Bioinformatics Advance Access published August 18, 2016

Sequence Analysis

Higher classification sensitivity of short metagenomic reads with CLARK-*S*

Rachid Ounit and Stefano Lonardi *

Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.

Associate Editor: Dr. Inanc Birol

Abstract

Summary: The growing number of metagenomic studies in medicine and environmental sciences is creating increasing demands on the computational infrastructure designed to analyze these very large datasets. Often, the construction of ultra-fast and precise taxonomic classifiers can compromise on their sensitivity (i.e., the number of reads correctly classified). Here we introduce CLARK-*S*, a new software tool that can classify short reads with high precision, high sensitivity and high speed.
Availability: CLARK-*S* is freely available at http://clark.cs.ucr.edu/
Contact: stelo@cs.ucr.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

One of the primary goals of metagenomic studies is to determine the taxonomical identity of bacteria and viruses in a heterogenous microbial sample (e.g., soil, water, urban environment, human microbiome). This analysis can reveal the presence of unexpected bacteria and viruses in a newly explored microbial habitat, or in the case of the human body, elucidate relationships between diseases and imbalances in the microbiome.

Arguably, the most effective and unbiased methods to study these microbial samples is via high-throughput sequencing. The associated computational problem is to assign sequenced (short) reads to a taxonomic unit. Several methods and software tools are available, but faster and more accurate algorithms are needed to keep pace with the increasing throughput of modern sequencing instruments. In Ounit *et al.*, 2015 we introduced CLARK, a taxonomy-dependent binning method whose classification speed is currently unmatched. A recent independent evaluation of fourteen taxonomic binning/profiling methods (Kraken by Wood and Salzberg, 2014, MEGAN by Huson *et al.*, 2007, and many others) showed that the classification precision of CLARK is comparable (sometimes better) than the state-of-the-art classifiers (Lindgreen *et al.*, 2016). While CLARK's speed and precision are very high, its classification sensitivity (i.e., the fraction of reads that it correctly classifies) can be significantly improved with the methods described next.

We recall that CLARK is a k-mer based method. Briefly, it assigns a read r to a reference genome G if r and G share more *discriminative* k-mers (i.e., k-mers that appear exclusively in one reference genome) than other genomes in the database. Here we show that the classification sensitivity can be increased by allowing mismatches between shared k-mers in a limited number of (carefully predetermined) positions, while maintaining the requirement for k-mers to be discriminative. The idea of allowing mismatches to improve the sensitivity of seed-and-extend alignment methods was pioneered in Ma *et al.*, 2002 with the notion of *spaced seed*. While spaced seeds have been used in some metagenomic binning/profiling methods (e.g., MEGAN), the use of *discriminative* spaced k-mers is novel. Here we describe an extension of the algorithmic infrastructure of CLARK based on spaced seed, called CLARK-*S*.

2 Methods

Given an integer k and m reference genomes $\{g_1, g_2, \ldots, g_m\}$, the *discriminative* k-mers D_i for genome g_i is the set of all k-mers in g_i that do not occur (exactly) in any other genome (Ounit *et al.*, 2015). A spaced seed s of *length* k and *weight* $w < k$ is a string over the alphabet $\{1,\star\}$ that contains w '1' and $(k - w)$ '$\star$'. Matches are required at a '1' positions, while mismatches are allowed at the '$\star$' locations. The set of discriminative spaced k-mers $E_{i,s}$ is the set of all k-mers of D_i that do not occur in any other set D_j $(j \neq i)$ when mismatches are allowed at '$\star$' positions in s.

The common model to align a short (Illumina) read r of fixed length to a reference genome G is not to allow gaps and assume that mismatches (due to genomic variations or sequencing errors) follow a Bernoulli distribution with parameter p (the similarity level between r and G) (Brown *et al.*,

2004). In such a model, the structure of the spaced seed (or multiple spaced seeds) is critical to achieve the highest possible precision and sensitivity (Ma *et al.*, 2002; Brown *et al.*, 2004). Finding an optimal set of spaced seeds through w, k and p is computationally difficult (Brown *et al.*, 2004), thus we decided to reduce the space search by judiciously setting w, k and p. Since CLARK is more precise for long contiguous k-mers (e.g., $k = 31$), but its highest sensitivity occurs for $k \in [19, 22]$, we considered spaced seeds of length $k = 31$ and weight $w = 22$. We set $p = 0.95$ to reflect the expected high similarity between genomic sequences at the species rank. Through an exhaustive search of all spaced seeds (with parameters $k = 31$, $w = 22$, $p = 0.95$) and using the dynamic programming method implemented by Ilie *et al.*, 2011 on a region of 100bp, we selected the three spaced seeds with the highest "hit probability" (Ma *et al.*, 2002), namely `1111*111*111**1*111**1*11*11111` (hit probability 0.99811), `11111*1**111*1*11*11**111*11111` (0.998099), and `11111*1*111**1*11*111**11*11111` (0.998093).

In the preprocessing stage, CLARK-*S* stores on disk, for each genome i and each spaced seed s, the set of discriminative spaced k-mers $E_{i,s}$. Compared to CLARK's classification phase, CLARK-*S* requires three look-ups for each k-mer in a read (one look-up per spaced seed).

3 Experimental Setup

As said, a recent independent evaluation of several published taxonomic binning methods showed that CLARK and Kraken are the two most accurate tools at the genus and phylum level (Lindgreen *et al.*, 2016). Instead of comparing CLARK-*S* to all published binning methods, it is therefore sufficient to compare it against CLARK and Kraken. To guarantee a consistent and fair evaluation, we ran CLARK-*S*, CLARK and Kraken on the same set of reference genomes, namely all microbial genomes in the NCBI RefSeq database (total of 5,747 species: 1,335 bacteria, 123 archaea and 4,289 viruses). Evaluations were carried out on simulated datasets and real metagenomic data, as explained next.

