Title: „Unmasking the Sampling Bias of Prokaryotic Genomes“

Author: Hannah Götsch (University of Vienna)

Coauthor: Franz Baumdicker (University of Tübingen)

Abstract:

Over the last decades the number of completely sequenced genomes in repositories such as NCBI is increasing rapidly. However, various analyses become difficult or even infeasible when dealing with very large datasets. In addition, datasets are often redundant and suffer from sampling biases. Especially bacteria datasets are often generated by biased sampling schemes. Excessive sequencing during disease outbreaks leads to clades of highly related samples. Strong sampling bias not only reduces the effective information in databases but can also result in misleading conclusions in various analyses. For example, it may lead to an excess in intermediate frequencies in site frequency spectra and obscure the real distribution of SNPs and genes.

We present a statistical tool that can detect oversampling in prokaryotic populations. Considering phylogenetic relationships between genomic samples, we estimate how many samples in the dataset are actually relevant (effective sample number), identify the „oversampled“ genomes, reduce the data to a subset that is closer to a random sample and consequently does not bias predictions.