MinCE: Fast Quantification of Large Metagenomic Datasets Along With Species and Strain Abundance

Thorhallur Audur Helgason
Mentor: Lior Pachter

Dramatic cost reductions in genome sequencing coupled with improvements in accuracy during the past two decades have facilitated broad sequencing efforts to catalog all genomes of living organisms. In particular, large public databases now house hundreds of thousands of bacterial genomes, most of which have been obtained through sequencing of cultured bacteria. However many bacteria are difficult to isolate and/or culture. To study them, "metagenome" sequencing approaches have been developed that rely on sequencing of short fragments from environmental samples to identify microbes in their natural habitat. The associated computational problems are manifold and complex, starting with the need for algorithms to align hundreds of millions, or even billions of DNA fragments to large existing databases.

We introduce MinCE, a method for quickly identifying bacterial species and strains in metagenomic samples. MinCE preprocesses a reference genome database to facilitate rapid lookups. Subsequently, the relevant genomes serving as the source for a collection of DNA fragments can be identified. At present, MinCE can be used to identify genomes from a 13.5Gb reference database containing 258,339 genomes of Eubacteria and Archaea. We present results on simulated data that suggest that MinCE has high sensitivity and specificity, making it suitable for metagenomics analyses.