cleavage enzyme, are evaluated, in particular, regarding their effect on the identification yield. In Section 4.3, the outcome of *de novo* sequencing as alternative peptide identification method is shown. In Section 4.4, the performance of diverse strategies for the grouping of protein results is examined. Furthermore, the developed grouping rules are evaluated by comparing results of replicate and multiple data sets from different experiments. Section 4.5 focuses on the assignment of identifications to taxonomic groups. Section 4.6 ends the chapter by evaluating different methods for functional annotation. Chapter 5 provides a detailed discussion of the outcome of this work and is structured in accordance to the outline of the previous results chapter. Finally, conclusion and outlook of the work are given in Chapter 6.

2

Theoretical Background

2.1 Analysis of Microbial Communities

Microorganisms represent the oldest and genetically most diverse life forms on earth. The total number of prokaryotes has been estimated to be around $5x10^{30}$ cells and this high amount of cells outnumbers by far all other organisms [16]. Most of these microorganisms are reported to occur in soil [17] and in global oceans [18], but are also widespread in terrestrial and oceanic subsurface regions – even in the most inhospitable locations on Earth [19, 20, 21]. By their omnipresence, prokaryotes influence the entire biosphere and play key roles for biogeochemistry, nutrient cycles and waste degradation on earth [1]. Their impact on human health was demonstrated by important findings proofing that many infectious diseases are caused by pathogenic microorganisms. For the healthcare sector, multidrug-resistant pathogens became a serious risk in the recent past [22, 23, 24]. Conversely, it was often demonstrated how microbes can be used beneficially for medical and biotechnological applications such as the production of antibiotics and industrial enzymes. The enormous population size and rapid changes by horizontal gene transfer contribute significantly to the vast microbial diversity and rapid evolution [25]. The findings how frequently genes are transferred from one organism to another even put the concept of individual microbial species into question [26, 27]. Accordingly, genomes may not be regarded as discrete and independent entities, but rather units with strong capabilities to reconstruct themselves with respect to their environment and the metabolic flux.

The analysis of microorganisms faces the issue that only a low proportion of microorganisms

are readily culturable in a laboratory [28, 29, 30]. As a consequence, most environmental microbiota have not been studied or described in detail. Moreover, the cultivation methods are limited and additionally bias the approaches to investigate the potential of microbial communities [31]. Consequently, these severe challenges stand against the immense microbial diversity which researchers attempt to investigate. Therefore, innovative methods and methodological improvements are required to study complex microbial communities.

2.1.1 Role of Microbial Communities in Humans

The human body harbors 10^{14} microbial cells [16] and a quadrillion viruses [32]. Thus, the number of bacterial cells in humans exceeds the number of human body cells by a factor of ten [33]. In particular, the quantity of microbial genes in the human gut is impressive, since it is estimated to exhibit a magnitude of more than 100 times that of the human genome [34]. Together, the microbial associates residing in and on the human body constitute the microbiota, whereas the collective genome they encode is called the microbiome. Although a diverse ensemble of microorganisms provides humans with beneficial genetic and metabolic characteristics, studies in microbiology were mostly performed with the focus on pathogenic organisms rather than investigating the benefits of resident microbes. The endogenous microbiota of humans was poorly understood for a long period of time [35], however, recent studies began to characterize the driving factors which influence the distribution of microbial communities to fully understand the human genetic and metabolic diversity [36, 37].

The microbiota is essential for health and disease in humans and was therefore also called a virtual organ with its own metabolic activities [38]. Microbial symbionts fulfill important functions, such as nutrition uptake, pathogen resistance and immune response [39]. The majority of microorganisms present in humans can be found in the gastrointestinal tract. The human intestine is mostly composed of Gram-positive and anaerobic microbes which are responsible for the processing and uptake of nutrients otherwise inaccessible to humans [34]. The gut flora has a strong impact on metabolic processes of the host, in particular, by the provision of energetic substrates. [40, 41].

To understand the role of the human microbiota in health and disease, large consortia, such as the Human Microbiome Project [37] and MetaHIT [42] were established. The goals of these collaborative initiatives were the characterization of the human microbiota and the identification of criteria influencing the evolution and distribution of involved microorganisms. Respective projects provided a hint on the diversity at the genetic level and also showed a large variability of microbial species and abundance even within closely related healthy individuals. While the human body holds an immense variety of human and microbial cells, it has also been found that a

conserved set of microbial genes and species is shared among different persons. Clearly, this core microbiome is essential for the metabolism and health of the hosts. However, each person is also able to carry a distinct microbiota and species abundance can vary strongly between individuals [42]

2.1.2 Microbial Analysis Techniques

In 1977, Carl Woese and George Fox revolutionized the field of microbiology by defining Archaea as a third domain of life [43]. This pioneering work became feasible using the 16S ribosomal ribonucleic acid (rRNA) technique which was then extensively applied to study microbial communities [44, 45]. Later on, the phylogenetic tree of life was divided into 23 main divisions under the three domains Archaea, Bacteria and Eucarya [46]. Due to the age and the critical role of ribosomes for protein synthesis, rRNA genes represent evolutionary chronometers [47, 48]. Additionally, the analysis of 16S rRNA gene sequences provides insights into the composition and diversity in environmental samples without culturing [49]. Along with the application of the polymerase chain reaction (PCR) to 16S rDNA sequences this approach accelerated the description of uncultured organisms in mixed microbial communities [50]. Nowadays, the Ribosomal Database Project contains more than 2.8 million archaeal and bacterial small subunit (SSU) rRNA gene sequences, reflecting the high microbial diversity on Earth [51]. Despite its wide application, one major shortcoming of 16S rRNA sequencing presents the limited information content about the functional role of the microbes within the community [31]. Microbes rarely live in single species communities, but interact with each other in their habitats and host organisms. Therefore, a clonal culture does not represent real conditions in nature with respect to molecular interactions, biological functions and resulting genomic diversity of microbial communities [52].

At the beginning of this century, a milestone was reached by the development of DNA shotgun sequencing methods that shifted from the expensive and labor-intensive Sanger sequencing technology to more affordable next-generation sequencing approaches with rather short read lengths, such as high-throughput pyrosequencing [53, 54, 55]. Using these modern technologies, genomic sequence information can directly be inferred from the microbial communities in their natural environment. Retrieving sequence data obtained from multiple species of an entire microbial community is called metagenomics [56]. By examining the genetic material of a whole consortium, metagenomic analysis allows to characterize the most dominant community members. Due to the overwhelming majority of uncultured organisms in microbial niches, metagenomic analyses are likely to uncover novel sequences from previously unknown genes. On the one hand, the relationship between a microbial community and its habitat can be investigated, on the other hand, the adaption of microbes to different environments, such as host

animals or other microbial members, and the related manifestation in the microbial genomes can be studied.

The resulting wide application of whole-genome and metagenomic sequencing studies provided completely new perspectives on the role of environmental microbial consortia [57, 58]. The Sargasso Sea project by Venter *et al.* encompassed an extensive environmental metagenome-based analysis which resulted in the identification of various novel genes [59]. Due to advances in sequencing technologies, the amount of available genomic and metagenomic sequence information has rapidly increased in the recent past and will probably grow further. It has been speculated that the number of population genomes stored in public databases will even outnumber those from pure culture and single cells [60]. Recently, instead of analyzing single snapshot metagenomes, researchers even moved forward to biologically replicated series of several metagenomes [61]. On the genomic level, single-cell genomes could already be obtained from uncultivated archaeal and bacterial cells [62].

Despite the outlined benefits and progress, one major drawback of genome-level approaches is the missing link between genomic presence and functional level. For example, Tringe *et al.* used environmental DNA data from different ecosystems for the clustering of functional groups and concluded that the predicted protein complement of a community is influenced by its environment [58]. With regard to rapid environmental changes, it is therefore required to determine the abundance of actually expressed genes within a microbial community. In contrast to metagenomics, metatranscriptomics determines the gene expression by providing the complete set of transcriptional profiles within a microbial community at the time of sampling [63, 64]. However, it was reported that the expression levels of mRNAs and proteins are only poorly correlated [65, 66]. Regarding the analysis of microbial communities, the short half-life of mRNAs in bacteria [67, 68] and ineffective mRNA enrichment [69] are the major challenges for metatranscriptome studies. As described in the following section, proteomic approaches address these limitations by directly characterizing the phenotypes as functional key players in microbial communities.

2.1.3 Beyond the Genome to the Proteome

According to the central dogma of molecular biology DNA is transcribed into messenger RNA entities that contain required information for the synthesis of particular proteins. Ribosomal cell structures translate each mRNA into a protein which itself undergoes various modifications before reaching its fully functioning state. Proteins are characterized as molecules consisting of one or more polymer chains of amino acids which are folded in a specific conformation. In biological systems, proteins are responsible for a plethora of important functions, such as cell

structure integrity, molecule transport, enzymatic regulation of metabolism, signal transduction and their own biosynthesis. Proteins exhibit a high temporal and spatial variability with respect to turnover and expression rates.

The proteome has been defined as the entire set of proteins in an organism including cells, tissues or subcellular components [70]. While the genome of an organism works as the static template being almost identically present in all cells, the proteome is a highly dynamic collection that spatially and temporarily varies between cell types or even within a cell, depending on environmental or physiological conditions. While the human genome has been estimated to comprise between 20 000–25 000 protein-coding genes [71, 72], the human proteome has been valued to contain over one million protein variants in total [73]. Furthermore, in eukaryotic cells, the majority of proteins is subject to post-translational modifications (PTMs) which cannot be observed at genome or transscriptome level. Phosphorylation, glycosylation, acetylation, methylation and ubiquitylation are modifications that frequently occur at amino acid side chains or peptide linkages by the mediation of activated enzymes. These PTMs are capable of directly influencing activity state, turnover, localization and interaction of proteins [74]. Consequently, protein modifications play a major role in various cellular processes and increase the functional diversity of the proteome [75].

Initially, the term proteomics was coined as an analogy to genomics and originally referred to the identification of proteins separated and visualized by two-dimensional gel electrophoresis [76]. The latter approach was modified by using multiple dyes on the same gel: the so-called difference gel electrophoresis (DIGE) allowed to reproducibly identify differences between protein samples [77]. Later on, these latter approaches have been widely replaced by shotgun proteomic methods which employ the enzymatic digestion of proteins into peptides prior to mass spectrometry (MS) or tandem mass spectrometry (MS/MS) analysis (see Section 2.1.5). Nowadays, proteomics is more generally associated with the comprehensive analysis of proteins with respect to their identification, quantification and functional classification. Furthermore, novel approaches to investigate protein structure and protein-protein interactions became an increasingly important part of the field [78]. Moreover, MS has been extensively used to determine a large variety of aforementioned occurring PTMs which are potentially able to influence the modulation of protein functions [79].

While an encompassing amount of knowledge has been gained from MS-based proteomic studies, the vast majority of experiments was performed on single cell or tissue samples of relatively low complexity. Over the last decade, the field of proteomics has immensely matured through technological advances with respect to sample preparation, instrumental techniques and data analysis [80]. Driven by these latter improvements and the provision of metagenomic sequence information (see Section 2.1.2), proteomic methods are nowadays increasingly applied

to investigate the protein expression and functional potential of microbial communities as described in the following section.

2.1.4 Microbial Community Proteomics

In the recent past, the cost-effectiveness and improved throughput of DNA sequencing technologies resulted in an extended availability of sequence information from single genomic and metagenomic experiments. While metagenomics and metatranscriptomics provide insights into the phylogenetic structure and functional potential of microbial communities, the extension of single-organism proteomics to the so-called metaproteomics or whole community proteomics is the large scale characterization of the entire protein complement of the environmental microbiota at a given point in time [8, 9]. The term metaproteome had been proposed by Rodriguez-Valera to characterize the most abundantly expressed genes and proteins in environmental samples [1]. By investigating microbial communities on the proteome level, the major goal of metaproteomic research is to find the link between microbial community composition and functional profile [81].

One of the early criticisms of proteomics was based on the assumption that this method would provide only limited depth in terms of identifying merely the most abundant housekeeping proteins. However, this argument has been lately disproved by studies reporting in-depth proteome coverage due to modern proteomics protocols and MS technologies, exemplified by the report of near-complete proteomes for mammalian cell lines [82, 83, 84] and findings on the human proteome [85, 86]. While the proteomic analysis of microorganisms enables to identify 50–70% of the predicted proteome for most bacteria grown in pure cultures, the diversity of the proteome is much higher in microbial communities than in single organisms. In this context, metagenome sequencing is able to obtain a content-rich catalog directly from a microbial community by translating metagenomic sequence information into a collection of predicted proteins (see Section 2.2.2).

A metaproteome study on dissolved organic carbon and soil particles was performed by Schulze *et al.* to investigate the phylogenetic groups and catalytic functions of species identified in microbial samples from four different environments [87]. Despite the low number of proteins identified, the presence and potential activity of different phyla could be demonstrated in this study. Hence, it was shown that microbial community proteomics presents a promising technique, in particular, when information about diversity and richness of the species is inferred. However, the taxonomic diversity reflecting the wide range of species within one sample remained challenging. Obviously, microbial communities with low complexity are easier to characterize than samples derived from complex ecosystems, such as soil and seawater. Consequently, early studies focused

on samples with low complexity, such as microbial consortia from acid mine drainage biofilms [88, 89] or activated sludge water bioreactors [3]. In addition, more complex environments, such as soil, present serious challenges regarding sample preparation and protein extraction, for instance, due to the presence of humic organic matter [90]. As one of the most complex microbial communities soil holds low protein abundance and accessibility due to its diversity and heterogeneous spatial distribution [91]. For freshwater and seawater samples, larger volumes are required to obtain sufficient material due to low cell densities [92]. While a wide range of potential fields of applications exist for metaproteomic research, the focus of this work was on biotechnological and medical applications. For this purpose, metaproteome samples originating from biogas plants (BGPs) as well as human gut microbiomes were studied. In the following, the essential background and the most relevant studies for both types of samples are provided.

Investigating composition and enzymatic activity of microbial communities in biogas plants. Over the recent past, BGPs became an important and reliable source of renewable energy in Germany [93]. The production of biogas is established by anaerobic digestion processes in which a complex microbial community converts organic material primarily to methane and carbon dioxide [94]. The produced biogas can be used as fuel or transformed into electricity in combined heat and power units. In general, various physicochemical and technical process parameters can influence the success of biogas production, including temperature, pH value, substrate composition and configuration of the fermenter [95]. Based on the impact of these parameters, the most important goals of BGP optimization address the biogas and methane yield, the efficiency of biomass degradation and the process stability within BGPs. Furthermore, the biogas producing microorganisms can be strongly affected by several process disturbances, such as acidification due to organic overloading [96] or increased ammonia concentrations due to protein-rich substrates [97]. To tackle such problems which impair biogas production and consequently cause economic losses, a deeper understanding of the composition and the metabolic state of microbial communities is required. In contrast to the community characterization which is based on metagenomic approaches [98, 99], metaproteomics can be used investigate the active role of individual species as functional key players in metabolic networks. In recent years, metaprotomic analyses were applied successfully to examine the taxonomic and functional profiles of microbial communities in BGPs [100, 101]. In the beginning, metaproteome analyses achieved only low amounts of protein identifications [5, 13]. However, in the recent past, due to experimental advances of high-resolution methods between 500 and 2 000 proteins were identified in BGP samples [101]. Notably, the majority of the identified proteins could be assigned to the main anaerobic process steps, namely hydrolysis, fermentation, acetogenesis and methanogenesis. In addition, the latter study covered the most important archaeal and bacterial taxa which are in-

volved in the production of biogas. Eventually, a long-term study of an agricultural BGP combined different analytical methods with metaproteomic analysis and evaluated the influence of process disturbances on the composition and activity of the microbial community [102]. While different BGPs may have a common set of certain methanogenic enzymes, such as methyl CoM reductase, that are expressed by particular dominant community members, each BGP provides an individual protein profile that is stable over longer time periods [93].

In this work, metaproteome data sets of BGP samples from different anaerobic digesters (see Section 3.2.1) were used to compare the performance of database search algorithms with respect to peptide-spectrum matching (see Section 4.1.1). Further goals were to evaluate the influence of the protein database on the identification yield (see Section 4.2.1) and taxonomic assignment (see Section 4.5.1), and to establish a metaproteomic data analysis workflow that allows to identify key enzymes and to link the most important microbial taxa to metabolic pathways (see Section 4.6.1).

Obtaining taxonomic and functional profiles of human intestinal microbiota. The highly complex ecosystem of 100 trillion bacterial cells in the human gut [16] is known interact with the innate immune system by providing signals to promote the maturation of immune cells and the normal development of protective functions [103, 104]. The human gut contains mainly anaerobic microbes that play an important role in the well-being of their host [105]. Microorganisms that interact with each other and the host also influence the development of several diseases [106]: for instance, alterations of the human gut microbiome have been associated with pathological states, such as obesity [107, 108], type-2 diabetes [109], cardiovascular disease [110] and inflammatory bowel disease [111]. While studies at the genomic level have shown the close connection between host and microbes [105, 112], the direct effects on the host proteome can only be detected by metaproteomic approaches. Consequently, human intestinal microbiomes have been also investigated in several proteomic studies that focused on the analysis of extracted microbial as well as host proteins from human fecal samples [14, 15, 113, 114, 115]. However, in comparison to the large number of metagenomic studies, relatively few investigations have been carried on gut samples at the proteome level. Several challenges, including sample heterogeneity, high abundance of host proteins and lacking database references, impede the analysis in these studies.

In this work, metaproteomic data sets originating from human intestine metaproteome (HIMP) samples (see Section 3.2.2) were used to evaluate the performance of different data analysis methods used for protein identification (see Section 4.2) and protein grouping (see Section 4.4.3). Since these samples originated from obese and non-obese individuals, another goal was to determine any characteristic taxonomic (see Section 4.5.3) and functional (see Section 4.6.3) profiles for both categories from the identified proteins.

2.1.5 Experimental Bottom-Up Workflow

Commonly, metaproteomic studies employ experiments using bottom-up proteomics, also referred to as shotgun proteomics [116]. The bottom-up approach is based on the proteolytic digestion of protein mixtures, followed by the chromatographic separation of the resulting peptides and eventually the mass spectrometric analysis via MS and MS/MS. Acquired peptide fragment spectra are then processed using bioinformatic methods (see Section 2.2) to identify peptides and to infer proteins present in the analyzed sample. Alternatively, top-down proteomic approaches can be used: in this case, protein mixtures are first separated on the protein-level, and whole single proteins are subsequently subjected to LC-MS/MS analysis [117, 118]. Eventually, the combination of top-down and bottom-up proteomics has been successfully applied in various studies [119, 120].

In the following, the most important experimental techniques of metaproteomic experiments are presented. This general overview is by no means exhaustive and for further detailed information regarding microbial analysis prior to MS, the reader is referred to a review of proteomic techniques in environmental and technical microbiology [121]. For more detailed information on general analysis techniques in bottom-up proteomics, the reader is further referred to comprehensive reviews in the literature [122, 123, 124]. While the experimental and analytical methods have advanced with respect to accuracy, resolution and speed, the high complexity and heterogeneity of microbial samples are the most severe challenges of the field of metaproteomics [10]. In the recent past, an increasing amount of protocols have been developed to tackle both experimental and data analysis issues [125, 126, 127]. The reader is referred to Section 2.2 for details on data analysis techniques in metaproteomics.

The classic workflow of bottom-up proteomics consists—with some variations—of the following five essential steps:

1. Sample acquisition and preparation
2. Protein separation
3. Enzymatic protein digestion
4. Peptide separation
5. Mass spectrometry

Sample acquisition and preparation. The first step of a metaproteomic experiment involves the sample acquisition and preparation. While variations have been reported to be caused by sample collection in proteomic experiments, the most severe challenges can be assigned to the preparation of samples from microbial communities: for environmental samples, the steps of cell lysis and protein extraction are often affected by impurities, such as humic and fulvic acids [90, 93]. Regarding the high sample complexity and individual characteristics of each microbial community, various sample preparation methods have been proposed [128]. Protocols for protein extraction in metaproteomics can be classified into methods for cell lysis with chemical reagents, mechanical cell disruption or thermal treatments [127]. Since few standard procedures are available, extensive method optimization steps are required, frequently by using a combination of the aforementioned protein extraction techniques. Another problem that is more pronounced for environmental than for pure culture samples concerns the limited availability of biomass due to difficulties during sample collection [10]. Finally, the effects of protein degradation during sample processing further impair the proteomic analysis by the generation of unwanted protein artifacts [129].

Protein separation. In metaproteomics, a common protein separation procedure is sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) by which proteins are separated according to their molecular weight. To further reduce the sample complexity, two dimensional polyacrylamide gel electrophoresis (2D-PAGE) is used by which proteins are separated in two dimensions. In the first dimension, the proteins are separated by their net charge using isoelectric focusing. The second dimension is a conventional SDS-PAGE that separates proteins by their molecular weight. The major advantages of gel-based methods are the low complexity of later on analyzed protein spots and the potential to identify protein isoforms. However, gel-based methods have limitations regarding hydrophobic membrane proteins or proteins present in low copy numbers [130]. As additional method in metaproteomics, centrifugal fractionation can be applied to separate crude fibers, suspended microorganisms and secreted proteins, as demonstrated for biogas sludge samples [131].

Enzymatic protein digestion. The idea of the bottom-up approach is to enzymatically degrade proteins into peptides which are analyzed and subsequently mapped back to the protein sequence. Thus, in the next step, proteins are denatured and proteolytically digested with a sequence specific protease. An important condition for such an enzyme is to be capable to cleave any protein inside its amino acid backbone. In proteomics, trypsin is most commonly used, since it presents a highly stable and efficient protease which specifically cleaves proteins into peptides ending with lysine or arginine residues [132]. Nevertheless, it should be noted that also tryptic cleavage rarely works perfectly and various studies reported the occurrence of missed cleavages and non-tryptic peptides [133, 134, 135]. Moreover, while the bottom-up technique facilitates the remaining experimental procedures, it also results in the loss of information on the original protein from which a peptide originated. This essential drawback eventually complicates matters related to the computational analysis, as described in more detail for the protein inference problem in Section 2.2.5.

Peptide separation. Liquid Chromatography (LC) is frequently employed to separate complex peptide mixtures, predominantly by the degree of hydrophobicity of the analyte [123]. The principle of LC is to bring the analyte into the mobile phase via a liquid solvent and pass it through the stationary phase being a chromatography column filled with adsorbent material. In LC, the compounds are separated in the mobile phase based on their affinity for the hydrophobic stationary phase. Consequently, the peptide compounds elute from the column at a specific time point, the so-called retention time, and are transferred separately into the MS instrument for further analysis. The online-coupling of LC methods to (tandem) mass spectrometry is called LC-MS (LC-MS/MS). Alternative chromatographic setups, such as a combination of reverse phase LC with strong cation exchange chromatography (SCX), can be used to achieve a multidimensional separation [136]. Furthermore, due to its beneficial properties for the analysis of post-translationally modified peptides hydrophilic interaction liquid chromatography (HILIC) was reported to be a valuable alternative separation method to SCX [137].

Mass spectrometry. The principle of MS presents the measurement of ionized compounds based on their mass-to-charge (m/z) ratios using controlled electromagnetic fields [138]. In the context of proteomics, the technique is used to identify peptides by measuring the m/z ratios of their ionized variants [122]. However, MS is also successfully applied in other fields, including metabolomics, glycomics and lipidomics.

In general, the MS peptide analysis consists of various steps of which the most important ones are highlighted here. Once the samples are loaded into the MS instrument (e.g. via direct online-coupling using an LC-system), gaseous particles are formed in the ion source, which transfers the

sample compounds from solution or solid media into the gas phase. Subsequently, the gaseous particles are ionized to produce charged species which are separated in the mass analyzer by applying electromagnetic fields according to their m/z ratios. In the last step, the ions hit the detector and an intensity signal for each specific ion m/z ratio is recorded. Before the detection process, ions can be subjected to multiple stages of mass analysis separation and fragmentation to increase the resolution of the analysis. While MS is realized via several analytical platforms, the outlined principles of ionization, mass analysis and detection are always applied.

While many ionization methods exist, the most commonly used ones belong to the category of so-called soft ionization to prevent unwanted fragmentation of biomolecules. In MS-based proteomics, the most relevant techniques are matrix-assisted laser desorption/ionization (MALDI) [139] and electrospray ionization (ESI) [140, 141]. Nowadays, ESI presents the favored ion source, since it has the advantage of working continuously in direct connection to an LC-based system.

The signals derived from MS instruments are recorded as mass spectra containing pairs of m/z ratios and intensity values based on the detected ion current. The most common operative mode presents tandem mass spectrometry (MS/MS) in which selected ionized compounds undergo fragmentation. Usually, a defined number of high-abundant precursor ions are selected and subjected to collision with an inert gas for subsequent fragmentation. As a consequence, two different kinds of mass spectra are produced: the MS1 spectra which contain the signals of analytes eluting from the column and the MS2 (or MS/MS) spectra which feature the corresponding fragment ion signals of analytes that were selected for fragmentation.

Since a shotgun proteomics experiment can result in tens of thousands of MS/MS spectra within a short time period, computational methods are required to provide a rapid analysis of the high-throughput data. As described in the following section, several techniques and software tools are available to derive peptide and protein identification from the fragment ion information of the spectra. In this context, the most severe challenges concerning the data analysis in bottom-up metaproteomics are highlighted specifically.

2.2 Data Analysis Workflow in Metaproteomics

This section outlines the most important computational methods used to process and analyze MS-based data from microbial community samples. In the following, parts of the original publication in Molecular BioSystems [142] are used with permission from The Royal Society of Chemistry.

The typical metaproteomic data analysis workflow can be regarded as a three-step process, consisting of pre-processing, protein identification and post-processing (Figure 2.1). In the displayed figure, the most relevant methods regarding the analysis of metaproteomic data are summarized for each step.

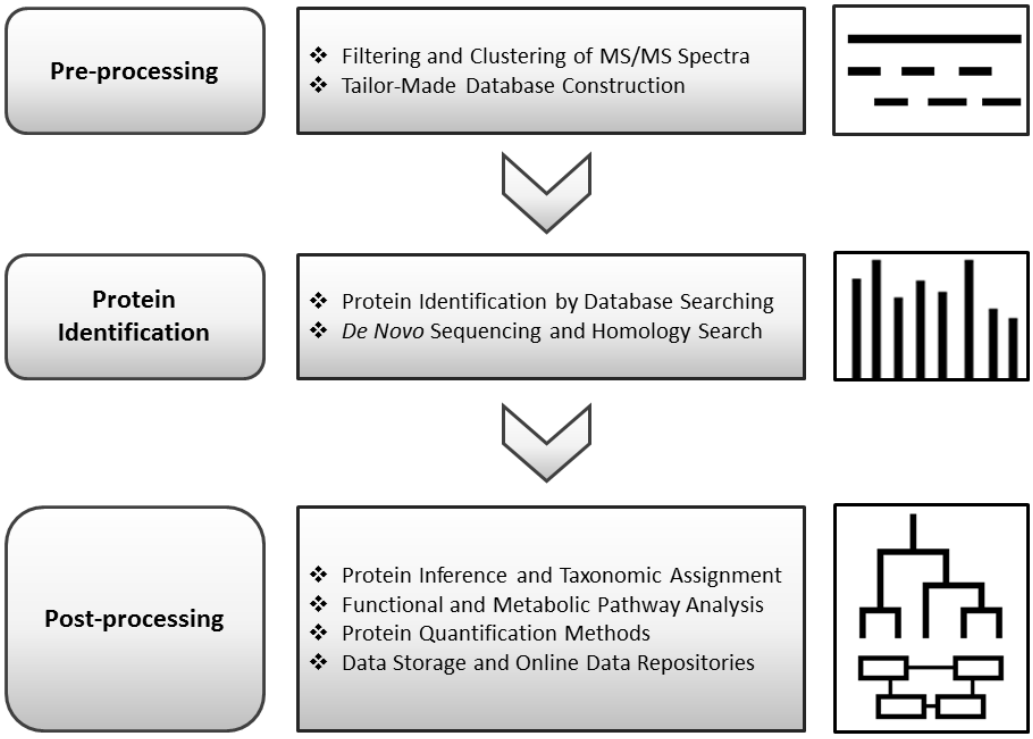

Figure 2.1: Metaproteomic data analysis workflow. Figure adapted from Muth *et al.* [142] with permission from The Royal Society of Chemistry.

2.2.1 Filtering and Clustering of MS/MS Spectra

The high amount of MS/MS spectra derived from metaproteomic experiments makes it useful to filter out noisy spectral data in order to accelerate the actual peptide and protein identification processing afterwards. Filtering criteria, such as the minimum number of peaks or the signal-to-noise ratio can be inferred from the spectra and algorithms can be applied to assess the overall quality of the spectral data [143, 144, 145, 146, 147]. Machine learning can be applied to separate low quality from high quality spectra including several spectral features [148].

As another method for reducing the total number of processed MS/MS spectra and improving the overall spectrum quality, spectral clustering can be applied. The clustering approach combines similar spectra into consensus spectra that serve as representatives for spectral clusters [149, 150]. In addition, clustering can be used for the identification of unexpected PTMs [151, 152]. Spectral clustering was also applied to find reliable identifications in heterogeneous proteomic data sets from the PRIDE database and to generate spectral libraries based on consensus spectra [153].

2.2.2 Tailor-Made Database Construction

The computational analysis of metaproteomic samples aims to identify and subsequently quantify proteins and peptides from MS/MS spectra. However, finding the optimal sequence space by which these identifications can be obtained is highly challenging for microbial communities: since the majority of organisms in a natural community is very heterogeneous or even unculturable (see Section 2.1.2), the restricted availability of suitable microbial sequences frequently results in a low number of identified proteins in previous metaproteomic studies: for instance, an environmental study analyzing sewage sludge from membrane bioreactors reported only 24 identified proteins due to missing sequences in the reference database [12]. In comparison to pure culture proteomics, metaproteomic studies also yield significantly fewer identified MS/MS spectra: for example, only 5% of the spectra could be identified in a study on the gut microbiome in mice [154]. Moreover, samples from microbial communities are affected by horizontal gene transfer and strain variability: a metaproteome study reported that small differences in the amino acid composition in comparison from experimental to theoretical data reduce the protein identification by a factor of two [155]. Consequently, the outcome of metaproteomic analyses depends strongly on composition and integrity of provided protein sequence databases. In the ideal case, the reference database covers exactly the sequences contained in the sample under study. Since the composition of a microbial community is unknown, however, mainly three approaches are employed to construct protein sequence databases. First, the full coding

potential of a sample can be retrieved by applying metagenomic sequencing. Next-generation DNA-sequencing technologies such as pyrosequencing [53, 156] and sequencing by synthesis [157] allow for the rapid generation of sample-specific databases as the produced reads cover the coding potential of the microbial community [158, 159]. In metagenomics, the prediction of whole genes from short sequence reads is more difficult as the traditional assembly applied for single genomes cannot be performed. For this purpose, specific gene prediction software tools exist [160, 161, 162, 163, 164]. Due to imperfect sequencing and assembly, however, protein databases derived from metagenome sequencing are prone to various sources of error resulting in partial or incorrect sequence information. Eventually, the generation of a metagenome database from the same sample is not always feasible due to experimental limitations. Therefore, the use of metagenomes that have been created from similar microbial communities in comparable conditions is an alternative—yet related—strategy [92]. Second, the protein database can be built by using published microbial reference genomes: as a consequence, it contains sequences from previously described organisms that are assumed to be present in the samples under investigation. For instance, this approach has been successfully applied in several studies on the human gut metaproteome [14, 15, 113]. Finally, protein databases can be derived from public repositories, such as UniProtKB [165] and National Center for Biotechnology Information (NCBI) RefSeq [166]. Although using predicted protein databases from metagenome sequencing may result in more identifications [92], the quality of the sample and the used metagenome still have a strong impact on the results: in particular cases, database searches against protein sequences from public repositories can be more effective than using metagenomes [167]. On the other hand, public databases can also result in a bias which is inferred by the overrepresentation of certain species, as in the case of clinical strains. Moreover, public databases may hold a high degree of redundancy which biases the results (see Section 2.2.5). In this work, different database types are evaluated with respect to their identification yield for varying metaproteomic analysis setups (see Section 4.2).

2.2.3 Protein Identification by Database Searching

The common principle of protein database search algorithms is to correlate acquired MS/MS spectra with theoretical fragment ion spectra. The theoretical spectra are calculated for each of the peptides derived from an *in silico* digested protein sequence database. SEQUEST [168] and Mascot [169] are the pioneering and still the most popular commercial database search algorithms. Freely available algorithms include X!Tandem [170], OMSSA [171], MyriMatch [172], Crux [173], InsPect [174], Comet [175], MS-GF+ [176], MS Amanda [177] and Andromeda [178] included in the MaxQuant software package [179].

Despite the immense variety of database search engines and different scoring techniques, each of these algorithms suffers from the problem of false positive (FP) identifications [180]. Therefore, procedures to control the false discovery rate (FDR) are essential to retrieve reliable search results. Various approaches have been developed to estimate the FDR for peptide and protein identifications, including algorithms based on statistical modeling, such as PeptideProphet [181]. This rescoring algorithm uses a mixture model-based approach to estimate the global FDR by assessing the probabilities of peptide identifications and was later updated by an expectation-maximization algorithm optionally including a decoy database [182]. This approach was reported to be robust in case of simulated partial sequence databases [183]. Post-processing algorithms such as MSblender [184] and iProphet [185] improve the yield of correct identifications. These methods benefit from the complementarity of the search engines and combine the results by calculating probabilities of correct identification based on individual algorithm scores.

The most commonly used method for estimating the FDR is the target-decoy approach (TDA) [186, 187] which has been implemented into database search engines such as MASCOT [169]. The usual way to generate a decoy database is to reverse or shuffle the protein sequences present in the input database [188]. The identifications from the decoy database search results are taken into consideration to estimate the number of FP matches in the target database search. Software tools, such as QVality [189] and FDRAnalysis [190] utilize the TDA to determine the FDR based on the scores from the individual search engines.

The software Percolator employs semi-supervised machine learning to increase the number of identifications at a constant FDR threshold [191]. Several identification features, such as score and precursor mass error, are extracted from both target and decoy results to train a support vector machine (SVM). The trained model is used to reevaluate each obtained peptide-spectrum match (PSM). Percolator was also adapted for the MASCOT search engine [192]. The binary classifier Nokoi is another machine learning technique that allows the distinction between correct and incorrect identifications [193]. The algorithm was trained on heterogeneous identifications from the MASCOT search engine and holds the benefit of circumventing the use of a decoy database: although the TDA provides reliable FDR estimations for the limited search space of a pure culture proteome, the scoring metric deteriorates when the database search space increases in size, as in case of metaproteomics and proteogenomics [194, 195, 196]. For incomplete databases, the conventional FDR estimation suffers from similar issues: despite their high quality, a significant number of MS/MS spectra are not identified in the target database, but a fraction of these spectra may find a match in the decoy database. Thus, for large and incomplete databases, problematic decoy hits can significantly affect the FDR estimation.

2.2.4 *De Novo* Sequencing and Homology Search

The major downside of database search engines is their dependence on protein sequences. For instance, in case of metaproteomics, the protein databases are often incomplete and the overhead of searching against a large search space covering all potential microbial species can be immense. The method of *de novo* sequencing fully circumvents the need of a protein database, since it infers the amino acid sequence directly from the information given in each mass spectrum. Additionally, unknown peptide sequences that have not been identified by conventional database searching can be found and certain post-translational modifications present on peptide sequences can be targeted. These advantages qualify *de novo* sequencing as promising tool for the research field of metaproteomics.

The drawback of this approach is the demand of high quality data to deduce the amino acid sequences correctly [197]. Furthermore, accurate processing methods, such as noise reduction, binning and filtering are mandatory steps for successful *de novo* sequencing [198]. Finally, in case protein sequences are neither available nor known, *de novo* searching remains the only valuable method to assign peptide sequences to MS/MS spectra [199]. The most popular *de novo* sequencing tools are the freely available PepNovo+ [200] and the commercial PEAKS software suite [201]. In addition, several other algorithms have been described in the literature [202].

Nevertheless, *de novo* sequencing has not been widely adopted for proteomic workflows: one of the major drawbacks is that the matching the *de novo* peptide sequencing to the protein level is not included in most available software tools. For instance, performing a heuristic search by means of the basic local alignment search tool (BLAST) presents an opportunity to map generated peptide sequences to a protein database [203]. In addition, candidate homology proteins can be identified by using an extension of the BLAST algorithm via the MS BLAST searching protocol [204]. However, the latter application should be used with caution as the process of *de novo* sequencing is prone to various errors which may result in incorrect protein sequence assignments in the end. Thus, a manual inspection of the results is still required in order to reduce the total error rate. The combined approach of *de novo* sequencing with BLAST similarity search is therefore not suited for high-throughput studies. Notably, the BLAST algorithm does not incorporate the information from the spectrum level and the identified amino acid sequences may be changed by allowing mutations in the search options without vitally adjusting the final scoring. All of these mentioned issues may be overlooked in case the BLAST algorithm is used in a naive approach as post-processing step for *de novo* sequencing.

Similar to the combination of multiple database search algorithms, a combined use of several algorithms in the context of *de novo* sequencing could be beneficial for the confidence of the derived peptide sequence suggestions. For example, Cantarel *et al.* merged the results of Pep-

Novo+ and PEAKS in a metaproteomic workflow to retrieve consensus sequence tags [199]. This approach may be promising for future whole-community proteomics studies, in particular, when it is coupled to an efficient method of receiving protein sequences by *de novo* sequencing results with a global FDR rate, as described in the PepExplorer software specification [205].

2.2.5 Protein Inference and Taxonomic Assignment

Bottom-up shotgun proteomics brings along the problem of protein inference that complicates the data analysis and interpretation [11]. Central to this issue is that the mapping of mostly tryptic peptide hits to the protein space can be affected in various ways: the same peptide sequence can be assigned to different protein splice isoforms or to homologues proteins from multiple different species or strains. Additionally, protein families with functional domains may share the same or at least similar sequences. Eventually, the shared peptides cannot be uniquely assigned and result in ambiguities in the identification of proteins. While the inference problem hampers the analysis and interpretation of experimental results in proteomic workflows [188, 206], in metaproteomics, the challenges are even higher when the samples contain hundreds of different organisms: the identified peptides can then be assigned to protein sequences from multiple species. This is a frequent situation in related organisms with high sequence similarity and also for conserved protein domains. Consequently, a unique taxonomic assignment and correct quantification of species is challenging for metaproteomic samples.

On the computational side, the database search algorithms were designed for pure-culture proteomics and often provide a limited output with respect to the total number of identified proteins: some search engines only consider the most probable hits for the display and this becomes an issue for samples from complex microbial communities. Conversely, the more recent versions of MASCOT [169] report all protein identifications for each PSM. However, this also leads to a high redundancy in the results output for metaproteomic experiments.

The software MEGAN, although it was originally tailored towards metagenomic analysis, supports a protocol for metaproteomic data and includes the lowest common ancestor (LCA) approach by constructing a phylogenetic tree based on the NCBI taxonomy database [207, 208]. The script-based PROPHANE workflow employs a protein grouping approach by merging those protein identifications that share common peptides to a group [4]. This approach also uses the phylogenetic information to find the common taxonomic level at which the assigned protein identifications are converging in the hierarchy. Although these methods are promising, both have not been included in a complete proteomic analysis workflow. Another disadvantage of the LCA method is that certain amino acids are more conserved than others and could introduce a bias into the analysis. Conversely, a different approach is to limit the analysis to a representa-

tive set of species [209]. Overall, both approaches reduce the taxonomic resolution and make it therefore difficult to differentiate between low-level taxa.

2.2.6 Functional and Metabolic Pathway Analysis

Proteomic workflows often end with lists of peptide and protein identifications. However, the analysis of metaproteomic samples always requires to put the obtained protein data into a semantic context. Here, various post-processing steps are described which can be applied to interpret metaproteome data at the functional level.

The functional annotation of the identified proteins can be achieved by accessing information in public databases. Therefore, the Universal Protein knowledgebase (UniProtKB) is an important resource that holds curated information content about proteins and also provides links to other repositories containing functional annotations on specific proteins [165]. The Cluster of Orthologous Groups (COG) database connects both eukaryotic and prokaryotic proteins to functional groups and respective categories [210, 211]. For example, COG categories were used in a metagenomic study on a biogas-producing microbial community in order to map coding sequences to predicted functions [98]. The COG classification was also used for the functional analysis of the human gut metaproteome [15]. The COG database was updated in 2014 and now contains an increased microbial genome coverage of 711 archael and bacterial genomes [212]. However, this database project is manually curated and therefore has the disadvantage of containing a relative low amount of entries. A related approach presents the Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (EggNOG) database that extends the original COG database by far more orthologous groups covering 3 686 organisms as underlying species set [213]. In addition, for those protein sequences that cannot be directly assigned to COGs, so-called non-supervised orthologous groups (NOGs) are created. Another important benefit of the EggNOG database comes with the feature of automatically linking the groups to further functional resources, such as the Gene Ontology (GO) database [214]. GO aims to describe gene products consistently across several databases and consists of three different ontologies as structured vocabularies that describe biological processes, cellular components and molecular functions. The disadvantage in the context of microbial community proteomics is that the GO database holds only the main reference genomes so far. Various other tools, such as Ontologizer [215] or the web-based DAVID [216] can also be used to connect protein information with ontologies and protein families. InterPro is an integrated documentation resource which holds catalogued information on protein domain, families and functional sites [217]. The repository also includes data from other resources on protein motifs, domains and functional sites, such as PRINTS [218], SMART [219], PROSITE [220], ProDom [221] and Pfam [222]. The InterPro database can thus

be used to gain insights into the functional context of the identified proteins. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is another valuable data resource that features genomic, molecular and functional information and also includes rich content on intermediate metabolic pathways [223]. Different kinds of expression data can be transferred into data models to investigate higher-order cellular processes. While the KEGG model is primarily based on enzymatic activities referenced by the Enzyme Commission (EC) nomenclature, proteins can be mapped via the KEGG automatic annotation server (KAAS) into corresponding KEGG orthology (KO) identifiers and related KEGG pathway maps [224]. As an alternative, the information from the COG database can be integrated to transfer the protein data indirectly onto the KEGG pathways for obtaining a whole metabolic pathway mapping as demonstrated in a study describing the functional core in human intestinal metaproteome samples [15]. The PROPHANE workflow [4] provides access to the COG and KEGG databases and also evaluates protein information by several secondary methods, such as BLAST [203], BioPerl [225] and Clustal W [226].

The Reactome database provides peer-reviewed and manually curated information on biological processes and pathways in human [227]. In addition, the repository holds orthologues events for non-human model organisms, such as mouse, rat or worm. For metaproteomic research, however, Reactome is a rather limited resource, as the knowledgebase contains only few higher eukaryotes. The MetaCyc project features a curated reference database of metabolic pathways from various species, but is mainly focused on plants and microorganisms [228]. The BioCyc database is connected to MetaCyc and provides organism-specific genome and pathway resources [229]. These resources are highly relevant for microbial community studies, in particular, as 891 genomes from the Human Microbiome Project have been successfully integrated over the past years [230]. Finally, metaproteomic data analysis can benefit from these resources by complementing the protein data with information on metabolic pathways, biological networks and protein-protein interaction maps [231].

2.2.7 Protein Quantification Methods

For the analysis and comparison of different protein expression profiles, methods for protein quantification are required to compare experiments and samples from microbial communities [232]. To measure the protein expression levels, gel-free protein quantification techniques, such as iTRAQ [233] or ICAT [234] are commonly used in traditional proteomics workflows, but cannot be directly applied to samples from complex microbial communities due to the high variability caused by sample preparation and protein separation: in particular for metaproteomics, these procedures are rather challenging and prone to errors due to the requirement of time-consuming optimization steps for different types of environmental samples. In addition, software for label-

based quantification is frequently outdated due to the rapidly evolving analytical methods [235].

As a valuable alternative, so-called label-free analysis techniques can be used for the quantification of LC/MS-MS data in metaproteomics [155]. These approaches have the benefit of directly using the protein identification data in order to provide a relative measure on the protein abundance. The most straightforward label-free quantification method is spectral counting: given the assumption that an increased spectral count for a specific protein correlates with higher protein abundance, the sum of the spectrum identifications for each protein is taken to estimate its abundance [236].

Various other label-free quantification methods either based on spectral counting or on the summed-up intensity of the matching peptides exist [237, 238]. For example, the normalized spectral abundance factor (NSAF) takes the protein sequence length into account [239]: longer protein sequences result in more tryptic peptides and an increased probability of being identified in comparison to shorter proteins. Various improvements have made over the past years, in particular in the robustness of the label-free quantification algorithms. For example, the robust intensity-based averaged ratio (RIBAR) and its extension xRIBAR correlate the intensity sum of the related MS/MS spectra in two experiments and thereby increase the reproducibility of the results [240].

2.2.8 Data Storage and Online Data Repositories

Although software tools are available for different data analysis steps, handling and integration of the upcoming data is often more difficult. Therefore, many research groups have built their own solutions in terms of in-house scripts and implemented database systems. Despite the clear benefits of customized workflows with respect to flexibility and user-definable settings, the outcome of the performed analyses can rapidly become user-dependent and difficult to compare. As a consequence, the reported results may not be replicated by any other laboratories. Most importantly, in-house solutions often lack standardized data formats and are not embedded into a server-based database architecture. However, to handle the amount of data from omics experiments, the integration of a laboratory information management system (LIMS) is highly useful. Such a framework often provides a relational database system which can be accessed by interfacing data analysis tools. In most cases, the management and querying of the database is performed by means of the Structured Query Language (SQL). On the commercial side, various available LIMS products facilitate the integration and storage of information obtained from various proteomics data analysis tools. Ms_lims [241], CPAS [242], MASPECTRAS [243], myProMS [244] and OpenMS [245] are MS-based data analysis systems that also hold comprehensive storage capabilities and are freely available.

Besides local database systems, so-called public repositories became popular within the last decade. The purpose of various initiatives is to share experimental proteomics data with the community via publicly accessible databases. The PRoteomics IDEntifications database (PRIDE) [246], Global Proteome Machine Database (GPMDB) [247], Mass spectrometry Interactive Virtual Environment (MassIVE) as part of ProteomeXchange [248] and PeptideAtlas [249] are the commonly used online repositories for proteomics data today. ProteomeCommons Tranche [250] and NCBI Peptidome [251] were popular databases, but both are no longer available while the Peptidome data have been saved to PRIDE [252]. The aim of these repositories is not only to be used as storage volumes, but to exchange data within the research community by giving access to both the original raw data and the final result files. Likewise, the stored data can be reanalyzed or used for benchmarking during the development of novel algorithms and data analysis techniques. Several publications describe extensively the operating modes of these public repositories [253, 254, 255].

3

Material and Methods

The most common issues for the analysis of microbial communities on the proteome level have been described in the last chapter. On the experimental side, the unknown complexity and intrinsic heterogeneity of the samples present enormous challenges for the scientists (see Section 2.1.4). Regarding the data analysis, which is the focus of this work, the size and redundancy of the protein sequence database have an impact on the results as introduced in Section 2.2.3. Furthermore, the taxonomic binning of peptide identifications is an issue on top of the protein inference problem (see Section 2.2.5). Finally, metaproteomic research intends to provide a perspective reaching beyond the stage of protein identification: researchers are not only interested in the identification of certain taxa, but also in the functional context of proteins and their role in metabolic pathways (see Section 2.2.6). Despite the importance and impact of these issues, the central question of the field "who is doing what" has not been addressed by any particular software workflow yet.

3.1 MetaProteomeAnalyzer

3.1.1 Software Workflow

The MetaProteomeAnalyzer (MPA) software was developed with the aim of analyzing and interpreting data from metaproteomic experiments, for instance, to investigate the composition and function of microbial communities in BGP or HIMP samples (see Section 2.1.4). The JAVA-based data analysis pipeline consists of a server application for processing experimental data sets and a graph-database driven interactive client for visualizing the results. The MPA software was published in Journal of Proteome Research [256] and parts of the original publication are used hereafter in this work.

A general overview of the client-server application is provided in Figure 3.1. In the following text, the most important steps of the data analysis are described in detail.

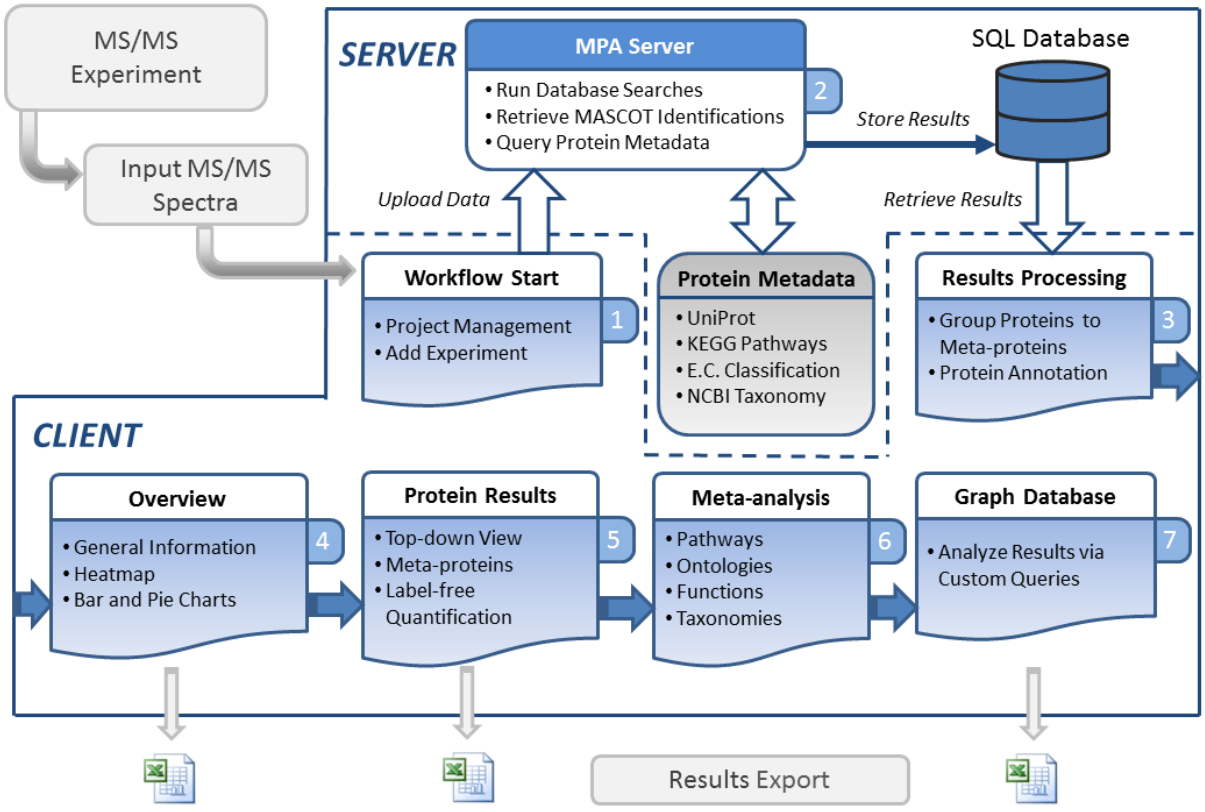

Figure 3.1: MetaProteomeAnalyzer software workflow. Figure adapted from Muth *et al.* [256].

The entry point into the workflow presents the creation of a new project (Step 1): experimental data, that is MS/MS spectra, are supplied by the user to the client application via the graphical user interface. These input data are sent to the MPA processing server which then runs up to

four of the supported database search algorithms X!Tandem [170], OMSSA [171], Crux [173] and InsPect [174] in sequential order (Step 2). As an additional option, the output of the popular commercial search engine MASCOT can be included for the identification results [169, 257]. Protein and peptide identifications are obtained by searching the MS/MS spectra against a supplied database in FASTA format. The MPA was mainly built for protein databases in UniProtKB format [165], but also supports customized databases. For example, references from metagenome sequencing can be included into the workflow. To combine the individual results of the used database search algorithms, the identification scores are transformed to q-values as comparable significance measures by using the QVality algorithm [189]. By definition, the q-value represents the minimum FDR and is used to filter the peptide-spectrum matches for a specific FDR level [258]. By using q-values, the individual confidence of each identification is measured and issues of p-values with respect to multiple hypothesis testing are avoided [259]. For each protein identification, additional information with respect to taxonomic and functional context is automatically queried from external resources via the UniProt remote API [260].

The MPA server holds a relational database based on SQL at the back-end to which all input and output data, such as MS/MS spectra, identification results and annotations, are stored during the processing. The whole pipeline is designed as LIMS with features for handling different projects and experiments. In addition, all stored information can be used for further analysis and re-analysis at a later time point.

After the processing has finished on the server side, the user can load the results onto the client application for the subsequent detailed analysis (Step 3). In this step, meta-information on the protein level, such as ontology keywords, taxonomic data from the NCBI database, metabolic pathway information from the KEGG public repository and enzymatic data via the numerical EC nomenclature is linked to the protein result set.

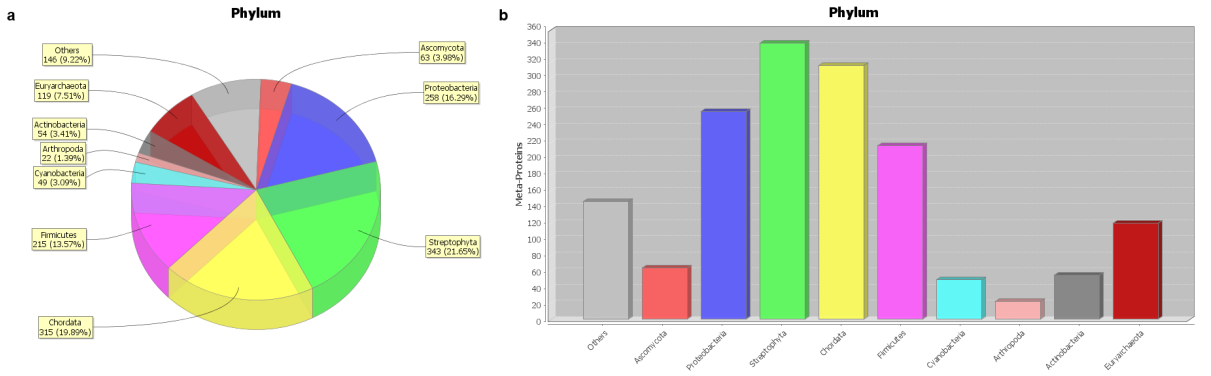

Figure 3.2: Pie and bar charts displaying protein distributions on taxonomic ranks. (a) Pie chart and (b) bar chart visualization for identified proteins, categorized into taxonomic ranks (e.g. phylum), and names (e.g. Proteobacteria in blue color). Figure adapted from Muth *et al.* [256].

The overview display on the client provides general information on the number of identified spectra, peptides and proteins (Step 4). Additionally, pie charts and bar charts show the relative identification yields for the different taxonomic and functional groups (Figure 3.2). The taxonomic levels can be adjusted by the user, ranging between superkingdom and species.

Figure 3.3 shows the main display of the MPA software with the database search result panel presenting the protein identifications in a top-down view: for each protein, the attributed peptides are shown in detail, and for each peptide, the supporting PSMs are displayed across the employed database search engines. In addition, an annotated spectrum display with fragment ion information can be inspected for each PSM (Step 5). The protein view shows relevant information, including for example protein description, protein taxonomy, sequence coverage, protein mass, spectral count, NSAF [239] and emPAI [261].

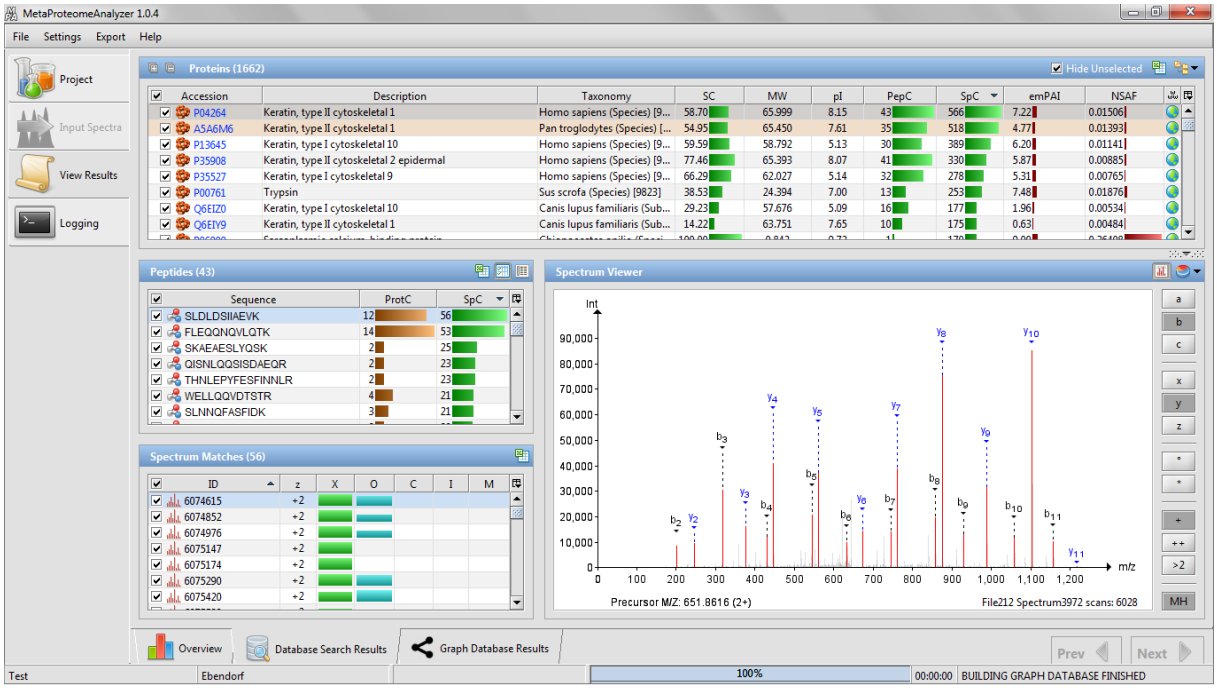

Figure 3.3: Search result panel of the MPA client user interface. The identified proteins are shown in the top panel, the identified peptides for the selected protein in the middle left panel, and the PSMs for the selected peptide in the lower left panel. The right panel displays the annotated fragment ion series of the currently selected PSM. Figure adapted from Muth *et al.* [256].

Further enzyme and pathway displays visualize proteins grouped by EC numbers and KEGG pathways (Step 6). These views directly support detailed investigations on proteins involved in certain microbial functions. The meta-protein view enables the inspection of protein groups of which the generation is described in the following section. In accordance to the general overview, the taxonomies and ontologies are visualized in detailed views with the related proteins. All views can be filtered for a specified entity, such as microbial group or metabolic pathway to allow for a

user-defined result inspection.

The MPA client application holds a graph database which can be used to ask specific questions in the result data by sending custom queries to the back-end (Step 7). The graph database will be further explained in Section 3.1.3. The software also provides a MPA project file export to share the results data with other researchers. Finally, the common exchange format comma-separated values is supported to transfer the results as plain text to conventional spreadsheet software.

3.1.2 Meta-Protein Generation

As already mentioned in Chapter 2.2, various issues impede the data analysis of microbial community proteomics experiments. An important problem concerns the non-unique relation of one peptide to many potential proteins, commonly formulated as the protein inference problem [11]. Transferring this paradigm to the field of metaproteomics, it becomes even more problematic as identified peptides can then be mapped to a large amount of expressed proteins originating from different species. As a consequence, the interpretation of the results is hindered by the redundancy of the reported protein identifications. To overcome this problem, various approaches have been proposed in metaproteomic studies to group redundant protein identifications [4, 209, 15].

In pure culture proteomics, the maximum parsimony approach attempts to explain the peptide identifications by a minimum protein set [262]. However, this strategy collides with heterogeneous data from microbial community samples, as the presence of a particular protein from a certain taxonomy can hardly be determined confidently. Consequently, from a given group of proteins, it is hard to decide which of the identifications should be highlighted or excluded completely. In addition, the quantification of individual proteins is complicated, since label-free quantitative measures, such as spectral counting or intensity-based methods, do not account for multiple protein candidates within a group.

To tackle this advanced protein inference problem, several rules for the protein grouping were implemented into the processing workflow (Section 3.1.1). In this work, the term *meta-protein* was defined as a protein group being generated based on one of the rules described in the following.

