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

Janaka N. Edirisinghe[1,*] (janakaed@anl.gov), Michael Shaffer[2,*] (michael.t.shaffer@colostate.edu), Mikayla A. Borton[2], Evan Stene[3], Lucia S. Guatney[3,4], Farnoush Banaei-Kashani[4], Kelly C. Wrighton[2], Christopher Henry[1], and **Christopher S. Miller**[3]

[1]Argonne National Laboratory, Argonne, IL; [2]Colorado State University, Ft. Collins, CO; [3]University of Colorado Denver, Denver, CO; and [4]University of Colorado Anschutz Medical Campus, Aurora, CO

Project Goals:

Advances in 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 of microbial communities of DOE relevance. We are developing a suite of improved computational tools for protein annotation, and integrating these tools with simultaneous advances in inference of community-level metabolic models that incorporate complex interactions between environment and microbes, and among microbial community members. By integrating these tools into the DOE Systems Biology Knowledgebase (KBase)[1], we aim to make these tools accessible to a broad user base of scientists.

Abstract:

In order to infer, understand, and model microbial and ecosystem traits and processes of relevance to biogeochemical cycling, protein and metabolite function need to be encoded, inferred, and studied in the context of community-level metabolic models. Our approach to inference of improved models relies on developing new computational tools for the microbial sciences community, in three main areas: 1) improved homology-based and non-homology-based protein annotations, 2) building an iterative cycle of gap-filling organism-level and metabolic models with improved protein annotations, and informing probabilistic protein annotations based on metabolic models, and 3) integrating improved protein annotations related to exchange metabolisms with community-level flux balance metabolic models, in order infer ecosystem-level processes and community-level interactions.

To improve the inference of protein annotations, we have recently developed a computational pipeline, Distilled and Refined Annotation of Metabolism (DRAM) that integrates annotations across sensitive sequence-homology searches from a variety of both broad and specialized databases[2]. The comprehensive approach yields an increased number of protein annotations, especially for difficult to annotate taxa with poor genomic sampling. DRAM is also unique in the synthesis it performs, returning annotations in metabolism-centric outputs and visual outputs that allow for quick comparisons of key metabolisms for expert curation. We have recently integrated DRAM into KBase as an additional annotation option, which allows for integration with a growing number of state-of-the-field Flux Balance Analysis (FBA) models in KBase.

DRAM is also being extended to automate non-homology methods common to expert curation, including using conserved gene neighborhood information across metagenome-assembled genomes for inference of protein function. Other planned non-homology methods for incorporation into a probabilistic annotation framework include co-expression across metatranscriptome, co-occurance across multiple samples, and co-occurrence with metabolite networks when metabolomics data is available.

Many of these non-homology methods require read-mapping-based counting across multiple samples. To make these approaches scalable, we are adopting a training approach for learning-enhanced read mapping. Using deep learning techniques and models borrowed from Natural Language Processing (NLP) applied to genome sequences (such as BERT[3]), we generate vectorized representations of assemblies. These same vectorizations can be generated from reads using the trained model and compared to assembly vectors as a form of mapping. The efficiency of the learned model is evaluated by its ability to match two related sequences (for example kmers from the same read, or adjacent sequences) as well as the agreement of learning-enhanced mapping and traditional read mapping tools.

The Argonne team has made significant progress towards enabling modeling tools to use improved annotations from DRAM and other algorithms. Specifically, an API was developed in KBase to support consistent reading and writing of ontology terms from KBase genomes. The KBase metabolic model reconstruction tool was adapted to permit users to build models from a variety of annotation ontologies, including EC, GO, KO, and RAST (in collaboration with the Stuart SFA). Additionally, the KBase model gapfilling tool is being adapted to permit the use of evidence from alternative annotation sources. This tool is available in prototype form in a Jupyter notebook. Finally, a new metagenome model reconstruction algorithm was added to KBase, permitting the reconstruction of metagenome models from annotated metagenome assemblies (AMA). The annotation ontology API was connected to AMA objects so alternative annotations (like DRAM) could eventually be performed on these objects as well. All of this lays the groundwork for the use of metabolic models on the levels of Metagenome-Assembled Genomes (MAGs) and entire metagenomes, to integrate multiple sources of annotation evidence and propose the most consistent metabolic representation of biological systems possible. This work builds towards our goal of community-level models for the inference of ecosystem-level processes and community-level interactions.

References

1. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol*. 2018;36:566–9. Available from: http://www.nature.com/articles/nbt.4163
2. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res*. 2020;48:8883–900. Available from: https://academic.oup.com/nar/article/48/16/8883/58847382.
3. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv Preprint*. 2018; Available from: http://arxiv.org/abs/1810.048053.

This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), grant no. DE-SC0021350 and Systems Biology Knowledgebase (KBase) was supported by Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.