CS-E5860 Computational genomics, Exam, April 7, 2017

Responsible teacher: Pekka Marttinen

You have three hours for the exam. The total number of points is 50. To pass the exam 25/50 points are required. You may use a scientific calculator with memory erased.

Q1) Term explanation:

Explain briefly (1-2 sentences) the following terms, 1p each.

a) translation
b) codon
c) open reading frame
d) GC content
e) BLAST
f) Jukes-Cantor model
g) indel
h) significance level
i) outgroup
j) accessory genome

Q2) Alignment:

Use the Smith-Waterman algorithm to find a local alignment between sequences CCAGCAT and GCAGA. Write down the computed dynamic programming table, the best local alignment score, and all best local alignments. Assume a substitution matrix that gives the score '+2' for matching symbols, '-1' for insertions and deletions and '-2' for mismatching symbols. (10p)

Q3) HMMs:

Using the tables F and B in Figure 1, output by the forward and backward algorithm, respectively, and the sequence of rolls $s = (1, 6, 6, 5)$, determine the following

a) $P(s(1) = 1, s(2) = 6, \pi_2 = \text{Loaded})$. (2p)

b) $P(s(3) = 6, s(4) = 5 | \pi_2 = \text{Fair})$. (2p) Π F

c) The probability $P(s)$ of sequence s given by the underlying HMM. (2p)

d) $P(\pi_i = \text{Loaded}|s)$ for each position i of the sequence s (4p)

	1	6	6	5
F	**1**	**2**	**3**	**4**
'Fair'	0.0833	0.0122	0.0029	0.00095
'Loaded'	0.0500	0.0300	0.0123	0.00095

B	**1**	**2**	**3**	**4**
'Fair'	0.0115	0.0351	0.1467	1.0000
'Loaded'	0.0190	0.0493	0.1200	1.0000

Figure 1: Forward and Backward tables for Question 3.

Q4) Neighbor-joining algorithm:

A distance matrix for four species (a,b,c,d) is given in Figure 2. Run one iteration of the neighbor-joining algorithm (i.e. merge a pair of nodes and compute the updated distances). (10p)

D =

	a	b	c	d
a		5	9	9
b			10	10
c				8
d				

Figure 2: Distance matrix for species in Question 4.

Q5) Miscellaneous:

a) The diagram in Figure 3 shows how genes A1, B1, B2, C1, C2, C3 have descended from a common ancestral gene following evolutionary events of speciation and gene duplication. Which genes are orthologs of A1? Which genes are paralogs of C3? Justify your answers. (2p, each)

b) Explain simulation-based inference. (3p)

c) Compare the first-order Markov model and the multinomial sequence model. (3p)

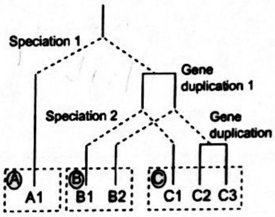

Figure 3: Relationships between genes in Question 5a.