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

Pieter Verschaffelt[1,2,*], Tim Van Den Bossche[2,3,*], Wassim Gabriel[4], Michał Burdukiewicz[5], Alessio Soggiu[6], Lennart Martens[2,3], Bernhard Y. Renard[7,8], Henning Schiebenhoefer[7,8,†,§], Bart Mesuere[1,2,†]

[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, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium, [4]Chair of Proteomics and Bioanalytics, Technical University of Munich, Freising, Germany, [5]Laboratory of Mass Spectrometry, Institute of Biochemistry and Biophysics Polish Academy of Sciences, Warsaw, Poland, [6]"One Health" Section, Department of Biomedical, Surgical and Dental Sciences, University of Milan, Milan, Italy, [7]Data Analytics and Computational Statistics, Hasso Plattner Institute for Digital Engineering, [8]Digital Engineering Faculty, University of Potsdam

** Shared first authorship, † Shared last authorship, § Corresponding author*

Abstract

The study of microbiomes has gained in importance over the past few years, and has led to 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 functional similarity between two data sets. MegaGO is highly performant: 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

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/.

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. While this approach allows a quite detailed investigation of the genomes that are present in such multi-organism samples, it only reveals their functional potential rather than their currently active functions[1]. To uncover these active functions within a given sample, the characterization of the protein content is often essential[2].

The growing focus on functional information as a complement to taxonomic information[3] is derived from the observation that two taxonomically similar microbial communities could have vastly different functional capacities, while taxonomically quite distinct communities could have remarkably similar functions. While 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[4].

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[5–14]. However, most tools are capable of directly or indirectly reporting functional annotations as a set of Gene Ontology[15] (GO) terms, regardless of the biological entity it is assigned to. In October 2020, there exist 44 264 of these terms in the complete GO tree. GO terms are organized in three independent domains: molecular function, biological process, and cellular component[16]. In each domain, terms are linked into a directed acyclic graph, an excerpt of which is shown in **Figure 1.** In the GO graph, a parent term can have one or more children (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 as parents GO:0043043, GO:0034645 and GO:0044267).