At the beginning of the protein grouping process, a meta-protein is generated for each protein identification holding the same features as the original protein. Similar as in hierarchical clustering strategies, the meta-proteins are merged when the rules are applied as explained in the following paragraphs. Each grouping method can be executed individually or in combination with other rules.

Leucine vs Isoleucin Distinction. MS instruments are not capable of distinguishing between the amino acids leucine and isoleucine due to their identical masses. Therefore, those peptides which differ only in these amino acids are considered equal (Figure 3.4a).

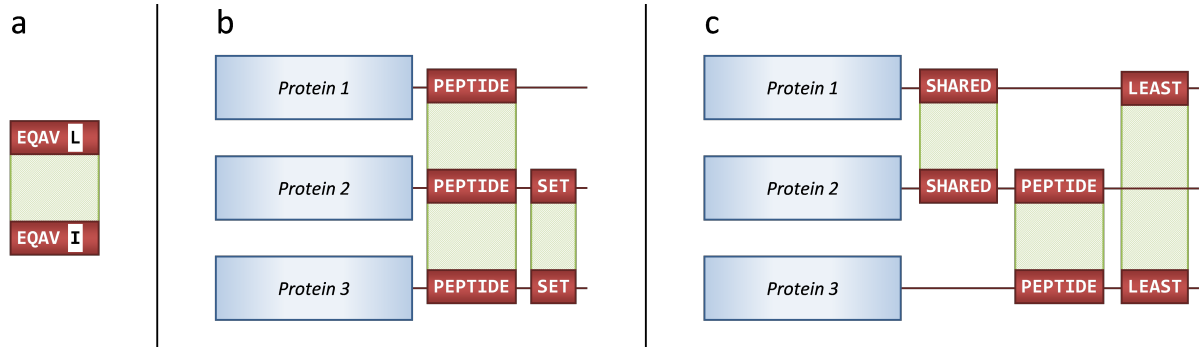

Figure 3.4: Peptide rules for meta-protein generation. Green areas visualize identical peptides. (a) Peptides differing in the amino acids leucine and isoleucine are considered identical. (b) Proteins 2 and 3 hold the same set of peptides. Protein 1 shares a subset. (c) All three proteins share exactly one single peptide with each other. Figure adapted from Muth *et al.* [256].

Protein Grouping based on Shared Peptides. In the data analysis of microbial community samples, it can be often observed that peptide identifications originate from homologous proteins expressed by organisms from different species. To reduce the size of the final protein result without excluding any valuable hits, the proteins can be grouped according to their shared peptides. In the MPA software, the rules for grouping of proteins differ with respect to the granted overlap of the peptide sets: In the *All Shared* rule, proteins are grouped when they have a whole peptide set or a subset in common (Figure 3.4b). Conversely, in the *Minimum One Shared* rule, proteins form a group when they share at least one single peptide (Figure 3.4c).

Mutation-Tolerant Grouping. Per definition, homologous proteins are related proteins that were derived from a common ancestor. Although proteins from different species might differ from one another by their sequence, they still fulfill similar or identical functions. As a consequence, an exact string matching of peptides during the grouping may not be sufficient with respect to the illustrated biological concept of homology. In addition, horizontal gene transfer frequently occurring in bacteria even enhances sequence mutation events. Since such changes are often translated from the genome to the proteome level, a high sequence variance leads to issues in the data analysis. As a consequence, a specific metric was required to account for the described sequence variability in metaproteomic data. To measure the similarity between peptides, the Levensthein edit distance (ED) was implemented in the MPA software. This straightforward string metric calculates the distance between two given peptide sequences a and b (Equation

3.1). The number of point mutations are counted which are required to transform one peptide sequence into another. The mutations are defined as deletion, insertion or substitution of one or more amino acids. In the MPA application, the user can specify the ED parameter value to refine the peptide-based grouping of proteins described in the previous paragraph. Eventually, proteins are merged when they share peptides which are considered equal according to the specified maximum ED value.

$$ED_{a,b}(i,j) = \begin{cases} max(i,j) & \text{if } min(i,j) = 0 \\ min \begin{cases} ED_{a,b}(i-1,j)+1 \\ ED_{a,b}(i,j-1)+1 \\ ED_{a,b}(i-1,j-1)+1 \end{cases} & \text{otherwise} \end{cases} \tag{3.1}$$

Protein Cluster Rule. The grouping of proteins can be also achieved based on the similarity of the sequences in the result set. For this purpose, the *Protein Cluster Rule* uses information from the UniRef (UniProt Reference Clusters) public resource that provides clustered sets of sequences from the UniProtKB [165, 263]. UniRef supports three sequence identity levels: the Uniref100 database merges identical sequences and subfragments from different organisms into a single entry. The UniRef90 and UniRef50 databases provide clustered sequences at 90% and 50% identity levels based on the UniRef100 entries. Using this resource, a UniRef cluster can be retrieved for each protein identification with a given UniProt accession. In the MPA software, the meta-proteins are then generated by reflecting the UniRef cluster assignments. Since proteins with a certain level of sequence homology also have a fair chance of sharing similar biological functions, the time and complexity of the data analysis can be reduced using this rule.

Taxonomy Rule. Given the information about the taxonomic origin of the protein species, the protein grouping can be further refined to merge only those proteins whose lineages converge below a specific taxonomic level. The idea behind the *Taxonomy Rule* is that proteins and their respective taxa are assumed to be more closely related at lower phylogenetic ranks. This method can be used to control the phylogenetic diversity during the meta-protein generation. Thus, limiting the taxonomic convergence level to a specified maximum rank results in closer relationships within the protein groups. This rule cannot be used in isolation, but in combination with other grouping rules described in the previous paragraphs.

Taxonomy Definition. As mentioned in the previous section, the taxonomic origin can have an impact on the protein grouping and the final evaluation of the results. In general, the proteins in public repositories, such as UniProtKB [165], are reported together with their species or subspecies from which they have been isolated. However, due to the enhanced protein inference issues with microbial community samples, it is difficult or even impossible to confidently select particular species that contributed mostly to the results, as described in more detail in Section 3.1.2. To better cover the uncertainties of the taxonomic origin, the individual protein taxonomies can be reconstructed by the following method: the so-called *Taxonomy Definition* process is based on the shared peptide associations of each individual data set and is therefore able to locally modify the taxonomic lineage of the proteins given from the original database. The algorithm is sequentially executed in three steps. In the first step, each peptide in the data set retrieves the information of the originating protein taxonomy (Figure 3.5a). Shared peptides obtain the LCA of all connected taxonomic lineages, as explained in the previous section. The second step transfers the retrieved peptide taxonomies back to the associated parent proteins (Figure 3.5b) and is mandatory to apply the previously described *Taxonomic Rule* for the meta-protein generation. Subsequently, on the protein level, a particular protein taxonomy can also be inferred by determining the LCA of all peptides linked to that protein. As an additional alternative, the most specific taxonomy (MST) is provided to preserve the peptide-level specificity. As final step, the taxonomy of the protein groups can be retrieved by performing the second step for meta-proteins and proteins (Figure 3.5c).

3.1.3 Graph Database System

In addition, as any software development process is dependent on the user requirements, the graphical interface and the output format are often tailored towards particular specifications. However, it is hard to predict all use cases and users also may develop novel questions during the data investigation. Therefore, instead of limiting the software during design and development to static interaction possibilities, it is advisable to provide a maximum flexibility for the analysis of the data to the user.

For the metaproteomic data analysis, the graph database system Neo4j (http://www.neo4j.com) (version 1.8) was integrated into the MPA software to enable a user-defined querying of the results based on a specific query language called Cypher. The graph database differs strongly from a classic relational database system. The relational model organizes data as tuples that represent ordered lists of elements grouped into relations [264]. A relation forms a table and describes a set of tuples in which each member is part of a data domain. Conversely, the graph database system consists of a graph structure with *nodes* (vertices) and *relationships* (edges). Both of these

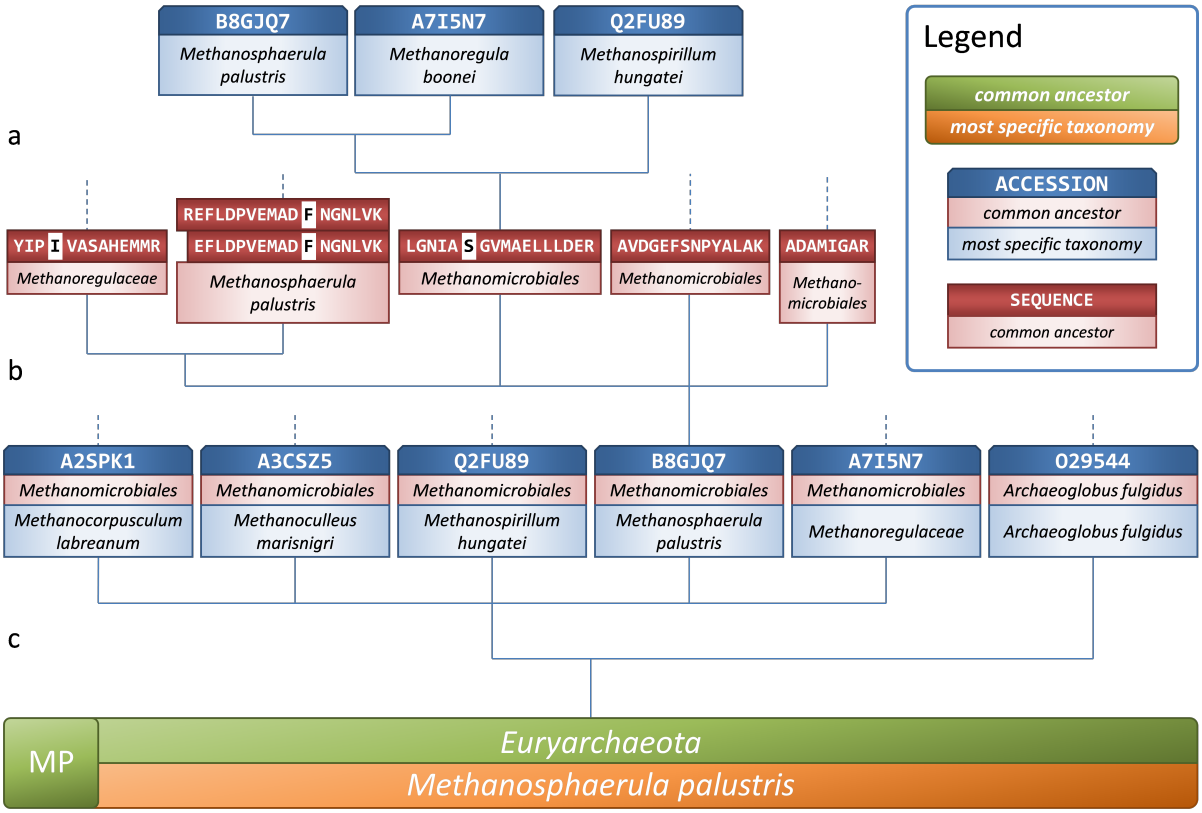

Figure 3.5: Example of the taxonomy definition process. (a) The taxonomy for the peptide *LGNIASGVMAELLLDER* is assigned as the LCA derived from all three protein hits sharing this peptide (*Methanomicrobiales*). (b) Protein *B8GJQ7* will receive either the common ancestor taxonomy (*Methanomicrobiales*) or the most specific taxonomy (*Methanosphaerula palustris*) of all peptides assigned to it. (c) A meta-protein will be classified in similar fashion using either LCA or MST of all its associated proteins (Euryarchaeota or Methanosphaerula palustris). Figure adapted from Muth *et al.* [256].

entities can be labeled with a name representing its function. The relationships always connect two nodes with each other. Additional key-value pairs, so-called *properties*, can be used to describe particular attributes for nodes and relationships. Similar to relational systems, the graph database is fully transactional to ensure data integrity.

While relational databases have been the workhorses for decades in almost any IT environment, one of the major disadvantages presents their rigid schema that makes it hard to add new relationships between entities. Moreover, various issues of scalability exist due to the high amount of upcoming data in modern applications, since the querying of the data involves many computationally expensive JOIN operations on the tables in the database. In contrast, the graph database inherently avoids such operations by accessing connected nodes directly in the structure. In particular for large data sets, the graph database is better scalable than relational systems. Another advantage of the Neo4j graph database is that it runs both in server and embedded mode. The MPA client application employs the embedded variant without the need of a separate

Table 3.1: Node types and descriptions for the graph database schema. Table adapted from Muth *et al.* [256].

Node type	Node description
Proteins	Identified proteins; properties include protein accession, description, sequence coverage, species and spectral count.
Peptides	Identified peptides; properties include peptide sequence and spectral count.
PSMs	Peptide-spectrum matches; properties include spectrum identifier and search engine score.
Taxonomies	Taxonomies; properties include taxonomy name, NCBI taxonomy ID, and rank.
Ontologies	UniProtKB ontologies; properties include ontology name and category (for example, biological process).
Pathways	KEGG pathways; properties include KO number and KEGG description.
Enzymes	EC-based enzymes; properties include EC number and description.

database server. This mode also offers low latency and complete control of the database life cycle.

Table 3.1 displays the node types and descriptions of the graph database schema. Moreover, the respective relationship types are described with outgoing and incoming relationship direction (Table A.1 in the appendix).

3.2 Experimental Data

In this work, different metaproteomic data sets were used to evaluate the performance of developed methods and software. The first data sets feature the metaproteomes of microbial community samples from different biogas plants (Section 3.2.1). The next data sets belong to human intestine metaproteomes from lean and obese individuals (Section 3.2.2). The third data set features a single-species proteomic sample which is employed as benchmark experiment (Section 3.2.3). All data sets were obtained by LC-MS/MS.

3.2.1 Biogas Plant Samples

Different BGP data sets were used for the evaluation of the developed metaproteomic data analysis workflow. The samples which were obtained from BGPs at different locations are described in the two following paragraphs.

EBENDORF. The technical replicate data sets EBENDORF01 and EBENDORF02 represent the metaproteome of a complex microbial community sample derived from an agricultural BGP located in Magdeburg/Ebendorf (Saxony-Anhalt, Germany). Details on the sample preparation and LC-MS/MS measurements can be found in the original publication in Journal of Proteome Research [256]. In the same study, the main process parameters and the substrate feed composition are summarized.

GENT. For more detailed investigations on different BGP data, three exemplary MS/MS data sets were used: The data sets GENT01, GENT07 and GENT16 with two technical replicates each were obtained from samples of different reactors located in Gent (Belgium). The samples were analyzed in a 16s RNA gene study by DeVrieze *et al.* [265]. The GENT01 and GENT07 samples were derived from continuously stirred tank reactors. The substrate used for the GENT01 reactor was an organic fraction of municipal solid waste, while the GENT07 sample was derived from a reactor which was fed with a mix of maize silage and chicken manure. In contrast, the GENT16 sample belonged to a steady-state operating reactor suspension in an industrial scale upflow anaerobic sludge blanket reactor using brewery waste water as substrate. The respective process parameters and the substrate feed composition of the fermenters are summarized in the original study [265]. Additional details regarding sample handling and LC-MS/MS measurements can be found in the study of Kohrs *et al.* [131].

3.2.2 Human Intestine Metaproteomes

MS/MS data sets derived from 29 HIMP samples were processed via the MPA software and used for the evaluation of the metaproteomic data analysis in this work. The processed data were also used for a comprehensive study that investigated signatures of bacterial and host proteins in the colon of obese and non-obese individuals [266]. Sample handling and LC-MS/MS measurement are described in detail for the respective samples in the aforementioned publication. Moreover, the individuals took part in a larger study as described in the publication of Verdam *et al.* [267]. HIMP10 presents a subset of ten HIMP data sets (P1, P3, P8, P11, P17, P23, P27, P28, P31, P34) that was used for the major computational analysis steps. For the functional analysis, the data from

all 29 HIMP samples were taken to separate the result sets from obese and non-obese individuals on the basis of bacterial proteins (Section 4.6.3).

3.2.3 Pyrococcus Furiosus

To further inspect issues related to the influence of database selection and parameters on the search results in the identification workflow, a *Pyrococcus furiosus* (PFU) sample was used which contained 14 467 MS/MS spectra as described in the study of Vaudel *et al.* [268]. The PFU data set was used for the benchmark analysis since it stands out prominently from the HIMP data: PFU is a hyperthermophilic archaeon featuring an unique biochemistry and high taxonomic distance to other species [269, 270].

3.2.4 Mixture of Nine Organisms

To evaluate the performance of the MPA software with respect to the reliability of taxonomic assignment process, data from a sample of known microbial composition was required. The 9MM sample was used from a study that evaluated the impact of using different protein sequence databases [126]. Table 3.2 displays detailed information on the nine bacterial and eukaryotic species contained in the mixture sample. In the original analysis of the study, two complementary methods were used for the sample preparation, namely filter-aided sample preparation (FASP) [271] and protein precipitation followed by in-solution digestion (PPID) [272]. The MS/MS data sets are denoted 9MM_FASP and 9MM_PPID in this work. Further details on sample handling and LC-MS/MS measurement can be found in the original publication by Tanca *et al.* [126].

Table 3.2: Microorganisms of the 9MM sample. Table adapted from Tanca *et al.* [126]

Species	Description	Genome size
Escherichia coli	Gram-negative bacillus	4 600 Kb
Pasteurella multocida	Gram-negative coccobacillus	2 250 Kb
Brevibacillus laterosporus	Gram-variable bacillus	5 180 Kb
Lactobacillus acidophilus	Gram-positive bacillus	1 993 Kb
Lactobacillus casei	Gram-positive bacillus	2 900 Kb
Enterococcus faecalis	Gram-positive coccus	3 128 Kb
Pediococcus pentosaceus	Gram-positive coccus	1 832 Kb
Rhodotorula glutinis	Yeast	20 300 Kb
Saccharomyces cerevisiae	Yeast	12 068 Kb

3.3 Protein Sequence Databases

3.3.1 UniProtKB (SwissProt/TrEMBL)

UniProtKB presents a public repository of protein sequence information and meta-information [165]. SwissProt is the manually annotated, non-redundant and curated part of UniProtKB. It is supported by information extracted from the literature and computational analysis of the curators. In this work, a local SwissProt database containing 547 599 entries (version 2013/02/20) was used in FASTA format for database searches with X!Tandem, OMSSA and MASCOT.

In addition, TrEMBL representing the unreviewed portion of UniProtKB was taken for the identification of MS/MS spectra from metaproteomic experiments. The main rationale was to compare the performance of both UniProtKB variants during the identification of metaproteomic data. Since the TrEMBL mainly covers computationally annotated protein information, far more sequences (27 122 814 entries) are contained in this database than in SwissProt. TrEMBL is also non-redundant with respect to full-length protein sequences occurring only once. However, due to the non-curated creation process, protein fragments, splicing isoforms and other variants are stored in separate entries. A local version of TrEMBL (version 2013/02/20) was used in FASTA format for the database searches with X!Tandem and OMSSA. Both UniProtKB variants were employed as target databases for the identification of MS/MS spectra from BGP samples (see Section 3.2.1).

3.3.2 Biogas Plant Metagenome (BGPMG)

The biogas plant metagenome (BGPMG) database consists of a combination of four metagenomes from full-scale and lab-scale BGPs. BGPMG features a total of 452 170 protein entries. Further details concerning the composition of this database can be found in the study by Kohrs *et al.* [131]. The BGPMG database was used as alternative resource to UniProtKB for the analysis of BGP metaproteome data sets (see Section 3.2.1).

3.3.3 Human Intestinal Metaproteome Database (HIMPdb)

The Human Intestinal Metaproteome database (HIMPdb) consists of 6 153 068 protein sequences from different sources, such as metagenomes, bacterial genomes, plant genomes and the human genome (Table 3.3). Thus, this manually created FASTA database covers a broad range of proteins which are expected to occur in human fecal samples. Hence, HIMPdb was used as the main target database for the identification of MS/MS spectra from human intestine data sets (see Section 3.2.2). Moreover, two particular subsets of HIMPdb were employed separately to study the effect of the database size: intestinal metagenome data obtained from 124 individuals (Qin2010db) [42] and a collection of 594 bacterial genomes (Bact594db). More detailed information on the exact composition of Bact594db can be found in the supplementary of the original publication [273].

Table 3.3: Composition of the human intestine metaproteome database. Name, description and number of protein entries are shown for each database. Table adapted from Muth *et al.* [273].

Database name	Description	Number of entries
HIMPdb	Concatenated target database	6 153 068
Bact594db	594 bacterial genomes	1 850 744
Qin2010db	124 metagenomes of European subjects [42]	3 267 604
Kurokawa2007db	13 metagenomes of Japanese subjects [274]	600 752
Human2010db	Human protein sequences (Integr8/Genbank)	69 879
Human2010altdb	Putative human protein sequences (Genbank)	116 718
FoodSourcesdb	Plant protein sequences (UniProtKB)	247 371
HIMPdb (refined)	Extracted proteins after first search	90 040

3.3.4 Pyroccoccus Furiosus Database (Pyrodb)

For benchmark experiments, MS/MS spectra from *Pyrococcus furiosus* strains (see Section 3.2.3) were searched against the FASTA database Pyrodb that contains 2 139 *Pyrococcus furiosus*, 7 325 *Saccharomyces cerevisiae* and 50 *Homo sapiens* protein entries (UniprotKB/SwissProt). In addition, the protein sequences of HIMPdb and Pyrodb were concatenated for a benchmark analysis. This database is referred to as PyroHIMPdb and contains 6 162 852 sequence entries in total.

3.4 Employed Software

3.4.1 X!Tandem

X!Tandem presents an open-source database search algorithm for identifying peptides and proteins from MS/MS spectra [170]. X!Tandem (version 2013.02.01) was used via the in-house MPA server for all conducted database searches throughout this work. Trypsin was the default enyzme cleavage parameter and a maximum of one missed cleavage was allowed. Carbamidomethylation of cysteine was selected as fixed, and oxidation of methionine as variable modification. For the HIMP data sets, the fragment ion tolerance was set to 0.4 Da and the precursor tolerance to 0.03 Da. For the BGP data sets, a fragment ion tolerance of 0.5 Da and a precursor tolerance of 10 ppm were used. For some explicitly mentioned experiments, the default parameters were modified with respect to the maximum of missed cleavages and the cleavage enzyme.

3.4.2 OMSSA

OMSSA constitutes another freely available protein database search engine developed by the NCBI [171]. OMSSA (version 2.1.8) was mounted into the MPA server application as second algorithm and used for the protein and peptide identification in this work. As for X!Tandem, the same parameters for the respective data sets were employed, only for some explicitly mentioned experiments, the default parameters were modified.

3.4.3 MASCOT

MASCOT is a commercial algorithm for matching MS/MS spectra against protein sequence databases [169]. Partly, BGP data sets were searched via MASCOT (version 2.2) using the following parameters: trypsin, one missed cleavage, monoisotopic mass, carbamidomethyl (C) as fixed and oxidation (M) as variable modification, precursor tolerance of 10 ppm , fragment ion tolerance of 0.5 Da, 1 ^{13}C and +2/+3 peptide charge. The resulting MASCOT DAT files containing peptide and protein identification information were uploaded via the MPA client to the server application.

3.4.4 DeNovoGUI

From a data analysis perspective, the most important requirement for the identification of proteins in MS-based proteomics is an appropriate protein sequence database as target for the search algorithms. In metaproteomics, however, the incompleteness of the protein sequence databases is problematic, since many bacterial strains have not been sequenced or are even considered unculturable (Section 2.2.3). The method of *de novo* sequencing yields the potential to overcome this issue by obtaining the peptide sequences directly from the MS/MS spectra (Section 2.2.4). Despite the potential of the *de novo* sequencing approach, algorithms, such as PEAKS [201] and PepNovo+ [200], are not widely applied by the community. While PEAKS presents a software package that is only commercially distributed, PepNovo+ is freely available, but has several shortcomings: Importantly, the PepNovo+ algorithm is only available as command line tool, which lowers the adoption of the software in proteomics labs. Moreover, this algorithm lacks further essential features, such as support of the standardized controlled vocabulary for PTMs, a multi-core implementation for running searches in parallel and an output format that provides fragment ion and spectrum annotation.

The software DeNovoGUI was developed to provide a front-end application for the PepNovo+ algorithm (Figure 3.6). The software was originally published in Journal of Proteome Research [275]. Besides the incorporated function of a graphical user interface, the feature of an automated parallelization across multiple compute cores was added to the tool to accelerate the *de novo* sequencing process for a large amount of MS/MS spectra, as commonly provided by metaproteomic experiments. The application also allows to add typical PTMs, such as oxidation of methionine, and further modifications that can be customized by the user.

For the *de novo* sequencing of metaproteomic data, the PepNovo+ algorithm (version 3.1) [200] was used via DeNovoGUI (version 1.2.0) in multi-threaded mode using four compute cores. The parameter values were chosen in accordance with the ones used for the database search algorithms: precursor ion tolerance of 0.03 Da, fragment ion tolerance of 0.5 Da, carbamidomethylation as fixed PTM and oxidation of methionine as variable PTM. The default fragmentation model (CID_IT_TRYP) was used which stands for CID fragmentation and tryptic cleavage. The maximum number of *de novo* peptide solutions was set to 20. As further described in the respective results section, the *de novo* peptides were classified according to their PepNovo+ score.

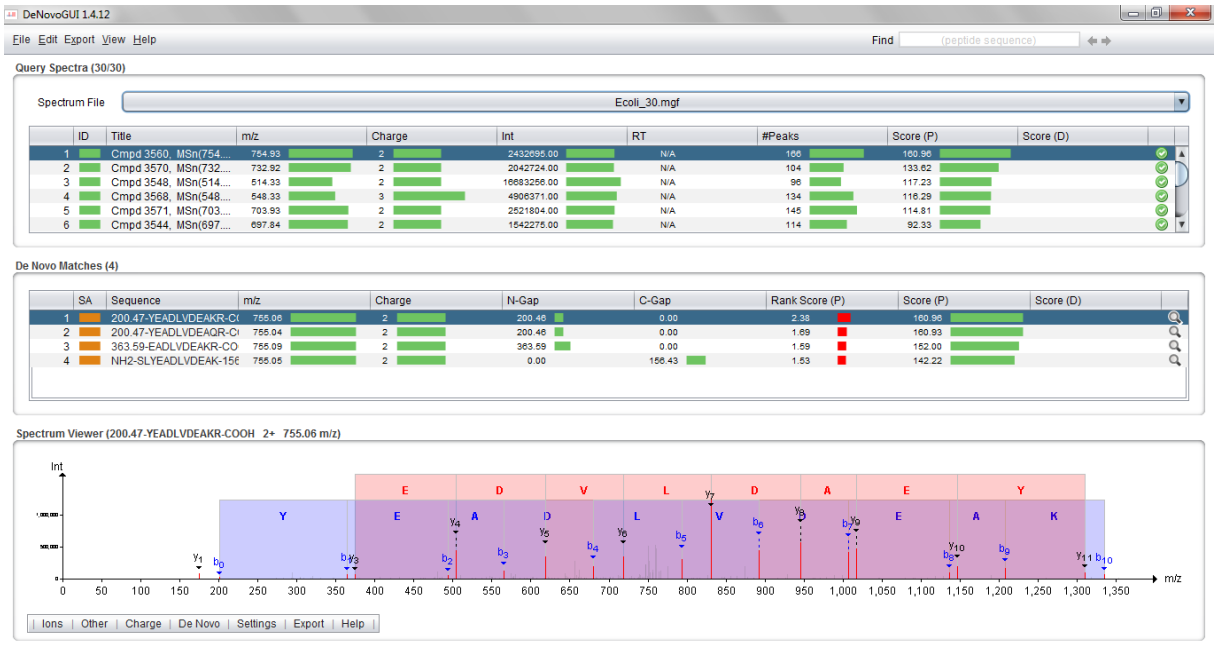

Figure 3.6: Main panel of the DeNovoGUI graphical user interface. Figure taken from Muth *et al.* [275].

3.4.5 Unipept

For the taxonomic assignment of peptides, the web-based tool Unipept (version 2.3) [276, 277] was used with the following parameters: equating isoleucine/leucin, filtering for duplicate peptides and advanced missed cleavage handling. The respective peptide sequences were exported using the MPA client application and subjected to the metaproteomic analysis feature on the Unipept website (http://unipept.ugent.be).

3.4.6 EggNOG

For the functional analysis of HIMP data, unannotated protein sequences were extracted from the MPA client application and subsequently matched to the Evolutionary Genealogy of Genes: Nonsupervised Orthologous Groups (EggNOG) version 4.1 containing bacterial orthologous gene clusters [213] using the HMMER algorithm (version 3.1) [278] with default parameters. The resulting nonsupervised orthologous groups (NOGs) were filtered by accepting the best search hit with an e-value below 10^{-30} for each query.

3.4.7 LEfSe

The linear discriminant analysis effect size (LEfSe) method was used to find characteristic microbial features at the taxonomic and functional level [279]. LEfSe is a supervised classification approach that first determines features by a non-parametric factorial Kruskal-Wallis sum-rank test that are statistically different among biological groups. Subsequently, the biological consistency of identified differences across subgroups is evaluated by employing an unpaired Wilcoxon rank-sum test. Finally, the effect size of each differentially abundant feature is estimated using linear discriminant analysis (LDA) to determine the magnitude of variation of the features between the groups. In this work, a significance level of alpha = 0.01 was chosen for both described statistical tests. An LDA $\log_{10}$ score threshold of 2.0 was applied to filter for markedly increased bacterial features.

3.5 Applied Methods

3.5.1 Target-Decoy Approach

In proteomic workflows, the statistical significance of peptide and protein identifications in a data set is usually assessed by estimating the FDR [280]. The FDR is defined as the expected proportion of *false positives* (FP) among multiple hypotheses. The target-decoy approach (TDA) presents the most common method of determining the FDR for an entire result data set [186, 187]. The TDA can be easily adopted to any database search workflow: it merely requires a target database that contains protein sequences appropriate to the protein mixture to be analyzed, and a decoy database which can be directly generated by reversing, shuffling or randomizing the protein sequences of the target database [206]. The aim is to minimize the amount of common peptides between target and decoy database. The major two strategies are the following: the first one is to append the decoy database to the target database resulting in a composite database twice the size of the original. The second option to search separately against the target and decoy database. The resulting number of identifications in the decoy database is then used to estimate the incorrect hits obtained when searching the original target database. Consequently, the FDR is calculated as the ratio of the number of decoy hits above a given threshold to the number of target hits above the threshold (Equation 3.2).

$$FDR = \frac{N_{decoy}(FP)}{N_{target}(FP + TP)} \tag{3.2}$$

While the FDR applies globally to a collection of PSMs, it is also useful to assign statistical

scores to individual hits: for this purpose, q-values are used to describe the minimum FDR at which a PSM is accepted [258, 259]. For example, a q-value of 0.01 stands for 1% estimated incorrect hits in a whole set of PSMs. In the MPA software, resulting PSMs from the database searches are therefore automatically passed to the QVality algorithm [189] (see Section 3.1.1). This method calculates q-values from a provided set of target and decoy PSM scores. Eventually, the hits can be filtered according to their q-values as FDR estimates which guarantee consistency across heterogeneous search engines. In this work, FDR thresholds of 1% and 5% were used, depending on the type of analysis.

3.5.2 Quality Control and Results Combination

Besides the TDA-based FDR estimation, also the original search scores were used to check for the quality of the identifications. For the search algorithm X!Tandem, the hyperscore was taken as scoring value for each PSM. The hyperscore is calculated by multiplying the preliminary dot product with the factorials of the amount of assigned b and y ions [281, 170] (Equation 3.3). The factorials are based on the assumption of a hypergeometric distribution for the matching fragment ions.

$$hyperscore = (\sum_{i=0}^{n} I_i * P_i) * N_b! * N_y! \tag{3.3}$$

Conversely, for OMSSA, the e-value was used as score basis for each PSM. The probabilistic value was then transformed according to Equation 3.4 [282]. The purpose of this transformation was to facilitate the comparison of the PSM scores from X!Tandem and OMSSA during the evaluation of the identification quality.

$$score(OMSSA) = -10 * log_{10}(e - value) \tag{3.4}$$

In addition, the search results from X!Tandem and OMSSA were combined in various metaproteomic analyses by using the individual PSMs which were filtered by a predefined FDR threshold (see Section 3.5.1). Furthermore, the union set of identifications was retained from the search algorithms to increase the overall sensitivity of the results.

3.5.3 Identification Rescoring

Since results from both database searching and *de novo* sequencing algorithms were analysed, another scoring metric was required to evaluate the quality of identified spectra independently of the original scores from the output of the methods. Therefore, a rescoring technique was employed by calculating the relative matched ion count (RMIC) that is based on the intensities of the matched fragment ions (a/b/c, x/y/z, y-NH_3, y-H_20, b-NH_3, b-H_20, precursor MH, MH-NH_3, and MH-H_2O) of the peptide divided by the total ion current (TIC) that is defined here as the sum of all peak intensities from the respective experimental MS/MS spectrum (Equation 3.5).

$$RMIC = \frac{\sum_{p \in S} I(p)}{TIC} \tag{3.5}$$

Accordingly, an RMIC value of 0.5 implies that 50% of the MS/MS peak intensities were covered by matching fragment ion peaks. The RMIC was used as a quality control mechanism independent of the applied identification strategy, since it is solely based on the provided spectrum and peptide information.

3.5.4 Two-Step Searching

The two-step searching approach proposed by Jagtap *et al.* [125] features an iterative database search strategy. The first step involves conventional protein identification search against a target database. Without limitation to a certain FDR threshold, the protein identifications are extracted and used to generate a refined sequence database for a subsequent search. Finally, the results from this second search are validated by a TDA-based FDR estimation. The main advantage of two-step searching is to reduce the protein database size in proteomic identification workflows. In this work, this approach was evaluated with respect to its applicability on metaproteomic data.

3.5.5 Jaccard Index

The Jaccard index, also called Jaccard similarity coefficient, constitutes a statistical measure which is applied to compare and assess the similarity of different sample sets. The metric is calculated as the size of the intersection divided by the size of the union of the sample sets (Equation 3.6).

$$Jaccardindex(A,B) = \frac{A \cap B}{A \cup B} \tag{3.6}$$

Analogous to the visualization technique of a Venn diagram, the Jaccard similarity coefficient presents a commonly used numerical indicator of the similarity between two sets.

4

Results

According to a typical data analysis workflow in metaproteomics, the following chapter is divided into six parts starting with the algorithms employed for peptide and protein identification: Section 4.1 begins with comparing the results of various database search engines on the basis of different metaproteomic data sets. In the following Section 4.2, several important algorithmic parameters are investigated that influence the outcome of database searches. As an alternative to techniques requiring a protein database, the method of *de novo* sequencing is evaluated in Section 4.3. The following sections then deal with post-processing approaches beyond the level of identification: in Section 4.4, a variety of protein grouping methods is examined to handle issues of protein inference and redundancy. Since metaproteomics is strongly focused on the semantic context of protein data, methods for assigning hits to taxonomic groups are investigated in Section 4.5. Finally, the functional annotation and pathway mapping of metaproteomic results are addressed in Section 4.6.

4.1 Search Algorithm Comparison

The objective was to evaluate the performance of the workflow used for the identification of peptides and proteins in the MPA software. First, a preliminary analysis is performed on an exemplary metaproteome BGP sample using three database search algorithms. In the second part of this section, an extended identification analysis is carried out by which two algorithms were employed to process on a larger collection of ten data sets from HIMP samples. In this section,

data from the original publications in Journal of Proteome Research [256] and Proteomics [273] are partly used.

4.1.1 Preliminary Analysis

In general, differences in identification yield were expected between database search algorithms due to diverging scoring models and parameter sets. To evaluate the impact of using different algorithms in the data analysis workflow, the BGP data set EBENDORF01 (Section 3.2.1) was searched using the search algorithms X!Tandem (Section 3.4.1), OMSSA (Section 3.4.2) and MASCOT (Section 3.4.3) against the SwissProt protein database (Section 3.3.1). As the number of reported protein hits depends on the output format of the search engine [283], the identification analysis was restricted to the spectrum and peptide level to guarantee a meaningful performance comparison of the algorithms.

First of all, the spectrum identifications of each of the algorithms were compared at 5% FDR against each other: Figure 4.1a shows that each search engine provided an essential amount of unique spectrum identifications. X!Tandem turned out as the best-performing algorithm at the spectrum level by identifying 2 523 (76.6%) out of 3 295 spectra. The same search engine provided also the most unique identifications, as it identified exclusively 799 (24.2%) spectra. It can be recognized that X!Tandem also yielded the highest number of unique identifications at the peptide level since this algorithm exclusively identified 281 (28.3%) out of 992 peptides (Figure 4.1b). In total, all three search engines showed a total overlap of 309 (31.1%) identified peptides.

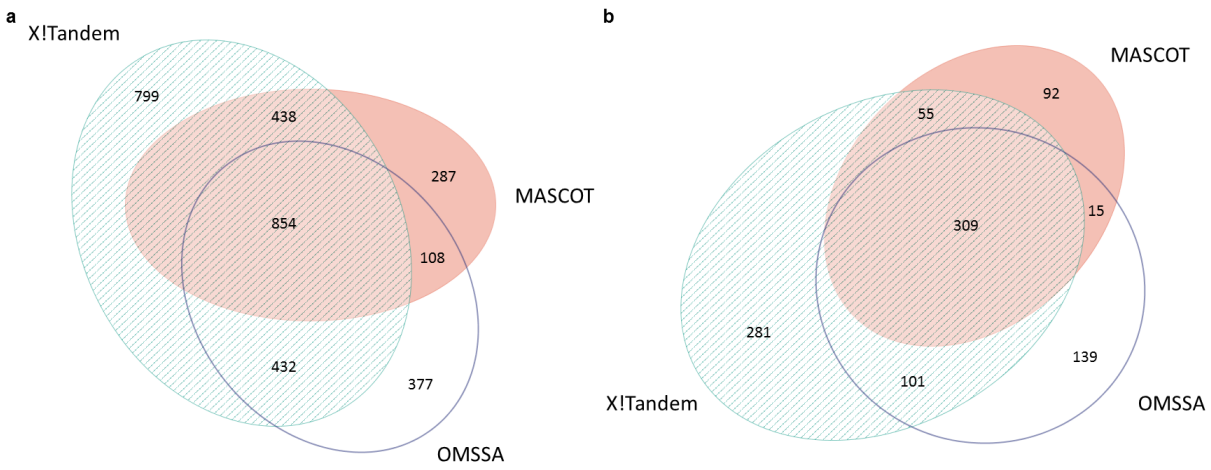

Figure 4.1: Comparison of identifications from three search engines for EBENDORF01 data set. The Venn diagrams show identified (a) spectra and (b) peptides being unique and shared for search algorithms X!Tandem, OMSSA and MASCOT at 5% FDR. Figure taken from Muth *et al.* [256].

4.1.2 Performance of X!Tandem and OMSSA

To further examine the question whether the use of multiple identification algorithms is beneficial for the analysis of metaproteomic data, database searches were performed on a representative data collection: in total, 317 375 MS/MS spectra from ten human intestine metaproteome data sets (HIMP10, see Section 3.2.2) were searched against a customized protein sequence database (HIMPdb, see Section 3.3.3) using the algorithms X!Tandem and OMSSA.

When combining both search engine results, 30.2% of the total MS/MS spectra and on average 7 322 peptides for each data set were identified at 5% FDR (Table 4.1). At 1% FDR, the percentage of identified spectra dropped to 21.2% and 6 737 peptides could be obtained on average (Table A.2 in the appendix). It can be also recognized that X!Tandem yielded significantly more identifications than OMSSA: on average, X!Tandem could identify 6 446 spectra and 4 624 peptides, while OMSSA obtained 4 084 spectral and 3 292 peptide hits. In the line with the findings of the previous investigation, it was found that a substantial part of the identifications was specific for each respective search engine: Table 4.1 shows that X!Tandem yielded 25.3% and OMSSA 10.8% exclusively identified spectra on average at 5% FDR. At the peptide level, 23% and 16% unique identifications were found by X!Tandem and OMSSSA, respectively.

Table 4.1: Total number of MS/MS spectra, percentage of identified spectra (ID), exclusive spectrum and peptide identification yield from X!Tandem and OMSSA for HIMP10 data sets P1-P34 (FDR < 5%). Table adapted from Muth *et al.* [273].

Dataset	Total	ID (%)	Peptides	Excl. Spectrum ID (%)		Excl. Peptide ID (%)	
				X!Tandem	OMSSA	X!Tandem	OMSSA
P1	35 179	31.7	8 473	21.5	11.4	19.8	16.4
P3	26 560	26.0	5 624	24.0	11.2	22.9	15.3
P8	31 891	31.6	7 640	19.1	11.2	17.5	17.1
P11	31 744	26.1	6 295	30.4	12.0	26.2	16.5
P17	32 203	31.7	8 082	19.5	11.5	18.9	16.3
P23	34 050	33.2	8 255	35.8	8.6	30.4	13.7
P27	27 339	24.8	5 266	24.8	11.2	22.6	14.9
P28	32 037	30.9	7 273	25.9	10.4	23.5	14.9
P31	35 848	34.6	9 084	30.1	9.9	25.9	15.4
P34	30 524	31.1	7 231	20.1	12.1	19.0	17.1
Average	31 737	30.2	7 322	25.3	10.8	22.7	15.8

4.2 Database Searching

For each of the algorithms investigated in the previous section, the search space is influenced by various parameters, including mass tolerance, protein database, number of missed cleavages, enzyme specificity and post-translational modifications. In this section, those parameters were examined that were expected to have the highest impact in a metaproteomic workflow. As introduced in Section 2.2.2, the protein database as target for metaproteomic samples is different in composition and size compared to the sequence databases used in pure-culture proteomic experiments. In metaproteomics, a FASTA database often needs to be manually constructed and consists of a collection of translated genomes and—preferably—also metagenomes to provide an adequate amount of target sequences for successful protein identification. Therefore, the relation between database composition and number of identifications is first examined for three exemplary BGP data sets. The second part of this section is devoted to the influence of the database size on the outcome for selected human intestine metaproteome samples. In the third paragraph, different parameter values for the number of missed cleavages are tested. The fourth part then involves the parameter evaluation of different cleavage enzymes. Finally, this section ends with analysis of a proteomic data set of known sample composition to evaluate the findings of the previous metaproteomic analyses. In the following, parts of the original publication in Proteomics [273] are used.

4.2.1 Influence of Protein Database

To investigate the impact of the database composition on the results in a metaproteomic experiment, the BGP data sets GENT01, GENT07 and GENT16 (see Section 3.2.1) were matched against three different databases: the MS/MS spectra of each of the respective samples were searched with X!Tandem and OMSSA against the databases SwissProt, TrEMBL and BGPMG. While both UniProtKB databases were publicly available (see Section 3.3.1), the BGP database was manually generated by assembling four different translated metagenomes from biogas fermenters (see Section 3.3.2).

The searches of GENT01 and GENT07 against the BGBMG metagenome database resulted in more identifications in comparison to SwissProt and TrEMBL (Table 4.2). Conversely, the highest fraction (22%) of identified spectra could be detected for GENT16 when searching against TrEMBL. For the latter data set, only a low percentage of identified spectra (2%) can be recognized in the BGPMG search result.

Table 4.2: Percentage of identified spectra (Spectrum ID) and number of peptides obtained from searching GENT01, GENT07 and GENT16 with X!Tandem and OMSSA against SwissProt, TrEMBL and BGPMG (FDR < 5%).

Dataset	Spectrum ID (%)			Peptides		
	SwissProt	TrEMBL	BGPMG	SwissProt	TrEMBL	BGPMG
GENT01	7.1	4.7	7.9	728	721	933
GENT07	2.9	2.0	10.7	508	417	1,342
GENT16	4.1	22.0	2.0	494	1 910	267

Portion of unique peptides. The goal of the next analysis was to review the portion of so-called unique peptides in the BGP result sets. In this context, a unique peptide is defined as peptide hit that exists solely in one protein of the whole result set. In contrast, a shared peptide can be ambiguously assigned to multiple proteins. The objective here was to test whether a higher fraction of unique peptides could be obtained for BGPMG, since such a metagenome database was expected to be more specific than the UniProtKB database variants which contain many homologous proteins from different species.

It was found that an average portion of 80.4% peptide identifications were classified as unique in the BGPMG searches, while this ratio was lower for SwissProt (69.9%) and TrEMBL (65.2%) searches (Table 4.3). From the investigated data sets, GENT16 resulted in the highest ratio of unique peptides: in particular, a portion of 80.1% could be obtained from TrEMBL and 87.6% from BGPMG searches.

Table 4.3: Number of peptide identifications and percentage of unique peptides obtained by searching GENT01, GENT07 and GENT16 with X!Tandem and OMSSA against SwissProt, TrEMBL and BGPMG (FDR < 5%).

Dataset	SwissProt		TrEMBL		BGPMG	
	Total	Unique (%)	Total	Unique (%)	Total	Unique (%)
GENT01	728	61.1	721	53.1	933	77.7
GENT07	508	72.8	417	62.4	1 342	75.9
GENT16	494	75.7	1 910	80.1	267	87.6
Average	577	69.9	1 016	65.2	847	80.4

Overlap between database search results. The previously reported results demonstrated that the identification yield strongly depends on the chosen protein database. The next objective was to examine this effect in more detail by reviewing the overlap of peptide identifications between the previously obtained search results for the BGP data sets.

Figure 4.2 summarizes the sets of peptides that are either commonly shared or exclusive to a particular database result for the data sets GENT01, GENT07 and GENT16 at 5% FDR. Overall, it can be found that identifications barely overlap between the metagenome database BGPMG and the public UniProtKB variants. The Venn diagrams also show that many database specific identifications were found in the searches against BGPMG: for instance, searching the GENT07 data set against BGPMG resulted in 1 208 (60.9%) exclusive peptide hits. Remarkably, the overlap was minimal between the results from the three search databases: 39 (GENT01), 42 (GENT07) and 50 (GENT16) peptide identifications were shared between SwissProt, TrEMBL and BGPMG. In comparison to GENT01 and GENT07, it can be recognized that GENT16 resulted in the highest number of database-specific peptide identifications: 1 566 (69.0%) out of 2 269 hits could be exclusively derived from the search against TrEMBL. In accordance with these findings, a low overlap was found when mapping the peptide sequences from TrEMBL searches to the protein sequences in BGPMG and SwissProt: for GENT16, only 7.4% and 14.8% of the TrEMBL peptides could be matched against BGPMG and SwissProt, respectively (Table A.4 in the appendix).

Comparison of target and decoy PSM scores. To find an explanation for the varying identification yield between the databases, the PSM scores from X!Tandem were further investigated for the BGP data searched against SwissProt, TrEMBL and BGPMG.

While the target and decoy PSM scores for TrEMBL were higher than the corresponding scores for SwissProt and BGPMG, compared to each other, the latter databases resulted in similar score ranges (Figure 4.3). The boxplots also show that target PSM scores for GENT16 searched against TrEMBL (Figure 4.3c) differed more from decoy PSM scores than respective distributions of scores for GENT01 (Figure 4.3a) and GENT07 (Figure 4.3b).

So far, the findings highlighted that the identification yield strongly depends on the used protein database. As the evaluated databases differed significantly in their composition, each of the database searches resulted in unique identifications. Also, the tested metagenome database was beneficial for an additional increase in hits. However, another important aspect became apparent when searching against a large database, such as TrEMBL that contains more than 27 million protein entries: the results suggested that the database size itself can have a major impact on the scoring and the FDR estimation during the identification process. Hence, this effect was further examined in the following section in more detail.

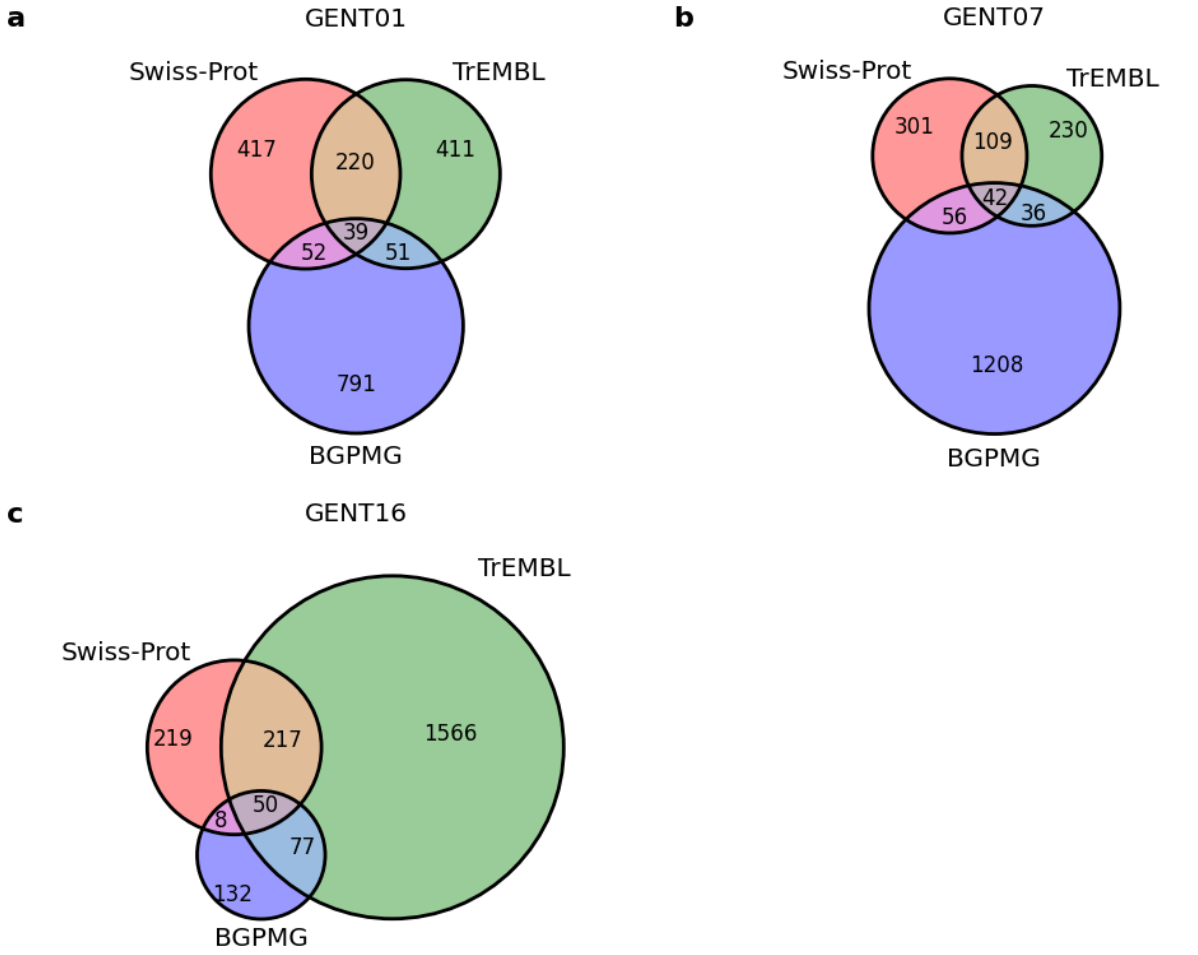

Figure 4.2: Venn diagram overlaps of peptides identified in BGP samples for three databases. The values in the circles show the numbers of identified peptides from combined database searches by using X!Tandem and OMSSA against SwissProt (red), TrEMBL (green) and BGPMG (blue) for BGP data sets (a) GENT01, (b) GENT07 and (c) GENT16 at 5% FDR.

4.2.2 Evaluation of Search Strategies

To systematically investigate the effects of the database size on the search results in metaproteomic analyses, three data sets (P1, P23 and P34) from the HIMP samples (see Section 3.2.2) were chosen for the investigations in the following. In this analysis, three different search strategies were evaluated for each of the HIMP data sets: (1) *Classic searching* was performed against a tailored protein database (HIMPdb) that had been manually constructed by integrating a variety of bacterial genome and metagenome information (see Section 3.3.3). (2) *Subset searching* was used by matching the aforementioned data sets against fractions of the HIMPdb database: Bact594db and Qin2010db were chosen as subset databases since they had delivered the most database-specific identification in preliminary analyses (Table A.3 in the appendix). (3) *Two-*

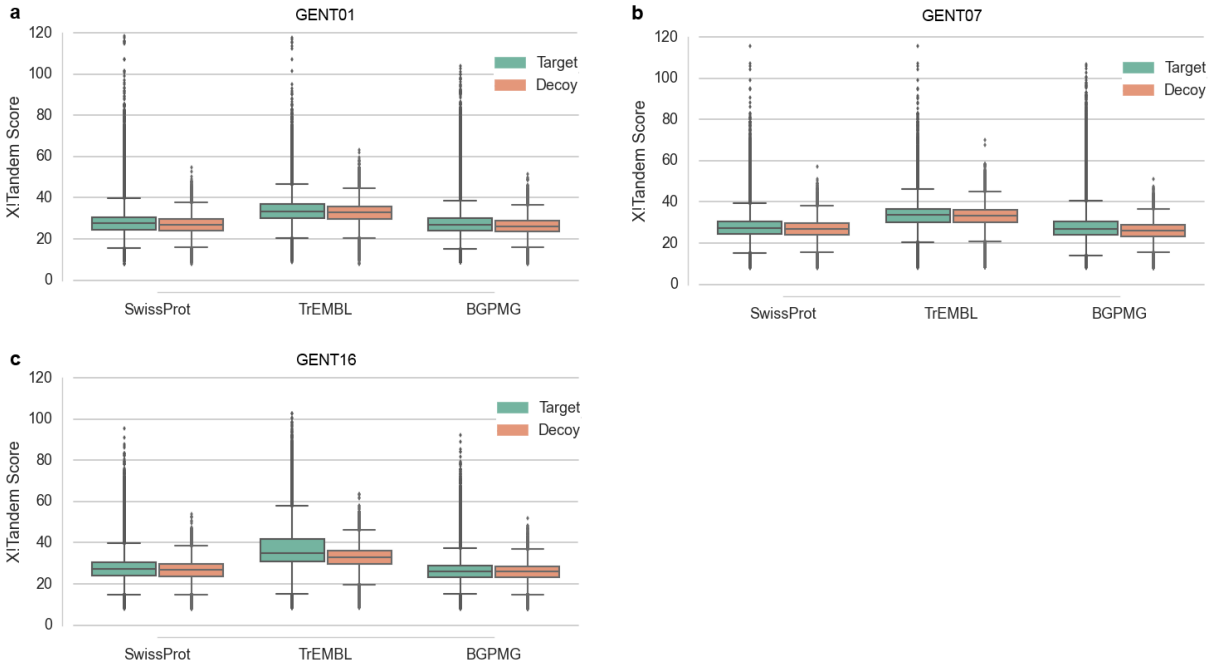

Figure 4.3: Comparison of scores from searches against three different databases for BGP data sets. Each of the grouped box plots shows target and decoy PSM scores for data sets (a) GENT01, (b) GENT07 and (c) GENT16. X!Tandem searches were performed against target and decoy databases of SwissProt, TrEMBL and BGPMG. Target PSM scores are displayed in green and decoy PSM scores in red.

step searching (see Section 3.5.4) was applied by searching the data in an initial round against HIMPdb without limiting the FDR and using the obtained proteins as target database in a second round by filtering the results with a stringent threshold (1% and 5% FDR). This method had been reported to improve the identification yield for large database searches in proteogenomics and metaproteomics [125].

Figure 4.4 shows an overview on the identification results for the data set P1: while classic searching against HIMPdb and subset searching against Qin2010db provided comparable numbers of PSMs and peptides, subset searching against Bact594db resulted in the lowest amount of identifications . It can be further recognized that two-step searching more than doubled the number of identifications in comparison to classic searching at 1% and 5% FDR. The bar plots also show that two-step searching resulted in more peptides than identified spectra at 5% FDR. It is worth noting that similar identification yields could be observed for the data sets P23 and P34 (Table A.5 in the appendix).

Analogous to the investigations on the BGP data sets in the previous text, the ratio of peptides that were uniquely identified for a single protein was also examined within the HIMP result sets for the three described search strategies.

Remarkably, two-step searching against HIMPdb resulted in the highest average fraction of

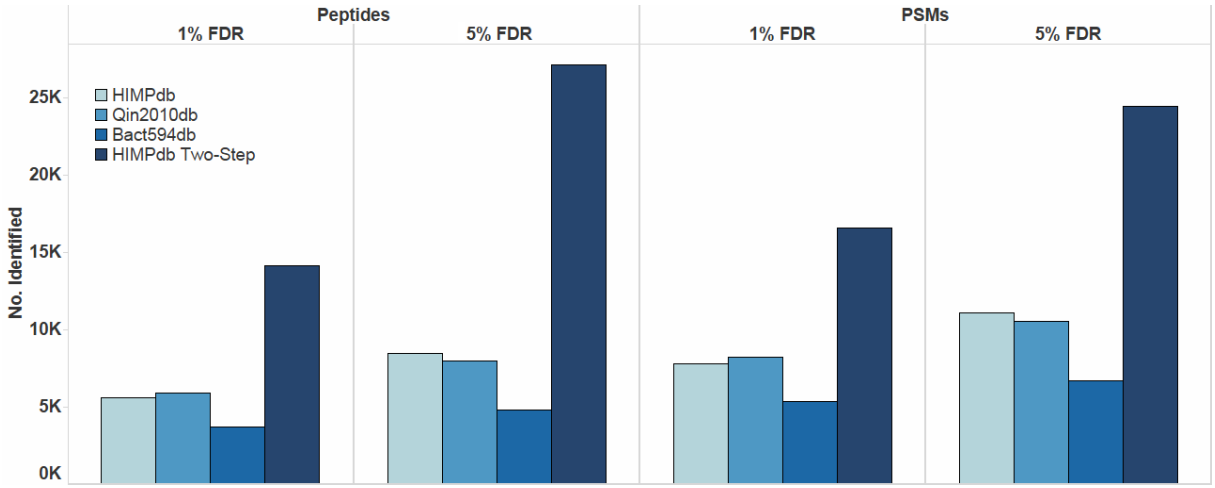

Figure 4.4: Overview on the identification results for data set P1. The bar plots display the total number of PSMs and peptides identified at 1% and 5% FDR. Classic searching was performed against HIMPdb, Bact594db and Qin2010db, while two-step searching was applied only against HIMPdb. Figure adapted from Muth *et al.* [273].

unique peptides (79.4%), while the lowest portion (51.2%) could be found for classic searching against HIMPdb (Table 4.4). Subset searching against Bact594db and Qin2010db resulted in an average of 65.3% and 68.2% unique peptides, respectively.

Table 4.4: Number of peptide identifications and percentage of unique peptides obtained by searching P1, P23 and P34 with X!Tandem and OMSSA (5% FDR). Classic searches were performed against HIMPdb, Qin2010db, Bact594db and two-step searches against HIMPdb. Table adapted from Muth *et al.* [273].

Dataset	HIMPdb		Qin2010db		Bact594db		Two-Step	
	No.	Unq. (%)	No.	Unq. (%)	No.	Unq. (%)	No.	Unq. (%)
P1	8 473	52.6	8 012	66.7	4 841	67.0	27 136	77.2
P23	8 255	51.8	7 730	69.5	4 617	66.8	31 913	81.4
P34	7 231	49.1	6 599	68.4	4 686	62.0	27 769	79.5
Average	7 986	51.2	7 447	68.2	4 715	65.3	28 939	79.4

Comparison of classic and subset searching. Using the results of the HIMP data, the numbers of PSMs and peptides that could be specifically assigned to a particular database were next investigated to illustrate the actual difference in the identification yield between classic and subset searching.

It was found that subset searching against Bact594db and Qin2010db resulted in significantly more database-specific identifications than classic searching against HIMPdb for the data set P1 (Figure 4.5). Furthermore, the number of database-specific identifications was almost one

magnitude higher for Qin2010db compared to Bact594db. Finally, similar findings illustrating the effects of subset searching could also be made for the data set P23 (Table A.6) and P34 (Table A.7 in the appendix).

