Improving environmental contamination monitoring through microbial genomics with the integration of machine learning and mechanistic knowledge

David B Bernstein[1,2], Hanqiao Zhang[1], Adam P Arkin[1,2], aparkin@Berkeley.edu
[1]University of California at Berkeley, [2]Lawrence Berkeley National Lab
Consortium for Monitoring, Technology, and Verification (MTV)

Introduction and Motivation

Abstract:

Microbial organisms are ubiquitous in nature, and their sensitivity to environmental conditions makes them suitable candidates for monitoring environmental contamination, such as radionuclides produced by nuclear reactors. Microbial communities can be readily surveyed through genome sequencing: amplicon sequencing identifies which microbes are present, and metagenomic sequencing identifies the presence of genes encoding biological functions. Either of these data types can be used as the input for machine learning algorithms that discriminate pristine vs. contaminated environmental samples. Mechanistic knowledge can also be leveraged to inform such predictions, as microbes may respond predictably to environmental contaminants. For example, an environmental contaminant may lead to an increase in the presence of metabolic pathways that degrade it. Our work is aimed at utilizing microbial genomics to classify environmental contamination through a combination of machine learning and mechanistic knowledge. We are pursuing this aim through three different approaches. First, translating amplicon sequencing data into functional profiles to improve machine learning feature selection. Second, utilizing mechanistic knowledge to pre-select features of interest in metagenomics data. Finally, developing general approaches to combine machine learning with genome-scale metabolic models to predict microbial phenotypes from genotypes. An effective integration of mechanistic modeling/knowledge with machine learning could ultimately lead to improved prediction accuracy, particularly when data is sparse.

- Microbial community taxonomy [1] and genomic functions [2] have been shown to be predictive of environmental contamination.
- We aim to develop improved methods for the detection of environmental contamination through microbial genomics with the integration of mechanistic knowledge (metabolic pathways) into machine learning algorithms.

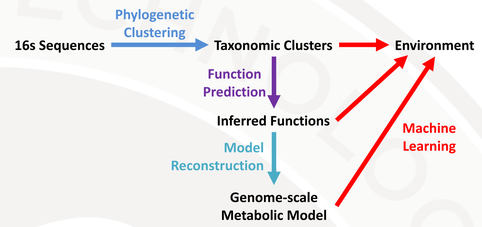
Mission Relevance

- This work will generally further the development of methods for the classification of environmental samples using microbial genomics.
- Relevant environmental contaminants include products of nuclear materials processing.

Technical Approach

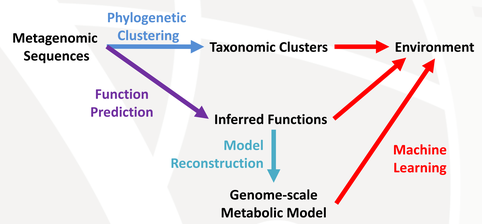
1. 16s Amplicon Sequencing

- 16s Sequencing data and geochemical measurements from 95 contaminated and pristine ground water wells at Bear Creek Watershed in Oak Ridge, Tennessee [1].
- Predict geochemical measurements including pH, Nitrate, Uranium, etc. using 16s data.
- Compare prediction for taxonomic, functional, or genome-scale metabolic model input features.



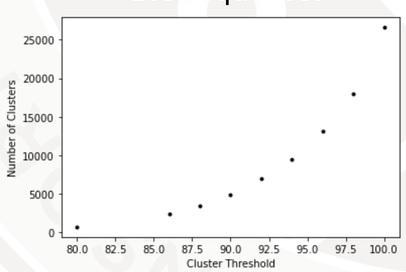
2. Metagenomics

- Compendium of 13,484 metagenomics samples from various environments ranging from human microbiome to marine [3].
- Predict environment from metagenomics data.
- Compare prediction for taxonomic, functional, or genome-scale metabolic model input features.

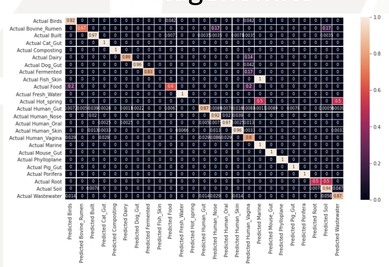


Results

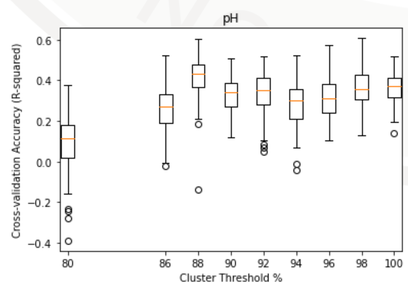
16s Sequences



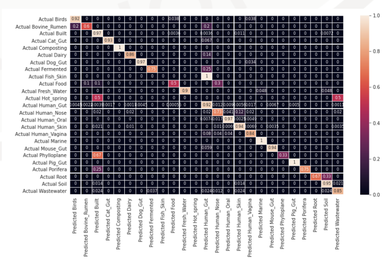
Metagenomics



Taxonomy



Functional Traits



MTV Impact

- Hired UC Berkeley undergraduate student through the Undergraduate Research Apprentice Program.
- Collaboration with Hazen (U Tennessee Knoxville, Oak Ridge National Lab), Alm (MIT), and Duff (Savanah River National Lab) labs for microbial community sampling and genomic data analysis.

Conclusion

- Prediction of environment from microbial genomics can yield different accuracy when using taxonomic vs functional input.

Next Steps

- Explore the sensitivity of our results to different steps in our pipeline
- Expand cross-validation with additional data
- Further integrate mechanistic knowledge into machine learning predictions (genome-scale metabolic modeling)

References

1. Smith, M. B., Rocha, A. M., Smillie, C. S., Olesen, S. W., Paradis, C., Wu, L., ... & Hazen, T. C. (2015). Natural bacterial communities serve as quantitative geochemical biosensors. MBio, 6(3).
2. He, Z., Zhang, P., Wu, L., Rocha, A. M., Tu, Q., Shi, Z., ... & Zhou, J. (2018). Microbial functional gene diversity predicts groundwater contamination and ecosystem functioning. MBio, 9(1).
3. Bahram, M., Netherway, T., Frioux, C., Ferretti, P., Coelho, L. P., Geisen, S., ... & Hildebrand, F. (2021). Metagenomic assessment of the global diversity and distribution of bacteria and fungi. Environmental Microbiology, 23(1), 316-326.

This work was funded in-part by the Consortium for Monitoring, Technology, and Verification under Department of Energy National Nuclear Security Administration award number DE-NA0003920

National Nuclear Security Administration