Building the Unipept Ecosystem

Empowering high-throughput metaproteomics data analysis for characterizing complex microbial communities

Pieter Verschaffelt

Acknowledgements

This is where I will thank a lot of people!

Summary

Samenvatting

List of publications

Robert Gurdeep Singh, Alessandro Tanca, Antonio Palomba, Felix Van der Jeugt, **Pieter Verschaffelt**, Sergio Uzzau, Lennart Martens, Peter Dawyndt, and Bart Mesuere. 2018. "Unipept 4.0: functional analysis of metaproteome data" *Journal of Proteome Research* 18 (2): 606-615.

Tim Van Den Bossche, **Pieter Verschaffelt**, Kay Schallert, Harald Barsnes, Peter Dawyndt, Dirk Benndorf, Bernhard Y Renard, Bart Mesuere, Lennart Martens, and Thilo Muth. 2020. "Connecting MetaProteomeAnalyzer and PeptideShaker to Unipept for seamless end-to-end metaproteomics data analysis" *Journal of Proteome Research* 19 (8): 3562-3566.

Pieter Verschaffelt, Philippe Van Thienen, Tim Van Den Bossche, Felix Van der Jeugt, Caroline De Tender, Lennart Martens, Peter Dawyndt, and Bart Mesuere. 2020. "Unipept CLI 2.0: adding support for visualizations and functional annotations" *Bioinformatics* 36 (14): 4220-4221.

Christopher Ashwood, Wout Bittremieux, Eric W Deutsch, Nadezhda T Doncheva, Viktoria Dorfer, Ralf Gabriels, Vladimir Gorshkov, Surya Gupta, Andrew R Jones, Lukas Käll, Dominik Kopczynski, Lydie Lane, Ludwig Lautenbacher, Marc Legeay, Marie Locard-Paulet, Bart Mesuere, Yasset Perez-Riverol, Eugen Netz, Julianus Pfeuffer, Timo Sachsenberg, Renee Salz, Patroklos Samaras, Henning Schiebenhoefer, Tobias Schmidt, Veit Schwämmle, Alessio Soggiu, Julian Uszkoreit, Tim Van Den Bossche, Bart Van Puyvelde, Joeri Van Strien, **Pieter**

Verschaffelt, Henry Webel, and Sander Willems. 2020. "Proceedings of the EuBIC-MS 2020 Developers' Meeting" *EuPA Open Proteomics* 24: 1-6.

Pieter Verschaffelt, Tim Van Den Bossche, Lennart Martens, Peter Dawyndt, and Bart Mesuere. 2021. "Unipept Desktop: a faster, more powerful metaproteomics results analysis tool" *Journal of Proteome Research* 20 (4): 2005-2009.

Pieter Verschaffelt, Tim Van Den Bossche, Wassim Gabriel, Michał Burdukiewicz, Alessio Soggiu, Lennart Martens, Bernhard Y Renard, Henning Schiebenhoefer, and Bart Mesuere. 2021. "MegaGO: A fast yet powerful approach to assess functional gene ontology similarity across meta-omics data sets" *Journal of Proteome Research* 20 (4): 2083-2088.

Robert-Bob Hettich, Richard J Giannone, Paul Abraham, Tim Van Den Bossche, Benoit Kunath, Kay Schallert, Stephanie S Schäpe, Jean Armengaud, Magnus Arntzen, Ariane Bassignani, Dirk Benndorf, Stephan Fuchs, Timothy Griffin, Live Hagen, Rashi Halder, Celine Henry, Robert Heyer, Pratik Jagtap, Nico Jehmlich, Marlene Jensen, Catherine Juste, Manuel Kleiner, Olivier Langella, Pedro Queiros, Udo Reichl, Theresa Lehmann, Emma Leith, Patrick May, Bart Mesuere, Guylaine Miotello, Bernhard Renard, Henning Schiebenhoefer, Alexander Sczyrba, Alessandro Tanca, Kathrin Trappe, Jean-Pierre Trezzi, Sergio Uzzau, **Pieter Verschaffelt**, Martin Von Bergen, Paul Wilmes, Maximilian Wolf, Lennart Martens, and Thilo Muth. 2021. "Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows" *Nature Communications* 12 (1).

Pieter Verschaffelt, James Collier, Alexander Botzki, Lennart Martens, Peter Dawyndt, and Bart Mesuere. 2022. "Unipept Visualizations: an interactive visualization library for biological data" *Bioinformatics* 38 (2): 562-563.

Kay Schallert, **Pieter Verschaffelt**, Bart Mesuere, Dirk Benndorf, Lennart Martens, and Tim Van Den Bossche. 2022. "Pout2Prot: an efficient tool to create protein (sub)groups from percolator output files" *Journal of Proteome Research* 21 (4): 1175-1180.

Felix Van der Jeugt, Rien Maertens, Aranka Steyaert, **Pieter Verschaffelt**, Caroline De Tender, Peter Dawyndt, and Bart Mesuere. 2022. "UMGAP: the Unipept MetaGenomics Analysis Pipeline" *BMC genomics* 23 (1): 1-16.

Pieter Verschaffelt, Alessandro Tanca, Marcello Abbondio, Tim Van Den Bossche, Tibo Vande Moortele, Peter Dawyndt, Lennart Martens, and Bart Mesuere. 2023. "Support for novel proteogenomics analysis in Unipept" *Journal of Proteome Research* submitted (under review).

List of conferences and research stays

January 2020: *Nyborg, Denmark.* **EuBIC Developers' Meeting 2020**. Together with Henning Schiebenhoefer, Tim Van Den Bossche and Bart Mesuere, I was hosting a hackathon called "Mapping proteins to functions: method and benchmark development". During this hackathon, we developed a tool called "MegaGO" which has also been published in the Journal of Proteome Research. The package consists of both an interactive web application, as well as a command line tool.

December, 2020: *Online.* **VIB: Research Software Developers Day**. The Research Software Developers Day, organised by VIB, was an online conference that consisted of software developers presenting their workflow and ideas about the development of research software. I was one of the presenters on this day and performed a 20-minute talk about the development of the Unipept Desktop application. This talk was followed by 10 minutes of questions and discussion afterwards. A recording of the talk is available on YouTube: https://youtu.be/2ftK-fJkcJeY.

June, 2021: *Online.* **Online Metaproteomics Symposium**. A quick overview and update of recent developments in the field of metaproteomics and an introduction to the recently started Metaproteomics Iniative. I was only a spectator at this online conference and did not actively participate in any of the activities.

September, 2021: *Luxembourg, Luxembourg*. **Fourth Metaproteomics Symposium**. The International Metaproteomics Symposium is the leading event in the field of metaproteomics and related microbiome studies. The aim of the symposium is to provide a platform for the participants to share their latest results in their respective fields using metaproteomic methods as well as discussing recent technologic innovations and presenting newly developed bioinformatic tools. I presented the Unipept Desktop tool at this conference during a 20-minute presentation (with a 10 minute questions session afterwards).

March, 2022: *Lisbon, Portugal*. **EuBIC Winter School 2022**. The EuBIC-MS Winter School on computational mass spectrometry (MS) takes place every two years. Its aim is to bring together the users and developers of computational mass spectrometry tools, as well as academia and industry. The winter school starts with an educational day dedicated to workshops and trainings in established computation MS tools and workflows. The following days, internationally renowned invited speakers give lectures and practical workshops covering the aspects of identification, quantification, result interpretation, and integration of MS data. I presented a poster during this conference and participated in the workshops and talks given during this week.

September - October, 2022: *Berlin, Germany*. **Research stay at the Bundesanstalt für Materialforschung und -prüfung (BAM)**. In fall of 2022, I spent one month at the BAM institute in Berlin in the e-sciences group of professor Thilo Muth. Together with Tanja Holstein, I worked on a project to integrate PepGM with Unipept. PepGM is a probabilistic graphical model for taxonomic inference of proteome samples with associated confidence scores. It started out as a tool for the analysis of viral proteome samples and by better integrating

Unipept with PepGM, we were able to expand the tool with support for proper metaproteomics datasets.

December, 2022: *Cancun, Mexico*. **HUPO 2022 World Congress**. HUPO is the Human Proteome Organization which organizes a world congress somewhere around the world each year. During these conferences, lectures by world renowned speakers and exciting networking opportunities are offered. I presented a poster about proteogenomics analysis using the Unipept Desktop application at this conference.

January, 2023: *Ascona, Switzerland*. **EuBIC Developers' Meeting 2023**. The EuBIC-MS Developers Meeting is organized each second year. This meeting is aimed at bringing together computer scientists and developers in the field of mass spectrometry-related bioinformatics to discuss and work together in an open and constructive spirit. The program is split between keynote lectures and multiple hackathon sessions where the participants develop bioinformatics tools and resources addressing outstanding needs in the mass spectrometry-related bioinformatics and user community. Together with Tibo Vande Moortele and Tim Van Den Bossche, I hosted a hackathon session at this conference titled "Exploring and solving functional analysis gaps in metaproteomics". During this week, we've had a lively discussion with the members of our hackathon team to discuss these functional analysis gaps in metaproteomics and settled on improving support for highlighting taxonomic diversity of metaproteomic samples in metabolic pathways. We started the development of a web application that allows users to upload a list of peptides, which will be taxonomically analysed. All identified taxa will be highlighted on a metabolic pathway by using the KEGG database.

Contents

Chapter 1

Introduction

Part I

Unipept Desktop

The introductory text for this part comes here…

Chapter 2

Unipept Desktop: a faster, more powerful metaproteomics analysis tool

One of the major tools that I developed as part of my PhD, is the Unipept Desktop application. Version 1.0 of the Unipept Desktop application was first released on January 14, 2021 and improves upon the Unipept Web application in a number of different ways. Since a desktop application has access to all compute resources available on a user's machine, it enables analysing much larger metaproteomics datasets and drastically improves organisation of samples and metadata. I started working on the Unipept Desktop application in August 2019 and I am still the lead developer up to this day. Therefore, I am the main author of this manuscript as well as the desktop tool itself.

This chapter contains a verbatim copy of the application note by (Verschaffelt, Van Den Bossche, Martens, et al.*, 2021) as published in Journal of Proteome Research.*

Abstract — Metaproteomics has become an important research tool to study microbial systems, which has resulted in increased metaproteomics data generation. However, efficient tools for processing the acquired data have lagged behind. One widely used tool for metaproteomics data interpretation is Unipept, a web-based tool that provides, amongst others, interactive and insightful visualizations. Due to its web-based implementation, however, the Unipept web application is limited in the amount of data that can be analyzed. In this manuscript we therefore present Unipept Desktop, a desktop application version of Unipept that is designed to drastically increase the throughput and capacity of metaproteomics data analysis. Moreover, it provides a novel comparative analysis pipeline and improves the organization of experimental data into projects, thus addressing the growing need for more performant and versatile analysis tools for metaproteomics data.

2.1 Introduction

Metaproteomics is a relatively young research field that focuses on the study of microbial environments and complex ecosystems, and of the interactions between the organisms involved, through the analysis of the proteins extracted from these environments. Over the past years, the technology to identify proteins from such complex samples has been greatly improved, allowing metaproteomics to transition from relatively small studies to large scale experiments (Rechenberger *et al.*, 2019; Wilmes *et al.*, 2015). The key enabling technologies for this transition are improved mass spectrometers and more powerful proteomics approaches, which have both come a long way since the introduction of metaproteomics analysis in 2004 (Rodríguez-Valera, 2004; Yates, 2019). To allow efficient processing of the resulting increase of acquired data, various dedicated tools have been made available to support metaproteomics data analysis (Muth *et al.*, 2015; Van Den Bossche *et al.*, 2020), but even with this increased bioinformatics support, many challenges still need to be overcome, especially regarding downstream analysis of the obtained identifications (Schiebenhoefer *et al.*, 2019).

Unipept is a leading tool for such downstream metaproteomics data analysis (Herbst *et al.*, 2016) that currently consists of a web application (Gurdeep Singh *et al.*, 2019), a web service, and a command line tool (Verschaffelt *et al.*, 2020). The Unipept web application provides users with the ability to analyze a metaproteomics sample and extract taxonomic and functional information from environmental samples derived from a variety of origins, ranging from the human gut to biogas plants. The Unipept web application provides users with interactive visualizations and allows them to, for example, filter out all functions that are associated with a specific taxon. Due to its web-based nature, however, the size and number of samples that can be analyzed by Unipept are limited. And while it is currently possible to analyze larger data sets using the Unipept CLI, this requires more sophisticated bioinformatics skills and does not provide the interactive link between taxa and functional annotations.

Because of the browser limitations, it can already take a substantial amount of time to process relatively small samples (e.g. containing up to a few thousand identified peptides) using Unipept, depending on the specific search configuration used. These limitations have become an issue, as the advances in metaproteomics have not only increased data set sizes, but have also increased the number of data sets that need to be processed (Zhang and Figeys, 2019).

In order to accommodate this evolution, the throughput of metaproteomics data analyses needs to increase as well, in turn requiring tools that are not constrained in the amount of memory and CPU resources they are allowed to consume. Moreover, analysis results also need to be retained for future reference, ideally in a project-based approach that can group multiple samples, and the corresponding results should be easily shareable with other researchers.

For specific applications, it is also important that all data is processed offline or on-site rather than being sent over the internet. For instance, sensitive medical data is often not allowed to be sent to external services for processing, but must be kept in-house to safeguard patient confidentiality and privacy.

All of the above issues need to be resolved in order to support the growing interest in, and reach of, metaproteomics. We therefore here present the Unipept Desktop Application, a novel cross-platform desktop application designed to specifically overcome these challenges while also retaining the functionality that exists in the current web app.

2.2 Implementation

The Unipept desktop application provides three different types of analyses: *i*) single assay analysis, *ii*) inter-assay comparative analysis, and *iii*) tryptic peptide analysis. The single assay analysis performs a full taxonomic and functional analysis of a single assay and corresponds to the default "metaproteomics analysis" as presented by the Unipept web application. The inter-assay comparative analysis on the other hand, provides the ability to explore similarities and differences between multiple assays. While the comparison of multiple assays was already possible with the Unipept web application, this was only available for a limited number of quite small assays due to strict memory constraints

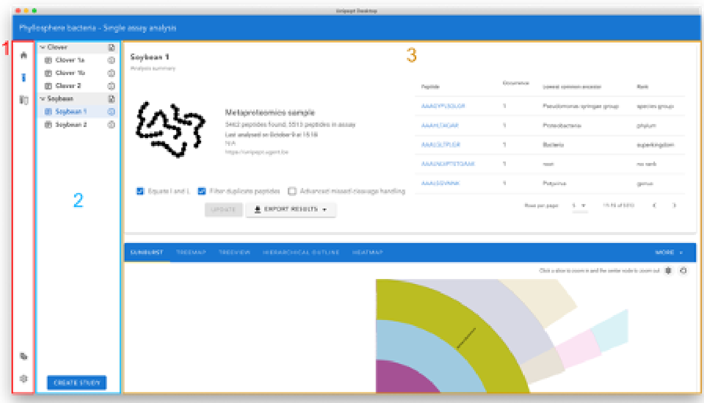


Figure 2.1: Screenshot of the Unipept Desktop application. The analysis page of the desktop application is depicted here and consists of three main parts: the sidebar that is used to navigate between the different analysis pipelines and functions of the application (1), the project explorer that displays a hierarchical view of the project (2) and the content view that renders analysis results (3).

posed by web browsers. The tryptic peptide analysis, lastly, can be used to look up which proteins, taxa and functions are associated with a given peptide.