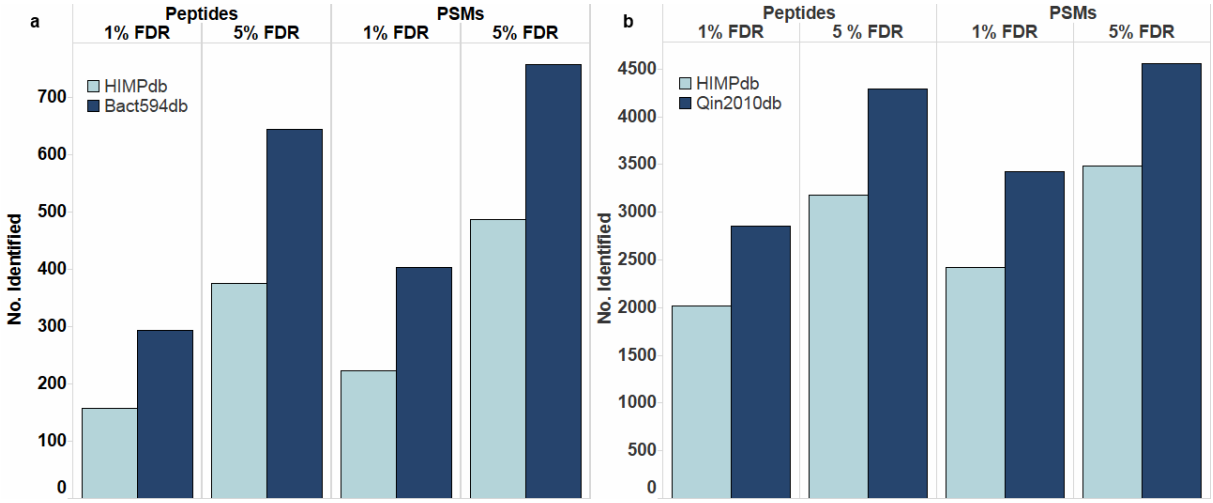

Figure 4.5: Database specific identifications from classic and subset searching for data set P1. (a) The total number of Bact594-specific PSMs and peptides are shown for classic searching against HIMPdb and subset searching against Bact594db. (b) The total number of Qin2010-specific PSMs and peptides are displayed for classic searching against HIMPdb and subset searching against Qin2010db. Figure adapted from Muth *et al.* [273].

Quality check for two-step searching results. The previous results indicate that considerable amounts of identifications were uniquely detected by subset searching against Bact594db and Qin2010db and would therefore have been lost when performing exclusively classic searching against HIMPdb. While databases of large size (> 10^6 entries) are commonplace in metaproteomics, a strategy is required to search in large databases without the issues of lacking sensitivity in classic searching. Although two-step searching was suggested as a reasonable search strategy for metaproteomics, two major findings in these data cast doubt on the reliability of this method: first, an unrealistically high identification yield was found in comparison to classic searching. Second, among the applied strategies, only two-step searching resulted in more peptides than spectra at 5% FDR. As a consequence, the next objective addressed a detailed evaluation on accuracy of the two-step searching method by comparing the PSM score distributions from classic and two-step searching for the data set P1.

Figure 4.6 displays the distributions of PSM scores from X!Tandem for classic and two-step searching filtered at 1% and 5% FDR. In comparison to classic searching, it can be recognized that the score distributions for two-step searching are shifted to the left at both FDR thresholds.

Eventually, to find an explanation for the diverging score distributions between classic and

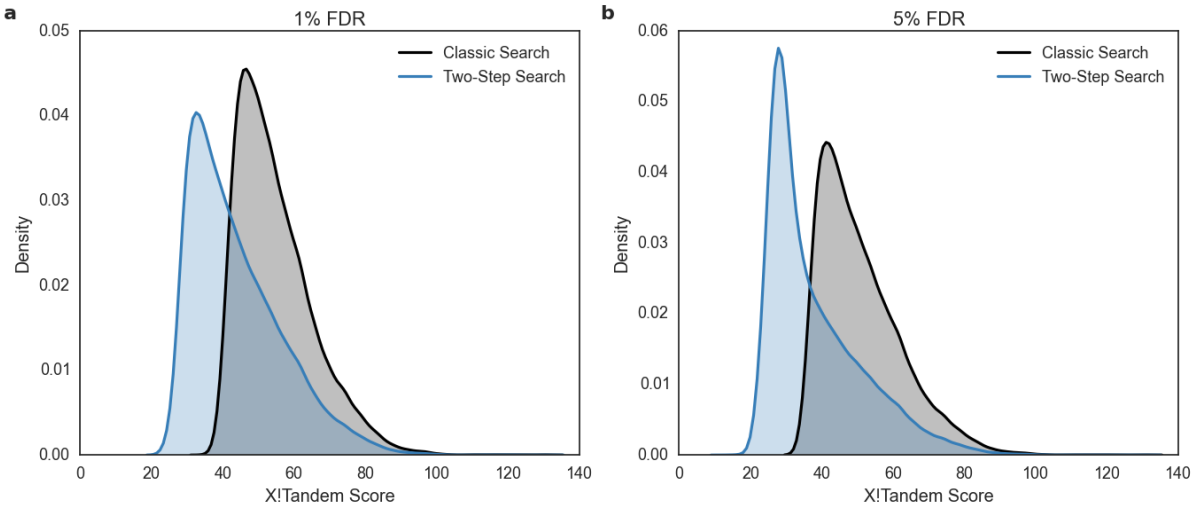

Figure 4.6: Comparison of scores between classic and two-step searching for data set P1. The density distributions of PSM scores from sample P1 are shown for classic database searching (in black) and two-step searching (in blue) against HIMPdb by using X!Tandem. Results are displayed for filtering by an (a) 1% and (b) 5% FDR threshold, respectively. Figure adapted from Muth *et al.* [273].

two-step searching, a reevaluation method to further assess the quality of the search hits was applied to the data set P1: for this purpose, the measure RMIC is calculated by summing up the intensities of the matched fragment ions for each PSM (see Section 3.5.3). To account for any particular influence of the chosen algorithm, the PSMs were investigated separately for X!Tandem and OMSSA. Such a rescoring method has the benefit of being independent of any particular search algorithm, since it only requires the input of the suggested peptide sequence and the experimental spectrum.

Figure 4.7 illustrates that the distributions of the RMIC values could be separated between both search strategies as two-step searching resulted in lower RMIC scores compared to the classic searching. It can be found that the separation of RMIC scores between the search strategies is stronger for X!Tandem (Figure 4.7a) in comparison to OMSSA (Figure 4.7b).

So far, it was demonstrated that both protein database and applied search strategy affect the outcome of data analysis workflows in metaproteomics. In the following, the influence of further search algorithm parameters are investigated in detail.

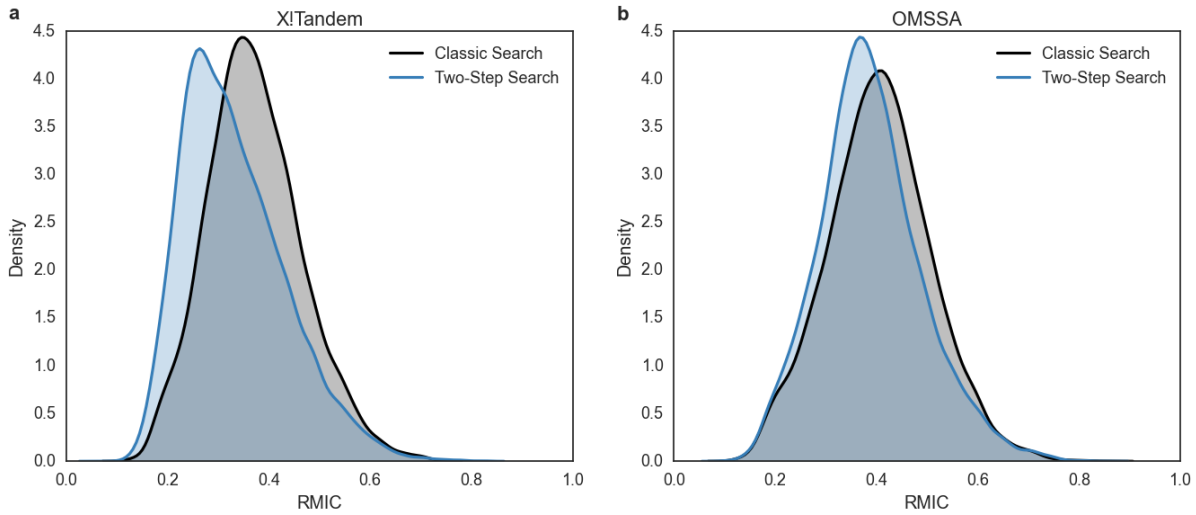

Figure 4.7: Reevaluation of identifications from classic and two-step searching for data set P1. The density distributions of RMIC values from (a) XTandem and (b) OMSSA results are shown for classic database searching (in black) and two-step searching (in blue). The identifications were obtained by classic and two-step searching against HIMPdb at 5% FDR.

4.2.3 Missed Cleavage Parameter Testing

Since the missed cleavage of peptide bonds is a common event in bottom-up proteomics due to insufficient specificity of the protease (see Section 2.1.5), the database search algorithms account for it by providing a parameter to specify the maximum number of allowed missed cleavages (MC). In the following investigation, the goal was to examine the effect of the MC parameter on the identification yield, since the sequence search space was expected to increase for each elevated MC value. Therefore, parameter values were chosen between $MC = 0$ and $MC = 3$ and database searches were performed for the HIMP data sets P1, P23 and P34 (Section 3.2.2).

The highest number of PSMs and peptides was reported for $MC = 0$ in each of the evaluated data sets at 1% FDR (Figure 4.8). At 5% FDR, most identifications could be obtained for either $MC = 0$ or $MC = 1$ depending on the chosen data set (Figure A.1 in the appendix).

Next, the influence of the MC parameter on the results was evaluated for the data set P1 in more detail: complementing the preceding analyses, the first objective was to examine the number of total peptides in dependence of the chosen MC parameter value. The second objective was to investigate how many peptide hits were exclusively found for a specific MC parameter value.

While the number of identified peptides for the evaluated MC values approached each other with increasing FDR threshold (Figure 4.9a), a significant amount of peptides could be identified exclusively for $MC = 0$ (Figure 4.9b). Similar findings could also be observed for P23 and P34 (Figures A.2 and A.3 in the appendix).

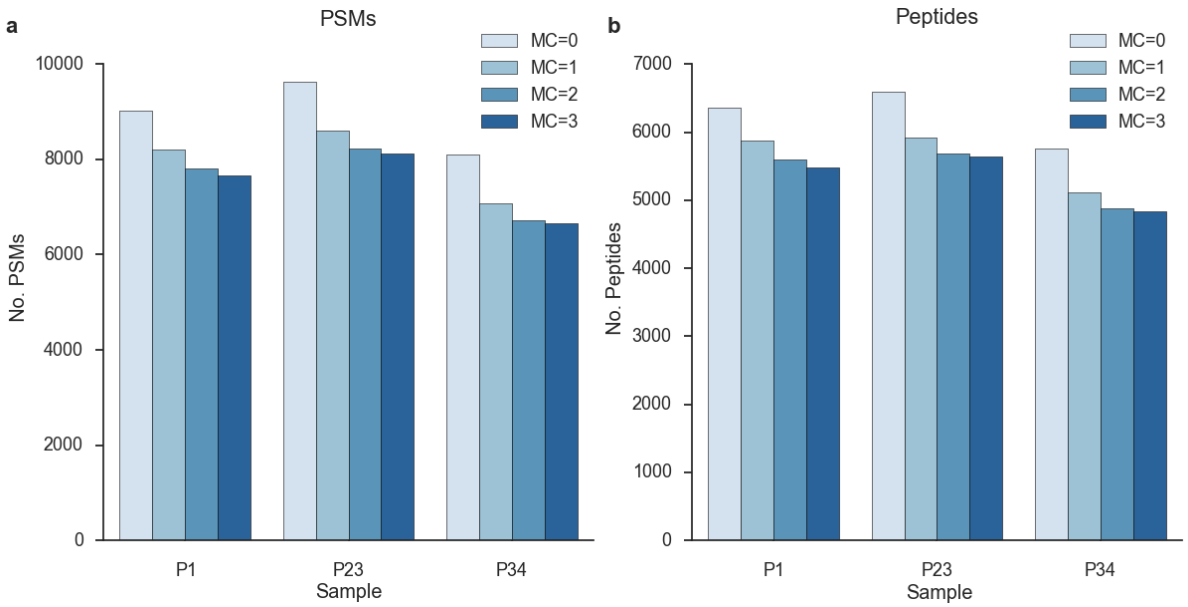

Figure 4.8: Comparative evaluation of the identification yield for different MC values (HIMP). The bar plots show the total number of (a) PSMs and (b) peptides for data sets P1, P23, and P34 when using missed cleavage parameter values *MC = 0 - 3* at 1% FDR. Identification results were combined from searching with X!Tandem and OMSSA against HIMPdb.

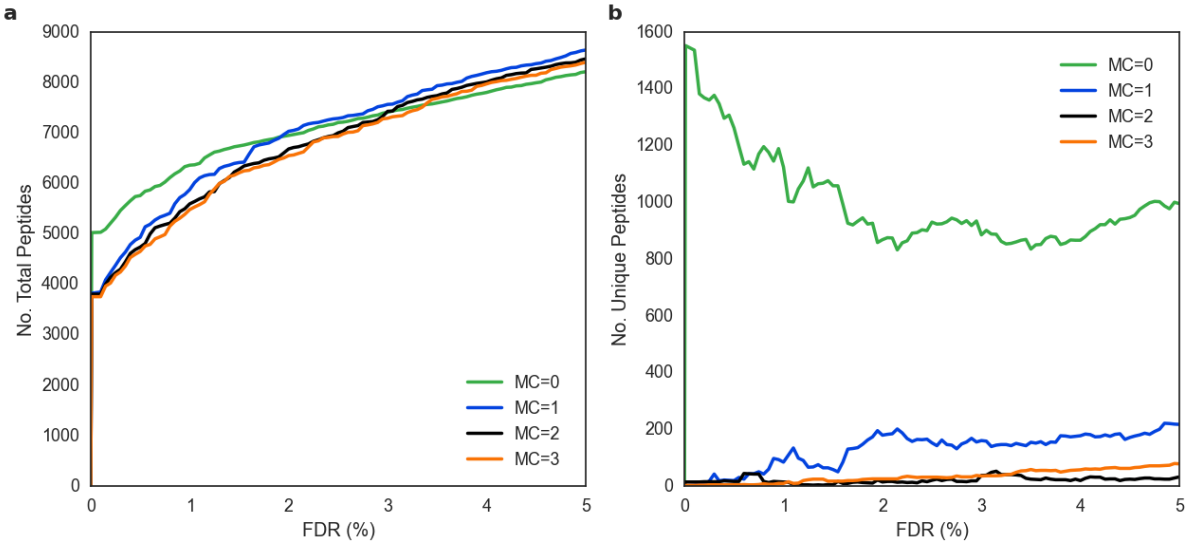

Figure 4.9: Comparison of total and exclusive peptides for different MC values (P1). The line charts display the number of (a) total and (b) exclusive peptides for data set P1 when using missed cleavage parameter values *MC = 0 - 3* as a function of the respective FDR threshold. Peptides were called exclusive when being identified uniquely for a particular MC parameter value.

4.2.4 Non-Tryptic Enzyme Settings

The final parameter under investigation was the cleavage enzyme since the hypothesis was that proteases occurring in the human intestinal tract may lead to the presence of various non-tryptic protein fragments in the metaproteome samples. Therefore, the HIMP data sets P1, P23 and P34 (see Section 3.2.2) were subjected to database searches by setting the enzyme parameter value to *semi-tryptic*. This parameter value allows one peptide terminus to be non-tryptic, while the other end remains tryptic. Since the default enzyme parameter value had been *tryptic* in the preceding analyses, the results could be directly compared for both enzyme parameter settings. In addition, *chympotrypsin* (pancreatic enzyme) and *pepsin A* (gastric enzyme) were tested as enzyme parameter values: it was speculated that the respective enzymes—representing common intestinal proteases—might be present in the HIMP samples.

Table 4.5 displays the identification results for *tryptic* and *semi-tryptic* cleavage settings at 5% FDR. It can be recognized that more PSMs and peptides were found for *tryptic* than for *semi-tryptic* as chosen cleavage parameter. The gain of using *semi-tryptic* as cleavage parameter was minimal: on average, an exclusive proportion of 1% PSMs and 7% peptides was found that could not be retrieved when using *tryptic*. Conversely, at 1% FDR, more identifications were obtained when using *semi-tryptic* as parameter (Table 4.6). Consequently, using both *tryptic* and *semi-tryptic* settings and combining the results from both search variants can serve to increase the identification yield in metaproteomic data analyses. However, a drawback presents the running time which increased five fold on average when using *semi-tryptic* instead of *tryptic* as cleavage parameter (data not shown).

Table 4.5: Number of identifications and percentage of exclusive hits for data sets P1, P23, and P34 using *tryptic* and *semi-tryptic* cleavage settings (FDR 5%). Table adapted from Muth *et al.* [273].

Dataset	Tryptic cleavage				Semi-tryptic cleavage			
	PSMs		Peptides		PSMs		Peptides	
	No.	Excl. (%)	No.	Excl. (%)	No.	Excl. (%)	No.	Excl. (%)
P1	11 133	8.1	8 473	12.5	10 354	1.1	7 959	6.9
P23	11 288	6.1	8 255	10.5	10 777	1.7	7 976	7.4
P34	9 491	8.6	7 231	13.1	8 743	0.8	6 678	5.9
Average	10 637	7.6	7 986	12.0	9 958	1.2	7 538	6.7

Finally, using the cleavage enzymes *chymotrypsin* and *pepsin A* for the database searches resulted in an insignificant number of less than 100 PSMs and peptides per data set (Table A.8 in the appendix).

Table 4.6: Number of identifications and percentage of exclusive hits for samples P1, P23, and P34 using *tryptic* and *semi-tryptic* cleavage settings (FDR 1%). Table adapted from Muth *et al.* [273].

Dataset	Tryptic cleavage				Semi-tryptic cleavage			
	PSMs		Peptides		PSMs		Peptides	
	No.	Excl. (%)	No.	Excl. (%)	No.	Excl. (%)	No.	Excl. (%)
P1	7 819	4.5	5 598	4.4	7 874	5.1	5 760	7.1
P23	8 228	1.8	5 685	1.7	8 581	5.8	6 109	8.4
P34	6 722	3.0	4 876	3.1	6 817	4.3	5 046	6.4
Average	7 590	3.1	5 386	3.1	7 757	5.1	5 638	7.3

4.2.5 Benchmark Evaluation of Proteomic Sample

To evaluate the influence of the database composition and size on the results observed in metaproteomic experiments, benchmark analyses were next performed by using a sample of known proteome content. Therefore, classic searching was performed for the PFU data set (see Section 3.2.3) against Pyrodb, a FASTA database that contained 9 514 protein sequences from *P. furiosus*, *S. cerevisiae* and *H. sapiens* (see Section 3.3.4). The main objective was to simulate the scenario of a search against the commonly large collection of protein sequences in metaproteomic experiments: for this purpose, Pyrodb was merged with HIMPdb, which served as target in preceding experiments. Subsequently, classic searching was used for the PFU data set against the aforementioned concatenated database, which is referred to PyroHIMPdb in the following. Finally, two-step searching against PyroHIMPdb was applied to further evaluate the method with respect to the identification quality.

From a total number of 14 467 MS/MS spectra within the PFU data set, classic searching against Pyrodb resulted in 10 576 PSMs and 6 408 peptides at 5% FDR (Figure 4.10). Using the same FDR threshold, classic searching against PyroHIMPdb resulted in a strong decrease in both PSMs and peptides: against the large concatenated database, 6 406 PSMs and 3 751 peptides could be obtained. In line with preceding findings of analyses on metaproteomic data, two-step searching achieved the most identifications for the PFU data set.

Figure 4.11 further illustrates the clear reduction in PSMs and peptides for the used search engines X!Tandem and OMMSA up to the FDR level of 5%. It can be also recognized that the combination of both algorithms increased the proportion of PSMs to a larger extent for PyroHIMPdb than for Pyrodb.

To understand the observed change in the amount of identifications between the applied search strategies, the target and decoy PSM scores of X!Tandem and OMSSA were investigated for classic searching (Pyrodb and PyroHIMPdb) and two-step searching (PyroHIMPdb).

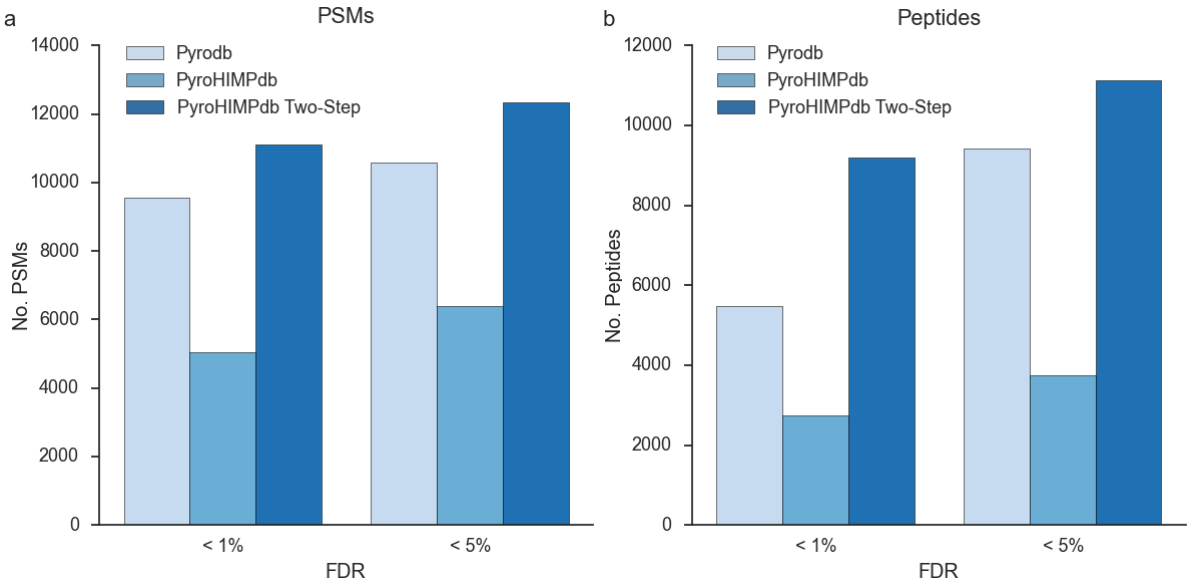

Figure 4.10: Comparative benchmark evaluation of different search strategies and databases for PFU data set. The bar plots show the total number of (a) PSMs and (b) peptides for classic searching against Pyrodb and PyroHIMPdb, and two-step searching against PyroHIMPdb at 1% and 5% FDR. The displayed identification amounts result from combined database searches by using X!Tandem and OMSSA. Figure adapted from Muth *et al.* [273].

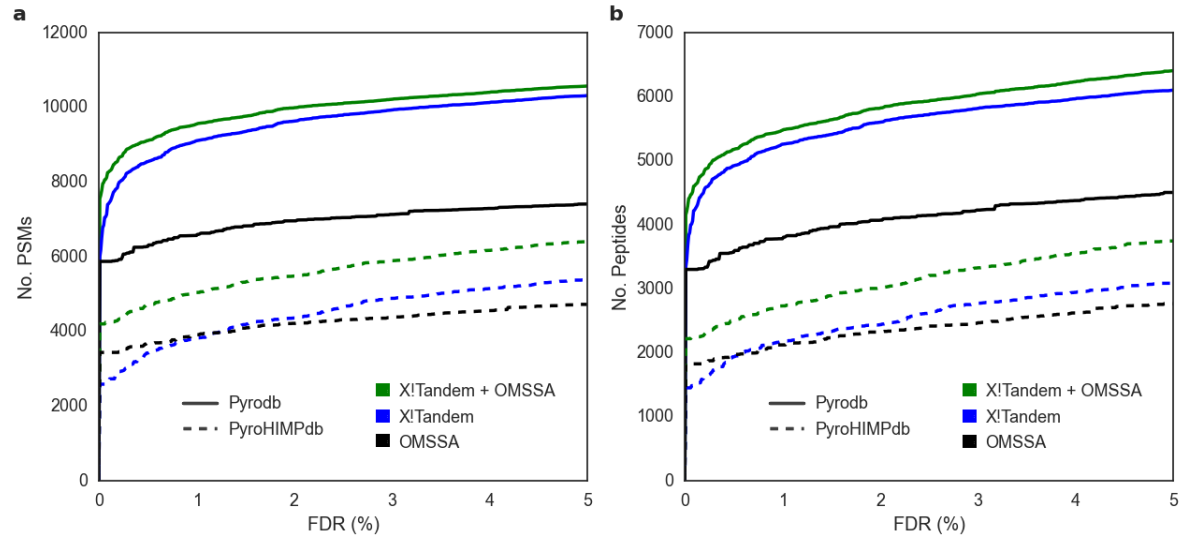

Figure 4.11: Evaluation of the identification yield for PFU searches against small (Pyrodb) and large (PyroHIMPdb) search space. The line plots show the total number of (a) PSMs and (b) peptides for database searching against Pyrodb (solid lines) and PyroHIMPdb (dashed lines) as a function of the respective FDR threshold. The single results are displayed for X!Tandem in blue and for OMSSA in black color. The identification amounts from combining X!Tandem and OMSSA searches are illustrated in green color. Figure adapted from Muth *et al.* [273].

Figure 4.12 shows that the distributions of target and decoy PSM scores for Pyrodb can be clearly distinguished in both search algorithms, the score distributions for PyroHIMPdb show

a stronger overlap. In comparison to Pyrodb, it can be further observed that the decoy score distributions for PyroHIMPdb are broader and have a larger tail to the right. Consequently, this effect results in an increased score threshold during the FDR estimation and explains the reduced number of total identifications in the PyroHIMPdb searches. Conversely, similar decoy score distributions were found for Pyrodb and PyroHIMPdb two-step searching results. Accordingly, the FDR score thresholds of these latter results are in the same range with a X!Tandem hyperscore of 21.1 and 21.8 at 5% FDR, while the cutoff values for PyroHIMPdb are increased with score values of 37.6 at 5% FDR and 43.4 at 1% FDR (Table A.9 in the appendix). Consequently, the total increase of identifications for two-step searching can be explained by a higher number of target PSMs above the FDR threshold, while no influence of the decoy hits could be observed in this case.

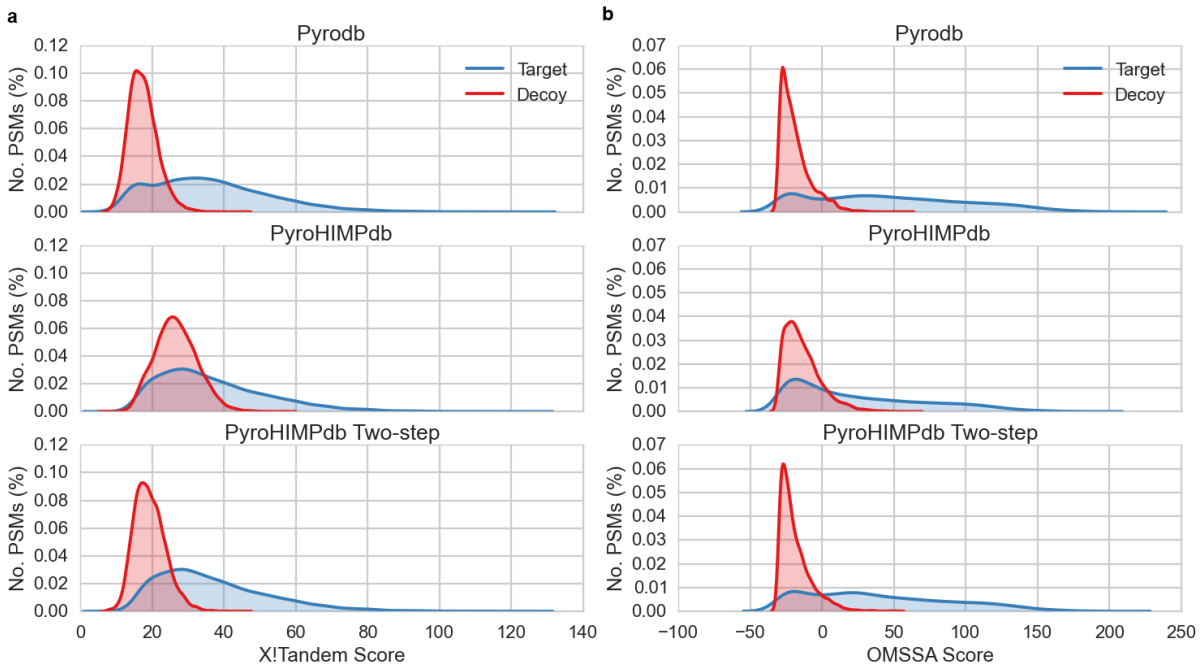

Figure 4.12: Evaluation of score distributions between classic and two-step searching for PFU data set. The line plots display the relative score distributions of target (in blue) and decoy (in red) PSMs identified by (a) X!Tandem and (b) OMSSA. The upper panel refers to classic searching against Pyrodb, the middle panel to classic searching PyroHIMPdb, and the lower panel to two-step searching PyroHIMPdb. Figure adapted from Muth *et al.* [273].

To evaluate previous findings from the metaproteomic analysis in Section 4.2.3, further benchmark investigations were performed by using varying parameter values of missed cleavages and cleavage enzyme for the PFU data set. Accordingly, parameter values between *MC = 0* and *MC = 3* were selected and *semi-tryptic* was chosen as alternative cleavage setting to *tryptic* for performing PFU database searches against Pyrodb.

Figure 4.13 displays that slightly more identifications were found for parameter values *MC = 1*, *MC = 2* and *MC = 3* in comparison to *MC = 0*. At 1% FDR, 9 054 PSMs and 5 019 peptides were obtained for MC = 0 in comparison to values of around 9 500 PSMs and 5 500 peptides for higher MC parameter values. Also at 5% FDR, an increase of around 500 PSMs and peptides could be observed when investigating the results of MC parameter values above zero. Furthermore, it can be recognized that markedly fewer PSMs and peptides were identified for *semi-tryptic* in comparison *tryptic* as chosen cleavage enzyme: for instance, only 6 707 PSMs and 3 964 peptides were identified at 1% FDR when using *semi-tryptic* cleavage.

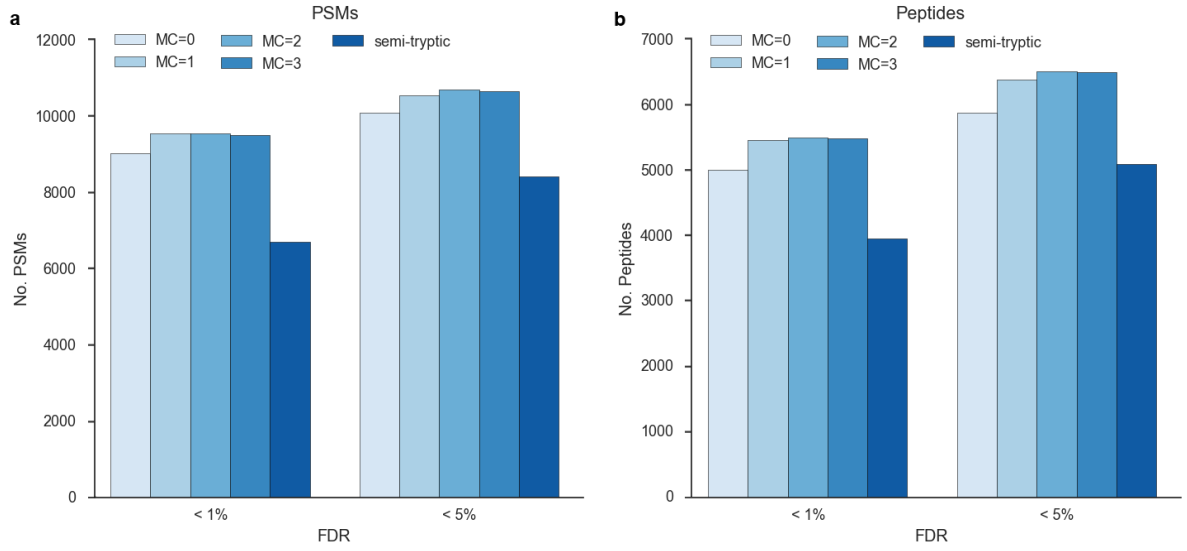

Figure 4.13: Comparative evaluation of the identification yield for different MC values (PFU). The bar plots show the total number (a) PSMs and (b) peptides for the PFU data set using missed cleavage parameter values *MC = 0 - 3* and *semi-tryptic* cleavage settings at 1% and 5% FDR.

Finally, the identification performance was assessed for FDR threshold values up to 5% by investigating the number of total and exclusive peptides found by each chosen MC parameter value for the PFU data set.

Figure 4.14a illustrates that less peptides were found for *MC = 0* in comparison to *MC = 1*, *MC = 2* and *MC = 3*. The latter three MC parameter values showed a comparable performance among each other. Figure 4.14b further shows that the number of exclusive peptides raised the most for *MC = 0* compared to higher MC values when elevating the FDR threshold.

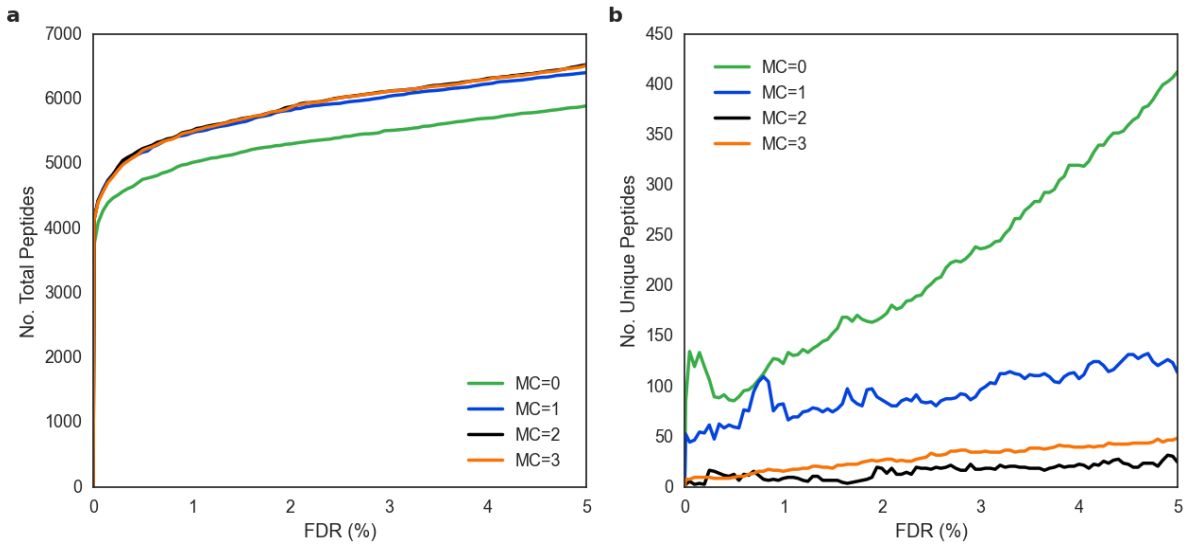

Figure 4.14: Comparison of total and exclusive peptides for different MC values (PFU). The line charts display the number of (a) total and (b) exclusive peptides for PFU searches using missed cleavage parameter values *MC = 0 - 3* as a function of the respective FDR threshold. Peptides were called exclusive when being identified uniquely for a particular MC parameter value.

4.3 *De Novo* Sequencing

The findings from previous sections indicate that search algorithm, protein database and cleavage parameters markedly affect the amount and quality of identifications in metaproteomic analysis workflows. To improve the identification yield for microbial community samples, tailored databases derived from metagenomic techniques can be engaged in addition to public databases. Alternatively, completely circumventing the use of protein databases is a conceivable option. For this purpose, the technique of *de novo* sequencing is able to infer peptide sequences directly from experimental MS/MS spectra (see Section 2.2.4). In this section, this method is applied on metaproteomic data sets and derived *de novo* sequences are compared to previous results from conventional database searching. In the following text, data of the original publication in Proteomics [273] are shown in parts.

4.3.1 Method Evaluation and Identification Recall

To test the eligibility of *de novo* sequencing for typical metaproteomic data, DeNovoGUI (see Section 3.4.4), which employs the PepNovo+ algorithm [200], was used to process the HIMP10 data sets (see Section 3.2.2). Using this software, the *de novo* sequencing algorithm was executed in parallel processes. Thereby, the same parameter values as for the database searches were used to guarantee a fair comparison of both methods. Furthermore, the *de novo* peptide sequences that could be exactly matched to the respective target were used for the following investigations. Also, the amino acids leucine and isoleucine were considered as equal.

Table 4.7 summarizes the results using *de novo* sequencing for the HIMP10 data sets: on average, 23% of the spectra could be identified at a score threshold of $S = 100$. It can be recognized that the average percentage of identified spectra raised to over 60% at $S = 50$.

Following on, various evaluation steps were applied: first, the sets of obtained *de novo* sequences were compared with the sets of identified peptides from database searching at 5% FDR (see Section 4.1.2). Thus, the peptide identifications of the database searches were used as references to assess the performance of the *de novo* sequencing method, since the actual ground truth of the identifications cannot be determined due the unknown microbial composition in the HIMP samples. However, the high percentage of identified spectra of 30% on average (see Section 4.1.2) justifies the use of the combined database search results for evaluating the performance of the *de novo* sequencing algorithm.

Table 4.7 displays that 1 689 (23%) out of 7 322 peptides from the database searches (Table 4.1) were identified using *de novo* sequencing on average at $S = 100$. Without the application of any score threshold, the complementarity between both techniques accounted for around 25%.

The second objective was to investigate whether any additional peptide hits can be gained by *de novo* sequencing that remained unidentified when using conventional database search algorithms. Therefore, the sets of *de novo* peptides were matched against an *in silico* digest of all protein sequences from HIMPdb.

It can be recognized that at *S = 100* an average of 1 557 peptides could be recalled successfully when matching the *de novo* sequences against the *in silico* digested database (Table 4.7). Remarkably, lowering the cutoff to *S = 50* resulted in a retrieval of 5 287 peptides.

Table 4.7: *De novo* sequencing results for the HIMP10 data sets (P1-P34). Peptide sequences have been matched against results from database searching (5% FDR) and an *in silico* digest of HIMPdb. Table adapted from Muth *et al.* [273].

Dataset	Spectrum IDs		Peptides (Recall)		Peptides (DB digest)	
	S = 100	S = 50	S = 100	S = None	S = 100	S = 50
P1	8 121	22 110	1 822	1 999	1 723	5 881
P3	6 196	16 617	1 628	1 791	1 345	4 960
P8	7 207	18 933	1 758	1 909	1 567	5 396
P11	6 766	18 612	1 510	1 638	1 406	5 076
P17	7 527	19 900	1 775	1 911	1 691	5 516
P23	7 952	21 543	1 814	1 948	1 666	5 480
P27	6 281	16 378	1 437	1 529	1 330	4 694
P28	7 217	19 927	1 591	1 733	1 491	5 085
P31	8 699	23 561	1 893	2 056	1 807	5 839
P34	6 677	18 360	1 665	1 802	1 539	4 940
Average	7 264 (23%)	19 614 (62%)	1 689 (23%)	1 831 (25%)	1 557	5 287

4.3.2 Comparison of Classic and Two-step Searching

In the following, the investigations were limited to the data set P1, since all ten data sets resulted in similar results from *de novo* sequencing as displayed in Table 4.1. To evaluate the reliability of the method, the *de novo* sequencing result from this sample was compared to the peptide identifications obtained from database searching in more detail. Hence, a comparable quality metric for the identifications from both techniques was required. In order to obtain such a measure, a binary classification scheme was employed a according to which *de novo* peptides were subdivided into low scoring (LS) and high scoring (HS) hits: as in the previous paragraph, a score cutoff of S = 100 was used here to classify the *de novo* sequences into these two categories. The peptide identifications originating from the database searching were divided into LS and HS hits using a score threshold of 40. Here, the score values were obtained as follows for the database

search algorithms: for X!Tandem, the hyperscore was chosen as quality metric from each PSM and for OMSSA, the probabilistic score was transformed to facilitate the comparison of the scores between X!Tandem and OMSSA (see Section 3.5.2). Although these metrics present only rough estimates, the peptides could be directly classified according to their identification quality when applying the respective thresholds.

Figure 4.15 displays the number of peptides that were found both by *de novo* sequencing and database searching for the P1 data set at 5% FDR. The results demonstrate that the majority of the peptide identifications overlapping between both techniques achieved a high score. Differences were found in the portion of overlapping peptides between classic and two-step searches: the number of identifications which received a low score from the respective search engine as well PepNovo+ was significantly increased for the two-step search approach. This was also the case for hits that received a high score from PepNovo+ and a low score from the search engine. In line with previous investigations, compared to X!Tandem (Figure 4.15a), the algorithm OMSSA (Figure 4.15b) was more stringent, since fewer low scoring identifications were obtained across the algorithms and search strategies.

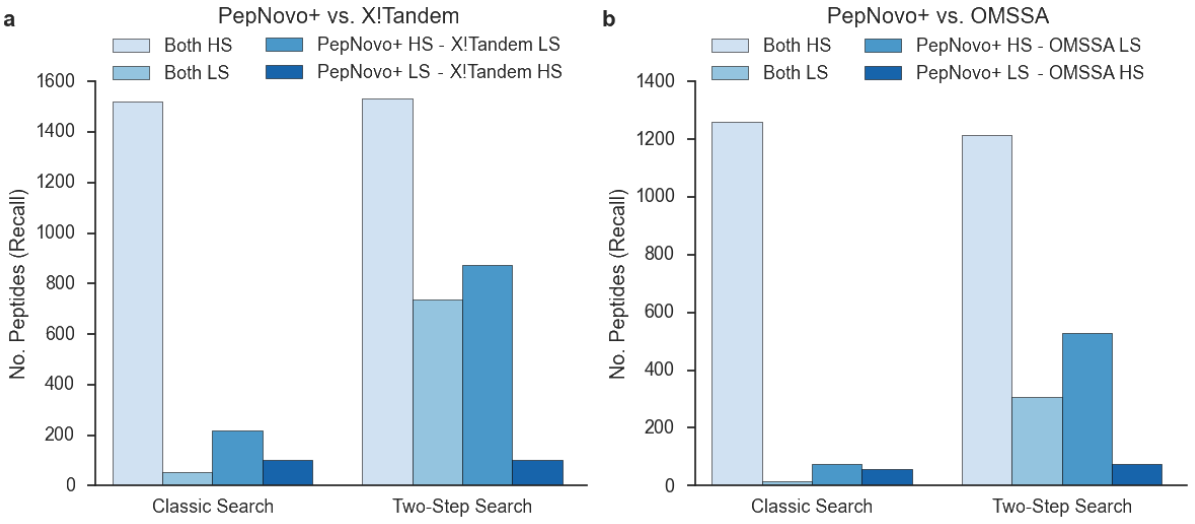

Figure 4.15: *De novo* sequencing recovery of peptides from classic and two-step searching for data set P1. The bar plots show the total amount of peptides that were identified both by *de novo* sequencing with PepNovo+ and database searching with (a) X!Tandem and (b) OMSSA. Classic and two-step searches were performed against HIMPdb and the results were filtered at 5% FDR threshold. The peptides are classified into categories of low scoring (LS) and high scoring (HS) identifications. In total, four categories are displayed which represent the combination of identification quality (LS/HS) and identification methods (PepNovo+ and X!Tandem/OMSSA). Figure adapted from Muth *et al.* [273].

As shown previously, various identifications from classic and two-step searching were recovered when using *de novo* sequencing. In the next analysis, the differences between both search strategies were evaluated with respect to the identification quality of the hits: therefore, the

PepNovo+ scores of the overlapping peptide identifications were examined between *de novo* sequencing and both database search strategies for the data set P1.

Figure 4.16 shows a trend towards lower scores for two-step searching at both applied FDR thresholds, since the respective distributions are shifted to the left in comparison to the scores corresponding to classic searching.

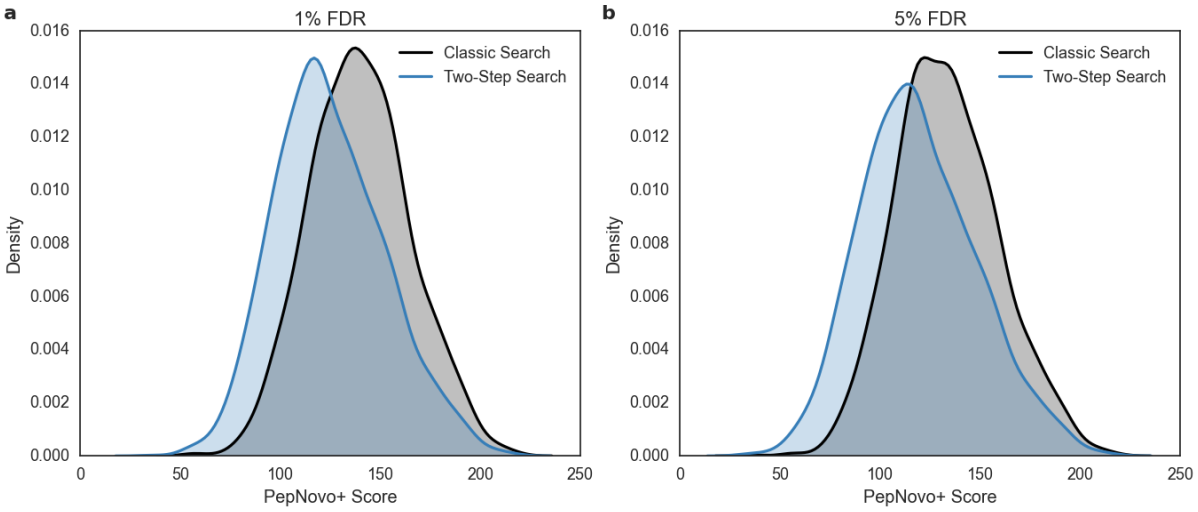

Figure 4.16: Comparison of PepNovo+ scores between classic and two-step searching for data set P1. The density distributions of PepNovo+ scores are shown for classic database searching (in black) and two-step searching (in blue). The peptide sets were taken from the intersections of hits from *de novo* sequencing and database searches at (a) 1% and (b) 5% FDR threshold. Classic and two-step searches were performed against HIMPdb and results from both search engines X!Tandem and OMSSA were combined. Figure adapted from Muth *et al.* [273].

4.4 Protein Grouping

In metaproteomics, peptide identifications are frequently linked to multiple homologous proteins that are expressed by organisms from different taxonomic origin: consequently, the results can be highly redundant at the protein level. However, instead of selecting solely one or a subset of proteins, an explorative analysis aims to maintain or even enhance the information given at this stage. One of the contributions in this work presents the method of protein grouping in the data analysis workflow of the MPA software (see Section 3.1.2). In this section, different rule-based strategies are evaluated for the generation of so-called meta-proteins which represent groups of proteins.

4.4.1 Testing Meta-Protein Generation Rules

To evaluate the rules developed for the meta-protein generation, the data sets from two BGP replicate samples EBENDORF01 and EBENDORF02 were used (see Section 3.2.1). The MS/MS data were processed using the MPA software by combining database searches with X!Tandem and OMSSA against SwissProt. In the unprocessed result sets, EBENDORF01 contained 1 324 (5% FDR) and 1 071 (1% FDR) proteins, while EBENDORF02 resulted in 1 138 (5% FDR) and 942 (1% FDR) protein identifications. In this analysis, different protein grouping rules were applied on both result sets using the meta-protein generation function.

Figure 4.17 shows that the highest protein result set reduction could be achieved using the *Minimum One Shared* rule , which merges proteins sharing at least one peptide (see Section 3.1.2). The second highest reduction was obtained by the *All Shared* rule, which combines proteins on the condition that they have all peptides or a subset in common (see Section 3.1.2). Since the latter presents a more stringent rule than *Minimum One Shared*, it could be expected that it would combine fewer protein identifications. In addition, around 5% more reduction can be observed on average for both rules at 1% FDR (Figure 4.17b) when compared to 5% FDR (Figure 4.17a). For the UniRef-based meta-protein generation rule (see Section 3.1.2), the protein grouping effect was stronger when lowering the sequence similarity threshold between the protein clusters [263]: while UniRef100 showed the lowest reduction of proteins (3.4–5.1%), UniRef50 displayed the highest reduction rates (34.6–38.0%).

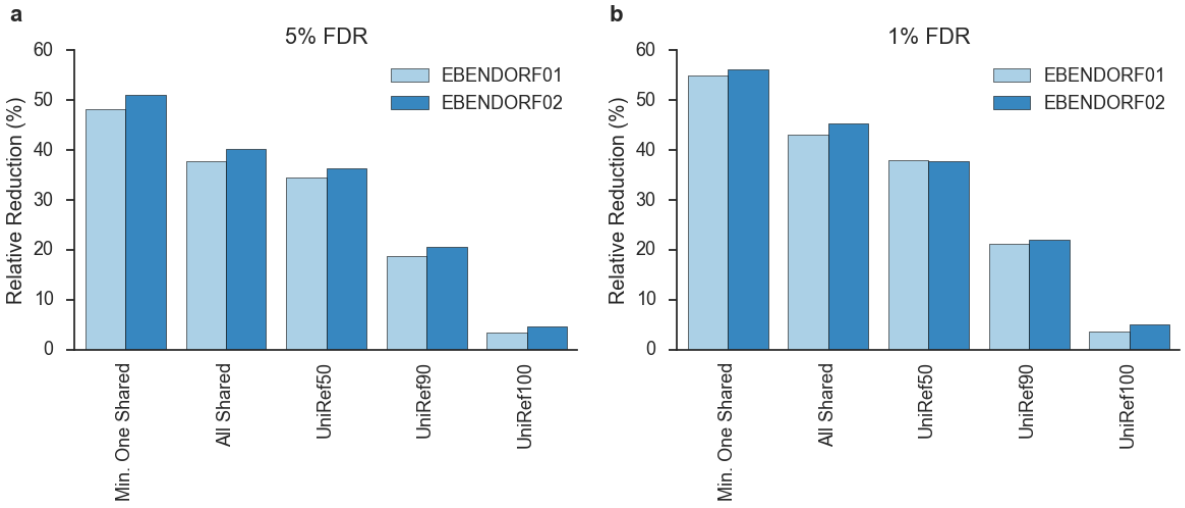

Figure 4.17: Protein result set reduction achieved by applying meta-protein generation rules. The percentage reduction of protein amounts in the results sets is displayed for EBENDORF01 and EBENDORF02 when using grouping rules *Minimum One Shared*, *All Shared*, *UniRef50*, *UniRef90*, *UniRef100* at (a) 5% and (b) 1% FDR. The proportions shown were calculated relative to the numbers of proteins from the unprocessed result sets.

Mutation-Tolerant Grouping. While the meta-protein generation is based on exact sequence string matching, the biological reality of frequently occurring sequence mutations is disregarded. Since proteins often fulfill the same functions after changing their sequence, a mutation-tolerant strategy is useful for the grouping process. Therefore, the Levenshtein edit distance (ED) was implemented as parameter representing the maximum allowed sequence transformations at the peptide level (see Section 3.1.2). Since the application of this method leads to more grouped sequences, a decrease of meta-proteins was expected depending on the chosen ED parameter value. The meta-protein generation was evaluated for EBENDORF01 using parameter values between $ED = 0$ and $ED = 4$. The latter was chosen as maximum value, since the length of tryptic peptide sequences typically ranges from 8 to 20 amino acids in bottom-up proteomics. Thus, the ED parameter was limited to permit at most 50% sequence variation. Once again, the grouping rules *Minimum One Shared* and *All Shared* were applied. To evaluate the effect of the ED parameter, the relative reduction of the meta-protein result set was calculated as ratio between the amount of meta-proteins for $ED > 0$ and $ED = 0$.

Figure 4.18 illustrates that the number of meta-proteins decreased with elevated ED parameter values. It can be recognized that for $ED = 2$, the reduction effect was less than 10% in both grouping rules. The bar plots also display that the ED parameter affected the protein grouping more at 5% FDR than at 1% FDR. Finally, the effect of the ED parameter was stronger for *Minimum One Shared* than for *All Shared*.

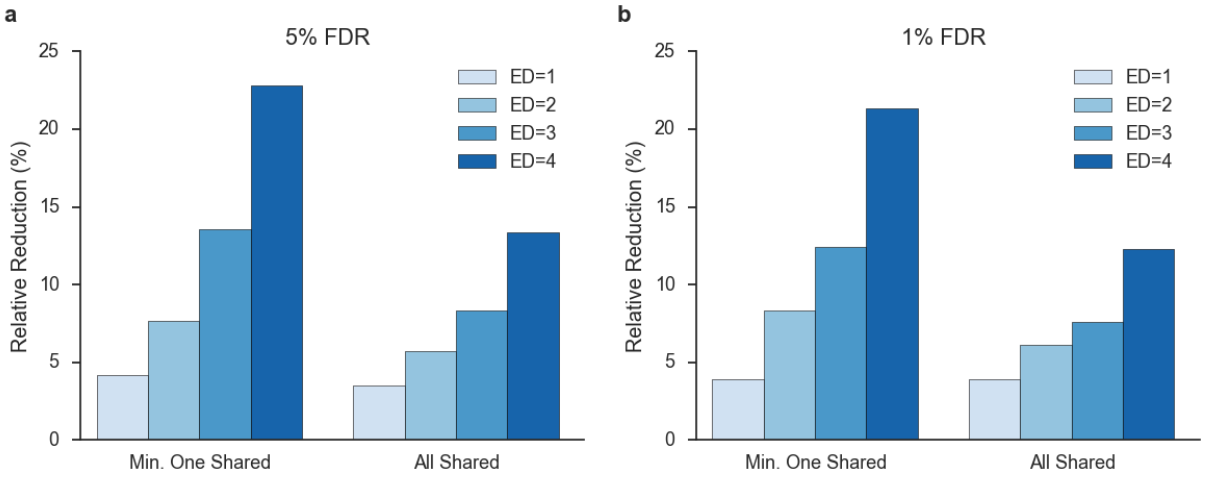

Figure 4.18: Meta-protein result set reduction achieved by mutation-tolerance grouping. The percentage reduction of the number of meta-proteins in the result set is displayed for EBENDORF01 achieved by applying the grouping rules *Minimum One Shared* and *All Shared* in combination with ED parameter values ranging from *ED = 1* to *ED = 4* at (a) 5% and (b) 1% FDR. The proportions shown were calculated relative to the number of meta-proteins for *ED = 0*.

Phylogenetic Diversity Control. The final evaluated parameter for the meta-protein generation presents the *Taxonomy Rule* (see Section 3.1.2). The aim of this method is to control the phylogenetic diversity by specifying a threshold for the maximum taxonomic convergence level at which proteins are grouped together. In comparison to the grouping used without a taxonomic threshold, an increase in meta-proteins was therefore expected depending on the chosen convergence level. The EBENDORF01 data was evaluated by using superkingdom, phylum, genus and species as taxonomic convergence levels. Note that the *Taxonomy Rule* cannot be used alone, since it presents rather a taxonomic filtering than a grouping method. Thus, this method was applied in combination with the grouping rules *Minimum One Shared* and *All Shared*. The relative increase of the meta-protein result set was calculated as ratio between the number of meta-proteins before and after applying taxonomic cutoffs during the grouping.

Figure 4.19 displays that the number of meta-proteins increased when lowering the taxonomic convergence level at both evaluated FDR thresholds. The strongest increase can be observed between the higher hierarchies superkingdom and phylum, while the lower taxonomic convergence levels genus and species less affected the outcome. The bar plots further illustrate a stronger effect for *Minimum One Shared* in comparison to *All Shared*. While a high diversity could be preserved at the species level, the relative increase of protein groups did not exceed 8%.

In previous investigations, the final size of the meta-protein result set from a single sample was considered as quality measure to evaluate the performance of the developed grouping rules. Thereby, it can be observed that *Minimum One Shared* and *All Shared* can be employed as methods to reduce the protein result set. However, monitoring the change in the absolute number

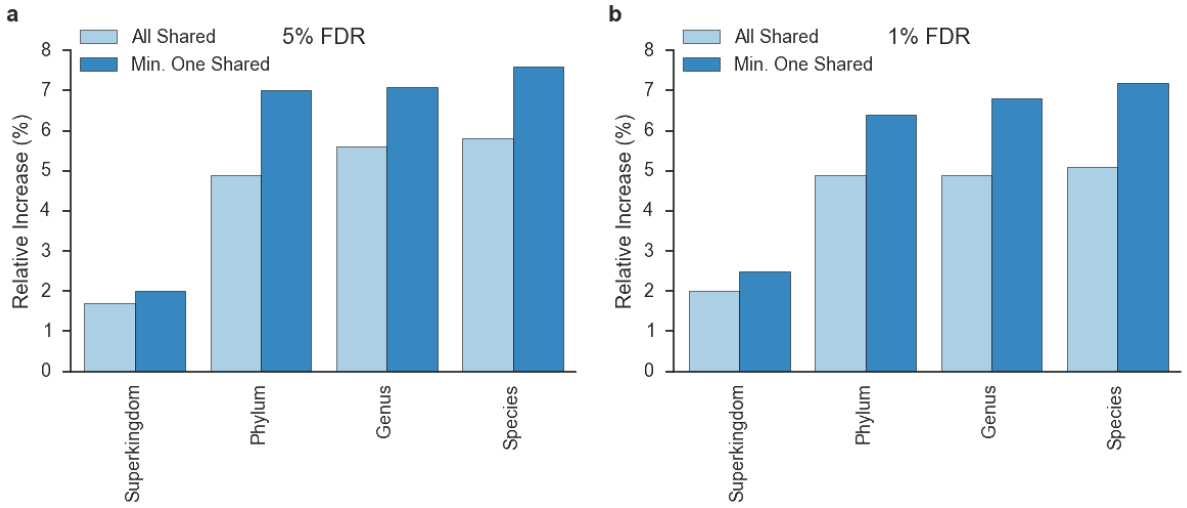

Figure 4.19: Meta-protein result set increase resulting from phylogenetic diversity control. The percentage increase of the number of meta-proteins in the result set is displayed for EBENDORF01 when using the grouping rules *Minimum One Shared* and *All Shared* and setting the maximum taxonomic convergence levels by the *Taxonomy Rule* to superkingdom, phylum, genus and species at (a) 5% and (b) 1% FDR. The proportions shown were calculated relative to the number of meta-proteins when no threshold for the taxonomic convergence level was applied.

of protein groups within one data set only provides a limited view on the accuracy of the meta-protein generation. In a typical metaproteomic experiment, samples from different origin or time point are compared against each other at the protein level. Therefore, it is also required to guarantee the consistency of the protein grouping across different data sets. Hence, the effects of the grouping rules on the comparability of results from replicate data were examined in the following paragraph.

4.4.2 Evaluating Reproducibility between Replicates

The next objective was to assess the reproducibility of the developed protein grouping methods. Therefore, the GENT01 and GENT16 samples were analyzed which were repeatedly measured resulting in two technical replicate data sets (see Section 3.2.1). These data sets were searched against SwissProt by using X!Tandem and OMSSA and the results were combined as in previous investigations. Subsequently, the peptide identifications between each of the respective data sets were compared based on the spectral count, thus, the number of identified spectra for each peptide. The purpose of this preliminary analysis was to determine whether the replicates form a comparable data basis. In this and each of the following investigations, for each pair of technical replicates, scatter plots were produced and the corresponding Pearson's correlation coefficients were computed.

Figure 4.20 shows a high correlation for the peptide identifications between the replicates of

GENT01 searched against SwissProt at 5% and 1% FDR. In addition, the GENT16 results showed also strong linear relationships between the replicates at both FDR thresholds (Figure A.4 in the appendix). Since these findings suggested a high reproducibility of the identifications between the replicates, it can be concluded that respective data sets would fit for evaluating the performance of the protein grouping methods.