We created six synthetic datasets, each representing a distinct microbial habitat and containing reads from the related dominant organisms (see Suppl. Note 1 for full details). We included samples from the human mouth (characterized by 12 dominant species), city parks (48 species), human gut (20 species), household (two datasets, 31 and 21 species) and soil (50 species). A seventh dataset included reads from 525 randomly chosen bacterial/archaeal species (see Suppl. Note 1). All these datasets are composed of 100bp reads generated by ART (Huang *et al.*, 2012) using the Illumina error model with default settings. Since two distinct species i and j can have sequence similarity as high as 98.8% (Stackebrandt and Goebel, 1994), a short read r generated from genome g_i may appear in another genome g_j for a given error rate or number of mismatches. Ignoring he possibility of ambiguity in reads classification is likely to lead to incorrect conclusions on precision and sensitivity. In order to carry out an unbiased evaluation, we created additional datasets (called "unambiguous", see Suppl. Note 2 for details) in which no read can be mapped to more than one species with the same error rate or number of mismatches. We tested the three tools on fourteen datasets containing a total of 23.5M reads from 647 species (see Suppl. Table 1). We also added three negative control samples containing short reads that do not exist in any genome in the NCBI/RefSeq database (see Suppl. Note 1). We used the precision and sensitivity metrics defined in Ounit *et al.*, 2015 to evaluate the classification performance.

For experiments on real metagenomes, we chose a large dataset from a recent study on the microbial profile of the NY City subway system, the Gowanus canal and public parks (Afshinnekoo *et al.*, 2015). We selected twelve datasets containing a total of 105M reads from various microbial habitat (e.g., bench, garbage can, kiosk, stairway rail, water, etc.), subway stations and riders usage (see Suppl. Table 3). While the ground truth for these data is unknown, the abundance of bacteria, eukaryotes and viruses present in these samples were provided in Afshinnekoo *et al.*, 2015. We trimmed raw reads as it was done in Afshinnekoo *et al.*, 2015 (see Suppl. Table 3), then compared the results of CLARK-*S* with the findings in Afshinnekoo *et al.*, 2015 (see Suppl. Table 4 and 5).

4 Results and Discussion

Observe in Supplemental Table 2 that the sensitivity achieved by CLARK-*S* on the fourteen simulated datasets is consistently higher than other tools, while maintaining high precision (the increase in sensitivity is even higher on unambiguous datasets). Also, note that CLARK-*S* did not classify any reads from the negative control samples. Supplemental Table 7 shows that CLARK-*S* classifies about 200 thousand short reads per minute (using one CPU), while CLARK classifies about 3.5M short reads per minute. If one can take advantage of eight cores, CLARK-*S* classifies about 1M short read per minute, which is sufficiently fast to process large metagenomic datasets in few minutes (see Suppl. Note 3). CLARK-*S* requires more time to build the database than CLARK or Kraken, but its RAM usage is comparable to the other tools (see Suppl. Table 8).

Observe in Supplemental Table 6 (real datasets) that CLARK-*S* classifies more reads than CLARK or Kraken. On average, CLARK-*S* classifies 10% more reads than Kraken, and 27% more reads than CLARK. Supplementary Table 5 indicates the reads count assigned by each tool to each species listed in Afshinnekoo *et al.*, 2015 and present in the database. CLARK-*S* achieves consistently the highest agreement with Afshinnekoo *et al.*, 2015 on all samples. For instance, in P00589 and P00720, CLARK-*S* detected the presence of the virus *Enterobacter* phage HK97 but CLARK/Kraken did not; in sample P01136, CLARK-*S* detected *Brucella ovis* but CLARK/Kraken failed to do so. In general, CLARK-*S* identified more relevant organisms than the other tested tools, as observed by a recent study focusing on water samples (Thompson *et al.*, 2016).

Acknowledgements

We thank Dr. Christopher E. Mason (Weill Cornell Medicine) for valuable discussions on the tools' evaluation.

Funding

Supported in part by the US National Science Foundation [IIS-1302134, IIS-1526742].

Conflict of interest: None.

References

Afshinnekoo, E., Meydan, C., *et al.* (2015). Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Systems*, **1**(1), 72–87.

Brown, D. G., Li, M., *et al.* (2004). A tutorial of recent developments in the seeding of local alignment. *Journal of Bioinformatics and Computational Biology*, **2**(04), 819–842.

Huang, W., Li, L., *et al.* (2012). ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**(4), 593–594.

Huson, D. H., Auch, A. F., *et al.* (2007). MEGAN analysis of metagenomic data. *Genome Research*, **17**(3), 377–386.

Ilie, L., Ilie, S., *et al.* (2011). SpEED: fast computation of sensitive spaced seeds. *Bioinformatics*, **27**(17), 2433–2434.

Lindgreen, S., Adair, K. L., *et al.* (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports*, **6**, 19233.

Ma, B., Tromp, J., *et al.* (2002). PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**(3), 440–445.

Ounit, R., Wanamaker, S., *et al.* (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, **16**(1), 236.

Stackebrandt, E. and Goebel, B. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16s rRNA sequence analysis in the present species definition in bacteriology. *Journal of Systematic and Evolutionary Microbiology*, **44**(4), 846–849.

Thompson, L. R., Williams, G. J., *et al.* (2016). Metagenomic covariation along densely sampled environmental gradients in the red sea. *The ISME journal*.

Wood, D. and Salzberg, S. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*, **15**(3), R46.