IITM-CS6024 : Algorithmic Approaches to Computational Biology — Release Date: Sep 19

Assignment 1 — Due Date: Sep 29, 23:59

Roll No: NGSEQ454 — Name: Gatacca

- Use LaTeX to write-up your solutions, and submit the resulting single pdf file at GradeScope (Note: Within GradeScope, indicate the page number where your solution to each question starts, else we won't be able to grade it!).
- Collaboration is encouraged, but all write-ups must be done individually and independently, and mention your collaborator(s) if any. Same rules apply for codes uploaded to HackerRank (i.e., write your own code; we will run plagiarism checks on codes).
- If you have referred a book or any other online material for obtaining a solution, please cite the source. Again don't copy the source *as is* - you may use the source to understand the solution, but write-up the solution in your own words.

1. (6 points) [Choose your DS!]:

   We saw the algorithm below to construct an Eulerian cycle in a graph (with n nodes and m edges). What data structures would you choose to implement Lines 4,5,6 in the algorithm below, so that the overall algorithm has $O(m)$ running time complexity? Please assume that Graph is a balanced and strongly connected directed graph provided as an adjacency list.

```
1 EULERIAN_CYCLE(Graph)
2  form a cycle Cycle by randomly walking in Graph(don't revisit same edge
3   while there are unexplored edges in Graph
4     select a node newStart in Cycle with still unexplored edges
5     form a cycle Cycle′ (starting at newStart) & then randomly walking
6     Cycle <- merge(Cycle, Cycle′)
7   return Cycle
```

2. (6 points) [Counting ambiguity]:
   (a) (3 points) Find a simple DNA sequence whose k-mer composition agrees with that of exactly 5 other DNA sequences. Justify. (Hint: What would its de Bruijn graph look like? Optional: Do "articulation points" help identify contigs in this graph, or do you need "maximal non-branching paths" as mentioned in the book?)
   (b) (3 points) How many 3-universal circular binary strings are there? Justify. (Optional: How would you count the number of k-universal circular strings?)

3. (11 points) [Coding warmup]:
   (a) (8 points) Implement the clump finding algorithm in the HackerRank contest at `https://www.hackerrank.com/aacb-assignment-1`. (Note: Create your `https://www.hackerrank.com` account using your smail id, and ROLL NUMBER as your user-name. Languages supported are C, C++, Java, Python, and R (use only one)).

(b) (3 points) In addition, write here in this document the key data structure you used to implement the code, and associated time and space complexity of your implementation? Does you algorithm work for any value of L, t and k?

4. (11 points) [Assembling SARS-CoV-2/COVID-19 genome] A tiny virius SARS-CoV-2 ($\tilde{}100nm$ diameter per virion) has been causing such havoc in the way our world operates and communicates, and you are naturally curious to know what genome sequence it holds. Thankfully, we've algorithms to assemble the viral genome from a set of whole genome sequence (WGS) reads obtained using an Illumina sequencer (see attached **SRR12638317.sra_[12].fastq** files). You can use Velvet, which is one of a number of *de novo* assemblers that uses short read sets as input and constructs de Bruijn graphs for genome assembly, inside the Galaxy framework.

   You can use any of these Galaxy Servers that host Velvet Assembler.

   - `https://usegalaxy.org.au/`
   - `http://bf2i-galaxy.insa-lyon.fr:8080/`
   - `https://usegalaxy.eu/`

   Helpful Link: `https://galaxyproject.github.io/training-material/topics/assembly/tutorials/general-introduction/tutorial.html` for genome assembly workflow of a different data.

   (a) (2 points) What is the average number of wrong base calls in the last 5 positions of the 1^{st} read in SRR12638317.sra_2.fastq file? (Hint: Use quality scores provided by the Illumina 1.9 Next Generation Sequencer (NGS).)

   (b) (2 points) The first step in any sequence analysis is to run FastQC tool. Using its output, report the length of each read, and calculate the average coverage of the genome by all reads (given a guess that the length of the coronavirus is 31.7 kilo bases; note that average genome coverage is the average number of reads that span a nucleotide in the genome)?

   (c) (3 points) Run velvet with a k-mer size of 29, and report how many contigs have been built, and what the mean, min and max length of the contigs are?

   (d) (4 points) Rerun velvet with the following k-mer sizes: 23, 57, 100, and report for each case the above contig metrics in table format. Comment on the trade-off between small and large k-mer values, and which k value looks optimal to you.

5. (6 points) [Research warmup]: Provide properly-formatted references for papers in this solution.

   (a) (3 points) Browse through the latest issue of top-ranked bioinformatics or systems biology journals (e.g., Bioinformatics, PLoS Computational Biology, Cell Systems, Molecular Systems Biology, etc.). Which one article caught your attention/interest the most based on only the titles and/or abstracts, and why?

   (b) (3 points) What is the latest research publication you could find on *de novo* genome assembly (i.e., genome assembly as seen in class without the knowledge of a reference genome), and what key algorithm did it use?