Unipept Visualizations: an interactive visualization library for biological data

Pieter Verschaffelt[1, 2] https://orcid.org/0000-0002-6675-1048
James Collier[4] https://orcid.org/0000-0002-0020-421X
Alexander Botzki[4] https://orcid.org/0000-0001-6691-4233
Lennart Martens[2, 3] https://orcid.org/0000-0003-4277-658X
Peter Dawyndt[1] https://orcid.org/0000-0002-1623-9070
Bart Mesuere[1, 2] https://orcid.org/0000-0003-0610-3441

1) Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium
2) VIB - UGent Center for Medical Biotechnology, VIB, Ghent, Belgium
3) Department of Biomolecular Medicine, Ghent University, Ghent, Belgium
4) VIB Bioinformatics Core, VIB, Ghent, Belgium

Abstract
Summary The Unipept Visualizations library is a JavaScript package to generate interactive visualizations of both hierarchical and non-hierarchical quantitative data. It provides four different visualisations: a sunburst, a treemap, a treeview and a heatmap. Every visualization is fully configurable, supports TypeScript, and uses the excellent D3.js library.

Availability and implementation 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.
Contact: unipept@ugent.be
Supplementary information: https://github.com/unipept/unipept-visualizations/wiki

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 et al., 2021), a command line interface (CLI) (Verschaffelt et al., 2020) and an application programming interface (API). 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 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 visualise hierarchical quantitative data, and can therefore also be used to visualise data from non-proteomics origins. To facilitate reuse of our broadly usable components, we have isolated the visualisations 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 visualisations and connect them to their own data sources. Currently, our visualisations are already incorporated in TRAPID 2.0, a web application for the analysis of transcriptomes (Bucchini et al., 2020) and UMGAP, the Unipept MetaGenomics Pipeline (Van der Jeugt et al., 2021).

