

Do problems 1-4 and then pick either 5 or 6. Please typeset your solutions.

Problem 1

Read the posted explanation by Gusfield on the LCP array construction. When it was argued that $LCP(Suff_{k+1}; Suff_{Pred(k+1)}) \geq h - 1$, it says “if the path to leaf $Pred(k+1)$ does not extend below v , that will be impossible”. Now, give a more expanded argument why this will be impossible.

Problem 2

Let T be a long string and P be a substring of T , and suppose P appears in T multiple times. Describe how these occurrence (i.e. starting position of P) in suffix array POS . Prove your claim.

Problem 3

Try to find a linear time algorithm for finding the MUMs in two strings T and T' using a suffix array POS and LCP array.

Problem 4

Given a string S , and consider all $|S|$ rotations of S . That is, let $S = PQ$ (P is a prefix and Q is a suffix of S), $S' = QP$ is a rotation of S . Now design an algorithm that outputs a list of lexicographically sorted rotations in time linear to the length of S .

Problem 5

Read carefully the posted tandem repeat writeup by Gusfield. Find a simple way to compute longest common extensions (LCE) as needed in it. Recall LCE asks for the length of longest common prefix of two given suffixes. You can not use complex tool like constant time lowest common ancestor in suffix trees. This looks not easy, but some simple algorithm we went over in class should help...

Problem 6

We are interested in the number of unique tandem repeats. For example, $S = \text{aaaab}$. Here, aa appears three times, but in terms of unique tandem repeat, we consider aa as one occurrence. There is a nice result that says given a string with n letters, the number of unique tandem repeats is bounded by $2 * n$. Note, if we drop the uniqueness requirement, the number of tandem repeats can be as large as $O(n^2)$. Now you are going to prove (part of) the claim.

Consider string S , where *three* TR start at position 0 of S , with length $2a$, $2b$ and $2c$ respectively (and $a < b < c$). I claim that at least one of three TR must also occur somewhere *inside* S .

- (1) Given a brief argument why the above claim leads to the $2n$ bound.
- (2) Now show if $c \geq 2a$, then the claim holds.
- (3) The case $c < 2a$ (and so $2a > b$). I claim that the TR of length $2a$ also appears in S , starting from $b - a$ and ending at $a + b - 1$. We need some case analysis and will just do one here. Show that for $0 \leq x < 2a - b$, $S[x] = S[x + b - a]$. Can you see how to finish the proof from here?