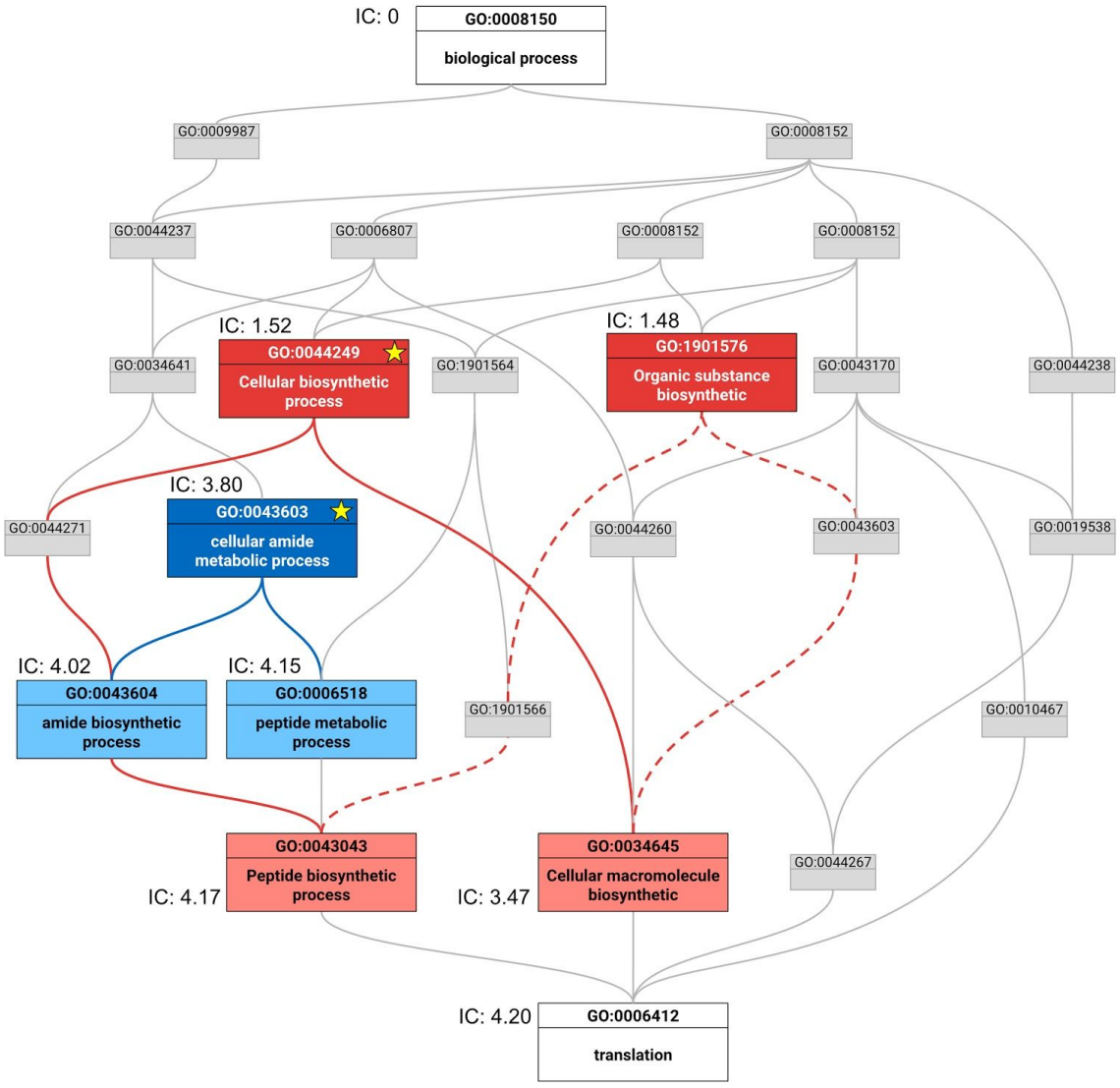


Figure 1. Excerpt of the biological process domain of the Gene Ontology showing all parent terms up to the root for “translation” (GO:0006412). The root GO term “biological process“ (GO:0008150) has multiple children. The most specific term “translation”, in contrast, has multiple parents. When comparing 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), i.e. the common ancestor with the highest information content for the terms in light red. Since 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 described earlier and only have one lowest common ancestor, which is also automatically the MICA for these terms: GO:0043603 (in dark blue). IC: information content, star: most informative common ancestor.

While 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 another data set. Indeed, even though two terms may be closely linked in the GO tree and are therefore highly similar (e.g., as parent and child terms, or as sibling terms), 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.[17], where a multi-sample 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 quite low. As explained above, 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[18–20]. 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 two data sets are.

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

Implementation

In order to measure the similarity of two 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[21], 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 occuring (**Supplementary Formula 1c**), including that of all of its children. Term frequencies are estimated based on the manually curated SwissProt database. 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 be more informative (**Figure 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 term’s individual information content, 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 1**).

MegaGO, however, can not only compare two terms but also two lists of GO terms. To achieve this, pairwise term similarities are aggregated using the Best Matching Average (BMA, **Supplementary Formula 2**)[22]. For each GO term in the first input data set, 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 2**.