Unipept Desktop delivers these core functions through a concise user interface (Figure 2.1) that consists of three main parts: the sidebar, the project explorer, and the content view. The sidebar on the far left allows the user to navigate between the different analysis pipelines and functions of this application. Directly to the right of the sidebar is the project explorer that allows the user to switch between assays, and to modify the project. The project explorer is only shown when performing single assay or comparative analyses. Assays and studies can be renamed or deleted by right clicking them, after which a context menu opens. Lastly, the content view takes up most of the application's visual space and presents either analysis results or the settings page.

The Unipept Desktop Application also allows offline analysis of data through a choice of the API endpoint in the settings menu. This endpoint, which uses the Unipept API and by default connects to the online Unipept system, can be configured to call any service that supports the Unipept API. By setting up a local instance of the Unipept backend system, the user can thus ensure that all data remains locally. Setting up a local Unipept back-end is possible by cloning the open source Unipept repository on GitHub, but requires advanced technical knowledge. We plan to make the installation process of these custom API endpoints even easier with future releases of Unipept.

Unipept Desktop is powered by the cross-platform Electron framework, which in itself is powered by Chromium browser technology. This means that the application is developed with web-centric technologies, such as the Vue frontend framework and TypeScript, and hence we were able to reuse large parts of the web app's codebase. The choice for the Electron platform was mostly driven by the extensive suite of different functionalities that can be integrated with minimum configuration efforts. Thanks to the Electron platform we can provide an automatic update mechanism, easily generate installation packages for all major platforms (Windows, macOS and Linux), and include automatic crash reporting, amongst others. Once installed, the Unipept Desktop application can thus update fully autonomously in the background, ensuring that users always have the latest functionality and bug fixes installed.

2.2.1 Project-centric analysis

The Unipept Desktop Application has full access to the local filesystem. Hence, it can store an arbitrary amount of data and does not need to worry about strict size limits; this in contrast to web applications that are only allowed to store up to a few megabytes using the local storage API. This allows us to improve upon the organization of data sets by introducing project-based data management capabilities. In accordance with the terminology introduced by the ISA-tab standard for experimental metadata annotation (Sansone *et al.*, 2012), we now refer to a data set derived from a sample as an “assay”, while a study is a grouping of multiple, related assays, and a Unipept project represents a collection of such studies.

On the file system, a project is stored in a single folder that contains an SQLite database file, a subfolder for each study and one text file per assay, located in the subfolder of the corresponding study. This folder can be modified outside of the application, using the default file explorer application of your operating system, thus providing maximum flexibility. All changes made to this project folder are automatically detected and imported by the application, granting users the ability to mass import assays and edit project properties with external applications. The application accepts simple text files with one peptide per line. In order to quantify peptide occurrence, a peptide can be included more than once in this file and the “filter duplicate peptides” option should be disabled for the analysis.

Because projects are folder-based, they can contain both the raw input data as well as the analysis results for an assay, making it practical for users to share projects with each other, for instance, in the form of compressed project folders. In addition, previously performed analyses do not need to be recomputed when the application is restarted, as opposed to analyses that were run on the Unipept website, which need to be recomputed every time the website is closed.

2.2.2 Comparative analysis

The Unipept Desktop Application provides both intra-assay and inter-assay comparative analyses that are rendered as heatmap visualizations. The intra-assay comparison can be started from the single assay analysis page by selecting the heatmap tab and provides a wizard to guide users through the set-up process of the comparison (Figure 2). Users are required to select two types of data sources (one for each axis of the heatmap) and indicate which items should be compared. Four different data sources are currently supported: NCBI taxa, GO terms (The Gene Ontology Consortium, 2019), EC numbers and InterPro entries (Finn *et al.*, 2017).

The inter-assay comparative analysis is designed to visualize differences and similarities in functional or taxonomic composition of multiple assays. Here too, users are presented with a wizard that is similar to the one found in the intra-assay comparison. For inter-assay comparisons, however, the horizontal axis of the heatmap is reserved for the set of selected assays, and users can therefore only select one collection of

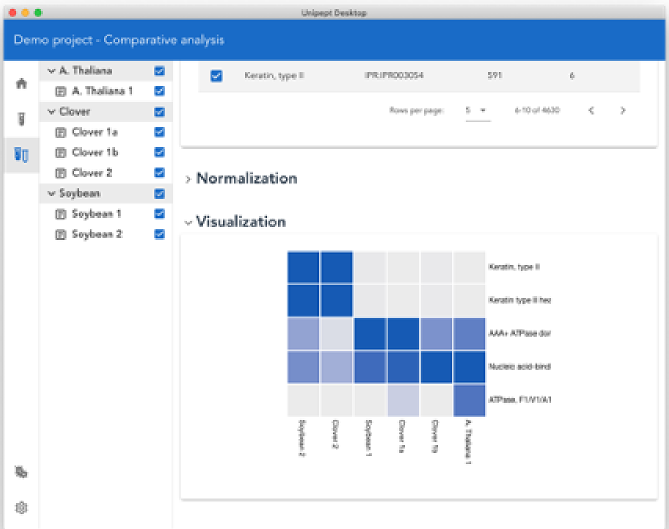


Figure 2.2: Screenshot of the inter-assay comparative analysis pipeline. Note that it is possible to select multiple assays from the project explorer. A heatmap is constructed from the set of items that were selected for comparison at the top of the page.

items that should be compared between the different assays.

Because the number of peptides can drastically differ between multiple assays, three different normalization techniques are provided to the user. The default setting normalizes the heatmap globally, i.e. the minimum and maximum values over the complete grid are computed and all grid values are normalized with respect to these values. The other two normalization techniques also normalize based on minimum and maximum values, but restricted within a row or column, respectively.

It is worth noting that, while the comparative analysis pipeline was originally designed for the Unipept Desktop Application, a slimmed-down version has meanwhile also been integrated into the Unipept web app.

With the advent of the Unipept Desktop Application, users now have a variety of ways in which they can use Unipept. A comparison between the various functionalities offered by these different services is provided in Table 2.1.

2.3 Conclusion

Unipept Desktop is a novel desktop application that extends upon the Unipept web application by eradicating the strict limitations posed by the web-based nature of this application to increase metaproteomics

	desktop app	web app	CLI	API
visualizations	✓	✓	∼	∼
basic metaproteomics analysis pipeline	✓	✓	✗	✗
tryptic peptide analysis pipeline	✓	✓	✗	✗
comparative analysis	✓	∼	✗	✗
metadata or projects	✓	✗	✗	✗
custom endpoint	✓	✗	✓	✗
store analysis results	✓	✗	∼	✗
process large samples	✓	✗	✓	✓
no command line knowledge required	✓	✓	✗	✗
no installation required	✗	✓	✗	✓

Table 2.1: Comparison of the functionalities provided by the different Unipept services.

data analysis throughput. Moreover, the Unipept Desktop Application adds new features such as allowing users to structure their data in a hierarchical project-based system, to keep track of their analysis results, and to share or distribute these results very easily. Whereas the Unipept web application is limited to assays with up to 50 000 peptides, the Unipept Desktop Application supports assays containing one million peptides or more. For reference, the desktop app can analyze between 250 and 2000 peptides per second (without advanced missed cleavage handling enabled), depending on the type of assay that's being analyzed.

In a future release of the Unipept Desktop Application, we plan to provide support for the preparation of custom reference databases and further improve support for offline analysis. This will allow us to gradually evolve to a tool that is not only suitable for metaproteomics data analysis, but also for novel proteogenomics analysis techniques for complex environmental samples.

Our choice for the Electron framework proves to be very valuable as well, as a large portion of Unipept's codebase can thus be shared between the new desktop application and the existing web application. This in turn allows us to easily migrate (a slimmed-down version of) specific desktop features to the web app, and vice versa.

2.4 Availability

The source code for Unipept Desktop is open source and provided under the MIT license as a repository on GitHub: https://github.com/unipept/unipept-desktop. Pre-generated installers for Windows, macOS and Linux (AppImage format) can be downloaded from the release page of our GitHub repository. Installation instructions and documentation for the Unipept Desktop Application can be found on our website: https://unipept.ugent.be/desktop.

2.5 Acknowledgements

This work was supported by the Research Foundation—Flanders (FWO) [1164420N to P.V.; 12I5220N to B.M.; 1S90918N to T.V.D.B.; G042518N to L.M.].

Chapter 3

Unipept Desktop 2.0: construction of targeted reference protein databases for proteogenomics analyses

This chapter contains a verbatim copy of the manuscript by (Verschaffelt et al., *2023) as submitted to the BioRXiv preprint server.*

Abstract — Unipept Desktop 2.0 is the most recent iteration of the Unipept Desktop tool that adds support for the analysis of proteogenomics datasets. Unipept Desktop now supports the automatic construction of targeted protein reference databases that only contain proteins associated with a predetermined list of taxa. This improves both the taxonomic and functional resolution of a metaproteomic analysis and yields several technical advantages. By limiting the proteins present in a reference database, it is now also possible to perform (meta)proteogenomics analyses. Since the protein reference database now lives on the user's local machine, they have complete control over the database used during an analysis. Data does no longer need to be transmitted over the internet, decreasing the time required for an analysis and better safeguarding privacy sensitive data. As a proof of concept, we present a case study in which a human gut metaproteome dataset is analyzed with Unipept Desktop 2.0 using different targeted databases based on matched 16S rRNA gene sequencing data.

3.1 Introduction

The metaproteomics research discipline has undergone a big transition since the term was first introduced in 2004 (Wilmes and Bond, 2004). We have witnessed the evolution of metaproteomics from very small-scale experiments, in which three distinct proteins (Ram *et al.*, 2005) could be identified in an ecosystem, to a mature technology that is able to analyze more than 100 000 protein fragments (or peptides) from various environments.

Unipept (Gurdeep Singh *et al.*, 2019) was one of the first major tools in this promising new research field that could be used to analyze tryptic peptide-based metaproteomics samples. Originally starting as a web application, Unipept was quickly accompanied by an application programming interface (Mesuere *et al.*, 2016) (API) and command line interface (Verschaffelt *et al.*, 2020) (CLI) that respectively allow for embedding Unipept's analyses in other tools and analyzing larger samples directly from the command line. API usage metrics currently indicate that more than 500 000 requests are handled by Unipept on a monthly basis, acknowledging the importance of the tool in this field.

The advent of recent technological improvements in mass spectrometry and more powerful proteomics approaches have allowed metaproteomics to transition from small studies to large scale experiments

(Wilmes *et al.*, 2015; Rechenberger *et al.*, 2019). Due to Unipept's inherent web-based nature, it was limited in the size of the samples that could be analyzed because of browser-imposed restrictions on available compute resources. This led to the development of the Unipept Desktop application (Verschaffelt, Van Den Bossche, Martens, *et al.*, 2021) in 2020.

Unipept Desktop (Verschaffelt, Van Den Bossche, Martens, *et al.*, 2021) paved the way for the analysis of large metaproteomics samples (containing 500k peptides or more) and completely overhauled the way these metaproteomics samples can be organized with the introduction of projects and studies. Projects can easily be shared with other researchers, who no longer need to reanalyze samples and wait for the results to become available, or can be archived for later use. The Unipept Desktop application also introduces a new inter- and intra-sample comparison pipeline which allows users to gain insight into the taxonomic and functional shift within and between multiple samples.

Over the last few years, interest in a new research area, proteogenomics, has been growing. Proteogenomics can be thought of as a logical next step in researching complex microbial ecosystems and combines information from both metagenomics and metaproteomics experiments in order to overcome a few key problems that arise when working with metaproteomics data in isolation (Schiebenhoefer *et al.*, 2019).

The first major issue that we need to consider is the ever-growing size of the protein reference databases that are being used to match peptides with proteins. UniProtKB (The UniProt Consortium, 2021, 2019), a freely accessible database containing protein sequences, has seen a rapid increase in size over the last decade and has grown from approximately 19 million proteins in 2012 to 227 million proteins in 2022. Compared to the early days of Unipept, we are able to identify increasingly diverse species as a direct result of the increased size of the reference database, but this also comes with a few drawbacks. Each peptide that is presented to Unipept will be matched with all proteins in which this peptide occurs. All of these proteins are associated with a specific organism and such a peptide-based analysis thus results in a set of organisms from which this peptide could potentially originate. In order to increase insight of researchers into the taxonomic composition of a sample, Unipept summarizes all of this information and calculates the lowest common ancestor (LCA) of this set of organisms (i.e., NCBI taxa) for each peptide. If all of the matched organisms are evolutionarily close to each other, this works very well and the LCA of our matches will be of value. If, however, one or more of the matched organisms is very different from the others, the LCA will typically end up at the root or another very general taxon within the NCBI taxonomy (Figure 3.1).

Proteogenomics tries to overcome this problem by combining information from prior metagenomics experiments from the same environment with metaproteomics experiments. The metagenomics experiment is used to explore the taxonomic composition of an ecosystem, which subsequently guides the researcher to query only a subset of the reference database. In the case of shotgun metagenomics, DNA sequences identified by a metagenomics experiment can be used to build a customized protein reference database (Tanca *et al.*, 2016).

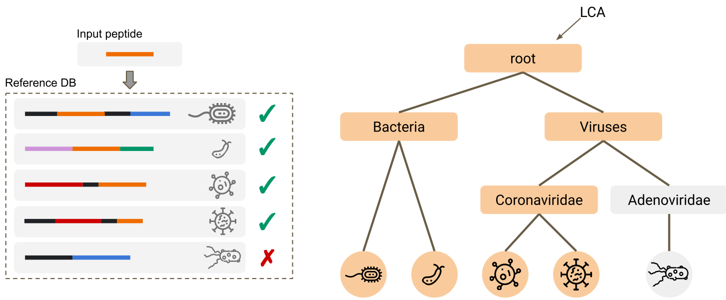
When using the Unipept Web application, analyses are always performed against the entire UniProtKB resource. We previously added the possibility to set up a local instance of Unipept and search against the database provided by this endpoint as part of the Unipept Desktop application, but setting up such a custom endpoint is often experienced as a big technical hurdle by our target audience. Most researchers are thus still dependent on the reference database provided by Unipept, including the database update scheme that Unipept dictates. This creates a set of problems and drawbacks that need to be overcome in order to properly support proteogenomics data analyses. Since users do not have control over the database that is being used, they cannot provide potential metagenomics information (such as taxa identified in the ecosystem under study) and restrict the search space of the reference database.

To solve this problem, we introduce version 2.0 of the Unipept Desktop application, which marks the beginning of a new era for the analysis of proteogenomics datasets. Unipept Desktop now provides support for the automatic construction of targeted protein reference databases on the user's local machine. Such targeted databases are based on a filtered version of UniProtKB and only contain UniProtKB records that are associated with the taxa provided by the user (Figure 3.2).

In this article, we discuss how these targeted protein reference databases are constructed by the Unipept Desktop application and how they can be queried efficiently on a user's machine. We also present a case study, based on a human gut metaproteome dataset obtained from 28 celiac patients, in which we investigate to what extent the accuracy of the analysis results improves by only taking into account a subset of the UniProtKB reference database.

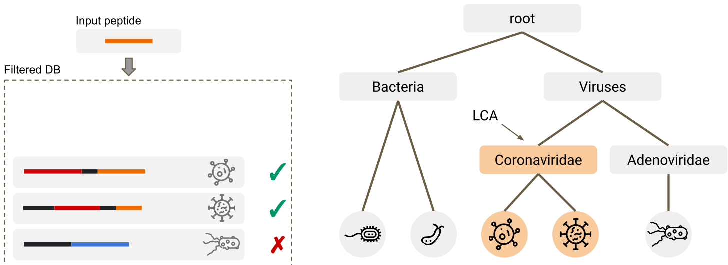
3.2 Unipept Desktop 2.0

In order to perform a proteogenomics experiment, researchers typically first perform a metagenomics experiment on the environment of interest, which then provides them with a set of organisms that are likely to be present in that environment. 16S rRNA gene sequencing and shotgun metagenomics are two widely used techniques in this



a) LCA computation performed against unfiltered UniProtKB reference database.

