582606 Introduction to bioinformatics
Course exam 16.10.2006, 16.00-18.30, A111, B123 and CK112
Esa Pitkänen, Elja Arjas, Samuel Kaski

Write the following information on the top of **each** answer paper: course name and date, your name, student number (or personal identity number, if you do not have or remember your student number). If you return more than one paper, number the pages and indicate the total number of answer papers in each paper.

You may answer either in Finnish, Swedish or English.

1. Suppose that DNA-sequence is modelled by a Markov chain, where the one-step transition probabilities are given by

$$P = \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix},$$

   where $p_{ij} = \frac{1}{3}$, if $i = j$, and $p_{ij} = \frac{2}{9}$ otherwise.

   (a) Find the stationary distribution of this Markov chain;

   (b) Determine the corresponding probability that a random 6-tuple of the DNA reads as the word ACGTTA.

2. (a) Perform local alignment for sequences `CGTCAAGTA` and `CAGTC`. Use match score 1, mismatch penalty $\mu = 1$ and indel penalty $\delta = 2$. What is the optimal alignment score? Report the optimal alignment or alignments.

   (b) Sketch an algorithm to report all optimal local alignments given a dynamic programming matrix with the values $M(i, j)$ computed. You do not have to provide a pseudocode, a general description of the necessary phases is enough.

3. Consider the following distance matrix for the four species a, b, c and d,

	a	b	c	d
a	0	4	6	10
b	4	0	6	10
c	6	6	0	10
d	10	10	10	0

   Find the phylogenetic tree corresponding to the distances. Is the tree ultrametric? Why or why not?

4. (a) Describe **briefly** (maximum of 1 page) how differential expression is measured with spotted microarrays. Be clear and include only the main steps.

   (b) If you want to measure differential expression in samples A and B, is it better to use a common reference for both, or to replicate the measurement of A against B? Why / why not? Justify statistically if possible.

Please give feedback on the course: http://ilmo.cs.helsinki.fi/kurssit/ servlet/Valinta?kieli=en.