Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis of human gut microbiota

Veronika B. Dubinkina[1,2], Alexander V. Tyakht,[1,2], Dmitry G. Alexeev[1,2]

[1]*Moscow institute of physics and technology (State University), Dolgoprudny, Russia,* dubinkina@phystech.edu

[2]*Research institute of physico-chemical medicine, Moscow, Russia*

During the last decade next-generation sequencing technologies developed explosively, and a large amount of short read metagenomic data has been accumulated. The first step in many metagenomic studies is beta-diversity (β-diversity) analysis - a measure of differences between two microbial samples, forming the basis for quantitative comparison of multiple samples. Therefore it is necessary to have effective methods for feature extraction for such challenging data: usually metagenomes consists of many different organisms (including bacteria, archaea, fungus, and viruses), genomic sequences of which may be unknown.

In recent years among methods for analysis of a genomic data, more interest is attracted to alignment-free methods for comparing samples based on work with *k*-mers (oligonucleotides of length *k,* also called *l*-tuples or *n*-grams) directly from metagenomic reads. This approach is highly effective for exploratory analysis (e.g. taxonomic identification and clustering) of large data sets, since it requires a relatively small number of additional calculations in comparison with other methods, includes the entire data set into analysis and does not depend on the reference. The most simple and effective for large data sets analysis is comparison of sequences by calculation of pairwise distances between them on the basis of *k*-mer spectra.

In this study we compared the most common reference-based methods (determination of beta-diversity of bacterial composition, genes, phylogeny-aware

methods) and *k*-mer approach - in order to see how different characteristics of the examined data influence the results of *k*-mer spectra analysis and to identify the advantage of a *k*-mer analysis compared with reference-based approaches. To evaluate the applicability of *k*-mer-based dissimilarity, metagenome of human gut microbiota was selected, the study of which has great biomedical importance and perspective.

In the first experiment, we simulated 100 metagenomes, consisting of 10 prevalent human gut bacterial genomes. On this dataset we investigated *k*-mer method performance and compared matrices of pairwise differences based on Bray-Curtis dissimilarity obtained using *k*-mer and using genus composition. Consequently k=11 was selected as the optimal value for further analysis.

In the second experiment, the database consisted of 280 real metagenomic samples of human gut sequenced in large-scale metagenomic projects: populations of China [1] and USA [2] were analysed. The *k*-mer dissimilarity matrix was compared with the matrix obtained via the following reference-based methods:

1) read mapping against the catalog of reference genomes for human gut from a previously published study [3]. Distance between metagenomes was built based on bacterial abundances using two ways: Bray-Curtis dissimilarity and whole-genome analog of UniFrac metric [3].

2) quantitative identification of reads using clade-specific marker genes - MetaPhlAn [4] (Bray-Curtis dissimilarity was used).

3) read mapping on the catalog of 3 million genes for gut microbiota [5] and then summing abundances across COG groups (Clusters of Orthologous Groups) (here Bray-Curtis dissimilarity was used, too).

Comparison of dissimilarity matrices showed that the closest correlation with the *k*-mer based metric belongs to functional COG composition of metagenomes. The analysis revealed specific differences between the cohorts from the two studies. To

identify the major factors contributing to this phenomenon, we analyzed the effect of differences in taxonomic composition, sequence quality scores and other factors. Interestingly, this investigation showed that *k*-mer approach can help to hint at the presence of dominant metagenomic components which are left unnoticed by the reference-based methods – which turned out to be a prevalent gut bacteriophage crAssphage in our case [6]. Therefore this approach is a useful addition to the «standard methods» of metagenomic analysis allowing quality control and rapid identification of a presence of novel organism.

References

1. J. Qin et al. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes, *Nature*, **7418**: 55–60.
2. Human Microbiome Project Consortium et al. (2012) Structure, function and diversity of the healthy human microbiome, *Nature*, **7402**: 207-214.
3. A. Tyakht et al. (2013) Human gut microbiota community structures in urban and rural populations in Russia, *Nature communications*, **T. 4**.
4. N. Segata et al. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes, *Nature methods*, **9**: 811-814.
5. J. Qin et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing, *Nature,* **7285**: 59-65.
6. B. Dutilh et al. (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes, *Nature communications*, **T. 5.**