Computational Molecular Biology
Final Exam

LIACS Room 402
Thursday June 5th 2014
14.00 – 17.00

- State your name and student number on every page of your answers.
- Every assignment has the same weight. There are 13 assignments.
- Always fully explain your answers,
- Please note that you have a total of 3 hours to answer the questions.
- It is a closed book exam, no books, notes, smart phones, etc. allowed.

1. Calculate the score of the DNA sequence alignment shown below using the following scoring rules: +2 for a match, -1 for a mismatch, -4 for opening a gap, and -1 for each position in the gap.

```
C T T A A C T G G T A - T C A C G T - - G
  I I   I   I I I I         I   I I I     I
- T T - A - T G G T G C C C G C G T G A G
```

2. Given two DNA sequences S and T of length N and M, respectively. Which algorithm can be used to find an optimal global alignment of these two sequences? What is the space-, and time-complexity of the algorithm you proposed? Is that the best known? Estimate the sizes of N and M for which your proposed algorithm will be capable of producing an optimal global alignment in a 'reasonable' amount of time. (Define 'reasonable' yourself.)

3. For algorithms like BLAST and FASTA several different scoring matrices could be used when applied to amino acid sequences. Name at least two different scoring matrices. Which scoring matrix would you advise to use? Give (high level) pseudo code for the calculation of a scoring matrix.

4. Assume a hidden Markov model H with L states is given that emitted a sequence X of length N ($X = x_1 x_2 \ldots x_N$).
    1) Which algorithm should be used to determine the most probable path in H while emitting X?
    2) Which algorithm should be used to determine for each i in {1,…,N} the most probable state s_i of H when emitting x_i? Is the 'path' $s_1 s_2 \ldots s_N$ always a valid path in H? (Explain.)

5. Next Generation Sequencing Technologies are used to sequence known and unknown genomic sequences of different organisms. Describe the 2 main

problems that arise in this context? Sketch for one of the problems a possible approach to solve it.

6. Describe the characteristics of homology based gene finding and ab initio gene finding methods, respectively.

7. Algorithms that solve the multiple alignment problem often use a special score, for example the sum-of-pairs (SP) score.
   1) What is the main drawback of the SP-score?
   2) Give a more reasonable example of a scoring function used in multiple alignment algorithms.
   3) Give pseudo code for the Center Star Algorithm.

8. The architecture of single-sequence RNA secondary structure prediction algorithms can be expressed using context free grammars. The *g6* grammar used in Pfold has the following production rules:
   1) $S \rightarrow L\ S$
   2) $S \rightarrow L$
   3) $L \rightarrow r\ F\ r'$
   4) $L \rightarrow r$
   5) $F \rightarrow r\ F\ r'$
   6) $F \rightarrow L\ S$

   Where r, and r' are terminals from the set {a, c, u, g}, and F, L and S are non-terminals.
   Explain each of the production rules in terms of the secondary structures it produces. Has each given secondary structure a unique order of production rules, if *g6* is used?

9. The original version of the BLAST algorithm did not allow *indels*. Describe how the improved version of BLAST has been adapted in order to allow also *indels* (Note: 2 lines of pseudo code should be sufficient). Why is this improved version of BLAST about 3 times faster than the original version?

10. In the algorithm of Knuth Morris and Pratt (KMP) the following pattern P = 'ACTAACACTACACTAAG' is preprocessed. Give the failure links for P after preprocessing P (use a drawing).

11. For scoring of an alignment of two sequences often a special gap function is used. Propose an appropriate gap scoring function for Human DNA sequence alignments. Explain your proposal.

12. Describe in high level pseudo code an ab initio method for Protein Tertiary Structure Prediction. What is the complexity of your method? In which case would you resort to ab initio methods for Protein Tertiary Structure Prediction?

13. Let T be a given text. The Burrow-Wheeler Transform of T is denoted by BWT(T). Assume that for T, BWT(T) = 'gc$aaac'. Determine the original text T. Note: '$' is the lexicographical smallest symbol at the end of the original text T.