Clulstering and Classification of Metagenomes with Sequence Features

Xuegong Zhang
Tsinghua University, Beijing, China zhangxg@tsinghua.edu.cn

Metagenomes are the mixture of DNAs from all microbial genomes (the microbiome) in samples of environment or human niches. The next-generation sequencing (NGS) technology has made large-scale study of metagenomes feasible, which opens a promising new way for understanding our “other self”: the microbiomes that live with us. Comparing and discriminating metagenome samples is a basic task on analyzing metagenome samples. The conventional approach based on mapping metagenome sequences to reference genomes and/or genes in databases is limited by the availability of microbial genomes and gene annotations. An alternative approach is to use sequence signatures as features to explore the relation among multiple metagenome samples. Typical sequence features are the relative frequency of different k-mer sequence strings in the metagenome. We conducted a systematic study on the application of unsupervised and supervised machine learning methods based on sequence features for clustering and classifying metagenomic samples and illustrated the effectiveness of such reference-free methods in revealing underlying relationships of the studied samples based on metagenomic sequence features of the microbiomes they host.

Keywords: metagenome, sequence signatures, unsupervised learning, supervised learning