

Biosequence Algorithms, Spring 2005
Exercise 4, February 22, 2005, at 12.15–14 in MT2

1. For each $i = 2, \dots, m + 1$, the height of the implicit suffix tree \mathcal{I}_i constructed by Ukkonen's algorithm is at most $i - 1$. Explain why this holds. (The *height of a tree* is the maximum number of edges on any root-to-leaf path. Hint: the longest string path.)
2. Simulate how Ukkonen's algorithm constructs the suffix tree for the string *aababb*. Restrict to the explicitly computed extensions, and show how they create new nodes and suffix links. (There shouldn't be more than about ten extensions computed explicitly.)
3. Consider the suffix tree built for a string S . Give a procedure that computes for each node v of the suffix tree a *compact representation of its node label* $\mathcal{L}(v)$, as a pair of indices (i, j) such that $\mathcal{L}(v) = S[i \dots j]$. (Hint: string-depth.)
4. Draw a generalized suffix tree for the strings CATCA, ATCATA and GATA. Explain how the following questions could be answered with the help of the tree: How many times does a single base, say A or C, appear in the strings? In how many different strings do they appear? What is a maximal substring that is common to at least two of the strings?
5. Given a circular string S of n characters, the **circular string linearization problem** is to choose a place to cut S so that the resulting linear string is the lexically smallest of all the possible n linear strings created by cutting S . (See Gusfield, Sec. 7.13) Explain how the problem can be solved in linear time as an application of suffix trees.