Visualizing the Longest Path Through an Assembly Graph

July 20, 2016

1 Background

DNA assembly programs generate *assembly graphs* as their output: graphs in which each node represents a *contig*, an unambiguously contiguous region of DNA determined from sequencing data, and each edge represents a possible overlap between two contigs. The challenge of the processes undertaken after assembly, then—referred to as *scaffolding* and, finally, *finishing*—are interpreting this data to determine the correct path through the contigs that represents the correct DNA sequence (*genome*) of the organism being sequenced.

Modern visualization software of assembly graphs, including Bandage (by Wick et al. 2015) and ABySS-Explorer (by Nielsen et al. 2009) takes the general approach of "showing everything at once": presenting all the information in an assembly graph to the user from the start. While broadly informative, this approach can be confusing and difficult to interpret for many users, particularly for genomes containing a massive amount of contigs (e.g. for those of eukaryotic organisms, or of metagenome assemblies).

The approach we plan to take is a bit different. Our focus is less on displaying all information at once, but on displaying the most important features: the *longest paths* of contigs through the graph, separated by connected components, with certain common patterns in sequencing data highlighted for the user to observe. These features, along with many others, will be contained in a fully interactive web-based visualization tool that we hope will provide a novel way of examining assembly graph data.

2 Longest Paths

The problem of finding the longest path through a directed graph is NP-Hard, provable by reduction from the Hamiltonian Cycle Problem.

3 Highlighting Patterns in Assembly Data

We highlight what Miller, Koren, and Sutton (2010) describe as "bubbles," "frayed ropes," and "spurs." (TODO – highlight cycles, also.)