MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data

Sergey V. Kazakov[1], Vladimir I. Ulyantsev[1], Veronika B. Dubinkina[2,3], Alexander V. Tyakht[2], Dmitry G. Alexeev[2,3]

[1]*ITMO University, Saint-Petersburg, Russia,* svkazakov@rain.ifmo.ru

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

[3]*Moscow institute of physics and technology (State University), Dolgoprudny, Russia*

Since the emergence of high-throughput genomic sequencing technologies, a huge volume of data has been accumulated describing complex microbial communities (microbiota). Efficient statistical analysis of such data – including identification of taxonomic and functional composition, community richness and similarity – requires dimension reduction. Cost decrease and progress of sequencing technologies led to the trend that microbiota from previously unstudied environments are sequenced increasingly often. The novel microbiota have a large fraction of uncultivated bacteria. Accordingly, no representative genomes exist for such microbes that could serve as a reference. Such problem poses a significant drawback even for communities thoroughly examined during decades like human gut microbiota where unknown genomes represent a lion's share of total DNA sequences [1].

Common methods for feature extraction from shotgun metagenomic data include:

- reference-based methods implicating alignment of sequence reads against a catalog of reference sequences [2];
- methods based on *de novo* assembly with subsequent analysis of the long contigs [3];
- abstract composition-based methods including k-mer spectrum analysis [4], neural networks, Markov models, etc.

Despite the popularity of the available approaches, most of them have inherent disadvantages that limit their scope of applicability. For instance, reference-based methods require a representative database of known genomes, however, many microbial branches of

the tree of life still lack such sequences. Assembly-based methods are computationally intense and can hardly be applied to metagenomes with highly complex community structure. On the other hand, composition-based methods do not require a reference base and mostly are computationally efficient; however, the interpretation of the resulting features is often unclear.

We have developed a novel approach *MetaFast* for compact representation of metagenomes. It does not require a priori knowledge about the taxa possibly included in the microbiota. Other advantages over the above-mentioned methods are rather small system requirements and interpretability of the extracted features.

The algorithm consists of four steps.

1. Assembling short genomic sequences from reads for every metagenome separately (basing on de Bruijn graph). These sequences are similar to contigs, but several times shorter.
2. Constructing one combined de Bruijn graph for all assembled sequences from all metagenomes and searching for connected components in it. Large components are subdivided into small ones by removing vertices occurring in a small number of metagenomes.
3. Calculating a characteristic vector for every metagenome with a length equal to the number of connected components. Each vector element is the number of k-mers from a connected component that are present in reads of the metagenome.
4. Cross-comparing metagenomes by calculating the Bray-Curtis dissimilarity matrix based on characteristic vectors.

To test the applicability of our method, several datasets consisting of simulated and real metagenomes were used.

In the first experiment, 100 metagenomes were generated by randomly fragmenting 10 bacterial genomes of most common intestinal bacteria into “reads”. Matrices of pairwise dissimilarity obtained using the proposed *MetaFast* method and by the Bray-Curtis measure on the basis of preassigned bacterial compound were found to be highly correlated (Mantel test:

Spearman correlation r=0.96; p-value=0.001).

In the second experiment, we tested *MetaFast* on a dataset of 157 real metagenomic samples of the human gut microbiota from a recent large-scale metagenomic project [5]. This dataset includes diverse bacterial compositions from both healthy people and patients with type 2 diabetes. For this data we applied *MetaFast* to calculate the pairwise dissimilarity matrix. We also calculated the taxonomic and functional composition using mapping to reference set of genomes [2]. Comparison analysis showed that in general *MetaFast* produces a dissimilarity matrix similar to those provided by reference-based approaches. *MetaFast* has better conformity with taxonomic composition (Spearman correlation r=0.92; p-value=0.001) than with functional composition (Spearman correlation r=0.81; p-value=0.001). Fig. 1 shows Procrustes analysis for multidimensional scaling (MDS) of dissimilarity matrices based on taxonomic composition and *MetaFast* method. An example of *MetaFast* output on a subset of the data is shown in Fig. 2.