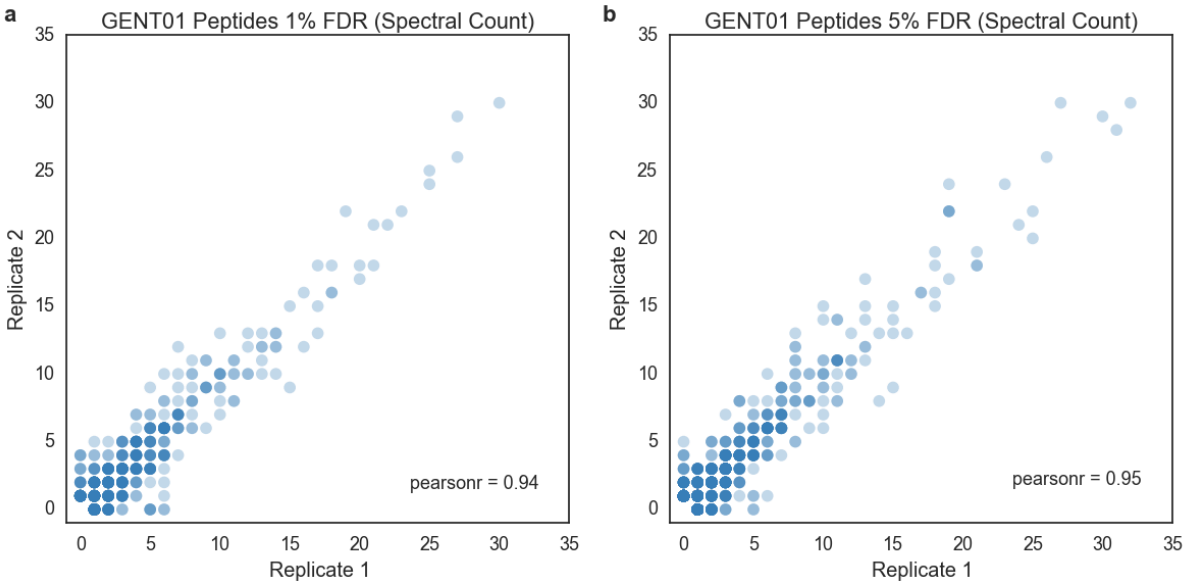

Figure 4.20: Reproducibility of peptide hits between technical replicates for GENT01. The plots compare the peptides that were reproducibly identified between GENT01 replicates on the basis of the spectral count at (a) 1% and (b) 5% FDR. The color scale represents the number of identified peptides; low amounts in bright blue and high amounts in dark blue, respectively. The Pearson correlation coefficient (pearsonr) is displayed in the lower right corner of each panel.

In the next analysis, the focus was shifted from the peptide to the protein level: while ascending the bottom-up hierarchy less consensus was expected between the replicate results due to protein inference issues (see Section 2.2.5). To estimate the reproducibility between the replicates at the protein level, their protein identifications were correlated on the basis of the number of assigned spectra and peptides for each protein.

In comparison to the initial analysis at the peptide level, weaker correlations of proteins were found between the replicates based on the peptide counts for GENT01 (Figure 4.21a) and GENT16 (Figure A.5a in the appendix). On the contrary, strong correlations of proteins can be recognized between the replicates on the basis of spectral counts for GENT01 (Figure 4.21b) and GENT16 (Figure A.5b in the appendix). These latter findings show that the majority of the identified spectra was assigned to proteins shared between the replicates.

Following on, a meta-protein generation was performed using the rules *Minimum One Shared* and *All Shared*. The goal was to examine the effect of the protein grouping rules on the compa-

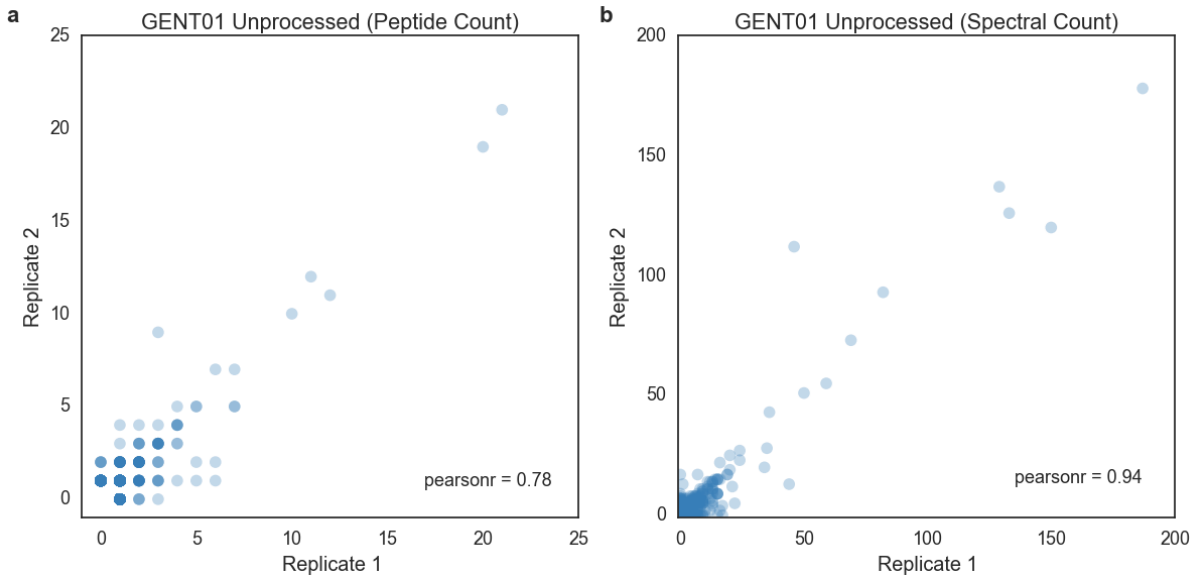

Figure 4.21: Reproducibility of protein hits between technical replicates for GENT01. The plots compare the proteins that were reproducibly identified between GENT01 replicates on the basis of their (a) peptide and (b) spectral count at 5% FDR. The color scale represents the number of identified proteins; low amounts in bright blue and high amounts in dark blue, respectively. The Pearson correlation coefficient (pearsonr) is displayed in the lower right corner of each panel.

rability of the replicate result sets. To estimate the reproducibility between the replicates at the meta-protein level, the generated protein groups were correlated on the basis of the number of assigned spectra and peptides for each meta-protein.

When applying *Minimum One Shared* for GENT01, very strong correlations for meta-proteins were found between the replicates with respect to identified spectra (pearsonr = 0.97) and peptides (pearsonr = 0.99) (Figure 4.22). In addition, comparably high correlation values could be observed for GENT16 (Figure A.6 in the appendix). In comparison to the unprocessed protein results (Figure 4.21), more peptides and spectra were shared between the respective meta-proteins for GENT01 (Figure 4.22).

Compared to the *Minimum One Shared* rule, the application of *All Shared* resulted in weaker correlations between the replicates for GENT01 (Figure 4.23). In particular, the agreement at the peptide level was low (pearsonr = 0.46), since various identifications could not be assigned to a common protein group (Figure 4.23a). In contrast, GENT16 data displayed strong correlations of meta-proteins for *All Shared* (Figure A.7 in the appendix).

To further illustrate the diverging performance of the rules, one exemplary case for GENT01 was inspected in detail. Therefore, a meta-protein called Acetyl-CoA decarbonylase/synthase complex subunit gamma was selected: for *Minimum One Shared*, it contained 20 peptides and 108 identified spectra in the first replicate, while 25 peptides and 128 spectra were assigned in the

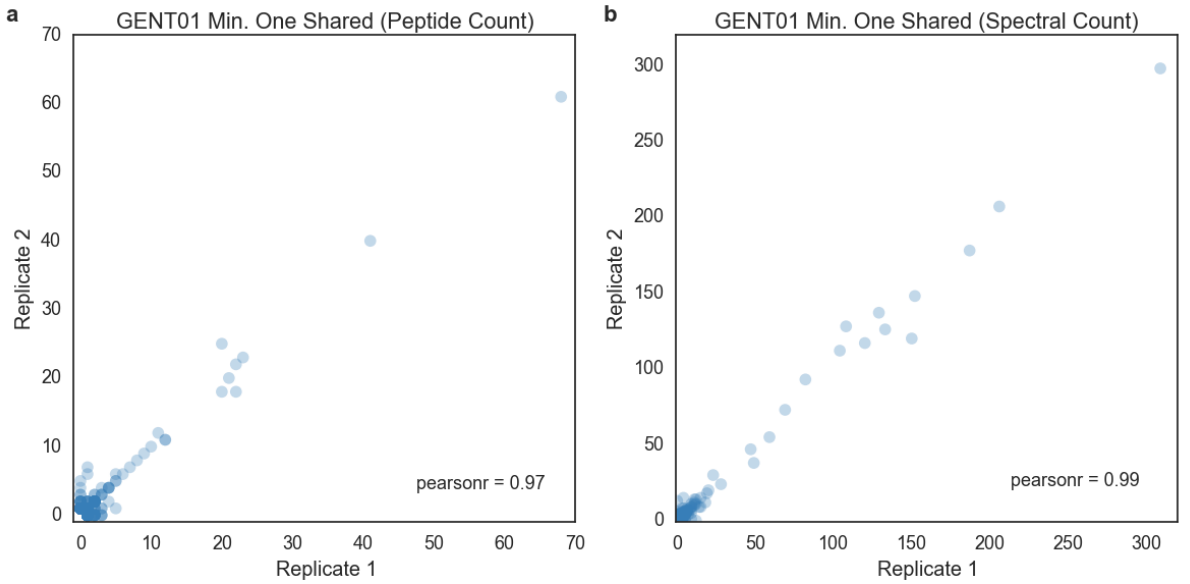

Figure 4.22: Reproducibility of meta-protein hits between technical replicates for GENT01. The plots compare the meta-proteins that were reproducibly identified between GENT01 replicates on the basis of their (a) peptide and (b) spectral count at 5% FDR. Meta-proteins were generated by using the *Minimum One Shared* rule. The color scale represents the number of identified meta-proteins; low amounts in bright blue and high amounts in dark blue, respectively. The Pearson correlation coefficient (pearsonr) is displayed in the lower right corner of each panel.

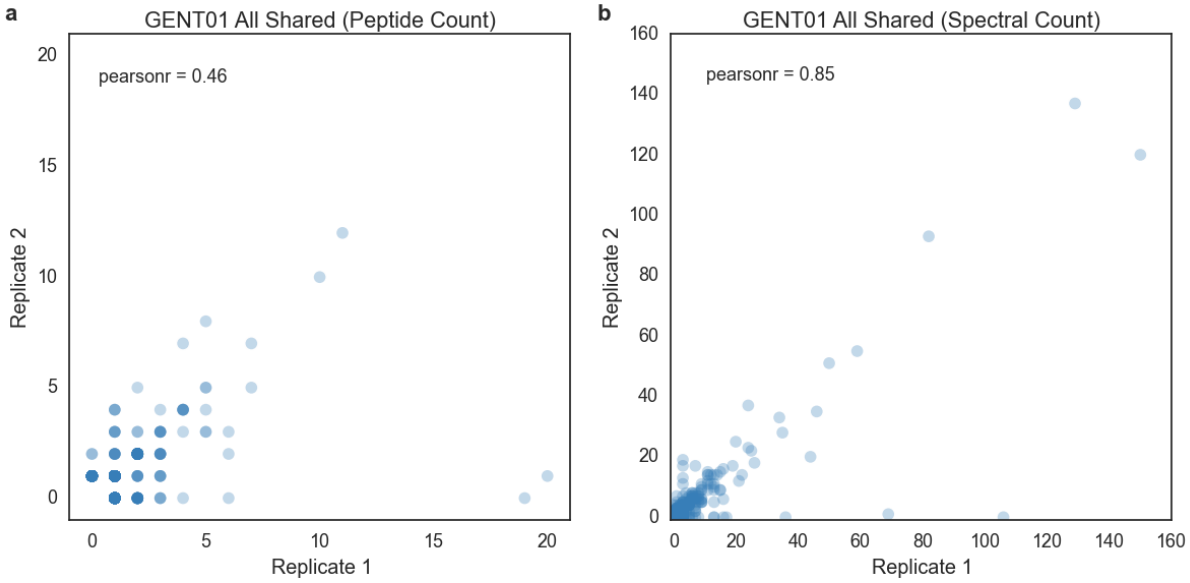

Figure 4.23: Reproducibility of meta-protein hits between technical replicates for GENT01. The plots compare the meta-proteins that were reproducibly identified between GENT01 replicates on the basis of their (a) peptide and (b) spectral count at 5% FDR. Meta-proteins were generated by using the *All Shared* rule. The color scale represents the number of identified meta-proteins; low amounts in bright blue and high amounts in dark blue, respectively. The Pearson correlation coefficient (pearsonr) is displayed in the lower right corner of each panel.

second replicate. Hence, this protein group exhibited a similar identification profile in both data sets. Conversely, for *All Shared*, the same protein group was found in four different forms with non-correlated amounts of identifications between the replicates: for instance, Acetyl-CoA decarbonylase/synthase complex subunit gamma 2 was identified with 19 peptides and 109 spectra the first replicate, while no identifications were assigned to this group in the second replicate. As a consequence, the demonstrated issues and low correlation values for GENT01 depreciate *All Shared* as protein grouping method as soon as results from multiple data sets are compared against each other.

In summary, these findings show that *Minimum One Shared* represents the most reproducible grouping rule when evaluating replicate data sets. However, to draw conclusions on the general performance of the protein grouping for a metaproteomic experiment, further investigations on multiple data sets from different samples are required. Consequently, the meta-protein generation procedure is regarded using HIMP samples from different individuals in the following.

4.4.3 Comparing Data Sets from Different Samples

A metaproteomic experiment typically involves various samples obtained from different subjects and varying time points. In particular, qualitative and quantitative protein profiles are investigated for each sample and—if required—changes across different measurements are recorded. In order to compare data sets from different samples, it is essential to detect protein identifications that are shared between respective result sets. However, reaching such a common basis for comparison is hampered by protein inference issues, since peptides are assigned inconsistently to proteins across different samples. For metaproteomic experiments, an aggravating circumstance concerns the presence of many species which leads to high amounts of candidate proteins (see Section 2.2.5). To tackle these challenges, the meta-protein generation procedure was developed as a peptide-centric protein grouping approach (see Section 3.1.2). In the following, the results from HIMP data sets are compared at the protein and meta-protein level to evaluate the applicability of the grouping methods for data sets from different sample origin. Eventually, the efficacy of the grouping rules is briefly illustrated by evaluating the numbers of peptides contained in the meta-protein results.

Comparability of result sets from different samples. Next, the identifications from each of the HIMP10 data sets were taken to evaluate the comparability of different results at the protein and meta-protein level. In this analysis, the Jaccard index was calculated as a similarity measure for each pair of result sets based on the shared peptide sequence information for proteins and meta-proteins (see Section 3.5.5). In addition, a threshold was used that filters the protein and meta-protein groups by their minimum amount of assigned peptides: it was assumed that the similarity between groups of different result sets could be related to the number of peptides per group. Hence, supposing the meta-protein generation method works correctly, the probability of obtaining groups that share the same peptides between the data sets increases with the amount of hits contained in each group.

At the meta-protein level, the Jaccard similarity coefficient between the result sets increased with the number of peptides per group (Figure 4.24a). Conversely, the average agreement decreased when more peptides were assigned to a group at the protein level. Figure 4.24b illustrates the average number of proteins and meta-proteins for each pair of result sets. In this graph, a steep decline in the average number of proteins and meta-proteins can be observed for groups with more than one assigned peptide. It can also be recognized that this event was more pronounced for the protein than for the meta-protein level. The average number of groups for both hierarchies was lower at 1% FDR when compared with the results at 5% FDR. Finally, the number of proteins and meta-proteins approached each other with increasing amount of peptides per group.

Evaluation of peptide frequencies for meta-protein generation rules. According to the findings on single (see Section 4.4.1) and replicate (see Section 4.4.2) result sets, *Minimum One Shared* determines the most reliable meta-protein generation rule. The goal of the final investigation was to examine the peptide frequencies of meta-proteins when applying *Minimum One Shared* and *All Shared* for multiple results from different data sets. In this analysis, the HIMP data sets P1, P23 and P34 were used to determine the number of peptides assigned to each of the meta-proteins.

Across the results for aforementioned data sets, *Minimum One Shared* performed better than *All Shared*: while the latter resulted in higher proportions of meta-proteins yielding one peptide, increased amounts of meta-proteins with a peptide frequency of more than one were obtained for *Minimum One Shared* (Figure 4.25). The relative amount of meta-proteins with more than five peptides was minimal for *All Shared*. It can be also recognized that the proportion of meta-proteins with one assigned peptides was lower at 1% FDR for both rules (Figure 4.25a) when compared to 5% FDR (Figure 4.25b). Vice versa, in comparison to 1% FDR, fewer meta-proteins with frequencies of two or more peptides were observed at 5% FDR.

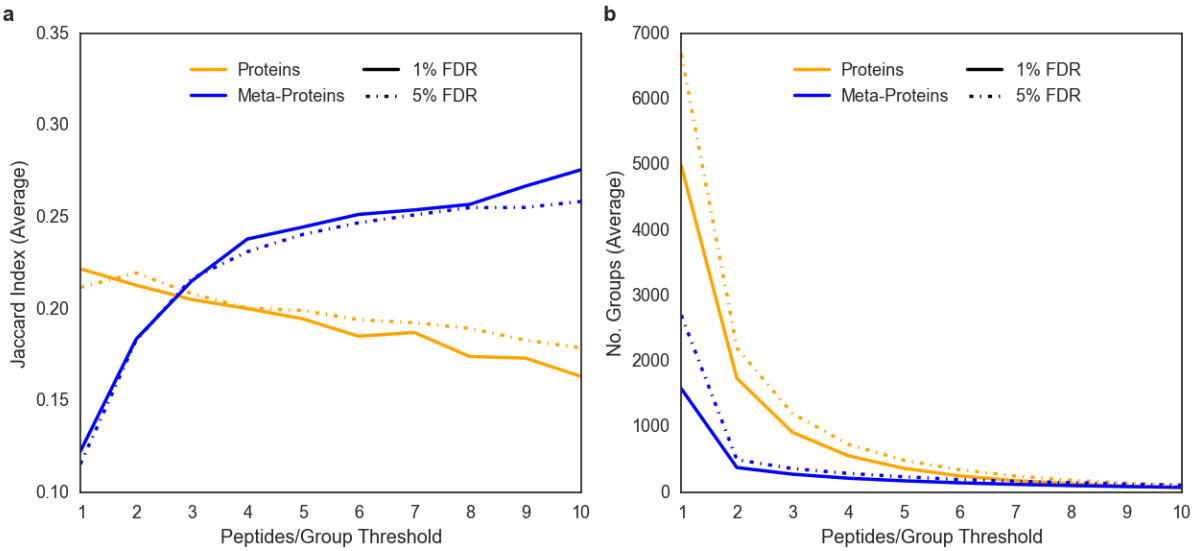

Figure 4.24: Evaluation of group similarity and size in dependence of assigned peptides. The line plots show (a) average Jaccard index and (b) average group size as a function of the number of assigned peptides per group. The identifications were obtained for HIMP10 data sets (P1-P34) and the groups were classified based on their shared peptides at the protein (orange) and at the meta-protein (blue) level. Jaccard similarity coefficients and group sizes were retrieved for each pair of results from HIMP10. The result sets were filtered at 1% (solid lines) and 5% FDR (dash-dot lines) threshold. The grouping rule *Minimum One Shared* was used for the meta-protein generation.

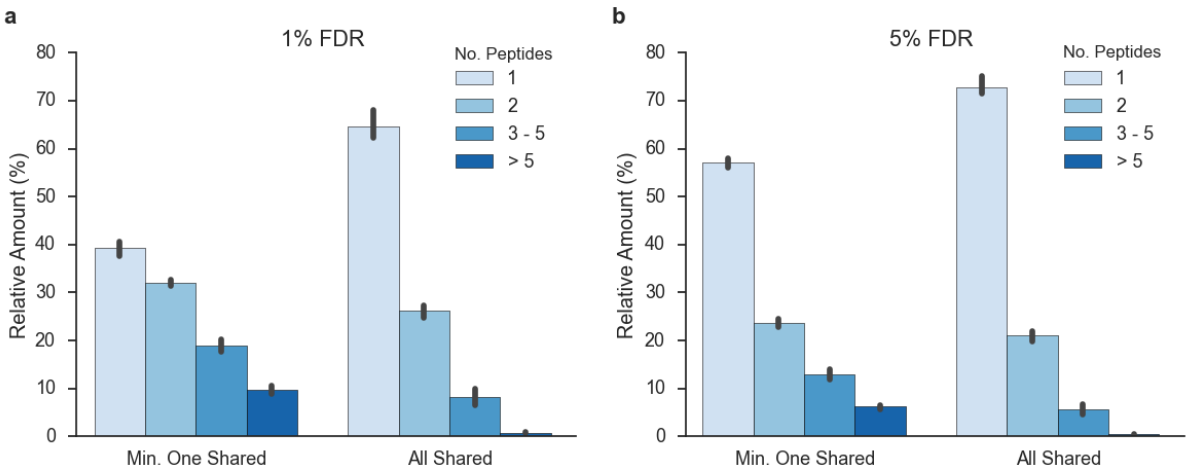

Figure 4.25: Comparative evaluation of peptide frequencies for meta-protein generation rules. The bar plots show the relative average amount of the number of peptides per meta-protein from the HIMP result sets P1, P23 and P34. The grouping rules *Minimum One Shared* and *All Shared* were used for the generation of meta-proteins at (a) 1% and (b) 5 % FDR.

4.5 Taxonomic Assignment

The previous sections have primarily dealt with various aspects of data processing related to the increase and validation of identified spectra, peptides and proteins. These investigations were performed with the aim to find an optimized data analysis workflow for metaproteomics ensuring the output of confident results for further postprocessing steps. Next, the effectiveness of the developed methods is evaluated for the taxonomic assignment of metaproteomic data. For this purpose, the MPA software is used to assign identifications to respective taxonomic ranks. First, the BGP data sets are used to evaluate the influence of the protein database on the taxonomic assignment process. Second, the developed methods are applied to data sets from a sample of known microbial composition. Based on these ground truth data, the reliability of the taxonomic assignment is subsequently examined in detail. Finally, as a proof of principle, a phylogenetic classification is performed based on result sets from the HIMP samples.

4.5.1 Influence of Protein Database

The results in Section 4.2 showed that the protein database plays an important role with respect to quality and yield of identifications in metaproteomic workflows. In principle, it could be observed that the two main influencing factors are the composition and the size of the database. As a consequence, both of these parameters were assumed to also affect the taxonomic assignment process. In a next step, the taxonomic origin of the identified peptides was therefore investigated for the BGP data sets GENT01, GENT07 and GENT16 (Section 3.2.1) which were searched against two databases varying in their composition and size: while SwissProt presents a manually curated and relatively condensed database, TrEMBL contains a large number of computationally annotated sequence entries (see Section 3.3.1). In this analysis, the peptides were regarded which could be uniquely assigned at the highest taxonomic rank of superkingdom.

Figure 4.26 displays that GENT16 results were significantly different with respect to the proportion of peptides assigned to Archaea between SwissProt (37%) and TrEMBL (62%). Hence, it can be also recognized that the fractions of assignments to Bacteria and Eukaryota varied between the database search results. While GENT16 was exceptional with respect to the high identification yield for TrEMBL, it can be observed that the taxonomic assignment process was also affected by the chosen database for GENT01 (Figure A.8 in the appendix) and GENT07 (Figure A.9 in the appendix).

In order to increase the phylogenetic resolution of the previous investigation, the unique peptide assignments were determined at the deeper taxonomic rank of phylum. In this case, peptides originating from eukaryotic taxa were filtered out to facilitate a systematic view on microbial

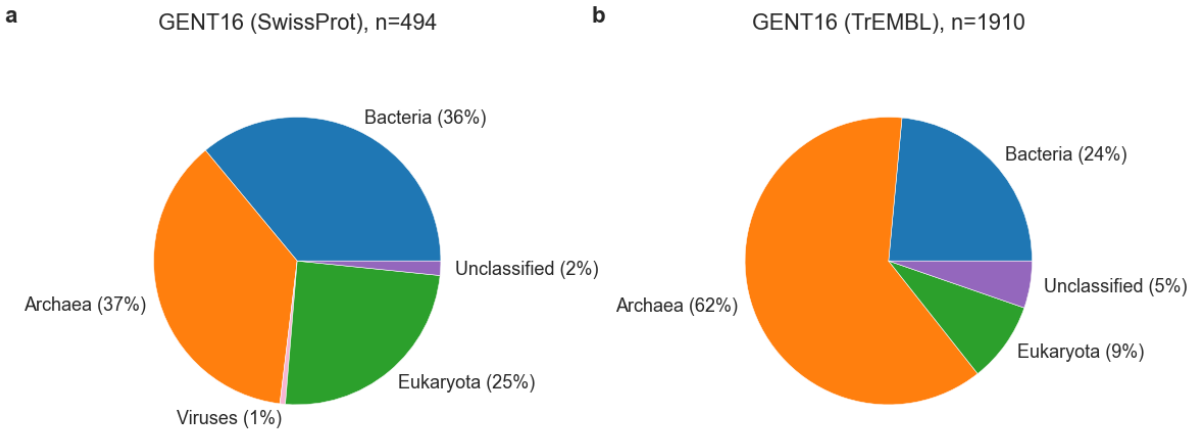

Figure 4.26: Phylogenetic classification of BGP data set GENT16 based on number of peptides per superkingdom. The pie charts display the relative distribution of total peptide hits retrieved from (a) SwissProt and (b) TrEMBL searches. The total number of assigned peptides is provided above each chart panel (n).

species.

It was found that each BGP result set provided a unique taxonomic profile in which Euryarchaeota, Proteobacteria and Firmicutes were the most abundant phyla (Figure 4.27). It can be observed that GENT07 resulted in more relative assignments to Actinobacteria in comparison to GENT01 and GENT16. For GENT16, decreased proportions of Firmicutes and increased levels of Euryarchaeota were found compared to the other BGP result sets. Again, the distribution of the peptides among the phyla differed significantly between the used protein databases: in line with the findings at the superkingdom level, the number of Euryarchaeota-specific peptides in GENT16 was particularly elevated for TrEMBL (73%) in comparison to SwissProt (52%) and other result sets. Due to the generally increased identification yield for GENT16 (see Section 4.2.1), the highest number of identifications (1 577 peptides) could be assigned to the phyla for TrEMBL.

4.5.2 Assignment Performance Evaluation

The aim of the next investigation was to evaluate the performance of the taxonomic assignment of the MPA software: therefore, MPA was compared against Unipept (see Section 3.4.5) with respect to the number of taxon-specific peptide identifications. In general, each of the tools follows a different assignment strategy: in Unipept, a central database provides information on each peptide that can be linked uniquely to a certain taxon. Conversely, the MPA relies on the peptide to protein relations, since the provided protein accessions are used to resolve the taxonomic origin. In this case, the NCBI taxonomy served as the reference database to retrieve the taxonomic lineage for each of the reported identifications in the MPA software. In these analyses, the same data

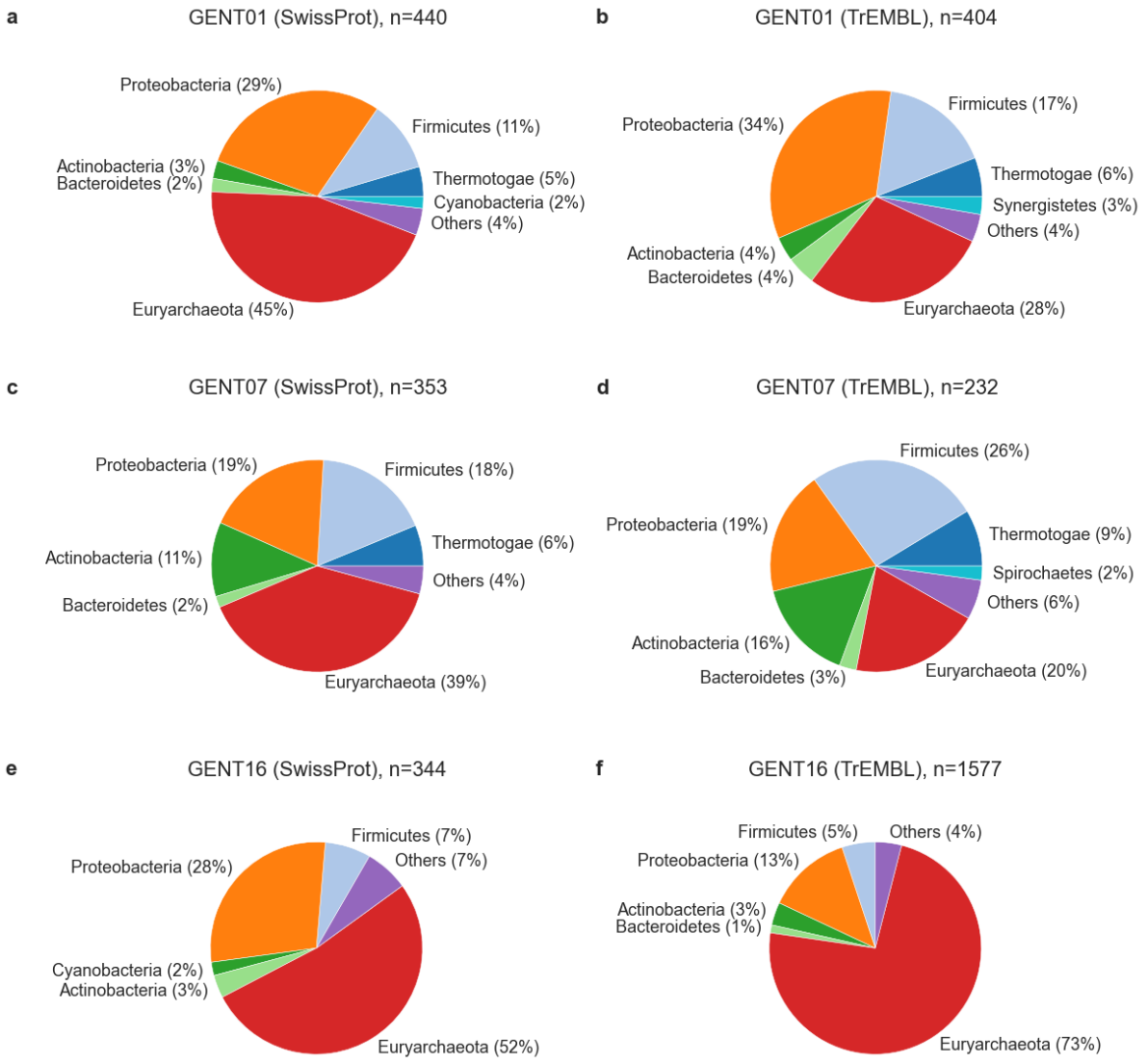

Figure 4.27: Phylogenetic classification of BGP data sets based on the number of peptides per phylum. The major phyla per data set and search protein database are presented in the pie charts that display the distribution of the peptide hits: (a) GENT01 against SwissProt, (b) GENT01 against TrEMBL, (c) GENT07 against SwissProt, (d) GENT07 against SwissProt, (e) GENT16 against SwissProt and (f) GENT16 against TrEMBL. All eukaryotic phyla were filtered out. The total number of assigned peptides is provided above each chart panel (n).

were used that resulted from searching GENT01, GENT07 and GENT16 against SwissProt and TrEMBL as described in the previous section. In addition, the distinct peptide sequences from these results were subjected to Unipept. Consequently, unique taxonomic assignments could be retrieved for both software tools.

Figure 4.28 provides an overview of peptide assignments to different taxonomic ranks for MPA and Unipept. The relative proportions of successfully assigned peptides are displayed in

the heatmaps. In comparison to Unipept, MPA achieved higher fractions of taxon-specific peptides across all data sets and taxonomic ranks for the results against SwissProt. For the TrEMBL searches, Unipept was in turn able to assign slightly more peptides for superkingdom. For each of the other taxonomic ranks, MPA outnumbered Unipept with few exceptions in case of data set GENT16. On average, Unipept could not assign 22% of the peptides to any taxonomic rank for the data sets searched against SwissProt.

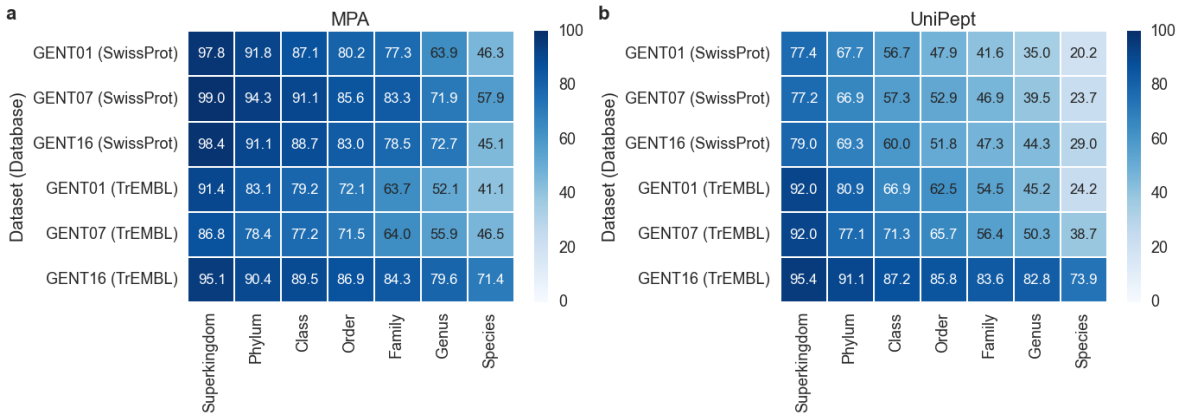

Figure 4.28: Taxonomic assignment performance of MPA and Unipept for BGP data sets. The heat maps show the relative percentage of peptides that could be assigned to a specific taxonomic rank for (a) MPA and (b) Unipept. The respective peptides were obtained from searching the data sets GENT01, GENT07 and GENT16 against SwissProt and TrEMBL (FDR < 5%). The white-blue color scale depicts the relative percentage of assigned peptides (white: low, blue: high).

Ground Truth Data of a Microbial Mixture. The previous investigation on the BGP data suffered from the shortcoming that it was not possible to evaluate the accuracy of the taxonomic assignment process due to the unknown microbial composition of the samples. To overcome this limitation, in the following analysis, metaproteomic data was used which originated from a lab-assembled microbial mixture containing nine bacterial and eukaryotic species (see Section 3.2.4) as published in a study by Tanca *et al.* [126]. Providing knowledge about the exact microbial composition, the ground truth data allowed to perform benchmark experiments evaluating the exact performance of the taxonomic assignment process in MPA and Unipept. Therefore, database searches were performed against SwissProt by using the microbial mixture data sets 9MM_FASP and 9MM_PPID originating from two different sample preparation steps. These samples were regarded as technical replicates, since the differences of the experimental setup were out of scope in this work. The MS/MS data sets were searched independently and thereafter the identifications from each result were merged. From these results, the taxon-specific peptides for the taxonomic ranks family, genus and species were exported at 1% and 5% FDR. As proposed in the original study, the peptides were classified as correct and incorrect taxonomic assignments according to

the information of the nine organisms contained in the sample [126]. In this step, the strategy of the authors was evaluated by which a filter was used that determines whether a set of peptides contributes significantly to a certain taxon: the proportion of these peptides to the total amount of taxon-specific identifications needs to be higher than a specified threshold. The authors recommended a value of 0.5% for this taxon significance threshold when testing the reliability of taxonomic assignment using Unipept [276] and MEGAN [207]. Eventually, different threshold values were applied to determine a robust default parameter for the MPA software.

It was found that the number of correct peptide assignments in the MPA results increased when elevating the taxonomic rank from species over genus to family (Figure 4.29). Furthermore, it can also be recognized that at a taxon significance threshold of above 3%, the amount of incorrect taxon-specific assignments is reduced to a minimum for the taxonomic ranks under investigation. Furthermore, it can be observed that an increase from 1% to 5% FDR resulted in a higher number of correct and incorrect assignments.

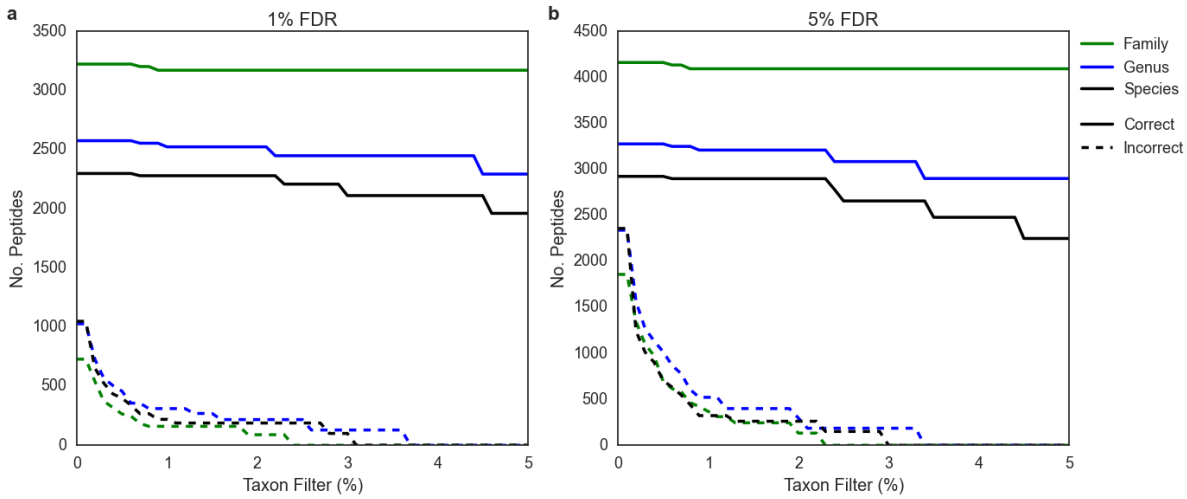

Figure 4.29: Taxonomic assignment performance of MPA for 9MM data set. The line plots show the number of correct (solid) and incorrect (dashed) taxon-specific peptide assignments of MPA for species (black), genus (blue) and family (green) as function of the taxon significance threshold at (a) 1% and (b) 5% FDR. The peptides were assigned according to the LCA approach.

Figure 4.30 shows that significantly fewer peptides could be assigned when using Unipept in comparison to previous results from MPA. In particular, the number of correct assignments at the species level was reduced. Remarkably, the application of a taxon significance threshold of up to 5% did not weed out the wrong assignments at the species level. Examining this result in detail, it can be recognized that all incorrectly assigned peptides were attributed to the eukaryotic species *Gallus gallus* by Unipept (data not shown). For the other taxa, the application of a taxon significance threshold of 3% could exclude most of the incorrect taxonomic assignments. In

addition, applying a threshold value of 0.5% could remove the majority of incorrectly assigned peptides at 1% FDR.

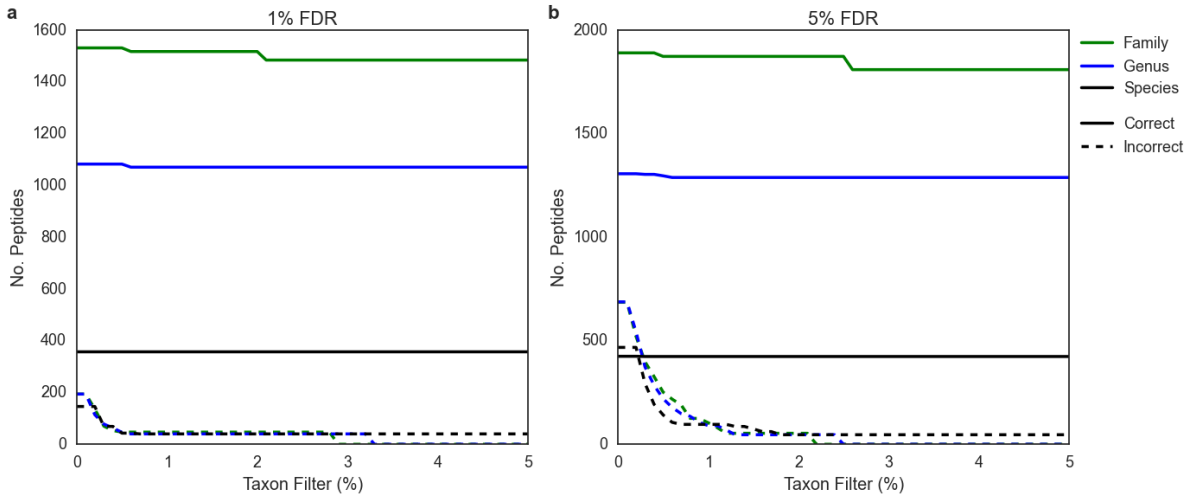

Figure 4.30: Taxonomic assignment performance of Unipept for 9MM data set. The line plots show the number of correct (solid) and incorrect (dashed) taxon-specific peptide assignments of Unipept for species (black), genus (blue) and family (green) as function of the taxon significance threshold at (a) 1% and (b) 5% FDR.

Finally, it should be noted that peptides of the species *Escherichia coli* and *Lactobacillus casei* were considered at a taxon significance threshold of 0.5% in the MPA (Table A.11 in the appendix). In contrast, Unipept did not report any assignments to these two species.

Performance Comparison of LCA and MST. Besides the conventional LCA approach, the alternative MST method was developed to preserve the specificity of the peptides at the phylogenetic level. Since both methods were implemented in the *Taxonomy Definition* process of the MPA software (see Section 3.1.2), the results between LCA and MST could be directly compared with respect to their correctness during the taxonomic peptide assignment for the microbial mixture data sets.

Figure 4.31 displays that the MST method resulted in a slightly better performance for the proportions of correct taxon-specific peptide assignments in comparison to the LCA method. Finally, no significant differences were found between the result sets from both replicates with respect to the relative number of correct taxonomic assignments.

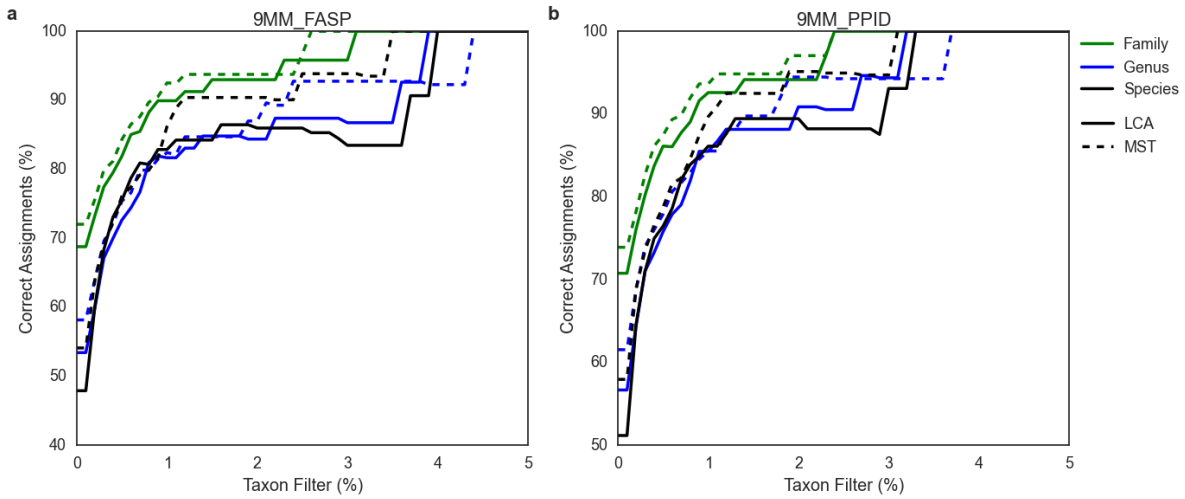

Figure 4.31: Performance comparison of LCA and MST taxonomic assignment methods. The line plots display the relative fraction of correct taxon-specific peptide assignments of MPA for species (black), genus (blue) and family (green) when using the LCA (solid) and MST (dashed) method as function of the taxon filter threshold for (a) 9MM_FASP and (b) 9MM_PPID data set results. The *Taxonomy Definition* feature of the MPA software was used to specify the LCA and MST method.

4.5.3 Phylogenetic Overview on Human Intestine Microbiota

Most studies on intestinal microbiota between lean and obese adults were conducted by techniques of 16S rRNA or metagenomic profiling [284]. Although a single study described the analysis of faecal samples taken from an obese and a lean adolescent [108], obese adults have not been analyzed nor compared to non-obese individuals at the proteome level. MS/MS data sets originating from 29 HIMP samples (see Section 3.2.2) were analyzed using the MPA software. The results of these analyses were used for a comprehensive study conducted by Kolmeder *et al.* in which the compositional and functional properties of the intestinal metaproteomes were compared between obese and non-obese adults [266].

While the focus of the above mentioned study was to find characteristic differences between the obese and non-obese group, the motivation of the next analysis was to establish a general phylogenetic overview on human intestine microbiota. For this purpose, ten HIMP data sets (HIMP10) were analyzed from which samples P1, P3, P8, P11 and P17 belonged to five non-obese subjects (BMI < 30 kg/m^2), while samples P23, P27, P28, P31 and P34 were derived from five obese individuals (BMI > 30 kg/m^2). For the taxonomic assignment of the HIMP10 data, the identification results were extracted from previous investigations (see Section 4.1.2) and thereupon the peptide sequences were subjected to Unipept [276, 277]. This procedure was required since HIMPdb contained an insufficient amount of protein accessions that were compatible with UniProtKB to perform the taxonomic assignment of peptides in the MPA software. In the first

investigation, the distribution of distinct peptides and identified spectra was examined at the highest taxonomic rank of superkingdom.

Figure 4.32 displays that the majority of the identifications in HIMP10 could be assigned to Bacteria with an average of 88.8% at the peptide and 86.2% at the spectrum level. It was found that 7.5% of the peptides and 10.4% of the spectra originated from Eukaryota on average. Moreover, it can be observed that the identification yield for Viruses and Archaea was minimal (0.1%).

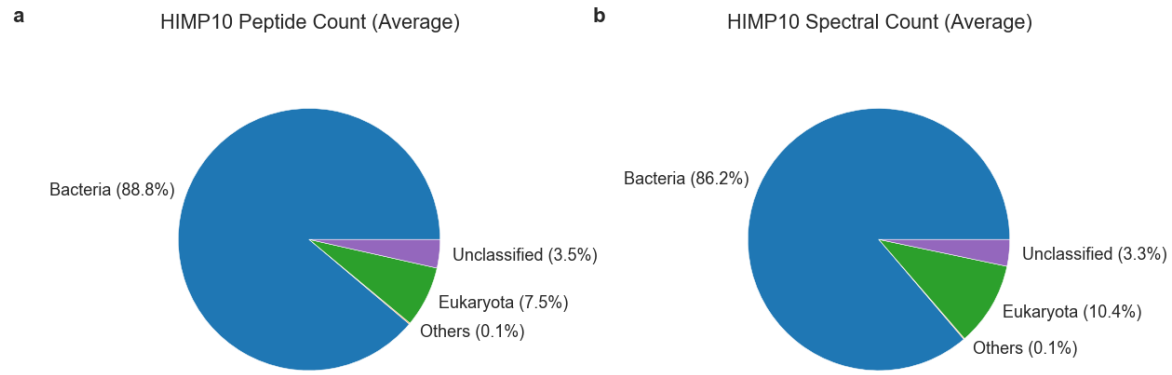

Figure 4.32: Superkingdom-level phylogenetic classification for HIMP10 data sets based on average amount of Unipept hits. The pie charts display the relative distribution of (a) distinct peptide sequences and (b) number of MS/MS spectra. The identifications were retrieved from searching HIMP10 samples against HIMPdb (FDR < 5%) and taxonomic assignments were obtained by subsequent use of Unipept. Peptides were called Unclassified when an assignment for superkingdom could not be found by Unipept.

Second, the proportion of peptides and spectra was examined that could be assigned to the most abundant phyla. For this analysis, peptides from non-bacterial origin were discarded and only those taxa were considered for which the ratio of identifications per phylum exceeded 1% in total.

Figure 4.33 illustrates that Actinobacteria (7.6%), Bacteroidetes (21.3%), Firmicutes (69.6%) and Proteobacteria (1.5%) were the major contributors of all bacterial phyla.

Third, a closer look was taken at the deeper taxonomic rank of genus. In general, lower ranks are expected to contain fewer assignments, since the LCA approach aims to converge at higher levels. However, the genus taxon allows a more detailed resolution than phylum-based investigations: therefore, the peptides assigned to the genera were counted for each of the four most prominent bacterial phyla.

It can be found that the most represented genera in Firmicutes were *Faecalibacterium* (21.5%), *Ruminococcus* (21.4%), *Eubacterium* (12.7%), *Roseburia* (11.1%), *Blautia* (8.8%) and *Clostridium* (7.6%) (Figure 4.34a). Moreover, it can be observed that the relative distribution of the genus-specific peptides in Firmicutes was more stable across the HIMP10 data sets than the one in

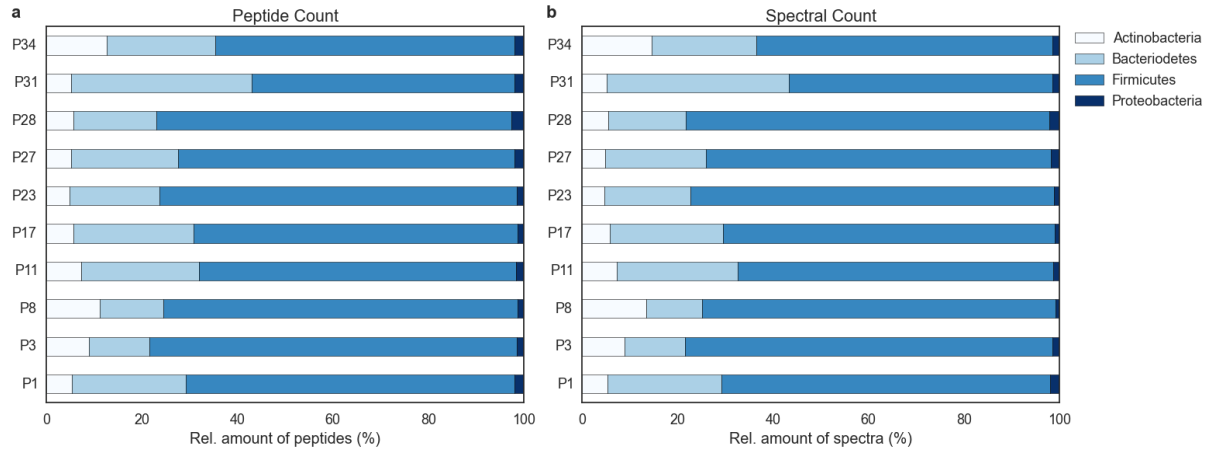

Figure 4.33: Phylum-level taxonomic classification of Unipept peptides for HIMP10 data sets. The stacked bar charts display the relative abundance of (a) peptides and (b) MS/MS spectra for Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria. The identifications were retrieved from searching HIMP10 samples against HIMPdb (FDR < 5%) and taxonomic assignments were obtained by subsequent use of Unipept. Non-bacterial hits and fraction below 1% of the total amount of identifications were filtered out.

Bacteroidetes (Figure 4.34b). In particular, the relative abundance of peptides in *Prevotella* was significantly increased for the data sets P1, P8, P11 and P27. In fact, the genus *Bacteroides* acted as the counterpart of *Prevotella* since a decreased amount of peptides were found in the same data sets.

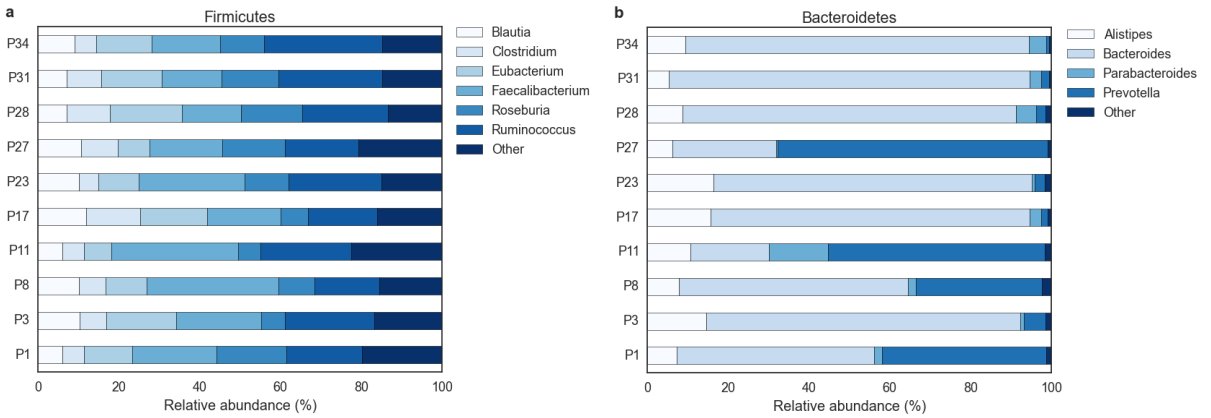

Figure 4.34: Genus-level taxonomic classification of Unipept peptides for HIMP10 data sets. The stacked bar charts display the relative abundance of genus-specific peptide hits for (a) Firmicutes and (b) Bacteroidetes. The identifications were retrieved from searching HIMP10 samples against HIMPdb (FDR < 5%) and taxonomic assignments were obtained by subsequent use of Unipept. Genera that contained less than 5% of the total identifications were classified as "Other".

Figure 4.35 displays the major difference between Actinobacteria and Proteobacteria: while Actinobacteria were supported to a large extent by peptides from two genera, namely *Bifidobacterium* and *Colinsella*, Proteobacteria contained assignments that were distributed among nu-

merous genera, including *Bilophila*, *Campylobacter*, *Desulfovibrio*, *Escherichia*, *Helicobacter*, *Pseudomonas* and *Sutterella*. Since the proportion of Proteobacteria-specific identifications was the lowest for the four considered phyla with 64 peptides on average, a more detailed analysis was renounced for this genus.

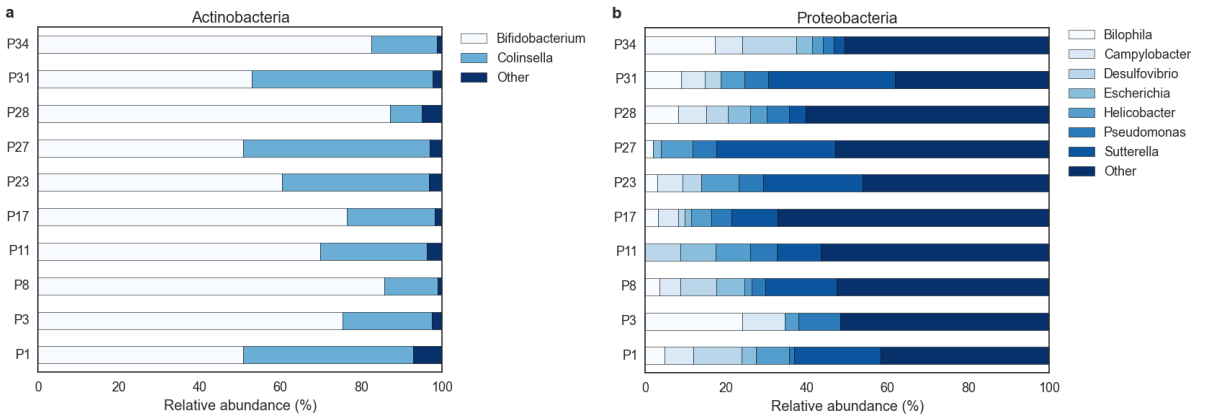

Figure 4.35: Genus-level taxonomic classification of Unipept peptides for HIMP10 data sets. The stacked bar charts display the relative abundance of genus-specific peptide hits for (a) Actinobacteria and (b) Proteobacteria. The identifications were retrieved from searching HIMP10 samples against HIMPdb (FDR < 5%) and taxonomic assignments were obtained by subsequent use of Unipept. Genera that contained less than 5% of the total identifications were classified as "Other".

4.6 Functional Analysis

The expressed proteins within a microbiome constitute the main target of research in metaproteomics. Notwithstanding the important requirement of the identification at the protein level, the ultimate goal presents the exploration of protein functions and detection of involved metabolic pathways in microbial communities. By using metaproteomic approaches, the composition of a microbial community can be linked to its function and the function of the community members can be investigated. If additional information on involved metabolites is available, models on active pathways can be constructed: combining data from metaproteomics and metabolomics holds the potential of gaining a comprehensive insight into the functioning of an ecosystem. In this section, the use of the MPA software is demonstrated for assigning metaproteomic data from BGP samples to functional ontologies, enzymatic activities and metabolic pathways. Subsequently, the reproducibility of the functional annotation is estimated on the basis of replicate samples. At last, alternative options for assigning unannotated protein sequences to functions are demonstrated using data set from HIMP samples.

4.6.1 Methods of Functional Annotation

An important goal of the analysis of BGPs is to understand the anaerobic digestion processes for subsequent optimization regarding robustness and yield of the biogas production (see Section 2.1.4). In this context, the metaproteome analysis of BGP samples investigates the anaerobic process steps in the process of converting biomass to methane, namely hydrolysis, fermentation, acetogenesis and methanogenesis. In this analysis, the main objective was to demonstrate the applicability of functional analysis methods in the MPA software. Therefore, the BGP data sets GENT01, GENT07 and GENT16 were first used to obtain detailed ontology information from UniProtKB for the identified proteins. Each ontological term thereby provides detailed information on the function of a particular protein. To study the processes inside the fermenters, the terms for the ontology *Biological Process* were derived. In addition, the ontology *Molecular Function* was investigated which describes the enzymatic activities of the microbial community inside the reactors.

Since a considerable bias of the chosen protein database was noticed in previous investigations (see Section 4.2.2, 4.2.1 and 4.5), the second goal was to evaluate this effect also for the functional analysis. Analogous to the previous section, the respective BGP result sets were used from searching against SwissProt and TrEMBL. In this analysis, identified spectra and peptides were taken as quantitative measures.

Figure 4.36 shows the peptide assignments of the BGP data sets for the ontology *Biological*

Process. The bar plots illustrate that most of the identifications in GENT01 and GENT07 were assigned to "Methanogenesis" for SwissProt, while this functional term was supported by fewer peptides in GENT16. The categories "Transport", "Hydrogen ion transport" and "ATP synthesis" were the most prominent in the GENT16 results for UniProt/TrEMBL. Notably, various peptide identifications for "Glycolysis" could be found for all three data sets. It can be also observed that GENT01 and GENT07 resulted in significantly more peptides identified from searches against SwissProt in comparison to TrEMBL, while the opposite was the case for GENT16.

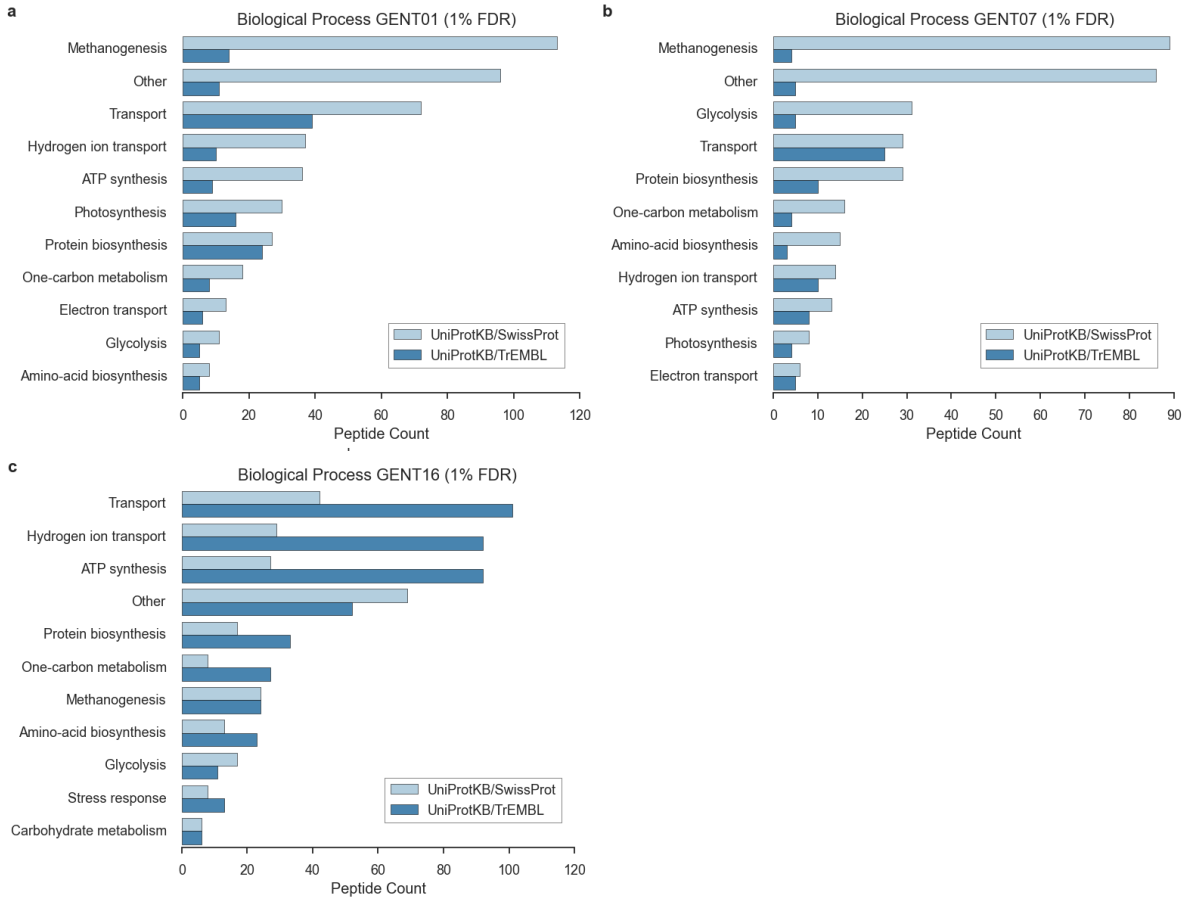

Figure 4.36: Total number of *Biological Process*-specific peptide hits for BGP data sets. The bar plots show the peptide assignments to ontological terms for (a) GENT01, (b) GENT07 and (c) GENT16. The identifications were obtained by searching with X!Tandem and OMSSA against SwissProt and TrEMBL (FDR < 1%). Fractions below 1% of the total identifications were classified as "Other".

