
Caleb Easterly1, Carolina Kolmender2, Thilo Muth3, Bart Mesure4, Subina Mehta4, Praveen Kumar4, James Johnson5, Shane Hulst6, Jaime Huerta-Cepas6, Bjoern Gruening7, Michael Riffel8, Damon May9, W. Judson Hervey10, Alessandro Tanca11, Brook L. Nunn12, Joel Nudel13, Timothy J. Griffin14, & Pratik D. Jagtap15

1Biochem, Mol Biol, Biophysics, University of Minnesota, MN; 2University of Helsinki, Helsinki, Finland; 3Robert Koch Institute, Berlin, Germany; 4Ghost University, Ghent, Belgium; 5Minnesota Supercomputing Institute, UMN, Minneapolis, MN; 6Eukaryotic Molecular Biology Laboratory, Heidelberg, Germany; 7University of Freiburg, Freiburg, Germany; 8University of Washington Genome Sciences, Seattle, WA; 9Naval Research Laboratory, Washington DC; 10Porto Conte Ricerche Science and Technology Park of Sardinia, Alghero, Italy

**Introduction**

Metaproteomics research involves large-scale characterization of the entire protein complement of the microbiome. Metaproteomics has the potential to unveil the mechanistic details of microbial interactions with the host/environment by analyzing microbial proteins under minimal enrichment. Many methods have been developed to determine the functional role of proteins expressed by the microbiome. However, the information on the functional content of the available microbiome proteomes often differs in emphasis, features, reproducibility, and other characteristics. Here, we evaluated the functional content of published metaproteomic data using various software tools to determine if any differences were due to more specific terms. To test this hypothesis, 24 different metaproteomic data sets were compared with each other using different tools. When mapped to the GO generic slim, the differences were reduced, suggesting that more specific terms drive the majority of the difference.

**Materials and Methods**

We analyzed the data with several functional analysis tools (see tool list in Introduction), using the standardized procedures for each tool and the input files in accordance with the Tool Features table. The inputs were compared using several methods:

- raw spectral data (WS over NS) were calculated based on spectral counts and ranked for the input proteins.
- The Jaccard index was defined as the size of the intersection divided by the size of the union. When mapped to the GO generic slim, the Jaccard index was defined as 0.693.

**Data**

- Mass spectral data (Rubino et al., 2015; PRIDE0030131) were acquired from plaque sampled from patient at high risk for denture caries and grown in biofilm reactor in the presence and absence of sucrose (WS and NS, respectively).
- Previous functional analysis of the data showed sucrose-induced changes in protein relative abundances that have biological relevance.
- Mass spectra were searched against the Human Oral Microbiome Database (HOMD) to obtain peptide sequences, and spectral counts were calculated for each peptide.

**Results**

- Comparison of number of identifications obtained by each tool: the "exclusive" column indicates the number of peptides identified by that tool only and the "shared" column gives the overlap with all other tools. The number of hits and the GO terms obtained for each tool are tabulated in the table below.

**Discussion**

- Different tools show very different results with the same data.
- When mapped to the GO slim, the differences were reduced, suggesting that more specific terms drive the majority of the difference.

**Conclusion**

- The tools showed a high level of similarity with each other and are not to be construed as those of the authors and are not to be construed as those of the authors.

**Acknowledgements**

This project is supported by National Science Foundation (NSF) grant 1458524 and National Institutes of Health (NIH) grant R01 AI136419. We thank Biocrates Life Sciences for providing software tools and the Bioinformatics Core Facility at the University of Minnesota Supercomputing Institute for providing essential infrastructure.