Rapid Functional Annotation Using Degenerate Kmers

Jason E. McDermott (Jason.McDermott@pnnl.gov)[1,2], William Nelson[1], Christine Chang[1], Bryan Killinger[1], Joon Lee[1], Arif Khan[3], Sayan Ghosh[3], Mahantesh Halappanavar[3], **Robert Egbert[1]**
[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland WA;
[2]Department of Molecular Microbiology and Immunology, Oregon Health & Science University, Portland, OR
[3]Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland WA

https://genomicscience.energy.gov/research/sfas/pnnlbiosystemsdesign.shtml

Project Goals:

The Persistence Control Science Focus Area at the Pacific Northwest National Laboratory is focused on developing fundamental understanding of factors governing the persistence of engineered microbial functions in rhizosphere environments. From this understanding, we will establish design principles to control the environmental niche of native rhizosphere microbes for the model bioenergy crop sorghum through data-driven genome reduction and engineered metabolic addiction to plant root exudates. These principles will lead to secure plant–microbe biosystems that promote secure, stress-tolerant, and highly productive biomass crops.

Abstract:

Proteins enact the functionality encoded by genomes and so understanding protein function is critical to understanding and predicting how natural and synthetic bacteria would interact with communities like native rhizosphere. Prediction of protein function from sequence is possible because of evolutionary relationships between proteins with similar functions, and existing algorithms can identify the corresponding sequence similarity. However, many proteins have similar functions but diverse sequences, which thwart existing methods, and driven by advances in sequencing technology the number of protein sequences with no known function or similarity to proteins of known function is large and growing rapidly. Additionally, non-genomic data from high-throughput functional screens and multi-omics approaches can be invaluable to providing information about protein function, but currently no methods exist that integrate such information with sequence-based approaches to provide functional annotations for proteins.

Previously we have developed machine learning methods to predict functions for problematic protein families including Type III secreted effectors (*1, 2*), multidrug resistant efflux pumps (*3*), and ubiquitin ligase mimics from bacteria and viruses (*4*). In the current work we describe development of a general, modular pipeline to represent protein sequences as vectors of short peptide sequences (called kmers), using both native and degenerate amino acid encoding to provide flexibility. The pipeline we describe creates these vectors, clusters them based on similarity, analyzes the network structure of the resulting graph, and creates signatures of kmers which best characterize protein families. The modular nature of the pipeline will allow incorporation of information and relationships derived from non-genomic sources of data such as fitness data for multiple environmental conditions being produced by our project using RB-TnSeq mutagenesis, metatranscriptomics and metaproteomics under different environmental

conditions, and functional relationships derived from other sources such as sequence co-evolution.

We describe two applications of our novel pipeline. In the first application we have used high-performance computing to assess the similarities between over 20 million bacterial protein sequences. We encoded the sequences using our pipeline, then calculated pairwise similarity using a GPU-based algorithm (*5*), and used exascale graph analytics to identify clusters of closely related sequences (*6*). The resulting analysis provides a landscape analysis of the bacterial protein universe. We show that this method can recapitulate known relationships between proteins based on traditional means (such as BLAST and hidden Markov models), highlight inconsistencies in the underlying protein database, and provide hypotheses for functions of novel proteins thus providing a large-scale sequence landscape. In the second, we have developed flexible kmer signatures for rapid and accurate classification of nitrogen cycling gene families from metagenomes. We evaluated the performance of our method relative to standard annotation methods in terms of sensitivity and specificity and speed of application.

Though still under development, our open-source, modular pipeline represents an advancement for analysis of large sets of protein sequences, and determination of complementation landscapes for complex microbiomes. The results we present suggest that our approach can provide expanded functional information from metagenomes, and will support integration of multiple other sources of information such as functional screens and omics data.

References

1. R. Samudrala, F. Heffron, J. E. McDermott, Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. *PLoS Pathog* **5**, e1000375 (2009).
2. J. E. McDermott *et al.*, Computational prediction of type III and IV secreted effectors in gram-negative bacteria. *Infect Immun* **79**, 23-32 (2011).
3. J. E. McDermott, P. Bruillard, C. C. Overall, L. Gosink, S. R. Lindemann, Prediction of multi-drug resistance transporters using a novel sequence analysis method. *F1000Res* **4**, 60 (2015).
4. J. McDermott *et al.*, Prediction of bacterial E3 ubiquitin ligase effectors using reduced amino acid peptide fingerprinting. *PeerJ* **7**, (2019).
5. J. Y. Lee, G. M. Fujimoto, R. Wilson, H. S. Wiley, S. H. Payne, Blazing Signature Filter: a library for fast pairwise similarity comparisons. *BMC Bioinformatics* **19**, 221 (2018).
6. S. Ghosh, M. Halappanavar, A. Tumeo, A. Kalyanaraman, A. Gebremedhin, in *2018 IEEE High Performance Extreme Computing Conference (HPEC)*. (Waltham, MA, 2018).

Funding statement: *This research was supported by the U.S. Department oEnergy (DOE), Office of Biological and Environmental Research (BER), as part of BER's Genomic Science Program (GSP), and is a contribution of the Pacific Northwest National Laboratory (PNNL) Secure Biosystems Design Science Focus Area "Persistence Control of Engineered Functions in Complex Soil Microbiomes". A portion of this work was performed in the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL), a national scientific user facility sponsored by BER and located at PNNL. PNNL is a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RL01830.*