Bioinformatics
Exercise sheet #1

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 pidgeonhole in Clare College (Old Court).

Background in genetics

1. Explain the structure of DNA (e.g. what nucleotides exist; how they are paired).
2. Do the same for RNA. Explain the biological function of the three largest groups of RNA (rRNA, tRNA, mRNA).
3. What do we mean when we say that the genetic code is degenerate? Put an example.
4. There are 64 possible codons, not all with the same role. What are these roles?
5. How are proteins created, and what is their structure?
6. Recap on 1–5: explain the Central Dogma of Biology (a diagram might be useful).

Sequence alignment

7. What do match, mismatch and gap situations correspond to in nature? (Explain it from the evolutionary point of view)
8. In which two ways could you penalise for gaps?

Dynamic Programming

9. Implement the Needleman-Wunsch algorithm for global alignment with your programming language of choice, aligning the sequences `PAWHEAE` and `HEAGAWGHEE`.[1] Parameters: check the BLOSUM50 matrix for match/mismatch scores, -8 for gaps.
10. What is the time and storage complexity of the N-W algorithm?
11. Whenever we are aligning subsequences, other methods are more fitting. How do the Smith-Waterman, the repeated matches and the overlap matches algorithms differ between each other and from N-W?
12. Would we be right to say that in a global alignment the first nucleotide of each sequence necessarily have to align, as well as the last ones?
13. *Optional*: How does the complexity change when you use affine gap penalties and how can we overcome this issue?

[1] This is an example shown in Chapter 2 of Biological Sequence Analysis, if you want to double-check your results.

Heuristic methods

14. The drawback of dynamic programming is time complexity, and here is where heuristic alignment algorithms – such as BLAST – come into play. Outline this algorithm.

Phylogenics

15. Implement the UPGMA algorithm. Show the successive clusters for the distance matrix below and draw the tree.

$$D = \begin{pmatrix} 0 & 3 & 4 & 4 & 4 \\ & 0 & 4 & 4 & 4 \\ & & 0 & 1 & 2 \\ & & & 0 & 2 \\ & & & & 0 \end{pmatrix}$$

16. How do the neighbour-joining and parsimony algorithms compare to each other and to UPGMA? For UPGMA *vs* neighbour-joining you can base your answer in the reconstruction of the tree corresponding to the matrix below.

$$D = \begin{pmatrix} 0 & 5 & 7 & 10 \\ & 0 & 4 & 7 \\ & & 0 & 5 \\ & & & 0 \end{pmatrix}$$

17. What is the difference between traditional parsimony and weighted parsimony, and what parameters render the two equivalent?

18. Myosins are a family of protein that allow the contraction of muscle. They are composed of chains – two of which are heavy and several others that are light.

    You can find 50 sequences of the heavy chain of a myosin of various species in the fasta file here.[2] Take a look at the format. Then feed it into Clustal Omega[3] and look at the multiple alignment and the phylogenetic tree.

[2] `https://www.ebi.ac.uk/Tools/examples/protein/sequence12.txt`
[3] `https://www.ebi.ac.uk/Tools/msa/clustalo/`