Journal of Microbiological Methods 122 (2016) 38–42

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Studying long 16S rDNA sequences with ultrafast-metagenomic sequence classification using exact alignments (Kraken)

Fabiola Valenzuela-González [a], Marcel Martínez-Porchas [b],
Enrique Villalpando-Canchola [b], Francisco Vargas-Albores [b,*]

[a] Instituto Tecnológico de Sonora, 5 de Febrero 818 Sur. Col. Centro, Ciudad Obregón, Sonora, Mexico
[b] Centro de Investigación en Alimentación y Desarrollo, A. C. Km 0.6 Carretera a La Victoria, Hermosillo, Sonora, |Mexico

ARTICLE INFO

Article history:
Received 16 November 2015
Received in revised form 22 January 2016
Accepted 22 January 2016
Available online 23 January 2016

Keywords:
16S metagenomics
Classification methods
Kraken applications
Long 16S rDNA sequences
Metagenomic sequence classification
Taxonomic classification

ABSTRACT

Ultrafast-metagenomic sequence classification using exact alignments (Kraken) is a novel approach to classify 16S rDNA sequences. The classifier is based on mapping short sequences to the lowest ancestor and performing alignments to form subtrees with specific weights in each taxon node. This study aimed to evaluate the classification performance of Kraken with long 16S rDNA random environmental sequences produced by cloning and then Sanger sequenced. A total of 480 clones were isolated and expanded, and 264 of these clones formed contigs (1352 ± 153 bp). The same sequences were analyzed using the Ribosomal Database Project (RDP) classifier. Deeper classification performance was achieved by Kraken than by the RDP: 73% of the contigs were classified up to the species or variety levels, whereas 67% of these contigs were classified no further than the genus level by the RDP. The results also demonstrated that unassembled sequences analyzed by Kraken provide similar or inclusively deeper information. Moreover, sequences that did not form contigs, which are usually discarded by other programs, provided meaningful information when analyzed by Kraken. Finally, it appears that the assembly step for Sanger sequences can be eliminated when using Kraken. Kraken cumulates the information of both sequence senses, providing additional elements for the classification. In conclusion, the results demonstrate that Kraken is an excellent choice for use in the taxonomic assignment of sequences obtained by Sanger sequencing or based on third generation sequencing, of which the main goal is to generate larger sequences.

1. Introduction

The study of genomic sequences directly obtained from a given environment, also known as “metagenomics”, has become a relevant genomic sub-discipline during the last several years. Metagenomics is a useful tool that allows the scientific community to imagine, hypothesize and elucidate a picture of the microbial diversity and dynamics without the need to perform microbial cultures in artificial media (Martínez-Porchas and Vargas-Albores, 2015). This approach, together with new sequencing technologies that generate massive amounts of data, promises to be a revolutionary technique for the study of microbial diversity and dynamics in almost any environment.

New programs based on creative algorithms have been developed to study the sequences elucidated by metagenomics. However, according to Wood and Salzberg (2014), the current programs to analyze the environmental DNA libraries are “*relatively slow and computationally expensive, forcing researchers to use faster abundance estimation programs which only classify small subsets of metagenomic data.*” Sensitivity has to be sacrificed to perform faster analyses or vice versa, putting researchers at a crossroads because they must make a choice depending on what is most important to them (even if both issues have paramount importance). Considering this difficulty, new taxonomic sequence classification systems have to be developed or the current systems must be improved.

Kraken is a novel computational program designed for the analysis of metagenomic DNA sequences and has been widely applied showing impressive promise as an analysis tool. This program is an ultrafast and highly accurate tool that assigns taxonomic labels to noisy and complex sets of DNA sequences. This taxonomic label assignment is an important part of metagenomic analysis, considering that the content of environmental samples is largely unknown at the time of sequencing. The sensitivity of Kraken is comparable to similar state-of-the-art tools, such as MetaPhlAn and MEGABLAST. Kraken is faster than any other program. It can perform more than 1.3 million reads min^{-1} at standard operation mode and 4.1 million reads min^{-1} at quick operation mode, i.e., 11-fold faster than its nearest competitor MetaPhlAn (370,770 reads min^{-1}) and more than 0.5 million-fold faster than the most sensible program at genus level (Naïve Bayes classifier; 7 reads min^{-1}) (Wood and Salzberg, 2014).

The development of Kraken could be regarded as a revolutionary tool for metagenomics because it can be used for the analysis of any

[*] Corresponding author.
E-mail address: fvalbores@ciad.mx (F. Vargas-Albores).

http://dx.doi.org/10.1016/j.mimet.2016.01.011
0167-7012/© 2016 Elsevier B.V. All rights reserved.

hypervariable region. Although particular regions have been established to overcome the classification problem of certain bacterial strains, such a strategy may not be adequate for random environmental samples because one or two hypervariable regions can provide useful information regarding a particular species that cannot be obtained from other regions. For instance, Chakravorty et al. (2007) demonstrated that region V1 was useful to differentiate among *Staphylococcus aureus* and coagulase-negative *Staphylococcus* sp. The authors also asserted that V2 and V3 contained better information for distinguishing all bacterial species to the genus level, except for the closely related Enterobacteriaceae. V2 is the best region to distinguish among *Mycobacterium* species, and V3 is the best region for *Haemophilus* species. In addition, the authors determined that V4, V5, V7 and V8 were less useful targets for genus or species-specific probes. This discrepancy could be reduced using Kraken because it use k-mers and makes multiple comparisons of single or assembled k-mers against all regions of the 16S gene.

