Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Sequence analysis

Radogest: random genome sampler for trees

Jacob S. Porter [1,*], Andrew S. Warren [1,*]

[1] Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, 22911, USA

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Radogest is a program for sampling k-mers by taxonomy from genomes. Radogest works with amino acid data, coding domain data, and whole genome data. Radogest currently supports downloading this data from the National Center for Biotechnology Information (NCBI). Radogest selects genomes to sample from with several strategies based on a post-order depth-first traversal of the NCBI taxonomic tree. It can hold out genomes based on species, and it can select genomes based on their quality. Radogest efficiently handles terabytes of data through parallelism. It can easily be deployed on a compute cluster for bioinformatics projects since it is written in Python3. One application of sampling k-mers labeled by their taxonomic id is to train metagenomics classifiers. The genome selection feature can be used to generate both 'hard' and 'easy' data for metagenomics classification.

Availability and Implementation: Available online: https://www.github.com/
Contact: jsporter@virginia.edu or aswarren@virginia.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A need for scalable data sampling methods is desired.

2 Features

Compare to CAMISIM **?**.

3 Methods

Opal **?**

4 Results

Text text

5 Conclusion

Text text

Acknowledgements

Text text

Funding

Text