pymgpipe: microbiome metabolic modeling in Python

Yoli Meydan ⓘ [1*], **Federico Baldini** ⓘ [1*], **and Tal Korem** ⓘ [1,2¶]

1 Program for Mathematical Genomics, Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA. **2** Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY, USA. ¶ Corresponding author * These authors contributed equally.

Introduction

Microbially-produced metabolites and microbiome metabolism in general are strongly linked to ecosystem-level phenotypes, including the health of the human host (Bar et al., 2020; Villanueva-Millán et al., 2015). To aid in the study of microbial metabolism from observational, human-derived data, a variety of computational methods that predict microbial community metabolic output from microbial abundances have been developed (Baldini et al., 2019; Diener et al., 2020; Mallick et al., 2019; Noecker et al., 2022). Several of these methods rely on community-scale metabolic models, which are mechanistic, knowledge-based models that enable the formulation and *in silico* testing of biological hypotheses regarding the metabolism of microbial communities (Baldini et al., 2019; Diener et al., 2020). Community-scale models primarily use Flux Balance Analysis, a technique that infers the metabolic fluxes in a system by optimizing an objective function, typically growth rate, subject to an assumption of a steady state and constraints imposed by the metabolic reactions present in the system (Orth et al., 2010). These metabolic reactions are obtained from genome-scale metabolic networks (GEMs), knowledge-based computational models encompassing the known biochemical reactions present within an organism (Thiele & Palsson, 2010). In recent years, curated GEMs for thousands of human-associated microbial organisms have become increasingly available, enabling a more in-depth exploration of the human microbiome (Heinken et al., 2023; Machado et al., 2018; Norsigian et al., 2020). In addition, several community-scale metabolic modeling methods specifically tailored to the human microbiome have emerged, such as MICOM and mgPipe (Baldini et al., 2019; Diener et al., 2020).

Statement of need

mgPipe is a method that combines individual GEMs into a shared compartment according to the microbial abundances observed in every sample to construct a community-level metabolic model. Input and output compartments are added to allow for a distinction between the uptake and secretion of metabolites by the community. After constructing a representative model for each sample, mgPipe computes the metabolic capacity for all present metabolites in the form of Net Maximal Production Capacities (NMPCs). NMPCs are calculated as the absolute difference between the maximum secretion through the output compartment and the maximal uptake through the input compartment (Baldini et al., 2019). To accomplish this, Flux Variability Analysis (FVA) (Mahadevan & Schilling, 2003) is used to compute reaction bounds (minimum and maximum fluxes) through metabolite exchange reactions.

mgPipe models can further be used to explore metabolic interactions among individual taxa, the contribution of these taxa to the overall community metabolism, and to raise hypotheses regarding the biochemical machinery underlying an observed phenotype. This utility of mgPipe has been demonstrated in various studies of the role of the human microbiome in complex

conditions such as preterm birth, inflammatory bowel disease, colorectal cancer, and Parkinson's disease (Baldini et al., 2020; Heinken et al., 2019; Hertel et al., 2019, 2021; Kindschuh et al., 2023). However, and despite its wide use and utility, only a MATLAB implementation of mgPipe is currently available, limiting its accessibility for those who are not proficient in MATLAB or cannot afford its license. Here, we provide a reliable, tested, open-source, and efficient Python implementation of mgPipe.

Implementation & Availability

pymgpipe is a Python implementation of mgPipe (Baldini et al., 2019). It utilizes COBRApy (Ebrahim et al., 2013) as its main constraint-based metabolic modeling interface, and optlang (Jensen et al., 2017) to formulate and modify the underlying mathematical optimization problem. pymgpipe merges individual GEMs into a single model following mgPipe's biologically-informed metabolic assumptions, such as the use of preordained diets, compartmentalized structure, abundance-scaled constraints on microbial flux contributions (Heinken et al., 2013), and community biomass optimization objective (Baldini et al., 2019). After building community-level models, metabolic profiles are computed in the form of NMPCs, as discussed above (Baldini et al., 2019). As part of this step, pymgpipe uses the VFFVA C package for a fast and efficient FVA implementation (Guebila, 2020). pymgpipe is compatible with both the Gurobi (Gurobi Optimization, LLC, 2023) and ILOG Cplex (IBM, Inc., 2023) solvers, which are both commercially available and free for academic use.