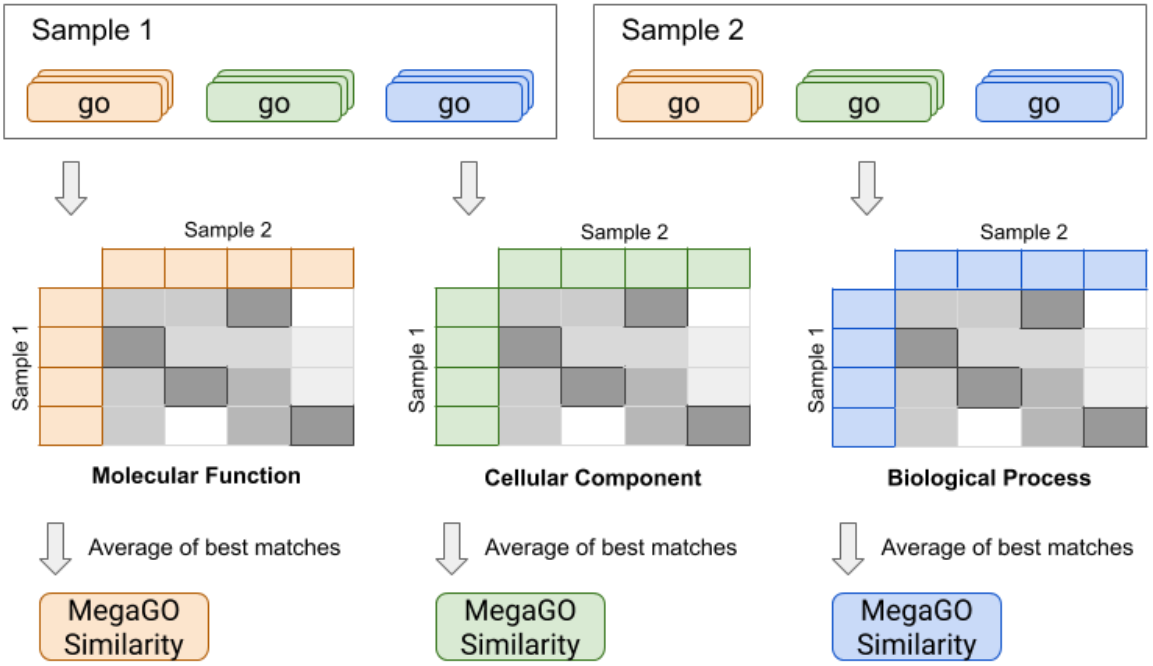


Figure 2. Overview of MegaGO workflow. Gene Ontology (GO) terms of each sample set are separated into three GO domains: molecular function, cellular component, and biological process. Each term of each sample set is compared to every term in the other set that is

from the same domain. The match with highest similarity for each term is then selected and the average across all these best matches is calculated.

MegaGO is implemented in Python and installable as a Python package from PyPi and can easily be invoked from the command line. The GOATOOLS[23] 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 bi-monthly 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, Opera), as well as Mozilla Firefox and Safari. The MegaGO application is also available as a Docker-container on Docker Hub 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 website: 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 or Linux systems. Systems requirements are at least 4GiB of memory and support for either Python 3.6 (or above), or the Docker runtime.

Validation

To validate MegaGO, we reprocessed the functional data from Easterly *et al.*[24]. 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. 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 a minute 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/.