Figure 3.1: An example of how the lowest common ancestor (LCA) for a set of identified taxa is computed in Unipept Desktop, using an unfiltered protein reference database. A single input peptide is matched against proteins in a reference database. The taxa associated with all matched proteins are then summarized as the LCA, which is the most specific node in the taxonomy tree that is a parent of all matched taxa. An unfiltered protein reference database is used in the matching process. Since more proteins from a more diverse range of species are matched, the LCA ends up at the root of the taxonomy tree, providing little to no information. b) The reference database is restricted to viral proteins only and a much more specific LCA will be found (mapping onto the Coronaviridae family).



b) LCA computation performed against targeted reference database of viral proteins.

Figure 3.2: An example of how the lowest common ancestor (LCA) for a set of identified taxa is computed in Unipept Desktop, using a filtered protein reference database. The reference database is restricted to viral proteins only and a much more specific LCA will be found (mapping onto the Coronaviridae family). This example illustrate the importance of targeted protein reference databases when analyzing metaproteomics samples.

research area that provide the required information, but other types of analysis can also be used. Sometimes the taxonomic composition of an environment can also be inferred from previous studies, if available.

Once the approximate taxonomic composition of a specific sample is known, a targeted protein reference database can be created that contains only those proteins that can be produced by the organisms of interest. This helps to drastically reduce the search space for a subsequent metaproteomics analysis (typically performed on the same environment) that is performed using the newly constructed targeted protein reference database.

Version 2.0 of the Unipept Desktop application is fully focused on the analysis of these proteogenomics datasets and therefore introduces an exciting new feature that allows users to construct targeted protein reference databases in two different ways.

To construct a targeted protein reference database with Unipept Desktop 2.0, researchers need some way of selectingUniProtKB proteins that can be present in the final result. The first selection method allows to specify which proteins are retained by providing a list of valid NCBI taxon identifiers. Unipept will then only include those proteins in the final database that are associated with a taxon that is present in this list, or that are associated with an (in)direct child of one of the provided taxon identifiers. Note that there's also the option to limit the database construction process to SwissProt (instead of SwissProt + TrEMBL) such that only manually-curated proteins are included. The second selection method requires the user to provide a set of UniProtKB reference proteome identifiers. Only the proteins that are found in these reference proteomes will then be selected for the construction of the reference database.

3.2.1 Construction of targeted protein reference databases

Protein reference databases that are required for the analysis of metaproteomic samples are typically very large. Depending on the number of proteins included, the size of these databases ranges from a few gigabytes to more than a terabyte. It is these huge size requirements that make it technically very hard to build targeted protein reference databases for multiple concurrent users on Unipept's servers.

Moving protein reference databases from Unipept's servers to a researcher's local machine opens up a world of new possibilities.

Firstly, researchers now have complete control over the database to be used for an analysis. They are no longer dependent on the update schedule dictated by Unipept, but can update their local reference database whenever they see fit. Previously, there was no way to roll back the database to the previous iteration of the UniProtKB resource after an update had been performed on Unipept's servers. This is important

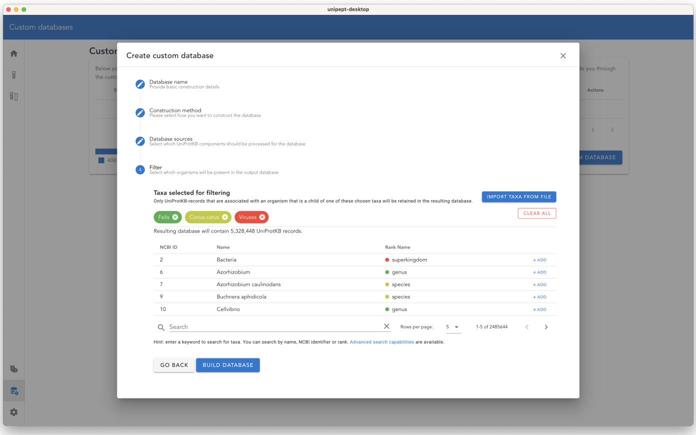


Figure 3.3: Database creation wizard in Unipept Desktop 2.0. This wizard guides the user through the process of building a targeted protein reference database. Researchers can select proteins in two different ways: by providing a set of taxon identifiers, or by providing a list of UniProtKB reference proteome IDs.

because researchers could no longer correctly compare metaproteomics analysis results for samples that were analyzed using a different database version.

Secondly, privacy sensitive data, which may be part of some metaproteomics experiments, is no longer sent over the internet to the Unipept servers for analysis. Some research institutes or applications do not allow sensitive data to be sent to remote services for analysis, but rather require it to be kept in-house to protect patient confidentiality and privacy.

Thirdly, in most cases, the runtime of analyses performed using a local database is improved because the data no longer needs to be transmitted over the Internet and the targeted protein reference databases are typically much smaller in size. The smaller the reference database, the faster the analyses can be performed.

Finally, the amount of false positive peptide matches that can occur is drastically reduced when compared to metaproteomics analyses against the complete UniProt database.

3.2.2 Implementation

3.2.2.1 Dependency management and portability

Unipept uses a custom format for storing protein reference databases, which includes a large amount of pre-computed data to speed up subsequent metaproteomics analyses. This database format is loaded into a relational database management system (RDBMS) such as MySQL. An RDBMS is a very specialized piece of software designed to query huge amounts of structured data as quickly as possible.

The Unipept Desktop application does not query this database directly, but instead relies on the Unipept API to do the hard work. The Unipept API provides a standardized set of HTTP REST endpoints that respond to queries from the Unipept Desktop (or Unipept Web) application with the desired information. This Unipept API in turn is a Ruby-on-Rails project that needs to be executed by a piece of software called a web server.

Both the installation and configuration of an RDBMS and a web server (to run the Unipept API) require a considerable amount of time, effort and technical skills, which is undesirable for users of the Unipept Desktop application. They need an application that is easy to install and that does not require a lot of user intervention to start and maintain.

To address these issues, we decided to encapsulate all required software dependencies in a Docker (Merkel) image. Docker is a free tool that allows to run predefined virtual containers (as defined by an image) on a variety of different operating systems. A Docker image contains a set of instructions to be executed by a virtual computing environment (controlled by Docker) and guarantees that these instructions will work deterministically on any supported system. By relying on Docker for the RDBMS and the web server, we have reduced the number of dependencies that are required by the Unipept Desktop to just one: Docker (which can be downloaded for free from its official website and supports all major operating systems).

Communication between the Unipept Desktop app and the software dependencies that are managed by Docker is completely transparent to the user. We use a NodeJS package called Dockerode, which handles all communication between both parties.

3.2.2.2 Filtering UniProtKB by NCBI taxon identifiers

We described earlier how a targeted protein reference database can be constructed by selecting a list of NCBI taxon identifiers. In this case, Unipept selects only those proteins from the UniProtKB resource that are associated with one of these taxa (or children of these taxa) and includes them in the targeted reference database. In this section, we describe how this is implemented so that efficient filtering can be performed on all 227 million UniProtKB proteins.

The first step in the database construction process is downloading and processing all proteins from the UniProtKB database sources. For each source (SwissProt and TrEMBL), the application constructs a database index structure that can be easily queried and reused in the future. This index structure consists of several "chunks", each containing a set of different proteins. These chunks are compressed line-based `tsv`-files that contain all the necessary information (protein identifier, linked NCBI taxon identifiers, functional annotations, etc.). The protein information is sorted numerically by NCBI taxon identifier, and each chunk contains only those proteins that correspond to a known subset of the NCBI taxon identifier space. Figure 3.4 shows a schematic representation of how this index structure is constructed and the impact it has on the final targeted database construction process.

Downloading and constructing the reusable database index structure is a process that only needs to be performed once for each version of the UniProtKB resource. Subsequent targeted protein reference databases will reuse an existing index or will automatically rebuild it if the UniProtKB resource has been updated since the previous time the index was constructed.

To efficiently query the index structure, we first determine which chunks need to be queried. This is simply a matter of looking up which ranges the various NCBI IDs provided fall into and only processing the chunks that correspond to those taxon ranges. All of these chunks are completely disjoint from each other and can be processed and filtered in parallel, maximizing the use of modern multi-core CPUs.

3.2.3 Case Study

To assess the strength of proteogenomics analyses in Unipept Desktop, we used Unipept Desktop 2.0 to perform the taxonomic annotation of a metaproteomic dataset obtained from 28 human fecal samples collected from celiac disease patients following a gluten-free diet and previously subjected to a 16S rRNA gene sequencing study (Bibbò *et al.*, 2020). Here, we re-analyzed and re-annotated the 16S rRNA gene sequencing data using a robust and up-to-date bioinformatics pipeline based on the amplicon sequence variant (ASV) approach and a newer database (Quast *et al.*, 2013) (as previously described (Palomba *et al.*, 2021)), to obtain accurate information about the set of bacterial taxa present in the environment under study. In parallel, the residues from the 28 fecal samples underwent protein extraction, filter-aided sample preparation and LC-MS/MS analysis, according to established procedures (Tanca *et al.*, 2022). Mass spectra were analyzed by Proteome Discoverer (version 2.4, Thermo Fisher Scientific) (Orsburn, 2021), using a publicly available collection of human gut metagenomes (Li *et al.*, 2014) as the

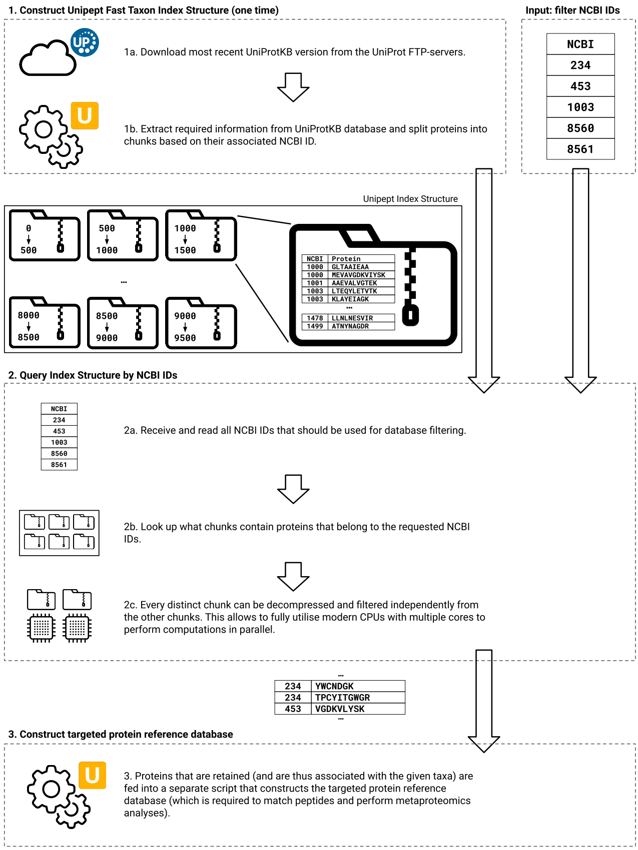


Figure 3.4: To efficiently filter the UniProtKB database, Unipept builds a custom index structure that can be easily filtered by taxon ID. This index structure only needs to be built the first time a targeted protein reference database is constructed. The index will be reused for all subsequent database builds. The NCBI IDs, which can optionally be provided by a user, are used to select which index chunks need to be queried. Each of these chunks is a compressed file containing protein information, and each of these chunks can be processed in parallel by multiple CPU-cores (speeding up the construction process).

sequence database, Sequest-HT as the search engine and Percolator for peptide validation (search parameters are detailed elsewhere). A total of 64 845 microbial peptides (of which 62 363 peptides remained after duplicate filtering) were identified (with 1% as FDR threshold) and used as input for Unipept Desktop annotation. We used the online NCBI Taxonomy Browser (Federhen, 2012) to convert between taxonomy IDs and taxa names where necessary.

The purpose of this case study was to demonstrate that the Unipept Desktop app is capable of analyzing a metaproteomic sample by matching the input peptides to only a specific subset of the proteins in the UniProtKB (The UniProt Consortium, 2021) database. Secondly, we investigated the extent to which the taxonomic profile of a metaproteomic dataset differs when annotated against different protein reference databases.

As the 16S rRNA gene is only present in bacterial species, we restricted our analysis to bacteria. All metaproteomic analyses in this study are performed using Unipept Desktop v2.0.0-alpha.7 with the search settings "filter duplicates" and "equate I/L" enabled (the "advanced missed cleavage handling" setting was disabled).

The experiment started by processing all 227 million proteins in UniProtKB 2022.3 (both SwissProt and TrEMBL) and constructing an initial, unfiltered, protein reference database. During the analysis of our metaproteomic sample using Unipept Desktop and this general-purpose database, we were able to match 56 070 peptides (out of a total of 62 363 peptides, or 89.9%). Of these matches, 30 248 peptides (53.9%) were annotated with a taxon at the rank of "family" (or lower) and 13 659 peptides (24.4%) were annotated with a taxon at the rank of "species" (or lower).

We repeated this analysis using three other targeted protein reference databases, constructed by including only all taxa identified by the 16S rRNA gene sequencing analysis at the family, genus or species ranks, respectively. The higher a rank in the NCBI taxonomy, the less specific the constructed reference databases will be and the more proteins they will still contain. For each NCBI rank of interest, we counted the number of peptides annotated with a taxon of that rank (or lower) and compared these counts across the different protein reference databases. As can be seen in Figure 3.5, the total number of peptides matched decreases as the size of the protein reference databases decreases, but the number of taxon matches at a lower rank of the NCBI taxonomy increases for all three targeted protein reference databases.

Looking at the higher ranks, we see that the larger reference databases have an advantage and provide a valuable annotation for more peptides (53 243 (95.0%), 48 637 (86.7%) and 28 877 (51.5%) taxon matches at the "Superkingdom" rank or lower for the "families", "genera" and "species" based databases, respectively). Note that, at this level, the targeted database constructed from a list of "families" already outperforms the unfiltered database (Figure 3.5).

It is important to note that there is a trade-off to be made between matches on a particular NCBI rank of interest and the input filter used to construct targeted protein reference databases. The more restrictive the input filter (assuming it is a good representation of the taxa in the environment under study), the fewer peptides will generally be matched, but the more likely it is that more detailed taxon matches will be found. This is easily explained. The more proteins there are in a reference database, the greater the chance that a purely random peptide match will occur by chance. These random matches are typically with proteins from organisms unrelated to the environment of interest, so the lowest common ancestor calculation ends up at a very high level in the NCBI taxonomy (Figure 3.1 and Figure 3.2).

Going down a few ranks in the NCBI taxonomy and counting the number of matches at the species level, we find 13 659 taxa (24.4%) when using the unfiltered database and 14 192 (25.3%), 14 593 (26.0%) and 13 960 matches (24.9%) for the "families", "genera" and "species" based databases, respectively. At this level, all three of the targeted reference databases outperform the unfiltered database while containing significantly fewer proteins. These significantly smaller protein reference databases require less storage space and fewer computing resources to work with, allowing the analyses to be performed on a simple laptop (rather than relying on the remote Unipept web servers). In most cases, the analysis is completed faster (especially when the missed cleavage handling option is enabled) and the end-user has full control over exactly which database is being used.

If we compare the targeted analysis with the analysis based on the unfiltered database, we see that 1 408 peptides that were previously annotated are now unmatched. As this is not an insignificant amount, it is important to investigate what the main differences between the two analyses at a taxonomic level are.

If we look at the taxa that were matched using the full database, and not when using the targeted database, we see that most of them belong to the *Firmicutes phylum* (813 matches). Looking a little further, we see that the species *Evtepia gabavorous* (80 matches), *Odoribacter splanchnicus* (44 matches), and *Turicibacter sanguinis* (34 matches) are the most represented within the *Firmicutes*.

