BITS :: Call for Abstracts 2019 - Oral communication

Type	Oral communication
Session	Metagenomics
Title	Metagenomic analysis through the eBWT
All Authors	Guerrini V (1), Rosone G (1)

Affiliation

(1) Dipartimento di Informatica, Università di Pisa, Pisa, Italy

Motivation

Metagenomics refers to the sequencing of microbial DNA collected directly from the environment, without isolation and lab cultivation of individual species. The analysis of environmental samples (i.e. metagenomes) are particularly important to figure out the microbial composition of different ecosystems and is used in several fields: for example, metagenomic studies in agriculture can help detecting crop diseases and understanding the relations between microbes and plants. Of fundamental importance is to identify with precision the microorganisms that are present in a metagenomic sample by comparing the biological sequences therein and assigning them to a specific taxon. At the very beginning the best strategy used to classify metagenomes was the aligner BLAST. Nevertheless, as the reference databases and the shotgun sequencing datasets have grown in size, the alignment strategy has become computationally expensive and alignment-free methods have been developed. Recent studies [Lindgreen et al, 2016] have attempted to benchmark the performance of several metagenomic classifiers by using simulated metagenomic datasets. From these studies Kraken and CLARK resulted top-performing tools in terms of both sensitivity and precision. Both are k-mer based methods, i.e. they classify a sequence by mapping its substrings of length k. An alternative approach to fixed k-mers is to use spaced seeds, for which the exact matching is required only for a subset of nucleotides. CLARK-S is the new version of CLARK that uses spaced k-mers and achieves higher sensitivity with the same high precision. The main drawback of these k-mer based approaches is that they are extremely memory-consuming. Other efficient alignment-free methods, such as Centrifuge, are based on a read-mapping strategy and use the FM-index to store and index the genome database.

Methods

The metagenomic classification problem has been formalized as follows: let S be a collection of biological sequences containing both reads and genomes, any read r is assigned to a genome g of provenance according to the similarity between r and g. In [1] we proposed a new lightweight approach for metagenomic classification which is not k-mer based and is assembly- and alignment-free. This approach is based on an extension of the Burrows-Wheeler transform (eBWT) and on the following intuitive idea: the greater is the similarity between two sequences, the greater is the number of substrings they share. Our method takes in input the eBWT of the entire collection S, enhanced with its document array and longest common prefix array. Our strategy works, by sequentially scanning them, in three steps: (1) detect and keep some “relevant” blocks of the eBWT; (2) analyze these blocks to evaluate a degree of similarity between any read and any genome in S; (3) perform the read assignment to a particular taxon. The notion of similarity between sequences has been defined by exploiting the clustering effect of the eBWT: this transformation tends to group together symbols that occur in similar contexts in the input string collection S. Now, we extend this notion of similarity so that our measure takes into account the degenerate base symbols ("ambiguity" characters associated with every possible combination of the four DNA bases). Moreover, we improve the steps (2) and (3), as to get a higher precision and sensitivity.
[1] V. Guerrini and G. Rosone. Lightweight Metagenomic Classification via eBWT. AlCoB 2019

Results

We evaluated our alignment-free strategy against two tools: CLARK-S and Centrifuge. To assess the performance of our sequence analysis method, we implemented a prototype C++ tool, LightMetaEbwt. Our tool is able to classify the reads to several taxonomic levels such as genomes, species or phylum, yet using a much smaller memory footprint. We perform the validation of our approach by using simulated metagenomes among those provided in benchmarking analysis. They reproduce the size, complexity and characteristics of real metagenomic samples containing around 20 millions of sequences (for the positive control) in addition to a negative control set of random shuffled reads which mimic sequences from unknown organisms that are likely to appear in metagenome samples. The overall accuracy for the three tools in the positive control was very similar, but the highest precision percentage (keeping high sensitivity) is obtained by LightMetaEbwt - e.g. 99,7%. The experiment results show that LightMetaEbwt achieves the best F1 score - the harmonic average of precision and recall - and the percentage of classified reads for the negative control is less than 0.01%. CLARK-S has lower precision than LightMetaEbwt and Centrifuge in the positive control, but the percentage of classified random shuffled reads is as much as LightMetaEbwt. While Centrifuge that keeps high sensitivity and precision in the positive control set classifies a higher number of random reads that should not be assigned to any taxon.

Info

Funding: GR is partially, and VG is totally, supported by the project MIUR-SIR CMACBioSeq ("Combinatorial methods for analysis and compression of biological sequences") grant n. RBSI146R5L.

Figure

-

Availability	https://github.com/veronicaguerrini/LightMetaEbwt.git

Corresponding Author

Name, Surname	Giovanna, Rosone
Email	giovanna.rosone@unipi.it
Submitted on	29.04.2019