Computational Molecular Biology
Final Exam

LIACS Room 403
Tuesday May 23rd 2017
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 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 best space-, and time-complexity of the algorithm you proposed? How does this compare to the best heuristic algorithms that solve this problem? Which gap penalty function would you use here?

2. Describe how Hidden Markov Models are used to determine a profile for a given family of protein sequences. Which algorithm is used to determine the parameters of the HMM? For a given protein sequence we would like to determine if it is a member of the given family of protein sequences using the HMM and a scoring-function. Define the scoring function and describe how its value is calculated?

3. When is the MAQ algorithm used? Depict templates that will be able to handle 3 mismatches.

4. Describe the characteristics of homology based and ab initio 3D protein structure prediction methods, respectively.

5. Draw a 2-dimensional De Bruijn Graph for the following reads: GCCAC, CACGG, GGACC, GCTCA, TCACG. Describe the problem related to next-gen sequencing that is solved using De Bruijn graphs.

6. Assume that in two pairwise sequence alignments the same score S has been obtained by the same optimal alignment algorithm. Say, the sequences A and B are aligned with the score S and the sequences C and D are also aligned with score S. All four sequences have comparable lengths, but different nucleotide contents: A and B have equal contents (25%) of all nucleotides, while C and D are biased: 50% C, 20% G, 15% A and 15% T. For which of the alignments, A vs. B or C vs. D, the obtained score is more significant?

7. Let T be a given text. The Burrow-Wheeler Transform of T is denoted by BWT(T). Assume that for T, BWT(T) = ‘gnthsorwii$s”. Determine the original text T using the UNPERMUTE algorithm (depict the steps). Note: ‘$’ is the lexicographical smallest symbol at the end of the original text T.

8. What is the reason for introducing the weights of sequences in the multiple sequence alignment algorithms? How are these weights used in the calculations?

9. One of the segments of the influenza A virus genome encodes two proteins, NS1 and NS2 with partially overlapping reading frames. The NS2 protein is produced by spliced mRNA after processing of an intron, whereas the NS1 is encoded by mRNA retaining the intron. Positions of the coding sequences and the intron are annotated in the nucleotide database as follows:

```
CDS             join(27..56,529..864)
                /gene="NS2"
...
CDS             27..719
                /gene="NS1"
```

   A researcher designs an experiment with the mutated protein NS1 at amino acid position 200. In this experiment it is important to keep the NS2 protein unchanged, which is possible if the mutation is introduced at a wobble position (the 3rd nucleotide in a codon) of the NS2 reading frame. What is the number of this nucleotide position?

10. A BLAST result for some nucleotide sequence query using BLAST database "nr/nt" (BLAST database containing entries from various species) yielded a human (*Homo sapiens*) sequence as a hit with relatively low E-value ("Expect"): say, E = 3e-13. Of course, a BLAST search with the same query in the same database, but with a constraint by organism name (*Homo sapiens*), yields the same hit. What is the most likely E-value in this case:
    (a) E = 3e-13;
    (b) E = 3e-12;
    (c) E = 2e-12;
    (d) E = 2e-14.

11. The folding free energies (ΔG) of the secondary structures formed by the following RNA sequences have been measured:

```
5'- CGCAAAAAGCG - 3'         ΔG = -1.7 kcal/mol;
5'- CGCCAAAAAGGCG - 3'       ΔG = -5.0 kcal/mol;
5'- CGCCGCAAAAAGCGGCG - 3'   ΔG = -10.8 kcal/mol.
```

    What is the folding free energy of this RNA sequence:

```
5'- CCCGCGCAAAAAGCGCGGG - 3' ΔG = ?
```

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

```
A  [ 1 17  0  0 33  0  0  0  0  1 11  1 ]
C  [ 0  0  0  0  0 33  0  0  1  0 20 32 ]
G  [31 16  0 33  0  0 33  0 32 22  1  0 ]
T  [ 1  0 33  0  0  0  0 33  0 10  1  0 ]
```

    A motif-searching program scans both strands of DNA using this PSSM matrix with the threshold score of 350. How many binding sites will be found in the DNA fragment shown below? Give the binding site nucleotide positions.

```
CGAGCCCAGA TGACGTGGGG GGCCACGTCA CCATTGCGCG CGGCAAAGCG
```