According to the NCBI taxonomy, *Evtepia gabavorous* belongs to the *Eubacteriales incertae sedis* “family”, which is actually an unclassified taxon and is therefore not included in the 16S SILVA database. Secondly, *Odoribacter splanchnicus* belongs to the *Marinifilaceae* family according to SILVA, whereas it is assigned to the *Odoribacteraceae* family by the NCBI taxonomy (which is why this family is not present in the taxon list that was used to construct a protein reference database and is therefore not matched during the Unipept analysis). Finally, a similar explanation applies to *Turicibacter*, which belongs to *Erysipelotrichaceae* according to SILVA and to the *Turicibacteraceae* family according to NCBI. Therefore, this family is also missing from the filtered taxon list and its proteins are not included in the database construction.

Most of the other remaining *Firmicutes* assignments are mainly attributed to *Clostridiales bacterium* and *Firmicutes bacterium* (413 matches in total), two taxa for which the rank is unspecified in NCBI and which are therefore not taken into account when constructing the targeted database. A further 122 peptide matches using the full database were taxonomically annotated at the root level, which does not provide any valuable information other than the fact that the peptide was indeed matched.

After this detailed analysis, we can conclude that most of the mismatches when using a targeted reference database are due to taxonomic inconsistencies between the SILVA database (used for the 16S rRNA gene taxonomic annotation) and the NCBI taxonomy that is being used by Unipept, or even to incomplete or provisional annotations within the NCBI taxonomy itself. This problem could be overcome by also using NCBI for the taxonomic classification of 16S data (instead of SILVA), but this choice is entirely up to the end-user, and is outside the scope of the analysis performed by Unipept Desktop.

All of these experiments have been conducted on a normal modern computer with a 6-core CPU (AMD Ryzen 3600X),16GiB of RAM, a SATA-6 SSD and a 100Mbps internet connection. On this machine, it took approximately 21 hours and 25 minutes to construct a custom database containing 46 million proteins or approximately 20 minutes for a targeted database containing 1 million proteins. All proteins from the UniProtKB resource are preprocessed the first time a targeted database is constructed and took approximately 5 hours and 30 minutes on this machine (this preprocessing step will automatically be performed when the UniProtKB resource updates). The final size of the database cache is 52GiB, and the 46 million and 1 million protein databases are respectively 255GiB and 5.76GiB in size.

3.2.4 Concluding Remarks

The newest iteration of the Unipept Desktop app builds upon the strength of the existing Unipept infrastructure to enable support for the analysis of proteogenomics samples. Leveraging taxonomic information of the environment under study (e.g. generated from a metagenomics experiment), it is possible to construct targeted protein reference databases that include only a (relevant) subset of proteins

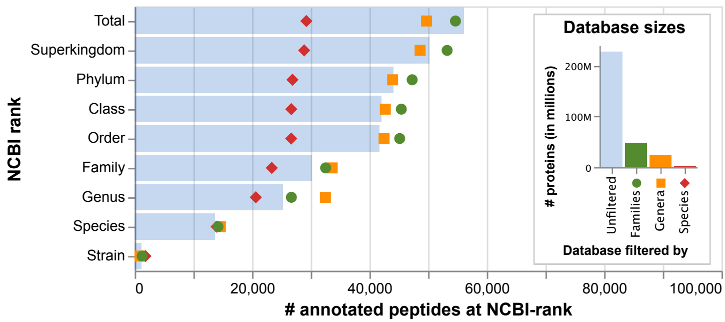


Figure 3.5: Four analyses have been performed on the same metaproteomic sample, but with a different underlying protein reference database. In the main visualization (left), we compare the number of peptides with a taxon match at a specific rank (or lower) in the NCBI taxonomy. In most cases the analyses performed by using a reference database constructed from families or genera performs as well as or better than the full UniProtKB database. However, these databases contain only a fraction of the proteins (right) and therefore require much less computational resources. At the species level, all protein reference databases behave almost identically. At higher ranks, we can see that the reference database constructed from a set of species is probably too restricted.

from the UniProtKB resource. These significantly smaller reference databases drastically improves the time and computational resources required to subsequently analyze metaproteomics samples, which ultimately makes it possible to perform these analyses on a local machine.

Since Unipept Desktop 2.0 makes it possible to perform metaproteomics analysis on a local machine, a range of new possibilities opens up. Privacy-sensitive data no longer needs to be transmitted over the internet and users now control which reference database is used. We have shown that using targeted protein reference databases can even lead to a metaproteomics analysis with a higher taxonomic resolution (assuming that the selected taxa suits the environment under study). No scientific software package is ever completed and we can still think about future improvements that could be beneficial for the Unipept Desktop application. First of all, at this point it is not yet possible to construct targeted protein reference databases from UniProtKB versions other than the current one. This is a consequence of the fact that previous versions of the UniProtKB database are provided in a different file format for which a new parser needs to be implemented.

Secondly, since all targeted protein reference databases are always constructed by filtering the UniProtKB resource, only proteins that are included in UniProtKB can be matched using Unipept. This can be problematic for some research disciplines (such as protein research of ocean water) that are investigating proteins of organisms that are not well represented in UniProtKB. This problem could be overcome by allowing Unipept to construct protein reference databases from external sources (e.g., represented by FASTA or PEFF files). These additions are considered for future versions of the Unipept Desktop app.

3.2.5 Acknowledgements

This work has benefited from collaborations facilitated by the Metaproteomics Initiative (https://metaproteomics.org/) whose goals are to promote, improve, and standardize metaproteomics (Van Den Bossche, Kunath, *et al.*, 2021). This work has furthermore been supported by the Research Foundation - Flanders (FWO) [1164420N to P.V., 12I5220N to B.M.] and by the University of Sassari [Fondo di Ateneo per la Ricerca 2020 to A.T.]. We thank the Flemish Supercomputer Center (VSC) funded by the Research Foundation - Flanders (FWO) and the Flemish Government for providing the infrastructure to build the Unipept database and to run the experiments from this manuscript. Part of this work was also supported by the Research Foundation - Flanders (FWO) for ELIXIR Belgium [I002819N].

Part II

The Unipept ecosystem

The introductory text for this chapter comes here…

Chapter 4

Unipept CLI 2.0: adding support for visualisations and functional annotations

This chapter contains a verbatim copy of the technical note by (Verschaffelt et al., *2020) as submitted to Bioinformatics.*

Abstract — Unipept (Mesuere *et al.*, 2012) is a collection of tools developed for fast metaproteomics data analysis. The Unipept ecosystem consists of a web application, an application programming interface (API) as a web service (Mesuere *et al.*, 2016) and a command-line interface (CLI) (Mesuere *et al.*, 2018). The key strengths of Unipept are its speed, its ease-of-use and the extensive use of interactive data visualization in the analysis results. The Unipept database is derived from the UniProt (UniProt, 2019) KB and consists of tryptic peptides linked with taxonomic and functional annotations. Unipept initially launched with support for taxonomic analysis of metaproteomics data in 2012. Version 4.0 (Gurdeep Singh *et al.*, 2019) of the Unipept web application was launched in November 2018 and extended the web interface with support for functional annotations such as Gene Ontology (GO) terms (Ashburner *et al.*, 2000), Enzyme Commission (EC) numbers (Webb, 1992) and InterPro entries (Hunter *et al.*, 2009).

4.1 Introduction

Unipept (Mesuere *et al.*, 2012) is a collection of tools developed for fast metaproteomics data analysis. The Unipept ecosystem consists of a web application, an application programming interface (API) as a web service (Mesuere *et al.*, 2016) and a command-line interface (CLI) (Mesuere *et al.*, 2018). The key strengths of Unipept are its speed, its ease-of-use and the extensive use of interactive data visualization in the analysis results. The Unipept database is derived from the UniProt (The UniProt Consortium, 2019) KB and consists of tryptic peptides linked with taxonomic and functional annotations. Unipept initially launched with support for taxonomic analysis of metaproteomics data in 2012. Version 4.0 (Gurdeep Singh *et al.*, 2019) of the Unipept web application was launched in November 2018 and extended the web interface with support for functional annotations such as Gene Ontology (GO) terms (Ashburner *et al.*, 2000), Enzyme Commission (EC) numbers (Webb, 1992) and InterPro entries (Hunter *et al.*, 2009).

The GO terms are organized into three different domains: 'cellular components', 'molecular functions' and 'biological processes'. Every GO-term is associated with exactly one domain and consists of a name, an identifier and an exact definition. The EC numbers can be used to classify enzymes, based on the chemical reactions that they catalyze. Every EC number consists of four numbers, separated by a dot, yielding a hierarchical classification system with progressively finer enzyme classifications. InterPro is a database that consists of predictive models collected from external databases that can be classified into five different categories. More information about functional annotation in metaproteomics can be found in the study by (Schiebenhoefer *et al.*, 2019).

For each input peptide, Unipept finds all functional annotations associated with all of the UniProt entries in which the peptide occurs. All found functions are listed in order of decreasing number of peptides associated with this function.

In this article, we present several new additions to the Unipept API and CLI which allow third-party applications [such as Galaxy-P (Jagtap *et al.*, 2015)] to integrate the new functional analysis capabilities provided by Unipept.

4.2 Materials and methods

The Unipept API is a high-performance web service that responds in a textual format (JSON) to HTTP-requests from other applications or tools and allows to integrate the services provided by Unipept into other workflows. Unipept's CLI is a Ruby-based application and high-level entry point which allows users to actively query Unipept's database. Compared to the API, users do not need to compile API-requests manually but can rely on the CLI to automatically do so in a parallelized way. In addition, it supports multiple input and output formats such as FASTA and CSV.

The Unipept database and web application were recently expanded to include GO terms, EC numbers and InterPro entries. These new annotations are now also available from newly developed API-endpoints and CLI-functions, providing structured access to this functional information.

Most of the newly developed endpoints support batch retrieval of information for a list of peptides. In this case, the API returns a list of objects where each object in the response corresponds with information associated with one of the input peptides. Every API-endpoint is accompanied by an identically named CLI-function, which provides the user with the ability to import data from or export data to various specifically formatted files. In addition, version 2 of the Unipept CLI introduces the ability to produce interactive visualizations directly from the command line.

Among other information, the Unipept tryptic peptide analysis lists functional annotations associated with a given tryptic peptide. These data are aggregated because a peptide can occur in multiple proteins that each can have multiple functional annotations. For each annotation, we also return the amount of underlying proteins that match with this specific annotation.

Some applications require all known information for a list of tryptic peptides. The 'pept2funct' function is a combination of the preceding three endpoints and returns all functional annotations associated with the given tryptic peptide. 'peptinfo' on the other hand, returns all the available information for one or more tryptic peptides. All functional annotations for this peptide are part of the response, as well as the lowest common ancestor for this peptide. Both functions also support splitting the GO terms and InterPro entries over the respective domains, and naming information can optionally be retrieved.

The 'taxa2tree' function constructs a tree from a list of NCBI taxon ids. This tree is an aggregation of the lineages that correspond with the given taxa and can be exported as three distinct output formats: JSON, HTML and as a URL. The HTML and URL representation of a taxonomic tree both provide three interactive data visualizations, albeit with different possibilities. A generated HTML-string first needs to be stored in a file before it can be rendered by a browser and cannot be easily shared with other people but is easily editable. A URL on the other hand is simply a shareable link to an online service that hosts all interactive visualizations.

4.3 Conclusion

Version 2.0 of the Unipept API and CLI is a significant update that provides fast and easy access to the powerful analysis pipeline of Unipept.

In addition to the existing taxonomic analysis, it now features multiple functional annotations which will enable users to gain new insights into complex ecosystems. These new features can easily be integrated into third-party tools such as the MetaProteome Analyzer (Muth *et al.*, 2018). Galaxy-P, a highly used workflow integration system, is already successfully making use of the novel analysis functions that are introduced with this new release.

4.4 Funding

This work was supported by the Research Foundation—Flanders (FWO) [1164420N to P.V.; 12I5220N to B.M.; 1S90918N to T.V.D.B.; G042518N to L.M.; 12S9418N to C.D.T.].

Chapter 5

Unipept Visualizations: an interactive visualization library for biological data

This chapter contains a verbatim copy of the application note by (Verschaffelt et al., *2022) as submitted to Bioinformatics.*

Abstract — The Unipept Visualizations library is a JavaScript package to generate interactive visualizations of both hierarchical and non-hierarchical quantitative data. It provides four different visualizations: a sunburst, a treemap, a treeview and a heatmap. Every visualization is fully configurable, supports TypeScript and uses the excellent D3.js library. The Unipept Visualizations library is available for download on NPM: https://npmjs.com/unipept-visualizations. All source code is freely available from GitHub under the MIT license: https://github.com/unipept/unipept-visualizations.

5.1 Introduction

Unipept is an ecosystem of software tools for the analysis of metaproteomics datasets that consists of a web application (Gurdeep Singh *et al.*, 2019), a desktop application (Verschaffelt, Van Den Bossche, Martens, *et al.*, 2021), a command line interface (Verschaffelt *et al.*, 2020) and an application programming interface. It provides taxonomic and functional analysis pipelines for metaproteomics data and highly interactive data visualizations that help interpret the outcome of these analyses.

We developed custom visualizations used for Unipept from scratch because existing libraries, such as Krona (Ondov *et al.*, 2011), were lacking essential features or were hard to integrate. They were designed as generic tools to visualize hierarchical quantitative data and can therefore also be used to visualize data from nonproteomics origins. To facilitate reuse of these broadly usable components, we have isolated the visualizations from the main Unipept project and made them available as a standalone package that can easily be reused by other software tools. We released this package under the permissive MIT open-source license, so researchers from other disciplines are free to reuse these visualizations and connect them to their own data sources. Currently, our visualizations are already incorporated in TRAPID 2.0, a web application for the analysis of transcriptomes ((Bucchini *et al.*, 2021) and UMGAP, the Unipept MetaGenomics Pipeline (Van der Jeugt *et al.*, 2022).

5.2 Visualizations

We currently provide four highly interactive data visualizations that are all designed for a specific purpose: a sunburst, a treeview, a treemap and a heatmap. The sunburst (Figure 5.1a), treeview (Figure 5.1d) and treemap (Figure 5.1b) can be used to visualize quantitative hierarchical data and are designed to depict the parent–child relationship of a hierarchy of nodes as clearly as possible, while still incorporating the strength of the relationship between, or the counts associated with, connected nodes. The heatmap (Figure 5.1c), conversely, is not suitable to visualize hierarchical information but displays a magnitude in two dimensions, including optional clustering and dendrogram rendering.

5.2.1 Quantitative hierarchical data visualizations

Hierarchical data occurs throughout a variety of bioinformatics disciplines. In the metaproteomics research area alone, many examples of hierarchical data exist, such as the hierarchical structure of the NCBI taxonomy (Schoch *et al.*, 2020), the hierarchy imposed by the enzyme commission numbers and the gene ontology terms (The Gene Ontology Consortium, 2019). In most cases, quantitative data are available for multiple nodes at many levels in the hierarchy. For example, Unipept assigns peptide counts to taxa that are scattered around the NCBI taxonomy, including identifications that are highly specific (near leaves of the tree) or lack deep taxonomic resolution (near the root of the tree). Being able to interactively zoom in and out on the hierarchical data enables exploratory analysis.