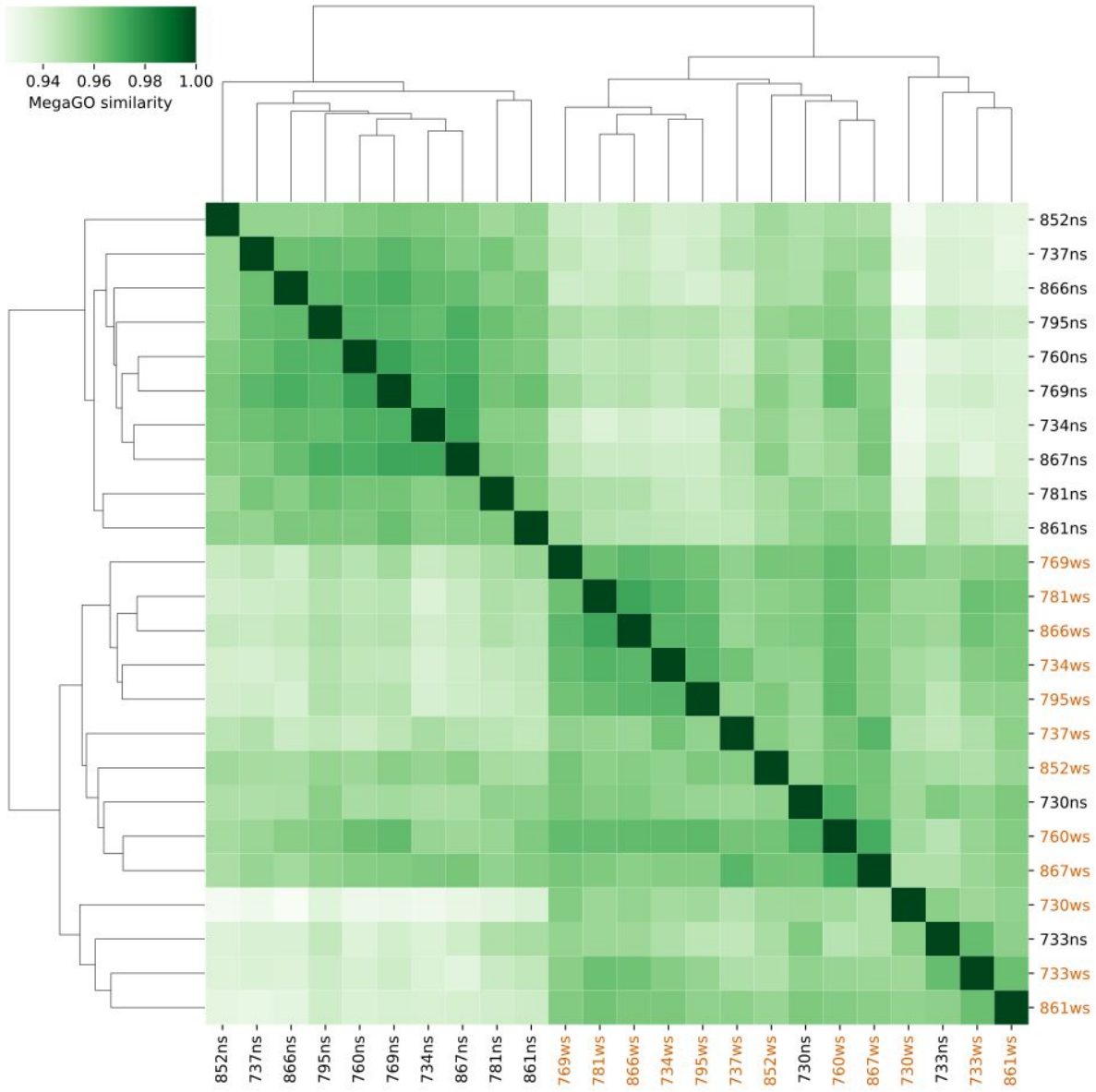


Figure 3. Hierarchically clustered heatmap comparing MegaGO similarities for the GO domain 'biological process' for each of the samples from Easterly *et al*.[24]. Samples that are treated with sucrose pulsing are labeled as "ws" and displayed in orange

In the heatmap (**Figure 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.*[24], and 733ns was originally also identified as both a taxonomic and functional outlier in Rudney *et al.*[25] Similar results can be observed for

the GO domain 'molecular function' (**Supplementary Figure 1**). 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 PCA of HOMINGS analysis by Rudney *et al.*, 861ns is closest to 730 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.

Conclusion

MegaGO enables the comparison of large sets of GO-terms, allowing users to efficiently evaluate multi-omics data sets containing thousands of terms. MegaGO calculates a similarity for each of the three GO-domains separately (biological process, molecular function, and cellular component). In the current version of MegaGO, quantitative data is 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, or protein based methods as well as combinations thereof.

Acknowledgements

We would like to acknowledge the European Bioinformatics Community for Mass Spectrometry (EuBIC-MS). This project found its origin at the EuBIC Developers' 2020 meeting in Nyborg, Denmark. We would like to thank Thilo Muth and Stephan Fuchs for their support. PV, TVDB, LM and BM would like to acknowledge Research Foundation - Flanders (FWO) [grants 1164420N, 1S90918N, G042518N and 12I5220N]. LM also acknowledges support from the European Union's Horizon 2020 Programme under Grant Agreement 823839 [H2020-INFRAIA-2018-1]. HS and BYR acknowledge support by Deutsche Forschungsgemeinschaft (DFG; grant number 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).

