Comparative Analysis of High-Throughput Metaproteomics Data in Unipept

Pieter Verschaffelt

Contents

Acknowledgements

Summary

Samenvatting

Chapter 1

Introduction

Chapter 2

Unipept Desktop

The introductory text for this chapter comes here...

2.1 Unipept Desktop: a faster, more powerful metaproteomics analysis tool

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.

2.1.1 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.

2.1.2 Implementation

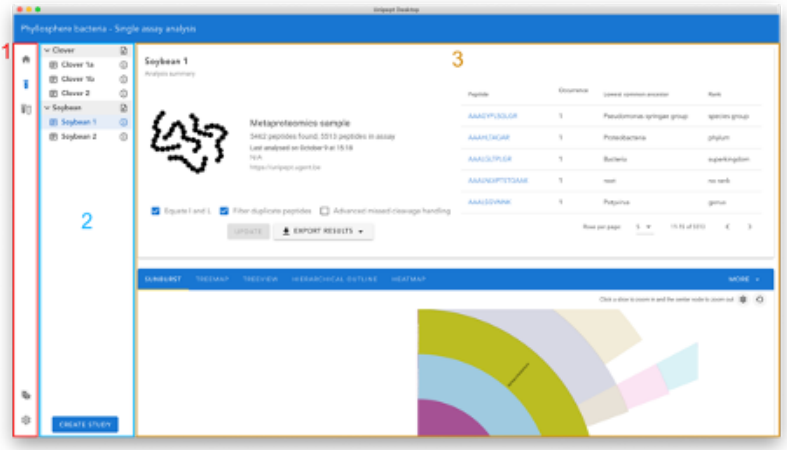

Figure 2.1: Screenshot of the Unipept Desktop application. The analysis page of the desktop application is depicted here and consists of three main parts: the sidebar that is used to navigate between the different analysis pipelines and functions of the application (1), the project explorer that displays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

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 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.

2.1.2.1 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 web site is closed.

2.1.2.2 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).

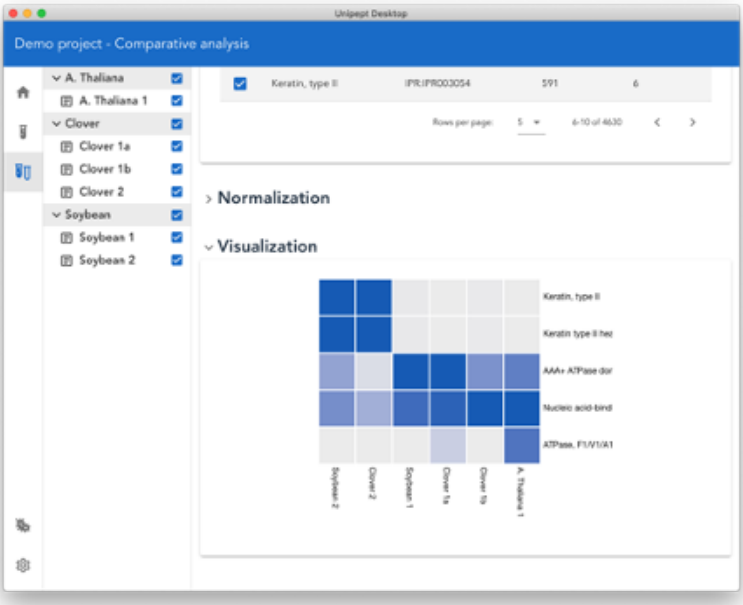

Figure 2.2: Screenshot of the inter-assay comparative analysis pipeline. Note that it is possible to select multiple assays from the project explorer. A heatmap is constructed from the set of items that were selected for comparison at the top of the page.

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 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 1 below.

2.1.3 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 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.

2.1.4 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.

2.1.5 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.].

2.2 Support for novel proteogenomics analysis in Unipept

Chapter 3

Functional analysis of metaproteomics datasets

The introductory text for this chapter comes here...

3.1 Unipept CLI 2.0: adding support for visualisations and functional annotations

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).

3.1.1 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 (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).

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.

3.1.2 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.

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.

3.1.3 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.

3.1.4 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.].

References

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

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.

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.

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.

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

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

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.

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

D330–D338.

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.

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

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

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.