Basic Info

Genome Size: 5,227,294 bp (10,454,588 if we include 3'-5')
Given Query Size 100
Chosen Seed Size: 23
Nodes for 5'-3': 62,311,380.
Estimated trie size: 3.25 GB
Max recurrences of Seeds: 13

Basic Idea

This is fundamentally a BLAST algorithm that is going to be forced to be exhaustive. It will take a genome (G) and break it down into seeds of size 23 (explanation for seed size is below). It then creates a prefix trie from the seeds - recording the location in the genome for each seed in the trie. The routine then executes each query as follows:

1. Initialize a bitarray that is $G - 23 + 1$ long.
2. Break query into seeds of size 23
3. Search trie with seeds. If seed is in trie, flip the bits in the bit array associated with the seed locations.
4. When all query seeds have been checked against the trie, examine a moving total (of the same length as number of seeds in query) of the bitarray.
5. Check location with largest total using Smith-Waterman. Keep alignment is it has 3 or less SNPs (single nucleotide polymorphisms).
6. If no alignment found in previous step, decrease moving total by one and repeat last step. Stop when moving total less than or equal to 9 (See discussion below about exhaustive criteria).

Exhaustiveness