Final Project

CSCI 5481, Computational Techniques for Genomics

Project overview

There are three project options. Each project should culminate in the submission of a short written report in the form of a scientific journal paper (see deliverables below), and in a short presentation on the last day of class.

This is intended to be a group-based project. Groups should be no more than 4 people. You may work on your own if you wish. Each member of a group must contribute something critical to the project and you must document what each person contributed.

Project option 1

Bring your own data set to analyze.

1. Use one or more techniques that we learned in the class to attempt to answer a biological question that is important to you.
2. Justify your choice of tools and your parameter settings.
3. Include an informative schematic diagram of the steps that you took, and at least 2 other visualizations.
4. For each analysis step that you perform, you must try several different settings of the parameters for the algorithm involved, and explain why the results change in the way that they do.

Project option 2

Download raw data for the COVID genome from the files section Canvas (here: https://canvas.umn.edu/courses/194169/files/folder/Final_Project), then:

1. Clean up the raw data using a quality trimmer/filtering tool like trimmomatic. Note that the raw data are FASTQ format, where lines 2, 6, 10, 14...are the sequences, and lines 4, 8, 12, 16… are the quality scores.
2. Correct sequencing errors. You can either count how many times each k-mer appears (using your *k* in step 3) and simply throw away k-mers that only appear once or twice, or run an external error-correction software program.
3. Break the remaining sequences/*k-mers* into a list of *k-mers* for some small *k* (*k*=30, perhaps). Treat read 1 and read 2 separately.
4. Create the De Bruijn graph using *(k-1)-mers* as nodes, in which an Eulerian path would traverse all *k-mers*. Visualize the graph with a graph visualization package (i.e. igraph in R).
5. Implement an algorithm to find an Eulerian path/circuit, and report the assembled genome (or all contigs if not complete). There will be certain issues that arise, such as extra edges that cannot be traversed, and duplicate edges. Document your decisions on how to deal with these.
6. For at least one assembled contig, align it to the reference genome (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/858/895/GCF_009858895.2_ASM985889v3/GCF_009858895.2_ASM985889v3_genomic.fna.gz) and plot the coordinates of the assembled location of each base against the known location of each base (i.e. a scatterplot where x is coordinate in the reference genome and y is coordinate in your assembled contig). If it doesn't align well as a single whole contig, break it into smaller sequences of length 100 and align each of those, then plot them all together.
7. Try several values of *k*; show and describe the effects of *k* on the final assembly.

Important: Try this first with a short section of 25-50 base pairs only with several different values of *k*. Then try it on larger and larger portions of the genome.

Project option 3

COVID phylogenies.

1. Find a source of assembled COVID-19 genomes that have some interesting metadata, such as timing or location of when the virus was found.
2. These could be many viruses from the same location over a long time, or perhaps many viruses from different locations around the world at the same time, or something else. Download at least 100 COVID genomes, and build phylogenies from them.
3. You will likely need to do a multiple sequence alignment first to calculate genetic distance. You can use external software. You can use external software (SeaView w/cd-hit or muscle as the aligner?) but you must explain your choice of parameters.
4. Try at least two different tree-building algorithms; compare their results visually; explain the differences as best you can based on how each algorithm works. You can use external software but you must explain your choice of parameters.
5. Use the resulting trees and other information you need to make a biological interpretation or conclusion from the data. For example, you could attempt to answer questions like, how fast is the virus mutating? How much community spread was there in a particular location versus travel-based spread bringing multiple unique genomes in from somewhere else? What was the original strain that seeded an outbreak in a particular location?

Deliverables

Presentation, due in class on Monday 12/14 (10%). Present a 5-minute presentation on your project. Each member of the group should participate.

Write-up, due by 5PM Wednesday 12/16 (90%). Please submit a description of your project formatted as a short research article of 1000-1200 words, containing the following:

1. **Abstract/Summary paragraph** (200 words)
Follow the structure of the example *Nature* summary paragraph here:
https://cbs.umn.edu/sites/cbs.umn.edu/files/public/downloads/Annotated_Nature_abstract.pdf

2. **Summary of previous findings** (200 words)
Describe the previous analysis and findings without critique.

3. **Results** (300-500 words)
In 1-2 paragraphs, describe what you did and what you found. It is important to consider and discuss alternative approaches that you could have used (or tried to use) and why your eventual choices were justified.

4. **Conclusion** (100 words)
Restate the purpose of your re-analysis. Summarize your findings. Briefly describe future work.

5. **Methods** (100-200 words)

Describe your analysis with enough detail that it could be reproduced by another researcher.

6. **Figures** (2-3 figures)
Include two or more figures supporting your findings.

7. **Acknowledgements** (Not graded)
Describe precisely what parts of the project were contributed by each member of the group.