

## Biosequence Algorithms, Spring 2005

Exercise 5, Wednesday March 2, 2005, at 12.15–14 in MT2

1. Consider the computation of frequent common substrings discussed at the lecture. Explain how the computation of the  $l(i)$  values can be extended to give also pointers to the corresponding substrings.
2. Explain how the *matching statistics* are computed for the pattern  $P = \text{“ANANAS”}$  and the target  $T = \text{“KANASANA”}$ , applying the method explained at the lecture. Show the suffix tree, and explain what tree traversals are performed.
3. The substrings of a newly sequenced string  $S_1$  that are potential contaminations from some strings in a set  $\mathcal{S}$  can be efficiently reported as an application of matching statistics. Explain the idea and sketch the algorithms.
4. Present the *suffix array* for the text *baababbac*. Explain how it is used to locate occurrences of the pattern *ba* in the text.
5. Present the *edit distance matrix*  $D(i, j)$  for strings  $S_1 = \text{aaabb}$  and  $S_2 = \text{bbaaab}$ . What are the optimal alignments and the optimal edit transcripts for the strings?