Computational Molecular Biology
Final Exam

LIACS Room 405
Tuesday June 5th 2018
14.00 – 17.00

- State your name, student number and affiliation on every page of your answers.
- Every assignment has the same weight. There are 12 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. Given two Protein sequences (amino acid 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 best space-, and time-complexity of the algorithm you proposed? Which scoring function and which gap-penalty function would you use? Explain why.

2. Describe how Hidden Markov Models can be used for gene prediction in DNA sequences. Which algorithm is used to determine the coding regions for a given DNA sequence? There is a fundamental problem when using HMMs for determining coding regions. Describe this problem and its solution.

3. Determine the suffix tree with the positions of the suffixes for the following sequence: ATCTATCTC and use it to find the occurrences of the read ATCT.

4. Give pseudo-code for a 3D protein structure prediction algorithm that uses threading.

5. Draw a 3-dimensional De Bruijn Graph for the following reads: GCAAAC, CAACAA, AAACGCA. De Bruijn Graphs are used to process next-gen sequencing data. In which cases? What would a typical dimension of such a De Bruijn Graph be? Why?

6. Below a fragment of the optimal alignment of two protein-coding DNA sequences is given (here the stretches of n aer: arbitrary nucleotides similar in the two sequences).

   ```
   nnnnnnnGACGAT---------CAAACTGnnnnnn... sequence A
   |||||||||||||         |||||||||||||
   nnnnnnnGACGATACATGACATCAAACTGnnnnnn... sequence B
   ```

   This alignment fragment suggests an insertion of 9 nucleotides in the sequence B vs. sequence A, which corresponds to 3 codons. However, the second codon of the insertion is TGA: a stop-codon. Thus, this insertion means that the downstream part of sequence B cannot code for a protein. Give a possible explanation for this discrepancy.

7. Let T be a given text. The Burrow-Wheeler Transform of T is denoted by BWT(T). Assume that for T, BWT(T) = ‘G$CCAATCTACAA. Determine the number of occurrences of the string ‘AATC’ in the original text T. Also determine the last 5 characters of T. Note: ‘$’ is the lexicographical smallest symbol at the end of the original text T.

8. A database entry for a mRNA contains the following annotation of its coding sequence (CDS) and 6 exons:

```
CDS  151..750
exon   1..180
exon 181..280
exon 281..427
exon 428..580
exon 581..631
exon 632..850
```

This mRNA encodes for a protein that contains a functional motif at amino acid positions 150-160. Other four isoforms of this protein are yielded by alternative splicing with skipping one of the exons 181-280, 281-427, 428-580 or 581-631. Which of these proteins still contain amino acids of the functional motif?

9. The score of the optimal local alignment of two sequences turned out to be higher than the optimal global alignment score for the same sequences, produced with the same scoring parameters. What can be concluded about the sequences?

10. The algorithms for sequence database similarity search like BLAST and FASTA exploit the strategy of identifying "word" similarities in query and subject sequences. Initial steps of these algorithms use the thresholds of word lengths, adjusted for the optimal algorithm performance. Where should these thresholds be higher: in the algorithms designed for nucleic acids or in those for proteins? Motivate your answer.

11. What is the advantage of using the "two-hit" approach in the sequence database similarity searches e.g. by BLAST program, as compared to the single hit strategy? (Two-hit approach: initial search for two non-overlapping "words" with scores higher than a threshold, located close to each other on a diagonal in the sequence space.)

12. Below a position-specific score matrix (PSSM) for a protein binding site is given:

```
A  [  0  52   0  25 ]
C  [  5   0   0   7 ]
G  [ 48   1   0  15 ]
T  [  0   0  53   6 ]
```

A motif-searching program scans <u>both strands of double-stranded DNA</u> using this PSSM matrix with the threshold score of 165. How many binding sites will be found in the DNA fragment shown below? Give the binding site nucleotide positions.

```
AGGATGCTTG AGATGTGGTC CGATGCAGCC TATCTGATCG
```