References

(1) Jansson, J. K.; Baker, E. S. A Multi-Omic Future for Microbiome Studies. *Nat Microbiol* **2016**, *1* (5), 507.
(2) Lohmann, P.; Schäpe, S. S.; Haange, S.-B.; Oliphant, K.; Allen-Vercoe, E.; Jehmlich, N.; Von Bergen, M. Function Is What Counts: How Microbial Community Complexity Affects Species, Proteome and Pathway Coverage in Metaproteomics. *Expert Rev. Proteomics* **2020**, *17* (2), 163–173.
(3) Louca, S.; Parfrey, L. W.; Doebeli, M. Decoupling Function and Taxonomy in the Global Ocean Microbiome. *Science* **2016**, *353* (6305), 1272–1277.
(4) Schiebenhoefer, H.; Van Den Bossche, T.; Fuchs, S.; Renard, B. Y.; Muth, T.; Martens, L. Challenges and Promise at the Interface of Metaproteomics and Genomics: An Overview of Recent Progress in Metaproteogenomic Data Analysis. *Expert Rev. Proteomics* **2019**, *16* (5), 375–390.
(5) Muth, T.; Behne, A.; Heyer, R.; Kohrs, F.; Benndorf, D.; Hoffmann, M.; Lehtevä, M.; Reichl, U.; Martens, L.; Rapp, E. The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation. *J. Proteome Res.* **2015**, *14* (3), 1557–1565.
(6) Muth, T.; Kohrs, F.; Heyer, R.; Benndorf, D.; Rapp, E.; Reichl, U.; Martens, L.; Renard, B. Y. MPA Portable: A Stand-Alone Software Package for Analyzing Metaproteome Samples on the Go, 2018. https://doi.org/10.1021/acs.analchem.7b03544.
(7) Van Den Bossche, T.; Verschaffelt, P.; Schallert, K.; Barsnes, H.; Dawyndt, P.; Benndorf, D.; Renard, B. Y.; Mesuere, B.; Martens, L.; Muth, T. Connecting MetaProteomeAnalyzer and PeptideShaker to Unipept for Seamless End-to-End Metaproteomics Data Analysis, 2020. https://doi.org/10.1021/acs.jproteome.0c00136.
(8) Verschaffelt, P.; Van Thienen, P.; Van Den Bossche, T.; Van der Jeugt, F.; De Tender, C.; Martens, L.; Dawyndt, P.; Mesuere, B. Unipept CLI 2.0: Adding Support for Visualizations and Functional Annotations. *Bioinformatics* **2020**, *25*, 25.
(9) Gurdeep Singh, R.; Tanca, A.; Palomba, A.; Van der Jeugt, F.; Verschaffelt, P.; Uzzau, S.; Martens, L.; Dawyndt, P.; Mesuere, B. Unipept 4.0: Functional Analysis of Metaproteome Data, 2019. https://doi.org/10.1021/acs.jproteome.8b00716.
(10) Riffle, M.; May, D.; Timmins-Schiffman, E.; Mikan, M.; Jaschob, D.; Noble, W.; Nunn, B. MetaGOmics: A Web-Based Tool for Peptide-Centric Functional and Taxonomic Analysis of Metaproteomics Data. *Proteomes* **2017**, *6* (1), 2.
(11) Schneider, T.; Schmid, E.; de Castro, J. V., Jr.; Cardinale, M.; Eberl, L.; Grube, M.; Berg, G.; Riedel, K. Structure and Function of the Symbiosis Partners of the Lung Lichen (Lobaria Pulmonaria L. Hoffm.) Analyzed by Metaproteomics. *Proteomics* **2011**, *11* (13), 2752–2756.
(12) Schiebenhoefer, H.; Schallert, K.; Renard, B. Y.; Trappe, K.; Schmid, E.; Benndorf, D.; Riedel, K.; Muth, T.; Fuchs, S. A Complete and Flexible Workflow for Metaproteomics Data Analysis Based on MetaProteomeAnalyzer and Prophane. *Nat. Protoc.* **2020**, *15* (10), 3212–3239.
(13) Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S. K.; Cook, H.; Mende, D. R.; Letunic, I.; Rattei, T.; Jensen, L. J.; von Mering, C.; Bork, P. eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology

Resource Based on 5090 Organisms and 2502 Viruses. *Nucleic Acids Res.* **2019**, *47* (D1), D309–D314.

