582606 Introduction to bioinformatics
Separate exam Tuesday 5.6.2007, 16.00-19.30, A111
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. (15 p) Suppose that a colleague of yours makes the claim that the "G-C content", i.e., the proportion of G and C bases, in a certain part of the DNA of a considered organism is 35 percent. Suppose then that, as a toy example, you have counted the number of G's and C's in a 100 bases long sequence from that region and found that their total number is 40. Explain how to devise a simple statistical test for testing the above hypothesis of your colleague. Give also a rough idea of whether the difference between the hypothesized 35 percent and the observed count 40 would be considered as "statistically significant" so that the hypothesis could be rejected.

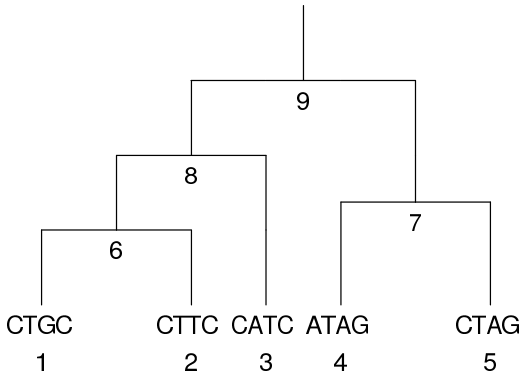
2. (10 p) Perform local alignment for sequences `ACTAACTCGG` and `ACCTAAGG`. Use match score 1, mismatch penalty $\mu = 1$ and indel penalty $\delta = 2$. What is the optimal alignment score? Report all optimal alignments.

3. (8 p) Consider the following distance matrix for the four species a, b, c and d,

	a	b	c	d
a	0	6	6	8
b	6	0	2	8
c	6	2	0	8
d	8	8	8	0

   Find an (additive) phylogenetic tree corresponding to the distances. Is the tree ultrametric? Why or why not?

Turn the page please!



4. (12 p) Find out the parsimony score for the phylogenetic tree and sequences given in the above figure using the parsimony algorithm presented in the lectures. Indicate the set F of each vertex of the tree.

5. (15 p)

   (a) Describe the main differences between oligonucleotide (affymetrix) and spotted (cDNA) microarrays. Be brief (max 1 page) and clear.

   (b) How would you preprocess or normalize spotted microarray data? Be brief (max 1 page), focus on the main issues, and justify your choices statistically if possible.