ABSTRACT

NOVEL COMPUTATIONAL APPROACHES TO INVESTIGATE MICROBIAL DIVERSITY

By

Qingpeng Zhang

Species diversity is an important measurement of ecological communities. Scientists believe that there is a strong relationship between species diversity and ecosystem processes. In almost all metagenomics projects, diversity analysis plays an important role in supplying information about the richness of species, the species abundance distribution in a sample, and the similarity/difference between samples, all of which are crucial to draw insightful and reliable conclusions. Since we have limited sequencing power and financial constraints, the metagenomics data sets from high diversity samples like soil only correspond to a tiny fraction of the actual genomic content in the sample. The large size of data sets and low coverage make the assessment of microbial diversity in complex samples even harder. With novel applications of data structures and the development of novel algorithms, my research provides the necessary and highly desired computational methods to enable scalable microbial diversity analysis of the complex metagenomes, with further potential to facilitate other analysis like assembly, annotation.

This dissertation covers an overview of existing approaches of doing microbial diversity analysis of metagenomic samples, especially based on the concept of OTU, including the steps in the procedure, like contigs binning, statistical analysis of OTU abundance information to estimate the microbial diversity. As the foundation of the IGS based framework, we described a novel method to count k-mers efficiently and a scalable approach to retrieve the coverage of a read in a data set based on efficient and online k-mer counting. We also introduced

the applications of this approach in reducing the redundancy of metagenomic reads dataset and analyzing sequencing error, which is beneficial to other tasks in metagenomic data analysis, like assembly or error trimming. Next, we discussed how we developed the concept of IGS based on the methods of efficient k-mer counting and digital normalization discussed before. The application of IGS to analyze microbial diversity of metagenomic data sets was discussed and the performance of the IGS method on simulated data sets and real data sets were demonstrated in the final chapters. Since this method is totally binning-free, assembly-free, annotation-free, reference-free, it is specifically promising to deal with the highly diverse samples, while we are facing large amount of dark matters in it, like soil.