Abstract

In this thesis, we have studied two computational problems in the statistical analysis of genomic data. First of them is the estimation of proportions of variants of SARS-CoV-2 virus in mixed sequencing samples. This allows to monitor the variants circulating in a given community using wastewater sequencing. We have proposed a mixture model of the sequencing process (including sequencing errors), using the frequencies of mutations for the selected variants. The design of the model allows to compute the estimates in a parallelizable manner. We have then experimentally assessed the precision of our approach on in silico and in vitro mixtures and also on wastewater samples from France and Slovakia. The second problem is the computation of the statistical significance of the number of overlaps between two genomic annotations (also called colocalisation analysis). A genomic annotation is a set of intervals on a genome, e.g. genes or telomeres. We have first shown that one of the popular definitions of the null hypothesis leads to an $\mathcal{NP}$-hard task. We have then proposed a direct sampling scheme for annotations under such null hypothesis. The main contribution is the reformulation of the null hypothesis using Markov chains and a fully polynomial algorithm for the calculation of p-values under it. The running time of the algorithm is independent of the length of the genome and of individual intervals. This allows to analyse annotations with tens of thousands elements (e.g. all human exons) on a common laptop in a reasonable time span. We have then experimentally assessed the efficiency and fidelity of our approach on both simulated and real datasets.

Keywords: SARS-CoV-2, variants, wastewater sequencing, genomic annotation, colocalisation analysis, Markov chains