Technische Universität München
Department of Informatics
Chair for Efficient Algorithms
Prof. Dr. Ernst W. Mayr/Dr. Jens Ernst
Johannes Nowak

Summer Term 2005
Exercise Sheet 3
10. Mai 2005

Selected Topics in Computational Biology

Due: 18.05.2005 **after** *the lecture*

Exercise 1 (10 points)

Construct a suffix tree for the string $CTGCCTGA$ with McCreight's algorithm. Describe all steps of the construction in detail.

Exercise 2 (10 points)

Consider the description of McCreight's algorithm as given in the lecture. In the rescanning-phase of the i-th step, we identify the extended locus of $\overline{xy}$. (Recall that according to our notation $head_{i-1}(t) = \alpha xy$.)

a) Suppose there exists no node $d = \overline{xy}$ in T_{i-1}. What consequences has this for the scanning phase in the i-th step?

b) Suppose conversely that $\overline{xy}$ exists in T_{i-1}. What can we derive from that for the form of $head_i(t) = xyz$.

Proof your statements.

Exercise 3 (10 points)

Consider again McCreight's algorithm as described in the lecture. In the i-th step the algorithm creates a new suffix link $f(\overline{\alpha xy}) = d(= \overline{xy})$ if $f(\overline{\alpha xy})$ is undefined. Give a generic example where $f(\overline{\alpha xy})$ is already defined in the i-th step of the algorithm.

Exercise 4 (10 points)

Proof that the running time of McCreight's algorithm is bounded by $O(n)$. The analysis can be done as follows:

- Show that the total number of nodes visited in all rescanning phases of the algorithm is bounded by $O(n)$.
- Show that the total number of character comparisons in all scanning phases of the algorithm is bounded by $O(n)$.
- Conclude that the total running time of the algorithm is $O(n)$.