Journal of Proteomics 171 (2018) 11–22

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

High-throughput metaproteomics data analysis with Unipept: A tutorial

Bart Mesuere [a,b,c,*], Felix Van der Jeugt [a], Toon Willems [a], Tom Naessens [a], Bart Devreese [d], Lennart Martens [b,c], Peter Dawyndt [a]

[a] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium
[b] VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium
[c] Department of Biochemistry, Ghent University, Ghent, Belgium
[d] Laboratory for Protein Biochemistry and Biomolecular Engineering, Ghent University, Ghent, Belgium

ARTICLE INFO

Article history:
Received 31 January 2017
Received in revised form 15 May 2017
Accepted 22 May 2017
Available online 24 May 2017

Keywords:
Metaproteomics
Tutorial
Biodiversity
Data analysis

ABSTRACT

In recent years, shotgun metaproteomics has established itself as an important tool to study the composition of complex ecosystems and microbial communities. Two key steps in metaproteomics data analysis are the inference of proteins from the identified peptides, and the determination of the taxonomic origin and function of these proteins. This tutorial therefore introduces the Unipept command line interface (http://unipept.ugent.be/clidocs) as a platform-independent tool for such metaproteomics data analyses. First, a detailed overview is given of the available Unipept commands and their functions. Next, the power of the Unipept command line interface is illustrated using two case studies that analyze a single tryptic peptide, and a set of peptides retrieved from a shotgun metaproteomics experiment, respectively. Finally, the analysis results obtained using these command line tools are compared with the interactive taxonomic analysis that is available on the Unipept website.

1. Introduction

The field of microbial ecology has evolved dramatically through the introduction of omics technologies. The introduction of metagenomics in particular has allowed growing insight into the composition and functional diversity of complex ecological systems. At the same time, metaproteomics is quickly gaining relevance as a complementary and powerful analytical tool in the study of entire (bacterial) communities [1]. The main benefit of protein analysis in metaproteomics is similar to the benefit it provides for single species studies: knowledge on the protein level provides improved insight into the actual functional components of each species in the population. Building on top of advances in instrumentation, sample preparation, and computational approaches [2,3,4], metaproteomics analyses are now within reach of an ever-growing number of researchers. Consequently, a broad range of applications have already cropped up in the recent literature, ranging from ecological studies [5], over livestock [6] and human healthcare [7,8,9], to applications in food production and safety [10,11]. However, despite the progress already made in the field, and its corresponding growth in popularity, a key issue in shotgun metaproteomics analyses remains the mapping of the identified peptides into intact proteins, and from there to species, taxa and functions [12,13].

The Unipept tool suite (http://unipept.ugent.be) aims to address this issue, and can thus help researchers analyze their metaproteomics data [14,15]. Originally developed as a web application with a focus on the analysis of the biodiversity of metaproteomics samples, the user-friendly Unipept web application allows researchers to easily search a list of tryptic peptides against UniProtKB [16]. The application then assigns a consensus taxon from the NCBI taxonomy [17] to each of the input peptides and presents the results using interactive visualizations. These consensus taxa are calculated using a Lowest Common Ancestor (LCA) algorithm, which in turn uses taxonomic annotations obtained from all UniProt records that match against one (or more) of the input peptides. To ensure fast and accurate results, the peptide-to-protein matching is based on a fast index built from UniProtKB [16], while taxa are derived from a decluttered version of the NCBI taxonomy. Recent review papers about the current state-of-the-art in metaproteomics praise Unipept for the performance and accuracy of its taxonomic identification pipeline, and for its interactive visualization framework that helps to explore the biodiversity in complex environmental samples [18,19,4].

As metaproteomics has become an established analysis technique for complex ecosystems, the focus of metaproteomics experiments

* Corresponding author.
E-mail address: bart.mesuere@ugent.be (B. Mesuere).

http://dx.doi.org/10.1016/j.jprot.2017.05.022
1874-3919/

has shifted from pilot studies to larger, comparative studies [20,21]. Of course, the latter studies generate quite a bit more data, which in turn requires more automated means for data processing. Unipept therefore offers web services and platform-independent command line tools that can be readily automated as part of larger overall workflows. This tutorial presents this Unipept command line interface, and introduces the available features step by step based on two easy to follow case studies.

2. The Unipept command line tools

The Unipept command line tools (http://unipept.ugent.be/clidocs) provide a command line interface to the Unipept web services [22], along with a few utility commands for handling proteins using the command line. All tools support fasta and plain text as input formats, and csv, xml, and json as output formats. Parallel web requests have been implemented transparently to ensure optimal performance. Similar to the Unipept web application, coding best practices were followed and the entire code base is covered with unit and integration tests. All code is released as open source under the permissive MIT License and is available on GitHub in a separate Unipept-cli repository (https://github.com/unipept/unipept-cli).

2.1. Installation

The Unipept command line tools are written in Ruby and work on Windows, macOS, and Unix. The only requirement is an installed Ruby run-time environment. Ruby 2.2 or higher is recommended, but all versions since Ruby 2.0.0, as well as JRuby are supported. The command line tools are made available as a Ruby *gem*, a packaged version of the code that can be used in combination with the RubyGems package manager. As a result, the Unipept tools can be installed with a single command:

```
$ gem install unipept
Fetching: unipept-1.3.0.gem (100%)
Successfully installed unipept-1.3.0
Parsing documentation for unipept-1.3.0
Installing ri documentation for unipept-1.3.0
Done installing documentation for unipept after 0 seconds
1 gem installed
```

After successful installation, the unipept command should be available. To check if the *gem* was installed correctly, run unipept –version. This should print the version number:

```
$ unipept --version
1.3.0
```

An update to the latest version of the command line tools is carried out using the *gem* update Unipept command.

Each Unipept command also has a built-in help function that can be displayed using the –help argument.

```
$ unipept --help
NAME
unipept - Command line interface to Unipept web services.

USAGE
unipept subcommand [options]

DESCRIPTION
The unipept subcommands are command line wrappers around the Unipept web
services.

Subcommands that start with pept expect a list of tryptic peptides as
input. Subcommands that start with tax expect a list of NCBI Taxonomy
Identifiers as input. Input is passed
- as separate command line arguments
- in a text file that is passed as an argument to the -i option
- to standard input

The command will give priority to the first way the input is passed, in
the order as listed above. Text files and standard input should have one
tryptic peptide or one NCBI Taxonomy Identifier per line.

COMMANDS
config        Set configuration options.
help          Show help.
pept2lca      Fetch taxonomic lowest common ancestor of UniProt entries
              that match tryptic peptides.
pept2prot     Fetch UniProt entries that match tryptic peptides.
pept2taxa     Fetch taxa of UniProt entries that match tryptic peptides.
taxa2lca      Compute taxonomic lowest common ancestor for given list of
              taxa.
taxonomy      Fetch taxonomic information from Unipept Taxonomy.

OPTIONS
-f --format=<value>        define the output format (available: json,
                           csv, xml) (default: csv)
-h --help                  show help for this command
   --host=<value>          specify the server running the Unipept web
                           service
-i --input=<value>         read input from file
-o --output=<value>        write output to file
-q --quiet                 disable service messages
-v --version               displays the version
```