Issue Date: 07.07.2015
Version: 2015-07-06 17:02

Exercise Sheet 6 for Computational Biology (Part 1), SS 15

Hand In: *Until Tuesday, 21.07.2015, 10:00 am, email to* `r_muelle@cs...` *or in lecture.*

Problem 16

4 points

For two string $S \in \Sigma^n$ and $T \in \Sigma^m$, we define the *matching statistics* of S w.r.t. T as

$$ms(i) = \max\Big(\Big\{ j - i + 1 \;\Big|\; j \in \{1, \ldots, n\} \;\wedge\; k, l \in \{1, \ldots, m\} \;\wedge\; S_{i,j} = T_{k,l} \Big\} \cup \{0\}\Big),$$

for $i = 1, \ldots, n$, i.e. $ms(i)$ is the length of the longest substring of S starting at index i that matches a substring (somewhere) in T.

Design an algorithm to compute the matching statistics of S w.r.t. T, for given $S \in \Sigma^n$ and $T \in \Sigma^m$. The running time of your algorithm should be in $\mathcal{O}(m + n)$.

Hint: If you use a suffix tree, you may keep the suffix links after its construction.

Problem 17

2 + 4 points

a) Design an algorithm to compute a longest common *subsequence*[1] of two given strings $S \in \Sigma^m$ and $T \in \Sigma^n$. The runtime of your algorithm should be in $\mathcal{O}(m \cdot n)$.

b) For two strings $R = aR'$ and $S = bS'$, $a, b \in \Sigma$, $R', S' \in \Sigma^*$, we define the *shuffle* of R and S as

$$R \sqcup\!\sqcup S \;:=\; a(R' \sqcup\!\sqcup S) \cup b(R \sqcup\!\sqcup S')$$

where $w \sqcup\!\sqcup \epsilon = \epsilon \sqcup\!\sqcup w = \{w\}$. For example, the shuffle of $R = ab$ and $S = cd$ is $R \sqcup\!\sqcup S = \{abcd, acbd, cabd, acdb, cadb, cdab\}$.

Consider a text $T \in \Sigma^l$ and two strings $R \in \Sigma^m$ and $S \in \Sigma^n$. Design an algorithm which decides whether T contains any interleaved (possibly with spaces) occurrence of R and S, i.e. any $w \in R \sqcup\!\sqcup S$ as a subsequence. The runtime of your algorithm should be in $\mathcal{O}(l \cdot m \cdot n)$.

[1] Remember, a subsequence need not consist of contiguous characters in S.

Problem 18

3 points

Prove Lemma 16 (page 128) in the German lecture notes; restated here for convenience:

Let K_1 and K_2 be two cycles of a minimal cycle cover $\mathcal{K}$ and let $w_1 \in K_1$ and $w_2 \in K_2$ be two elements of the cycles. It then holds

$$ov(w_1, w_2) \; < \; cost(K_1) + cost(K_2) \; .$$

Recall that we assume the set of strings $\mathcal{S}$ for the SCSP to be substring-free, i. e., no string is a substring of another.