Figure 4.37 illustrates that "Transferase" and "Oxidoreductase" were the most abundant functional terms for the ontology *Molecular Function* in each of the BGP result sets. It can be recognized that a portion of the identifications was assigned to the specific term "Methyltransferase". The enzyme "Acetyltransferase" belonging to the acetoclastic pathway was also found in the results of GENT01 and GENT07. In line with the findings for the ontology *Biological Process*,

more assignments were obtained for GENT01 and GENT07 from SwissProt than from TrEMBL searches, while the ratio of assigned peptides was inverted for GENT16 between the database variants.

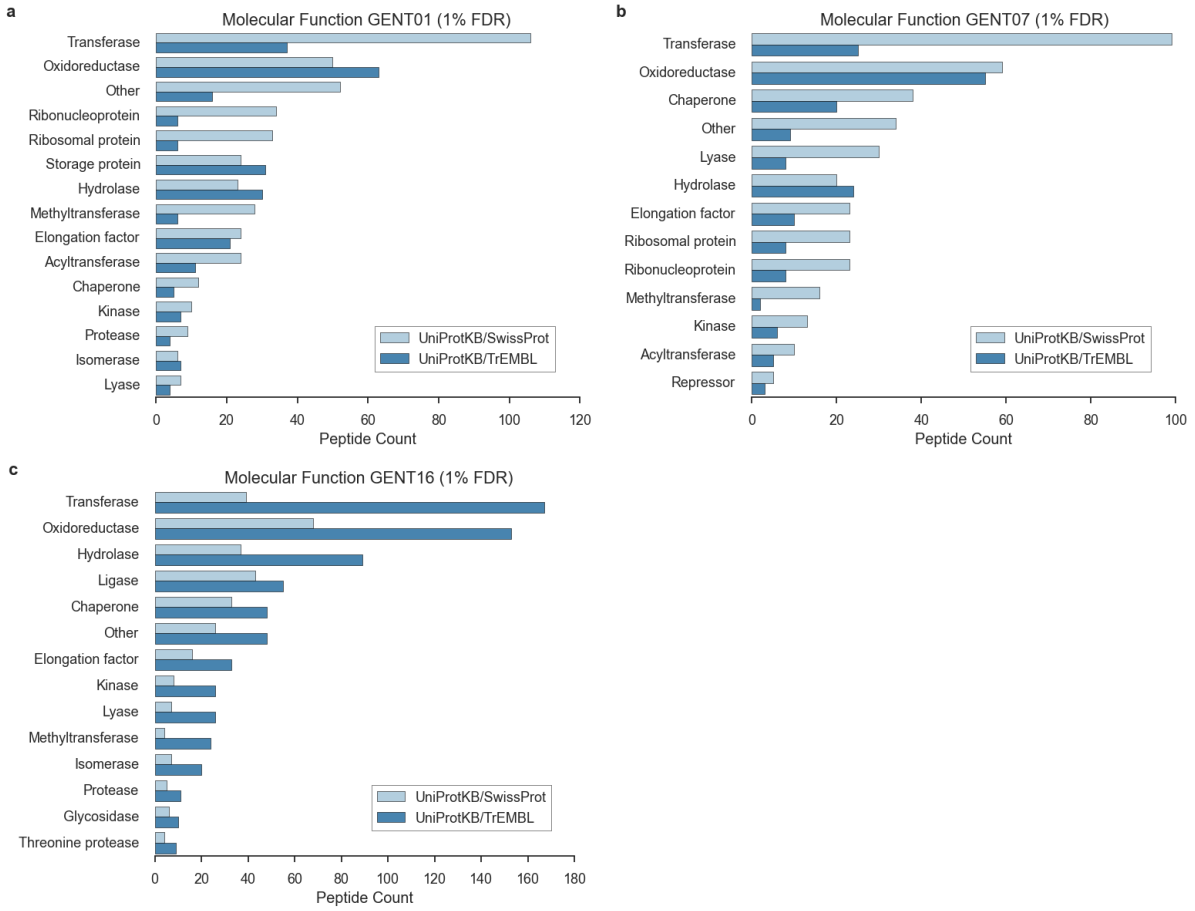

Figure 4.37: Total number of *Molecular Function*-specific peptide hits for BGP data sets. The bar plots show the peptide assignments to ontological terms for (a) GENT01, (b) GENT07 and (c) GENT16. The identifications were obtained by searching with X!Tandem and OMSSA against SwissProt and TrEMBL (FDR < 1%). Fractions 1% of the total identifications were classified as "Other".

Enzyme classification. To investigate potential enzymatic activities of the microbial consortia inside the BGPs more thoroughly, the data were further analyzed by using the EC numbers that specify enzyme-catalyzed reactions. This numeric nomenclature consists of four consecutive numbers: each number represents an increasingly finer classification of the enzyme. The main EC categories are oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5) and ligases (EC 6). For the following analysis, only the most detailed level of enzyme classification was inspected. In addition, a unique descriptive label was retrieved for each EC number. The number of assigned spectra and peptides were used to quantify the observed

enzymes. For the sake of clarity, assignments were filtered out which obtained less than 1% of the total identified spectra.

"Coenzyme-B sulfoethylthiotransferase" (EC 2.8.4.1) was the most abundant enzyme in both SwissProt and TrEMBL results for GENT01 (Figure 4.38). The observed enzyme is also called "Methyl-coenzyme M reductase" and catalyzes the final step in the formation of methane. In general, the overlap at the enzyme level was low between the protein databases: in comparison to SwissProt, fewer identifications were assigned to EC numbers in TrEMBL. While "Co-methyltransferase" (EC 2.1.1.245), an enzyme from the acetoclastic pathway was observed for SwissProt, it was absent for TrEMBL. Vice versa, "5,10-Methylenetetrahydromethanopterin reductase" (EC 1.5.99.11) could be identified only in the TrEMBL results.

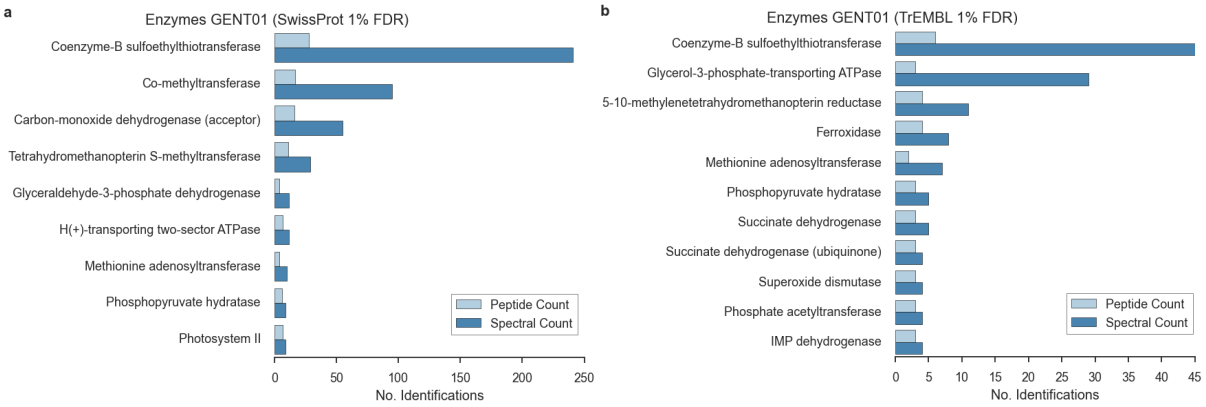

Figure 4.38: Total number of enzyme-specific spectrum and peptide hits for GENT01. The identifications were obtained by searching with X!Tandem and OMSSA against (a) SwissProt and (b) TrEMBL at 1% FDR. Assignments with less than 1% of the total identified spectra were filtered out.

Moreover, a difference could be observed between the results of GENT01 and GENT07, since only the latter yielded a significant abundance of "Coenzyme F420 hydrogenase" (EC 1.12.98.1) which represents an essential enzyme for the final processing step of methanogenesis (Figure 4.39). It can be also recognized that—compared to GENT01—fewer identifications could be assigned to specific enzymes for GENT07. Conversely, the abundance of enzyme-specific assignments was higher for GENT16 than for the other data sets (Figure 4.40). The most prominent enzyme was "Carbon-monoxide dehydrogenase" (EC 1.2.99.2) which belongs to the acetoclastic methane production route. Furthermore, "Acetate-CoA ligase" (EC 6.2.1.1) could be detected exclusively in the GENT16 results. It should be also noted that "Glyceraldehyde-3-phosphate dehydrogenase" (EC 1.2.1.12) was found in all data sets. Finally, "Alcohol dehydrogenase" (EC 1.1.1.1) was found exclusively in the TrEMBL results for GENT16. Analogous to previous observations for GENT16, the total number of assignments was markely higher for TrEMBL in comparison to SwissProt.

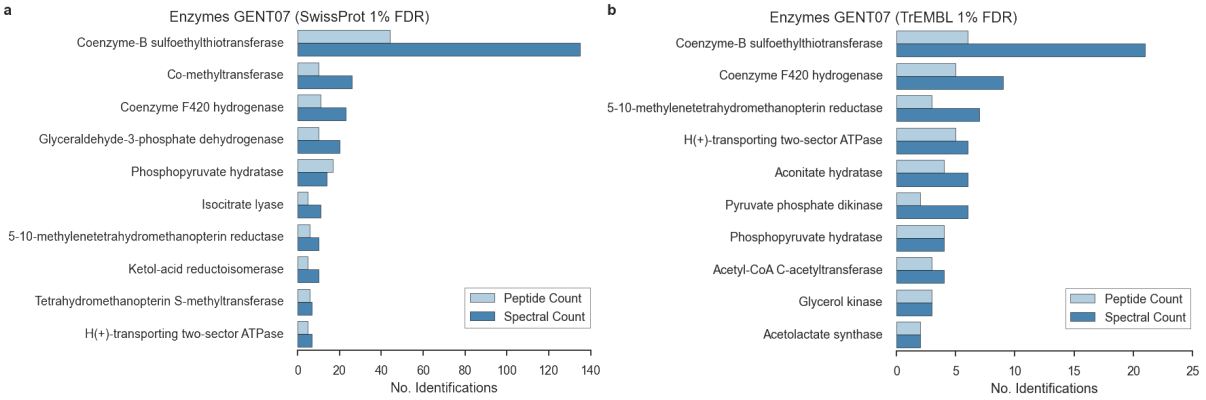

Figure 4.39: Total number of enzyme-specific spectrum and peptide hits for GENT01. The identifications were obtained by searching with X!Tandem and OMSSA against (a) SwissProt and (b) TrEMBL at 1% FDR. Assignments with less than 1% of the total identified spectra were filtered out.

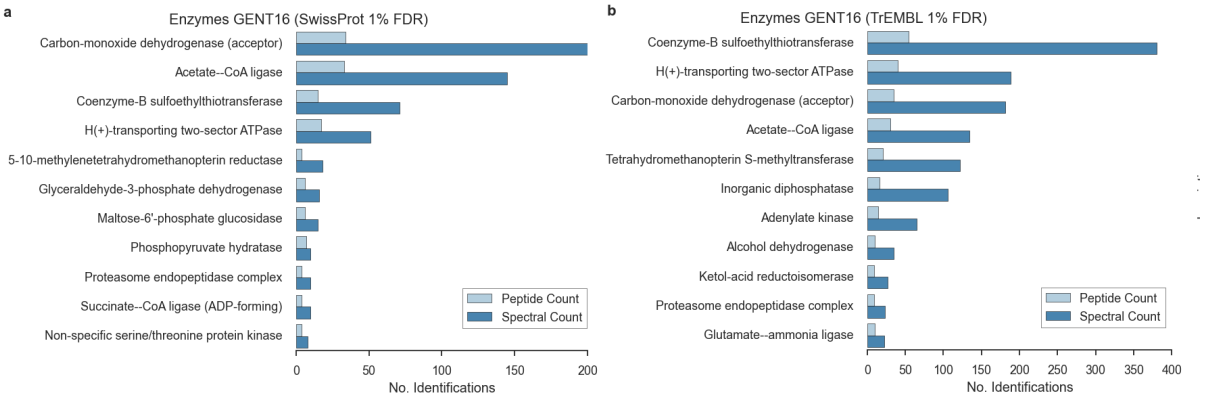

Figure 4.40: Total number of enzyme-specific spectrum and peptide hits for GENT16. The identifications were obtained by searching with X!Tandem and OMSSA against (a) SwissProt and (b) TrEMBL at 1% FDR. Assignments with less than 1% of the total identified spectra were filtered out.

The performed analysis illustrates a systematic way to gain an overview on potentially active enzymes in metaproteomic samples. However, an analysis solely based on EC nomenclature has the limitation of disregarding the context in which the enzymes are working. Hence, another key aspect for a functional data analysis presents the role of enzymes in biochemical and metabolic pathways (see also Section 2.2.6). To fully cover the functional space, the mapping of relevant protein identifications into such networks is demonstrated in the next paragraph.

Mapping into KEGG pathways. To gain knowledge about the functioning of microbial communities in the environment, a global view on molecular routes is required. Therefore, the MPA software allows to map identified proteins into KEGG pathways by direct submission to the KEGG website. To demonstrate the value of this feature, the protein results for GENT01 at 5% FDR were mapped into the major pathway of carbon metabolism.

Figure 4.41 shows the respective KEGG reference pathway (map01200) including involved metabolites, intermediates and enzymes. In this case, the identified proteins from GENT01 are highlighted in the network and represent enzymes relevant for the production of methane: this example shows that both pathways describing the conversion of acetate and carbon dioxide to methane are covered by the data.

In another pathway analysis, the taxonomic range was restricted to the level of superkingdom. Thereby, the protein identifications that were unique to Archaea and Bacteria were assigned to the previous pathway map of carbon metabolism. It can be observed that the pathway of methanogenesis is only present in Archaea (Figure A.10 in the appendix), while the one of glycolysis/gluconeogenesis is mainly represented by Bacteria (Figure A.11 in the appendix).

Another distinction at the taxonomic level could be found when comparing the results of the two superkingdoms in the KEGG reference pathway of amino acid biosynthesis (map01230): few proteins could be assigned Archaea (Figure A.12 in the appendix), in contrast to Bacteria which are heavily involved in this route (Figure A.13 in the appendix). More detailed findings on this analysis are provided in the publication of the MPA software [256].

In summary, the advantage of this implementation is to access the KEGG pathway maps directly from the MPA software. Hence, this approach enables the researcher to examine the coverage of essential pathways for any metaproteomic result sets. In general, the developed workflow provides several methods to investigate the functional profile of the microorganisms within a sample.

4.6.2 Quantifying the Functional Profile

In the foregoing functional analysis, the number of identified spectra and peptides was used as measure of quantification. While the focus was on the functional assignment of identifications in single data sets, the robustness of the quantitative methods needs to be further evaluated for consistency between multiple different results sets. Consequently, the next objective was to investigate the reproducibility of the results for different quantitative measures: in particular, the number of identified spectra, peptides, protein and meta-proteins was reviewed for their applicability to the functional analysis of results from different experiments. As already addressed in Section 4.4.2, the results from the GENT01 and GENT16 data sets were used that each originated

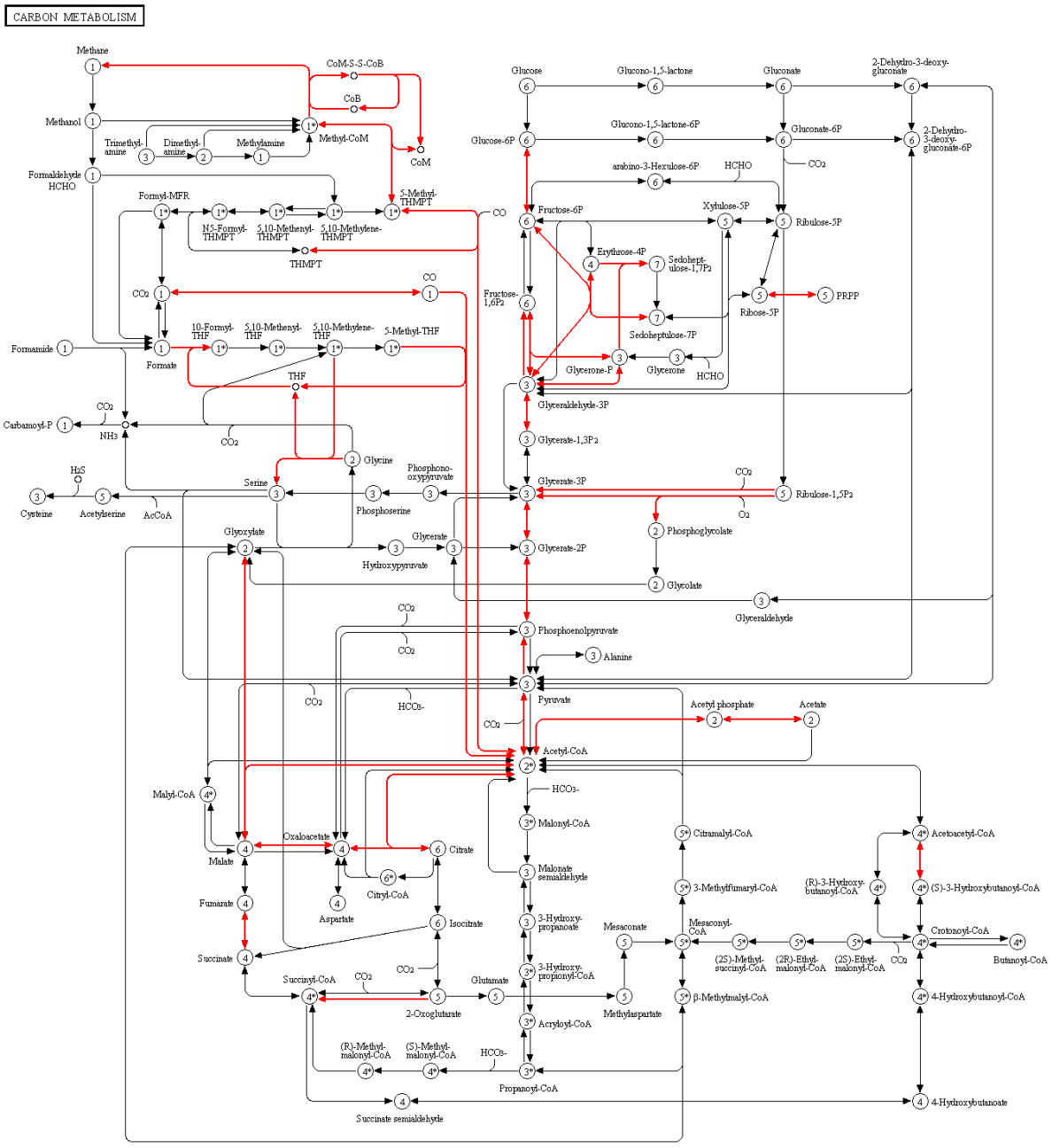

Figure 4.41: KEGG reference pathway of carbon metabolism (map01200) for GENT01 protein identifications. The edges represent enzymes required for the conversion of one metabolite into another. The identified proteins of the data set are highlighted in red after submission to the KEGG website. The identifications were obtained by searching with X!Tandem and OMSSA against SwissProt at 5% FDR.

from two technical replicates. The ontologies *Biological Process* and *Molecular Function* were correlated between the replicates on the basis of the number of assigned identification peptides for each functional term. The correlation analyses were performed using Pearson's correlation coefficients.

Figure 4.42 displays that the correlation between the results from both GENT01 replicates was very strong for each of the evaluated quantitative measures (pearsonr >= 0.96). At the meta-protein level, the functional assignments were close to the maximum (pearsonr = 0.99). High correlations could also be found for the GENT16 replicates (Figure A.14 in the appendix). In this case, the spectrum assignments showed a perfect linear relationship (pearsonr = 1.0).

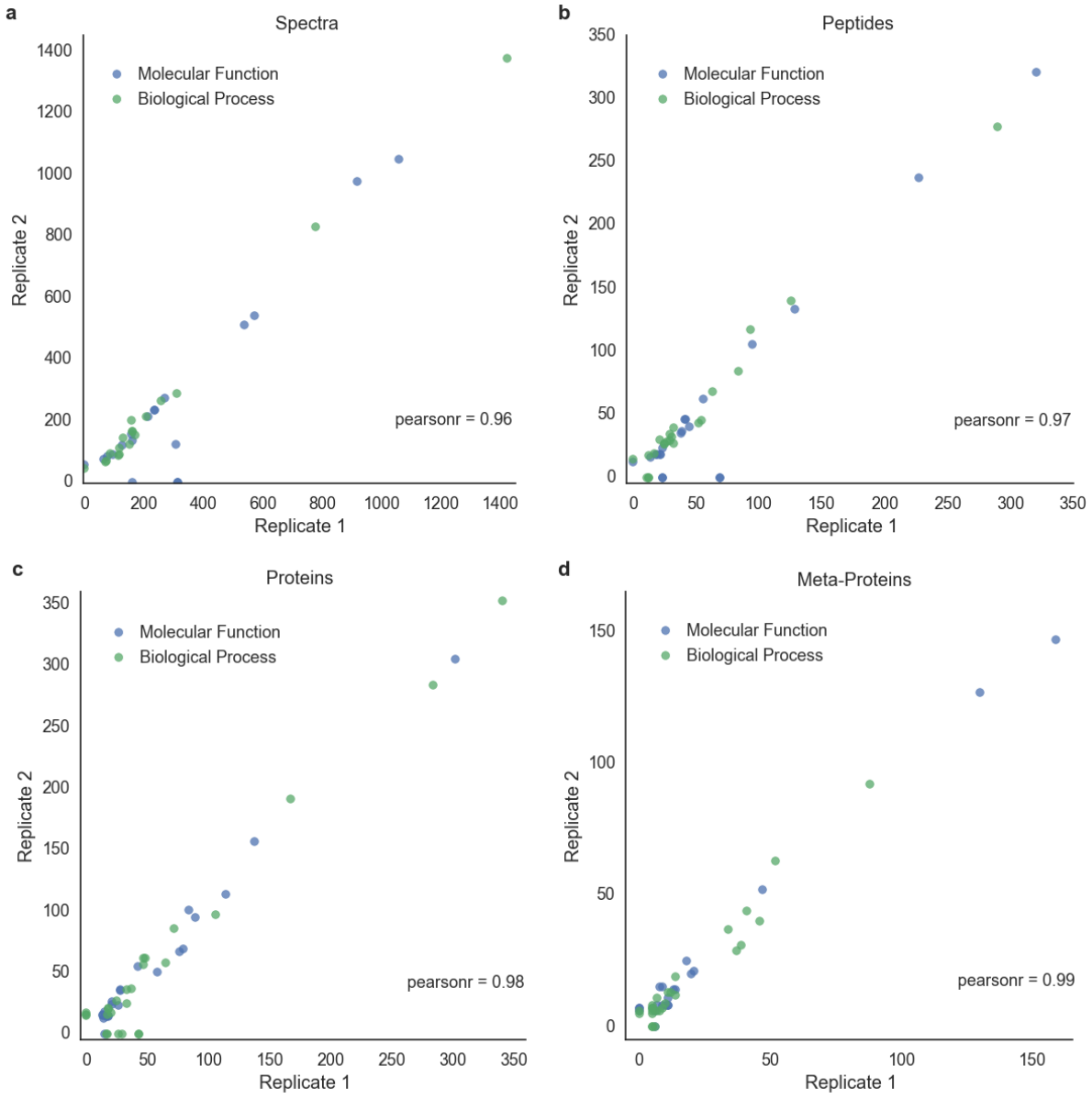

Figure 4.42: Reproducibility of ontology-specific assignments across technical replicates for GENT01. Each scatter plot compares either the number of (a) spectra, (b) peptides, (c) proteins and (d) meta-proteins that were reproducibly assigned across two replicate experiments to the functional ontologies *Molecular Function* (blue) and *Biological Process* (green). The data set GENT01 was searched against SwissProt (5% FDR). Meta-proteins were generated by using the *Minimum One Shared* rule. The Pearson correlation coefficient (pearsonr) is displayed in the lower right corner of each panel.

Since the comparison of replicates merely displays the technical variation, functional data of different BGP samples was investigated in the following analysis. Therefore, the identifications that could be uniquely assigned to the functional term "Methanogenesis" from the ontology *Biological Process* were compared between GENT01, GENT07 an GENT16 data searched against SwissProt at 1% FDR.

It could be found that spectra and peptides (Figure 4.43a) had a larger impact on the abundance of "Methanogenesis" than proteins and meta-proteins (Figure 4.43b). For instance, while the number of identified spectra varied significantly between GENT01 (786) and GENT16 (113), the number of meta-proteins was similar between GENT01 (12) and GENT16 (14).

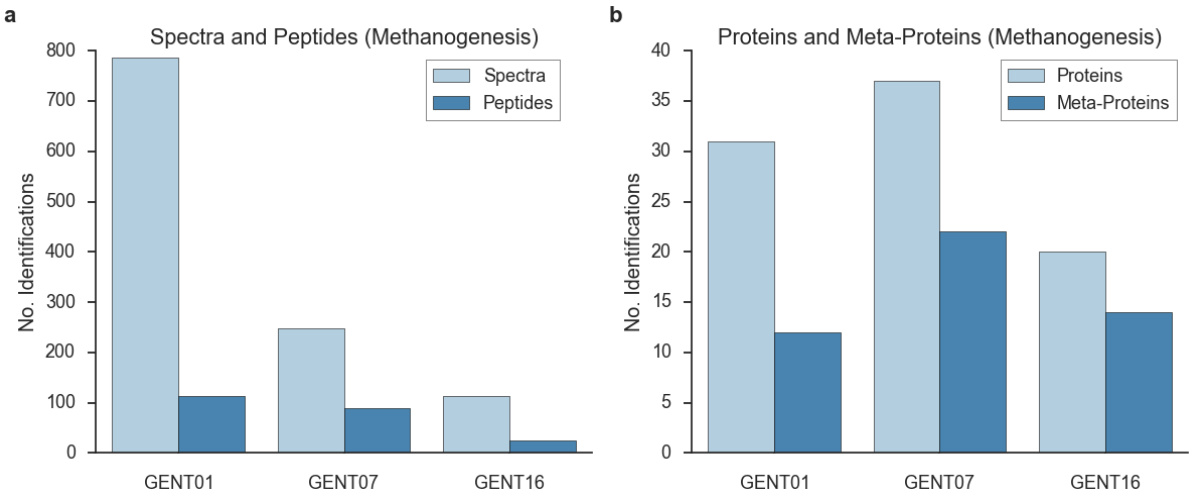

Figure 4.43: Total number of "Methanogenesis"-specific identifications for BGP data sets. (a) The absolute number of assignments to the ontology term "Methanogenesis" are shown at the spectrum and peptide level for GENT01, GENT07 and GENT16. (b) In addition, the "Methanogenesis"-specific assignments are displayed at the protein and meta-protein level. The results were obtained by searching with X!Tandem and OMSSA against SwissProt at 1% FDR.

In the previous section, the variance of different quantitative measures for the functional analysis was investigated. To infer functional processes from metaproteomic data, however, the proportion of identified spectra and peptides may be less important than the quantity of different proteins: for instance, the more enzymes are found in the result set, the better the metabolic network can be explained for a particular pathway. Therefore, the total number of proteins which could be mapped to KEGG pathways was examined for the BGP results in the following.

To investigate the coverage of identified enzymes in the pathways, the KO and EC numbers from the GENT01, GENT07 and GENT16 protein results were used that were identified by searching against SwissProt and TrEMBL at 5% FDR. The identified proteins were mapped to the KEGG reference pathway for carbon metabolism (map01200). In a preliminary analysis, it was found that around five times more peptides in GENT16 were assigned to carbon metabolism for

TrEMBL in comparison to SwissProt (Figure A.15 in the appendix). Consequently, the hypothesis was that the choice of the database would also affect the number of found KO and EC numbers.

Figure 4.44a shows that searching the GENT16 data set against TrEMBL doubled the number KO identifiers for carbon metabolism in comparison to SwissProt. An increase in EC identifiers can be also observed from SwissProt to TrEMBL for GENT16 (Figure 4.44b). For the evaluated result sets, clear differences in the number of assigned KO and EC numbers were found between the two chosen databases.

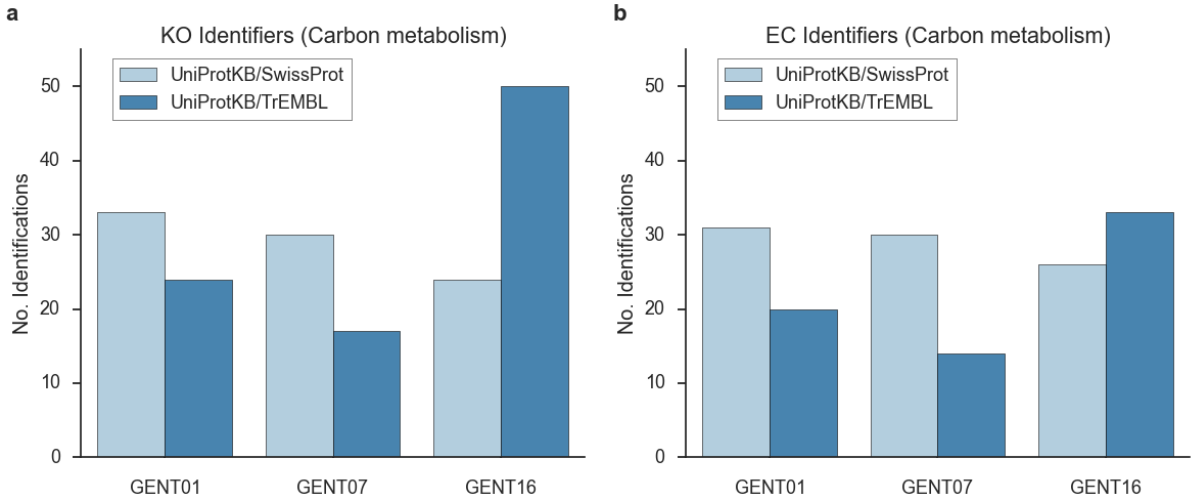

Figure 4.44: Total number of "Carbon metabolism"-specific identifiers for BGP data sets. The bar plots show the total number of (a) KO and (b) EC identifiers covering the pathway carbon metabolism (KEGG map01200) for GENT01, GENT07 and GENT16. The results were obtained by searching with X!Tandem and OMSSA against SwissProt and TrEMBL at 5% FDR.

The previous investigations relied exhaustively on metadata from the UniProtKB databases. However, in metaproteomics, customized protein databases that frequently contain unannotated sequences are used either as alternative or supplement to public databases. In the next section, additional methods are demonstrated to perform functional studies on sparsely annotated proteins in the data for the HIMP samples.

4.6.3 Postprocessing Unannotated Data

To investigate potential functions of proteins without given sequence annotation, various tools and databases are available (see Section 2.2.6). In the following, the HIMP10 results sets were processed by matching identified protein sequences against the EggNOG database (see Section 3.4.6) [213]. EggNOG provides a large repository of non-supervised orthologous groups (NOGs) from complete genomes. The EggNOG analysis was limited to sequences from Bacteria, since it had been previously observed that the majority of the results in the HIMP data were of bacte-

rial origin (see Section 4.5.3). The main purpose of the following analysis was to find bacterial NOGs that could separate result sets between non-obese and obese individuals. In order to pass protein sequences to EggNOG, the respective protein identifications that yielded a minimum of two peptides were exported using the MPA software. To search against the bacterial EggNOG database, the software package HMMER (ver. 3.1) [278] was used, since the EggNOG website allowed only one search query at a time. The results were filtered by using the best search hit with an e-value below 10^{-30} for each query.

The first objective was to obtain an overview on the functional profile of the HIMP results. Therefore, the data sets P1 and P23 were analyzed using the MPA and EggNOG annotation pipeline as described above. Sample P1 originates from a non-obese (BMI = 22.0) and P23 from an obese subject (BMI = 50.0). The distinct peptides for each EggNOG category were used as quantitative measures for the functional analysis. In addition, the influence of the chosen FDR threshold on the number of assigned peptides was evaluated.

Although the absolute numbers of assignments moderately increased at higher FDR values, the relative proportions remained constant (Figure 4.45). For instance, the proportion of P1 peptides assigned to "Carbohydrate transport and metabolism" accounted for 42.1% at 5% FDR and 43.0% at 1% FDR. In general, it can be observed that the highest proportion of assignments could be related to the aforementioned category. At 5% FDR, 55–60% of the 20 most abundant NOGs belonged to this functional class (Table A.12 and A.13 in the appendix). For P1, 1 189 (965) proteins could be assigned and for P23, 1 156 (986) proteins of this category were found at 5% (1%) FDR (Table A.14 and A.15 in the appendix). The second most abundant functional class was "Amino acid transport and metabolism" at 5% FDR. Next to the previous category, proteins with unknown function were in a similar abundance range. Furthermore, the forth most abundant class was "Energy production and conversion".

While the first aim was to gain general insights into the functional potential of two representative samples, the major goal of metaproteomics is to examine the relation between microbial community and function. To increase the information gain, functional data is coupled to knowledge about the taxonomic origin of proteins in the study. Therefore,the HIMP10 result sets were further analyzed by combining EggNOG functional annotation with taxon-specific peptide assignments that had previously retrieved by means of the Unipept analysis software (see Section 4.5.3). In the following, the functionally classified data were grouped into different bacterial taxa for the phylum rank. The taxa Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria were used, since they were the four most abundant phyla according to previous findings in this work.

Figure 4.46a displays that the major part of the peptides originating from the Gram-positive Firmicutes was linked to "Carbohydrate transport and metabolism" (53.9%). The next most abundant categories were "Amino acid transport and metabolism" (19.2%) and "Energy produc-

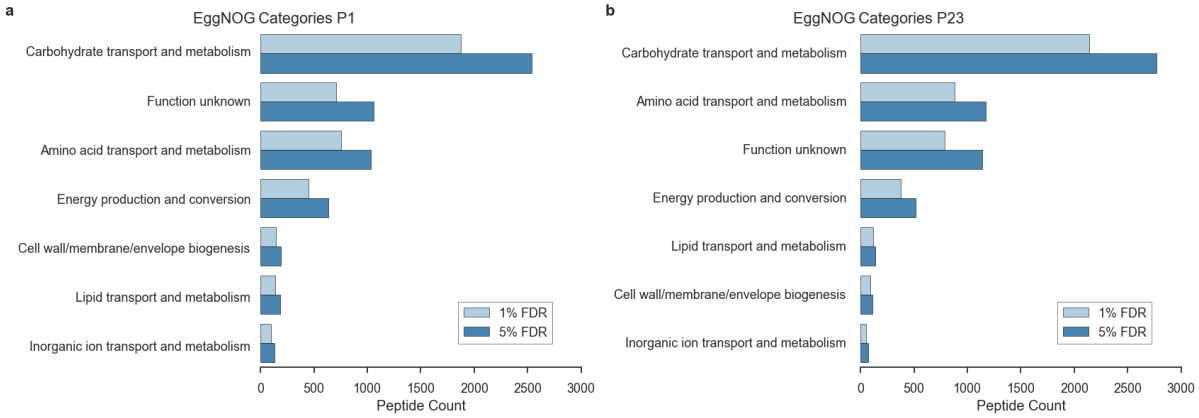

Figure 4.45: Total number of peptides assigned to EggNOG categories for HIMP data sets at 1% and 5% FDR. The identifications for (a) P1 and (b) P23 were obtained by searching with X!Tandem and OMSSA against HIMPdb. Fractions below 1% of the total assignments were filtered out.

tion and conversion" (14.4%). Less frequently found were peptides linked to functional classes "Function unknown" (5.0%) and "Lipid transport and metabolism" (3.6%). In contrast, 37.5% of Bacteroidetes-specific peptides could not be associated with a known function (Figure 4.46b). Similar as in Firmicutes, "Carbohydrate transport and metabolism" (30.0%) was the most represented category of known function. In addition, "Amino acid metabolism and transport" (9.1%) and "Energy production" (7.7%) play an important role in the Gram-negative Bacteroidetes. Moreover, only "Inorganic ion transport/metabolism" (8.8%) and "Cell wall/membrane/envelope biogenesis" (5.1%) were represented in significant amounts for this phylum.

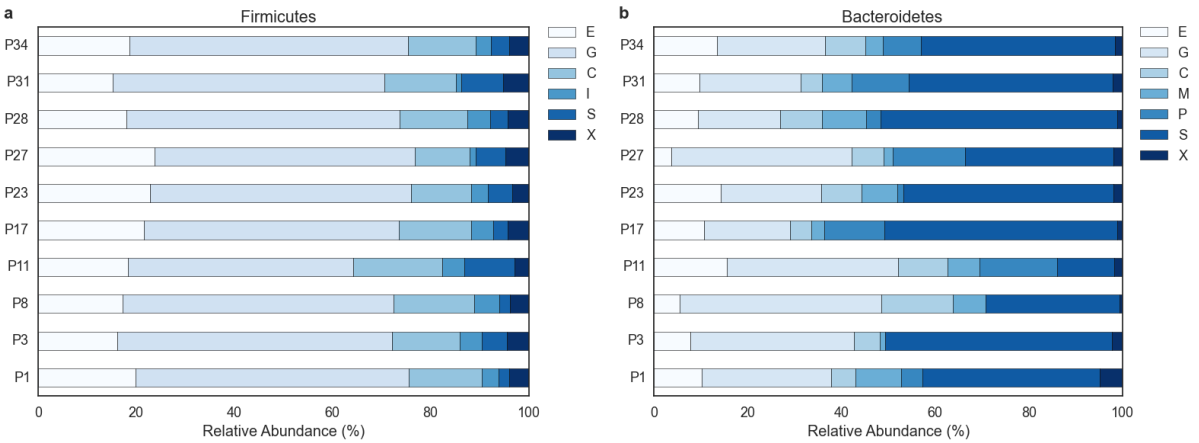

Figure 4.46: Phylum-level functional classification of peptides from Firmicutes and Bacteroidetes for HIMP10. The stacked bar charts display the relative abundance of (a) Firmicutes and (b) Bacteroidetes peptides that could be assigned to the following EggNOG categories: (E) Amino acid transport and metabolism; (G) Carbohydrate transport and metabolism; (C) Energy production and conversion; (I) Lipid transport and metabolism; (S) Function unknown; (P) Inorganic ion transport and metabolism; (M) Cell wall/membrane/envelope biogenesis; (X) Other. Less than 2% of the total identifications were classified as Other.

Figure 4.47a shows that the major part of Actinobacteria is associated with functions related to carbohydrate metabolism (60.3%). The second most abundant category represents "Amino acid transport and metabolism" (33.8%), while "Energy production and conversion" (2.8%) plays a minor role in this phylum. While anaerobic and Gram-positive Actinobacteria that mainly reside in the large intestine were dominated by two functional categories, Gram-negative Proteobacteria provided more mixed functions (Figure 4.47b). The functional diversity in Proteobacteria fits well with the different observed taxonomic genera (see Section 4.5.3). However, it should be considered that only an insignificant number of Proteobacteria-specific peptide assignments was found in the results (Table A.16 in the appendix). While the relative functional distribution was relatively stable across the data sets for Firmicutes, Bacteroidetes and Actinobacteria, the absolute amount of the phylum-specific assignments was highly variable (Table A.16 in the appendix).

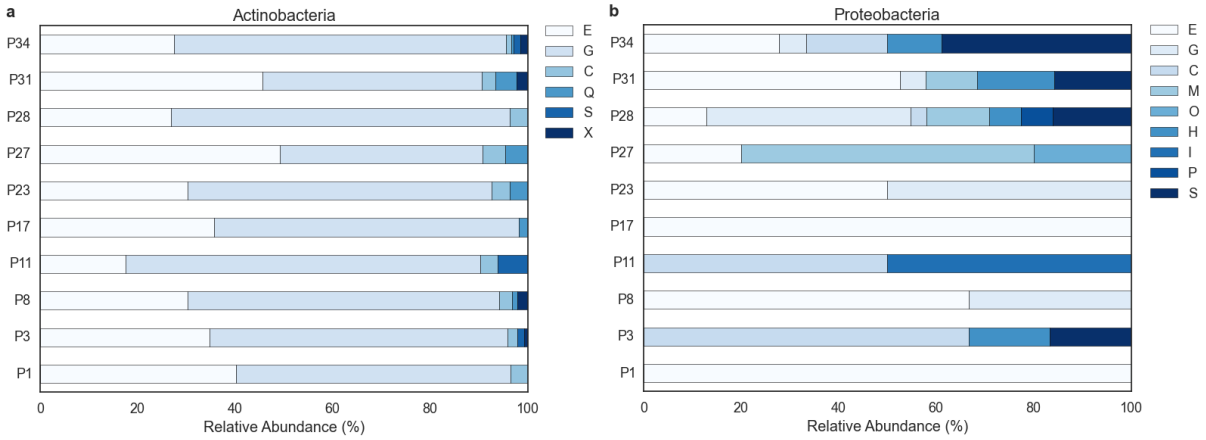

Figure 4.47: Phylum-level functional classification of peptides from Actinobacteria and Proteobacteria for HIMP10. The stacked bar charts display the relative abundance of (a) Actinobacteria and (b) Proteobacteria peptides that could be assigned to the following EggNOG categories: (C) Energy production and conversion; (E) Amino acid transport and metabolism; (G) Carbohydrate transport and metabolism; (H) Coenzyme transport and metabolism; (I) Lipid transport and metabolism; (M) Cell wall/membrane/envelope biogenesis; (O) Posttranslational modification, protein turnover, chaperones; (P) Inorganic ion transport and metabolism; (Q) Secondary metabolites biosynthesis, transport and catabolism; (S) Function unknown; (X) Other. Less than 2% of the total identifications were classified as Other.

Supervised classification based on bacterial functional groups. In the original study on the HIMP data [266], bootstrap aggregated (bagged) redundancy analysis (RDA) was used as supervised classification method to identify each non-supervised orthologous group (NOG) that differed significantly between obese and non-obese individuals. The idea behind this approach was to detect characteristic functional patterns in the invenstigated data by which a new sample could be reliably classified into a specific category (obese or non-obese). To evaluate the general validity of supervised classification on the metaproteomic data, another method called LEfSe

was applied which is frequently used for classifying high-dimensional data of microbial samples (see Section 3.4.7). Overall, the scope of this analysis was to compare the performance of both supervised methods in order to justify their application for the classification of metaproteomic data based on their functional annotation. For this analysis, the complete set of EggNOG assignments for 13 obese and 16 non-obese subjects was used to maximize the sensitivity of the supervised method. The feature vectors were constructed by extracting the peptide assignments for each EggNOG identifier from the results. In accordance with the above mentioned study, NOGs with less than seven assignments were filtered out.

Table 4.8 summarizes 27 NOGs that were identified as significantly different between lean and obese samples. The six NOGs that were characteristic for the obese group were two lipoprotein types, aminoacyl-histidine dipeptidase, arginine deiminase, a LacI transcription factor and pectate lyase. The last enzyme is produced by bacteria to degrade plant material and points to a pectin rich diet. Conversely, from the NOGs identified for the non-obese individuals, enzymes that are mainly involved in carbohydrate metabolism were overrepresented, including alpha-glucuronidase, carbohydrate kinase, enolase, glycoside hydrolase, phoshoketolase, riboulokinase, xyluokinase. In same group of data, the abundant proteins from amino acid metabolism were aminotransferase, cysteine desulfurase and cystathionine. Consequently, the use of bacterial NOGs led to a separation of the two groups, in particular, due to increased abundance of enzymes characteristic for the non-obese class.

Table 4.8: The supervised classification method LEfSe reports 27 NOGs that were significantly different (alpha = 0.01) between non-obese and obese individuals. For each NOG the corresponding group and LDA score are given. The LDA score can be regarded as a measure of consistent difference in relative abundance between the NOGs in the groups. An LDA score threshold of 2 was used.

EggNOG Description	Group	LDA Score
Alpha-glucuronidase (EC 3.2.1.139)	non-obese	2.40
Alpha-keto-beta-hydroxylacyl reductoisomerase	non-obese	3.05
Amino acid ABC transporter	non-obese	2.61
Aminoacyl-histidine dipeptidase	obese	3.51
Aminotransferase	non-obese	3.06
Arginine deiminase (EC 3.5.3.6)	obese	3.11
Arylsulfotransferase	non-obese	2.50
Bifunctional purine biosynthesis protein purh	non-obese	2.69
Carbohydrate kinase	non-obese	2.62
Catalyzes reaction R1*	non-obese	2.53
Catalyzes reaction R2*	non-obese	3.95
Catalyzes reaction R3*	non-obese	3.61
Cystathionine	non-obese	3.08
Cysteine desulfurase	non-obese	2.85
Dihydro-orotase (EC 3.5.2.3)	non-obese	2.58
Enolase N	non-obese	3.03
Formate acetyltransferase	non-obese	3.45
Galactose glucose-binding lipoprotein	obese	3.37
Glycoside hydrolase,family 3 domain protein	non-obese	3.23
Lipoprotein	obese	3.13
Periplasmic binding protein LacI transcriptional regulator	obese	3.50
Pectine lyase	obese	2.49
Phosphate-selective porin O and P	non-obese	2.34
Ribulokinase	non-obese	2.49
Phosphoketolase	non-obese	3.26
Promotes the GTP-dependent binding of aminoacyl-tRNA	non-obese	2.34
Xylulokinase (EC 2.7.1.17)	non-obese	2.84

R1*: Condensation reaction of fatty acid synthesis by the addition to an acyl acceptor of two carbons from malonyl-ACP. R2*: Reversible conversion of 2-phosphoglycerate into PEP. R3*: Reversible conversion of 3-phosphohydroxypyruvate to phosphoserine and of 3-hydroxy-2-oxo-4-phosphonooxybutanoate to phosphohydroxythreonine.

5 Discussion

5.1 Combining Multiple Search Algorithms

For the evaluated metaproteomic data, each database search engine resulted in a significant proportion of algorithm-specific identifications. As a consequence, when combining the results by addition of these individual search engine hits, the overall identification yield could be increased compared to the single use of database search algorithms.

The results from the preliminary analysis of BGP data sets advocate the application of multiple search engines, since each of the three algorithms reported a considerable fraction of algorithm-specific hits: for instance, X!Tandem identified exclusively 24% of the spectrum and 28% of the peptide hits (Figure 4.1). While OMSSA and MASCOT retrieved a lower absolute and exclusive number of hits, X!Tandem exhibited the highest identification yield. This result can be explained by the refinement mode used in the algorithm which performs an additional analysis on previously selected candidate proteins to detect missed cleavages and modifications during the peptide spectrum matching [281, 170]. While this feature greatly reduces the running time of the identification process, limitations caused by the induced statistical bias of such multi-pass searches have been described in the literature [285, 286]. Therefore, a reliable FDR estimation is mandatory to ensure a high quality of the results [287] (see Section 5.2.2).

The combined use of X!Tandem and OMSSA for the analysis of ten data sets from HIMP samples resulted in spectrum identification rates of 30% at 5% FDR and 21% at 1% FDR (Table 4.1 and A.2). These proportions are higher than the values commonly reported in other metaproteomic

studies: while a study on the human intestine using SEQUEST [168] as single search engine listed between 8% and 17% identified spectra for different databases at 5% FDR [199], a mouse metaproteome analysis in which the commercial Scaffold software was used resulted in only 5% spectrum identifications below 1% FDR [154]. Hence, the findings in this work demonstrate that the use of more than one search algorithm can significantly increase the fraction of identified spectra for metaproteomic experiments. Furthermore, the results in this work show that the complementary use of both algorithms is justified, since X!Tandem yielded 25% and OMSSA 11% exclusive spectrum identifications. In line with these findings, another study on human cell line data reported a complementarity of 12% for X!Tandem and OMSSA at 1% FDR [288]. Moreover, the authors suggested using a combination of up to five search engines to increase the sensitivity of database searching. However, it was beyond the scope of this study to evaluate the use of such a high number of algorithms, since each additional search engine commonly doubles the running time and increases the overhead for combining the results. Frequently, the use of multiple search algorithms is impractical for many metaproteomic research groups due to constraints in time and computational resources. To efficiently improve the identification yield for large amounts of upcoming MS/MS data in metaproteomics, more advanced computational solutions with a reasonable scalability and cost efficiency, such as using cluster and cloud computing, are recommended [289, 290, 291].

Another benefit of using multiple search engines is the option to validate questionable hits, so-called one hit wonders, i.e. protein identifications that are based on a single peptide hit [292]. Gupta and Pevzner argued against the common practice of experimentalists to overhastily exclude potentially valuable single-peptide hits [293]. To avoid early elimination of correct hits, the combination of multiple search engine results can be a useful backup method: while one algorithm may miss one hit or assign a low score, another search engine retrieves a high scoring peptide for the same spectrum. Consequently, this approach retains the individual strength of each algorithm without suffering from the disadvantage of losing identifications due to inaccurate scoring or inappropriate parameter choice.

Summary. In conclusion, the application and combination of multiple search algorithms increases the sensitivity and specificity of the identification process. While respective methods have been investigated in numerous proteomic studies [294, 295, 296, 297], it was here demonstrated that combining different identification algorithms is particularly useful for the data analysis in metaproteomics. The MPA software lowers the adoption threshold of such an approach by fully integrating multiple search engines and automatically combining hits identified by individual algorithms. Although the algorithm choice is vital for the successful outcome of a metaproteomic experiment, an equally important aspect presents the selection of appropriate search engine pa-

rameters as discussed in the following section.

5.2 Evaluating Parameters of Database Searching

The objective of the investigations in Section 4.2 was to evaluate the determining factors of database searching that impact the identification of peptides and proteins in metaproteomic experiments. The influence of various parameters, including protein database, search strategy and cleavage settings, on the identification yield was shown by analyzing exemplary data sets from metaproteomic samples.

5.2.1 Influence of Protein Database

In contrast to pure-culture proteomics, low numbers of available genomes led to a decreased identification yield in metaproteomic experiments in the past [5, 12, 13]. To cover the widest possible search space for protein identification in microbial communities, manually constructed as well as publicly available databases are used in metaproteomic analysis workflows [14, 88, 89, 113]. While latter investigations focused on the biology behind the data, the goal of the computational analysis in this work was to inspect the relation between protein database composition and identification yield. For this purpose, three exemplary metaproteomic data sets from BGP samples (GENT01, GENT07 and GENT16) were searched against two public (SwissProt and TrEMBL) and one manually generated protein database (BGPMG) containing entries of sequenced metagenomes from biogas plants.

The findings suggest that the use of the metagenome database BGPMG improves the identification yield for particular metaproteomic samples: regarding the total number of identifications, the results for GENT01 and GENT07 show that BGPMG provides a higher portion of identified spectra than the public databases (Table 4.2). Notably, the overall agreement between the results from three search databases was very low (Figure 4.2): this result indicates that few protein sequences were shared between the original databases. Among all data sets, the percentage of database-specific identifications was the highest for the search of GENT07 against BGPMG by which around 61% of the peptide hits were obtained exclusively. While 2.6 times more peptides were found here for BGPMG than for SwissProt, a study investigating microbial consortia on a oceanic scale reported six times more identified peptides using a metagenomic database in comparison to the search against a public one [92]. Moreover, the highest fraction of unique peptides was observed for the BGPMG searches with an average of approximately 80% (Table 4.3). These latter findings show that BGPMG contains more specific protein targets than the public databases: thus, the metagenome database suffers less from the issue of protein inference ac-

cording to which one peptide is found in multiple proteins. Tanca *et al.* had reported also higher portions of unique peptides in their analyses for experimental databases (96–97%) in comparison to public resources (86–88%) [126]. In general, the probability of obtaining unique peptide hits is higher for more specific (e.g. metagenomic) databases due to a lower sequence similarity when compared to public databases. However, the single use of a metagenome sequence database could be also considered detrimental: searching the data set GENT16 against BGPMG resulted in the lowest fraction of identified spectra (2.0%) in comparison to the higher yield from the searches against SwissProt (4.1%) and TrEMBL (22.0%) (Table 4.2). Due to the high fraction of identified spectra against the latter database, any severe problems originating from the experimental analysis, such as sample preparation or protein extraction, could be excluded. Although the reason for such a low number of matches against BGPMG for GENT16 could not be determined, these findings point to a potential risk for analysis workflows that are exclusively based on metagenomic information: when a sample is searched against a metagenome database that does not fit to the sample under investigation, the chances of successful identification are rather limited: for instance, a different community composition may lead to sequence changes which then reduces the identification yield. While it is recommended to evaluate the fit of the entries in a reference database, this issue might become obsolete when metagenomic analyses are performed routinely along with metaproteomic experiments.

As mentioned above, the public databases provided a significant number of unique identifications that cannot be retrieved when matching the MS/MS spectra against the metagenome reference background. In particular, a high identification yield was observed for GENT16 searched against TrEMBL: 22% identified spectra and 69% of the total number of peptides could be uniquely found here (Table 4.2). From a computational perspective, this result can be explained by a stronger disparity of target and decoy PSM score distributions for TrEMBL when compared to SwissProt and BGPMG (Figure 4.3). In addition, it was also found that only a limited proportion of TrEMBL peptides could be mapped back to SwissProt and BGPMG, underlining the benefit of obtaining unique identifications from public databases. From a biological perspective, the findings might be related to a specific composition of the GENT16 sample that is derived from a biogas plant with process parameters that vary significantly from GENT01 and GENT07: for example, the substrate for GENT16 is based on brewery waste water which is different from the feed in GENT01 and GENT07 (see Section 3.2.1). Moreover, other important process parameters, such as pH value, reactor type and temperature, are different between the BGPs for the respective samples. Overall, the results clearly indicate that the GENT16 sample contains strains which protein sequences are better covered by TrEMBL when compared to SwissProt and BGPMG.

It could be found that public databases, such as SwissProt and TrEMBL, can support the analysis of metaproteomic data for two main reasons: first, they contain protein sequence informa-

tion that can be easily accessed and integrated into any data analysis pipeline. While SwissProt presents a well-curated and relatively condensed resource, mainly unreviewed protein entries are found in TrEMBL. Remarkably, the latter led to a significant number of additional identifications in the investigated data sets. Second, another important advantage of public databases presents the provided annotation of proteins. Manually curated entries provide high-quality meta-information and cross-references to taxonomic origin and functional context. With respect to these latter aspects, further strengths and limitations of public databases will be discussed in Section 5.5 and 5.6.

Summary. Based on the results of the evaluations for the BGP data sets, it is definitely recommended to include metagenome sequences into the data analysis workflow in metaproteomics. However, it is advisable to rely on a customized metagenomic database with exact or at least close origin to the samples under investigation to provide a sequence-specific basis for the protein identification. Since the exclusive use of a metagenomic database increases the risk of missing peptide sequences that are not contained exactly in the database, it is useful to include sequence entries from public databases. While metagenome sequences are commonly unannotated, public resources have the additional benefit of containing well annotated protein entries by which useful meta-information can be retrieved (see Section 5.5 and 5.6). Finally, another important challenge for a metaproteomic analysis workflow concerns the influence of the database size as discussed in the following.

5.2.2 Evaluation of Search Strategies

In a computational benchmark study [287], Jeong *et al.* stated that the impact of the database size is often not addressed by the community, and reported a significant decrease in identifications with an increasing number of protein sequences using the standard ISB [298] and CPTAC [299] data sets. Vice versa, it had been observed in a study with an artificial mixture sample of nine different microorganisms that searching against subsets of a large database led to an increase in PSMs and peptides [126]. However, in this latter metaproteomic study, the reasons for the observed effects had not been examined by the authors in detail. In order to systematically investigate the influence of the database search space on the identification yield in a metaproteomic experiment, exemplary data sets from human intestine metaproteome samples (P1, P23 and P34) were used and three different search strategies were applied: (1) classic searching against a large metaproteomic database (HIMPdb), (2) subset searching against two portions of the latter database (Bact594db and Qin2010db) and (3) two-step searching against HIMPdb.

In general, the results indicate that the identification yield is strongly influenced by the num-

ber of entries in the protein search database. In particular, the data suggest that subset searching against Bact594db and Qin2010db markedly increases the number of database-specific identifications in comparison to classic searching against HIMPdb (Figure 4.5). These findings raise concerns that valuable identifications might be lost in searches against metaproteomic databases that typically contain millions of protein entries. Moreover, it can be observed that two-step searching more than doubled the number of identifications (Figure 4.4). However, the disproportionately high amount of hits in the latter strategy points to an underestimation of the FDR. The hypothesis of an increased number of false positives in the results of two-step searching is corroborated by evaluations of the underlying PSM score distributions suggesting a decrease in identification quality when compared to classic searching (Figure 4.6). These findings are confirmed when employing a method to rescore the PSMs: the identifications from two-step searching show lower RMIC score values indicating less reliable hits when compared to classic searching (Figure 4.7).

The next goal was to put preceding observations in a more general context by performing benchmark analyses on a proteomic PFU sample in three steps: first, classic searching was applied against a targeted PFU-specific database (Pyrodb) corresponding to a common proteomic analysis. Second, classic searching was used against a large and unspecific database (PyroHIMPdb) to simulate a typical database search in metaproteomics. Finally, two-step searching was performed against PyroHIMPdb to evaluate the performance of the method by investigating the quality of the results.

The results of the PFU benchmark investigations confirm previous findings on subset searching against HIMPdb (Figure 4.5): a markedly decreased identification yield can be recognized for classic searching against PyroHIMPdb when compared to Pyrodb results (Figure 4.10 and 4.11). Moreover, the investigation of respective target and decoy PSM scores revealed a much broader overlap of both score distributions for PyroHIMPdb when compared to Pyrodb (Figure 4.12). Although both databases contained the same amount of PFU target sequences, the absolute number of protein entries differed significantly between Pyrodb (around 9 000 sequences) and PyroHIMPdb (over 6 million sequences). These findings indicate a serious issue of the TDA-based FDR estimation: the probability of obtaining decoy hits —representing estimates for false positive identifications—increases with the search space. Consequently, the method of classic searching against large databases is prone to an overestimation of the FDR. This observation is in line with previous proteomic [194, 287] and proteogenomic [196] studies that reported a reduced resolution when increasing the search space in TDA-based identification workflows. Although PyroHIMPdb served only as a representative database for a typical metaproteomic analysis, the observed statistical bias is important for general proteomic workflows as well due to a constantly increasing number of protein sequences in public repositories. Unlike pure culture proteomics,

however, it is difficult to exclude certain proteins in a metaproteomic experiment *a priori* due to the high number of potentially occurring species in the samples. The benchmark analyses further show that two-step searching against PyroHIMPdb resulted in more identifications in comparison to classic searching against Pyrodb (Figure 4.10). Although a significantly increased identification yield of two-step searching indicated an FDR underestimation in preceding results of the HIMP experiments, the score distributions from the PFU experiments did not show any peculiarities (Figure 4.12).

