Bioinformatics
Exercise sheet #2

Lena Morrill*

Lent 2018

Please submit your work by noon on the day before the supervision, either by email at lm687@cam.ac.uk, or leaving it in my pigeonhole in Clare College (Old Court).

Assembly of genomes

De novo assembly

1. Explain briefly how reads are sequenced using Next Generation Sequencing (e.g. Illumina)
2. What are k-mers, preffixes and suffixes? How do they relate to Hamiltonian and de Bruijn graphs?
3. What is an Eulerian cycle?
4. Use the sequence `ATTACGGTACCCCTACA` for the following two questions.
   (i) Construct its de Bruijin graph with $k = 3$.
   (ii) Construct the paired de Brujin graphs with $k = 3$ and $d = 1$.
5. Why do we assign the k-mers to the edges rather than the nodes?
6. What are the runtime complexities for Hamiltonian paths, and for Eulerian cycles?
7. Build the Hamiltonian and de Brujin graphs given the set of k-mers {`"ATC"`, `"TGG"`, `"GGC"`, `"GCG"`, `"CGT"`, `"GTG"`, `"TGC"`, `"GCA"`, `"CAA"`, `"AAT"`}.

*Some of the questions are based on those by Petar Veličković and Sebastian Müller.

BWT

8. What are suffix tries, suffix trees and suffix arrays?
9. Construct the suffix trie of the word `chitchat`. How do repeated substrings appear in the trie?
10. Construct the BWT of some 6-letter word. Show every step.
11. Implement the BWT and test it with the word above.
12. Invert the BWT for `nnmyeoid`, showing your reasoning.
13. How can you use the BWT for patttern matching?

Clustering

14. Outline the algorithm for k-Center and k-Means clustering. State the complexity for the latter. Implement either.
15. Heuristics for k-means:
    i) Describe the center of gravity of a set of n-dimensional points.
    ii) Explain the steps in Lloyd's algorithm.
    iii) How can you prevent Lloyd's algorithm from clustering your data incorrectly due to an unfortunate initialisation step?
16. Outline the Expectation-Maximisation (EM) algorithm.
17. What are soft and hard clustering? Why might we favour the former over the latter?
18. How does the EM algorithm relate to soft clustering?
19. Explain how hierarchical clustering organises the data points, and the common distance functions. Have you seen any sort of hierarchical clustering elsewhere in the course?
20. Explain the idea behind the Markov Clustering Algorithm, the random walk on the graph, and the parameters.