(14) Huson, D. H.; Auch, A. F.; Qi, J.; Schuster, S. C. MEGAN Analysis of Metagenomic Data. *Genome Res.* **2007**, *17* (3), 377–386.

(15) The Gene Ontology Consortium. The Gene Ontology Resource: 20 Years and Still GOing Strong. *Nucleic Acids Res.* **2019**, *47* (D1), D330–D338.

(16) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000**, *25* (1), 25–29.

(17) Sajulga, R.; Easterly, C.; Riffle, M.; Mesuere, B.; Muth, T.; Mehta, S.; Kumar, P.; Johnson, J.; Gruening, B.; Schiebenhoefer, H.; Kolmeder, C. A.; Fuchs, S.; Nunn, B. L.; Rudney, J.; Griffin, T. J.; Jagtap, P. D. Survey of Metaproteomics Software Tools for Functional Microbiome Analysis. *PLOS ONE.* **2020**.

(18) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* **2009**, *4* (1), 44–57.

(19) Waardenberg, A. J.; Bassett, S. D.; Bouveret, R.; Harvey, R. P. CompGO: An R Package for Comparing and Visualizing Gene Ontology Enrichment Differences between DNA Binding Experiments. *BMC Bioinformatics* **2015**, *16* (1), 25.

(20) Fruzangohar, M.; Ebrahimie, E.; Ogunniyi, A. D.; Mahdi, L. K.; Paton, J. C.; Adelson, D. L. Comparative GO: A Web Application for Comparative Gene Ontology and Gene Ontology-Based Gene Selection in Bacteria, 2013. https://doi.org/10.1371/journal.pone.0058759.

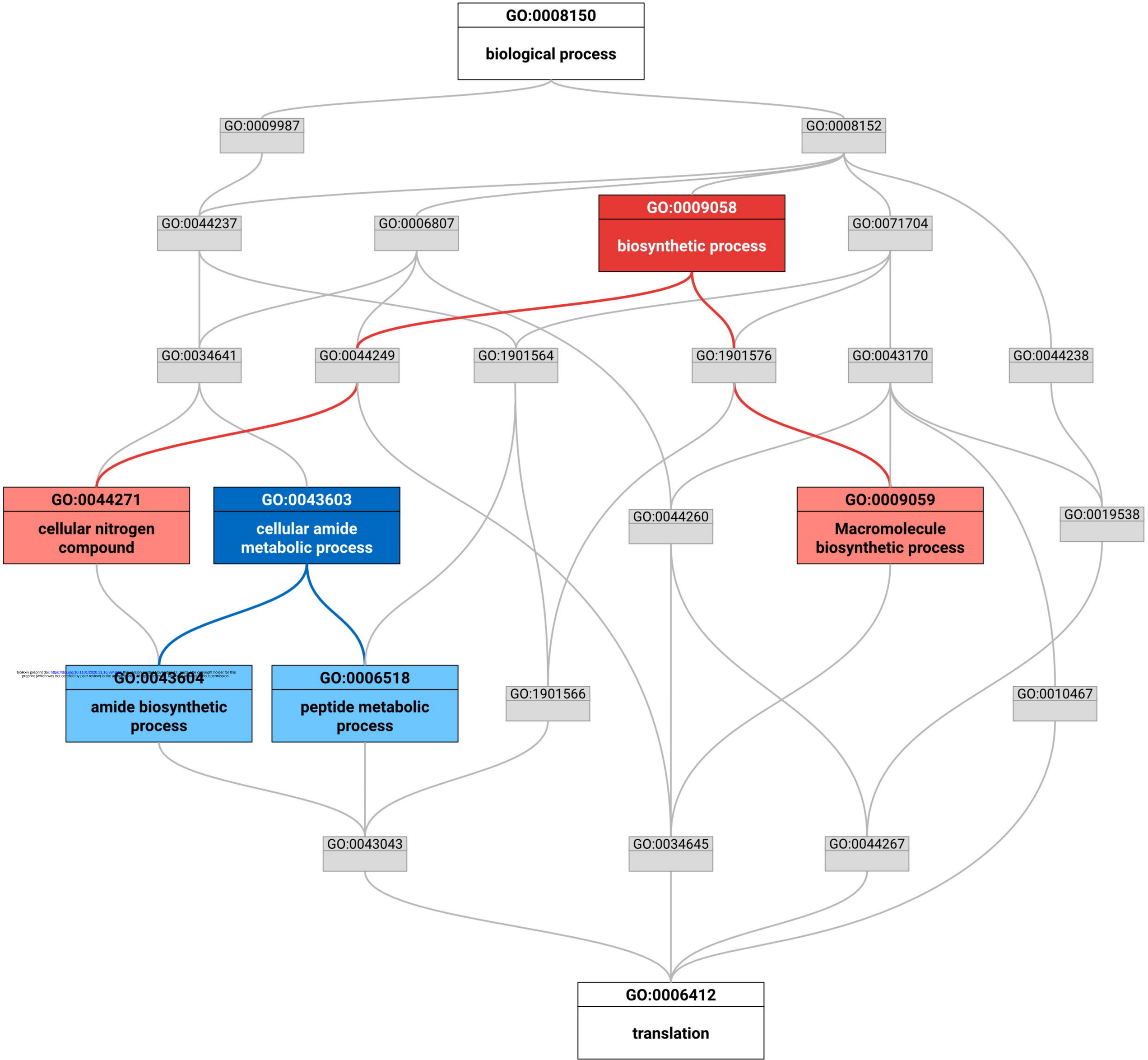
(21) Lin, D. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning*; 1998; Vol. 98, pp 296–304.

