AAC-GG AAAGGA
AAAGGA AACGAG

Computational Molecular Biology
Final Exam

LIACS Room 174
Friday June 3rd 2016
14.00 – 17.00

- State your name and student number 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. Below a fragment of the dynamic programming matrix calculated by an algorithm for global pairwise sequence alignment is shown for the 3'ends of two nucleotide sequences: ...AACGAGGCA-3' (seqX) and ...AAAGGA-3' (seqY). The scoring rules are as follows:
match +1; mismatch -1; gap penalty -2 for each nucleotide in a gap.

```
... A   A   C   G   A   G   G   C   A  3' (seqX)
A  20  18  16  14  12  10   8   6   4
A  18   .   .   .   .   .   .   .   .
A  16   .   .   .   .   .   .   .   .
G  14   .   .   .   .   .   .   .   .
G  12   .   .   .   .   .   .   .   .
A  10   .   .   .   .   .   .   .   .
3'
(seqY)
```

   (a) Fill in the missing numbers in the matrix (dots).
   (b) What is the score of the global alignment?
   (c) Show the sequence alignment in this region.

2. Given two DNA sequences S and T of length N and M, respectively. Which algorithm can be used to find an optimal local 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?

3. Describe how Hidden Markov Models can be used to find the optimal alignment for a set of sequences? The Viterbi Algorithm is an important algorithm when working with HMMs. What does it compute? Mention one important application.

4. When is the MAQ algorithm used? What are the important characteristics of the MAQ algorithm? Depict templates that will be able to handle 2 mismatches.

5. HMM's have been very successful in gene-finding algorithms. Nevertheless HMM's have a major short-coming when modelling introns and exons. Describe this shortcoming and how this can be resolved.

6. A sequence database similarity search using a DNA sequence as a query in the standard BLASTN program (searching a nucleotide database) has yielded no results. However, using the same query in the BLASTX program (searching a protein database using a translated nucleotide query) yielded a number of significant hits. How can this be explained and what can be concluded from this result?

7. Below a position-specific score matrix (PSSM) for binding sites of a transcription factor is given:

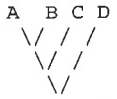
```
A [15  0  0 35  0  0  0 ]
C [15  0  0  0 35  0  0 ]
G [ 5  0 35  0  0 35  0 ]
T [ 0 35  0  0  0  0 35 ]
```

CTGACGT

For a DNA sequence of 30 base pairs, shown below, suggest a single mutation that would create a binding site of the transcription factor on one of the DNA strands. The site should have the optimal score according to the PSSM.

```
1 - ctgctgctaa cgtctgaaac ttcatatcca - 30
```

8. Four homologous sequences A, B, C and D have been aligned by a progressive multiple alignment algorithm (e.g. ClustalW). The guide tree produced by the algorithm is shown below.



If the optimal global pairwise alignments would be calculated in this dataset by a dynamic programming algorithm, which of them (A vs. B; A vs. C; A vs. D; B vs. C; B vs. D or C vs. D) would likely be equivalent to the alignment of the sequence pair deduced from the multiple alignment?

9. Draw a 2-dimensional De Bruijn Graph for the following reads: ACAAC, CACTA, CTACC, ACTCA, TACAC.

10. Multiple alignments of protein amino acid sequences are used in the algorithms for protein secondary structure prediction. What is the idea behind this?

11. In order to search for the genes coding for structured non-coding RNAs, it is possible to scan a genomic sequence with a window of, say, 200 nucleotides, computing the lowest folding free energy value within the window. In such an algorithm, the window locations with the lowest free energies should be considered as the putative non-coding RNA genes. What are the deficiencies of such an approach and what could you suggest to improve the search?

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$agcgtctata'. Determine the original text T using the UNPERMUTE algorithm. Note: '$' is the lexicographical smallest symbol at the end of the original text T.