Special course in Computer Science:

Advanced Text Algorithms

Exercise set 2

Due: 27.11.2018

1. Show the construction of the suffix tree for the string S =“REVERENCE”, and explain how it would be used to locate occurrences of the patterns
   (a) “REN”,
   (b) “ENCE”, and
   (c) “REC”
in string S.

2. Show how Ukkonen’s algorithm constructs the suffix tree for the string “BABAABA”, including the construction of the suffix links.

3. Draw a generalized suffix tree for the strings ATGC, TGCTA, GTACTA. Using the tree answer the following questions:
   i) How many times does letter G appear in the strings? How about T?
   ii) In how many of the three strings does letter G appear?
   iii) What is a maximal substring that is common to at least two of the given strings?

4. Present the edit distance matrix D(i, j) for strings S_1 = winter and S_2 = writer. What are the optimal alignments and the optimal edit transcripts for the strings?

5. Consider locating approximate occurrences of the pattern pat =“dad” in the text “notabadidea”. Score matches by 5, and mismatches and insertions/deletions by −1. Present the dynamic programming table. How do you find the approximate occurrences of pat having similarity at least 3 with the corresponding substring in the text.

6. In a dynamic programming table for edit distance, must the entries along a row be nondecreasing? What about down a column or down a diagonal of the table? Discuss the same questions also for the optimal global alignment.