(22) Schlicker, A.; Domingues, F. S.; Rahnenführer, J.; Lengauer, T. A New Measure for Functional Similarity of Gene Products Based on Gene Ontology. *BMC Bioinformatics* **2006**, *7* (1), 493.

(23) Klopfenstein, D. V.; Zhang, L.; Pedersen, B. S.; Ramírez, F.; Warwick Vesztrocy, A.; Naldi, A.; Mungall, C. J.; Yunes, J. M.; Botvinnik, O.; Weigel, M.; Dampier, W.; Dessimoz, C.; Flick, P.; Tang, H. GOATOOLS: A Python Library for Gene Ontology Analyses. *Sci. Rep.* **2018**, *8* (1), 10872.

(24) Easterly, C. W.; Sajulga, R.; Mehta, S.; Johnson, J.; Kumar, P.; Hubler, S.; Mesuere, B.; Rudney, J.; Griffin, T. J.; Jagtap, P. D. metaQuantome: An Integrated, Quantitative Metaproteomics Approach Reveals Connections Between Taxonomy and Protein Function in Complex Microbiomes. *Mol. Cell. Proteomics* **2019**, *18* (8 suppl 1), S82–S91.

(25) Rudney, J. D.; Jagtap, P. D.; Reilly, C. S.; Chen, R.; Markowski, T. W.; Higgins, L.; Johnson, J. E.; Griffin, T. J. Protein Relative Abundance Patterns Associated with Sucrose-Induced Dysbiosis Are Conserved across Taxonomically Diverse Oral Microcosm Biofilm Models of Dental Caries. *Microbiome* **2015**, *3* (1), 89.



