Bioinformatics, 32(11), 2016, 1746–1748
doi: 10.1093/bioinformatics/btw039
Advance Access Publication Date: 27 January 2016
Applications Note

Gene expression

Unipept web services for metaproteomics analysis

Bart Mesuere[1,]*, Toon Willems[1], Felix Van der Jeugt[1], Bart Devreese[2], Peter Vandamme[3] and Peter Dawyndt[1]

[1]Department of Applied Mathematics, Computer Science and Statistics, [2]Laboratory for Protein Biochemistry and Biomolecular Engineering, Faculty of Sciences and [3]Laboratory for Microbiology Faculty of Sciences, Ghent University, B-9000, Ghent, Belgium

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on 17 July 2015; revised on 4 January 2016; accepted on 19 January 2016

Abstract

Summary Unipept is an open source web application that is designed for metaproteomics analysis with a focus on interactive datavisualization. It is underpinned by a fast index built from UniProtKB and the NCBI taxonomy that enables quick retrieval of all UniProt entries in which a given tryptic peptide occurs. Unipept version 2.4 introduced web services that provide programmatic access to the metaproteomics analysis features. This enables integration of Unipept functionality in custom applications and data processing pipelines.
Availability and implementation: The web services are freely available at http://api.unipept.ugent.be and are open sourced under the MIT license.
Contact: Unipept@ugent.be
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Unipept is a web application for biodiversity analysis of complex metaproteomics samples (Mesuere *et al.*, 2012). The application is powered by a fast index built from UniProtKB (UniProt Consortium, 2015) and a cleaned up version of the NCBI taxonomy (Federhen, 2012). This index enables quick retrieval of all UniProt entries in which a given tryptic peptide occurs. Using the taxonomic annotations of UniProt entries, Unipept also returns the complete set of organisms in which a given peptide occurs. This set of organisms is then processed using a Lowest Common Ancestor (LCA) algorithm to determine the taxonomic specificity of the peptide. All results are presented in clear overview tables and in an interactive treeview.

Fast computation of LCAs for given lists of peptides also enables interactive biodiversity analysis of metaproteomics datasets. The biodiversity in complex samples can then be inspected using multiple interactive visualizations such as a treeview (Supplementary Fig. S2), a sunburst view (Supplementary Fig. S3) and a treemap (Supplementary Fig. S4). All visualizations on the Unipept website can be saved as publication-grade graphics, and all analysis results can be exported as Microsoft Excel-compatible CSV files.

To guarantee optimal performance and correctness, the Unipept project pursues excellence regarding best practices for modern web application development. One example of this is automatic correctness testing by over 1000 tests after each code change. The entire application including the web services is open source and licensed under the terms of the MIT license. The source code can be found at http://github.com/unipept/unipept.

In this article, we present the latest addition to the Unipept toolbox: a set of web services that expose the Unipept analysis functions for use in other applications and data processing pipelines.

2 Methods

Unipept version 2.4 introduced the Unipept web services. These web services allow access to all Unipept peptide analysis features through a RESTful API. This means that all communication with the web services can be done using simple stateless HTTP requests, to which the server answers in JSON. JSON is an open standard for transmitting data that is both human readable and has wide support in developer tools and programming languages.

In the next sections, we discuss the available API functions by drawing parallels between usage of the Unipept website and the Unipept API. Supplementary Figure S1 displays a schematic overview of the functions, along with the expected input and output. The full documentation can be found at http://api.unipept.ugent.be and as supplementary material. Next to the documentation, the website also offers an interactive API explorer (Supplementary Fig. S5) where API requests can be composed and tested with just a few clicks.

2.1 pept2prot

The fundamental component in the Tryptic Peptide Analysis feature of Unipept is fast retrieval of all UniProt entries in which a given tryptic peptide occurs. All subsequent calculations are based on this result, and therefore the database indexes are heavily optimized to return it as fast as possible. When doing a Tryptic Peptide analysis in the web interface, the set of all matching UniProt entries is listed on the Protein Matches tab.

Its web service counterpart, pept2prot, takes a single tryptic peptide as input and returns the list of all UniProt entries containing the given tryptic peptide. By default, for each entry, the UniProt accession number, protein name and associated NCBI taxon ID are returned. Optionally, users can also request additional information fields such as the name of the organism associated with the UniProt entry, a list of cross-referenced EC numbers (Bairoch, 2000) and a list of cross-referenced GO terms (Gene Ontology Consortium, 2015). Users can also choose to equate the isobaric amino acids isoleucine (I) and leucine (L) when matching peptides to proteins, a typical option for mass spectrometry-related queries. Batch retrieval of multiple peptides at once is also supported.

2.2 pept2taxa

After matching the UniProt entries, Unipept uses the cross-referenced NCBI taxon IDs to compile a set of organisms in which the queried peptide occurs. These organisms are then mapped to their taxonomic lineages using a cleaned up version of the NCBI taxonomy database. Using the web interface, the list of organisms along with their lineage can be found in the Lineage Table tab and an interactive visualization is available in the Lineage Tree tab (Supplementary Fig. S2).

Similarly, the API function pept2taxa takes a tryptic peptide as input and returns the set of organisms associated with the UniProt entries containing the given tryptic peptide. By default, the taxon ID, name and rank are returned for each of the matched organisms. Optionally, the full lineage of each organism can be requested as a sequence of taxon IDs and/or taxon names. Batch requests and equating isoleucine and leucine are also supported.

