Bioinformatics, 37(19), 2021, 3349–3350
doi: 10.1093/bioinformatics/btab196
Advance Access Publication Date: 23 March 2021
Applications Note

Genome analysis

KEC: unique sequence search by K-mer exclusion

Pavel Beran [1,*], Dagmar Stehlíková[1], Stephen P. Cohen[2] and Vladislav Čurn[1]

[1]Department of Genetics and Agricultural Biotechnology, Biotechnological Centre, University of South Bohemia, Faculty of Agriculture, 37005 České Budějovice, Czech Republic and [2]Department of Plant Pathology, The Ohio State University, Columbus, OH 43210, USA

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on December 3, 2020; revised on February 17, 2021; editorial decision on March 14, 2021; accepted on March 22, 2021

Abstract

Summary: Searching for amino acid or nucleic acid sequences unique to one organism may be challenging depending on size of the available datasets. K-mer elimination by cross-reference (KEC) allows users to quickly and easily find unique sequences by providing target and non-target sequences. Due to its speed, it can be used for datasets of genomic size and can be run on desktop or laptop computers with modest specifications.

Availability and implementation: KEC is freely available for non-commercial purposes. Source code and executable binary files compiled for Linux, Mac and Windows can be downloaded from https://github.com/berybox/KEC.

Contact: beranp02@jcu.cz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the availability of modern online databases and sufficiently inexpensive high-throughput sequencing, the accumulation of large amounts of DNA sequence data of many organisms is trivial. Using these data for subsequent analysis, however, usually requires high-performance computers and bioinformatically skilled personnel. Even routine PCR primer design, if based on genomic data, is mostly conducted using several Linux-based tools for sequence trimming, assembly, alignment, filtering, etc.—a process which requires considerable time and experience (Larrea-Sarmiento et al., 2018).

In this work, we present a new software application, K-mer elimination by cross-reference (KEC), which allows for a simple search of unique sequences in large datasets. The main advantages of the software are its speed, and ease of use. The output provided is useful for subsequent detailed analyses like primer design, comparative genomics, motif finding, etc. KEC is a command-line application that is available for all major platforms (Linux, Mac, Windows) as a stand-alone executable binary file without any dependencies.

2 Software description

KEC is written in the Go programming language which allows for compilation on all major modern operating systems.

The original motivation for KEC was to find sufficiently unique DNA sequences within bacterial genomes for diagnostic PCR primer design, which requires the identification of unique sequences among many closely related bacterial species. After preliminary testing, it was determined that KEC can quickly and efficiently process larger eukaryotic genomes as well. Moreover, the software can be useful for other genome, transcriptome or proteome related tasks. In the case of primer design, other tools or pipelines like UniqPrimer (Karim et al., 2019) are available. Those, however, use a different approach and only the specific purpose of primer design. Comparing DNA K-mers is also used to solve specific tasks like microbial identification (Panyukov et al., 2017). Our approach is unique and has a wide range of applications beyond these singular tasks.

The strategy used for the unique sequence search is based on the creation of K-mers, with K being a user-specified parameter, of target and non-target sequences. The K-mers are stored in native Go maps, a data type analogous to key/value dictionary in other programming languages (e.g. Python). In the second step, target and non-target maps are cross-referenced, producing a unique set of K-mers which are not present in the non-target map. Target specific K-mers are placed back to their original position within the original sequence to find regions of overlapping K-mers that are merged to create larger sequences (Fig. 1). The larger sequences have, on average, at least 1 out of K nucleotides unique to the target; e.g. if K is set to 12, then at least 1 nucleotide of 12 is different from non-target sequences through the whole target. Users can specify the minimum length of unique larger sequences returned by KEC.

Alternatively, KEC allows the cross-referencing in step two to keep, rather than exclude, the K-mers present in non-target, while excluding all target sequences not present in the non-target. In this case, the resulting merged sequences are those present in both target and non-target sequence pools.

Thanks to very high speed of Go maps, KEC operates very fast. So far, KEC has been successfully tested on bacterial genomes. With an average office computer or laptop, the time to compare one target bacterial genome to a non-target genome is usually under one

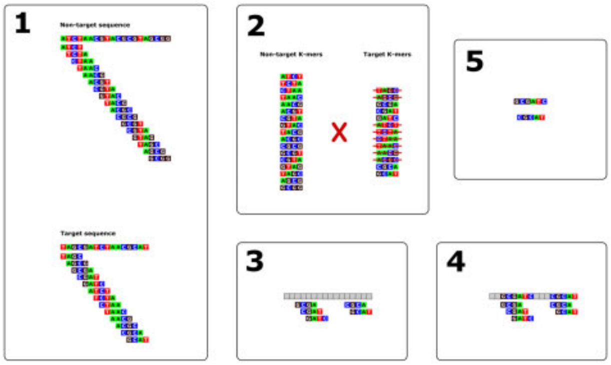

Fig. 1. KEC unique sequence search strategy. (1) K-mers are created based on target and non-target sequences. (2) Target and non-target K-mers are cross-referenced, target K-mers present in non-target pool are removed. (3) Surviving target K-mers are placed to the original position within original sequence. (4) Overlapping K-mers are merged into larger sequences. (5) Resulting larger sequences are recovered

second (around 500 ms with an average bacterial genome size of 5 Mb).

KEC does not aim to be all-in-one solution for any specific task. Instead, it provides a convenient, simple and fast alternative in established workflows. In the case of primer design, even a researcher without specialization in bioinformatics can move from short, single sequences (e.g. a single gene) to genomic sequences without the need to learn new operating system specifics or complex tools. More information about the software and a tutorial for a simple use case scenario is available as a Supplementary Material.

3 Conclusion

KEC is a new tool for searching for unique sequences in large datasets. It quickly compares K-mers of target and non-target sequences and returns sequences that are not present in the non-target sequences. The main advantages of the software are its speed, memory efficiency and ease of use. KEC enables researchers working on any major operating system to easily compare multiple sequences to find unique regions.

Acknowledgements

The authors thank Emily Luna, Prof. Jan Leach (both Colorado State University) and her lab members for providing selfless support and useful comments and ideas at multiple occasions.

Funding

This work was financially supported by the Ministry of Education, Youths, and Sports of the Czech Republic within INTER-EXCELLENCE/INTER-COST programme [LTC19014 (MSMT-15739/2019-8)]. This article was based on work from European Cooperation in Science and Technology (COST) action CA16107 EuroXanth, supported by COST. S.P.C. was supported by the AFRI Education and Workforce Development Postdoctoral Fellowship [2018-08122] from the U.S. Department of Agriculture, National Institute of Food and Agriculture.

Conflict of Interest: none declared.

References

Karim,S. *et al.* (2019) Development of the automated primer design workflow uniqprimer and diagnostic primers for the broad-host-range plant pathogen *Dickeya dianthicola*. *Plant Dis*., **103**, 2893–2902.

Larrea-Sarmiento,A. *et al.* (2018) Development of a genome-informed loop--mediated isothermal amplification assay for rapid and specific detection of Xanthomonas euvesicatoria. *Sci. Rep*., **8**, 14298.

Panyukov,V.V. *et al.* (2017) Short unique sequences in bacterial genomes as strain- and species-specific signatures. *Math. Biol. Bioinf*., **12**, 547–558