SeqBIM

Abstract

Phylo k-mers: constructing phylogenetically-informed k-mers for phylogenetic placement and recombination detection

Nikolai Romashchenko[1]*, Benjamin Linard[1,2], Fabio Pardi[1], Eric Rivals[1]

[1]*LIRMM, University of Montpellier, CNRS, Montpellier, France*
[2]*SPYGEN, 17 Rue du Lac Saint-André, 73370 Le Bourget-du-Lac, France*
Corresponding author: *nromashchenko@lirmm.fr

Abstract

Multiple sequence alignment is an essential preliminary step for a large number of different algorithms in bioinformatics. With the decrease of the sequencing cost, the need to process more and more data increases, making alignment-based approaches computationally expensive in practice. This led to the emergence of many alignment-free algorithms and the widespread adoption of k-mer based approaches.

We describe phylogenetically-informed k-mers, or *phylo k-mers*, a concept recently introduced in the context of phylogenetic placement in metagenomics [1], and successfully applied for viral recombination detection [2]. A phylo k-mer is a k-mer that is present with a non-negligible probability in unknown relatives of the sequences contained in an alignment. While the calculation of these probabilities is computationally heavy and requires the reference alignment as an input, it has to be done only once per alignment. Once calculated, phylo k-mers can be applied in alignment-free algorithms that require a massive input of new query sequences: in RAPPAS [1] for phylogenetic placement, and SHERPAS [2] for viral recombination detection.

We discuss methods of calculation, or *construction* of phylo k-mers, and present `xpas`, the phylo k-mer construction library. It allows for fast and memory-efficient construction of databases of phylo k-mers. Those databases are used by SHERPAS and the new version of RAPPAS, which is currently under development.

Keywords: *alignment-free*, *phylogenetics*, *k-mers*, *phylo k-mers*, *recombination*

References

[1] Benjamin Linard, Krister Swenson, and Fabio Pardi. Rapid alignment-free phylogenetic identification of metagenomic sequences. *Bioinformatics*, 35, 01 2019.

[2] Guillaume E. Scholz, Benjamin Linard, Nikolai Romashchenko, Eric Rivals, and Fabio Pardi. Rapid screening and detection of inter-type viral recombinants using phylo-k-mers. *bioRxiv*, 2020.