Shibuya *et al.*
Algorithms for Molecular Biology (2022) 17:5
https://doi.org/10.1186/s13015-022-00212-0

Algorithms for Molecular Biology

RESEARCH

Open Access

Space-efficient representation of genomic k-mer count tables

Yoshihiro Shibuya[1], Djamal Belazzougui[2] and Gregory Kucherov[1,3*]

Abstract

Motivation: *k*-mer counting is a common task in bioinformatic pipelines, with many dedicated tools available. Many of these tools produce in output *k*-mer count tables containing both *k*-mers and counts, easily reaching tens of GB. Furthermore, such tables do not support efficient random-access queries in general.

Results: In this work, we design an efficient representation of *k*-mer count tables supporting fast random-access queries. We propose to apply Compressed Static Functions (CSFs), with space proportional to the empirical zero-order entropy of the counts. For very skewed distributions, like those of *k*-mer counts in whole genomes, the only currently available implementation of CSFs does not provide a compact enough representation. By adding a Bloom filter to a CSF we obtain a Bloom-enhanced CSF (BCSF) effectively overcoming this limitation. Furthermore, by combining BCSFs with minimizer-based bucketing of *k*-mers, we build even smaller representations breaking the empirical entropy lower bound, for large enough *k*. We also extend these representations to the approximate case, gaining additional space. We experimentally validate these techniques on *k*-mer count tables of whole genomes (*E. Coli* and *C. Elegans*) and unassembled reads, as well as on *k*-mer document frequency tables for 29 *E. Coli* genomes. In the case of exact counts, our representation takes about a half of the space of the empirical entropy, for large enough *k*'s.

Keywords: *k*-mers, Counts, Compression, Compressed static function, Bloom filter

Background

Nowadays, many bioinformatics pipelines rely on *k*-mers to perform a multitude of different tasks. Representing sequences as sets of words of length k generally leads to more time-efficient algorithms than relying on traditional alignments. For these reasons, alignment-free algorithms have started to replace their alignment-based counterparts in a wide range of practical applications, from sequence comparison and phylogenetic reconstruction [1–4] to finding SNPs [5, 6] and other tasks. These algorithms often require to associate some kind of information to *k*-mers involved in the analysis, that is, to build maps where keys are *k*-mers. Typical values to associate to *k*-mers are their frequencies in a particular dataset. Actual counting can be performed by one of several available *k*-mer counting tools developed in recent years [7–10]. Count tables generally include both *k*-mers and counts requiring considerable amounts of disk space to be stored. For example, the output generated by KMC [7] for a human genome, with $k = 32$ weights in at around 28GB.

In many applications, space can be significantly reduced by representing the mapping without actually storing *k*-mers. Having two independent data structures allows for more aggressive space optimizations. For example, the original sequence dataset can be used as the primary source of *k*-mers while a random-access data structure will then allow retrieving their counts efficiently. One application of such a data structure is the efficient representation of *k*-mer counts for read correction [11]. More generally, information about *k*-mer counts is increasingly

*Correspondence: gregory.kucherov@univ-eiffel.fr
[1] LIGM, Université Gustave Eiffel, Marne-la-Vallée, France
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.