Hierarchical Sets: Analyzing Pangenome Structure through Scalable Set Visualizations - DTU Orbit (11/12/2018)

Hierarchical Sets: Analyzing Pangenome Structure through Scalable Set Visualizations

The increase in available microbial genome sequences has resulted in an increase in the size of the pangenomes being analyzed. Current pangenome visualizations are not intended for the pangenome sizes possible today and new approaches are necessary in order to convert the increase in available information to increase in knowledge. As the pangenome data structure is essentially a collection of sets we explore the potential for scalable set visualization as a tool for pangenome analysis. We present a new hierarchical clustering algorithm based on set arithmetics that optimizes the intersection sizes along the branches. The intersection and union sizes along the hierarchy are visualized using a composite dendrogram and icicle plot, which, in pangenome context, shows the evolution of pangenome and core size along the evolutionary hierarchy. Outlying elements, i.e. elements whose presence pattern do not correspond with the hierarchy, can be visualized using hierarchical edge bundles. When applied to pangenome data this plot shows putative horizontal gene transfers between the genomes and can highlight relationships between genomes that is not represented by the hierarchy. We illustrate the utility of hierarchical sets by applying it to a pangenome based on 113 Escherichia and Shigella genomes and find it provides a powerful addition to pangenome analysis. The described clustering algorithm and visualizations are implemented in the hierarchicalSets R package available from CRAN (https://cran.r-project.org/web/packages/hierarchicalSets) CONTACT: Thomas Lin Pedersen (thomasp85@gmail.com) Supplementary information Supplementary data are available at Bioinformatics online.

General information
State: Published
Organisations: Center for Biological Sequence Analysis, Department of Biotechnology and Biomedicine
Contributors: Pedersen, T. L.
Number of pages: 9
Pages: 1604-1612
Publication date: 2017
Peer-reviewed: Yes

Publication information
Journal: Bioinformatics
Volume: 33
Issue number: 11
ISSN (Print): 1367-4803
Ratings:
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Scopus rating (2017): CiteScore 7.84
Web of Science (2017): Impact factor 5.481
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 2
Scopus rating (2016): CiteScore 6.42
Web of Science (2016): Impact factor 7.307
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 2
Scopus rating (2015): CiteScore 6.06
Web of Science (2015): Impact factor 5.766
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 2
Scopus rating (2014): CiteScore 5.5
Web of Science (2014): Impact factor 4.981
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 2
Scopus rating (2013): CiteScore 5.78
Web of Science (2013): Impact factor 4.621
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 2
Scopus rating (2012): CiteScore 6.73