The three visualizations for hierarchical data provided by our package take input data in the same hierarchical format, making it trivial to switch between the different types of visualization once the input data

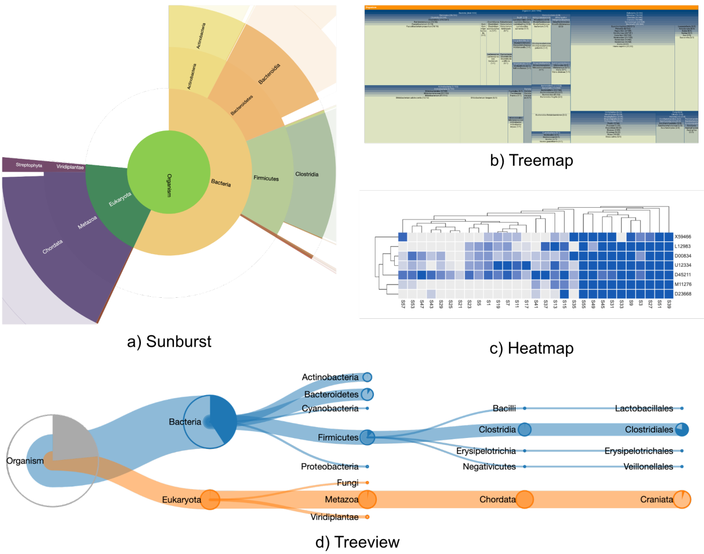


Figure 5.1: Overview of the visualizations currently provided by the Unipept Visualizations library. All examples were generated with default configuration settings, except for the heatmap for which the setting 'dendrogramEnabled' was set to 'true'.

are formatted correctly.

5.2.2 Quantitative non-hierarchical data visualizations

A heatmap (Figure 5.1c) is a well-known visualization that consists of a two-dimensional grid of cells in which each cell is assigned a specific color from a scale corresponding to its magnitude. The heatmap implementation in our package provides this functionality in an extensively customizable form. Users can reorganize elements, change the color scheme and update label information, among other operations. All values are also automatically normalized to a $[0, 1]$-interval.

As neighboring rows and columns in the input data can have very distinct values, and as this can interfere with reasoning about the heatmap, it is important to group similar values. Our implementation achieves this through hierarchical clustering based on the UPGMA algorithm (Sokal and Michener, 1958). The produced grouping of rows and columns is further clarified by an optional dendrogram that can be plotted alongside each axis of the heatmap.

However, after clustering, it can still occur that two consecutive leaves in a dendrogram are quite dissimilar due to the $2^n - 1$ possible linear orderings that can be derived from a dendrogram (a dendrogram contains $n - 1$ flipping points for which both children can be switched). This can be addressed by reordering the leaves of the tree, as the orientation of the children of all n nodes in a dendrogram can be flipped without affecting the integrity of the dendrogram itself. Our heatmap implementation uses the Modular Leaf Ordering technique (Sakai *et al.*, 2014) to reorder all leaves of the dendrogram such that the distance between consecutive leaves is minimized. This technique is a heuristic that performs very well in comparison to the more resource-intensive Optimal Leaf Ordering (Bar-Joseph *et al.*, 2001) or Gruvaeus–Wainer algorithms (Gruvaeus and Wainer, 1972).

5.3 Implementation

The visualization package has been developed with D3 (Bostock *et al.*, 2011) and TypeScript (Bierman *et al.*, 2014) and every visualization is displayed in the web browser with one of two technologies: SVG or HTML5 canvas. SVG's are easy-to-use and are scalable by nature but often lack necessary performance for complex interactive visualizations. HTML5 canvas, in contrast, provides much better performance using a rasterized image.

Every visualization is presented as a single JavaScript class and provides a full set of configuration options to extend and configure the visualization. New versions of the package will automatically be published on NPM (https://npmjs.org) and GitHub (https://github.com/unipe pt/unipept-visualizations), so that any project depending on it package can always use the latest version.

We also provide an extensive set of documentation resources that ease the adoption process of our package, as well as a collection of live notebooks (see https://observablehq.com/collection/@unipept/unipept-visualizations). These notebooks provide interactive and editable examples that demonstrate the full potential and guide users through the different configuration options. The code and resources that make up the live notebooks can be modified online and provide a very convenient way to try out the package.

5.4 Funding

This work has been supported by the Research Foundation—Flanders (FWO) [1164420N to P.V.; 12I5220N to B.M.; G042518N to L.M.].

Chapter 6

MegaGO: a fast yet powerful approach to assess functional Gene Ontology similarity across meta-omics data sets

This chapter contains a verbatim copy of the manuscript by (Verschaffelt, Van Den Bossche, Gabriel, et al.*, 2021) as submitted to Journal of Proteome Research.*

Abstract — The study of microbiomes has gained in importance over the past few years and has led to the emergence of the fields of metagenomics, metatranscriptomics, and metaproteomics. While initially focused on the study of biodiversity within these communities, the emphasis has increasingly shifted to the study of (changes in) the complete set of functions available in these communities. A key tool to study this functional complement of a microbiome is Gene Ontology (GO) term analysis. However, comparing large sets of GO terms is not an easy task due to the deeply branched nature of GO, which limits the utility of exact term matching. To solve this problem, we here present MegaGO, a user-friendly tool that relies on semantic similarity between GO terms to compute the functional similarity between multiple data sets. MegaGO is high performing: Each set can contain thousands of GO terms, and results are calculated in a matter of seconds. MegaGO is available as a web application at https://megago.ugent.be and is installable via pip as a standalone command line tool and reusable software library. All code is open source under the MIT license and is available at https://github.com/MEGA-GO/.

6.1 Introduction

Microorganisms often live together in a microbial community or microbiome where they create complex functional networks. These microbiomes are therefore commonly studied to reveal both their taxonomic composition as well as their functional repertoire. This is typically achieved by analyzing their gene content using shotgun metagenomics. Whereas this approach allows a detailed investigation of the genomes that are present in such multiorganism samples, it reveals only their functional potential rather than their currently active functions (Jansson and Baker, 2016). To uncover these active functions within a given sample, the characterization of the protein content is often essential (Lohmann *et al.*, 2020).

The growing focus on functional information as a complement to taxonomic information (Louca *et al.*, 2016) is derived from the observation that two taxonomically similar microbial communities could have vastly different functional capacities, whereas taxonomically quite distinct communities could have remarkably similar functions. Whereas the investigation of the active functions is thus increasingly seen as vital to a complete understanding of a microbiome, the identification and comparison of these detected functions remains one of the biggest challenges in the field (Schiebenhoefer *et al.*, 2019).

Several omics tools exist to describe functions in microbial samples, although these tools link functionality to different biological entities such as genes, transcripts, proteins, and peptides (Muth *et al.*, 2015, 2018; Van Den Bossche *et al.*, 2020; Verschaffelt *et al.*, 2020; Gurdeep Singh *et al.*, 2019; Riffle *et al.*, 2018; Schneider *et al.*, 2011; Schiebenhoefer *et al.*, 2020; Huerta-Cepas *et al.*, 2019; Huson *et al.*, 2007). However, most tools are capable of directly or indirectly reporting functional annotations as a set of Gene Ontology (The Gene Ontology Consortium, 2019) (GO) terms, regardless of the biological entities they are assigned to. In October 2020, there were 44264 of these terms in the complete GO tree. GO terms are organized into three independent domains: molecular function, biological process, and cellular component (Ashburner *et al.*, 2000). In each domain, terms are linked into a directed acyclic graph, an excerpt of which is shown in Figure 6.1. In the GO graph, a parent term can have one or more child (e.g., the root node "biological process" is the parent of the children GO:0009987 and GO:0008152), and children can have multiple parents (e.g., the most specific term "translation" has GO:0043043, GO:0034645, and GO:0044267 as parents).

Whereas this highly branched graph structure of GO allows flexible annotation at various levels of detail, it also creates problems when the results from one data set are compared to those of another data set. Indeed, even though two terms may be closely linked in the GO

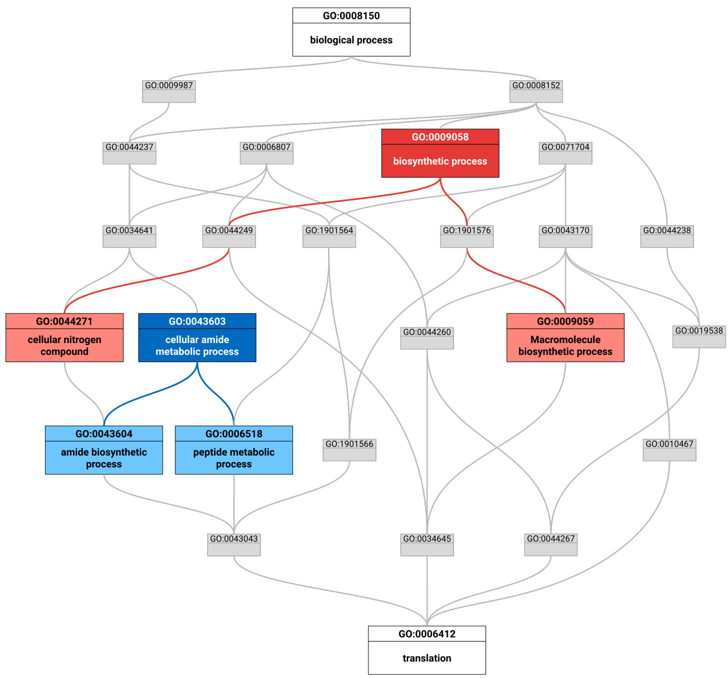


Figure 6.1: Excerpt of the biological process domain of the Gene Ontology showing all parent terms up to the root for "translation" (GO:0006412). The root GO term "biological process" (GO:0008150) has multiple children. The most specific term "translation", in contrast, has multiple parents. When comparing the two terms GO:0044267 and GO:0034645 (portrayed in light red), we find two different lowest common ancestors: GO:0044249 and GO:1901576 (dark red). Only one of these, however, can be the most informative common ancestor (MICA), that is, the common ancestor with the highest information content for the terms in light red. Because an IC of 1.52 is larger than 1.48, the GO:0044249 is the MICA. The terms GO:0043604 and GO:0006518 (in light blue) are more similar than the two terms we previously described and have only one lowest common ancestor, which is also automatically the MICA for these terms: GO:0043603 (in dark blue). IC, information content; *, most informative common ancestor.

tree and are therefore highly similar (e.g., as parent and child terms or as sibling terms), the typically employed exact term matching will treat these terms as wholly unrelated, as the actual GO terms (and their accession numbers) are not identical. This problem is illustrated in a study by (Sajulga *et al.*, 2020), where a multisample data set was analyzed using several metaproteomics tools. The resulting GO terms were then compared using exact matching. The overlap between the result sets was quantified using the Jaccard index and was found to be low. As previously explained, this low similarity is likely the result of the limitations of the exact term matching approach.

There is thus a clear need for a more sophisticated GO term comparison that takes into account the existing relationships in the full GO tree. However, most existing tools that provide such comparison are based on enrichment analyses (Huang *et al.*, 2009; Waardenberg *et al.*, 2015; Fruzangohar *et al.*, 2013). In such analyses, a list of genes is mapped to GO terms, which are then analyzed for enriched biological phenomena. As a result, to the best of our knowledge, no tools allow the direct comparison of large functional data sets against each other, nor are these able to provide metrics to determine how functionally similar data sets are.

We therefore present MegaGO, a tool for comparing the functional similarity between large sets of GO terms. MegaGO calculates a pairwise similarity score between multiple sets of GO terms for each of the three GO domains and can do so in seconds, even on platforms with limited computational capabilities.

6.2 Implementation

To measure the similarity between sets of GO terms, we first need to measure the similarity of two individual terms. We compare two terms using the Lin semantic similarity metric, which can take on a value between 0 and 1 (Supplementary Formula 1a). The Lin semantic similarity is based on the ratio of the information content of the most informative common ancestor (MICA) to the average of the terms' individual information content.

The information content (Supplementary Formula 1b) is computed by estimating the terms' probability of occurrence (Supplementary Formula 1c), including that of all of their children. Term frequencies are estimated based on the manually curated SwissProt database (The UniProt Consortium, 2019). As a result, a high-level GO term such as "biological process" (through its many direct or indirect child terms) will be present in all data sets and thus carries little information. A more specific term such as "translation" (or any of its potential child terms) will occur less frequently and thus will be more informative

(Figure 6.1). To finally calculate the similarity of two terms, we compare their information content with that of their shared ancestor that has the highest information content, the MICA. If the information content of the MICA is similar to the terms' individual information content, then the terms are deemed to be similar. The dissimilar terms "peptide biosynthetic process" and "cellular macromolecule biosynthetic" are situated further from their MICA "cellular biosynthetic process" than the similar terms "amide biosynthetic process" and "peptide metabolic process" with their respective MICA "cellular amide metabolic process" (Figure 6.1).

MegaGO, however, can compare not only two terms but also sets of GO terms. More specifically, two sets of GO terms can be compared via the web application, but an unlimited number of sets can be compared via the command line tool. Note that in these sets, duplicate GO terms will be removed so that each GO term will be equally important, regardless of how often it is provided by the user. To compare the sets of GO terms, pairwise term similarities are aggregated using the Best Matching Average (BMA, Supplementary Formula 2) (Schlicker *et al.*, 2006). For each GO term in the first input data set, the BMA finds the GO term with the highest Lin semantic similarity in the second data set and averages the values of these best matches. Moreover, MegaGO calculates the similarity for each of the three domains of the gene ontology (molecular function, biological process, and cellular component), as GO terms from distinct domains do not share parent terms. The general overview of MegaGO is shown in Figure 6.2.

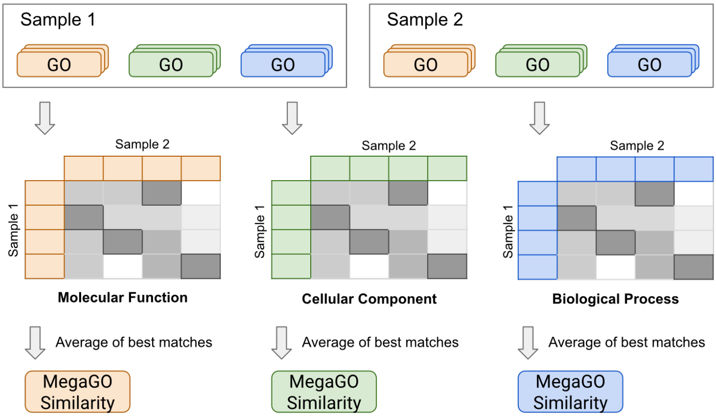


Figure 6.2: Overview of MegaGO workflow. The Gene Ontology (GO) terms of each sample set are separated into three GO domains: molecular function, cellular component, and biological process. Each term of each sample set is compared to every term in the other set that is from the same domain. The match with highest similarity for each term is then selected, and the average across all of these best matches is calculated.

MegaGO is implemented in Python, is installable as a Python package from PyPi, and can easily be invoked from the command line. The

GOATOOLS (Klopfenstein *et al.*, 2018) library is used to read and process the Gene Ontology and to compute the most informative common ancestor of two GO terms, which are both required to compute the information content value (Supplementary Formula 1, p(go)). GO term counts are recomputed with every update of SwissProt, and a new release is automatically published bimonthly to PyPi, which includes the new data set. Automated testing via GitHub Actions is in place to ensure correctness and reproducibility of the code. In addition, we also developed a user-friendly and easily accessible web application that is available on https://megago.ugent.be. The backend of the web application is developed with the Flask web framework for Python, and the frontend uses Vue. Our web application has been tested on Chromium-based browsers (Chrome, Edge, and Opera) as well as Mozilla Firefox and Safari. The MegaGO application is also available as a Docker-container on Docker Hub (https://hub.docker.com/repository/docker/pverscha/mega-go) and can be started with a single click and without additional configuration requirements. Our Docker container is automatically updated at every change to the underlying MegaGO code. All code is freely available under the permissive open source MIT license on https://github.com/MEGA-GO/. Documentation for our Python script can be found on our Web site: https://megago.ugent.be/help/cli. A guide on how to use the web application is also available: https://megago.ugent.be/help/web.

