David Bernstein, University of California, Berkeley

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

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.