Bioinformatics, 2017, 1–2
doi: 10.1093/bioinformatics/btx385
Advance Access Publication Date: 15 June 2017
Applications Note

OXFORD

Sequence analysis

kmerPyramid: an interactive visualization tool for nucleobase and *k*-mer frequencies

Jochen Kruppa[1], Erhard van der Vries[2], Wendy K. Jo[2], Alexander Postel[3], Paul Becher[3], Albert Osterhaus[2] and Klaus Jung[1,*]

[1]Institute for Animal Breeding and Genetics, [2]Research Center for Emerging Infections and Zoonoses (RIZ) and [3]Department of Infectious Diseases, Institute of Virology, University of Veterinary Medicine Hannover, Hannover, Germany

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on February 28, 2017; revised on June 7, 2017; editorial decision on June 8, 2017; accepted on June 12, 2017

Abstract

Summary: Bioinformatics methods often incorporate the frequency distribution of nulecobases or *k*-mers in DNA or RNA sequences, for example as part of metagenomic or phylogenetic analysis. Because the frequency matrix with sequences in the rows and nucleobases in the columns is multi-dimensional it is hard to visualize. We present the R-package 'kmerPyramid' that allows to display each sequence, based on its nucleobase or *k*-mer distribution projected to the space of principal components, as a point within a 3-dimensional, interactive pyramid. Using the computer mouse, the user can turn the pyramid's axes, zoom in and out and identify individual points. Additionally, the package provides the *k*-mer frequency matrices of about 2000 bacteria and 5000 virus reference sequences calculated from the NCBI RefSeq genbank. The 'kmerPyramid' can particularly be used for visualization of intra- and inter species differences.
Availability and implementation: The R-package 'kmerPyramid' is available from the GitHub website at https://github.com/jkruppa/kmerPyramid.
Contact: klaus.jung@tiho-hannover.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Many methods in bioinformatics for the analysis of DNA and RNA sequences make use of the frequency distribution of the four nucleic bases or higher *k*-mers. For example, the frequency distribution plays a role in metagenomic binning of sequence reads (Dodsworth *et al.*, 2013; Imelfort *et al.*, 2014) and phylogenetic analysis (Podar *et al.*, 2013). Reference sequences of multiple species are available from public data bases or are generated for example by next-generation sequencing (NGS). Frequency matrices of *k*-mer distributions derived from sequences are often multi-dimensional and thus hard to visualize.

Here, we present the R package 'kmerPyramid' as an interactive visualization tool that can be used for visualization of clustering results, of comparisons between genomic regions, of horizontal gene transfer as well as the display of low complexity regions. The kmerPyramid is based on principal component analysis (PCA) that is used to project the multi-dimensional matrix of nucleobase and *k*-mer frequencies in the 3-dimensional space. PCA, as a method for dimension reduction, has already been demonstrated to preserve relevant information when exploring *k*-mer frequencies (Dodsworth *et al.*, 2013; Imelfort *et al.*, 2014; Podar *et al.*, 2013). Our package provides a more comfortable environment for exploring the projected frequency data and allows the user to compare the frequency of his sequences with frequencies of thousands of viral and bacterial reference sequences. For further background we refer to the Supplementary Data.

2 Functionality and examples of application

In total, the package comprises two main functions, several utility functions and four datasets. The main function `pyramid_3d()` performs the PCA and plots the projected data of the first three