MegaGO is cross-platform and runs on Windows, macOS, and Linux systems. The system requirements are at least 4 GiB of memory and support for either Python 3.6 (or above) or the Docker runtime.

6.3 Validation

To validate MegaGO, we reprocessed the functional data from (Easterly *et al.*, 2019). This data set consists of 12 paired oral microbiome samples that were cultivated in bioreactors. Each sample was treated with and without sucrose pulsing, hereafter named ws and ns samples, respectively. Each sample contains mass-spectrometry-based proteomics measurements, and all samples were annotated with 1718 GO terms on average. We calculated the pairwise similarity for each of the 300 sample combinations, which took less than 1 min for a single sample pair on the web version of MegaGO. This resulted in a MegaGO similarity score for each of the three GO domains for each sample combination. These similarities were then hierarchically clustered and visualized in a heatmap. All data and intermediate steps of our data analysis are available at https://github.com/MEGA-GO/manuscript-data-analysis/ and can be reproduced with the command line tool using the –heatmap option.

In the heatmap (Figure 6.3), we can observe that the two sample groups cluster together, except for 730ns and 733ns that are clustered in the ws sample group. These two samples were also identified as outliers in (Easterly *et al.*, 2019) and 733ns was originally also identified as both a taxonomic and functional outlier in (Rudney *et al.*, 2015). Similar results can be observed for the GO domain "molecular function" (Supplementary Figure 1). The MegaGO similarity-based clustering of "cellular component" GO terms (Supplementary Figure 2) has two additional samples clustered outside of their treatment group: 852ws in the ns cluster and 861ns in the ws group. Again, these patterns can also be found in previous analyses: 852ws is placed in direct proximity of the ns samples in the principal component analysis (PCA) of the HOMINGS analysis by Rudney et al., and 861ns is closest to 730ns and 733ns in PCA of Rudney et al.'s taxonomic analysis. Interestingly, subjects 730 and 852 were the only ones without active carious lesions, which could cause their divergence in the similarity analyses.

Results produced by MegaGO are thus in close agreement with prior analyses of the same data, showing that MegaGO offers a valid and very fast approach for comparing the functional composition of samples.

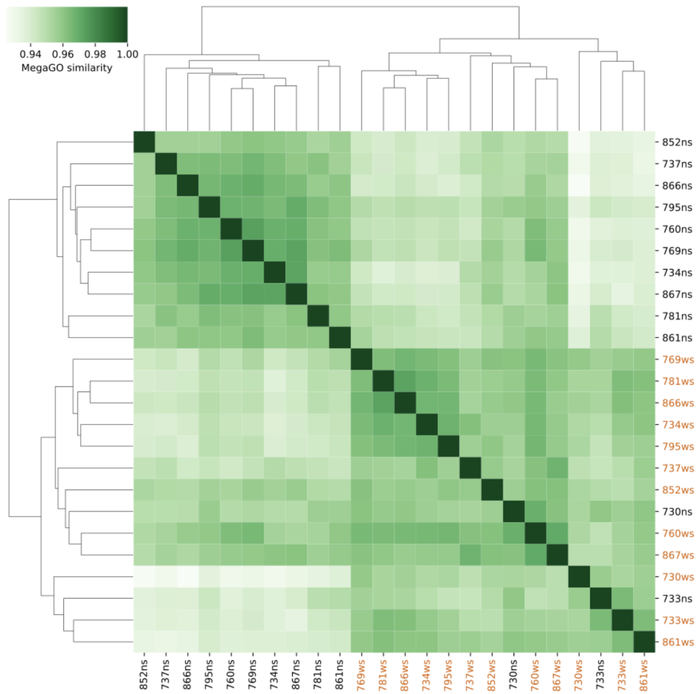


Figure 6.3: Hierarchically clustered heatmap comparing MegaGO similarities for the GO domain "biological process" for each of the samples from (Easterly *et al.*, 2019) Samples that are treated with sucrose pulsing are labeled as "ws" and are displayed in orange.

6.4 Conclusions

MegaGO enables the comparison of large sets of GO terms, allowing users to efficiently evaluate multiomics data sets containing thousands of terms. MegaGO separately calculates a similarity for each of the three GO domains (biological process, molecular function, and cellular component). In the current version of MegaGO, quantitative data are not taken into account, thus giving each GO term identical importance in the data set.

MegaGO is compatible with any upstream tool that can provide GO term lists for a data set. Moreover, MegaGO allows the comparison of functional annotations derived from DNA-, RNA-, and protein-based methods as well as combinations thereof.

6.5 Acknowledgements

We acknowledge the European Bioinformatics Community for Mass Spectrometry (EuBIC-MS). This project found its origin at the EuBIC Developers' 2020 meeting (Ashwood *et al.*, 2020) in Nyborg, Denmark. We thank Thilo Muth (Bundesanstalt für Materialforschung und -prüfung, Berlin, Germany) and Stephan Fuchs (Robert Koch Institute, Berlin, Germany) for their support. P.V., T.V.D.B., L.M., and B.M. acknowledge the Research Foundation - Flanders (FWO) (grants 1164420N, 1S90918N, G042518N, and 12I5220N). L.M. also acknowledges support from the European Union's Horizon 2020 Programme under grant agreement 823839 (H2020-INFRAIA-2018-1). H.S. and B.Y.R. acknowledge support by Deutsche Forschungsgemeinschaft (DFG; grant numbers RE3474/5-1 and RE3474/2-2) and the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI; 031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, and 031A532B).

Chapter 7

Other projects

During the course of my career as a PhD student, I have also been working on a lot of different research projects for which I was not the main contributor, but for which I, nonetheless, provided a significant addition. I have selected two of these projects and included them as sections in this chapter.

7.1 Pout2Prot: An efficient tool to create protein (sub)groups from Percolator output files

This section contains a verbatim copy of the manuscript by (Schallert et al.*, 2022) as submitted to Journal of Proteome Research.*

Abstract — In metaproteomics, the study of the collective proteome of microbial communities, the protein inference problem is more challenging than in single-species proteomics. Indeed, a peptide sequence can be present not only in multiple proteins or protein isoforms of the same species, but also in homologous proteins from closely related species. To assign the taxonomy and functions of the microbial species, specialized tools have been developed, such as Prophane. This tool, however, is not directly compatible with post-processing tools such as Percolator. In this manuscript we therefore present Pout2Prot, which takes Percolator Output (.pout) files from multiple experiments and creates protein group and protein subgroup output files (.tsv) that can be used directly with Prophane. We investigated different grouping strategies and compared existing protein grouping tools to develop an advanced protein grouping algorithm that offers a variety of different approaches, allows grouping for multiple files, and uses a weighted spectral count for protein (sub)groups to reflect abundance. Pout2Prot is available as a web application at https://pout2prot.ugent.be and

is installable via pip as a standalone command line tool and reusable software library. All code is open source under the Apache License 2.0 and is available at https://github.com/compomics/pout2prot.

7.1.1 Introduction

In metaproteomics, the study of the collective proteome of whole (microbial) ecosystems, it is important to learn about the taxonomy and functions represented in the community. For this purpose, tools such as Unipept (Verschaffelt, Van Den Bossche, Martens, *et al.*, 2021) and Prophane (Schiebenhoefer *et al.*, 2020) have been made available to specifically perform downstream annotation of metaproteomic data, while other, more generic tools also provide connections to downstream annotation tools (Schiebenhoefer *et al.*, 2020; Van Den Bossche *et al.*, 2020; Muth *et al.*, 2015). These tools, however, work very differently: while Unipept relies on identified peptides without inferring the corresponding proteins (a peptide-centric approach), Prophane uses protein groups as input (a protein-centric approach). Recently, these two tools were compared in the first multilab comparison study in metaproteomics (CAMPI), (Van Den Bossche, Kunath, *et al.*, 2021) which indicated that the choice between these approaches is a matter of user preference.

The process of grouping proteins is unfortunately not as straightforward as it might first appear (Martens and Hermjakob, 2007; Uszkoreit *et al.*, 2015; Audain *et al.*, 2017; Nesvizhskii and Aebersold, 2005). Identified peptide sequences have to be assembled into a list of identified proteins, but when a peptide can be mapped to multiple proteins, this leads to the protein inference problem (Nesvizhskii and Aebersold, 2005). In metaproteomics, this problem is exacerbated due to the presence of homologous proteins from multiple species in its necessarily large protein databases (Schiebenhoefer *et al.*, 2019). Protein grouping is therefore commonly used to generate a more manageable list of identified protein groups that can be used for further downstream analysis. However, different protein grouping algorithms can be chosen, leading to different lists of protein groups from a single set of identified peptides (Martens and Hermjakob, 2007). In the past, many protein grouping methods have been developed, as reviewed in Audain et al., (Audain *et al.*, 2017) but these typically do not interface well with post-processing tools like Percolator, (Käll *et al.*, 2007) which are able to increase the number of peptide-to-spectrum matches (PSMs) due to a better separation of true and false matches (Bouwmeester *et al.*, 2020). Moreover, the common strategy used by these tools is the Occam's razor strategy, which is not always ideal (Van Den Bossche, Kunath, *et al.*, 2021). We here therefore present a new tool, Pout2Prot, which provides users with two relevant protein inference options that are tailored toward metaproteomics use cases: Occam's razor and anti-

Occam's razor. Occam's razor is based on the principle of maximum parsimony and provides the smallest set of proteins that explains all observed peptides. Here, however, proteins that are not matched by a unique peptide are discarded, and their associated taxonomy and functions, which might actually be present in the sample, are lost. This algorithm is for example used in the X!TandemPipeline (Langella *et al.*, 2017). On the other hand, anti-Occam's razor is based on the maximal explanatory set of proteins, where any protein that is matched by at least one identified peptide will be included in the reported protein list. This algorithm is used in, for example, MetaProteomeAnalyzer (MPA) (Muth *et al.*, 2015). Unfortunately, there is no simple way to determine a priori which algorithm will be optimal, as this can differ from sample to sample (Muth *et al.*, 2015). These strategies are visually represented in Figure 7.1.

Moreover, as proteins are grouped based on their identified peptides, carefully defined rules are required on when and how to group these proteins. There are two possible approaches here: the first approach consists of grouping all proteins that share one or more identified peptides (i.e., the shared peptide rule), while the second approach consists of only grouping proteins that share the same set (or subset) of identified peptides (i.e., the shared peptide set rule). These two approaches can also be interpreted as grouping at two different levels: the protein group level (based on the shared peptide rule) and the

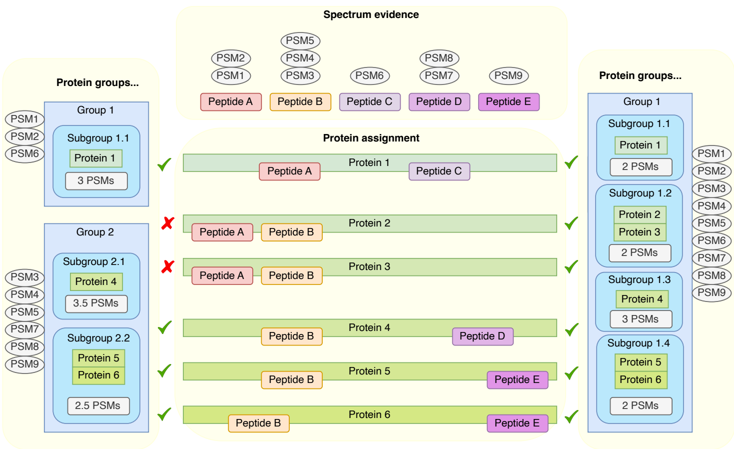


Figure 7.1: Protein grouping algorithms Occam's razor (left) and anti-Occam's razor (right). Groups can be based on shared peptide rule (protein groups) or on shared peptide set rule (protein subgroups). This figure also illustrates how PSMs are assigned to protein (sub)groups and shows the weighted PSM count for subgroups. When a PSM is assigned to multiple subgroups, it will be calculated as one divided by the number of subgroups, which can result in fractional PSM counts.

protein subgroup level (based on the shared peptide set rule). These two approaches are also visualized in Figure 7.1.

Pout2Prot implements all of these approaches: Occam's razor and anti-Occam's razor, and both of these at the protein group and protein subgroup level. During conceptualization and testing, we discovered challenges with the naive description of these algorithms. First, different protein subgroups can have the same peptide and therefore have the same spectrum assigned to them, leading to distorted spectrum counts. Second, when removing proteins using Occam's razor or when assigning subgroups using anti-Occam's razor, "undecidable" cases can occur as illustrated in Figure 7.2. In these undecidable cases, the naive approach might produce inconsistent results when the algorithm is run multiple times.

In this manuscript, we describe a new command line tool and web application that can convert .pout files from different experiments into two files containing protein groups and subgroups either as .tsv for direct use with Prophane or as human readable .csv files. Furthermore, we include a file converter that turns Proteome Discoverer output files into the .pout file format. Thus, Pout2Prot enables Percolator (or Proteome Discoverer users) to use Prophane for downstream functional and taxonomic analysis.

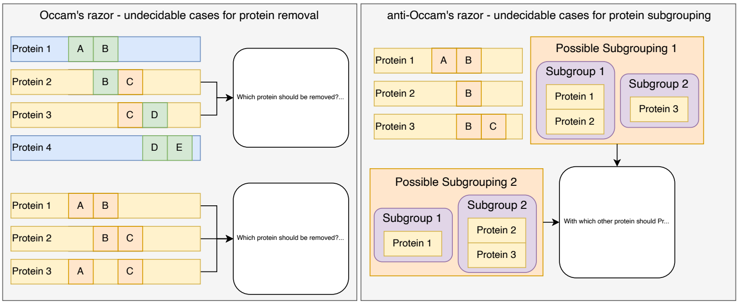


Figure 7.2: Illustration of undecidable cases. Undecidable cases are situations where peptides and proteins are matched in such a way that the naive interpretation of the algorithm cannot make a clear decision. Specifically, this occurs in Occam's razor when one of two or more proteins can be removed to explain the remaining peptides (left), and this occurs in anti-Occam's razor when a protein can be put into a subgroup with two or more other proteins that cannot be subgrouped together (right).

7.1.2 Implementation

Pout2Prot is implemented in Python and installable as a Python package from PyPI. It can then be invoked from the command line. We also provide a user-friendly and easily accessible web application of our tool (https://pout2prot.ugent.be). The transpiler Transcrypt (https://www.transcrypt.org/) was used to convert our Python package into JavaScript-compatible code and reuse it in our web application. Protein grouping analysis is efficient and can, consequently, be performed entirely on the user's local machine. Moreover, the web application processes all data locally, so that no data is sent to our servers. This safeguards user data and allows researchers to analyze confidential information more safely.

The detailed implementation of the protein grouping algorithms is visualized in the Supporting Information (Figure S1 and S2) and consists of four sub-algorithms: the creation of protein groups, the removal of proteins using the rule of maximum parsimony, and a subgroup algorithm each for Occam's razor and anti-Occam's razor.

7.1.3 Evaluation

Pout2Prot converts .pout files to protein (sub)group files that can be immediately imported in Prophane for further downstream analysis. This Prophane input file consists of four tab-separated fields: sample category, sample name, protein accessions, and spectrum count. The sample category allows users to divide their experiment in different categories (e.g., “control” and “disease”). If no sample categories are provided, these values will be identical to the sample name, which results in individual quantification by Prophane. The sample name is identical to the name of the .pout file, so each protein (sub)group can be traced back to its origin file. The protein accessions will contain the proteins present in the protein (sub)group, based on the chosen strategy. Finally, the spectrum count contains the weighted spectrum count from all PSMs present in that protein (sub)group, with PSMs present in multiple subgroups counted as fractional values in each subgroup.