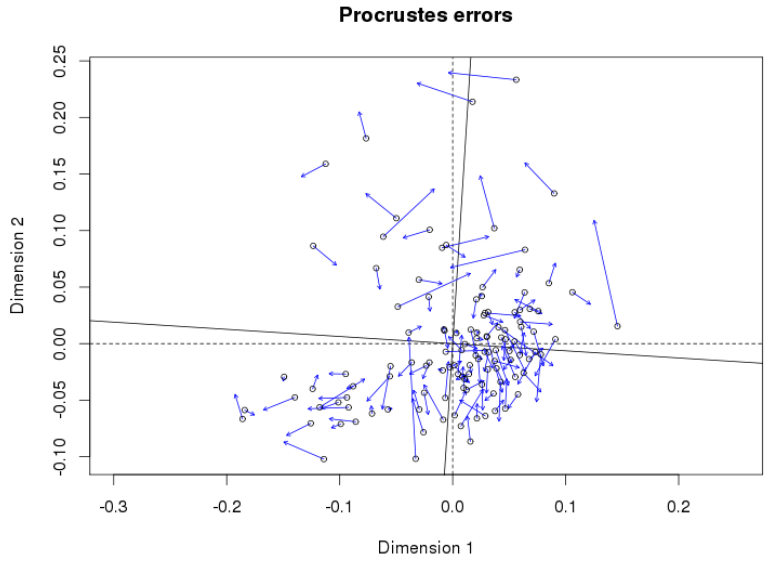

Fig 1. Procrustes analysis for MDS based on taxonomic and *MetaFast* dissimilarity matrices.

The largest dataset (157 real metagenomes) consisted of 7.8 billion reads with an average length 90 bp, a total of 580 Gb archived FASTQ files. The complete analysis by *MetaFast* took approximately 30 hours on a supercomputer, using 120 Gb of memory and 20 CPUs (AMD Opteron™ Processor 6176 - 800MHz).

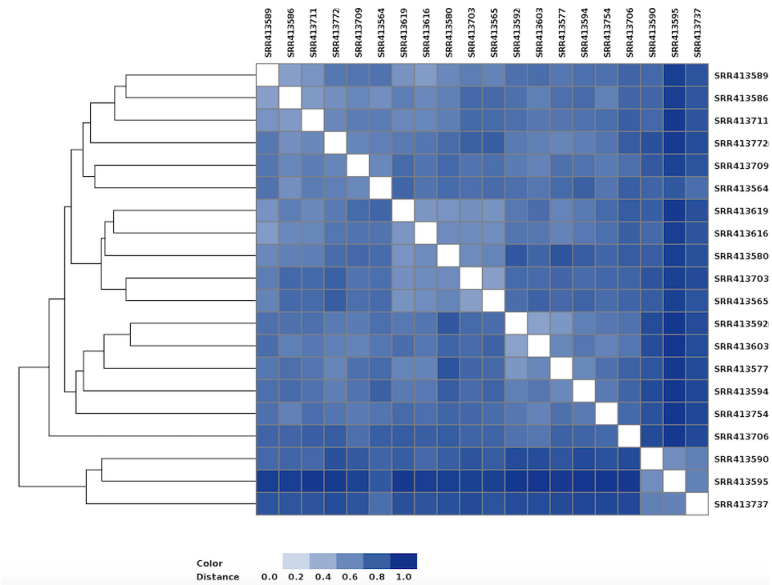

Fig 2. Example of *MetaFast* output on a set of real data.

The software *MetaFast* was implemented in Java and can be run on any operating system (tested on Linux 2.6.32 x86_64). Source code, binaries and documentation can be found at github.com/ctlab/metafast.

References

1. H. Nielsen et al. (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes, *Nature biotechnology*, **32.8 (2014)**: 822-828.
2. A. Tyakht et al. (2013) Human gut microbiota community structures in urban and rural populations in Russia, *Nature communications*, **T. 4**.
3. J. Nijkamp et al. (2013) Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold, *Bioinformatics*, **29(22)**:2826-2834.
4. R. Edwards et al. (2012), Real Time Metagenomics: Using k-mers to annotate metagenomes, *Bioinfomatics*, **28(24)**:3316-3317.
5. J. Qin et al. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes, *Nature*, **7418**: 55–60.