Stephanie DeMane, sdemane

Simulator

1. If a genome is supplied, the sequencer simulator begins by determining the coverage c, mean read length l, and genome size G, which should be given by the command-line input. The standard deviation for the read lengths, which are normally distributed, is $\sigma^2 = 3$ (in accordance with the spec). From there, the total number of reads to be generated is calculated as $N = \frac{cG}{l}$. If k is not supplied, it is set to $l - 3\sigma^2$ or $l - 3\sigma^2 - 1$, whichever is odd. This will ensure that less than 1% of naturally generated read lengths are less than k.

2. The sequencer generates each read individually. First it samples a read length r from a normal distribution with mean l and standard deviation σ^2. Then it samples a starting position x from a uniform distribution with a domain of $[0, G)$. If $x + r > G$, then the sample is an edge case and is discarded. If $r < k$, then the sample is also discarded. Otherwise, the read is added to the list of reads.

3. Generate reads individually until you have N reads.

Assembler

1. Determine the list of reads and the value of k.

2. Take a read and separate it into all possible k-mers. Add the k-mers to a genome-wide set of k-mers, associated with the appropriate read and the starting position within the read.

3. Do the same for the reverse complement of the read. Store the lists of k-mers for the original read and its reverse complement as a pair.

4. Repeat 2 and 3 for all reads.

5. For each read, separate the forward lists of k-mers into chains: i.e., into smaller lists of adjacent k-mers which are part of the same set of reads.

6. Take each chain and generate the matching reverse-complement list and store it as a block. Map each block to both the first k-mer in the forward chain and the first k-mer in the reverse-complement chain.

7. Iterate through the blocks and record the possible connections between them, searching the blocks by initial k-mers relative to the preceding final k-mers. This forms a directed graph.

8. If any two blocks are connected by a forward edge only to each other, combine them.

9. Trace (an) Eulerian path(s) through the graph. From this, contigs can be assembled by taking the sequences of final characters of the k-mers along the path.