7.1.3.1 Qualitative comparison to other tools

To develop a protein grouping algorithm and to truly compare different protein grouping tools, the behavior of the algorithm must be validated against a set of well-defined data, where differences between expected and observed behavior (i.e., the composition of the groups) can be clearly distinguished. During the development of Pout2Prot, it quickly became clear that multiple algorithms can solve certain test cases, but fail at others. This also led to the discovery of the undecidable cases outlined in Figure 7.2. Therefore, we created 14 test cases (Supporting Information, Figures S3–S16) that capture all possible pitfalls of protein grouping algorithms, and solved those cases by using both Occam’s razor and anti-Occam's razor at the protein group and subgroup level. To resolve the issue of undecidability, we propose that no choice should be made at all. For undecidable cases for protein removal (Occam's razor), no protein should be removed, and for undecidable cases of protein subgroups (anti-Occam's razor), the protein in question should remain in its own subgroup.

shows the result of the comparison between five protein grouping tools: PIA, Fido (integrated into Percolator), MetaProteomeAnalyzer (MPA), X!TandemPipeline, and Pout2Prot. To run tests with each tool, appropriate input files that reflect the test cases were created manually, and these are all available on the Pout2Prot GitHub repository. If a test case did not produce the expected output, it was investigated more closely to ensure this was not the result of differences between, or potential errors in, these input files. For undecidable cases, it was verified that the random choice behavior could be observed (i.e., multiple analyses, different results). For anti-Occam's razor subgrouping Cases 3 and 10, a difference in behavior was observed for PIA and Fido that can be attributed to a different conception of what a protein group is. Specifically, if a protein's peptide set is a strict subset of another protein's peptide set, PIA and Fido will not group these two proteins, while MPA and Pout2Prot will. Of all the tests that could be run, one resulted in an error: the algorithm for X!TandemPipeline for Case 13. In this case, only one of the six proteins was put into a single group, which leads to a situation where one of the three peptides was not explained by the resulting groups.

While we tried to make a fair comparison, it should be noted that PIA also offers and even recommends another option that falls in between Occam's razor and anti-Occam's razor. This method called SpectrumExtractor uses spectrum level information to determine which proteins should be removed or grouped together. Furthermore, Fido offers an option similar to Occam's razor that operates at the level of the protein database. Percolator and other tools (e.g., Triqler (The and Käll, 2019)) assign probabilities to proteins instead of making a binary choice for each protein. In contrast, Pout2Prot is based on the binary model in which a peptide or protein is either identified or not. This choice is influenced by the fact that a probabilistic approach makes the assignment of taxonomies and functions in metaproteomics very difficult.

7.1.3.2 Performance evaluation

To evaluate the performance of Pout2Prot, we tested it on a metaproteomics data set, derived from the six selected SIHUMIx (Schäpe *et al.*, 2019) data sets used in the Critical Assessment of Metaproteome Investigation (CAMPI) study (Van Den Bossche, Kunath, *et al.*, 2021). Here, we used the X!Tandem (Craig and Beavis, 2004) files available on PRIDE (Perez-Riverol *et al.*, 2019) (PXD023217) to (i) convert

Tool	Occam grouping	Occam subgrouping	anti-Occam grouping	anti-Occam subgrouping
PIA		case 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14 successful		case 1, 2, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14 successful
		case 12 undecidable		case 3, 10 different approach
Fido (Percolator)				case 1, 2, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14 successful
				case 3, 10 different approach
MPA			all successful	case 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12 successful
				case 8, 13, 14 undecidable
X!TandemPipeline	case 1, 2, 3, 4, 5, 6, 8, 10, 11, 14 successful	case 1, 2, 3, 4, 5, 6, 8, 10, 11, 14 successful		
	case 7, 9, 12 undecidable	case 7, 9, 12 undecidable		
	case 13 incorrect	case 13 incorrect		
Pout2Prot	all successful	all successful	all successful	all successful

Figure 7.3: Comparison of the outcome of test cases for five protein grouping tools. The 14 test cases were run with the PIA, Fido (Percolator), MetaProteomeAnalyzer (MPA), X!TandemPipeline, and Pout2Prot. Test cases producing the expected outcome are marked as "successful" (green). Otherwise, these are either categorized as "undecidable" (yellow) if a random choice was made in case of undecidability, "incorrect" (red) if the result cannot be explained logically, and as "different approach" for PIA and Fido, because the anti-Occam protein subgrouping approach used here follows different rules (blue). If a tool does not implement a certain grouping method it is marked as "not implemented" (grey).

these files to Percolator Input (.pin) files with tandem2pin, (ii) process the .pin files with Percolator resulting in Percolator Output (.pout) files, and (iii) convert these .pout files to protein (sub)grouping files with Pout2Prot, once using Occam's razor, once using anti-Occam's razor.

Interestingly, the identification rate (the number of identified spectra divided by the total number of spectra measured) at 1% False Discovery Rate (FDR) increases on average by 7% when using Percolator (Figure 7.4a, blue bars (X!Tandem) vs red bars (Percolator)). It is important to notice that Pout2Prot takes into account the PSM FDR, not the protein FDR. As expected and described before, the semisupervised machine learning algorithm Percolator is able to increase the number of PSMs due to the better separation of true and false matches (Käll *et al.*, 2007; Bouwmeester *et al.*, 2020). More interestingly, we examined the effect of Percolator on the number of protein groups and subgroups. To establish the number of protein (sub)groups before Percolator analysis, we reanalyzed the publicly available raw files of the selected data sets with MPA, also using X!Tandem with identical search settings. Note here that MPA is only able to group proteins according to the anti-Occam's strategy, so only those numbers were compared in the section below.

In Figure 7.4b, we observe that after Percolator analysis, the number

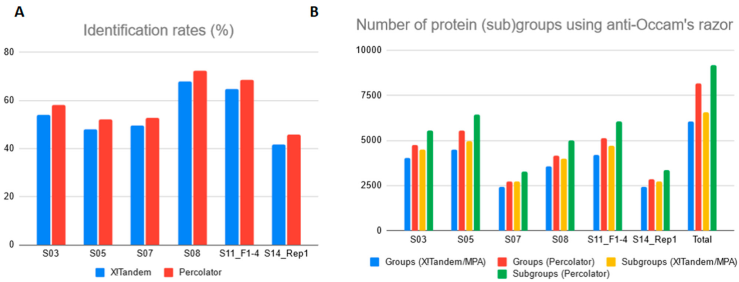


Figure 7.4: (A) Identification rates per sample for X!Tandem and Percolator analyses. Here, the identification rate was defined as the number of identified spectra divided by the total number of spectra measured. (B) Number of protein (sub)groups compared between X!Tandem and Percolator for the anti-Occam's razor strategy, and number of protein (sub)groups using Percolator for the Occam's razor strategy. S03, S05, S07, S08, S11_F1–4, and S14_Rep1 refer to the six SIHUMIx samples.

of protein groups per sample increased by 18.5% on average (blue vs red bars) and the number of protein subgroups per sample increased by 25.3% on average (yellow vs green bars). The total number of groups and subgroups across all samples increased more drastically (by 34.7% and 39.9%, respectively) in comparison to the averages per sample. All raw data is available in Supporting Information (Tables S1 and S2).

Furthermore, we also investigated the effect on the number of protein (sub)groups of combining different fractions at different places in the workflow. We combined (i) the Mascot Generic Format (.mgf) files before the X!Tandem search, (ii) before the Percolator search, and (iii) before Pout2Prot protein inference. Since the range for the number of protein (sub)groups constitute a 2–3% difference, the point in the workflow where the different files are combined, is of minimal impact (Supporting Information, Table S3). For completeness, an example result file for taxonomic and functional analysis after processing of Pout2Prot output in Prophane can be found in Supporting Information, Figures S17 and S18). In addition, the time for a Pout2Prot analysis (Occam's razor) for the complete SIHUMIx experiment via the web service was less than 5s.

7.1.4 Conclusion

Pout2Prot enables the conversion of Percolator output (.pout) files to protein group and protein subgroup files, based on either the Occam's razor or anti-Occam's razor strategy, and therefore closes an important gap in the bioinformatic workflow of metaproteomics data analysis. Moreover, Pout2Prot also allows the user to create protein (sub)groups across experiments. The output of Pout2Prot can be imported directly into Prophane, which in turn allows users to perform downstream taxonomic and functional analysis of metaproteomics samples.

7.1.5 Acknowledgements

This work has benefited from collaborations facilitated by the Metaproteomics Initiative (https://metaproteomics.org/) whose goals are to promote, improve, and standardize metaproteomics (Van Den Bossche, Arntzen, *et al.*, 2021). TVDB, PV, LM, and BM would like to acknowledge the Research Foundation - Flanders (FWO) [grants 1S90918N, 1164420N, G042518N, and 12I5220N]. LM also acknowledges support from the European Union's Horizon 2020 Programme under Grant Agreement 823839 [H2020-INFRAIA-2018-1]. KS and DB would like to acknowledge the German Federal Ministry of Education and Research (BMBF) of the project "MetaProteomanalyzer Service" within the German Network for Bioinformatics Infrastructure (de.NBI) [031L103]. The authors declare no conflict of interest.

7.2 Highlighting taxonomic diversity of metaproteomic samples in metabolic pathways

Chapter 8

Future work

No scientific work is ever complete and there remains a lot of challenges in the field of metaproteomics and proteogenomics that still need to be solved. In this section, I discuss some issues and challenges that currently arise and what needs to be overcome in order to solve these.

8.1 Modelling the inherent ambiguities in the Unipept matching system

8.1.1 Current situation

In a process prior to database construction, peptide sequences are reconstructed by spectral search engines. During the construction of the protein reference database, Unipept aggregates the functional and taxonomic annotations of proteins by grouping them by exact matching of peptide sequences. Since a single mass spectrum can, however, be explained by different peptide sequences, a search engine sometimes needs to pick and output the most probable peptide sequence that explains this spectrum. This inherent ambiguity is ignored during the Unipept database construction process by only considering peptide sequence similarity when grouping and summarising the functional and taxonomic annotations of peptides.

Right now, all proteins that are fed into Unipept during database construction will be in-silico tryptically digested into peptides. Then, a similarity function is used to compare peptides with each other, and the protein-level annotations of peptides that are found to be equal, according to this similarity function, are grouped and summarised. Currently, two peptides are considered to be equal when their peptide sequences are exactly the same. In this case, we rely on the search engines that are matching experimental mass spectra with peptide sequences in order to get a reliable result. In practice, it is common for multiple peptide sequences to correspond with the same mass spectrum, causing ambiguity in the spectral matching process. But, since Unipept only looks at sequence identity, part of this spectral ambiguity is (potentially unjustified) ignored.

8.1.2 Proposed work plan

Instead of grouping together different peptides by only looking at the sequence similarity, we can predict the mass spectrum of each peptide in the Unipept protein reference database using a tool such as MS2PIP (Degroeve and Martens, 2013). MS2PIP employs the XGBoost machine learning algorithm in order to predict MS2 signal peak intensities from peptide sequences and has proven to produce very reliable results. Tryptic peptides that were identical when using the sequence-based similarity metric, will also be identical with this new metric. But important to note is that the spectrum-based similarity metric will be "less strict" than sequence-based similarity, meaning that peptides that were different under the sequence-based similarity can now be considered identical. Consequently, this also means that resolution of the taxonomic and functional profile for a metaproteomics sample will go down. By now taking into account the spectral ambiguity that was previously masked by Unipept's analysis, we can design an experiment to investigate to what extent this ambiguity proves to be an issue.

8.1.2.1 Update the Unipept database construction process

The construction process of the Unipept database is currently not designed to work with different similarity metrics when it comes to grouping peptides. A first, big change, will have to be made to this construction process in order to allow it to accept arbitrary similarity metrics. This will, in turn, allow us to implement the spectrum based similarity metric (as well as variants) and easily plug them into the database construction process. No other major changes will have to be made to the finalised Unipept database, the underlying database structure will be more or less the same.

8.1.2.2 Perform a first experiment

In order to test the hypothesis that we proposed at the start of this section, we will have to perform an experiment in which we compare the end result of a metaproteomics analysis using the updated spectrum-based similarity approach for Unipept against using the traditional sequence-based similarity approach. For this experiment, we can analyse the SIHUMI sample using each of the approaches and compare the end results. SIHUMI is a sample that was artificially constructed for the renowned CAMPI study (Van Den Bossche, Kunath, *et al.*, 2021) and for which the exact taxonomic composition is known. We expect to find that the spectrum-similarity based Unipept leads to a lower, but more realistic, taxonomic and functional resolution for the provided sample.

8.1.2.3 Predict retention times

Instead of only looking at the predicted MS2 peak intensities of a peptide sequence using MS2PIP, we can go one step further and also predict retention times and take these into account and expand the similarity metric that was developed during the previous steps. Retention times can be predicted using the DeepLC tool (Bouwmeester *et al.*, 2021) and will cause some peptides, that are similar when we compare them solely by spectra, to be different if we also take into account this predicted retention time. Since the Unipept database construction process has already been updated to be compatible with hot-pluggable similarity metrics at this point, we only need to implement a new similarity metric and rebuild the Unipept database.

8.1.2.4 Perform a second experiment

Finally, we can augment the experiment designed earlier to compare the results between the updated spectrum-based similarity approach for Unipept against using the traditional sequence-based similarity approach with the analysis using the spectrum-similarity based Unipept that also takes into account retention times for the tryptic peptides. For this comparison, we expect the taxonomic and functional resolution of the end result to have increased when comparing it to the spectrum-based similarity of before.

8.2 Identification and analysis of arbitrary peptides, including variants

8.2.1 Current situation

Unipept requires that all input peptides are tryptic in order to be able to match them with peptides in its reference database. However, researchers are transitioning to experiment with datasets that contain other peptide formats that Unipept currently can not use for downstream analysis. In this work package, I will therefore design a new index structure for Unipept based on bidirectional FM-indices and search schemes. This new index will no longer require the input peptides to adhere to a fixed format.

Over the last 10 years, a lot of research has gone into the development and improvement of efficient data structures for sequence alignment. One such highly-used data structure that offers excellent performance is the FM-index (Ferragina and Manzini, 2000). An FM-index is produced by computing the Burrows-Wheeler transform of a specific string and allows to look up if (and where) a pattern occurs in the preprocessed text in a very efficient manner. By adjusting the FM-index and its accompanying query algorithms, we are not limited to matching exact strings but we can also detect if a specific sequence (with up to a certain number of k mismatches) is present in a longer string.

A big advantage of these FM-indices over the approach that Unipept currently follows for matching input peptides with proteins in the protein reference database is that the FM-index allows us to match arbitrary peptide sequences with proteins (instead of only directly matching tryptic peptides as we do today). This opens up the possibility to go and analyse semi-tryptic peptides using Unipept, or even matching tryptic peptides with missed cleavages. At this point, it is already possible to analyse peptides with missed cleavages, but this drastically slows down the analysis since a lot of Unipept's precomputed aggregations are not available in this case.

In order to efficiently match a peptide (with up to k mismatches) with a protein, an FM-index by itself does not suffice and we need to look for improved data structures. This is where search schemes come into play. A search scheme is a strategy that describes how a bi-directional FM-index can be queried such that patterns with up to k mismatches can efficiently be matched with a long string (such as a protein). Search schemes were first proposed by (Lam et al.) and were further generalised by (Kucherov *et al.*, 2014).

8.2.2 Proposed work plan

8.2.2.1 Design and implement a new index structure for Unipept

The first step that should be performed in order to allow Unipept to match peptides of arbitrary format, is to design a new index structure for our database, using FM-indices, and to implement this new index structure. For each protein in the protein reference database, we can construct an FM-index of the protein sequence, which allows us to match any kind of peptide with this protein and keep it in our new index structure. We can use the Rust programming language since it is designed with performance and parallelization in mind, and it already provides a very good, open-source implementation of the FM-index data structure.[1] These changes will allow us to match arbitrary peptides

[1] https://docs.rs/fm-index/latest/fm_index/

and peptides with missed cleavages using Unipept.