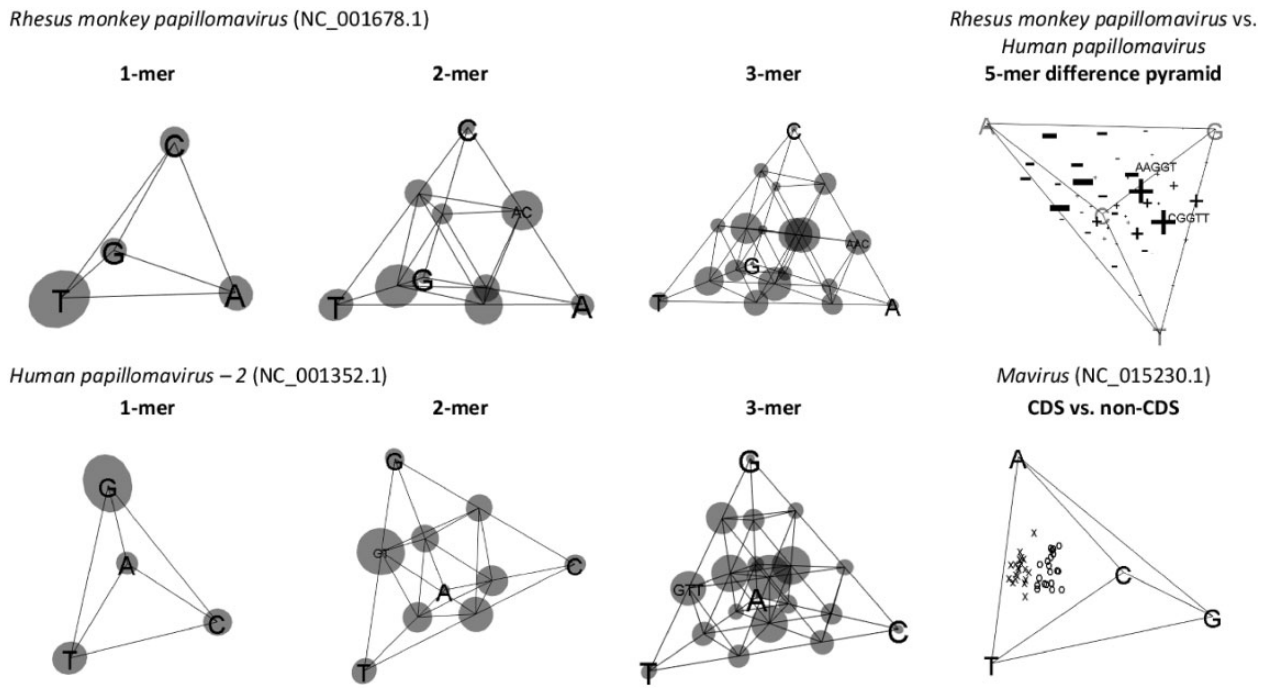

Fig. 1. First three columns: Grid based variant of the kmerPyramid showing 1 to 3-mer frequency distribution of two Papilloma virus strains coming from Rhesus monkey and Human. Each grid point represents all permutations of one *k*-mer, and the frequency of these permutations in the whole sequence of this *k*-mer is represented by the size of the bubble at this grid point. Top right: 5-mer difference pyramid between the Rhesus and Human Papilloma virus. A (+)-symbol indicates an increase of the specific 5-mer, while a (–)-symbol indicates a decrease. The size of the symbols correlates with the size of differences. Bottom right: 1-mer distribution of coding (o) and non-coding regions (x) of the Mavirus

principal components in an interactive plot based on its individual nucleobase frequency distribution. As main argument, the raw ($n \times 4^k$) data matrix X must be provided, representing the frequencies of the k-mers for n sequences. Functionality to estimate the k-mer frequencies from sequences is offered, too. In addition, a list of colors and labels can be provided. This pyramid plot can then be turned in each direction, the user can zoom into the point cloud and identify individual points by clicking onto them. The second main function `pyramid_3d_grid()` allows to estimate the k-mer frequencies for a fix k using a sliding window approach, and to plot the projected frequency data in a grid sectioned pyramid.

Since nucleobase frequencies play an important role in the analysis of viruses and bacteria, we determined frequency matrices for ca. 2000 bacteria and 5000 viruses from sequences retrieved from the NCBI genbank (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank) including species and taxonomic information. The resulting frequency matrices were also added to the package. In addition, two small example datasets including coding and non-coding sequences from the Mavirus and ten exemplary virus sequences are provided.

The kmerPyramid can be used in different fields of application. If the user is interested to visualize bases or k-mers that are very frequent in a sequence, the grid based version of the pyramid can be used. The first three columns of Figure 1 show the grid based pyramid of two papillomavirus sequences (human and monkey) for $k = 1, 2, 3$. For $k = 5$, the permutation frequency differences between the two papillomavirus sequences are displayed in another pyramid (top right plot). Those k-mers that are more/less frequent in the human are labelled by a +/– symbol. Thus, the ordered CGGTT is more frequent in the human strain and the ordered AAGGT in the monkey strain. Bottom right of Figure 1, each point in the basic ACGT-pyramid represents either a coding (o) or a non-coding (x) region of the Mavirus. The user can see that the coding regions have a larger GC-content. A more detailed methods description and a comprehensive list of applications is presented in the Supplementary Material.

3 Summary

The kmerPyramid package, as a 4-dimensional coordinate system, is a useful tool for visualizing and exploring word frequencies in DNA and RNA sequences. It can be used for visualizing a variety of biological aspects on the sequence level. The two variants of the pyramid can be used to either display the relation of multiple sequences to each other based on their nucleobase frequencies or to visualize the frequency of k-mers within a sequence. The grid based variant can also be used to compare k-mer frequencies between two sequences. Further conclusions and remarks are given in the Supplementary Material.

Funding

This work was supported by the Niedersachsen-Research Network on Neuroinfectiology (N-RENNT) of the Ministry of Science and Culture of Lower Saxony.

Conflict of Interest: none declared.

References

Dodsworth,J.A. *et al.* (2013) Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat. Commun.*, **4**, 1854.

Imelfort,M. *et al.* (2014) GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, **2**, e603.

Podar,M. *et al.* (2013) Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol. Direct*, **8**, 9.