When comparing the performance of the different search strategies, the results unveil striking shortcomings of classic searching against large databases: an FDR overestimation leads to reduced accuracy and resolution of the results. In contrast, the results of subset searching suggest a better performance of the method, since it produced exclusive sets of database-specific identifications. The findings also indicate that the sensitivity of a metaproteomic analysis can be increased by combining results from multiple subset searches. While the method of subset searching was employed for metagenomic (Qin2010db) and genomic (Bact594db) databases, the strategy could be also be applied to public databases: to minimize the negative effects of the database size, the large TrEMBL database could be split for specific taxonomic ranks (e.g. superkingdom or phylum) into taxon-specific subset databases. Analogous to the application of multiple search algorithms (see Section 5.1), however, the use of subset searching is resource intense and the combination of distinct results is technically challenging. While Rooijers *et al.* had previously reported an 2-fold increase of the identification yield for an iterative search method [113], this approach suffers from technical issues by requiring two external BLAST routines in addition to the database searching. In this work, the method of two-step searching was therefore evaluated, as it automatically reduces the protein search space by solely employing conventional database searches [125]. This method can be directly applied for any metaproteomic analysis, however, an excessively high identification yield and low-quality hits in the presented results points to an underestimation of the FDR. Based on these findings, the application of two-step searching can only be recommended when using a stringent FDR filtering to minimize the number of false positive hits. Accordingly, Gupta *et al.* recommended to disable the second pass feature in X!Tandem that may lead to the underestimation of the FDR due to insufficient compliance with the TDA [300].

Summary. In conclusion, the path of finding an optimal search strategy in metaproteomics is paved with various obstacles. A general recommendation is to downsize the database as much as possible by retaining only the protein taxa of interest to decrease the risk of losing identifications when the database size is increased [301]. Due to the unknown sample composition in metaproteomic experiments, however, it is not feasible to exclude particular sequence entries from the protein database without sacrificing potentially valuable hits. Therefore, the use of subset searching against multiple database variants is recommended to increase the resolution for complex metaproteomic and proteogenomic search spaces. The findings indicate that the method of two-step searching should be used with utmost caution to avoid the danger of underestimating the FDR. Finally, it is important to critically evaluate the use of TDA-based FDR estimation due to the significantly decreased identification yields for searches against large databases. To circumvent the latter issue, solutions using a decoy-free result validation, for instance, based on mixture modeling [181, 302, 303] or machine learning [193], might be an alternative. Moreover, an improved procedure for obtaining more accurate FDR estimates has also been proposed recently [304]. In the next part, the impact of other parameters, including missed cleavages and enzyme specificity, on the outcome of metaproteomic analyses will be discussed.

5.2.3 Missed Cleavages and Enzyme Specificity

Throughout this work, the parameter for the maximum of allowed missed cleavages was set according to general recommendations for database searches in proteomics [305, 306]. Since tryptic digestion works rarely perfect in a proteomic experiment [133, 134, 135], it is useful to choose the corresponding MC parameter values above zero. However, it can be expected that this search parameter affects the results, as higher MC parameter values increase the search space of generated *in silico* sequences during the peptide spectrum matching process. So far, the effect of this parameter on the identification yield has not been addressed by any metaproteomic study in detail. Therefore, MC parameter values between zero and three were evaluated for database searches of three exemplary HIMP data sets (P1, P23 and P34). In addition, the same MC parameters were tested in the benchmark evaluation for the PFU sample.

For each of the HIMP result sets, the somewhat counter-intuitive effect of retrieving lower numbers of identifications in total was observed for increasing MC parameter values. It was found that a parameter value of $MC = 0$ resulted in the highest number of identified PSMs and peptides at 1% FDR , whereas $MC = 3$ yielded the lowest number of hits (Figure 4.8). Remarkably, the number of identified peptides was more similar for all MC parameter values at 5% FDR compared to 1% FDR (Figure 4.9a). Moreover, a considerable number of exclusive identifications was found for $MC = 0$ at all tested FDR levels (Figure 4.9b). In contrast, results for the proteomic PFU

benchmark experiment indicate a clear increase in identifications towards higher MC parameter values (Figure 4.13 and 4.14a). The contradictory findings between the metaproteomic (HIMP) and proteomic (PFU) results can be explained—similar to previous observations—by the search space complexity of the target databases that is several times higher for the metaproteomic experiments. Hence, it can be suggested that the decreased identification yields for MC values above zero in the HIMP searches are a result of an FDR overestimation. While the issues for the TDA-based FDR estimation prevail in the investigations on the HIMP data, the PFU analyses show the expected behavior of an increased identification yield for higher MC parameter values. In this case, the PFU results also show that significant proportions of exclusive peptides with multiple missed cleavages were found (Figure 4.14b). From a biological perspective, this finding could be expected since *Pyrococcus furiosus* presents a hyperthermophile bacterium that synthesizes various proteins resistant to an enzymatic digest. Nevertheless, the number of exclusive peptides in the PFU searches was the highest for *MC = 0* among all tested values. From a computational perspective, this result can be attributed to the preference of the search algorithms for longer peptides: for *MC = 1*, one spectrum can be explained by two peptide variants, thus a shorter sequence with no missed cleavage and a longer sequence with one missed cleavage. Since the former receives a lower score due to a lower number of matched fragment ion peaks in comparison to the latter, only the longer peptide hit is considered for *MC = 1*. Consequently, a high number of exclusive peptide hits for *MC = 0* is replaced by longer peptides for *MC = 1*.

The next evaluated parameter was the choice of the cleavage enzyme, since various proteomic studies emphasized—despite the high specificity of trypsin [132]—the occurrence of non-tryptic peptides [133, 134, 135, 307]. In addition, it was assumed that the presence of various proteases in the human intestinal tract may lead to non-tryptic protein fragments. To increase the overall peptide detectability, semi-tryptic or non-tryptic searches can be employed [308]. In the investigations, *tryptic* and *semi-tryptic* were used as search settings, and the number of PSMs and peptides obtained from separate searches with these parameters were evaluated.

Depending on the chosen FDR threshold, it was found that the highest number of identifications could be retrieved using either *tryptic* (Table 4.5) or *semi-tryptic* (Table 4.6) as cleavage parameter. The additional gain of using *semi-tryptic* as search option was rather minimal with approximately 7% exclusive peptides at 1% and 5% FDR. In the PFU benchmark analyses, significantly fewer peptides were detected for *semi-tryptic* in comparison to *tryptic* (Figure 4.13). These findings stand in contrast to the results of another metaproteome study focusing on host proteins in which the authors reported a similar number of tryptic and semi-tryptic peptides [115]. The benchmark results indicate that *semi-tryptic* is not recommended to be applied as the only cleavage parameter. Notwithstanding considerably higher running times, however, *semi-tryptic* searches can still be used to complement results from *tryptic* searches. Thereby, the data pro-

cessing can be accelerated by methods effectively reducing the computational time of the search algorithms, such as aforementioned cluster computing approaches [289, 290] and algorithms predicting either peptide truncatability [309] or tryptic cleavage [310].

Lastly, the evaluation of *chymotrypsin* and *pepsin A* as cleavage parameters resulted in a very low number of identifications. Since trypsin was used as cleavage enzyme in the experimental setup, it can be assumed that non-tryptic peptides were low abundant in the samples. Consequently, it could not be demonstrated that particular gastrointestinal enzymes were active in the investigated samples.

Summary. The findings indicate that increasing the MC parameter values impairs the identification yield due to the increased search space for metaproteomic data. While the use of up to four missed cleavages has been reported in a metaproteomic study [311], lower parameter values are recommended based on the observed results. Similar to the previously discussed combination of searches against different databases, a combination of multiple searches with varying MC values between $MC = 0$ and $MC = 2$ can be used to increase the specificity and sensitivity. Consequently, an automated integration of database search results for different MC parameter values might be useful in the future to improve the overall reliability of results in metaproteomics. Moreover, the results demonstrated that *semi-tryptic* may not be sufficient as exclusive parameter setting, but was able to complement the results for the tryptic cleavage parameter. Finally, the activity of any suggested gastrointestinal enzymes could not be confirmed by the results from the HIMP samples, which is probably due to their low abundance in comparison to trypsin.

5.3 Testing Performance of *De Novo* Sequencing

In the last section, various obstacles of metaprotomic data analysis with respect to database searching and statistical validation of the identifications were discussed. As an alternative to the database-driven methods regarded in the last sections, *de novo* sequencing bypasses the following two major problems at once: by not relying on a protein database, this technique circumvents the issue of missing sequence information and also the statistical bias caused by the complex search spaces of metaproteomic databases. However, a mapping of *de novo* sequences to the protein level is eventually required to infer taxonomic and functional information. While various studies applied *de novo* sequencing for the analysis of metaproteomic data [12, 199, 312, 313, 314], the method has not been contrasted to database searching in detail. For this purpose, the approach was applied for the HIMP10 data and the identified *de novo* peptide sequences were compared to previously identified hits from database searching. To obtain protein identifications, the same *de novo* peptides were also mapped to a protein sequence database (HIMPdb).

The results demonstrate that *de novo* sequencing insufficiently recovers hits from database searching: only 23% of all database search peptides could be obtained using *de novo* sequencing at a score threshold of $S = 100$ (Table 4.7). Since this ratio increased only to 25% when no score threshold was applied, the identification quality seems to decrease drastically below a score threshold of $S = 100$. It can be further observed that the percentage of identified spectra increased significantly from 23% to 60% when the score threshold was decreased from $S = 100$ to $S = 50$. This atypically high fraction of identified spectra suggests that such a low score threshold is not an appropriate parameter choice, since its application would considerably increase the number of false positive identifications. This analysis also reveals a major issue related to the applicability of the scoring scheme as error control mechanism: it seems impractical to set a fixed cutoff value for such an arbitrary score.

Furthermore, the results show that significantly less *de novo* peptides could be mapped to the *in silico* digested protein database in comparison to database searching at 5% FDR (Table 4.1). While it can be also observed that over three times more *de novo* peptide sequences could be matched against the database when decreasing the score threshold from $S = 100$ to $S = 50$ (Table 4.7), these findings indicate that the overall cost of using *de novo* sequencing and mapping of peptides to proteins exceeds the benefit of potentially gaining additional identifications. Moreover, in accordance with preceding observations, the undesired side effect of adding false positives is inherently given by this strategy. The error control is difficult in this case, as the target-decoy approach cannot be directly applied to the method of matching *de novo* peptides against an *in silico* digest of a protein database.

In addition, *de novo* sequencing was used to check for the quality of database search hits: when comparing the results from classic and two-step searching, a clear trend towards more low scoring peptide identifications can be observed in the latter. While the results indicate that most overlapping peptides between *de novo* sequencing and both database searching techniques obtained a high score, the number of low-scoring identifications was significantly increased for two-step searching when compared to classic searching (Figure 4.15). Furthermore, a clear distinction can be found between distributions of *de novo* peptide scores for two-step searching in comparison to classic searching (Figure 4.16). Consequently, these observations confirm previous results that attested results from two-step searching a lower identification quality (see Section 5.2.2). These investigations suggest that *de novo* sequencing could be regarded as method for estimating the quality of doubtful peptide identifications derived from database searching. However, due to high computational effort and limited statistical validation, *de novo* sequencing is yet impractical to be used as quality control mechanism for high-throughput studies in metaproteomics.

An important prerequisite for the application of *de novo* sequencing is to establish an *a posteriori* error control, since relying on the scoring of the *de novo* technique involves the risk of

including many false positives. In fact, the actual benefit of using *de novo* sequencing can only be estimated by investigating the quality of suggested peptides sequences in detail. To that end, the software DeNovoGUI assists the user with a visualization of *de novo* sequences and corresponding fragment ion peaks for each considered spectrum [275]. The tool also provides the opportunity of exporting the results and using an external BLAST feature for mapping the sequences onto the protein level. Another promising tool presents PepExplorer which aligns result sequences from various *de novo* algorithms against a target-decoy database [205, 315]. The software employs pattern recognition and gives the user the option to filter the results based a certain FDR threshold.

While the application of *de novo* sequencing in high-throughput studies was also limited due to high running times of the algorithms, the latest release of an algorithm featuring the real-time processing of hundreds of spectra per second might increase the popularity of the method [316]. Together with advances in mass spectrometry instrumentation and resolution, *de novo* sequencing might contribute significantly to the identification of unknown peptides in future proteomic analyses. With respect to the specific application in microbial community proteomics, Cantarel *et al.* reported that *de novo* sequencing adds valuable information to the analysis even if protein sequence information is not available or incomplete [199]: by the combined use of PEAKS [201] and PepNovo+ [200], 754 non-redundant proteins were identified that had not been found with conventional database searching using SEQUEST [168]. However, it was not evaluated in this study whether any additional taxonomic or functional information could be obtained by the use of *de novo* sequencing.

Summary. Based on these findings, it can be concluded that the overhead of using *de novo* sequencing as complementary method to conventional database searching is currently higher than the benefit. For the investigated metaproteomic data, only a moderate fraction of database search results could be recovered when using the *de novo* sequencing algorithm PepNovo+ and the gain of mapping *de novo* sequences to an *in silico* digested protein database could also not justify the involved efforts. Moreover, *de novo* sequencing still requires more reliable error control mechanisms and further algorithmic improvements concerning the accuracy of the obtained peptide sequences. Particularly due to these limitations, *de novo* sequencing has not been widely employed in the community yet. In the context of metaproteomics, it also remains to be demonstrated whether the results of *de novo* sequencing provide additional insights regarding protein function or taxonomic assignment.

5.4 Generating Meta-Proteins by Protein Grouping

Various strategies have been proposed for solving the issue of protein inference [11]: while some approaches rely on a fixed set of PSMs propagated by the search algorithms [317, 318], other probabilistic methods reassess the identifications and also seek for homologous hits [319, 320, 321, 322]. A common solution presents a parsimony-based strategy that attempts to explain the minimum amount of proteins from a given set of peptide hits [262]. In contrast to pure-culture proteomics, however, the data analysis of microbial communities usually involves databases which contain a large collection of homologous proteins across species from different organisms. Besides the fact that the inference problem cannot be easily solved in the presence of these complex communities, such an anticipating algorithmic solution may also not be useful for the typical metaproteomic analysis workflow: a result set with as many proteins as possible effectively preserves the maximum amount of information with respect to taxonomic composition and functional potential of the sample. For this purpose, the MPA software provides the meta-protein generation feature that allows to group proteins according to particular rules.

Three different grouping rules for the meta-protein generation were implemented in the MPA software (see Section 3.1.2). The comparative evaluation of these rules applied to BGP protein result sets indicates that each rule has a profound effect on the number of protein groups (Figure 4.17). *Minimum One Shared* displays the highest reduction of proteins, while *All Shared* results in a weaker grouping effect. The latter is more conservative, because it combines only proteins with at least a common subset of multiple peptides. Accordingly, the findings from the evaluation of protein grouping for multiple data sets suggest that for *Minimum One Shared*, the majority of meta-proteins was linked to more than one peptide, while *All Shared* contained more than 50% single peptide assignments (Figure 4.25).

In all investigations, a more stringent threshold of 1% FDR in comparison to 5% FDR leads to a stronger effect of protein grouping resulting in more assigned peptides per group. These observations can be explained by the circumstance that peptide-centric grouping rules profit from more reliable peptide identifications leading to fewer conflicts during the protein assignment. Conversely, the results were less affected when using the UniRef-based grouping rules: these rules are based on the sequence similarity of the proteins in the database and are therefore less susceptible to changes of the PSM-based FDR [263]. Although the UniRef-based grouping methods show the lowest relative reduction in the regarded results (Figure 4.17), their advantage is to rely on clustered sets of protein sequences from UniProtKB: while a single protein may feature a limited amount of information, more biologically relevant data, such as protein function or cross-references, can be provided by other members of the corresponding UniRef cluster. Thus, the UniRef-based protein grouping might also achieve a gain of semantic information by

reducing the final result set.

For *Minimum One Shared* and *All Shared*, the Levenshtein edit distance was evaluated as parameter that permits sequence variations, such as point mutations, insertion or deletions, in the grouping process at the peptide level (Figure 4.18). In this case, a higher ED parameter value leads to increased protein grouping. Since *Minimum One Shared* is based on a single peptide, a higher ED parameter value results in a stronger grouping effect for this rule compared to the *All Shared* rule. Although the total reduction for *ED = 2* relative to *ED = 0* was less than 10% for both grouping rules, the mutation-tolerant strategy considers amino acid substitutions that frequently occur in sequences of metaproteome samples. However, the Levenshtein metric has the limitation that it only reckons the similarity of two sequences without considering the biological background: for instance, a hydrophilic amino acid is more likely to be replaced by another hydrophilic residue than to be transformed into a hydrophobic variant. Therefore, the mutation-tolerant protein grouping could be extended by using substitution matrices that are based on evolutionary models. Consequently, the use of amino acid matrices such as PAM [323] or BLOSUM [324] would enable to assess the probability of sequence transformations.

The last evaluated method presents the *Taxonomy Rule* which can be employed as additional filtering step for other grouping rules to set the maximum taxonomic level at which proteins converge. A maximum increase of 8% in the number of total meta-proteins at the evaluated taxonomic convergence levels suggests that the *Taxonomy Rule* can be used efficiently to control the microbial diversity (Figure 4.19). Thereby, it allows to refine the results of a metaproteomic analysis by classifying the protein groups based on the phylogenetic hierarchy.

Furthermore, the reproducibility of protein groups between results from replicate samples was evaluated for *Minimum One Shared* and *All Shared*. It was found that *Minimum One Shared* led to the best performance in the correlation analysis that compared meta-proteins on the basis of their spectral and peptide count (Figure 4.22). When picking out a particular protein group, it could be observed that *All Shared* failed to merge proteins which belonged together according to manual inspection. These findings indicate that *Minimum One Shared* was more reproducible and less error-prone than *All Shared*.

Moreover, it can be recognized that the generation of meta-proteins using *Minimum One Shared* rule increases the average similarity between different result sets (Figure 4.24a). It was also found that meta-protein sets contain more shared peptides than protein sets (Figure 4.24b) which consequently reduces the amount of groups to be compared. As the number of comparable groups is decreased after the meta-protein generation for each result set, different samples can be better compared and quantified than in unprocessed protein result sets.

Overall, the described protein grouping methods are tailored towards metaproteomic data, because they feature particular rules for allowing mutations and setting taxonomic convergence

limits. In addition, the grouping of proteins might be applied for proteomic analyses in general to tackle the protein inference problem. On the one hand, these approaches are directly applicable to large data sets, on the other hand, they only rely on the confidence of the assigned spectra and peptides. At the protein level, further features such as occurrence rate, sequence coverage and protein length are important. Therefore, the current approach lacks a qualitative confidence estimate for the generated meta-proteins. For instance, a recently developed software called ProteinInferencer assesses the confidence issue by maintaining a controlled protein FDR [325]. Another problem arises at the quantitative level: the calculation of label-free quantification measures, such as NSAF [239], is often based on a defined sequence length. As a consequence, the calculation of such quantitative measures is frequently biased by partial protein sequences in the databases. In metaproteomics, the high sequence similarity between conserved proteins of related organisms makes matters worse. Consequently, problems are forwarded to the meta-protein level, when measures for the quantification of protein groups are calculated on the basis of the assigned protein hits. In the MPA, this issue is addressed by providing the feature of so-called aggregate functions: for instance, the average of all protein NSAF values can be calculated for each meta-protein within a sample. This method has the advantage that it can be directly employed for the label-free quantification of meta-proteins. However, an essential improvement to this straightforward approach could be a strategy which also takes the high number of shared peptides between homologous proteins into account [326].

Summary. In this work, the meta-protein generation is presented to reduce the redundancy of very large protein sets and also to preserve the diversity of the protein species in metaproteomic analyses. Evaluating different rule-based protein grouping methods, it was found that the approach of grouping a set of proteins based on one shared peptide (*Minimum One Shared*) shows the highest reduction of the protein result set. Besides tackling the redundancy issues, a refined analysis of metaproteomic data is established by additional grouping features, including mutation-tolerant grouping and taxonomic diversity preservation. In addition to successful grouping of proteins within one result set, these findings indicate that the comparability of results from replicates and completely different experiments can be improved by using meta-proteins instead of proteins as common data basis. While meta-proteins have been directly quantified by the peptide and spectral counts in these investigations, it is recommended to rely on more accurate measures reflecting the varying sequence length and high sequence similarity within the generated protein groups.

5.5 Investigating Techniques of Taxonomic Assignment

The focus of Section 4.5 was the assignment of identifications to particular taxonomic ranks. First, the influence of the protein database on the retrieved taxa was elucidated by analyzing representative metaproteomic data from BGP samples. Moreover, the MPA software was compared to another metaproteomic analysis tool with respect to the taxon-specific peptide assignment ratio. Subsequently, ground truth data from a microbial mixture sample of nine organisms was used to evaluate the performance of both tools. In the last part of this section, results from HIMP samples were classified according to their taxonomic composition.

5.5.1 Influence of Protein Database

Previous investigations in this work revealed that the choice of the protein database significantly affects the identification yield (see Section 4.2). These findings were the rationale to investigate the impact of this parameter on the taxonomic assignment process in more detail.

The analysis carried out on data sets from three exemplary BGP samples reveals a strong effect of the protein database on superkingdom-specific assignments: for instance, the proportion of peptides from GENT16 assigned to Archaea increased from 37% in SwissProt to 62% in TrEMBL (Figure 4.26). In contrast, the eukaryotic assignments dropped from 25% in SwissProt to 9% in TrEMBL. While these differences occur for GENT16 with an almost four times increase in identifications for TrEMBL, a strong influence of the protein database on the taxonomic assignments can be recognized across all evaluated data sets. These findings can be explained by the diverging content of the database variants in UniProtKB: SwissProt contains curated information with a particular focus on eukaryotic proteins from clinical strains, whereas TrEMBL provides a large number of non-curated entries including a significant proportion of proteins from Archaea and Bacteria. These findings also suggest that the combined use of SwissProt or TrEMBL could provide more taxon-specific annotations for the protein results than the use of only one of them. Although such a database combination increases processing time and overhead for the data integration, it improves the resolution of the taxonomic assignment process. Analogous to guideline statements made in Section 5.2, it is generally recommended to include multiple databases into a data analysis workflow for metaproteomics. However, it should be also considered that SwissProt features manually reviewed annotations in contrast to automatically annotated sequences in TrEMBL (see Section 3.3.1): thus, more comprehensive and reliable meta-information can be expected by the protein annotations in SwissProt.

Another important finding relates to the unique taxonomic distribution observed for each BGP result set. For instance, the results at the phylum level show that GENT16 provided more

Euryarchaeota and fewer Firmicutes assignments than the other investigated BGP data sets. An explanation can be that the BGP samples varied in their microbial composition due to differing process parameters as already discussed in Section 5.2.1. For instance, reactor type and substrate feed of GENT16 were significantly different in comparison to GENT01 and GENT07 (see Section 3.2.1). Furthermore, the temperature was different for each of the three BGPs: a long-term study on a continuously operated BGP demonstrated that the process temperature is an important factor that affects microbial activity and community composition in anaerobic digesters [102]. A desired goal of the BGP analysis is to link metaproteome results with such external parameters to determine novel relationships between the community and its environment [93]. While early metaproteomic studies resulted in the detection of only few proteins for the analysis of sludge [12] and BGP [13] samples, latest developments in sample preparation and instrument technique led to a high amount of identifications that render such a taxonomic analysis possible in the first place [101, 131, 327].

Summary. The findings show that the chosen database has a significant influence on the assigned identifications at the taxonomic levels of superkingdom and phylum. While the microbial community structure had been shown to remain relatively stable in a single BGP under changing environmental conditions over time [102], it was found in this work that each sample from a different fermenter showed a characteristic taxonomic distribution. While the aim of these preliminary investigations on a limited number of three data sets was to highlight the factors influencing the taxonomic data analysis of metaproteomic samples, several comprehensive studies investigated the composition of microbial communities for BGP samples in more detail [101, 131, 327]. The latter studies also showed that sequence information derived from public and metagenomic databases are equally important for the successful taxonomic assignment of proteins.

5.5.2 Assignment Performance Evaluation

Next, the performance of the taxonomic assignment was assessed for the MPA software in comparison to the metaproteomic analysis software Unipept [276, 277]. For this analysis, the same BGP result sets from GENT01, GENT07 and GENT16 were used as in the previous investigation.

In comparison to the peptide-centric Unipept software, it was found that the protein-based MPA application provided increased proportions of taxonomic assignments across the data sets and taxonomic ranks for the majority of the results. In particular, the relative amount of SwissProt assignments was considerably higher in the MPA software when compared to Unipept (Figure 4.28). For example, the average portion of assigned peptides accounted for 78% in Unipept at the superkingdom level for SwissProt, but the MPA could assign more than 98% of the iden-

tifications. While few ranks in the GENT16 data set could be better resolved for TrEMBL when using Unipept, the analysis shows a higher overall performance of MPA.

While the preceding findings suggest that the MPA can be recommended for the taxonomic assignment of metaproteomic data, it was not possible to check errors of assignment process due to the unknown microbial composition of the BGP samples. Therefore, ground truth data from a mixture sample containing nine particular species was used in the next analysis to assess the exact assignment performance of MPA and Unipept. In the original study by Tanca *et al.* [126], the identifications from database searching were classified into correct and incorrect taxonomic assignments based on the knowledge about the actual origin of species within the microbial mixture sample. In addition, a taxon significance threshold was used to filter for taxon-specific assignments that were provided in considerable amounts. When analyzing two microbial mixture replicate data sets from the above mentioned study, the accuracy and precision of the taxonomic assignment methods could be assessed for the tools Unipept and MPA.

In line with the findings of the previous analysis on the BGP data, a markedly increased number of taxon-specific peptide hits was found for MPA (Figure 4.29) in comparison to Unipept (Figure 4.30). The results also show that increasingly more correct assignments are available when the taxonomic rank is increased from species over genus to family. While the amount of incorrect assignments is then proportionately increased, the application of the described taxon significance threshold can reduce or remove false attributions completely for MPA (Figure 4.29). Conversely, for Unipept, even a filter threshold of up to 5% could not avoid wrong assignments at the rank of species. The findings also confirm that the threshold value of 0.5% for Unipept was well chosen by Tanca *et al.* in their study, since it filters out the great majority of incorrect identifications (Figure 4.30). In addition, the comparison of results between 1% and 5% FDR shows that—in particular for MPA—a higher amount of correct assignments can be gained by a less stringent FDR threshold. While more incorrect assignments are retrieved, applying the taxon significance threshold can also be used to increase the accuracy. Consequently, these findings suggest to apply the described threshold instead of relying solely on the FDR threshold.

As alternative to the LCA-based taxonomic assignment [207, 328], the MST method was developed to preserve the peptide-level specificity. Based on the results of the microbial mixture data sets, the relative proportion of correct taxon-specific peptide identification were compared between LCA and MST in the following.

The results indicate a trend that the developed MST method obtains more correct taxon-specific peptide assignments than the conventional LCA approach (Figure 4.31). These findings were consistent between both regarded replicate data sets. Since the LCA-based peptide assignments to high-level taxa are often caused by conserved sequences, a noticeable bias is imposed. In that case, the taxonomic resolution is reduced and strains or related species cannot be dis-

tinguished anymore. On the contrary, the idea of the MST method is to preserve the peptide specificity and by resolving differences between closely related organisms. This approach follows a similar principle as methods that only use unique peptides being specific for a single organism, as conducted in previous metaproteomic studies [113, 329]. However, the exclusive use of unique peptides does often not provide enough confidently identified peptides to discriminate between particular organisms due to low individual species coverage in metaproteomic experiments [10]. As a compromise, MST therefore also includes shared peptides for the taxonomic assignment at the protein and meta-protein level. Consequently, unique peptides are decisive for the species- or strain-level specificity, while shared peptides contribute to the confidence of the taxonomic assignment. Eventually, common issues of metaproteomic data analysis, such as high sequence similarity and bias in the statistical validation of identifications also affect the MST method. As potential extension of this approach, the Pipasic algorithm by Penzlin *et al.* uses a similarity and abundance correction strategy to identify and quantify identifications at the species level [326].

Summary. When comparing the overall performance between MPA and Unipept, the MPA is more successful with respect to the total number of taxon-specific peptides, because significantly more identifications could be assigned to the examined taxonomic levels. Furthermore, the application of a taxon significance threshold improves the results at elevated FDR levels: consequently, the number of false assignments is decreased, while the total amount of correct assignments is increased in comparison to more stringent FDR thresholds. Furthermore, the findings indicate that MST has a slightly better taxonomic assignment performance than LCA by resolving species-related differences at the peptide level.

5.5.3 Phylogenetic Overview on Human Intestine Microbiota

In the next analysis, the MPA software was used to process and analyze MS/MS data from HIMP samples of 29 obese and non-obese individuals. The entire taxonomic analysis of the identifications was performed by the supplementary use of Unipept [276], since the majority of the results contained protein identifications from unannotated sequences originating from HIMPdb. The exported results from MPA and Unipept provided the basis for a detailed study in which the HIMP samples were compared based on the abundance of bacterial and host proteins at the taxonomic and functional level [266]. While samples from a lean and an obese adolescent had been analyzed at the proteome level [108], metaproteomic analysis on a representative cohort of adult individuals has not been performed before. In this work, the analyses of the original study [266] were complemented by investigating a subset of ten HIMP data sets (HIMP10) to obtain a general phylogenetic overview on the community composition in human gut samples.

The first analysis involved the phylogenetic classification of the hits at the superkingdom level. It was found that the vast majority of the identifications was assigned to bacterial origin. The results show that an average portion of approximately 86% identified spectra and 89% peptides were matched to Bacteria (Figure 4.32). On the contrary, only 7.5% peptides and around 10% identified spectra originated from eukaryotic taxa. Although the literature pronounces the bacterial predominance in human intestine samples [42], it was also demonstrated that food and host proteins are highly relevant in the context of the interaction between host metabolism and gut microbiota [330].

The following investigation on bacterial phyla shows that the results are mainly in line with previous findings reported on the taxonomic bacterial distribution in the human intestine. In detail, it was found that Firmicutes (69.6%) and Bacteroidetes (21.3%) are the most abundant taxa in the faecal samples, while Actinobacteria (7.6%) and Proteobacteria (1.5%) are less frequently found (Figure 4.33). While the observed distribution corresponds well to findings of previous studies on the composition and diversity of the human gut [105, 331, 332, 333], different fractions were reported by Kolmeder *et al.* [15]: in comparison to results in this work, more Actinobacteria (33%) but lower amounts of Firmicutes (60%), Bacteroidetes (6%) and Proteobacteria (0.2%) were reported in their work. The differences in abundances might be related to the fact that only three subjects at two different time points were regarded in the latter study.

To gain a better resolution in the taxonomic analysis, the taxonomic rank was lowered by counting the identified peptides at the genus level for each of the previously described phyla.

While Firmicutes resulted in a balanced distribution on the observed genera, Bacteroidetes display either a high abundance of *Prevotella* and low abundance of *Bacteroides* or vice versa in between the samples (Figure 4.34). Although the results did not indicate any correlation to the obese or lean group, determining factors with more direct influence on the microbial composition may play a role here: a study by Wu *et al.* reported a classification of enterotypes based on levels of *Bacteroides* and *Prevotella* [334]. While the *Bacteroides* enterotype can be associated with a diet rich in animal and protein fat, *Prevotella* is linked to subjects with a carboydrate and fiber rich diet. Moreover, a study between African and European children revealed that dietary habits influence significantly the *Bacteroides*/*Prevotella* balance of human gut bacteria [335].

Summary. The phylogenetic overview analysis, which was performed by the combined use of MPA and Unipept, confirms previous findings on the microbial composition of human intestine samples: the largest proportion of the bacterial peptide assignments related to Firmicutes and Bacteroidetes. While no indications can be observed that particular taxa were significantly more abundant in either the obese or the non-obese group of samples, a potential connection can be guessed between dietary factors of the subjects and microbial abundance of the genera *Bacteroides* and *Prevotella* in respective samples. However, additional information regarding diet and lifestyle habits would be required to accurately link these results to traits and conditions. In the next section, the assignment of metaproteomic data to protein function and metabolic pathway will be discussed.

5.6 Assessing Methods of Functional Analysis

The investigations of Section 4.6 dealt with the evaluation of functional analysis methods for metaproteomic data. First, the MPA software was applied for mapping the results of BGP metaproteomic data to functional ontologies, enzyme classifiers and metabolic pathways. In order to quantify the functional profile of metaproteomic samples, the reproducibility of the functional assignments was tested between replicate samples in the next step. Finally, by using the protein identifications of the HIMP samples, a procedure was evaluated for assigning results which contain unannotated sequences (e.g. from a metagenome database) to functions.

5.6.1 Elucidating Functional Annotation Methods

The detailed analysis of BGP result sets shows that several functional terms and enzymes from the methane production process could be found. The analysis of the ontology *Biological Process* reveals that "Methanogenesis" was the functional term with the most identifications for GENT01 and GENT07 (Figure 4.36). The high abundance of processes and enzymes involved in the production of biogas and methane had already been observed in metagenome and metaproteome studies on BGP samples [5, 13, 100]. In contrast, the ontological term for the biogas production was less frequently found in the results of GENT16. As discussed previously (see Section 5.2.1), an explanation for this suspicious event could be that the GENT16 sample may deviate from GENT01 and GENT07 in its microbial composition due significantly different process parameters (see Section 3.2.1): although methanogenesis is important without any doubt, the identified proteins for GENT16 might be less well annotated due to their origin from TrEMBL. In addition, a significant number of identifications could be assigned to "Glycolysis" for all three data sets. In this case, an explanation can be that various enzymes from glycolysis catalyze the break down of

carbohydrates in the BGP fermentation process [93].

The results for the ontology *Molecular Function* show that the term "Acetyltransferase" representing an important enzyme in the acetoclastic pathway is detected only for GENT01 and GENT07 in significant amounts (Figure 4.37). On the contrary, the ontological keyword "Transferase" features the highest number of assigned peptides in all three data sets. In this case, using the plain UniProtKB keyword ontology has the limitation of concealing the hierarchy of the controlled vocabulary, meaning that the relationship between proteins and ontological terms is often *one-to-many*: in the BGP result sets, several "Transferase"-specific proteins might also belong to "Acetyltransferase". Another disadvantage of the use of ontological keywords is that the terms are necessarily subjected to manual review [336]. In particular, most of the protein entries in TrEMBL have been computationally annotated by using inference from sequence similarity: this procedure led to a significant level of misannotation in this and other public databases as reported by Schnoes *et al.* [337]. Finally, the data analysis workflow developed in this work currently lacks the feature for an enrichment analysis to identify significantly under- or overrepresented functional terms [338].

Beyond the ontology-based analysis features, the MPA software provides a finer granularity of describing enzyme-catalyzed reactions for the identified proteins using the EC nomenclature. The results show that several relevant enzymes from the methane producing pathway could be identified in the three BGP data sets (Figure 4.38, 4.39 and 4.40). For instance, it was found that "Coenzyme-B sulfoethylthiotransferase" (EC 2.8.4.1) was the enzyme with the highest number of assignments. This enzyme is also called methyl-coenzyme M reductase and catalyzes the final reaction in the biogas production process. Since previous studies had successfully correlated the abundance of methyl-coenzyme M reductase genes and mRNA transcripts to methanogenic activity by means of measured methan production [339, 340, 341], it was speculated that this key enzyme might serve as predictive biomarker candidate for an early detection of process disturbances inside BGPs [93, 102]. Yet, it is unclear, whether this approach would have significant advantages over conventional BGP process monitoring methods. In line with previous findings on "Glycolysis"-specific peptide hits, the protein "Glyceraldehyde-3-phosphate dehydrogenase" (EC 1.2.1.12) was found in all data sets: this enzyme catalyzes the sixth step of glycolysis and might serve to degrade carbohydrates which are the major substrate in BGPs. While the EC nomenclature serves to obtain an overview on the potential enzymatic activities in metaproteomic samples, EC numbers do not provide a wider context of protein-substrate interactions in functional networks. For this purpose, pathway databases are more useful, since they fully describe the biochemical reactions in which enzymes are involved.

The KEGG mapping feature within the MPA software allows to directly transfer protein identifications into metabolic pathways. By applying this function to the GENT01 result data, the

conversion of acetate and carbon dioxide to methane could be directly visualized (Figure 4.41). Moreover, a separation at the level of superkingdom via the taxonomic view in the tool enabled to display the specific activity of Archaea in methanogenesis and the predominance of Bacteria in the glycolysis/gluconeogenesis pathway. The latter application emphasizes how information at the taxonomic level can be combined with the functional annotation of proteins in the MPA software. Moreover, the integrated graph database system serves to provide answers to user-defined queries by profiting from the connectivity of such meta-information.

Similar to previous observations, it was found that the outcome of the functional analysis was affected by the chosen database. Remarkably, particular important enzymes could only be observed in the results from either SwissProt or TrEMBL. Thus, the findings indicate that complementary searches against multiple databases are beneficial for the identification and quantification of functional ontologies and enzymes.

Summary. The described automated integration of semantic information reaching beyond protein identification is valuable to investigate functional aspects of metaproteomic data. When analyzing three exemplary BGP samples, typical functional terms and enzymes of the methane production process were identified. The findings also show considerable sample-specific differences that might be attributed to varying BGP process parameters. Even more importantly, since the quality of protein annotations differs significantly between SwissProt and TrEMBL, the functional analysis is also influenced by the chosen database. Besides identifying and quantifying enzymatic key players, the functional role of the microbial community can be studied by integrating the results of a metaproteomic analysis into metabolic pathways. By using the MPA for the automated mapping of identified proteins into KEGG pathways, important enzymes could be visualized inside the context of the methanogenesis metabolism. Furthermore, the benefit of combining phylogenetic information with pathway data was exemplified by demonstrating the methanogenesis-specific occurrence of Archaea proteins in the data. It should be finally noted that several metaproteomic studies on BGP samples applied the functional analysis features of the developed software [131, 101, 327, 102].

5.6.2 Quantifying the Functional Profile

The functional classification of metaproteomic data frequently involves multiple samples from different conditions or time points. However, the comparison of functional profiles from such diverse samples demands that reproducible quantitative measures have been defined. In proteomic literature, absolute and relative quantification strategies with labeling and label-free techniques have been described [234, 233, 232]. In metaproteomics, label-free techniques, such as spectral counting are often used to circumvent experimental issues occurring during labeling of complex samples [236, 342]. Hence, the reproducibility of functional assignments was evaluated next between result sets from replicate samples using the number of identified spectra, peptides, protein and meta-proteins as quantitative units. The ultimate goal was to assess the applicability of these measures for the functional analysis of results from different metaproteomic samples or experiments.

It could be observed that all investigated quantitative measures were highly reproducible between the technical replicate data. When correlating the number of assignments to the functional ontologies *Biological Process* and *Molecular Function*, very strong correlations (0.96-1.0) were observed for the four quantitative measures across the replicate data sets from GENT01 and GENT16 (Figure 4.41). Overall, the use of meta-proteins resulted in the highest correlation values in comparison to the other quantitative measures. Furthermore, the comparison of results from different BGP data sets for the biological process of methanogenesis indicated that spectra and peptides are more sensitive to changes compared to proteins and meta-proteins, since their relative abundances varied the most across the BGP samples (Figure 4.43). Nevertheless, a varying number of identified proteins and meta-proteins can have a strong impact on the functional analysis: the more different enzymes are found in the results of a sample, the higher is the coverage in the respective pathways and the more reactions in a metabolic network can be explained.

In order to further examine the influence of the protein database on the enzyme coverage in metabolic pathways, the KO and EC identifiers were compared between the BGP data sets for SwissProt and TrEMBL. It was found that the number of identified KO numbers doubled in the results for TrEMBL when compared to SwissProt (Figure 4.44). On the one hand, these results confirm the findings in previous sections regarding the influence of the chosen protein database, on the other hand, the consequences from this impact may be more grave in this case, since few different assignments can already affect conclusions which are drawn based on the enzymes found in the metabolic pathway analysis.

Summary. It was found that the reproducibility between data sets from replicate samples was consistently high regarding for the investigated quantitative measures (i.e. identified spectra, peptides, proteins and meta-proteins) in the exemplary functional assignment to the ontologies *Biological Process* and *Molecular Function*. Furthermore, the evaluation of results from different metaproteomic samples suggested that changes across samples are easier to detect when counting the spectra and peptides that were specific for a particular functional process. When comparing the KO and EC numbers between the result sets from searches against two different public databases, it was observed again by the varying amount of assignments that the chosen database can strongly influence the analysis at the functional pathway level.

5.6.3 Postprocessing Unannotated Data

The functional analysis of results containing unannotated protein sequences (e.g. obtained from a metagenomic database) requires additional steps to match the identifications against resources that provide relevant meta-information. In this work, to enhance the results from the HIMP samples with functional annotations, the HMMER software [278] was used for searching identified proteins against a bacterial EggNOG database [213] that provides orthologous groups of proteins from complete genomes.

The aim of the first analysis was to obtain a general overview on the functional profiles of the HIMP samples. In a second investigation, the functionally annotated data were also grouped into bacterial taxa at the taxonomic rank of phylum; Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria were considered here, since these had been the most abundant bacterial phyla according to previous analyses carried out at the phylogenetic level (see Section 4.5.3).

The findings of the first analysis show that a considerable proportion of functionally annotated results correspond to previous findings in studies on the human gut microbiome. The results for two exemplary HIMP data sets (P1 and P23) show that the highest number of peptides were assigned to the EggNOG category "Carbohydrate transport and metabolism" (Figure 4.45). Since more than 50% of the 20 most abundant NOGs belonged to this category, these data confirm findings from previous studies that highlighted the importance of carbohydrate degradation and fermentation processes performed by gut microbiota [343, 344]. Metagenomic studies had also shown an enrichment of genes related to carbohydrate metabolism in the gut [274, 345]. A metatranscriptomic study analyzed fecal samples from two healthy subjects and reported that most expressed genes could be related to the metabolism of carbohydrates [346]. Gosalbes *et al.* also found functional groups from the latter category overrepresented in their metatranscriptome data [69]. Furthermore, the absolute number of more than 1 000 identified proteins per sample corresponds to the findings of an analysis of gut metaproteome samples from three healthy in-

dividuals that were measured over a time period of 6–12 months [15]. The results in this work show that "Amino acid transport and metabolim" was the second most abundant category. The relevance of the latter category is referenced in a recent study that describes the ability of gut bacteria to modulate the amino acid provision by using food and host proteins [347]. Eventually, "Energy production and conversion" was another highly abundant functional class in this analysis. Likewise, comparable abundance estimates for an analogous functional category had been previously reported in numerous metatranscriptome and metaproteome studies on the human gut [14, 15, 113, 69, 348].

The second analysis combined functional information with taxonomic assignments at the phylum level of bacterial taxa: the results show that "Carbohydrate transport and metabolism" was the most abundant category in Firmicutes (Figure 4.46a), while this category alternated with "Function unknown" for most assignments in Bacteroidetes (Figure 4.46b), indicating that Firmicutes are better characterized and annotated by their functions than Bacteroidetes. While Kolmeder *et al.* counted spectral instead of peptide hits for each functional class in their study, they also observed that the aforementioned categories were the most abundant for Firmicutes and Bacteroidetes. The results of this work also suggest that the functional diversity of Bacteroidetes and Firmicutes is higher than of Actinobacteria, because the latter phylum was mainly represented by assignments associated with carbohydrate metabolism and amino acid transport (Figure 4.47a). Proteobacteria provided more different functions than the other regarded phyla, but also held considerably less assignments (Figure 4.47b). In addition, three obese samples exhibited a higher abundance and functional diversity for Proteobacteria when compared to the other data sets. However, the low number of total assigned peptides would only allow speculations about any association between sample group and represented phylum in this case.

Supervised classification based on bacterial functional groups. In order to find significantly different bacterial NOGs in the samples of obese and lean subjects, the supervised classification method LEfSe was applied on the complete result set of 29 HIMP samples. In their original study, bootstrap aggregated (bagged) RDA was employed as method to identify the separating NOGs in the same data [266]. While LEfSe is based on LDA as supervised classification technique and is frequently used for metagenomic data, bagged RDA had been previously applied in a study on phylogenetic microarray and qPCR data [349].

It was found that the application of LEfSe could separate obese and non-obese samples on the basis of bacterial functional groups. The results show that 21 out of 27 significantly different bacterial NOGs were characteristic for the non-obese group and could be mainly related to carbohydrate and amino acid metabolism (Table 4.8). These findings are in line with the original study [266]: comparing the results of both supervised methods in detail, it can be observed that

20 out of 25 bacterial NOGs from the latter study could be confirmed by the results in this work. Due to the observed reproducibility of the results, both supervised methods are justified to be applied at the functional level. Since both of them can be used on moderate sample sizes and high variable counts, they can be recommended for the classification of high-dimensional functional data in metaproteomics. However, two main advantages advise the use of LEfSe instead of other supervised classification techniques: first, an effect size is provided that constitutes a quantitative measure of the strength of an observed phenomenon. Second, LEfSe can be more easily accessed, as it is embedded in the web-based Galaxy resource that unifies bioinformatic solutions for the analysis of omics data [350, 351]. Since recently, the Galaxy pipeline also offers more computational methods for the analysis of metaproteomic data [352].

Summary. To sum up, it can be constituted that the functional profile analysis on HIMP samples confirms previous findings reported by studies on the gastrointestinal tract: in particular, the EggNOG analysis revealed a high abundance of peptides from enzymes responsible for carbohydrate transport and metabolism. Furthermore, the combination of functional and taxonomic information highlighted that carbohydrate degradation was the most abundant enzymatic group in Firmicutes. In addition, the application of the supervised classification method LEfSe showed that 27 bacterial NOGs were significantly different between the obese and the non-obese group of individuals. Overall, the retrieval of these key enzymes presents the starting point for further more detailed investigations that determine which microbial species interact in the intestine of healthy individuals to fulfill the most important tasks, for instance, accomplishing a efficient digestion of carbohydrates. In practice, the ultimate goal of longitudinal studies with a larger sample size would be to examine whether the gut microbiome can be altered to restore the health status of the host. Eventually, modulations of composition and metabolism of the gut microbiota could be induced by diet [353], tailor-made probiotics [354] or fecal bacteriotherapy [355].

6

Conclusion and Outlook

The general aim of this work was to develop and evaluate computational methods for analyzing MS-based proteomic data sets derived from microbial community samples. While metaproteomics presents a promising technique for identifying and quantifying the whole set of proteins in a microbial community, it could be recognized in the first place that adequate bioinformatic analysis methods and dedicated software tools were lacking in this research field. The first objective was therefore to develop a data analysis pipeline geared towards metaproteomic data sets. While various computational methods have been developed and are commonly used for the proteomic analysis of pure-culture samples, a thorough testing of protein identification search algorithms and respective parameter settings had not been performed on the basis of metaproteomic data. Therefore, the second objective was to investigate available computational methods and to identify their limitations in relation to the data analysis in metaproteomics. Subsequently, alleviating these bottlenecks guarantees the eligibility of data analysis strategies to complex samples of microbial communities. The third objective was to analyze and interpret metaproteomic results with the focus on evaluating methods beyond protein identification, namely (i) protein grouping, (ii) taxonomic assignment and (iii) functional analysis.

Development of metaproteomic analysis software. The first part of this work presents the development of an analysis software that is tailored towards metaproteomic data sets. The MPA software was a major milestone in this work, since it allowed to process and analyze a high number of MS/MS data sets derived from diverse metaproteomic samples. The tool integrates several commonly used search algorithms and combines their results to increase the identification yield. While the identification of peptides and proteins constitutes an essential step in any proteomic analysis workflow, it is more the starting point than the final goal regarding the analysis of microbial community composition and function. For further detailed investigations, public databases harbor a large quantity of relevant meta-information at the protein level. Since it can be tedious and error-prone to manually link search results of a metaproteomic experiment to these resources, the developed software retrieves such metadata from various sources in an automated fashion and annotates each protein hit in the result set. In this procedure, the processed data are categorized meaningfully: for instance, proteins are classified according to their enzymatic function or taxonomic origin. Another advantage of the MPA application presents the meta-protein generation function that allows the grouping of redundant protein identifications obtained from sequence databases which contain many homologous entries from different organisms. This feature enables the user to dynamically choose one or multiple grouping methods; each rule presents a refinement strategy, ranging from taxonomy-based common ancestor approaches to methods based on protein sequence similarity. Therefore, instead of providing a static one-fits-all solution to the protein inference problem, the developed feature of meta-protein generation aims to be flexible enough to consider diverse objectives when grouping protein results.

The graphical user interface of the MPA software includes numerous visualization features that help to address questions related to the analysis of metaproteomic data by categorizing the results into different functional or phylogenetic classes. Another degree of usability is added by the graph-based database system that allows the user to submit specific queries accounting for more complex questions that are not covered by the default representation of the results in the software. Moreover, the implementation of a direct mapping of proteins into pathways can be used to gain a comprehensive overview on the potential metabolic activities of the involved microbial species. While the software has been developed with the main focus on analyzing metaproteomic data, it is also a powerful data analysis tool for pure-culture proteomics in general: the meta-protein generation and the automated integration of meta-information at the protein level are innovative and useful techniques that had not been realized in any proteomic software workflow.

Although the current MPA software provides a project management system which includes the flexible storage of multiple searches, it lacks the possibility of comparing different result sets between stored experiments. A straightforward solution could be to extend the existing graph database structure by including additional nodes for experiments and projects. However, such

a useful feature would also require additional development with respect to statistical methods for the qualitative and quantitative comparison of the results. It should also be noted that the current software architecture presents a compromise between solid server infrastructure and light-weight desktop software. A more powerful cluster- or cloud-based solution could therefore handle the huge amount of upcoming high-throughput data in metaproteomics more efficiently. The use of multiple cores and also server instances would highly reduce the running times by executing the database search algorithms in parallel tasks. Moreover, the current requirement of a server system impedes the portability of the application. In this context, a single software package that bundles algorithms and storage database might be of use to laboratories without access to extensive resources of IT hardware and expertise.

Evaluation of identification methods and parameters. After the development of the MPA software which provided the required computational framework for the major part of this work, the second milestone was to evaluate different methods and selected parameters used to identify peptides and proteins from MS/MS-based metaproteomic data. The rationale behind this task was to detect particular issues occurring during the computational analysis. Based on the identified bottlenecks, guidelines can be proposed for the optimal use of data analysis methods in metaproteomic research to increase the reliability of individual experiments and the comparability between different studies: besides varying sample preparation techniques, much bias is introduced by the heterogeneous use of data analysis procedures between different labs. As a summary of this work, an schematic overview provides detailed recommendations for an optimized data analysis workflow in the context of MS/MS-based metaproteomics (Figure 6.1). In the following, reference is made to the corresponding steps of this proposed workflow.

By analyzing representative metaproteomic data sets from biogas plant and human intestine samples (Figure 6.1; Step 1), it is demonstrated that sequence database, search algorithm and search parameters each strongly affect the identification yield. Among these factors, the most important one presents the choice of the sequence database (Figure 6.1; Step 2). It was found that searching against metagenome databases can improve the performance of a metaproteomic experiment with respect to the number of identified spectra and unique peptides. However, the findings also indicate that the success of a performed analysis depends on the relationship between the microbial community of the sample under investigation and the specimen that served as template for the metagenome. In addition, it can be observed that public databases, such as SwissProt and TrEMBL, also provide significant amounts of exclusive identifications that cannot be retrieved when matching MS/MS spectra against a metagenome background. To increase the total identification yield, it is therefore recommended to complement metagenome sequences by protein entries from public databases which hold the additional advantage of providing mostly

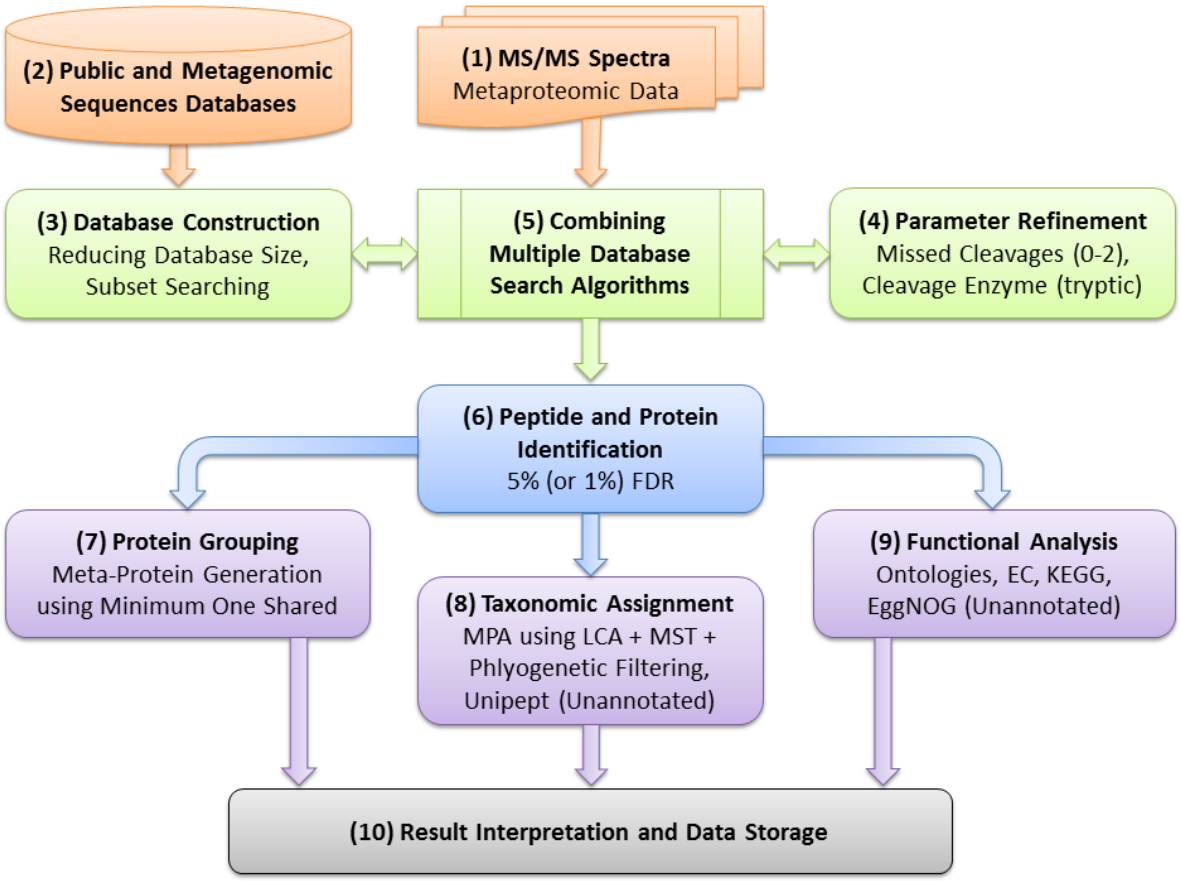

Figure 6.1: Optimized data analysis workflow for metaproteomics. The overview figure includes recommendations for a data analysis workflow in metaproteomics. Details on the steps 1--10 are given in the text.

well-annotated information. Moreover, the findings of the HIMP data analysis also advocate the use of single genomes from organisms that are suspected to be present in the samples. Finally, a balanced customization of the database is required to minimize the risk of biasing towards particular strains.

The next observed obstacle refers to the frequently neglected influence of the database size on the identification yield: the most severe issue is caused by applying the target-decoy approach for estimating the FDR to search hits from large metaproteomic databases: the benchmark experiment in which *Pyrococcus furiosus* data were searched against a database with many unrelated sequences reveals that valuable identifications are lost, since an increase of the protein search space increases the number of decoy hits as false positive estimates. This results in an overestimation of the FDR and decreased identification yields in metaproteomic analyses. Conversely, it could be observed that a two-step searching method results in a drastically increased number of identifications, but underestimates the FDR due to a overly reduced database size after the first

database search round. While a reduction of the database is generally recommended, it is often not feasible to exclude particular sequence *a priori* when the actual composition of the microbial sample is unknown. Therefore, searching against subsets of a large databases is proposed in this work as an alternative to two-step searching (Figure 6.1; Step 3). Subset searching reduces the search space for each identification round and significantly increases database-specific identifications in comparison to classic database searching. Moreover, it can be recognized in this work that search algorithms and methods for the statistical validation of database search results need to be further improved to account for issues resulting from complex search spaces, as in the case of metaproteomics or proteogenomics. In particular, the target-decoy approach should be carefully applied for the FDR estimation. The results also show that using different missed cleavage settings and combining the results increases the number of identifications (Figure 6.1; Step 4). To that end, searching with varying parameter values can also help to validate doubtful results by comparing the respective outcome of different settings. Furthermore, the findings in this work show that the use of multiple search engines and the subsequent combination of the results significantly increase the number of identifications (Figure 6.1; Step 5). According to the results in this work, an FDR threshold of 5% can be recommended to increase the identification yield in a metaproteomic experiment (Figure 6.1; Step 6). If two-step searching is applied, however, a more stringent threshold of 1% FDR should be considered to counteract the potential increase of false positive identifications.

Since the influence of the database and the frequent lack of sequence information present severe challenges in metaproteomics, the method of *de novo* sequencing becomes an interesting alternative. However, the findings show that the overlap between results from this technique and database searching is low. Moreover, the number of *de novo* peptides that were matched successfully against a provided protein sequence database could not justify the effort of using *de novo* sequencing to validate questionable peptide hits from conventional database searching. As a consequence, a clear benefit of using *de novo* sequencing as complementary method to database searching cannot be determined.

Besides the issues found during the evaluation of the data analysis methods, each additional method that is carried out raises computational running time as well as practical effort with respect to combination and validation of the results. As a consequence, primarily to the actual study, performing pilot experiments with different settings and a subsequent cost-benefit analysis represent recommended means to approximate an optimal workflow for the analysis of metaproteomic data. Finally, another promising strategy might be the alternative use of spectral libraries: while this method could not be evaluated due to the lack of metaproteome reference data, a development in the future could be the construction of customized spectral libraries on the basis of validated metaproteome results from previous database searches. This technique would be

particularly time-saving for repeated measurements or samples from which the same or related microbial communities have already been characterized in previous experiments.

Methods beyond protein identification. In the last part of this work, the importance of methods for postprocessing protein data is highlighted, since much valuable semantic context is required for metaproteomic studies. During the analysis of microbial communities, identifications often occur redundantly due to many homologous entries in protein databases. To reduce this redundancy, a contribution in this work presents the feature of meta-protein generation based on different rules that enable the grouping of proteins in the result set. The comparative analysis of these rules shows that the highest reduction of proteins is achieved by the *Minimum One Shared* method which merges a set of proteins sharing at least one peptide (Figure 6.1; Step 7). While the protein grouping is already useful for the application within a single result set, it also increases the comparability of results when providing a common set of meta-proteins shared between data sets from different samples. Additional rules, for instance, allowing point mutations at the peptide level or using a fixed maximum taxonomic convergence level, are viable options for user-definable grouping of proteins in metaproteomic experiments.

When investigating the influence of the protein database on the taxonomic assignment of well-annotated protein data from BGP samples, significant phyla-related differences can be observed between the results of searches against SwissProt and TrEMBL. While the resolution of the taxonomic analysis might be increased by searching against both public database variants, in case of doubt, it is recommended to rely on the manually curated SwissProt rather than the automatically annotated TrEMBL. In the ideal case, metagenomic sequences ensure the assignments by providing phylogenetic information in addition to public databases.

The correct taxonomic assignment of the results is essential for drawing conclusions on the composition of a studied microbial sample (Figure 6.1; Step 8). When using data sets of known microbial composition, MPA outperforms Unipept by yielding significantly more correct taxon-specific peptides. Instead of relying on a fixed FDR cutoff value, a phylogenetic filtering threshold is recommended to reduce the fraction of incorrect taxonomic assignments. As alternative to the commonly used LCA approach, the developed MST method shows a trend towards more correct taxonomic assignments. The idea of the latter method is to differentiate more accurately between closely related species or strains by preserving the peptide specificity. Finally, the taxonomic analysis of data from HIMP samples suggests a link between dietary factors and microbial abundance of the taxonomic genera *Bacteroides* and *Prevotella*, but more information on diet and lifestyle habits of the individuals is required to investigate these findings in sufficient detail. In this analysis, a combination of MPA and UniPept assists to process and interpret unannotated protein sequence data.

