Optimization of alignment-based methods for taxonomic binning of metagenomics reads using the Design Of Experiments methodology

Metagenomics is the study of the genetic content of a sample. Sequencing all the DNA a sample contains ideally accesses the whole microbial diversity of the sample, and in particular micro-organisms that cannot be grown on a culture media. Taxonomic binning aims to assign each individual read from a metagenomic sample to a particular taxon within the microbial diversity. Alignment-based methods proceeds generally in two steps: first align reads against a reference database, and then for each read, identify the lowest common ancestor among the best hits retrieved. Beyond the sequencing technology and the completeness of the RDB, the performance of the workflow depends on the mapper, the mapper parameters, and the best hit selection threshold, and selecting the optimal configuration of the workflow remains quite empirical. Here, we will present how the experimental design methodology can be used to find the optimal taxonomic binning workflow configuration, at a minimal computational cost. First, we simulated metagenomic samples under different sequencing error models, and evaluated the workflow configurations given by a screening experimental plan. Several performance indicators were estimated and prediction models were used to identify the optimal workflow configuration. This optimal configuration was then validated using independent datasets.