Systems Biology Research Symposium
Oral Presentation Session

Grand Ballroom
Tuesday, June 5th
7:00-8:30pm

Disease State Inference using Microbiome Informatics

Huzefa Rangwala[1,2,4], Ammar Naqvi[3,4], and Patrick Gillevet[2,3,4]

[1]Department of Computer Science, George Mason University, Fairfax, VA, USA, [2]Department of Environmental Science and Policy, George Mason University, Prince William, VA, USA, [3]Microbiome Analysis Center, Prince William, VA, USA, [4]Bioinformatics and Computational Biology Department, Prince William, VA, USA
Presenter's email address: rangwala@cs.gmu.edu

The human digestive tract is one of the most densely populated microbial environments on the earth. The influence of these microbial communities on the human development, immunity, and metabolism is largely unstudied. Using the new sequencing technologies we are able to effectively interrogate the the composition of the microbial communities in human mucosal samples. The scope of this paper is to characterize the composition of bacterial microorganisms within the gut and vaginal flora in relation to IBD and HIV, respectively using computational approaches.

To differentiate the healthy human control and disease suffering samples based on the composition and abundance of the 16S rRNA sequences, we assemble the different sequence reads from the 16S rRNA gene into well separated contigs representing the sequences for the different microorganisms. The contigs are then classified into a hierarchical phylogenetic taxa for bacterial species. We perform the taxa classification with three different methods: (i) using BLAST to search against existing in rRNA data in Genbank (ii) use a Bayesian classifier that uses a posterior probability to identify query sequence based on the occurrence of seven-length base pair subsequences in the RDP rRNA database, and (iii) a hierarchy-aware support vector machine (SVM) based classification system developed within the context of protein sequence classification.

We train SVM-based classification models to differentiate the healthy versus disease states. This is done using features generated from three sources: (i) the raw sequence reads obtained from the samples, (ii) the contig sequences after using the sequence assembly algorithm in SeqMan, (iii) and the predicted taxa information from the hybrid system developed above.

The broader impact of this project is that the computational pipeline developed for Metabiomic studies (i.e. the interactions between the microbiome and host) for IBD and HIV can be extended to other disease states, and physiological conditions like obesity. We will also extend the project to include the functional classification of the microbial organisms within the human body to explain the pathogenic roles played by such complex communities.

Key words: metabiome, classification, taxonomy, microbiome