Based on annotated protein data from public resources, the proposed metaproteomic analysis workflow provides detailed functional information (Figure 6.1; Step 9). Besides the use of functional ontology data, the MPA features EC- and KEGG-based categorization views that allow to directly map relevant protein identifications into metabolic pathway routes. In line with previous findings during the taxonomic analysis of annotated protein data, the chosen sequence database has a high influence on the functional assignment: the investigated BGP samples show a significantly different abundance of key enzymes from the methanogenesis pathway between the public database variants. Due to the complementary detection of important functional entities, parallel searches against SwissProt and TrEMBL can increase the overall information content of the analysis. In line with previous recommendations in this work, the strategy of using a less strict FDR threshold of 5% can be useful here to increase the number of functional assignments and mappings to KEGG pathways. However, the user is then in charge to resolve uncertain or incorrect assignments of the proteins manually. Furthermore, the results of this work show that taxonomic and functional information should be combined to unveil properties that remain hidden when investigated separately: for example, it can be detected which species perform a particular function in the community. Finally, the functional analysis of unannotated protein sequences is demonstrated by matching protein identifications of the HIMP data sets against the EggNOG database. Remarkably, the findings of this exemplary analysis correlate well with reports on functional categories from previous omics-based studies on the human gastrointestinal tract. Ultimately, the work is concluded using a supervised method for identifying bacterial enzymes that differ significantly between the result data from lean and obese subjects. In this analysis, 27 functional assignments which are identified to vary significantly between the two groups confirm the findings of a study in which a more detailed functional analysis has been carried out [266].

Eventually, an optimized analysis workflow should also provide features for a detailed interpretation of the results and offer a reliable data storage (Figure 6.1; Step 10). While the MPA software meets these latter requirements, it has been designed to be extended by more external resources, such as further databases containing enzymatic or ontological information. Another future development presents the implementation of a strategy which automatically annotates protein sequences from metagenome and other non-reference databases: currently, UniProtKB reference entries are supported to retrieve comprehensive meta-information that enables the taxonomic and functional analysis. For instance, an additional protein mapping feature would allow to match unannotated sequences against a well-annotated public database. In this scenario, each identified protein could be updated within an automated processing step by including relevant annotations. In addition, a full integration of Unipept can increase the information content of the taxonomic analysis, in particular, when identifications are obtained from unannotated protein

sequences, as commonly given by metagenome databases (Figure 6.1; Step 8). Finally, the automated retrieval of meta-information from the EggNOG database can add essential knowledge about protein functions to the analysis workflow (Figure 6.1; Step 9).

List of Figures

2.1 Metaproteomic data analysis workflow 17

3.1 MetaProteomeAnalyzer software workflow 28
3.2 Pie and bar charts displaying the protein distributions on taxonomic ranks . . . 29
3.3 Search result panel of the MPA client user interface 30
3.4 Peptide rules for the meta-protein generation 32
3.5 Example of the taxonomy definition process 35
3.6 Main panel of the DeNovoGUI graphical user interface 43

4.1 Comparison of identifications from three search engines for EBENDORF01 . . . 48
4.2 Venn diagrams of peptides identified in BGP samples for three different databases 53
4.3 Comparison of scores from BGP searches against three different databases . . . 54
4.4 Overview on the identification results for data set P1 55
4.5 Database specific identifications from classic and subset searching for data set P1 56
4.6 Comparison of scores between classic and two-step searching for data set P1. . . 57
4.7 Reevaluation of identifications from classic and two-step searching for data set P1 58
4.8 Comparative evaluation of the identification yield for different MC values at 1% FDR (HIMP) 59
4.9 Comparison of total and exclusive peptides for different MC values (P1) 59
4.10 Comparative benchmark evaluation of different search strategies (PFU) 62

4.11 Evaluation of the identification yield for PFU searches against small (Pyrodb) and large (PyroHIMPdb) search space 62
4.12 Evaluation of score distributions between classic and two-step searching (PFU) 63
4.13 Comparative evaluation of the identification yield for different MC values (PFU) 64
4.14 Comparison of total and exclusive peptides for different MC values (PFU) . . . 65
4.15 *De novo* sequencing recovery of peptides from classic and two-step searching for data set P1 68
4.16 Comparison of PepNovo+ scores between classic and two-step searching for data set P1 69
4.17 Protein result set reduction achieved by meta-protein generation rules 71
4.18 Meta-protein result set reduction achieved by allowing point mutations 72
4.19 Meta-protein result set increase resulting from phylogenetic diversity control . . 73
4.20 Reproducibility of peptide hits between technical replicates for GENT01 74
4.21 Reproducibility of protein hits between technical replicates for GENT01 75
4.22 Reproducibility of meta-protein hits between technical replicates for GENT01 (*Minimum One Shared*) 76
4.23 Reproducibility of meta-protein hits between technical replicates for GENT01 (*All Shared*) 76
4.24 Evaluation of group similarity and size in dependence of shared peptides 79
4.25 Comparative evaluation of peptide frequencies for meta-protein generation rules 79
4.26 Phylogenetic classification of BGP data set GENT16 based on number of peptides per superkingdom 81
4.27 Phylogenetic classification of BGP data sets based on the number of peptides per phylum 82
4.28 Taxonomic assignment performance of MPA and Unipept for BGP data sets . . 83
4.29 Taxonomic assignment performance of MPA for 9MM data set 84
4.30 Taxonomic assignment performance of Unipept for 9MM data set 85

4.31 Performance comparison of LCA and MST taxonomic assignment methods . . 86
4.32 Phylogenetic classification of HIMP10 (superkingdom level) 87
4.33 Phylum-level taxonomic classification of Unipept peptides for HIMP10 88
4.34 Genus-level taxonomic classification of Unipept peptides for HIMP10 88
4.35 Genus-level taxonomic classification of Unipept peptides for HIMP10 89
4.36 Biological Process-specific peptide assignments for BGP data sets 91
4.37 Molecular Function-specific peptide assignments for BGP data sets 92
4.38 Total number of enzyme-specific identifications for GENT01 93
4.39 Total number of enzyme-specific identifications for GENT01 94
4.40 Total number of enzyme-specific identifications for GENT16 94
4.41 KEGG pathway of carbon metabolism for GENT01 protein identifications 96
4.42 Reproducibility of ontology-specific assignments across technical replicate experiments for GENT01 97
4.43 Total number of "Methanogenesis"-specific identifications for BGP data sets . . 98
4.44 Total number of "Carbon metabolism"-specific identifications for BGP data sets 99
4.45 Total number of peptides assigned to EggNOG categories for HIMP data sets at 1% and 5% FDR. 101
4.46 Phylum-level functional classification of peptides from Firmicutes and Bacteroidetes for HIMP10 101
4.47 Phylum-level functional classification of peptides from Actinobacteria and Proteobacteria for HIMP10 102

6.1 Optimized data analysis workflow for metaproteomics 136

A.1 Comparative evaluation of the identification yield for different MC values at 5% FDR (HIMP) 187
A.2 Comparison of total and exclusive peptides for different MC values (P23) 187
A.3 Comparison of total and exclusive peptides for different MC values (P34) 188

A.4 Reproducibility of peptide hits between technical replicates for GENT16 188
A.5 Reproducibility of protein hits between technical replicates for GENT16 189
A.6 Reproducibility of meta-protein hits between technical replicates for GENT16 (*Minimum One Shared*) 189
A.7 Reproducibility of meta-protein hits between technical replicates for GENT16 (*All Shared*) 190
A.8 Phylogenetic classification of BGP data set GENT01 based on number of peptides per superkingdom 190
A.9 Phylogenetic classification of BGP data set GENT07 based on number of peptides per superkingdom 191
A.10 KEGG reference pathway of carbon metabolism (map01200) for GENT01 protein hits from Archaea 192
A.11 KEGG reference pathway of carbon metabolism (map01200) for GENT01 protein hits from Bacteria 193
A.12 KEGG reference pathway of amino acid synthesis (map01230) for GENT01 protein hits from Archaea 194
A.13 KEGG reference pathway of amino acid synthesis (map01230) for GENT01 protein hits from Bacteria 195
A.14 Reproducibility of ontology-specific assignments across technical replicate experiments for GENT01 196
A.15 Total number of "Carbon metabolism"-specific peptide assignments for BGP data sets 197

List of Tables

3.1 Node types and descriptions for the graph database schema 36

3.2 Microorganisms of the 9MM sample 38

3.3 Composition of the human intestine metaproteome database 40

4.1 Number of MS/MS spectra, percentage of identified spectra and exclusive identification yield for HIMP10 49

4.2 Percentage of identified spectra and number of peptides obtained from searching GENT01, GENT07 and GENT16 against SwissProt, TrEMBL and BGPMG (FDR 5%) 51

4.3 Number of identified peptides and percentage of unique peptides (BGP) 51

4.4 Number of identified peptides and percentage of unique peptides (HIMP) . . . 55

4.5 Number of identifications and percentage of exclusive hits for data sets P1, P23, and P34 using tryptic and semi-tryptic cleavage settings (FDR 5%) 60

4.6 Number of identifications and percentage of exclusive hits for samples P1, P23, and P34 using tryptic and semi-tryptic cleavage settings (FDR 1%) 61

4.7 *De novo* sequencing results for the HIMP10 data sets 67

4.8 27 NOGs classified by LEfSe as significantly different between non-obese and obese individuals 104

A.1 Relationship types and descriptions for the graph database schema 179

A.2 Number of MS/MS spectra, percentage of identified spectra and number of peptides for HIMP10 180
A.3 Number of identified peptides from HIMP data sets P1-P34 that could be uniquely mapped against an *in silico* digest of the respective databases (FDR < 5%) 180
A.4 Proportion of UniProtKB/TrEMBL peptide identifications for GENT01, GENT07 and GENT16 (FDR < 5%) matched against UniProtKB/SwissProt and BGPMG . 180
A.5 Number of PSMs and peptides identified by searching data sets P1, P23 and P34 with X!Tandem and OMSSA against HIMPdb, Bact594db and Qin2010db. In addition, two-step searching against HIMPdb was performed (FDR < 5%) . . . 181
A.6 Number of Bact594db-specific PSMs and peptides for data sets P1, P23 and P34 181
A.7 Number of Qin2010db specific PSMs and peptides for data sets P1, P23 and P34 181
A.8 Number of identifications for data sets P1, P23, and P34 using chymotrypsin and pepsin A 181
A.9 PFU result score threshold values using X!Tandem and OMSSA for searching against Pyrodb, PyroHIMPdb and PyroHIMPdb (two-step searching) at 1% and 5% FDR 182
A.10 Number of species-specific peptides for data sets 9MM_FASP and 9MM_PPID 182
A.11 Number of genus-specific peptides for data sets 9MM_FASP and 9MM_PPID . 182
A.12 The 20 most abundant NOGs for HIMP data set P1 183
A.13 The 20 most abundant NOGs for HIMP data set P23 184
A.14 Peptide and protein assignments to EggNOG categories for data set P1 185
A.15 Peptide and protein assignments to EggNOG categories for data set P23 186
A.16 Number of phylum-level EggNOG peptide assignments for HIMP10 186

List of Contributions

This work partially contains content and material that has been previously published elsewhere.

Publications

Kolmeder C.A., Ritari J., Verdam F.J., **Muth T.**, Keskitalo S., Varjosalo M., Fuentes S., Greve J.W., Buurman W.A., Reichl U., Rapp E., Martens L., Palva A., Salonen A., Rensen S.S., de Vos W.M. (2015) Colonic metaproteomic signatures of active bacteria and the host in obesity. Proteomics 15(20):3544–3552. T.M. performed the data analysis with respect to MS/MS peptide and protein identification and edited the manuscript. *Copyright 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.*

Muth T.*, Kolmeder C.A.*, Salojärvi J., Keskitalo S., Varjosalo M., Verdam F.J., Rensen S.S., Reichl U. de Vos W.M., Rapp E., Martens L. (2015) Navigating through metaproteomics data: A logbook of database searching. Proteomics 15(20):3439–53. T.M. performed the data analysis and edited the manuscript. *Copyright 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.*

Muth T.*, Behne A.*, Heyer R., Kohrs F., Benndorf D., Hoffmann M., Lehtevä M., Reichl U., Martens L., Rapp E. (2015) The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. J Proteome Res. 14(3):1557–1565. T.M. designed the bioinformatic workflow, developed modules of the software and wrote the manuscript. *Copyright 2015 American Chemical Society.*

Oveland E., **Muth T.**, Rapp E., Martens L., Berven F.S., Barsnes H. (2014) Viewing the proteome: How to visualize proteomics data? Proteomics 15(8):1341–1355. T.M. took part in the literature research and wrote parts of the manuscript. *Copyright 2014 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.*

Muth T., Weilnböck L., Rapp E., Huber C.G., Martens L.,Vaudel M., Barsnes H. (2014) DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. J Proteome Res. 13(2):1143–1146. T.M. developed the software and wrote the manuscript. *Copyright 2013 American Chemical Society.*

Kohrs F.*, Heyer R.*, Magnussen A., Benndorf D., **Muth T.**, Behne A., Rapp E., Kausmann R., Heiermann M., Klocke M., Reichl, U. (2014) Sample prefractionation with liquid isoelectric focusing enables in depth microbial metaproteome analysis of mesophilic and thermophilic biogas

plants. Anaerobe 29:59–67. T.M. gave conceptual advice and edited the manuscript. *Copyright 2013 Elsevier Ltd.*

Behne A., **Muth T.**, Borowiak M., Reichl U., Rapp E. (2013) glyXalign: high-throughput migration time alignment pre-processing of electrophoretic data retrieved via xCGE-LIF-based glycoprofiling. Electrophoresis 34(16):2311–2315. T.M. supervised the software development and edited the manuscript. *Copyright 2013 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.*

Muth T., Benndorf D., Reichl U., Rapp E., Martens L. (2013) Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. Mol Biosyst. 9(4):578–585. T.M. reviewed the literature and wrote the manuscript. *Copyright 2013 The Royal Society of Chemistry.*

* Equal authorship.

Students supervision

Behne A. (2013) A performance evaluation of spectral library searching in metaproteomics. Student research project, Faculty for Process and Systems Engineering, Otto von Guericke University, Magdeburg.

Espe J. (2014) Development of a graph-based algorithm for the identification of selected glycoproteins and implentation of a glycopattern library. Bachelor thesis, Faculty for Process and Systems Engineering, Otto von Guericke University, Magdeburg.

Behne A. (2014) Development of algorithms and methods for the annotation of CGE-LIF-based N-glycan profiles. Diploma thesis, Faculty for Process and Systems Engineering, Otto von Guericke University, Magdeburg.

Dorl S. (2015) Clustering of tandem mass spectrometry data from complex microbial samples. Master thesis, Faculty for Process and Systems Engineering, Otto von Guericke University, Magdeburg.

Talks

Muth T., Kolmeder C.A., Rapp E., Reichl U., Martens L. (2015) A closer look into the maze of metaproteomics data. MidWinter Proteomics Bioinformatics Seminar, Semmering, Austria.

Muth T., Behne A., Heyer R., Kohrs F., Hoffmann M., Benndorf D., Lehtevä M. Reichl U., Martens L., Rapp E. (2014) MetaProteomeAnalyzer: a software suite for the functional and taxonomic characterization of (meta)proteome data. 12th Austrian Proteomics Research Symposium, Salzburg, Austria.

Muth T., Behne A., Heyer R., Kohrs F., Benndorf D., Lehtevä M., Reichl U., Martens L., Rapp E. (2014) MetaProteomeAnalyzer: a graph database backed protein analysis software. 4th Linked Data Benchmark Council Technical User Communiy Meeting, Amsterdam, Netherlands.

Muth T., Hennig R., Rapp E., Reichl U. (2013) glyXtool - a software tool for high-throughput processing of glycoanalysis data. 24th Joint Glycobiology Meeting, Wittenberg, Germany.

Muth T., Hennig R., Behne A., Rapp E., Reichl U. (2013) glyXtool and glyXalign: software for high-throughput processing of glycoanalysis data. 3rd Beilstein Symposium on Glyco-Bioinformatics, Potsdam, Germany.

Workshop organization

Benndorf D., **Muth T.**, Heyer R., Rapp E., Reichl U. (February 2016) Symposium on Advances and Applications in Metaproteomics, Magdeburg, Germany.

Muth T., Hunger M., Van Bruggen R., Martens L. (November 2012) Workshop: Graph Databases in Life Sciences, Ghent, Belgium.

Posters

Muth T., Behne A., Heyer R., Kohrs F., Hoffmann M., Benndorf D., Lehtevä M., Reichl U., Martens L., Rapp E. (2015) MetaProteomeAnalyzer - a software suite for the functional and taxonomic characterization of (meta)proteome data. Proteomic Forum 2015, Berlin, Germany.

Muth T., Espe J., Rapp E., Reichl U. (2014) Glyco Profiler – CGE-LIF-based recognition of glycoproteins. 25th Joint Glycobiology Meeting, Ghent, Belgium.

Muth T., Weilnböck L., Rapp E., Huber C.G., Martens L.,Vaudel M., Barsnes H. (2013) DeNovoGUI: an open-source graphical user interface for de novo sequencing of tandem mass spectra. German Conference on Bioinformatics 2013, Göttingen, Germany.

Muth T., Behne A., Hennig R, Reichl U., Rapp E. (2013) Software for automated high-throughput processing of xCGE-LIF based glycoanalysis data. 7th Glycan Forum, Berlin, Germany.

Muth T., Heyer R., Behne A., Kohrs F., Benndorf D., Rapp E., Reichl U. (2012) MetaProteomeAnalyzer: A software tool specifically developed for the functional and taxonomic characterization of metaproteome data. German Conference on Bioinformatics 2012, Jena, Germany.

Hennig R., **Muth T.**, Reichl U., Rapp E. (2012) glyXtool – a software for automated high-throughput processing of xCGE-LIF based glycoanalysis data. 6th Glycan Forum, Berlin, Germany.

Bibliography

[1] Rodriguez-Valera F (2004) Environmental genomics, the big picture? FEMS Microbiol Lett 231: 153–158.

[2] Leininger S, Urich T, Schloter M, Schwark L, Qi J, et al. (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature 442: 806–809.

[3] Wilmes P, Wexler M, Bond PL (2008) Metaproteomics provides functional insight into activated sludge wastewater treatment. PLoS One 3: e1778.

[4] Schneider T, Keiblinger KM, Schmid E, Sterflinger-Gleixner K, Ellersdorfer G, et al. (2012) Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. ISME J 6: 1749–1762.

[5] Hanreich A, Schimpf U, Zakrzewski M, Schluter A, Benndorf D, et al. (2013) Metagenome and metaproteome analyses of microbial communities in mesophilic biogas-producing anaerobic batch fermentations indicate concerted plant carbohydrate degradation. Syst Appl Microbiol 36: 330–338.

[6] Guarner F, Malagelada JR (2003) Gut flora in health and disease. Lancet 361: 512–519.

[7] Singh J, Behal A, Singla N, Joshi A, Birbian N, et al. (2009) Metagenomics: Concept, methodology, ecological inference and recent advances. Biotechnol J 4: 480–494.

[8] Wilmes P, Bond PL (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environ Microbiol 6: 911–920.

[9] Banfield JF, Verberkmoes NC, Hettich RL, Thelen MP (2005) Proteogenomic approaches for the molecular characterization of natural microbial communities. OMICS 9: 301–333.

[10] Hettich RL, Pan C, Chourey K, Giannone RJ (2013) Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. Anal Chem 85: 4203–4214.

[11] Nesvizhskii AI, Aebersold R (2005) Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 4: 1419–1440.

[12] Kuhn R, Benndorf D, Rapp E, Reichl U, Palese LL, et al. (2011) Metaproteome analysis of sewage sludge from membrane bioreactors. Proteomics 11: 2738–2744.

[13] Hanreich A, Heyer R, Benndorf D, Rapp E, Pioch M, et al. (2012) Metaproteome analysis to determine the metabolically active part of a thermophilic microbial community producing biogas from agricultural biomass. Can J Microbiol 58: 917–922.

[14] Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, et al. (2009) Shotgun metaproteomics of the human distal gut microbiota. ISME J 3: 179–189.

[15] Kolmeder CA, De Been M, Nikkilä J, Ritamo I, Mättö J, et al. (2012) Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. PloS one 7: e29913.

[16] Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A 95: 6578–6583.

[17] Torsvik V, Ovreas L, Thingstad TF (2002) Prokaryotic diversity–magnitude, dynamics, and controlling factors. Science 296: 1064–1066.

[18] Karner MB, DeLong EF, Karl DM (2001) Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature 409: 507–510.

[19] Gold T (1992) The deep, hot biosphere. Proc Natl Acad Sci U S A 89: 6045–6049.

[20] Krumholz LR, McKinley JP, Ulrich GA, Suflita JM (1997) Confined subsurface microbial communities in Cretaceous rock. Nature 386: 64–66.

[21] Roussel EG, Bonavita MA, Querellou J, Cragg BA, Webster G, et al. (2008) Extending the sub-sea-floor biosphere. Science 320: 1046.

[22] Enright MC, Robinson DA, Randle G, Feil EJ, Grundmann H, et al. (2002) The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). Proc Natl Acad Sci U S A 99: 7687–7692.

[23] Alekshun MN, Levy SB (2007) Molecular mechanisms of antibacterial multidrug resistance. Cell 128: 1037–1050.

[24] Magiorakos AP, Srinivasan A, Carey RB, Carmeli Y, Falagas ME, et al. (2012) Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. Clin Microbiol Infect 18: 268–281.

[25] Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304.

[26] Rossello-Mora R, Amann R (2001) The species concept for prokaryotes. FEMS Microbiol Rev 25: 39–67.

[27] Goldenfeld N, Woese C (2007) Biology's next revolution. Nature 445: 369.

[28] Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59: 143–169.

[29] Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276: 734–740.

[30] Rappe MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57: 369–394.

[31] Streit WR, Schmitz RA (2004) Metagenomics–the key to the uncultured microbes. Curr Opin Microbiol 7: 492–498.

[32] Haynes M, Rohwer F (2011) The Human Virome. Metagenomics of the Human Body : 63–77.

[33] Savage DC (1977) Microbial ecology of the gastrointestinal tract. Annu Rev Microbiol 31: 107–133.

[34] Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-bacterial mutualism in the human intestine. Science 307: 1915–1920.

[35] Relman DA, Falkow S (2001) The meaning and impact of the human genome sequence for microbiology. Trends Microbiol 9: 206–208.

[36] Lederberg J (2000) Infectious history. Science 288: 287–293.

[37] Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The human microbiome project. Nature 449: 804–810.

[38] Shanahan F (2002) The host-microbe interface within the gut. Best Pract Res Clin Gastroenterol 16: 915–931.

[39] Dethlefsen L, McFall-Ngai M, Relman DA (2007) An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature 449: 811–818.

[40] O'Hara AM, Shanahan F (2006) The gut flora as a forgotten organ. EMBO Rep 7: 688–693.

[41] Tremaroli V, Backhed F (2012) Functional interactions between the gut microbiota and host metabolism. Nature 489: 242–249.

[42] Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59–65.

[43] Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74: 5088–5090.

[44] Woese CR, Olsen GJ (1986) Archaebacterial phylogeny: perspectives on the urkingdoms. Syst Appl Microbiol 7: 161–177.

[45] Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221–271.

[46] Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87: 4576–4579.

[47] Vergin KL, Urbach E, Stein JL, DeLong EF, Lanoil BD, et al. (1998) Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order Planctomycetales. Appl Environ Microbiol 64: 3075–3078.

[48] Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, et al. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289: 1902–1906.

[49] Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, et al. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A 82: 6955–6959.

[50] Theron J, Cloete TE (2000) Molecular techniques for determining microbial diversity and community structure in natural environments. Crit Rev Microbiol 26: 37–57.

[51] Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, et al. (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42: D633–642.

[52] Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6: e1000667.

[53] Ronaghi M (2001) Pyrosequencing sheds light on DNA sequencing. Genome Res 11: 3–11.

[54] Ahmadian A, Ehn M, Hober S (2006) Pyrosequencing: history, biochemistry and future. Clin Chim Acta 363: 83–94.

[55] Marsh S (2007) Pyrosequencing applications. Methods Mol Biol 373: 15–24.

[56] Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, et al. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 66: 2541–2547.

[57] Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37–43.

[58] Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554–557.

[59] Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66–74.

[60] Waldor MK, Tyson G, Borenstein E, Ochman H, Moeller A, et al. (2015) Where next for microbiome research? PLoS Biol 13: e1002050.

[61] Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, et al. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature biotechnology 31: 533–538.

[62] Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499: 431–437.

[63] Gilbert JA, Field D, Huang Y, Edwards R, Li W, et al. (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One 3: e3042.

[64] Gilbert JA, Hughes M (2011) Gene expression profiling: metatranscriptomics. Methods Mol Biol 733: 195–205.

[65] Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19: 1720–1730.

[66] Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, et al. (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science 329: 533–538.

[67] Deutscher MP (2006) Degradation of RNA in bacteria: comparison of mRNA and stable RNA. Nucleic Acids Res 34: 659–666.

[68] Reck M, Tomasch J, Deng Z, Jarek M, Husemann P, et al. (2015) Stool metatranscriptomics: A technical guideline for mRNA stabilisation and isolation. BMC Genomics 16: 494.

[69] Gosalbes MJ, Durbán A, Pignatelli M, Abellan JJ, Jiménez-Hernández N, et al. (2011) Metatranscriptomic approach to analyze the functional human gut microbiota. PLoS One 6: e17447.

[70] Wilkins M (2009) Proteomics data mining. Expert Rev Proteomics 6: 599–603.

[71] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.

[72] Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945.

[73] Jensen ON (2004) Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr Opin Chem Biol 8: 33–41.

[74] Mann M, Jensen ON (2003) Proteomic analysis of post-translational modifications. Nat Biotechnol 21: 255–261.

[75] Walsh CT, Garneau-Tsodikova S, Gatto J G J (2005) Protein posttranslational modifications: the chemistry of proteome diversifications. Angew Chem Int Ed Engl 44: 7342–7372.

[76] Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, et al. (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Biotechnology (N Y) 14: 61–65.

[77] Ünlü M, Morgan ME, Minden JS (1997) Difference gel electrophoresis. A single gel method for detecting changes in protein extracts. Electrophoresis 18: 2071–2077.

[78] Rappsilber J (2011) The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. J Struct Biol 173: 530–540.

[79] Olsen JV, Mann M (2013) Status of large-scale analysis of post-translational modifications by mass spectrometry. Mol Cell Proteomics 12: 3444–3452.

[80] Altelaar AFM, Munoz J, Heck AJR (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet 14: 35–48.

[81] Wilmes P, Bond PL (2006) Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol 14: 92–97.

[82] Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, et al. (2011) The quantitative proteome of a human cell line. Mol Syst Biol 7: 549.

[83] Moghaddas Gholami A, Hahne H, Wu Z, Auer FJ, Meng C, et al. (2013) Global proteome analysis of the NCI-60 cell line panel. Cell Rep 4: 609–620.

[84] Branca RMM, Orre LM, Johansson HJ, Granholm V, Huss M, et al. (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat Methods 11: 59–62.

[85] Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, et al. (2014) A draft map of the human proteome. Nature 509: 575–581.

[86] Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, et al. (2014) Mass-spectrometry-based draft of the human proteome. Nature 509: 582–587.

[87] Schulze WX, Gleixner G, Kaiser K, Guggenberger G, Mann M, et al. (2005) A proteomic fingerprint of dissolved organic carbon and of soil particles. Oecologia 142: 335–343.

[88] Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, et al. (2005) Community proteomics of a natural microbial biofilm. Science 308: 1915–1920.

[89] Lo I, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, et al. (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. Nature 446: 537–541.

[90] Benndorf D, Balcke GU, Harms H, von Bergen M (2007) Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. ISME J 1: 224–234.

[91] Nannipieri P (2006) Role of stabilised enzymes in microbial ecology and enzyme extraction from soil with potential applications in soil proteomics. In: Nucleic acids and proteins in soil, Springer. pp. 75–94.

[92] Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, et al. (2010) Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. ISME J 4: 673–685.

[93] Heyer R, Kohrs F, Reichl U, Benndorf D (2015) Metaproteomics of complex microbial communities in biogas plants. Microbial biotechnology 8: 749–763.

[94] Weiland P (2010) Biogas production: current state and perspectives. Appl Microbiol Biotechnol 85: 849–860.

[95] Ward AJ, Hobbs PJ, Holliman PJ, Jones DL (2008) Optimisation of the anaerobic digestion of agricultural resources. Bioresource technology 99: 7928–7940.

[96] Blume F, Bergmann I, Nettmann E, Schelle H, Rehde G, et al. (2010) Methanogenic population dynamics during semi-continuous biogas fermentation and acidification by overloading. Journal of applied microbiology 109: 441–450.

[97] Chae K, Jang A, Yim S, Kim IS (2008) The effects of digestion temperature and temperature shock on the biogas yields from the mesophilic anaerobic digestion of swine manure. Bioresource Technology 99: 1–6.

[98] Schlüter A, Bekel T, Diaz NN, Dondrup M, Eichenlaub R, et al. (2008) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. J Biotechnol 136: 77–90.

[99] Wirth R, Kovács E, Maróti G, Bagi Z, Rákhely G, et al. (2012) Characterization of a biogas-producing microbial community by short-read next generation DNA sequencing. Biotechnology for biofuels 5: 1.

[100] Heyer R, Kohrs F, Benndorf D, Rapp E, Kausmann R, et al. (2013) Metaproteome analysis of the microbial communities in agricultural biogas plants. N Biotechnol 30: 614–622.

[101] Kohrs F, Heyer R, Magnussen A, Benndorf D, Muth T, et al. (2014) Sample prefractionation with liquid isoelectric focusing enables in depth microbial metaproteome analysis of mesophilic and thermophilic biogas plants. Anaerobe 29: 59–67.

[102] Theuerl S, Kohrs F, Benndorf D, Maus I, Wibberg D, et al. (2015) Community shifts in a well-operating agricultural biogas plant: how process variations are handled by the microbiome. Appl Microbiol Biotechnol 99: 7791–7803.

[103] Berg RD (1996) The indigenous gastrointestinal microflora. Trends Microbiol 4: 430–435.

[104] Chow J, Lee SM, Shen Y, Khosravi A, Mazmanian SK (2010) Host-bacterial symbiosis in health and disease. Adv Immunol 107: 243–274.

[105] Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635–1638.

[106] Clemente JC, Ursell LK, Parfrey LW, Knight R (2012) The impact of the gut microbiota on human health: an integrative view. Cell 148: 1258–1270.

[107] Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022–1023.

[108] Ferrer M, Ruiz A, Lanza F, Haange SB, Oberbach A, et al. (2013) Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. Environ Microbiol 15: 211–226.

[109] Larsen N, Vogensen FK, van den Berg FW, Nielsen DS, Andreasen AS, et al. (2010) Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PLoS One 5: e9085.

[110] Howitt MR, Garrett WS (2012) A complex microworld in the gut: gut microbiota and cardiovascular disease connectivity. Nat Med 18: 1188–1189.

[111] Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, et al. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A 104: 13780–13785.

[112] Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355–1359.

[113] Rooijers K, Kolmeder C, Juste C, Dore J, de Been M, et al. (2011) An iterative workflow for mining the human intestinal metaproteome. BMC Genomics 12: 6.

[114] Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, et al. (2012) Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One 7: e49138.

[115] Lichtman JS, Marcobal A, Sonnenburg JL, Elias JE (2013) Host-centric proteomics of stool: a novel strategy focused on intestinal responses to the gut microbiota. Mol Cell Proteomics 12: 3310–3318.

[116] Washburn MP, Wolters D, Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature biotechnology 19: 242–247.

[117] Sze SK, Ge Y, Oh H, McLafferty FW (2002) Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue. Proc Natl Acad Sci U S A 99: 1774–1779.

[118] Kelleher NL (2004) Top-down proteomics. Anal Chem 76: 197A–203A.

[119] Borchers CH, Thapar R, Petrotchenko EV, Torres MP, Speir JP, et al. (2006) Combined top-down and bottom-up proteomics identifies a phosphorylation site in stem–loop-binding proteins that contributes to high-affinity RNA binding. Proceedings of the National Academy of Sciences of the United States of America 103: 3094–3099.

[120] Wu S, Lourette NM, Tolić N, Zhao R, Robinson EW, et al. (2009) An integrated top-down and bottom-up strategy for broadly characterizing protein isoforms and modifications. J Proteome Res 8: 1347–1357.

[121] Benndorf D, Reichl U (2014) Proteomics in environmental and technical microbiology. Engineering in Life Sciences 14: 27–46.

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

[123] Yates JR, Ruse CI, Nakorchevsky A (2009) Proteomics by mass spectrometry: approaches, advances, and applications. Annual review of biomedical engineering 11: 49–79.

[124] Zhang Y, Fonslow BR, Shan B, Baek MC, Yates III JR (2013) Protein analysis by shotgun/bottom-up proteomics. Chemical reviews 113: 2343–2394.

[125] Jagtap P, Goslinga J, Kooren JA, McGowan T, Wroblewski MS, et al. (2013) A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics 13: 1352–1357.

[126] Tanca A, Palomba A, Deligios M, Cubeddu T, Fraumene C, et al. (2013) Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. PLoS One 8: e82981.

[127] Tanca A, Palomba A, Pisanu S, Deligios M, Fraumene C, et al. (2014) A straightforward and efficient analytical pipeline for metaproteome characterization. Microbiome 2: 49.

[128] Canas B, Pineiro C, Calvo E, Lopez-Ferrer D, Gallardo JM (2007) Trends in sample preparation for classical and second generation proteomics. J Chromatogr A 1153: 235–258.

[129] Bodzon-Kulakowska A, Bierczynska-Krzysik A, Dylag T, Drabik A, Suder P, et al. (2007) Methods for samples preparation in proteomic research. Journal of Chromatography B 849: 1–31.

[130] Gevaert K, Van Damme P, Ghesquière B, Impens F, Martens L, et al. (2007) A la carte proteomics with an emphasis on gel-free techniques. Proteomics 7: 2698–2718.

[131] Kohrs F, Wolter S, Benndorf D, Heyer R, Hoffmann M, et al. (2015) Fractionation of biogas plant sludge material improves metaproteomic characterization to investigate metabolic activity of microbial communities. Proteomics 15: 3585–3589.

[132] Olsen JV, Ong SE, Mann M (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Mol Cell Proteomics 3: 608–614.

[133] Strader MB, Tabb DL, Hervey WJ, Pan C, Hurst GB (2006) Efficient and specific trypsin digestion of microgram to nanogram quantities of proteins in organic-aqueous solvent systems. Anal Chem 78: 125–134.

[134] Picotti P, Aebersold R, Domon B (2007) The implications of proteolytic background for shotgun proteomics. Mol Cell Proteomics 6: 1589–1598.

[135] Burkhart JM, Schumbrutzki C, Wortelkamp S, Sickmann A, Zahedi RP (2012) Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. J Proteomics 75: 1454–1462.

[136] McDonald WH, Ohi R, Miyamoto DT, Mitchison TJ, Yates JR (2002) Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. International Journal of Mass Spectrometry 219: 245–251.

[137] Boersema PJ, Mohammed S, Heck AJ (2008) Hydrophilic interaction liquid chromatography (HILIC) in proteomics. Analytical and bioanalytical chemistry 391: 151–159.

[138] Glish GL, Vachet RW (2003) The basics of mass spectrometry in the twenty-first century. Nature Reviews Drug Discovery 2: 140–150.

[139] Karas M, Bachmann D, Bahr Ue, Hillenkamp F (1987) Matrix-assisted ultraviolet laser desorption of non-volatile compounds. International journal of mass spectrometry and ion processes 78: 53–68.

[140] Whitehouse CM, Dreyer RN, Yamashita M, Fenn JB (1985) Electrospray interface for liquid chromatographs and mass spectrometers. Analytical chemistry 57: 675–679.

[141] Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246: 64–71.

[142] Muth T, Benndorf D, Reichl U, Rapp E, Martens L (2013) Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. Mol Biosyst 9: 578–585.

[143] Purvine S, Kolker N, Kolker E (2004) Spectral quality assessment for high-throughput tandem mass spectrometry proteomics. OMICS 8: 255–265.

[144] Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, et al. (2006) Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics 5: 652–670.

[145] Salmi J, Moulder R, Filen JJ, Nevalainen OS, Nyman TA, et al. (2006) Quality classification of tandem mass spectrometry data. Bioinformatics 22: 400–406.

[146] Wong JW, Sullivan MJ, Cartwright HM, Cagney G (2007) msmsEval: tandem mass spectral quality assignment for high-throughput proteomics. BMC Bioinformatics 8: 51.

[147] Xu H, Freitas MA (2010) A dynamic noise level algorithm for spectral screening of peptide MS/MS spectra. BMC Bioinformatics 11: 436.

[148] Flikka K, Martens L, Vandekerckhove J, Gevaert K, Eidhammer I (2006) Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. Proteomics 6: 2086–2094.

[149] Flikka K, Meukens J, Helsens K, Vandekerckhove J, Eidhammer I, et al. (2007) Implementation and application of a versatile clustering tool for tandem mass spectrometry data. Proteomics 7: 3245–3258.

[150] Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, et al. (2008) Clustering millions of tandem mass spectra. J Proteome Res 7: 113–122.

[151] Falkner JA, Falkner JW, Yocum AK, Andrews PC (2008) A spectral clustering approach to MS/MS identification of post-translational modifications. J Proteome Res 7: 4614–4622.

[152] Guthals A, Watrous JD, Dorrestein PC, Bandeira N (2012) The spectral networks paradigm in high throughput mass spectrometry. Mol Biosyst 8: 2535–2544.

[153] Griss J, Foster JM, Hermjakob H, Vizcaino JA (2013) PRIDE Cluster: building a consensus of proteomics data. Nat Methods 10: 95–96.

[154] Daniel H, Moghaddas Gholami A, Berry D, Desmarchelier C, Hahne H, et al. (2014) High-fat diet alters gut microbiota physiology in mice. ISME J 8: 295–308.

[155] VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009) Systems biology: Functional analysis of natural microbial consortia using community proteomics. Nat Rev Microbiol 7: 196–205.

[156] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.

[157] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.

[158] Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133–141.

[159] Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31–46.

[160] Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15: 387–396.

[161] Hoff KJ, Lingner T, Meinicke P, Tech M (2009) Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res 37: W101–105.

[162] Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 38: e191.

[163] Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38: e132.

[164] Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res 40: e9.

[165] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32: D115–119.

[166] Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501–504.

[167] Seifert J, Herbst FA, Halkjaer Nielsen P, Planes FJ, Jehmlich N, et al. (2013) Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities. Proteomics 13: 2786–2804.

[168] Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5: 976–989.

[169] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551–3167.

[170] Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20: 1466–1467.

[171] Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, et al. (2004) Open mass spectrometry search algorithm. J Proteome Res 3: 958–964.

[172] Tabb DL, Fernando CG, Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res 6: 654–661.

[173] Park CY, Klammer AA, Kall L, MacCoss MJ, Noble WS (2008) Rapid and accurate peptide identification from tandem mass spectra. J Proteome Res 7: 3022–3027.

[174] Tanner S, Shu H, Frank A, Wang LC, Zandi E, et al. (2005) InsPecT: identification of post-translationally modified peptides from tandem mass spectra. Anal Chem 77: 4626–4639.

[175] Eng JK, Jahan TA, Hoopmann MR (2013) Comet: an open-source MS/MS sequence database search tool. Proteomics 13: 22–24.

[176] Kim S, Pevzner PA (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5: 5277.

[177] Dorfer V, Pichler P, Stranzl T, Stadlmann J, Taus T, et al. (2014) MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. J Proteome Res 13: 3679–3684.

[178] Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, et al. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10: 1794–1805.

[179] Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26: 1367–1372.

[180] Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L (2011) SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics 11: 996–999.

[181] Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74: 5383–5392.

[182] Choi H, Nesvizhskii AI (2008) Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J Proteome Res 7: 254–265.

[183] Victor B, Gabriel S, Kanobana K, Mostovenko E, Polman K, et al. (2012) Partially sequenced organisms, decoy searches and false discovery rates. J Proteome Res 11: 1991–1995.

[184] Kwon T, Choi H, Vogel C, Nesvizhskii AI, Marcotte EM (2011) MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines. J Proteome Res 10: 2949–2958.

[185] Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, et al. (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics 10: M111–007690.

[186] Elias DA, Monroe ME, Marshall MJ, Romine MF, Belieav AS, et al. (2005) Global detection and characterization of hypothetical proteins in Shewanella oneidensis MR-1 using LC-MS based proteomics. Proteomics 5: 3120–3130.

[187] Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4: 207–214.

[188] Martens L, Hermjakob H (2007) Proteomics data validation: why all must provide data. Mol Biosyst 3: 518–522.

[189] Käll L, Storey JD, Noble WS (2009) QVALITY: non-parametric estimation of q-values and posterior error probabilities. Bioinformatics 25: 964–966.

[190] Wedge DC, Krishna R, Blackhurst P, Siepen JA, Jones AR, et al. (2011) FDRAnalysis: a tool for the integrated analysis of tandem mass spectrometry identification results from multiple search engines. J Proteome Res 10: 2088–2094.

[191] Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 4: 923–925.

[192] Brosch M, Yu L, Hubbard T, Choudhary J (2009) Accurate and sensitive peptide identification with Mascot Percolator. J Proteome Res 8: 3176–3181.

[193] Gonnelli G, Stock M, Verwaeren J, Maddelein D, De Baets B, et al. (2015) A decoy-free approach to the identification of peptides. J Proteome Res 14: 1792–1798.

[194] Colaert N, Degroeve S, Helsens K, Martens L (2011) Analysis of the resolution limitations of peptide identification algorithms. J Proteome Res 10: 5555–5561.

[195] Krug K, Nahnsen S, Macek B (2011) Mass spectrometry at the interface of proteomics and genomics. Mol Biosyst 7: 284–291.

[196] Blakeley P, Overton IM, Hubbard SJ (2012) Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. J Proteome Res 11: 5221–5234.

[197] Pevtsov S, Fedulova I, Mirzaei H, Buck C, Zhang X (2006) Performance evaluation of existing de novo sequencing algorithms. J Proteome Res 5: 3018–3028.

[198] Ning K, Ye N, Leong HW (2008) On preprocessing and antisymmetry in de novo peptide sequencing: improving efficiency and accuracy. J Bioinform Comput Biol 6: 467–492.

[199] Cantarel BL, Erickson AR, VerBerkmoes NC, Erickson BK, Carey PA, et al. (2011) Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. PLoS One 6: e27173.

[200] Frank A, Pevzner P (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem 77: 964–973.

[201] Ma B, Zhang K, Hendrie C, Liang C, Li M, et al. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 17: 2337–2342.

[202] Allmer J (2011) Algorithms for the de novo sequencing of peptides from tandem mass spectra. Expert Rev Proteomics 8: 645–657.

[203] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

[204] Shevchenko A, Sunyaev S, Loboda A, Bork P, Ens W, et al. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. Anal Chem 73: 1917–1926.

[205] Leprevost FV, Valente RH, Lima DB, Perales J, Melani R, et al. (2014) PepExplorer: a similarity-driven tool for analyzing de novo sequencing results. Mol Cell Proteomics 13: 2480–2489.

[206] Nesvizhskii AI (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics 73: 2092–2123.

[207] Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377–386.

[208] Huson DH, Mitra S (2012) Introduction to the analysis of environmental sequences: metagenomics with MEGAN. Methods Mol Biol 856: 415–429.

[209] Chourey K, Nissen S, Vishnivetskaya T, Shah M, Pfiffner S, et al. (2013) Environmental proteomics reveals early microbial community responses to biostimulation at a uranium- and nitrate-contaminated site. Proteomics 13: 2921–2930.

[210] Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637.

[211] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.

[212] Galperin MY, Makarova KS, Wolf YI, Koonin EV (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43: D261–269.

[213] Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, et al. (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res 42: D231–D239.

[214] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.

[215] Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24: 1650–1651.

[216] Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, et al. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8: R183.

[217] Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. (2000) InterPro–an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16: 1145–1150.

[218] Attwood TK, Beck ME, Bleasby AJ, Parry-Smith DJ (1994) PRINTS–a database of protein motif fingerprints. Nucleic Acids Res 22: 3590–3596.

[219] Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A 95: 5857–5864.

[220] Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, et al. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 3: 265–274.

[221] Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, et al. (2005) The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 33: D212–215.

[222] Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, et al. (2000) The Pfam protein families database. Nucleic Acids Res 28: 263–266.

[223] Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.

[224] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35: W182–185.

[225] Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611–1618.

[226] Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.

[227] Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39: D691–697.

[228] Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, et al. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 32: D438–442.

[229] Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36: D623–631.

[230] Caspi R, Altman T, Billington R, Dreher K, Foerster H, et al. (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 42: D459–471.

[231] Goh WW, Lee YH, Chung M, Wong L (2012) How advancement in biological network analysis methods empowers proteomics. Proteomics 12: 550–563.

[232] Vaudel M, Sickmann A, Martens L (2010) Peptide and protein quantification: a map of the minefield. Proteomics 10: 650–670.

[233] Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, et al. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 3: 1154–1169.

[234] Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, et al. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17: 994–999.

[235] Choi H, Fermin D, Nesvizhskii AI (2008) Significance analysis of spectral count data in label-free shotgun proteomics. Mol Cell Proteomics 7: 2373–2385.

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

[237] Colaert N, Vandekerckhove J, Gevaert K, Martens L (2011) A comparison of MS2-based label-free quantitative proteomic techniques with regards to accuracy and precision. Proteomics 11: 1110–1113.

[238] Nahnsen S, Bielow C, Reinert K, Kohlbacher O (2013) Tools for label-free peptide quantification. Mol Cell Proteomics 12: 549–556.

[239] Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, et al. (2006) Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J Proteome Res 5: 2339–2347.

[240] Colaert N, Gevaert K, Martens L (2011) RIBAR and xRIBAR: Methods for reproducible relative MS/MS-based label-free protein quantification. J Proteome Res 10: 3183–3189.

[241] Helsens K, Colaert N, Barsnes H, Muth T, Flikka K, et al. (2010) ms_lims, a simple yet powerful open source laboratory information management system for MS-driven proteomics. Proteomics 10: 1261–1264.

[242] Rauch A, Bellew M, Eng J, Fitzgibbon M, Holzman T, et al. (2006) Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. J Proteome Res 5: 112–121.

[243] Hartler J, Thallinger GG, Stocker G, Sturn A, Burkard TR, et al. (2007) MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data. BMC Bioinformatics 8: 197.

[244] Poullet P, Carpentier S, Barillot E (2007) myProMS, a web server for management and validation of mass spectrometry-based proteomic data. Proteomics 7: 2553–2556.

[245] Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, et al. (2008) OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics 9: 163.

[246] Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, et al. (2005) PRIDE: the proteomics identifications database. Proteomics 5: 3537–3545.

[247] Craig R, Cortens JP, Beavis RC (2004) Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 3: 1234–1242.

[248] Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, et al. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol 32: 223–226.

[249] Deutsch EW, Lam H, Aebersold R (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep 9: 429–434.

[250] Hill JA, Smith BE, Papoulias PG, Andrews PC (2010) ProteomeCommons.org collaborative annotation and project management resource integrated with the Tranche repository. J Proteome Res 9: 2809–2811.

[251] Ji L, Barrett T, Ayanbule O, Troup DB, Rudnev D, et al. (2010) NCBI Peptidome: a new repository for mass spectrometry proteomics data. Nucleic Acids Res 38: D731–735.

[252] Martens L (2013) Resilience in the proteomics data ecosystem: how the field cares for its data. Proteomics 13: 1548–1550.

[253] Vizcaino JA, Foster JM, Martens L (2010) Proteomics data repositories: providing a safe haven for your data and acting as a springboard for further research. J Proteomics 73: 2136–2146.

[254] Csordas A, Wang R, Rios D, Reisinger F, Foster JM, et al. (2013) From Peptidome to PRIDE: public proteomics data migration at a large scale. Proteomics 13: 1692–1695.

[255] Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaino JA (2015) Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. Proteomics 15: 930–950.

[256] Muth T, Behne A, Heyer R, Kohrs F, Benndorf D, et al. (2015) The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. J Proteome Res 14: 1557–1565.

[257] Helsens K, Martens L, Vandekerckhove J, Gevaert K (2007) MascotDatfile: an open-source library to fully parse and analyse MASCOT MS/MS search results. Proteomics 7: 364–366.

[258] Storey JD (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64: 479–498.

[259] Granholm V, Käll L (2011) Quality assessments of peptide-spectrum matches in shotgun proteomics. Proteomics 11: 1086–1093.

[260] Patient S, Wieser D, Kleen M, Kretschmann E, Jesus Martin M, et al. (2008) UniProtJAPI: a remote API for accessing UniProt data. Bioinformatics 24: 1321–1322.

[261] Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, et al. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics 4: 1265–1272.

[262] Zhang B, Chambers MC, Tabb DL (2007) Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. J Proteome Res 6: 3549–3557.

[263] Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23: 1282–1288.

[264] Codd EF (1970) A relational model of data for large shared data banks. Communications of the ACM 13: 377–387.

[265] De Vrieze J, Saunders AM, He Y, Fang J, Nielsen PH, et al. (2015) Ammonia and temperature determine potential clustering in the anaerobic digestion microbiome. Water Res 75: 312–323.

[266] Kolmeder CA, Ritari J, Verdam FJ, Muth T, Keskitalo S, et al. (2015) Colonic metaproteomic signatures of active bacteria and the host in obesity. Proteomics 15: 3544–3552.

[267] Verdam FJ, Fuentes S, de Jonge C, Zoetendal EG, Erbil R, et al. (2013) Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. Obesity (Silver Spring) 21: E607–615.

[268] Vaudel M, Burkhart JM, Breiter D, Zahedi RP, Sickmann A, et al. (2012) A complex standard for protein identification, designed by evolution. J Proteome Res 11: 5065–5071.

[269] Verhees CH, Kengen SWM, Tuininga JE, Schut GJ, Adams MWW, et al. (2003) The unique features of glycolytic pathways in Archaea. Biochem J 375: 231–246.

[270] Ettema TJG, de Vos WM, van der Oost J (2005) Discovering novel biology by in silico archaeology. Nat Rev Microbiol 3: 859–869.

[271] Wisniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. Nat Methods 6: 359–362.

[272] Chourey K, Jansson J, VerBerkmoes N, Shah M, Chavarria KL, et al. (2010) Direct cellular lysis/protein extraction protocol for soil metaproteomics. J Proteome Res 9: 6615–6622.

[273] Muth T, Kolmeder CA, Salojärvi J, Keskitalo S, Varjosalo M, et al. (2015) Navigating through metaproteomics data: A logbook of database searching. Proteomics 15: 3439–3453.

[274] Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, et al. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res 14: 169–181.

[275] Muth T, Weilnböck L, Rapp E, Huber CG, Martens L, et al. (2014) DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. J Proteome Res 13: 1143–1146.

[276] Mesuere B, Devreese B, Debyser G, Aerts M, Vandamme P, et al. (2012) Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. J Proteome Res 11: 5773–5780.

[277] Mesuere B, Debyser G, Aerts M, Devreese B, Vandamme P, et al. (2015) The Unipept metaproteomics analysis pipeline. Proteomics 15: 1437–1442.

[278] Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7: e1002195.

[279] Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, et al. (2011) Metagenomic biomarker discovery and explanation. Genome Biol 12: R60.

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

[281] Craig R, Beavis RC (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom 17: 2310–2316.

[282] Sennels L, Bukowski-Wills JC, Rappsilber J (2009) Improved results in proteomics by use of local and peptide-class specific false discovery rates. BMC bioinformatics 10: 179.

[283] Paulo JA (2013) Practical and Efficient Searching in Proteomics: A Cross Engine Comparison. WebmedCentral 4: WMCPLS0052.

[284] Walters WA, Xu Z, Knight R (2014) Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Lett 588: 4223–4233.

[285] Everett LJ, Bierl C, Master SR (2010) Unbiased statistical analysis for multi-stage proteomic search strategies. J Proteome Res 9: 700–707.

[286] Bern M, Kil YJ (2011) Comment on "Unbiased statistical analysis for multi-stage proteomic search strategies". J Proteome Res 10: 2123–2127.

[287] Jeong K, Kim S, Bandeira N (2012) False discovery rates in spectral identification. BMC Bioinformatics 13 Suppl 16: S2.

[288] Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW (2013) Combining results of multiple search engines in proteomics. Mol Cell Proteomics 12: 2383–2393.

[289] Lewis S, Csordas A, Killcoyne S, Hermjakob H, Hoopmann MR, et al. (2012) Hydra: a scalable proteomic search engine which utilizes the Hadoop distributed computing framework. BMC Bioinformatics 13: 324.

[290] Pratt B, Howbert JJ, Tasman NI, Nilsson EJ (2012) MR-Tandem: parallel X!Tandem using Hadoop MapReduce on Amazon Web Services. Bioinformatics 28: 136–137.

[291] Muth T, Peters J, Blackburn J, Rapp E, Martens L (2013) ProteoCloud: a full-featured open source proteomics cloud computing pipeline. J Proteomics 88: 104–108.

[292] Veenstra TD, Conrads TP, Issaq HJ (2004) What to do with "one-hit wonders"? Electrophoresis 25: 1278–1279.

[293] Gupta N, Pevzner PA (2009) False discovery rates of protein identifications: a strike against the two-peptide rule. J Proteome Res 8: 4173–4181.

[294] Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, et al. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 5: 3475–3490.

[295] Balgley BM, Laudeman T, Yang L, Song T, Lee CS (2007) Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. Mol Cell Proteomics 6: 1599–1608.

[296] Searle BC, Turner M, Nesvizhskii AI (2008) Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. J Proteome Res 7: 245–253.

[297] Jones AR, Siepen JA, Hubbard SJ, Paton NW (2009) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. Proteomics 9: 1220–1229.

[298] Klimek J, Eddes JS, Hohmann L, Jackson J, Peterson A, et al. (2008) The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. J Proteome Res 7: 96–103.

[299] Paulovich AG, Billheimer D, Ham AJL, Vega-Montoto L, Rudnick PA, et al. (2010) Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. Mol Cell Proteomics 9: 242–254.

[300] Gupta N, Bandeira N, Keich U, Pevzner PA (2011) Target-decoy approach and false discovery rate: when things may go wrong. J Am Soc Mass Spectrom 22: 1111–1120.

[301] Noble WS (2015) Mass spectrometrists should search only for peptides they care about. Nat Methods 12: 605–608.

[302] Choi H, Ghosh D, Nesvizhskii AI (2008) Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res 7: 286–292.

[303] Renard BY, Timm W, Kirchner M, Steen JAJ, Hamprecht FA, et al. (2010) Estimating the confidence of peptide identifications without decoy databases. Anal Chem 82: 4314–4318.

[304] Keich U, Kertesz-Farkas A, Noble WS (2015) Improved False Discovery Rate Estimation Procedure for Shotgun Proteomics. J Proteome Res 14: 3148–3161.

[305] Ossipova E, Fenyö D, Eriksson J (2006) Optimizing search conditions for the mass fingerprint-based identification of proteins. Proteomics 6: 2079–2085.

[306] Stead DA, Preece A, Brown AJP (2006) Universal metrics for quality assessment of protein identifications by mass spectrometry. Mol Cell Proteomics 5: 1205–1211.

[307] Vandermarliere E, Mueller M, Martens L (2013) Getting intimate with trypsin, the leading protease in proteomics. Mass Spectrom Rev 32: 453–465.

[308] Alves G, Yu YK (2013) Improving peptide identification sensitivity in shotgun proteomics by stratification of search space. J Proteome Res 12: 2571–2581.

[309] Alves P, Arnold RJ, Clemmer DE, Li Y, Reilly JP, et al. (2008) Fast and accurate identification of semi-tryptic peptides in shotgun proteomics. Bioinformatics 24: 102–109.

[310] Fannes T, Vandermarliere E, Schietgat L, Degroeve S, Martens L, et al. (2013) Predicting tryptic cleavage from proteomics data using decision tree ensembles. J Proteome Res 12: 2253–2259.

[311] Sowell SM, Abraham PE, Shah M, Verberkmoes NC, Smith DP, et al. (2011) Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. ISME J 5: 856–865.

[312] Kan J, Hanson TE, Ginter JM, Wang K, Chen F (2005) Metaproteomic analysis of Chesapeake Bay microbial communities. Saline Systems 1: 7.

[313] Klaassens ES, de Vos WM, Vaughan EE (2007) Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. Appl Environ Microbiol 73: 1388–1392.

[314] Lacerda CMR, Choe LH, Reardon KF (2007) Metaproteomic analysis of a bacterial community response to cadmium exposure. J Proteome Res 6: 1145–1152.

[315] da Veiga Leprevost F, Barbosa VC, Carvalho PC (2015) Using PepExplorer to Filter and Organize De Novo Peptide Sequencing Results. Curr Protoc Bioinformatics 51: 13.27.1–13.27.9.

[316] Ma B (2015) Novor: Real-Time Peptide de Novo Sequencing Software. J Am Soc Mass Spectrom 26: 1885–1894.

[317] Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, et al. (2009) IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. J Proteome Res 8: 3872–3881.

[318] Meyer-Arendt K, Old WM, Houel S, Renganathan K, Eichelberger B, et al. (2011) IsoformResolver: A peptide-centric algorithm for protein inference. J Proteome Res 10: 3060–3075.

[319] Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75: 4646–4658.

[320] Li YF, Arnold RJ, Li Y, Radivojac P, Sheng Q, et al. (2009) A bayesian approach to protein inference problem in shotgun proteomics. J Comput Biol 16: 1183–1193.

[321] Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, et al. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics 8: 2405–2417.

[322] Searle BC (2010) Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics 10: 1265–1269.

[323] Dayhoff MO, Schwartz RM (1978) A model of evolutionary change in proteins. In: In Atlas of protein sequence and structure. volume 5, pp. 345–358.

[324] Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89: 10915–10919.

[325] Zhang Y, Xu T, Shan B, Hart J, Aslanian A, et al. (2015) ProteinInferencer: Confident protein identification and multiple experiment comparison for large scale proteomics projects. Journal of proteomics 129: 25–32.

[326] Penzlin A, Lindner MS, Doellinger J, Dabrowski PW, Nitsche A, et al. (2014) Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. Bioinformatics 30: i149–i156.

[327] Püttker S, Kohrs F, Benndorf D, Heyer R, Rapp E, et al. (2015) Metaproteomics of activated sludge from a wastewater treatment plant–A pilot study. Proteomics 15: 3596–3601.

[328] Jagtap P, McGowan T, Bandhakavi S, Tu ZJ, Seymour S, et al. (2012) Deep metaproteomic analysis of human salivary supernatant. Proteomics 12: 992–1001.

[329] Karlsson R, Davidson M, Svensson-Stadler L, Karlsson A, Olesen K, et al. (2012) Strain-level typing and identification of bacteria using mass spectrometry-based proteomics. J Proteome Res 11: 2710–2720.

[330] Musso G, Gambino R, Cassader M (2011) Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes. Annu Rev Med 62: 361–380.

[331] Cani PD, Delzenne NM (2009) Interplay between obesity and associated metabolic disorders: new insights into the gut microbiota. Curr Opin Pharmacol 9: 737–743.

[332] Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, et al. (2005) Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 102: 11070–11075.

[333] Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, et al. (2008) Evolution of mammals and their gut microbes. Science 320: 1647–1651.

[334] Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334: 105–108.

[335] De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, et al. (2010) Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A 107: 14691–14696.

[336] Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet 9: 509–515.

[337] Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5: e1000605.

[338] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.

[339] Munk B, Bauer C, Gronauer A, Lebuhn M (2012) A metabolic quotient for methanogenic Archaea. Water Sci Technol 66: 2311–2317.

[340] Traversi D, Villa S, Lorenzi E, Degan R, Gilli G (2012) Application of a real-time qPCR method to measure the methanogen concentration during anaerobic digestion as an indicator of biogas production capacity. J Environ Manage 111: 173–177.

[341] Morris R, Schauer-Gimenez A, Bhattad U, Kearney C, Struble CA, et al. (2014) Methyl coenzyme M reductase (mcrA) gene abundance correlates with activity measurements of methanogenic H2 /CO2 -enriched anaerobic biomass. Microb Biotechnol 7: 77–84.

[342] Pan C, Banfield JF (2014) Quantitative metaproteomics: functional insights into microbial communities. Methods Mol Biol 1096: 231–240.

[343] Chassard C, Lacroix C (2013) Carbohydrates and the human gut microbiota. Current Opinion in Clinical Nutrition & Metabolic Care 16: 453–460.

[344] Koropatkin NM, Cameron EA, Martens EC (2012) How glycan metabolism shapes the human gut microbiota. Nature Reviews Microbiology 10: 323–335.

[345] Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480–484.