8.2.2.2 Implement a bi-directional FM-index

A second step consists of updating the FM-index data structure that was used during the previous step such that supports matching patterns in two directions (backwards *and* forwards). This so-called bi-directional FM-index is extensively described in (Lam *et al.*) and is required for efficiently approximate pattern matching using search schemes. We can improve and expand the existing, open-source Rust FM-index implementation from the previous step such that it allows searching in two directions. By contributing to this open-source project we do not need to start from scratch, and we can share our improvements with other researchers around the globe.

8.2.2.3 Implement and validate search scheme prototypes

A lot of different proposals for search schemes already exist at this point. During this step, we can take a look at a selection of search schemes such as the Pigeon H.S. (Fletcher and Patty, 1996), 01*0 seeds (Vroland *et al.*, 2016), schemes proposed by Kucherov (Kucherov *et al.*, 2014), and Man_{best} (Pockrandt, 2019). All of these search schemes will have to be benchmarked for performance and applicability for our needs. The search scheme that comes out as best from this comparison can then be tweaked and refined further.

8.2.2.4 Integrate the best search scheme into Unipept

Finally, the search scheme that was selected during the previous step needs to be integrated into the Unipept database index structure. By doing so, Unipept will effectively support matching peptides with up to k mismatches with the proteins from a reference database.

8.3 Towards a meta-, multi-omics Unipept

8.3.1 Current situation

Over the last few years, we have been working very hard to build the Unipept Desktop application which provides a first step in the integration of metagenomics information in metaproteomics experiments. By first performing a metagenomics experiment on a sample, researchers are able to derive the taxonomic profile of the ecosystem under study. This taxonomic profile can then be used in a subsequent step as a guide for constructing a targeted protein reference database (that only contains proteins that are associated with the taxa that were detected during the metagenomics experiment).

As a possible future addition to Unipept, I propose to further integrate data from different "omics" sources such as transcriptomics and metagenomics into Unipept Building on the individual strength of these techniques, an aggregated view enables researchers to gain a much deeper insight into and understanding of what exactly is taking place in a complex ecosystem. By augmenting Unipept with support for both metagenomics and metatranscriptomics analyses, it has the potential to become the "go-to" tool for all analyses related to the "meta-omics" research disciplines. The ultimate goal of this work package is to transform Unipept into the first tool that provides a complete global overview of multi-disciplinary "meta-omics" experiments.

8.3.2 Proposed work plan

8.3.2.1 Allow Unipept to directly load metagenomics reads

By improving Unipept with the capability of loading metagenomics reads directly into the software, we can allow users to construct a fully custom protein reference database from these reads. At this time, a targeted reference database is always constructed by extracting and filtering proteins from the UniProtKB resource. This works very well when the organisms under study have been analysed before and their proteomic profile is available in the UniProtKB resource. By providing support for the construction of protein reference databases from metagenomic reads instead, we can also allow organisms that are not present in the UniProtKB resource to be analysed.

8.3.2.2 Design and implement new visualizations for metagenomics experiments

Unipept provides a lot of valuable and interactive data visualisations that increase the insight of researchers into the taxonomic and functional profile of a metaproteomics ecosystem. These visualisations are implemented in a very generic way that allows them to be applicable to other situations as well. I propose to expand Unipept with the ability to visualise the taxonomic profile determined by a metagenomics experiment, and the potential functional profile that is determined by performing a metatranscriptomics experiment. This will increase the insight of users into the ecosystem that they are currently investigating.

8.4 Differential analysis of metaproteomics data

References

Ashburner,M. *et al.* (2000) Gene Ontology: Tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Ashwood,C. *et al.* (2020) Proceedings of the EuBIC-MS 2020 Developers' Meeting. *EuPA Open Proteomics*, **24**, 1–6.

Audain,E. *et al.* (2017) In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *Journal of Proteomics*, **150**, 170–182.

Bar-Joseph,Z. *et al.* (2001) Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, **17**, S22–S29.

Bibbò,S. *et al.* (2020) Fecal Microbiota Signatures in Celiac Disease Patients With Poly-Autoimmunity. *Frontiers in Cellular and Infection Microbiology*, **10**, 349.

Bierman,G. *et al.* (2014) Understanding TypeScript. In, Jones,R. (ed), *ECOOP 2014 Object-Oriented Programming*, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 257–281.

Bostock,M. *et al.* (2011) D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, **17**, 2301–2309.

Bouwmeester,R. *et al.* (2021) DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods*, **18**, 1363–1369.

Bouwmeester,R. *et al.* (2020) The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *PROTEOMICS*, **20**, 1900351.

Bucchini,F. *et al.* (2021) TRAPID 2.0: A web application for taxonomic and functional analysis of de novo transcriptomes. *Nucleic Acids Research*, **49**, e101.

Craig,R. and Beavis,R.C. (2004) TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.

Degroeve,S. and Martens,L. (2013) MS2PIP: A tool for MS/MS peak intensity prediction. *Bioinformatics*, **29**, 3199–3203.

Easterly,C.W. *et al.* (2019) metaQuantome: An Integrated, Quantitative Metaproteomics Approach Reveals Connections Between Taxonomy and Protein Function in Complex Microbiomes *. *Molecular & Cellular Proteomics*, **18**, S82–S91.

Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Research*, **40**, D136–D143.

Ferragina,P. and Manzini,G. (2000) Opportunistic data structures with applications. In, *Proceedings 41st Annual Symposium on Foundations of Computer Science.*, pp. 390–398.

Finn,R.D. *et al.* (2017) InterPro in 2017beyond protein family and domain annotations. *Nucleic Acids Research*, **45**, D190–D199.

Fletcher,P. and Patty,C.W. (1996) Foundations of Higher Mathematics PWS Publishing Company.

Fruzangohar,M. *et al.* (2013) Comparative GO: A Web Application for Comparative Gene Ontology and Gene Ontology-Based Gene Selection in Bacteria. *PLOS ONE*, **8**, e58759.

Gruvaeus,G. and Wainer,H. (1972) Two Additions to Hierarchical Cluster Analysis. *British Journal of Mathematical and Statistical Psychology*, **25**, 200–206.

Gurdeep Singh,R. *et al.* (2019) Unipept 4.0: Functional Analysis of Metaproteome Data. *Journal of Proteome Research*, **18**, 606–615.

Herbst,F.-A. *et al.* (2016) Enhancing metaproteomicsThe value of models and defined environmental microbial systems. *PROTEOMICS*, **16**, 783–798.

Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*,

4, 44–57.

Huerta-Cepas,J. *et al.* (2019) eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, **47**, D309–D314.

Hunter,S. *et al.* (2009) InterPro: The integrative protein signature database. *Nucleic Acids Research*, **37**, D211–D215.

Huson,D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377–386.

Jagtap,P.D. *et al.* (2015) Metaproteomic analysis using the Galaxy framework. *PROTEOMICS*, **15**, 3553–3565.

Jansson,J.K. and Baker,E.S. (2016) A multi-omic future for microbiome studies. *Nature Microbiology*, **1**, 1–3.

Käll,L. *et al.* (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, **4**, 923–925.

Klopfenstein,D.V. *et al.* (2018) GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports*, **8**, 10872.

Kucherov,G. *et al.* (2014) Approximate String Matching Using a Bidirectional Index. In, Kulikov,A.S. *et al.* (eds), *Combinatorial*

Pattern Matching, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 222–231.

Lam,T.W. *et al.* High Throughput Short Read Alignment via Bi-directional BWT.

Langella,O. *et al.* (2017) X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *Journal of Proteome Research*, **16**, 494–503.

Li,J. *et al.* (2014) An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, **32**, 834–841.

Lohmann,P. *et al.* (2020) Function is what counts: How microbial community complexity affects species, proteome and pathway coverage in metaproteomics. *Expert Review of Proteomics*, **17**, 163–173.

Louca,S. *et al.* (2016) Decoupling function and taxonomy in the global ocean microbiome. *Science*, **353**, 1272–1277.

Martens,L. and Hermjakob,H. (2007) Proteomics data validation: Why all must provide data. *Molecular BioSystems*, **3**, 518–522.

Merkel,D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. 5.

Mesuere,B. *et al.* (2018) High-throughput metaproteomics data analysis with Unipept: A tutorial. *Journal of Proteomics*, **171**, 11–22.

Mesuere,B. *et al.* (2012) Unipept: Tryptic Peptide-Based Biodiversity Analysis of Metaproteome Samples. *Journal of Proteome Research*, **11**, 5773–5780.

Mesuere,B. *et al.* (2016) Unipept web services for metaproteomics analysis. *Bioinformatics*, **32**, 1746–1748.

Muth,T. *et al.* (2018) MPA Portable: A Stand-Alone Software Package for Analyzing Metaproteome Samples on the Go. *Analytical Chemistry*, **90**, 685–689.

Muth,T. *et al.* (2015) The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation. *Journal of Proteome Research*, **14**, 1557–1565.

Nesvizhskii,A.I. and Aebersold,R. (2005) Interpretation of Shotgun Proteomic Data. *Molecular & Cellular Proteomics*, **4**, 1419–1440.

Ondov,B.D. *et al.* (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.

Orsburn,B.C. (2021) Proteome DiscovererA Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes*, **9**, 15.

Palomba,A. *et al.* (2021) Time-restricted feeding induces Lactobacillus- and Akkermansia-specific functional changes in the rat fecal microbiota. *NPJ biofilms and microbiomes*, **7**, 85.

Perez-Riverol,Y. *et al.* (2019) The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Research*, **47**, D442–D450.

Pockrandt,C.M. (2019) Approximate String Matching: Improving Data Structures and Algorithms.

Quast,C. *et al.* (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, **41**, D590–D596.

Ram,R.J. *et al.* (2005) Community Proteomics of a Natural Microbial Biofilm. *Science*, **308**, 1915–1920.

Rechenberger,J. *et al.* (2019) Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae. *Proteomes*, **7**, 2.

Riffle,M. *et al.* (2018) MetaGOmics: A Web-Based Tool for Peptide-Centric Functional and Taxonomic Analysis of Metaproteomics Data. *Proteomes*, **6**, 2.

Rodríguez-Valera,F. (2004) Environmental genomics, the big picture? *FEMS Microbiology Letters*, **231**, 153–158.

Rudney,J.D. *et al.* (2015) Protein relative abundance patterns associated with sucrose-induced dysbiosis are conserved across taxonomically diverse oral microcosm biofilm models of dental caries. *Microbiome*, **3**, 69.

Sajulga,R. *et al.* (2020) Survey of metaproteomics software tools for functional microbiome analysis. *PLOS ONE*, **15**, e0241503.

Sakai,R. *et al.* (2014) Dendsort: Modular leaf ordering methods for dendrogram representations in R. *F1000Research*, **3**, 177.

Sansone,S.-A. *et al.* (2012) Toward interoperable bioscience data. *Nature Genetics*, **44**, 121–126.

Schallert,K. *et al.* (2022) Pout2Prot: An Efficient Tool to Create Protein (Sub)groups from Percolator Output Files. *Journal of Proteome Research*, **21**, 1175–1180.

Schäpe,S.S. *et al.* (2019) The Simplified Human Intestinal Microbiota (SIHUMIx) Shows High Structural and Functional Resistance against Changing Transit Times in In Vitro Bioreactors. *Microorganisms*, **7**, 641.

Schiebenhoefer,H. *et al.* (2020) A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane. *Nature Protocols*, **15**, 3212–3239.

Schiebenhoefer,H. *et al.* (2019) Challenges and promise at the interface of metaproteomics and genomics: An overview of recent progress in metaproteogenomic data analysis. *Expert Review of Proteomics*, **16**, 375–390.

Schlicker,A. *et al.* (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, **7**, 302.

Schneider,T. *et al.* (2011) Structure and function of the symbiosis partners of the lung lichen (Lobaria pulmonaria L. Hoffm.) analyzed by metaproteomics. *PROTEOMICS*, **11**, 2752–2756.

Schoch,C.L. *et al.* (2020) NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database*, **2020**, baaa062.

Sokal,R.R. and Michener,C.D. (1958) A Statistical Method for Evaluating Systematic Relationships University of Kansas.

Tanca,A. *et al.* (2022) Metaproteomic Profile of the Colonic Luminal Microbiota From Patients With Colon Cancer. *Frontiers in Microbiology*, **13**.

Tanca,A. *et al.* (2016) The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome*, **4**, 51.

The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, **47**, D330–D338.

The,M. and Käll,L. (2019) Integrated Identification and Quantification Error Probabilities for Shotgun Proteomics * [S]. *Molecular & Cellular Proteomics*, **18**, 561–570.

The UniProt Consortium (2019) UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**, D506–D515.

The UniProt Consortium (2021) UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, **49**, D480–D489.

Uszkoreit,J. *et al.* (2015) PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. *Journal of Proteome Research*, **14**, 2988–2997.

Van Den Bossche,T. *et al.* (2020) Connecting MetaProteomeAnalyzer and PeptideShaker to Unipept for Seamless End-to-End Metaproteomics Data Analysis. *Journal of Proteome Research*, **19**, 3562–3566.

Van Den Bossche,T., Kunath,B.J., *et al.* (2021) Critical Assessment of MetaProteome Investigation (CAMPI): A multi-laboratory comparison of established workflows. *Nature Communications*, **12**, 7305.

Van Den Bossche,T., Arntzen,M.Ø., *et al.* (2021) The Metaproteomics Initiative: A coordinated approach for propelling the functional characterization of microbiomes. *Microbiome*, **9**, 243.

Van der Jeugt,F. *et al.* (2022) UMGAP: The Unipept MetaGenomics Analysis Pipeline. *BMC Genomics*, **23**, 433.

Verschaffelt,P., Van Den Bossche,T., Gabriel,W., *et al.* (2021) MegaGO: A Fast Yet Powerful Approach to Assess Functional Gene Ontology Similarity across Meta-Omics Data Sets. *Journal of Proteome Research*, **20**, 2083–2088.

Verschaffelt,P. *et al.* (2020) Unipept CLI 2.0: Adding support for visualizations and functional annotations. *Bioinformatics*, **36**, 4220–4221.

Verschaffelt,P. *et al.* (2023) Unipept Desktop 2.0: Construction of targeted reference protein databases for proteogenomics analyses. 2023.02.09.527820.

Verschaffelt,P., Van Den Bossche,T., Martens,L., *et al.* (2021) Unipept Desktop: A Faster, More Powerful Metaproteomics Results Analysis Tool. *Journal of Proteome Research*, **20**, 2005–2009.

Verschaffelt,P. *et al.* (2022) Unipept Visualizations: An interactive visualization library for biological data. *Bioinformatics*, **38**, 562–563.

Vroland,C. *et al.* (2016) Approximate search of short patterns with high error rates using the 01⊠0 lossless seeds. *Journal of Discrete Algorithms*, **37**, 3–16.

Waardenberg,A.J. *et al.* (2015) CompGO: An R package for comparing and visualizing Gene Ontology enrichment differences between DNA binding experiments. *BMC Bioinformatics*, **16**, 275.

Webb,E.C. (1992) Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.*

Wilmes,P. *et al.* (2015) A decade of metaproteomics: Where we stand and what the future holds. *PROTEOMICS*, **15**, 3409–3417.

Wilmes,P. and Bond,P.L. (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental Microbiology*, **6**, 911–920.

Yates,J.R. (2019) Recent technical advances in proteomics. *F1000Research*, **8**, F1000 Faculty Rev–351.

Zhang,X. and Figeys,D. (2019) Perspective and Guidelines for Metaproteomics in Microbiome Studies. *Journal of Proteome Research*, **18**, 2370–2380.