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.

Project option 1

Bring your own data set to analyze. Use two or more techniques that we learned in the class to attempt to answer an important biological question. Justify your choice of tools and your parameter settings. Include an informative schematic diagram of the steps that you took, and 3 or more other visualizations. For each analysis step that you perform, be sure to try at least one alternative way to do the analysis to determine if the results still hold.

Project option 2

Download a chromosome from the human genome (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/). Simulate whole-genome shotgun data by breaking it into random overlapping substrings of length *k*. Break it down further into a list of *k-mers* for some small *k*. Create the De Bruijn graph using *k-mers* as nodes, where a Hamiltonian circuit would be needed to find a valid path through the genome. Then convert this to a *k-1*-dimensional De Bruijn graph in which an Eulerian circuit would traverse all *k-mers*. Implement an algorithm to find an Eulerian circuit, and report the assembled chromosome. Compare it to the original chromosome by plotting the coordinates of the assembled location of each base against the known location of each base.

Try this first with a short section of 25-50 base pairs with several different values of *k*. Generate plots of the *k*-dimensional and *k-1*-dimensional graphs. Then try it on larger and larger portions of the chromosome. Report the performance (fraction correctly assembled and runtime). Then try simulating sequencing error in the data (1 random error per 100 bases) and report the effects.

Project option 3

We have around 1 million paired-end RNA-Seq sequences generated from total RNA in two human fecal samples (500k sequences each). Data available here: https://www.dropbox.com/s/kxte655hgckwdon/Final-project-option-3-trimmed-fasta.zip?dl=1. The majority of the sequences are ribosomal RNA. Many are from Bacteria. The goals are to identify and characterize both known and possibly novel species in the data. You can try to follow these steps, or you can try another way to analyze the data, but please use at least two of the techniques that we have studied in class. Some suggested analyses. Assembly is highly recommended as the first step:

1. Use an existing assembly tool (e.g. SPAdes assembler or Meta-SPAdes) to assemble the reads into contigs. Paired-end reads will be extremely important to help resolve ambiguities. If possible, retain only high-confidence contigs for the next steps.
2. Use an existing alignment tool (BURST, BWA, Bowtie2) to identify sequences that match known ribosomal

sequences (See file “SILVA_123_SSURef_tax_silva_trunc.fasta.gz” here: https://www.arb-silva.de/no_cache/download/archive/release_132/Exports/). Report the species that were identified.

3. Choose a subset of sequences that do not match the database (say at 95% or above), and search them against all data in BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to determine if parts of them do match other ribosomes. Perform a global multiple sequence alignment of these in SeaView (http://doua.prabi.fr/software/seaview) along with a subset of the known sequences to visualize what regions are or are conserved in the known and unknown ribosomal sequences.
4. Choose a subset of sequences that do match the database and a subset that do not, and use RNA secondary structure prediction software to make 2D visualizations of the predicted secondary structure for the representatives of known and novel species for comparison. Do the “known” and “novel” species have the same secondary structure over sections of the sequences that should align?

Deliverables

Write-up, due by 10PM Wednesday 12/11 (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.

Presentation, due in class on Monday 12/9 (10%). Present a 5-minute presentation on your project. Each member of the group should participate.