Title: Scalable Computational Tools For Inference Of Protein Annotation And Metabolic Models In Microbial Communities

Authors: Janaka N. Edirisinghe[1*] (janakae@anl.gov), Mikayla A. Borton[2*] (mikayla.borton@pnnl.gov), Michael Shaffer[3], Zahmeeth Sakkaff[1], Filipe Liu[1], Rory M. Flynn[3], Saeedeh Davoudi[4], Lucia S. Guatney[4,5], Derick Singleton[4], Tobby Lie[4], James C. Stegen[2], Byron C. Crump[6], Jim Davis[1], Farnoush Banaei-Kashani[4], Kelly C. Wrighton[3], Christopher Henry[1], and **Christopher S. Miller**[4]

Institutions: [1] Argonne National Laboratory, Argonne, IL; [2] Pacific Northwest National Laboratory, Richland, WA; [3] Colorado State University, Ft. Collins, CO; [4] University of Colorado Denver, Denver, CO; [5] University of Colorado Anschutz Medical Campus, Aurora, CO; and [6] Oregon State University, Corvallis, Oregon.

Website URLs: https://www.kbase.us ; https://github.com/WrightonLabCSU/DRAM

Project Goals: High-throughput omics technologies have made the assembly of microbial genomes recovered from the environment routine. Computational inference of the protein products encoded by these genomes, and the associated biochemical functions, should allow for the accurate prediction and modeling of microbial metabolism, organismal interactions, and ecosystem processes. However, a lack of scalable, probabilistic protein annotation tools limits the full potential of metabolic modeling. Our approach to inference of improved models relies on developing new computational tools in three main areas: 1) improved protein annotations, 2) iterative cycles of gap-filling metabolic models with improved protein annotations and informing probabilistic protein annotations based on metabolic models, and 3) integrating improved protein annotations with community-level flux balance metabolic models. We aim to make these tools broadly accessible via the DOE Systems Biology Knowledgebase (KBase)[1].

Abstract Text: To improve the inference of protein annotations, we have made improvements over the past year to the comprehensive annotation pipeline, Distilled and Refined Annotation of Metabolism (DRAM)[2]. We have significantly improved DRAM to incorporate new functionality, bug fixes, and speed gains, to improve the speed and sensitivity of viral annotations, to allow for custom HMM sets, and to expand the metabolic repertoire (e.g. polymeric carbons / polyphenols, carbohydrates, organic nitrogen / methylamines, and bile salts). We have also improved annotation precision by including gene-customized cutoffs for specific functional genes that are problematically annotated using homology, as well as visual validation with pre-constructed phylogenetic trees. In summary, DRAM now provides faster, phylogenetically informed community profiles and genome annotations.

We have also improved the model reconstruction pipelines in KBase to directly utilize improved annotations from the DRAM KBase app[3]. Metabolic models are now capable of representing many ecologically important metabolisms, including methanogenesis, nitrogen cycling, and sulfur reduction. DRAM annotations of viral genomes are now in the KBase infrastructure and integrated to automatically ingest outputs from other virus-specific applications on KBase (e.g.VirSorter). Ultimately, our improvements to DRAM alongside tight KBase integration will lead to more accurate community metabolic models of microbiome systems.

We are also evaluating these improved annotations and model reconstructions with experimental data. We computationally predicted growth of 350 genomes on 60 carbon sources, comparing the predicted phenotypes with experimentally observed data from Biolog studies. Although DRAM and RAST individually performed with similar accuracy, RAST displayed more false negatives while DRAM displayed more false positives in comparison with experimental data. However, a gap-filling approach *combining* RAST and DRAM annotations boosted accuracy dramatically, while also identifying gene candidates for many gapfilled reactions. This gapfilling approach requires *a priori* knowledge of phenotypes to work; to provide this knowledge for a much broader set of genomes (including MAGs), we also developed machine learning classifiers to predict phenotypes based on annotated gene functions genome-wide.

Now that we have an established cyberinfrastructure with integration of high-quality annotations and models, we can next evaluate the scalability and performance using more complex datasets. As a use case, we demonstrate these improvements to genome annotation and modeling using Metagenome Assembled Genomes (MAGs) from the Genome Resolved Open Watersheds database (GROWdb). This growing resource samples a large number (currently 250) of river microbiomes worldwide, and offers a metadata-rich, complex, real-world dataset with which to evaluate the utility and scalability of our annotation and modeling efforts in KBase. We have ingested a snapshot of 163 samples from US surface waters into KBase, including genome-resolved metagenomics, extensive geospatial metadata, metatranscriptomics, and metabolomics (FT-ICR). To inform metabolic modeling, we used DRAM to genomically inventory the 2,093 nonredundant surface-water-derived MAGs for complete energy biosynthesis systems and analyze these results organized by phylogeny and geospatial metadata.

Finally, over the last year, we have focused on developing improved representations for biological sequence data by adapting universal language models from the Natural Language Processing (NLP) community for two specific applications: protein annotation and taxonomic binning. We are currently extending the ProteinBERT model[4] to 1) use the primary protein sequences to derive representations for the sequences in an unsupervised manner, and 2) utilize Enzyme Commission (E.C.) annotations for supervised training of the model to capture global representations of function. These models can be extended with other global features, such as E.C. numbers of adjacent genes, or inferred 3D protein structure classes. Encouraged by the improvements to metabolic modeling we have shown above when using multiple annotation sources, we will add an additional, orthogonal NLP-based annotation using these approaches.

References:

1. Arkin AP, Cottingham RW, Henry CS, et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol*. 2018;36:566–9.
2. Shaffer M, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res*. 2020;48:8883–900.
3. Shaffer M, et al. kb_DRAM: Distilled genome annotations and metabolic modeling in KBase. *submitted*
4. Brandes N, et al. ProteinBERT: A universal deep-learning model of protein sequence and function. *bioRxiv.* 2021; 2021.05.24.445464.

Funding Statement: This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), grant no. DE-SC0021350, and the DOE Joint Genome Institute Community Science Program. KBase was supported by Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.