EN 600.439/639: Computational Genomics

Homework 5

Fall 2014, Prof. Ben Langmead

Use Piazza when asking for clarifications or reporting typos. All the usual programming and submission guidelines apply, per the class announcements on the Piazza site. There are a total of 35 points.

1. (3 pts) We have visualized the De Bruijn graph in two similar ways:

   (a) A directed multigraph with one node per distinct $(k-1)$-mer and one edge per k-mer.

   (b) A directed graph with one node per distinct $(k-1)$-mer and one labeled edge per distinct k-mer. The edge label equals the number of times the k-mer occurs.

   The representations are equivalent, so why is the second representation useful? Your answer should discuss the big-O space bounds of the De Bruijn graph assuming error-free reads. Let N be the total length of all the reads and let G be the length of the genome from which the reads are drawn.

2. (2 pts) Estimate the fraction of all possible DNA 22-mers that are present in the human genome. Write the answer in scientific notation (not as a fraction). Assume the genome is 3 billion nucleotides long and that all 22-mers in the genome are distinct.

3. We discussed error correction in class.

   (a) (3 pts) Why does low average coverage impede our ability to correct errors?

   (b) (2 pts) Is it possible for a sequencing error to mutate a k-mer into a k-mer that occurs *more often* in the input than the original k-mer?

4. (20 pts) Solve the problem here: `http://bit.ly/CG_Frankengenome`.

   (a) Submit your code, along with a README describing how we should run it.

   (b) Submit a file with labels for every nucleotide of the Frankengenome, formatted as discussed on the problem page.

5. (5 pts) Do rosalind.info problem "Open Reading Frames":

   `http://rosalind.info/problems/orf/`