Title: LVQ-KNN: Composition-based DNA/RNA binning of short nucleotide sequences utilizing a prototype-based k-nearest neighbor approach

Authors: Ariane Belka, Mareike Fischer, Anne Pohlmann, Martin Beer, Dirk Höper

PII: S0168-1702(18)30384-8
DOI: https://doi.org/10.1016/j.virusres.2018.10.002
Reference: VIRUS 97498

To appear in: *Virus Research*

Received date: 29-6-2018
Revised date: 25-9-2018
Accepted date: 2-10-2018

LVQ-KNN: Composition-based DNA/RNA binning of short nucleotide sequences utilizing a prototype-based k-nearest neighbor approach

Ariane Belka[a], Mareike Fischer[b], Anne Pohlmann[a], Martin Beer[a] and Dirk Höper[a,*]

[a]Institute of Diagnostic Virology, Friedrich-Loeffler-Institut, Südufer 10, D-17493 Greifswald - Insel Riems, Germany
[b]Institute for Mathematics & Computer Science, Ernst-Moritz-Arndt University, Walther-Rathenau-Straße 47, D-17489 Greifswald, Germany

* dirk.hoeper@fli.de, to whom correspondence should be addressed.

Highlights

- LVQ-KNN bins sequences based on their oligonucleotide composition
- LVQ-KNN bins sequences derived from functional DNA and RNA molecules into DNA/RNA
- oligonucleotide frequencies differentiate functional DNA and RNA, e.g. viral genomes

Abstract

Unbiased sequencing is an upcoming method to gain information of the microbiome in a sample and for the detection of unrecognized pathogens. There are many software tools for a taxonomic classification of such metagenomics datasets available. Numerous of them have a satisfactory sensitivity and specificity for known organisms, but they fail if the sample contains unknown organisms, which cannot be detected by similarity-based classification employing available databases. However, recognition of unknowns is especially important for the detection of newly emerging pathogens, which are often RNA viruses. Here we present the composition-based analysis tool LVQ-KNN for binning unclassified nucleotide sequence reads into their provenance classes DNA or RNA. With a 5-fold cross-validation, LVQ-KNN reached correct classification rates (CCR) of up to 99.9% for the classification into DNA/RNA. Real datasets gained CCRs of up to 94.5%. Comparing the method to another composition-based analysis tool, similar or better classification results were reached. LVQ-KNN is a new tool for DNA/RNA classification of sequence reads from unbiased sequencing approaches that could be applicable for the detection of yet unknown RNA viruses in metagenomic samples. The source-code, training and test data for LVQ-KNN is available at Github (https://github.com/ab1989/LVQ-KNN).

Keywords: composition-based analysis, oligonucleotides, metagenomics, learning vector quantization algorithm, k-nearest neighbor method, cross validation

1. Introduction

Metagenomics is the challenge of analyzing the community of organisms in a sample using unbiased genomic techniques like next-generation sequencing (NGS) bypassing the need of lab cultivation and isolation of individual species (Chen and Pachter, 2005). Because of the new, fast and cost-effective sequencing methods, a huge amount of sequence data is