CMPE 549 - Assignment 3

Spring 2022

Due 06.06.2022, 09:00

Description

In this assignment, we will be working on a genome sequencing problem and will be using a dataset provided by the Rosalind platform. Please download the dataset from this link.

Requirements

- You are expected to use `Python` programming language for this assignment.
- Make sure you have installed the `Biopython` package (version 1.79) for your `Python` interpreter.
- You are expected to submit a single `.ipynb` file (i.e. Jupyter notebook) that is runnable. Name your notebook with your student id (e.g. `2019400XXX.ipynb`). Note that this notebook will be your report as well, so explain your work in the related sections of the notebook.
- Clearly separate answers to different questions and subquestions with headings.
- Remember to use the markdown feature of the Jupyter notebook for your headings and textual answers.
- Late submission policy: You can submit until 11.06.2022, 09:00 with 20% penalty. After this, no submissions will be accepted.

Question

1. In the dataset provided for this question, k = 50 and d = 200, where d refers to the number of nucleotides between two read pairs (see the link in the Tips section below for more information). To simplify the problem, we assume perfect coverage and error-free reads. Do not use a third-party software for solving this question (except for trivial tasks such as holding the data in `numpy` arrays).
   - **a.** Load the read pairs from the data file (each read pair is in a separate line and each pair has been separated by a "|" sign), and remove the lines corresponding to the metadata (the first two lines). Afterwards, construct the paired de Bruijn graph from this data. Print the upper left 20 x 20 part of the de Bruijn adjacency matrix you constructed.
   - **b.** Find the unique Eulerian path in the de Bruijn graph and reconstruct the associated genome sequence accordingly. Print the length of the final reconstructed genome.
   - **c.** Print the first 200 and last 200 characters of your reconstructed genome string.
   - **d.** Instead of having N paired reads, imagine we only had the same reads as $2N$ independent reads, i.e. without knowing which ones are paired. How would we proceed in this case for genome reconstruction? What would be harder/easier in obtaining a final genome sequence? Please discuss.

Tips

To test your algorithm, you can use this toy data provided in the same web site: https://rosalind.info/problems/ba3j/ (This is only for your convenience, do not submit the assignment with only results on the toy data.) This link also can serve as a refresher for genome reconstruction with read-pairs.

If you cannot obtain reasonable results despite your best efforts on the original dataset, you may provide a discussion of why you think this is the case for partial credit.