## 582606 Introduction to bioinformatics

Separate exam 20.11.2007, 16.00-19.30, Exactum 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. (12 p) Suppose N is binomial with n = 1000 trials and success probability p for each. What approximation for probabilities of the form  $P(N \le k)$  do you get from the Central Limit Theorem? How can you use this result for approximating probabilities of the form  $P(a \le \hat{p} \le b)$ , where  $\hat{p} = N/1000$  is the relative frequency of successes in the data? (*Hint:* Note that  $E(\hat{p}) = p$  and  $Var(\hat{p}) = p(1-p)/1000$ .)
- 2. (8 p) Perform global alignment for sequences GTCAAG and TACAGT. 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.
- 3. (8 p) Perform local alignment for sequences ATATCGATCG and CGAATATCA. Use match score 1, mismatch penalty  $\mu = 1$  and indel penalty  $\delta = 1$ . What is the optimal alignment score? Report the optimal alignment or alignments.
- 4. (8 p) Design an algorithm which takes a dynamic programming matrix with the values M(i, j) computed, match score, and mismatch and indel penalties as input and prints all optimal global alignments. Describe your algorithm with pseudocode.
- 5. (12 p) Consider the following distance matrix for the five species a, b, c, d and e,

	a	b	c	d	e
a	0	4	2	6	5
b	4	0	4	6	5
c	2	4	0	6	5
d	6	6	6	0	5
e	5	5	5	5	0

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

- 6. (12 p)
  - (a) Describe briefly (max 1 page) what is the basic idea underlying statistical tests of differential expression. Why is such testing needed?
  - (b) Explain the problem of multiple testing. Why is a solution needed? Sketch one solution and justify it.