pymgpipe models are backwards-compatible with the MATLAB mgPipe models to ensure cross-software compatibility. Additionally, pymgpipe offers multithreading capabilities for both model construction and simulation, making it scalable to studies with a large sample size. The pymgpipe python package, as well as all associated documentation, tests, and example workflows, can be found at https://github.com/korem-lab/pymgpipe.
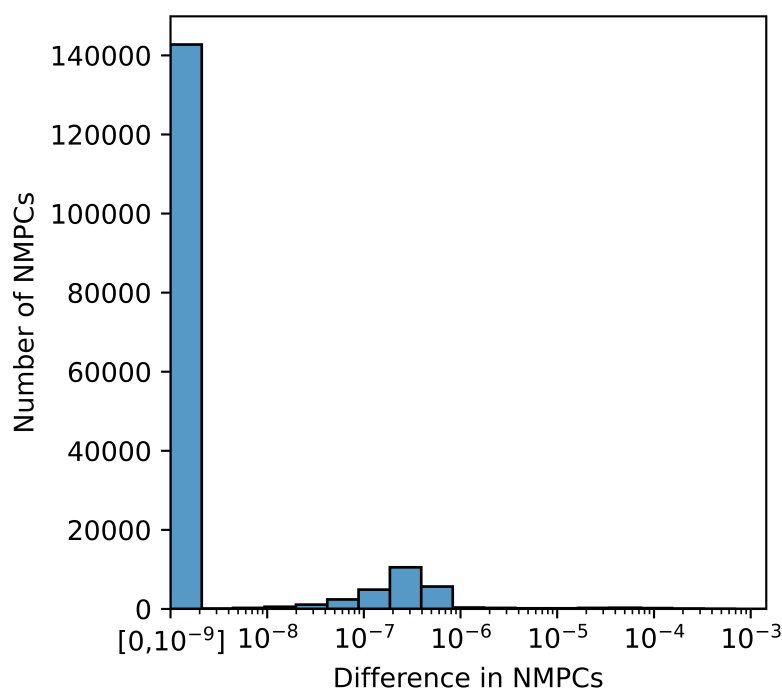
Comparison to mgPipe

Figure 1: Histogram of magnitude of differences in NMPCs between mgPipe and pymgpipe.

To assess the accuracy of pymgpipe we compared its models and predictions with mg-Pipe, as implemented in the Microbiome Modeling Toolbox, Cobra Toolbox commit: 71c117305231f77a0292856e292b95ab32040711 (Baldini et al., 2019). We generated community-scale models for a vaginal microbiome dataset consisting of 232 samples, each composed of between 2 to 50 taxa (94 unique taxa), as previously described (Kindschuh et al., 2023). The models exhibited identical metabolic networks and structure between the two implementations (not shown). Additionally, metabolic profiles (NMPCs) output by pymgpipe exhibited only minor differences (mean±sd. 5.37e-7±1.23e-5; difference is below 1e-5 for 99.4% of all data points; Figure 1). These differences are negligible (within solver tolerance) and are most likely due to variations in FVA implementations (Guebila, 2020), solver versions, and tolerances. Overall, pymgpipe presents as an accurate Python implementation of the mgPipe pipeline.

Acknowledgments

References

Baldini, F., Heinken, A., Heirendt, L., Magnusdottir, S., Fleming, R. M., & Thiele, I. (2019). The microbiome modeling toolbox: From microbial interactions to personalized microbial

communities. *Bioinformatics*, *35*(13), 2332–2334. https://doi.org/10.1093/bioinformatics/bty941

Baldini, F., Hertel, J., Sandt, E., Thinnes, C. C., Neuberger-Castillo, L., Pavelka, L., Betsou, F., Krüger, R., & Thiele, I. (2020). Parkinson's disease-associated alterations of the gut microbiome predict disease-relevant changes in metabolic functions. *BMC Biology*, *18*, 1–21. https://doi.org/10.1186/s12915-020-00775-7

Bar, N., Korem, T., Weissbrod, O., Zeevi, D., Rothschild, D., Leviatan, S., Kosower, N., Lotan-Pompan, M., Weinberger, A., Le Roy, C. I., & others. (2020). A reference map of potential determinants for the human serum metabolome. *Nature*, *588*(7836), 135–140. https://doi.org/10.1038/s41586-020-2896-2

Diener, C., Gibbons, S. M., & Resendis-Antonio, O. (2020). MICOM: Metagenome-scale modeling to infer metabolic interactions in the gut microbiota. *MSystems*, *5*(1), e00606–19. https://doi.org/10.1128/mSystems.00606-19

Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy: Constraints-based reconstruction and analysis for python. *BMC Systems Biology*, *7*, 1–6. https://doi.org/10.1186/1752-0509-7-74

Guebila, M. B. (2020). VFFVA: Dynamic load balancing enables large-scale flux variability analysis. *BMC Bioinformatics*, *21*, 1–13. https://doi.org/10.1186/s12859-020-03711-2

Gurobi Optimization, LLC. (2023). *Gurobi Optimizer Reference Manual*. https://www.gurobi.com

Heinken, A., Hertel, J., Acharya, G., Ravcheev, D. A., Nyga, M., Okpala, O. E., Hogan, M., Magnúsdóttir, S., Martinelli, F., Nap, B., & others. (2023). Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nature Biotechnology*, 1–12. https://doi.org/10.1038/s41587-022-01628-0

Heinken, A., Ravcheev, D. A., Baldini, F., Heirendt, L., Fleming, R. M., & Thiele, I. (2019). Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. *Microbiome*, *7*, 1–18. https://doi.org/10.1186/s40168-019-0689-3

Heinken, A., Sahoo, S., Fleming, R. M., & Thiele, I. (2013). Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut Microbes*, *4*(1), 28–40. https://doi.org/10.4161/gmic.22370

Hertel, J., Harms, A. C., Heinken, A., Baldini, F., Thinnes, C. C., Glaab, E., Vasco, D. A., Pietzner, M., Stewart, I. D., Wareham, N. J., & others. (2019). Integrated analyses of microbiome and longitudinal metabolome data reveal microbial-host interactions on sulfur metabolism in parkinson's disease. *Cell Reports*, *29*(7), 1767–1777. https://doi.org/10.2139/ssrn.3305554

Hertel, J., Heinken, A., Martinelli, F., & Thiele, I. (2021). Integration of constraint-based modeling with fecal metabolomics reveals large deleterious effects of fusobacterium spp. On community butyrate production. *Gut Microbes*, *13*(1), 1915673. https://doi.org/10.1080/19490976.2021.1915673

IBM, Inc. (2023). *ILOG cplex optimization studio*. https://www.ibm.com/products/ilog-cplex-optimization-studio/cplex-optimizer

Jensen, K., Cardoso, J. G., & Sonnenschein, N. (2017). Optlang: An algebraic modeling language for mathematical optimization. *J. Open Source Softw.*, *2*(9), 139. https://doi.org/10.21105/joss.00139

Kindschuh, W. F., Baldini, F., Liu, M. C., Liao, J., Meydan, Y., Lee, H. H., Heinken, A., Thiele, I., Thaiss, C. A., Levy, M., & others. (2023). Preterm birth is associated

with xenobiotics and predicted by the vaginal metabolome. *Nature Microbiology*, 1–14. https://doi.org/10.1038/s41564-022-01293-8

Machado, D., Andrejev, S., Tramontano, M., & Patil, K. R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*, *46*(15), 7542–7553. https://doi.org/10.1093/nar/gky537

Mahadevan, R., & Schilling, C. H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, *5*(4), 264–276. https://doi.org/10.1016/j.ymben.2003.09.002

Mallick, H., Franzosa, E. A., McIver, L. J., Banerjee, S., Sirota-Madi, A., Kostic, A. D., Clish, C. B., Vlamakis, H., Xavier, R. J., & Huttenhower, C. (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nature Communications*, *10*(1), 3136. https://doi.org/10.1038/s41467-019-10927-1

Noecker, C., Eng, A., Muller, E., & Borenstein, E. (2022). MIMOSA2: A metabolic network-based tool for inferring mechanism-supported relationships in microbiome-metabolome data. *Bioinformatics*, *38*(6), 1615–1623. https://doi.org/10.1093/bioinformatics/btac003

Norsigian, C. J., Pusarla, N., McConn, J. L., Yurkovich, J. T., Dräger, A., Palsson, B. O., & King, Z. (2020). BiGG models 2020: Multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Research*, *48*(D1), D402–D406. https://doi.org/10.1093/nar/gkz1054

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–248. https://doi.org/10.1038/nbt.1614

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), 93–121. https://doi.org/10.1038/nprot.2009.203

Villanueva-Millán, M., Pérez-Matute, P., & Oteo, J. (2015). Gut microbiota: A key player in health and disease. A review focused on obesity. *Journal of Physiology and Biochemistry*, *71*, 509–525. https://doi.org/10.1007/s13105-015-0390-3