Computational Molecular Biology
Final Exam

LIACS Room B03
Wednesday May 29th 2019
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. What kind of algorithm is used for the computation of RNA secondary structure conformation with the lowest free energy?

2. Mutual information (MI) values can be used to estimate statistical significance of nucleotide covariations in RNA secondary structure models. However, high MI values are frequently not sufficient to conclude about base-pairing. Why?

3. Assume that when using a homology modeling server to predict the structure of a protein, you get a message that no proper template has been found. However, when you submit this protein as a query to the BLAST program to search in the protein sequence database, you get many sequence hits that are highly similar to the query sequence. What can be concluded in this case?

4. In a threading algorithm of a fold recognition program, a query amino acid sequence Q has been aligned to the template sequence S with a known structure. Below the fragment of the alignment is shown, with the deletion of two amino acid residues within a conserved region:

```
Q: ...--V...
S: ...LAI... (here dots denote identical residues)
```

What is the most likely reason for the alignment of the residue V to I rather than to L or A ?

A. The I>V is a more conservative substitution (higher score) than L>V or A>V, according to a substitution matrix used in the algorithm.
B. The I>V is a more conservative substitution (higher score) than L>V, and the A>V substitution is rejected because such an alignment leads to two deletions instead of one.
C. If the residue V in the sequence Q is substituted by I from S, the most favourable combination of pairwise knowledge-based potentials is computed for the sequence Q.
D. If the residue I in the structure S is substituted by V from Q, the most favourable combination of pairwise knowledge-based potentials is computed for the structure S.
E. There is no reason, it is just a random choice of one of the three possibilities.

5. A programmer A develops a stochastic Monte Carlo algorithm to simulate the folding of proteins. The main core of the algorithm: after generating a random structure, randomly introduced iterative structure changes are always accepted if they improve (decrease) the free energy value ($E_2 < E_1$). If such a change increases the energy ($E_2 > E_1$), it may be accepted with a probability $p \sim \exp (E_1 - E_2 / c)$, where c is a constant value. A programmer B further suggests to make the parameter c variable, with a gradual decrease during the simulation.
   Why is the last suggestion important? What would happen with the algorithm, if c would have a constant value?

6. A researcher tries to design an RNA molecule that would form the following secondary structure:

```
     N N      N N
    N   N    N   N
     N-N      N-N
     N-N      N-N
     N-N      N-N
     N-N      N-N
5'-N   NNNN   N-3'    (here N is any nucleotide.)
```

   Which of the RNAs given below is the best design?

```
5'-AGGGGAAAAGGGGAAAACCCCAAAACCCCA-3'    (A)
5'-AGGGGAAAACCCCAAAAGGGGAAAACCCCA-3'    (B)
5'-AGCGCAAAAGCGCAAAAGGGGAAAACCCCA-3'    (C)
5'-AGCGCAAAAGCGCAAAAGCGCAAAAGCGCA-3'    (D)
```

   Motivate your answer.

7. Given four Protein sequences (amino acid sequences) S_1, S_2, S_3, and S_4 of length N_1, N_2, N_3 and N_4, respectively. Which algorithm can be used to find an optimal global alignment of these four sequences? What is the best space-, and time-complexity of the algorithm you proposed? Which of the following scoring matrices BLOSUM62 or PAM100 would you use? Explain why.

8. From CASP11 to CASP13 we have seen a remarkable improvement in performance in the Free Modelling Category. Give a short description of the two main reasons for this? The best performer during CASP13 was AlphaFold. Name 3 important characteristics of the AlphaFold algorithm. (Give a short explanation of the importance of each characteristic.)

9. Determine the suffix tree with the positions of the suffixes for the following sequence: TACTAATCTACTA and use it to find the occurrences of the read ACTA.

10. Assume that we would like to use Knuth-Morris-Pratt's algorithm to search for occurrences of string P in text T, both strings are taken over an alphabet {A, C, T, G}. Assume P is equal to the string 'ATTCATTGATTCA'. Give the failure links for P. Note, you may use a table or a clear drawing.

11. Draw a 4-dimensional De Bruijn Graph for the following reads: GCAACA, CAACAA, AACGCAA. 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 in case of DNA sequencing data? Why?

12. Let T be a given text. The Burrow-Wheeler Transform of T is denoted by BWT(T). Assume that for T, BWT(T) = 'C$CCCTTCATAAA'. Determine the number of occurrences of the string 'CAT' in the original text T, without reconstructing the original text T. Also determine the last 5 characters of T. Note: '$' is the lexicographical smallest symbol at the end of the original text T.