Augmenting taxonomic profiling with coverage information to improve sensitivity and specificity

Martin S. Lindner[1,2]

[1]Robert Koch Institut, Berlin
[2]4-Antibody, Basel

Metagenomic samples typically consist of mixture of genomic material from multiple (in particular microbial) organisms. One of the key challenges in metagenomics, denoted as Taxonomic Profiling, is to disentangle the genomic ravel and identify the organisms present in a sample. In this talk, I will show how genome coverage information can be used to circumvent typical pitfalls causing low sensitivity and specificity of current approaches in difficult situations.

Our idea was to fit mixtures of discrete probability distributions to genome coverage profiles with the Expectation Maximization algorithm. With this information, we can calculate the average coverage in the covered areas of the genomes, handle spike-like artifacts, and estimate the similarity between the reference genome and the organism in the sample. In our taxonomic profiling tool MicrobeGPS, we use this information to cluster reference genomes into groups, each representing one organism in the sample. In addition to quantitative measures such as number of reads and relative abundance, our approach provides further information on the identity and reliability of the observed organisms. This simplifies the interpretation as well as leads to higher sensitivity and specificity of the results.