T-61.5120 Computational Genomics, Exam 21.10.2015

You have 3 hours to complete exam. Total amount of points available is 50, to pass the exam 25/50 points is required. You may use a scientific calculator.

1. (10 points) Give short (2-3 sentences) definitions of the following concepts.

a) Multinomial sequence model	f) Phylogenetic tree
b) Open reading frame	g) Affine gap penalty
c) Optimal global alignment of 2 sequences	h) Transition probability
d) Randomization test	i) Orthologous genes
e) Substitution rate	j) Frameshift mutation

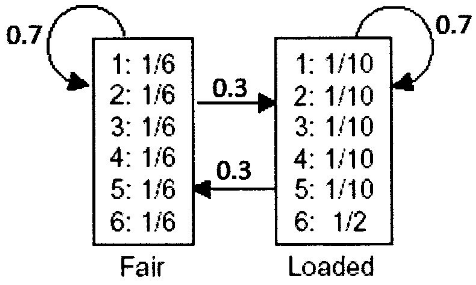
2. (10 points) Explain the principle behind the Smith-Waterman local alignment algorithm and apply the algorithm to the sequences 'AACTGACT' and 'TACTAA'. Write down the computed dynamic programming table, the best alignment score and all best local alignments.
Assume a substitution matrix that gives the score '+2' for matching symbols, and '-1' for mismatch, insertion and deletions.

3. (10 points) Viterbi algorithm
(a) (5 points) Explain how Viterbi algorithm works.
(b) (5 points) Consider the following HMM with 2 states: H (indicating high GC content) and L (indicating low GC content). The transition and emission probabilities are given in the following 2 tables:

$$T = \begin{array}{ccc} & H & L \\ H & 0.5 & 0.5 \\ L & 0.4 & 0.6 \end{array} \qquad E = \begin{array}{ccc} & H & L \\ A & 0.2 & 0.3 \\ C & 0.3 & 0.2 \\ G & 0.3 & 0.2 \\ T & 0.2 & 0.3 \end{array}$$

Simulate the Viterbi algorithm and find the most probable path of hidden states that produces the sequence: AGTAA.
Note: log2(0.5)=-1; log2(0.4)=-1,322; log2(0.6)=-0,737; log2(0.2)=-2,322; log2(0.3)= -1,737

4. (10 points) Forward and backward algorithms
(a) (5 points) Consider the Hidden Markov Model for the dishonest casino with the transition and emission probabilities depicted below. Compute the table F of forward probabilities for the sequence of rolls s=2,4,6.



(b) (2 points) Compute the probability P(s) of sequence s=2,4,6 given by the underlying HMM.

(c) (3 points) Using the table F computed at point a) and the table of backward probabilities B given below compute P(πi = Loaded|s) for each position i of the sequence s=2,4,6.

B	state 0	state 1 (emit 2)	state 2 (emit 4)	state 3 (emit 6)
Init	0.0056	0.0422	0.3333	1
Fair	0.0062	0.04312	0.26669	1
Loaded	0.005	0.04133	0.4	1

5. (10 points)

   a) (2 points) Explain the difference between the Jukes-Cantor correction and the Kimura 2 parameter model.

   b) (4 points) Explain how comparative genomics can be used to identify regulatory elements.

   c) (4 points) Describe the Pair Hidden Markov Models, their usage and their advantage compared with the Needleman-Wunsch algorithm.