Bioinformatics

Example Sheet 2

Petar Veličković

Michaelmas Term 2016

Genome sequencing

1. From a high level, explain the problem of *genome sequencing*, and what are the given inputs and desirable outputs. What limitations prevent us from having more informative inputs?
2. Define a k-mer, a *prefix* and a *suffix* of a string within this context. How are these individual components used within the Hamiltonian and de Bruijn graphs?
3. What is a necessary condition for a graph to have a Eulerian cycle?
4. There exists an $O(n^2 2^n)$-time algorithm for computing Hamiltonian paths (where n is the number of nodes). Conversely, what is the computational complexity of the best-known algorithm for computing Eulerian cycles? Provide pseudocodes for both of those algorithms.
5. Build the Hamiltonian and de Bruijn graphs over the following set of k-mers: {`"ATG", "TGG", "GGC", "GCG", "CGT", "GTG", "TGC", "GCA", "CAA", "AAT"`}, and highlight the reconstructed genome.
6. Explain how to sequence a genome using de Bruijn graphs constructed from *paired reads*. What is the limitation of the previous approach to sequencing that this approach tries to overcome?
7. Outline the key incorrect assumptions that these approaches make, and how to fix two of them.

Clustering

1. What is the output of a typical *gene expression* experiment, and why might one wish to do further processing on such a result?
2. Define the inputs and outputs of the *k-means clustering* algorithm, and state its complexity class.

3. Outline the steps taken by *Lloyd's algorithm*, which attempts to circumvent the issue from the above. State its time complexity, and provide an informal proof of its convergence. How might we use it to find *"good"* approximations for the k-means clustering solution?

4. Implement Lloyd's algorithm in a language of your choice, and demonstrate that it works by applying it for $k = 4$ on an easily separable set of 2D points. You may, for example, generate the points as $C + (\varepsilon_x, \varepsilon_y)$, where $\varepsilon_x, \varepsilon_y \sim U(-2, 2)$ (where U is a uniform real distribution) and $C \in \{(0, 0), (0, 5), (5, 0), (5, 5)\}$. The choice of C should then determine the resulting cluster.

5. Explain the two high-level steps taken by the *expectation-maximisation* (EM) algorithm, and then show how it relates to *soft k-means clustering* (giving particular care to the *stiffness parameter*).

6. What is the time complexity of the *hierarchical clustering* algorithm when using the *minimum-distance* metric between clusters?

7. Explain the inputs, outputs, steps and time complexity of the Markov Clustering (MCL) algorithm.