PanFR

An interactive tool for the visualization of iFR Approximated Pan-Genomic De Bruijn Graphs

Shubhang Kulkarni | REU Mentor - Prof. Brendan Mumey | Purdue University | August 2017

Abstract

We propose the novel tool **PanFR**, which accomplishes an interactive visualization of pan-genomes via approximating their de Bruijn Graphs with the corresponding iFR (interesting Frequented Region) clusters as the nodes, and the paths which the genomic sequences take between these clusters as the edges.

The tool is split into two applications, the serverlet and browser GUI, allowing for collaborative pan-genome exploration, along with rich interactive visual queries for data analysis.

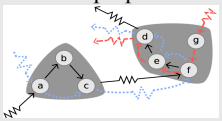
Introduction

- Pan-genomes represent collective genomic information of multiple organisms/individuals from a related group/species
- Representation via colored de Bruijn Graphs
- Recent rise in pan-genomic data has resulted in the increase in size and complexity of these graphs
- Visualizing these graphs becomes very challenging

Related Work

- Recent idea for a data filter, and thus graph simplification, using Frequented Regions (FRs) [1]
- FR – set of compressed de Bruijn Graph (cDBG) nodes traversed by many sequence paths *[see fig below]*
- Initially all cDBG nodes are singleton FRs.
- Merged in a bottom-up agglomerative fashion
- iFR – (interesting) FR with greater support than any ancestor in the tree(s) formed via mergers.
- Proposed use of iFRs for visualization purposes

A Frequented Region is a tuple (C,S)



Example Regions

Compressed Hierarchy Construction

The Algorithm:

- Clusters the iFRs into hierarchy groups [Union-Find]
- Reconstructs the hierarchies [exploiting iFR properties]
- Represents each hierarchy with single group node [initially the root]
- Adds paths through group nodes [with necessary bookkeeping]
- Supports dynamic interactive queries from the user*

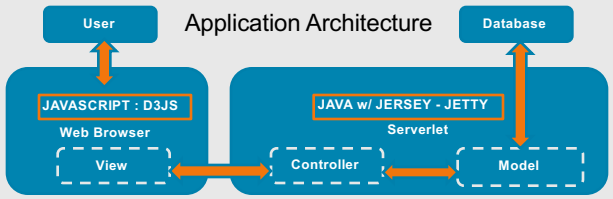
*Main Query – User Interaction (clicking) on a group node reveals all immediate children of the group node, which become the group nodes of their own sub-hierarchies

Visual Queries

Besides the compressed hierarchy construction algorithm, PanFRi supports various features for visualizing trends in the approximated pan-genomic data:

- Path Nodes & Edges Highlight [Fig 1, 2]
- Edge Filtration via Computed Scores
- Console Path Info Display [Fig 3]
- Console Node Info Display [Fig 3]
- Hierarchies Display [Fig 3]

Implementation



Results

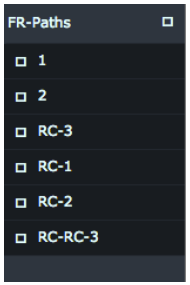


Fig 1

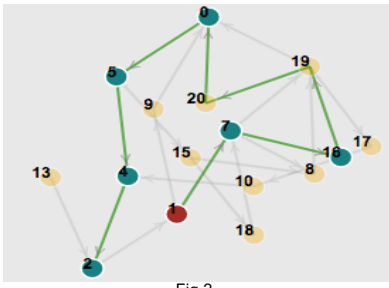


Fig 2

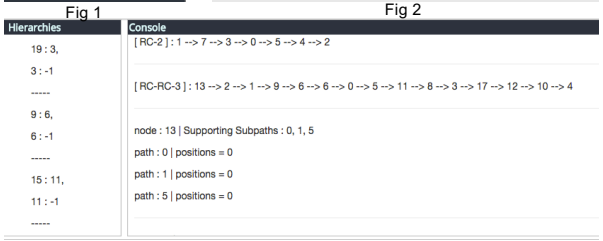


Fig 3

(Fig-1) A display of all the paths in the pan genome for E. Coli with appended reverse complement. (Fig-2) iFR approximated Graph displayed using the force layout. Path edges and nodes for RC-1 have been highlighted. Red node represents the start of the path. Nodes 19 and 20 are not highlighted as the path actually traverses through nodes 3 and 14, which fall in their respective hierarchies. (Fig-3) Shows the hierarchies and console display. Hierarchy of "-1" indicates leaf node. When a path [Fig-1] or a node[Fig-2] is clicked, the corresponding information is displayed on the console.
[Note – "RC" in the above examples stands for "Reverse Complement"]

Future Work

- Iterative addition of selected strains to visualization
- Private multi-user pan-genomic exploration support, so that a single server may support multiple explorations
- Implement node ordering for better layout

References

[1] A. Cleary, T. Ramaraj, I. Kahanda, J. Mudge, B. Mumey, Exploring Frequented Regions in Pan-Genomic Graphs, *ACM BCB 2017, Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017

[2]A. Cleary, T. Ramaraj, J. Mudge, B. Mumey, Approximate Frequent Subpath Mining Applied to Pangenomics, *BICoB 2017, International Conference on Bioinformatics and Computational Biology*, 2017.

MONTANA STATE UNIVERSITY