Marburg University

Department of Computer Science
Prof. Dr. Sebastian Wild

Date: 2025-12-05
Version: 2026-01-16 13:50

Exercise Sheet 8 for Algorithms of Bioinformatics (Winter 2025/26)

Hand In: *Until 2025-12-12 18:00, on ILIAS.*

Problem 1

30 points

Construct the Aho-Corasick automaton for

$$P[0..p) = [\texttt{aaa}, \texttt{abaaabbab}, \texttt{abab}, \texttt{abababab}, \texttt{bab}] \,.$$

Problem 2

40 points

Suppose that you are given a dictionary $S_1, \ldots, S_d$ of strings within $\{0,1\}^*$, with the total length of the dictionary being $D = |S_1| + \cdots + |S_d|$. You are also given a string S, of length N. In this question, we wish to compute a modified type of alignment between S and the dictionary $S_1, \ldots, S_d$. More precisely, we must write S as a concatenation of dictionary strings as well as extra 0s and 1s, minimizing the number of added 0s and 1s.

For example, if $S_1 = \texttt{000}, S_2 = \texttt{101}$ and $S = \texttt{000000010101}$, then a solution is $S = S_1 S_1 \texttt{0} S_2 \texttt{01}$, which has cost 3 due to the inserted `0` and `01`. This is optimal as well: one cannot use `000` and `101` more than twice resp. once, and we have in fact done this.

Design an algorithm to find an optimal such alignment. For full credit, your algorithm should use $O(D + N \min(d, \sqrt{D}))$ time; $O(D + N\sqrt{D})$ time or $O(ND)$ time will give partial credit.