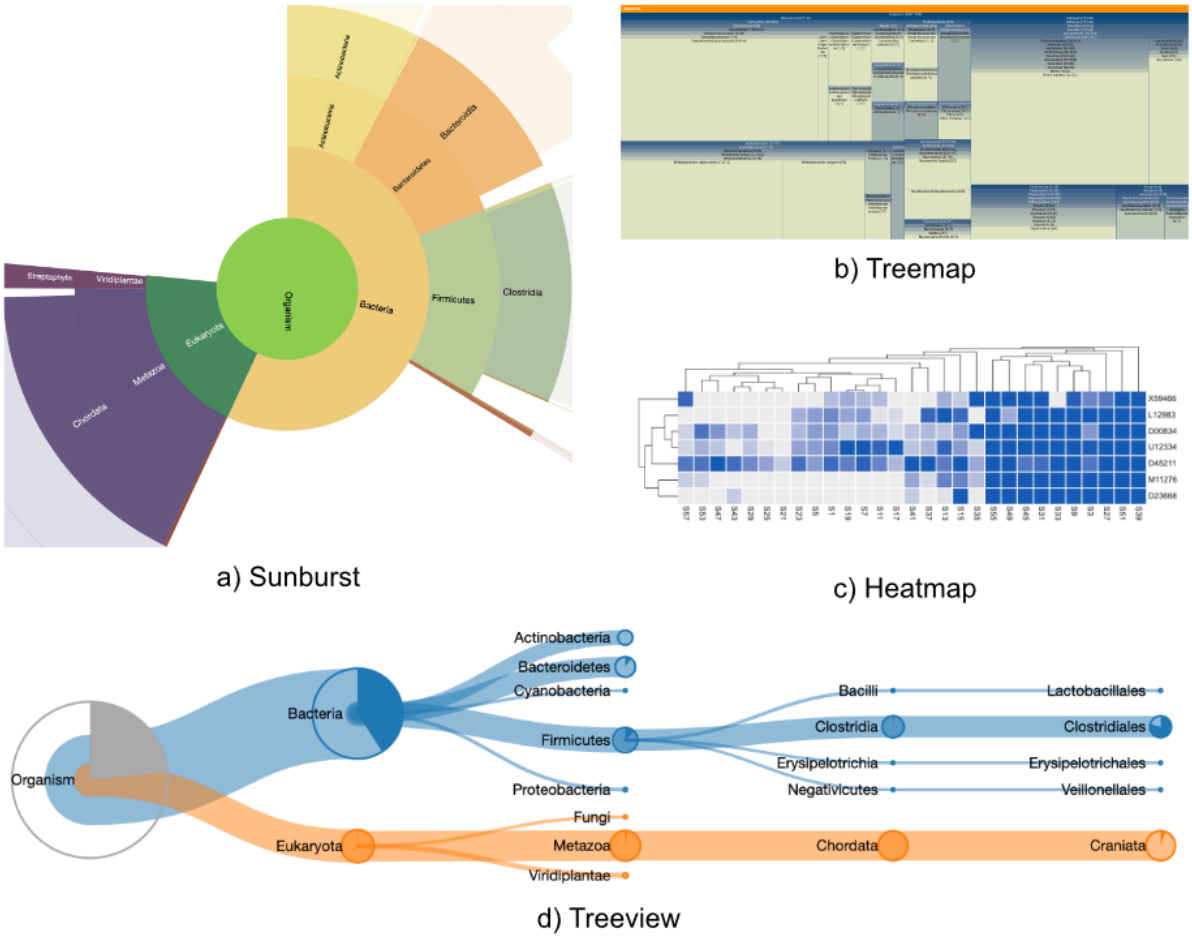

Figure 1: Overview of visualizations currently provided by the Unipept Visualizations library. All examples were generated with default configuration settings, except for the heatmap for which the setting `dendrogramEnabled` was set to `true`.

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 1a), treeview (Figure 1d) and treemap (Figure 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 1c), conversely, is not suitable to visualise hierarchical information but displays a magnitude in two dimensions, including optional clustering and dendrogram rendering.

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, 2018). In most cases, quantitative data is 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 is formatted correctly.

2.2 Heatmap

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

As neighbouring 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 et al., 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 (MOLO) 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 & Wainer, 1972).

3. Implementation

The visualization package has been developed with D3 (*D³ Data-Driven Documents*, n.d.) and TypeScript (Torgeson, 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/unipept/unipept-visualizations), so that any project depending on it 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.

References

Bar-Joseph, Z., Gifford, D. K., & Jaakkola, T. S. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* , *17 Suppl 1*, S22–S29.

Bucchini, F., Del Cortona, A., Kreft, Ł., Botzki, A., Van Bel, M., & Vandepoele, K. (2020). TRAPID 2.0: a web application for taxonomic and functional analysis of de novo transcriptomes. In *bioRxiv* (p. 2020.10.19.345835). https://doi.org/10.1101/2020.10.19.345835

D³ Data-Driven Documents. (n.d.). Retrieved April 30, 2021, from https://doi.org/10.1109/TVCG.2011.185

Gruvaeus, G., & Wainer, H. (1972). Two additions to hierarchical cluster analysis. *The British Journal of Mathematical and Statistical Psychology*, *25*(2), 200–206.

Gurdeep Singh, R., Tanca, A., Palomba, A., Van der Jeugt, F., Verschaffelt, P., Uzzau, S., Martens, L., Dawyndt, P., & Mesuere, B. (2019). Unipept 4.0: Functional Analysis of Metaproteome Data. *Journal of Proteome Research*, *18*(2), 606–615.

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in

a Web browser. *BMC Bioinformatics*, *12*, 385.

Sakai, R., Winand, R., Verbeiren, T., Moere, A. V., & Aerts, J. (2014). dendsort: modular leaf ordering methods for dendrogram representations in R. In *F1000Research* (Vol. 3, p. 177). https://doi.org/10.12688/f1000research.4784.1

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database: The Journal of Biological Databases and Curation*, *2020*. https://doi.org/10.1093/database/baaa062

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

The Gene Ontology Consortium. (2018). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338.

Torgeson, G. B. M. A. (2014). *Understanding TypeScript*. https://doi.org/10.1007/978-3-662-44202-9_11

Van der Jeugt, F., Maertens, R., Steyaert, A., Verschaffelt, P., De Tender, C., Dawyndt, P., & Mesuere, B. (2021). UMGAP: the Unipept MetaGenomics Analysis Pipeline. In *bioRxiv* (p. 2021.05.18.444604). https://doi.org/10.1101/2021.05.18.444604

Verschaffelt, P., Van Den Bossche, T., Martens, L., Dawyndt, P., & Mesuere, B. (2021). Unipept Desktop: A Faster, More Powerful Metaproteomics Results Analysis Tool. *Journal of Proteome Research*. https://doi.org/10.1021/acs.jproteome.0c00855

Verschaffelt, P., Van Thienen, P., Van Den Bossche, T., Van der Jeugt, F., De Tender, C., Martens, L., Dawyndt, P., & Mesuere, B. (2020). Unipept CLI 2.0: adding support for visualizations and functional annotations. *Bioinformatics* , *36*(14), 4220–4221.