Although Kraken appears to offer many advantages for the study of long sequence, the program and its corresponding algorithms have been designed for assigning taxonomic identity to k-mers or short DNA sequences, as shown in Fig. 1 of the Wood and Salzberg (2014) manuscript. This design is compatible with the next generation sequencing systems (NGS) which perform massive but short reads (Jeon et al., 2015); average from 200 to 600 bp, except for PacBio RS2 equipment that performs reads up to 1.4 kbp (Mosher et al., 2014), and the new applications should adjust to the output data of the emergent technologies. However, these technologies are expected to be capable of performing larger reads in the next several years.

The Kraken application has been designed to work with short sequences, i.e., the more sequences that are identified, the deeper the classification. Furthermore, best results are expected with longer sequence due to the availabilities of more k-mers, where Krakren can make a more precise classification, especially if these are consecutive k-mers. Therefore, the following question has been formulated: *What would the precision and sensitivity of Kraken be when using long sequence inputs*? This question leads to the following hypothesis: *The input of larger metagenomic DNA sequences into Kraken provides more accurate and deeper taxonomic classifications of environmental samples*. If this hypothesis can be proven, the potential of Kraken as a taxonomic classifier would be even greater, reaching new horizons in metagenomics.

2. Materials and methods

2.1. DNA extraction

Total DNA was extracted from a marine environmental sample by following the instructions of the commercial kit DNeasy Blood & Tissue Kit (Qiagen, USA). The purified DNA samples were then stored at −20 °C for further analyses.

2.2. 16S rRNA amplification

The amplification of almost all of the region coding for 16S rRNA structural genes (16S rDNA) was performed using the universal primers 27F (5′-AGAGTTTGATCMTGGCTC-3′; M = A or C) (Lane, 1991) and 1387R (5′-GGGCGGWGTGTACAAGGC-3′; W = A or T) (Marchesi et al., 1998) under the following thermal cycling conditions: one step at 94 °C for 1 min; 30 cycles of 94 °C for 1 min, 55 °C for 1 min and 72 °C for 2 min; and an extension cycle at 72 °C for 10 min. The amplification was confirmed by electrophoresing the PCR products in 1% agarose gels stained with ethidium bromide.

2.3. Cloning

The remaining primers, dNTPs, enzymes and possible short-failed PCR products were eliminated from the PCR amplicons using a PureLink kit (Invitrogen, USA). The purified products were then inserted into the cloning vector pGEM®-T Easy Vector (Promega, USA), and the successfully linked products were transformed into competent *Escherichia coli* cells, JM109 (Promega, USA). These cells were cultured in plates containing LB agar supplemented with ampicillin/X-gal/IPTG and were incubated at 37 °C for 24 h. Thereafter, the white colonies corresponding to colonies that had successfully assimilated the insert were selected and cultured in LB/ampicillin broth at 37 °C for 24 h with constant agitation (Sambrook and Russell, 2001).

Plasmid DNA was extracted following the specifications of the GenElute™ Plasmid Miniprep Kit (Sigma-Aldrich, USA). Thereafter, the concentration of plasmid DNA was adjusted for further sequencing. Samples were sent to the University of Arizona to be sequenced by Sanger in both directions using the T7 and SP6 primers that correspond to the insertion site of the cloning vector.

2.4. Sequence analysis

Sequences were debugged from their respective vector by using the software VecScreen (NCBI). Thereafter, each pair of sequences was assembled by the CAP3 Sequence Assembly Program (Huang and Madan, 1999). The resulting elongated sequences were introduced into the Ribosomal Database Project classifier (RDP classifier; (Wang et al., 2007)) and into the Kraken classifier installed as Illumina BaseSpace app (https://basespace.illumina.com/apps). For the RDP classifier, the sequences were submitted in FASTA format and only results with a confidence threshold ≥80% were considered. In other analysis, the same sequences were not previously assembled and were introduced as paired samples in the analysis by Kraken. Labels and quality values in Illumina format 1.8 (Phred + 33) were assigned to the sequences to operate Kraken.

Some sets (sense, antisense) of Sanger sequences were unable to form contigs due to their low quality or small size. These sets were also submitted to the Kraken program as paired samples. Sequences that could be read at only one end were also submitted to the taxonomic classifiers.

2.5. Taxonomic placement consistency vs. formed operational taxonomic units

The groups of sequences used for both the RDP and KRAKEN were aligned using the K-align software (Lassmann et al., 2009; Lassmann and Sonnhammer, 2005) located in the European Bioinformatics Institute portal (http://www.ebi.ac.uk/Tools/msa/kalign/). Thereafter, cluster comparisons were performed by MEGA6 (Tamura et al., 2013) considering the minimum evolution approach and analyzing the data

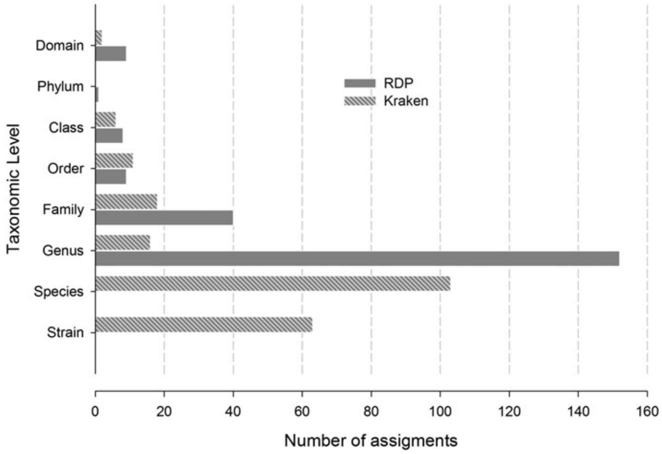

Fig. 1. Taxonomic level (Domain; Phylum; Class; Order, Family; Genus; Species; Strain (any level higher than species)) assignment of the 264 clones using the RDP and Kraken classifiers.