[346] Booijink CCGM, Boekhorst J, Zoetendal EG, Smidt H, Kleerebezem M, et al. (2010) Metatranscriptome analysis of the human fecal microbiota reveals subject-specific expression profiles, with genes encoding proteins involved in carbohydrate metabolism being dominantly expressed. Appl Environ Microbiol 76: 5533–5540.

[347] Neis EPJG, Dejong CHC, Rensen SS (2015) The role of microbial amino acid metabolism in host metabolism. Nutrients 7: 2930–2946.

[348] Leimena MM, Ramiro-Garcia J, Davids M, van den Bogert B, Smidt H, et al. (2013) A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. BMC Genomics 14: 530.

[349] Jalanka-Tuovinen J, Salojärvi J, Salonen A, Immonen O, Garsed K, et al. (2014) Faecal microbiota composition and host-microbe cross-talk following gastroenteritis and in postinfectious irritable bowel syndrome. Gut 63: 1737–1745.

[350] Goecks J, Nekrutenko A, Taylor J, et al. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11: R86.

[351] Johnson J, Gottschalk B, Onsongo G, Bandhakavi S, et al. (2014) The Galaxy Framework as a Unifying Bioinformatics Solution for omics Core Facilities. Journal of biomolecular techniques: JBT 25: S5–S6.

[352] Jagtap PD, Blakely A, Murray K, Stewart S, Kooren J, et al. (2015) Metaproteomic analysis using the Galaxy framework. Proteomics 15: 3553–3565.

[353] Scott KP, Gratz SW, Sheridan PO, Flint HJ, Duncan SH (2013) The influence of diet on the gut microbiota. Pharmacol Res 69: 52–60.

[354] Gerritsen J, Smidt H, Rijkers GT, de Vos WM (2011) Intestinal microbiota in human health and disease: the impact of probiotics. Genes Nutr 6: 209–240.

[355] Gupta S, Allen-Vercoe E, Petrof EO (2016) Fecal microbiota transplantation: in perspective. Therap Adv Gastroenterol 9: 229–239.

Appendices

Table A.1: The relationship types and descriptions for the graph database schema are shown. *Out* and *In* are defined as nodes with outgoing and incoming relationship direction. Table adapted from Muth *et al.* [256].

Relationship type	Relationship description
HAS_PEPTIDE	Proteins that share certain peptides; Out: Proteins; In: Peptides.
IS_MATCH_IN	PSMs that match for peptides. Out: PSMs; In: Peptides.
BELONGS_TO	Proteins that belong to certain taxonomies; Out: Proteins; In: Taxonomies.
BELONGS_TO_ENZYME	Proteins that fulfill an enzymatic function; Out: Proteins; In: Enzymes.
BELONGS_TO_PATHWAY	Proteins that are part of certain pathways; Out: Proteins; In: Pathways.
INVOLVED_IN_BIOPROCESS	Proteins that are involved in biological processes; Out: Proteins; In: Ontologies (Biological Process).
HAS_MOLECULAR_FUNCTION	Proteins that have certain molecular functions; Out: Proteins; In: Ontologies (Molecular Function).
BELONGS_TO_CELL_COMP	Proteins that belong to cellular components; Out: Proteins; In: Ontologies (Cellular Component).
IS_SUPERGROUP_OF	Relationship to reflect the enzyme (EC) hierarchy; Out: Enzymes; In: Enzymes.
IS_ANCESTOR_OF	Relationship for the taxonomic hierarchy (reaching from superkingdom to species); Out: Enzymes; In: Enzymes.
IS_METAPROTEIN_OF	Relationship between protein groups and proteins; Out: Proteins; In: Proteins.

Table A.2: Number of identified peptides that could be uniquely mapped against an *in silico* digest of the respective subset database (MS/MS spectra, percentage of identified spectra, spectrum and peptide identifications using X!Tandem and OMSSA for HIMP10 data sets searched against HIMPdb (FDR < 1%). Table adapted from Muth *et al.* [273].

Data set	Total	ID (%)	Peptides	Spectrum IDs		Peptide IDs	
				X!Tandem	OMSSA	X!Tandem	OMSSA
P1	35 179	22.2	6 274	7 353	4 915	5 268	3 930
P3	26 560	17.0	4 260	4 291	4 105	3 366	3 351
P8	31 891	21.6	5 651	6 665	4 750	4 692	3 752
P11	31 744	17.6	4 590	5 366	3 134	3 907	2 532
P17	32 203	22.3	5 997	6 849	4 058	5 140	3 425
P23	34 050	24.2	6 278	8 030	3 726	5 553	3 049
P27	27 339	17.2	3 966	4 475	3 512	3 338	2 745
P28	32 037	21.4	5 353	6 470	3 464	4 527	2 801
P31	35 848	24.4	6 987	8 504	5 807	5 766	4 532
P34	30 524	22.0	5 400	6 453	3 371	4 678	2 804
Average	31 737	21.2	5 476	6 446	4 084	4 624	3 292

Table A.3: Number of identified peptides from HIMP data sets P1-P34 that could be uniquely mapped against an *in silico* digest of the respective databases (FDR < 5%). Database search was performed by using X!Tandem and OMSSA against the listed databases representing the subsets of the HIMPdb database. Table adapted from Muth *et al.* [273].

Dataset	Bact594	Qin2010	Kurokawa2007	Human2010	Human2010old	Food
P1	376	3 177	98	3	10	40
P3	241	1 996	55	4	7	24
P8	315	2 765	87	4	10	42
P11	218	2 712	45	8	14	23
P17	286	2 887	94	3	10	26
P23	328	2 886	123	2	7	27
P27	220	2 091	71	5	7	25
P28	280	2 368	63	5	4	43
P31	370	2 577	113	3	11	50
P34	318	2 102	64	1	9	31

Table A.4: Proportion of UniProtKB/TrEMBL peptide identifications for GENT01, GENT07 and GENT16 (FDR < 5%) that could be matched against UniProtKB/SwissProt and BGPMG.

Dataset	Total	UniProtKB/SwissProt	BGPMG
GENT01	721	266 (36.9%)	95 (13.2%)
GENT07	417	158 (37.9%)	83 (19.9%)
GENT16	1 910	282 (14.8%)	141 (7.4%)

Table A.5: Number of PSMs and peptides identified by searching HIMP data sets P1, P23 and P34 with X!Tandem and OMSSA against HIMPdb, Bact594db and Qin2010db. In addition, two-step searching against HIMPdb was performed (FDR < 5%). Table adapted from Muth *et al.* [273].

Dataset	HIMPdb		Bact594db		Qin2010db		HIMPdb Two-step	
	PSMs	Peptides	PSMs	Peptides	PSMs	Peptides	PSMs	Peptides
P1	11 133	8 473	6 694	4 841	10 562	8 012	24 469	27 136
P23	11 288	8 255	6 655	4 614	10 428	7 730	28 132	31 913
P34	9 491	7 231	6 274	4 686	8 425	6 599	24 888	27 769

Table A.6: Number of Bact594db-specific PSMs and peptides identified by database searches against HIMPdb and Bact594db for HIMP data sets P1, P23 and P34 at 1% and 5% FDR. Table adapted from Muth *et al.* [273].

Dataset	1% FDR				5 % FDR			
	HIMPdb		Bact594db		HIMPdb		Bact594db	
	PSMs	Peptides	PSMs	Peptides	PSMs	Peptides	PSMs	Peptides
P1	224	158	404	293	488	376	758	645
P23	197	137	335	246	423	328	712	600
P34	157	109	314	227	395	318	625	549

Table A.7: Number of Qin2010db specific PSMs and peptides identified by database searches against HIMPdb and Qin2010db for data sets P1, P23 and P34 at 1% and 5% FDR. Table adapted from Muth *et al.* [273].

Dataset	1% FDR				5 % FDR			
	HIMPdb		Qin2010db		HIMPdb		Qin2010db	
	PSMs	Peptides	PSMs	Peptides	PSMs	Peptides	PSMs	Peptides
P1	2 851	2 013	3 424	2 423	4 296	3 177	4 556	3 487
P23	2 661	1 894	3 191	2 316	3 817	2 886	4 156	3 234
P34	1 687	1 333	2 071	1 612	2 608	2 102	2 841	2 346

Table A.8: Number of identifications for data sets P1, P23, and P34 using *chymotrypsin* and *pepsin A* cleavage settings at 1% and 5% FDR. Table adapted from Muth *et al.* [273].

Dataset	1% FDR				5 % FDR			
	chymotrypsin		pepsin A		chymotrypsin		pepsin A	
	PSMs	Peptides	PSMs	Peptides	PSMs	Peptides	PSMs	Peptides
P1	90	91	88	88	90	91	88	88
P23	63	62	66	67	63	62	66	67
P34	47	47	63	64	47	47	63	64

Table A.9: PFU result score threshold values using X!Tandem and OMSSA for searching against Pyrodb, PyroHIMPdb and PyroHIMPdb (two-step searching) at 1% and 5% FDR.

Database	X!Tandem		OMSSA	
	1% FDR	5% FDR	1% FDR	5% FDR
Pyrodb	25.9	21.1	7.4	-9.9
PyroHIMPdb	43.4	37.6	29.3	10.2
PyroHIMPdb (two-step searching)	27.6	21.8	7.3	-11.2

Table A.10: Number of species-specific peptides identifications for data sets 9MM_FASP and 9MM_PPID using MPA and UniPept at 1% and 5% FDR. Taxon filter (TF) thresholds of 0.5% and 3% were employed.

Species	1% FDR					5 % FDR			
		MPA		UniPept		MPA		UniPept	
	TF	0.5%	3%	0.5%	3%	0.5%	3%	0.5%	3%
Escherichia coli		71	0	0	0	117	0	0	0
Lactobacillus acidophilus		185	185	56	56	230	230	60	60
Lactobacillus casei		97	0	0	0	125	0	0	0
Pasteurella multocida		554	554	191	191	675	675	232	232
Pediococcus pentosaceus		150	150	58	58	178	178	63	63
Saccharomyces cerevisiae		1 222	1 222	54	54	1 573	1 573	69	69
Incorrect		338	0	43	40	632	0	106	46

Table A.11: Number of genus-specific peptides identifications for data sets 9MM_FASP and 9MM_PPID using MPA and UniPept at 1% and 5% FDR. Taxon filter (TF) thresholds of 0.5% and 3% were employed.

Genus	1% FDR					5 % FDR			
		MPA		UniPept		MPA		UniPept	
	TF	0.5%	3%	0.5%	3%	0.5%	3%	0.5%	3%
Brevibacillus		31	0	0	0	40	0	0	0
Escherichia		75	0	0	125	0	0	0	
Lactobacillus		512	512	259	259	645	645	309	309
Pasteurella		554	554	247	247	675	675	296	296
Pediococcus		155	155	94	94	184	184	103	103
Saccharomyces		1 227	1 227	471	471	1 580	1 580	581	581
Incorrect		356	129	40	40	867	185	178	0

Table A.12: The 20 most abundant NOGs with respect to assigned proteins for the HIMP data set P1. The identifications were obtained by searching with X!Tandem and OMSSA against HIMPdb at 5% FDR.

Category	Description	Proteins	Peptides
E	Glutamate dehydrogenase	149	337
G	Catalyzes the reversible conversion of 2PG into PEP	104	150
C	Phosphoenolpyruvate Carboxylase	92	119
G	Catalyzes the conversion of L-arabinose to L-ribulose	88	141
G	Uronic isomerase	85	151
G	Solute-binding protein	76	210
G	Converts the aldose L-fucose into ketose L-fuculose	71	142
G	Phosphohexose isomerase	69	132
G	Extracellular solute-binding protein family 1	67	186
G	Xylose Isomerase	66	99
I	Acetyl-coa acetyltransferase	52	87
C	Alcohol dehydrogenase	51	87
G	Catalyzes the interconversion of 2PG and 3PG	38	79
G	glyceraldehyde-3-phosphate dehydrogenase	37	50
C	Transferase	29	49
C	Dehydrogenase	28	63
G	Solute-binding protein	28	34
M	(No annotation provided)	24	84
E	M18 family aminopeptidase	24	49

Table A.13: The 20 most abundant NOGs with respect to assigned proteins for the HIMP data set P23. The identifications were obtained by searching with X!Tandem and OMSSA against HIMPdb at 5% FDR.

Category	Description	Proteins	Peptides
E	Glutamate dehydrogenase	113	228
G	Solute-binding protein	110	290
G	Xylose Isomerase	77	110
G	Uronic isomerase	69	119
G	Extracellular solute-binding protein family 1	59	187
G	Catalyzes the conversion of L-arabinose to L-ribulose	55	103
C	Alcohol dehydrogenase	54	93
G	Phosphohexose isomerase	53	99
G	Converts the aldose L-fucose into ketose L-fuculose	48	108
E	Aminoacyl-histidine dipeptidase	48	64
E	M18 family aminopeptidase	44	96
C	Phosphoenolpyruvate Carboxylase	41	69
G	glyceraldehyde-3-phosphate dehydrogenase	41	46
G	Catalyzes the interconversion of 2PG and 3PG	40	93
S	Basic membrane	32	65
G	Catalyzes the reversible conversion of 2PG into PEP	32	51
G	Extracellular solute-binding protein family 1	31	94
I	Acetyl-coa acetyltransferase	30	48
G	Alpha amylase, catalytic	27	80

Table A.14: Peptide and protein assignments to EggNOG categories for the HIMP data set P1. The identifications were obtained by searching with X!Tandem and OMSSA against HIMPdb at 1% and 5% FDR.

Category Name	1% FDR		5 % FDR	
	Proteins	Peptides	Proteins	Peptides
Carbohydrate transport and metabolism	965	1 873	1 189	2 540
Function unknown	298	755	373	1 035
Amino acid transport and metabolism	324	709	454	1 059
Energy production and conversion	250	451	316	635
Lipid transport and metabolism	85	144	96	189
Cell wall/membrane/envelope biogenesis	44	138	53	182
Inorganic ion transport and metabolism	33	99	38	133
Metabolites synthesis, transport and catabolism	15	53	17	61
PTM, protein turnover, chaperones	14	33	19	50
Cell motility	9	28	16	45
Coenzyme transport and metabolism	11	27	13	33
Transcription	8	14	9	19
Trafficking, secretion, and vesicular transport	7	12	7	13
Signal transduction mechanisms	1	8	1	8
Nucleotide transport and metabolism	3	5	7	12
Defense mechanisms	1	5	1	7
Translation, ribosomal structure and biogenesis	1	2	1	2
Cell cycle control and division	1	2	1	2

Table A.15: Peptide and protein assignments to EggNOG categories for HIMP data set P23. The identifications were obtained by searching with X!Tandem and OMSSA against HIMPdb at 1% and 5% FDR.

Category Name	1% FDR		5 % FDR	
	Proteins	Peptides	Proteins	Peptides
Carbohydrate transport and metabolism	986	2 140	1 156	2 768
Function unknown	369	883	450	1 171
Amino acid transport and metabolism	370	787	486	1 138
Energy production and conversion	191	377	235	515
Cell wall/membrane/envelope biogenesis	26	120	29	142
Lipid transport and metabolism	43	92	51	114
Inorganic ion transport and metabolism	15	53	21	75
PTM, protein turnover, chaperones	18	42	19	52
Nucleotide transport and metabolism	17	36	21	51
Metabolites synthesis, transport and catabolism	15	34	20	50
Cell motility	16	28	20	37
Transcription	6	18	6	20
Coenzyme transport and metabolism	2	4	2	4
Signal transduction mechanisms	1	2	2	4
Trafficking, secretion, and vesicular transport	0	0	2	4

Table A.16: Number of phylum-level peptide identifications for HIMP10 data sets that could be functionally assigned to the bacterial EggNOG database.

Data set	*Actinobacteria*	*Bacteroidetes*	*Firmicutes*	*Proteobacteria*
P1	117	370	782	6
P3	144	89	528	6
P8	278	144	717	3
P11	114	115	375	2
P17	115	333	736	3
P23	109	302	818	2
P27	65	104	447	5
P28	108	256	713	31
P31	138	688	856	19
P34	273	325	624	18
Average	146	273	660	10

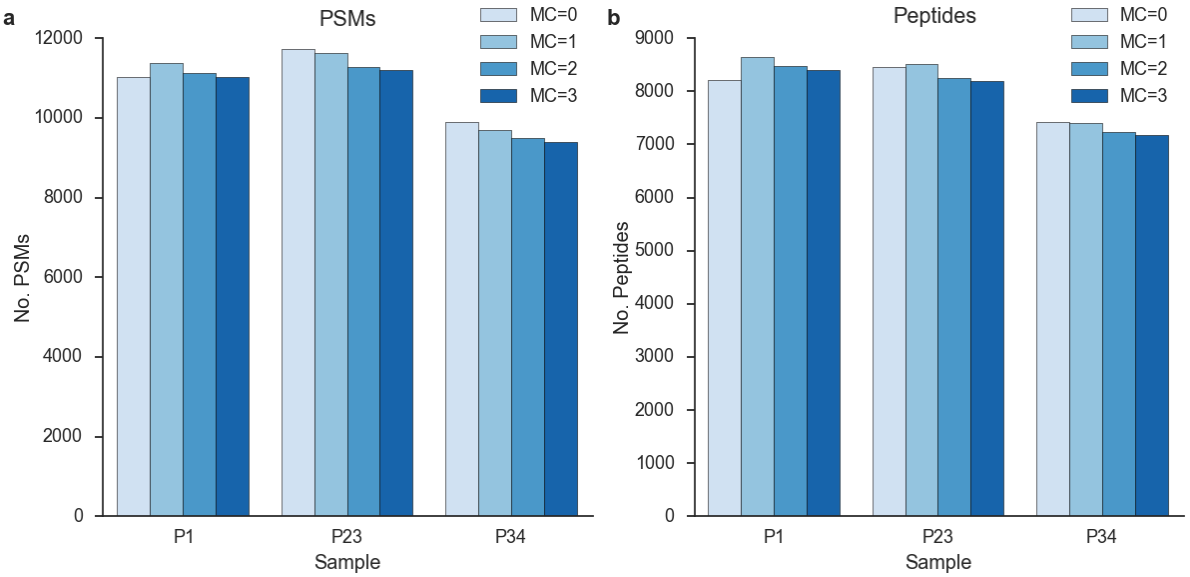

Figure A.1: Comparative evaluation of the identification yield for different MC values (HIMP). The bar plots show the total number of (a) PSMs and (b) peptides for data sets P1, P23, and P34 when using missed cleavage parameter values *MC = 0 - 3* at 5% FDR. Identification results were combined from searching with X!Tandem and OMSSA against HIMPdb.

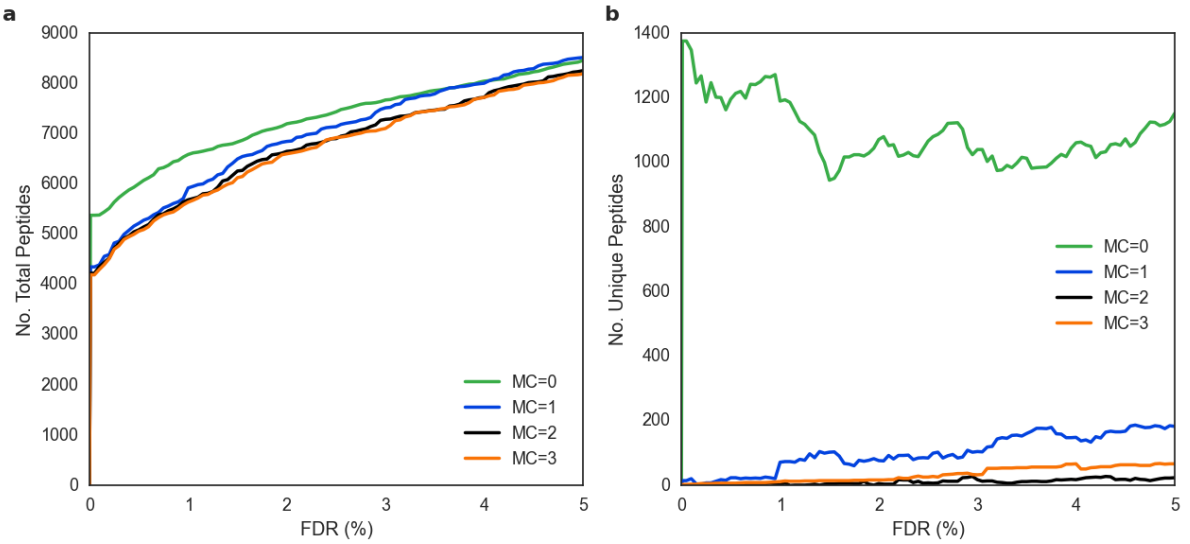

Figure A.2: Comparison of total and exclusive peptides for different MC values (P23). The line charts display the number of (a) total and (b) exclusive peptides for data set P23 when using missed cleavage parameter values *MC = 0 - 3* as a function of the respective FDR threshold. Peptides were called exclusive when being identified uniquely for a particular MC parameter value.

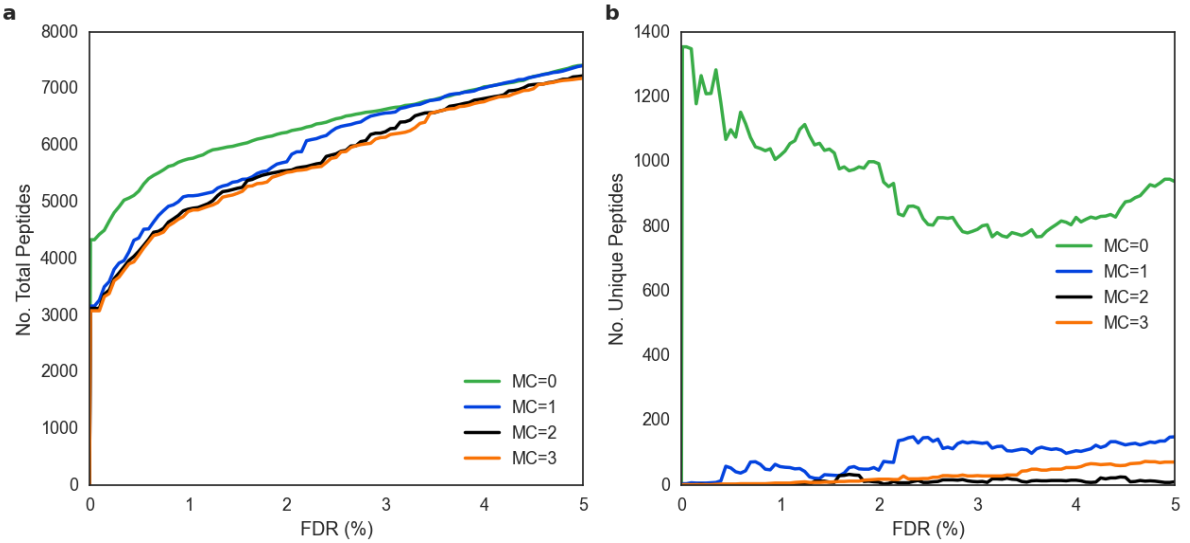

Figure A.3: Comparison of total and exclusive peptides for different MC values (P34). The line charts display the number of (a) total and (b) exclusive peptides for data set P1 when using missed cleavage parameter values *MC = 0 - 3* as a function of the respective FDR threshold. Peptides were called exclusive when being identified uniquely for a particular MC parameter value.

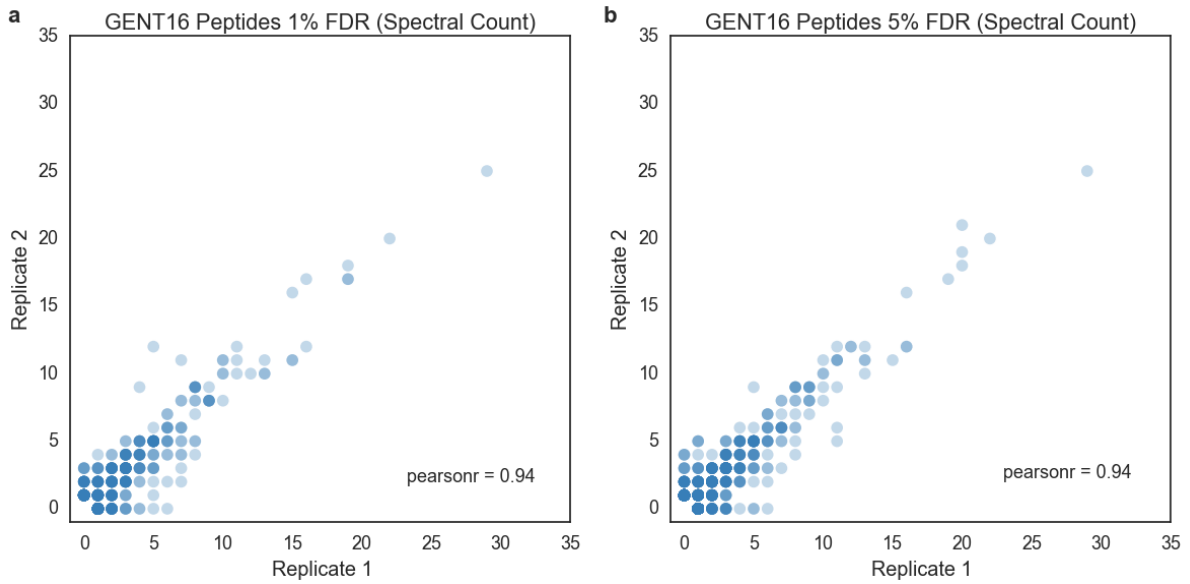

Figure A.4: Reproducibility of peptide hits between technical replicates for GENT16. The plots compare the peptides that were reproducibly identified between GENT16 replicates on the basis of the spectral count at (a) 1% and (b) 5% FDR. The color scale represents the number of identified peptides; low amounts in bright blue and high amounts in dark blue, respectively. The Pearson correlation coefficient (pearsonr) is displayed in the lower right corner of each panel.

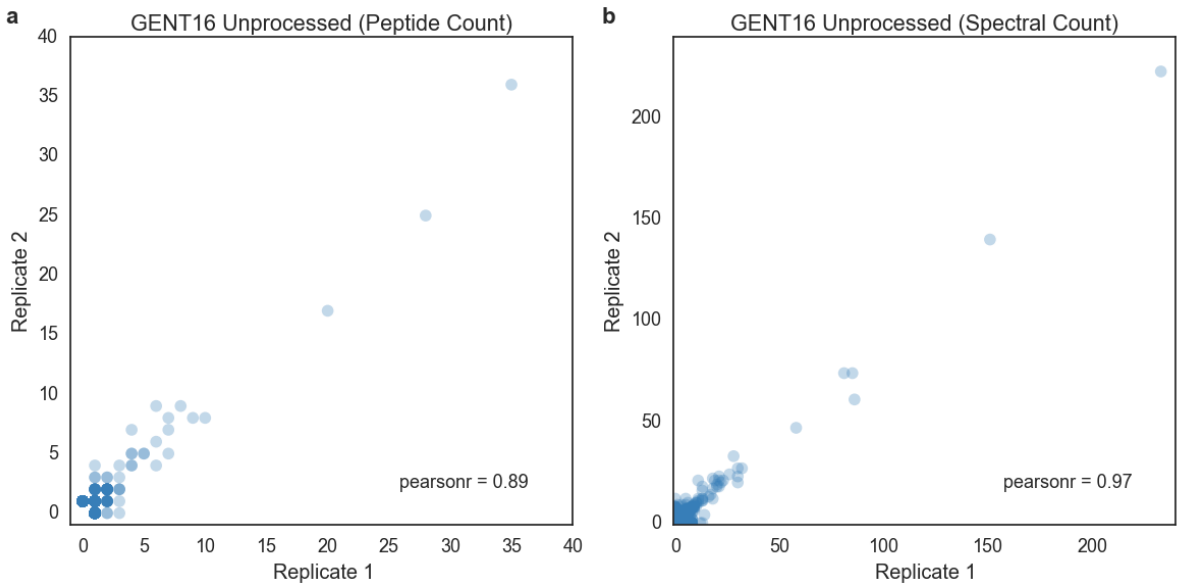

Figure A.5: Reproducibility of protein hits between technical replicates for GENT16. The plots compare the proteins that were reproducibly identified between GENT16 replicates on the basis of their (a) peptide and (b) spectral count at 5% FDR. The color scale represents the number of identified proteins; low amounts in bright blue and high amounts in dark blue, respectively. The Pearson correlation coefficient (pearsonr) is displayed in the lower right corner of each panel.

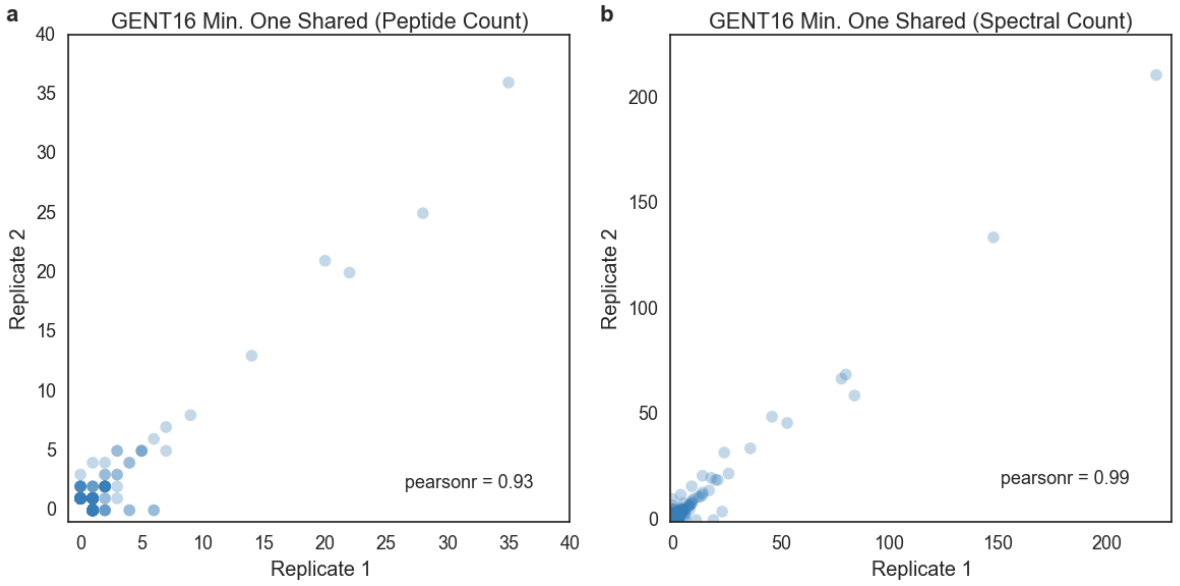

Figure A.6: Reproducibility of meta-protein hits between technical replicates for GENT16. The plots compare the meta-proteins that were reproducibly identified between GENT16 replicates on the basis of their (a) peptide and (b) spectral count at 5% FDR. Meta-proteins were generated by using the *Minimum One Shared* rule. The color scale represents the number of identified meta-proteins; low amounts in bright blue and high amounts in dark blue, respectively. The Pearson correlation coefficient (pearsonr) is displayed in the lower right corner of each panel.

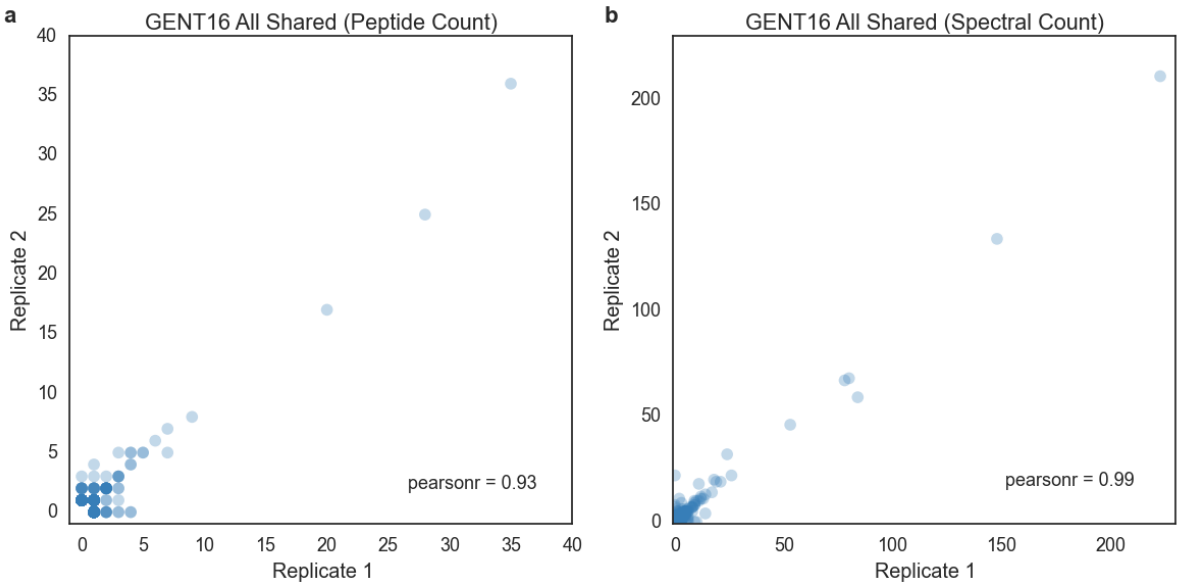

Figure A.7: Reproducibility of meta-protein hits between technical replicates for GENT16. The plots compare the meta-proteins that were reproducibly identified between GENT16 replicates on the basis of their (a) peptide and (b) spectral count at 5% FDR. Meta-proteins were generated by using the *All Shared* rule. The color scale represents the number of identified meta-proteins; low amounts in bright blue and high amounts in dark blue, respectively. The Pearson correlation coefficient (pearsonr) is displayed in the lower right corner of each panel.

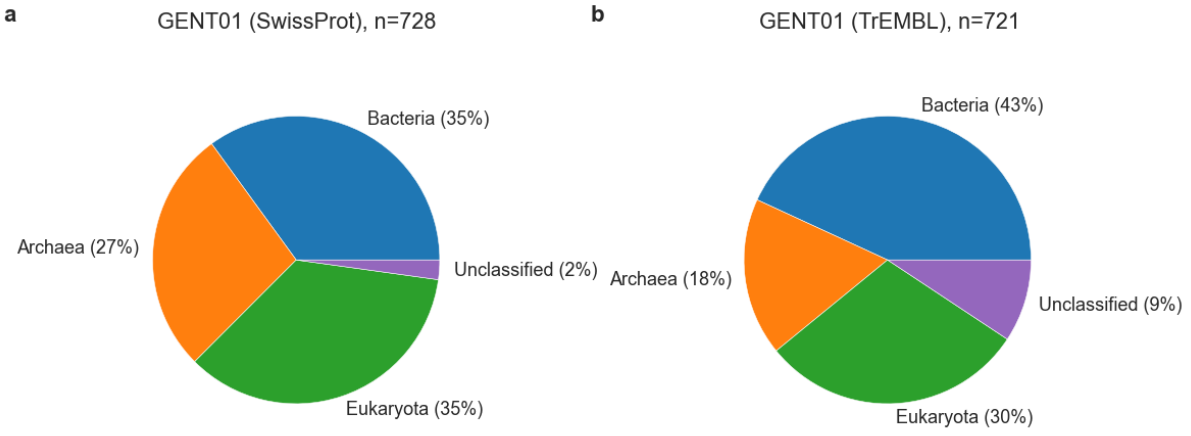

Figure A.8: Phylogenetic classification of BGP data set GENT01 based on number of peptides per superkingdom. The pie charts display the relative distribution of total peptide hits retrieved from (a) SwissProt and (b) TrEMBL searches. The total number of assigned peptides is provided above each chart panel (n).

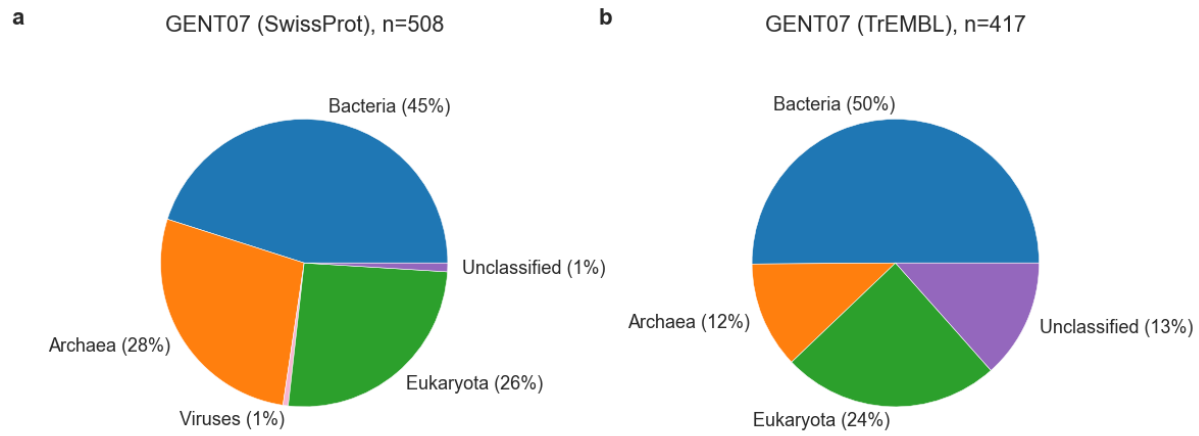

Figure A.9: Phylogenetic classification of BGP data set GENT07 based on number of peptides per superkingdom. The pie charts display the relative distribution of total peptide hits retrieved from (a) SwissProt and (b) TrEMBL searches. The total number of assigned peptides is provided above each chart panel (n).

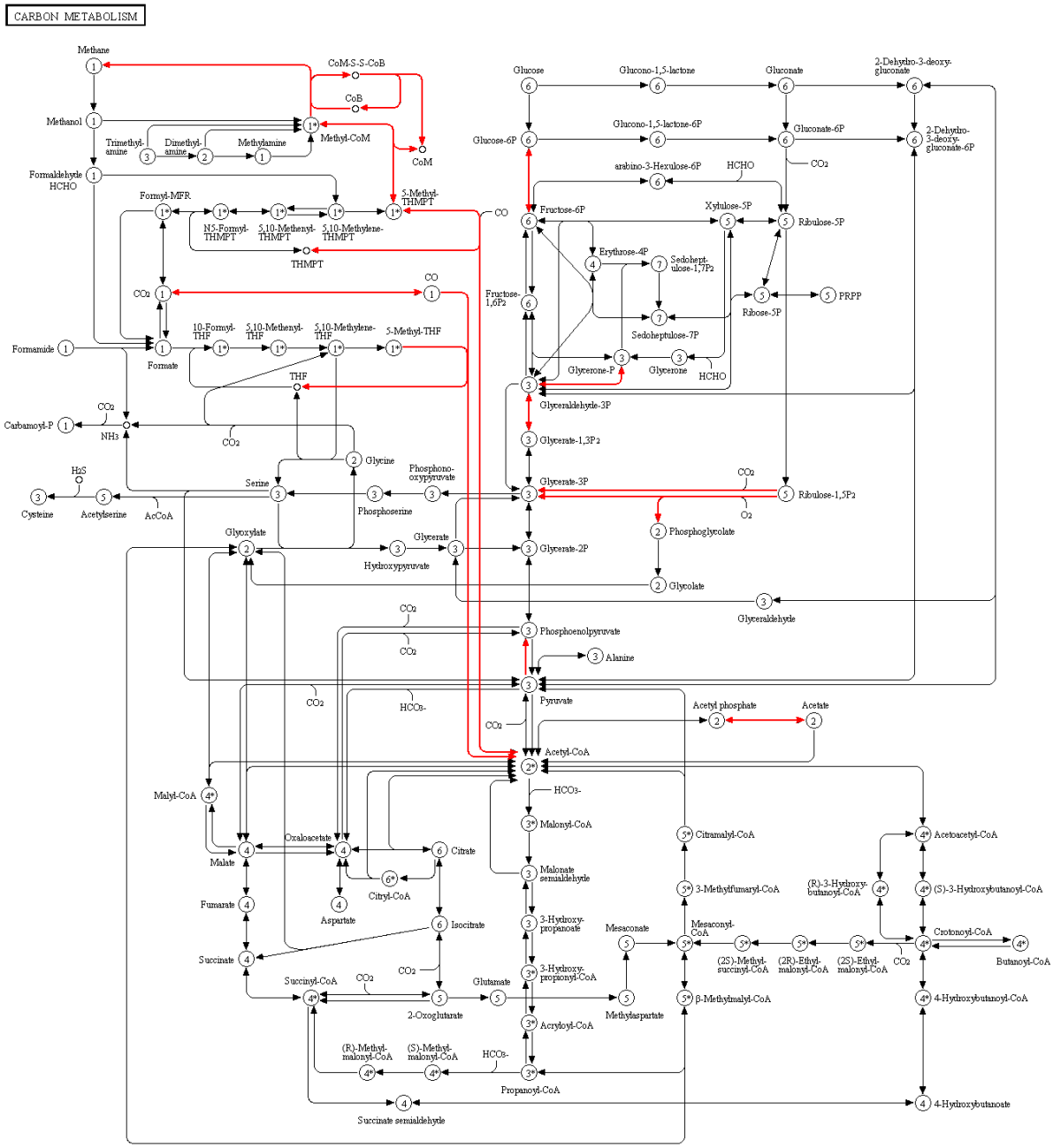

Figure A.10: KEGG reference pathway of carbon metabolism (map01200) for GENT01 protein hits from Archaea. The edges represent enzymes required for the conversion of one metabolite into another. The identified proteins of the data set are highlighted in red after submission to the KEGG website. The identifications were obtained by searching with X!Tandem and OMSSA against SwissProt at 5% FDR.

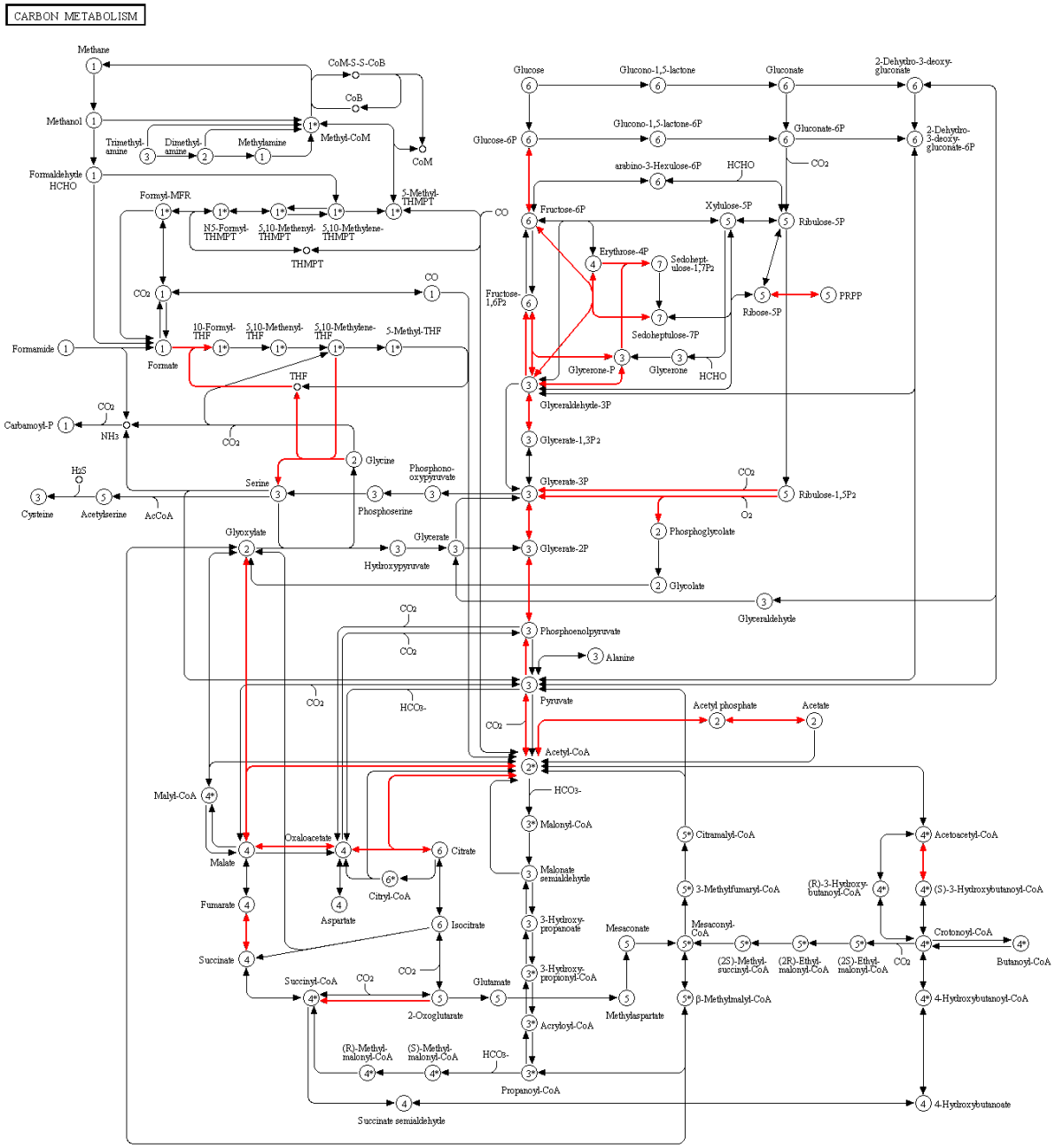

Figure A.11: KEGG reference pathway of carbon metabolism (map01200) for GENT01 protein hits from Bacteria. The edges represent enzymes required for the conversion of one metabolite into another. The identified proteins of the data set are highlighted in red after submission to the KEGG website. The identifications were obtained by searching with X!Tandem and OMSSA against SwissProt at 5% FDR.

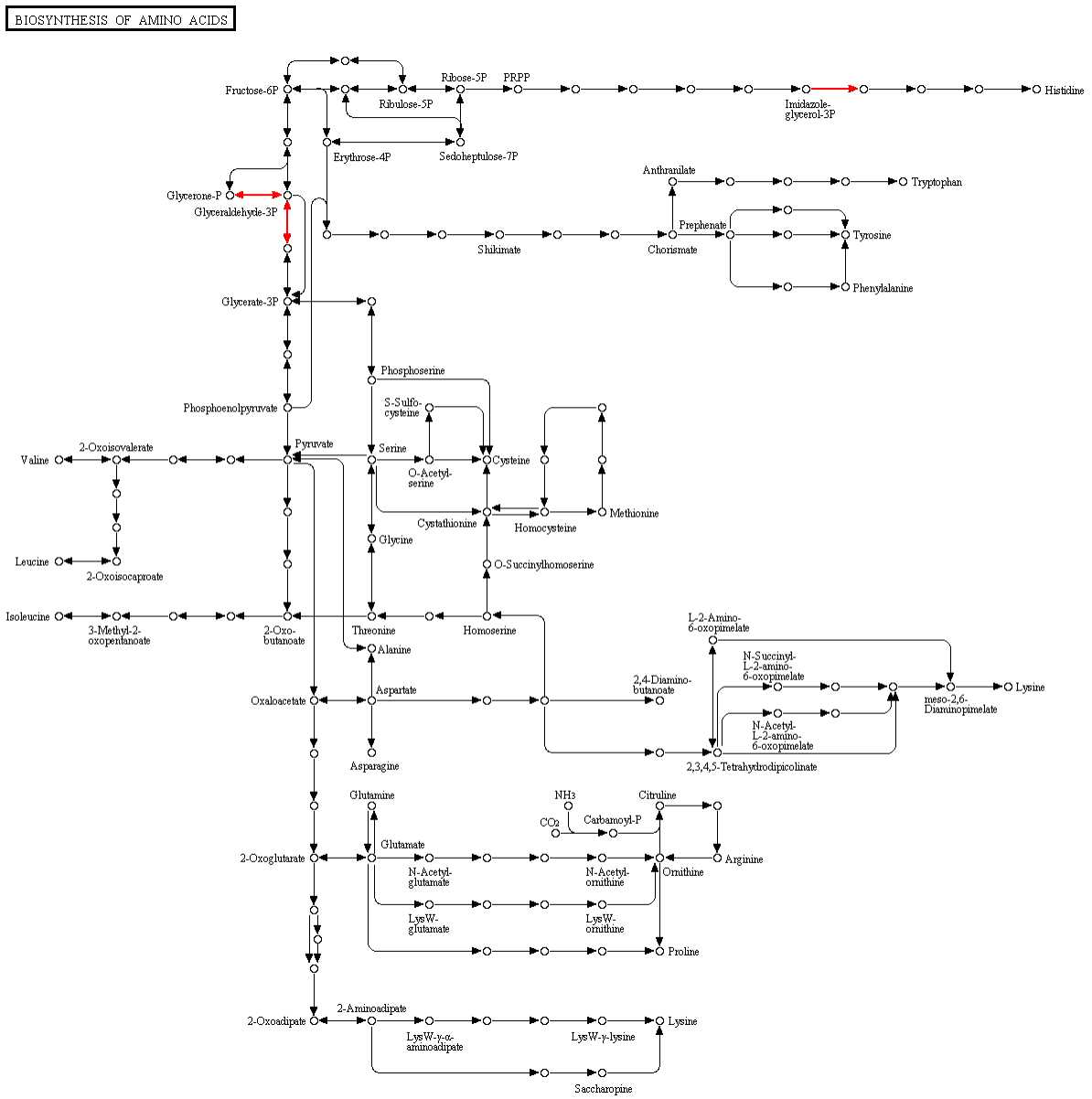

Figure A.12: KEGG reference pathway of amino acid synthesis (map01230) for GENT01 protein hits from Archaea. The edges represent enzymes required for the conversion of one metabolite into another. The identified proteins of the data set are highlighted in red after submission to the KEGG website. The identifications were obtained by searching with X!Tandem and OMSSA against SwissProt at 5% FDR.

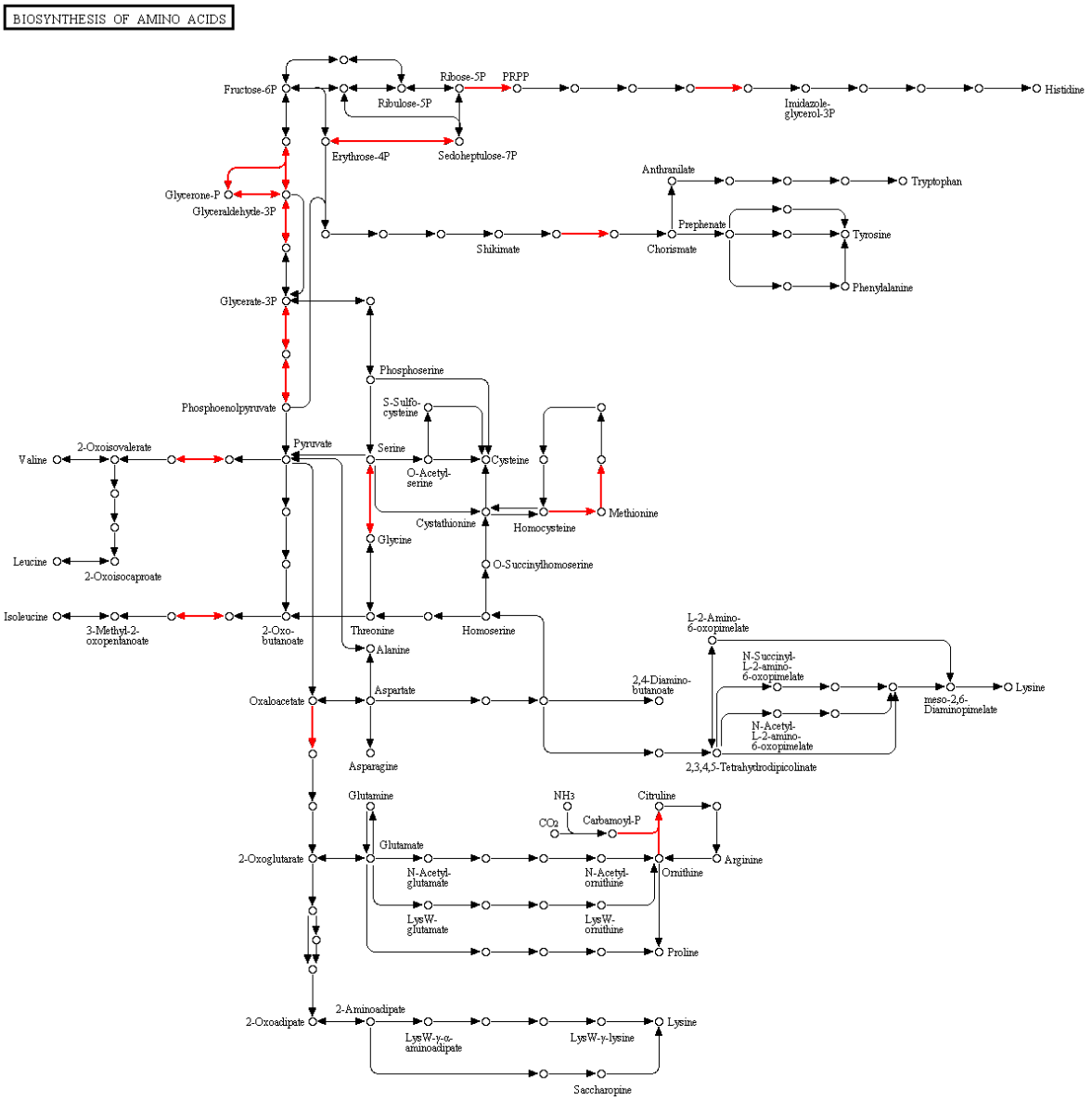

Figure A.13: KEGG reference pathway of amino acid synthesis (map01230) for GENT01 protein hits from Bacteria.. The edges represent enzymes required for the conversion of one metabolite into another. The identified proteins of the data set are highlighted in red after submission to the KEGG website. The identifications were obtained by searching with X!Tandem and OMSSA against SwissProt at 5% FDR.

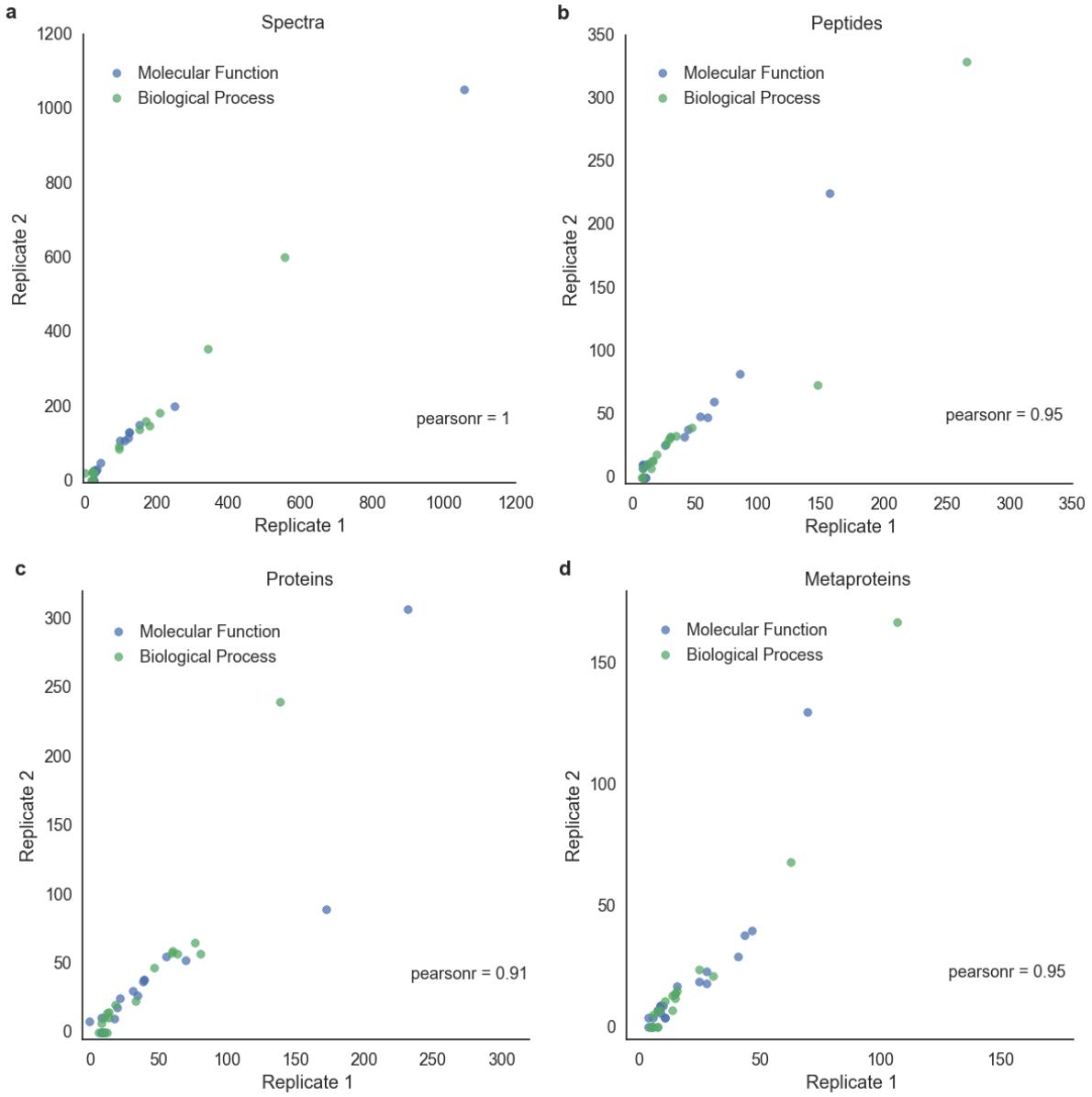

Figure A.14: Reproducibility of ontology-specific assignments across technical replicates for GENT16. Each scatter plot compares either the number of (a) spectra, (b) peptides, (c) proteins and (d) meta-proteins that were reproducibly assigned across two replicate experiments to the functional ontologies *Molecular Function* (blue) and *Biological Process* (green). The data set GENT16 was searched against SwissProt (5% FDR). Meta-proteins were generated by using the *Minimum One Shared* rule. The Pearson correlation coefficient (pearsonr) is displayed in the lower right corner of each panel.

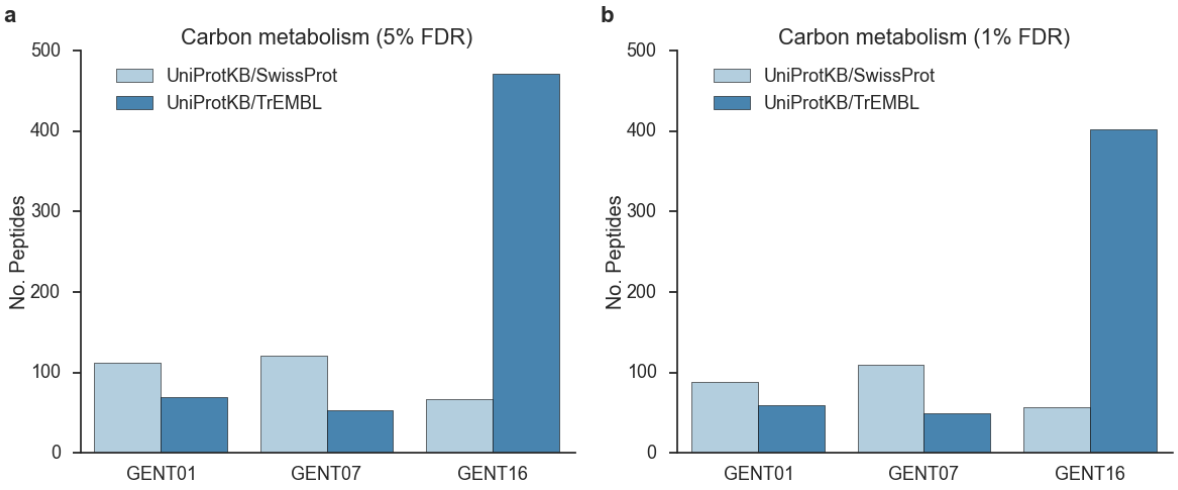

Figure A.15: Total number of "Carbon metabolism"-specific peptide assignments for BGP data sets. The bar plots show the total amount of peptide assignments to the pathway carbon metabolism (KEGG map01200) for GENT01, GENT07 and GENT16. The results were obtained by searching with X!Tandem and OMSSA against SwissProt and TrEMBL at (a) 5% and (b) 1% FDR.