Sample 1

go go go

Sample 2

go go go

Sample 2

Sample 1

Molecular Function

Average of best matches

MegaGO Similarity

Sample 2

Sample 1

Cellular Component

Average of best matches

MegaGO Similarity

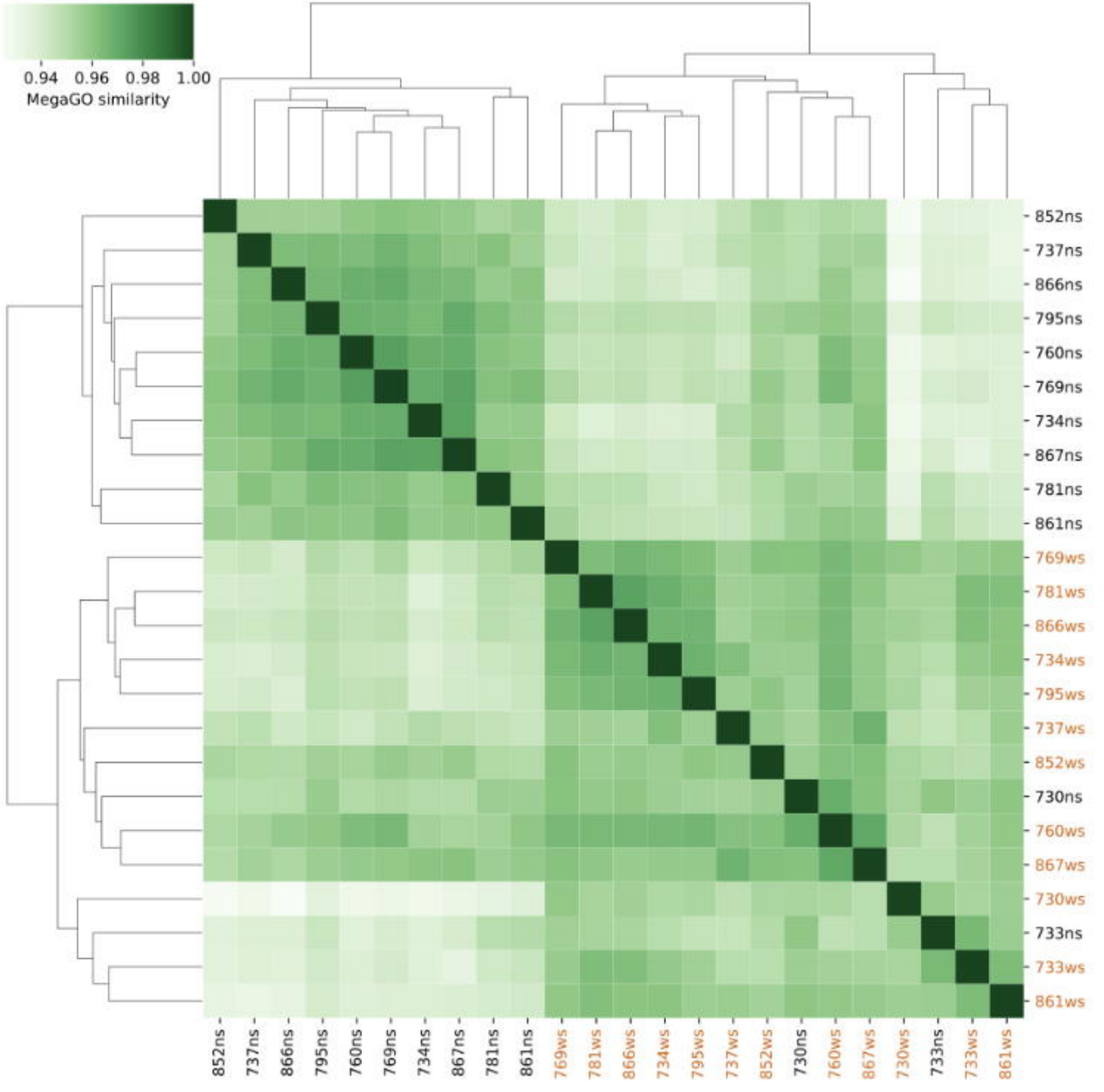
Sample 2

Sample 1

Biological Process

Average of best matches

MegaGO Similarity


0.94 0.96 0.98 1.00
MegaGO similarity
852ns
737ns
866ns
795ns
760ns
769ns
734ns
867ns
781ns
861ns
769ws
781ws
866ws
734ws
795ws
737ws
852ws
730ns
760ws
867ws
730ws
733ns
733ws
861ws