Bioinformatics Supervision

Michaelmas Term 2017

–Problem Sheet 1–

Supervisor: Sebastian Müller (Department of Plant Sciences)

Please hand in your work 24 hours prior to the supervision either to `sm934@cam.ac.uk` or at the Plant Scienes Department reception (make sure my name is on it). Feel free to team up with other group members, the main aim is to understand the material. Many of the examples and questions are based on the a book (2 Volumes) by Compeau and Pevzner [1]. I encourage you to borrow a copy (most college libraries should have it, let me or the lecturer know if not). If you hand in electronically, please name the file `group<x>_<crsid>_problemsheet<x>.<x>`

Introduction to Genetics

1. Describe the structure of the deoxyribonucleic acid (DNA), and highlight the ways in which it differs from the ribonucleic acid (RNA). Distinguish the concepts of a gene and a genome with respect to DNA structure.

2. Explain, with the aid of a diagram, the process of gene expression (synthesis of a protein based on the genetic information contained within DNA). Your answer should contain the following terms:

   - DNA
   - messenger RNA
   - codon
   - amino acid
   - protein
   - transcription
   - translation

3. How are different genes delimited within the DNA molecule? Can you relate this to a similar concept used within a programming language (covered within the Tripos)?

4. How many different codons exist? How about different amino acids? Provide an evolutionary explanation for the discrepancy between your two answers.

[1] Compeau, P. & Pevzner, P.A. (2015). Bioinformatics algorithms: an active learning approach (2nd Edition). Active Learning Publishers

Pairwise Sequence Alignments

5. Why do we use dynamic programming algorithms for sequence alignment problems?

6. Whats is the difference between *local* and *global alignment*? Try to think of applications as to when they should be used respectively.

7. Outline the key transformations that need to be made to the algorithm in order to find optimal *local alignments*, as well as incorporating *affine gap penalties*. Why are these features useful? Try to find examples when they are useful. Have you changed the time complexity of the algorithm by doing so?

8. Define the Longest Common Subsequence (LCS) problem between two strings and find a solution for the case of the two strings: `ACGT` and `GGTTTAAGCCGT`

9. Compute the *global alignment* and the best score of the following sequences `CGTGAA`, `GACTTAC` with the following parameters:

   - match score = +5
   - mismatch score = -3
   - gap penalty = -4

   Show the alignment graph including backtracking pointers or bring it to supervision.

10. If the sequences have different base composition (such as GC content) or length, what parameter values would you choose in order to determine multiple alignment of the sequences? Justify your answer.

11. **OPTIONAL** Implement the reduced-storage variant of the global alignment algorithm in a language of your choice, and verify that it provides the same result for the inputs in question 9.

Multiple Sequence Alignment

12. Compare and contrast the *dynamic programming*, *greedy*, and *progressive* approaches to aligning k sequences of length n, highlighting their respective time and memory complexities.

13. Explain the inputs and steps performed by the CLUSTALW algorithm.

14. Copy the entire text from a FASTA file (`http://www.cs.ukzn.ac.za/~hughm/bio/data/DinosaurCollagen.fasta`), containing the Collagen protein sequences from a number of different species. Then enter it into an online tool for multiple sequence alignment from the European Bioinformatics Institute (`http://www.ebi.ac.uk/Tools/msa/kalign/`). What does the Phylogenetic tree look like? Can you change the alignment options, so that the tyrannosaurus collagen is no longer paired with the newt collagen?