2.3 pept2lca

The matched organisms from the previous section are then used to calculate the taxonomic lowest common ancestor (LCA). Simply put, the LCA is the most specific taxonomic rank that all matched organisms have in common. However, the algorithm used by Unipept has several advancements to better cope with taxonomic noise and misclassifications (Mesuere *et al.*, 2012). One of these improvements is the invalidation of taxonomic nodes that provide little informational values, such as those containing words like 'uncultured', 'unspecified' or 'undetermined' in their name. Invalidated taxa are ignored during LCA calculation and mapped to their first valid ancestor. These invalidated taxa would otherwise result in a drastic loss of information when used for LCA calculation. Another example is mapping strain-specific taxon IDs to their first valid parent taxon to counter the, now abandoned, practice of creating strain-level taxon IDs (Federhen *et al.*, 2014).

Correspondingly, the pept2lca function returns the LCA (taxon ID, name and rank) for a given tryptic peptide. Optionally, the full lineage (IDs and/or names) can be requested and both equating isoleucine and leucine and batch requests are supported. The LCAs for all tryptic peptides are pre-calculated and stored in the database. Therefore, the peptide matching steps can be skipped for the pept2lca function, resulting in improved performance.

2.4 taxa2lca

The Unipept LCA algorithm can also be used outside a proteomics context by using the taxa2lca function. This API function takes a list of NCBI taxon IDs and calculates their LCA by using the advanced algorithm as applied by Unipept. The result is returned by listing the taxon ID, name and rank of the LCA. Additional lineage information is also available upon request. Note that the pept2lca function can be mimicked by chaining the pept2taxa and taxa2lca functions. This is however not recommended, as pept2lca makes use of precomputed data and is therefore several orders of magnitude faster.

2.5 Taxonomy

The taxonomy function provides access to the cleaned up version of the NCBI taxonomy as used by Unipept. This function can be used, for example, to compute more detailed statistics about taxon hits or implement alternative aggregation strategies next to the LCA computation as used by Unipept. The function takes one or more taxon IDs as input and returns the name and rank for each of the given IDs. Optionally, the full lineage can also be returned.

3 Results

The Unipept project consists of two main parts: a collection of scripts to construct the database and the web application. The first part of the database construction, the code to parse UniProt, was recently updated to use Berkeley DB (Olson *et al.*, 1999), a high performance key value store, to store intermediate results. This resulted in an enormous boost in parsing speed: where the old parser took over 30 days to parse UniProt, the new approach using Berkeley DB does the job in under 10 hours. The second part of the database construction is the pre-calculation of the LCAs of all the peptides in the database. The old Ruby code was rewritten in Java and computation time was reduced from over four weeks to just 15 min with the help of some new Java 8 features. The combination of these advancements allows us to consistently offer analysis results based on the latest UniProt release. The second part of Unipept is a Ruby on Rails web application that uses JavaScript for all client side interactions. All data visualizations (Mesuere *et al.*, 2015) are made in-house with the D3.js JavaScript library (Bostock *et al.*, 2011).

The GalaxyP project already takes advantage of the new Unipept web services to integrate Unipept functionality into the Galaxy Framework (Jagtap *et al.*, 2015). Exact matching of peptides to UniProt entries is also implemented by the Peptide Match application (Chen *et al.*, 2013) of the Protein Information Resource (PIR). Where Unipept is restricted for use with tryptic peptides, Peptide

Match has no such limitation. However, the advantage of accepting all peptides comes at the cost of reduced performance. For a test set of 500 tryptic peptides, the Unipept `pept2prot` function returned all matching UniProt entries in 1.5 s whereas Peptide Match took over 33 min. Since (meta)proteomics experiments almost exclusively use trypsin to digest proteins, resulting in a list of tryptic peptides, this is a reasonable compromise (Olsen *et al.*, 2004). All functions of the Unipept API are tweaked for optimal performance and usable for high throughput data analysis. The `pept2lca` function (no counterpart in PIR), can process over 10 000 peptides per second. For this reason, the information fields that are returned by default are limited to the subset of available fields that can be returned without performance penalty.

Acknowledgements

The authors would like to thank the Bioinformatics Institute Ghent (BIG N2N) and the Flemish Supercomputer Center (VSC).

Funding

This work was supported by the Research Foundation Flanders (FWO) [G.0638.10].

Conflict of Interest: none declared.

References

Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

Bostock,M. *et al.* (2011) D^3 data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, **17**, 2301–2309.

Chen,C. *et al.* (2013) A fast Peptide Match service for UniProt Knowledgebase. *Bioinformatics*, **29**, 2808–2809.

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

Federhen,S. *et al.* (2014) Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Stand Genomic Sci.*, **9**, 1275.

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.

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

Mesuere,B. *et al.* (2012) Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. *J. Proteome Res.*, **11**, 5773–5780.

Mesuere,B. *et al.* (2015) The Unipept metaproteomics analysis pipeline. *Proteomics*, **15**, 1437–1442.

Olsen,J.V. *et al.* (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics*, **3**, 608–614.

Olson,M.A. *et al.* (1999) Berkeley DB. In: *USENIX Annual Technical Conference, FREENIX Track*, pp. 183–191.

UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.