

Algorithms in Bioinformatics 2, SoSe2007
Assignment sheet # 5

Toibas Kloepper

May 14, 2007

1 Longest prefix repeat (2 points)

A substring q is called a *prefix repeat* of string t if q is a prefix of t and has the form pp for some string p . Give a linear-time algorithm to find the longest prefix repeat of an input string t . This problem was one of the original motivations for developing suffix trees.

2 Finding MUMs (5 points)

Write a program that determines all MUMs of a given minimum length in a pair of sequences. To solve this problem, please download www-ab.informatik.uni-tuebingen.de/teaching/ss07/albi2/java/java05.zip and modify the file `albi2.suffixtree.MUMfinder.java`. Please apply your program to the provided pairs of data files `human1.fa` and `mouse1.fa` etc. Use 25 as minimum length.

3 Size of the T data-array (1 points)

What is the maximum length required for the data-array T in the WOTD algorithm, as a function of n , the length of the text?

4 Memory requirement (1 point)

Why does the algorithm that builds the WOTD suffix tree require *two* copies of the auxiliary array *suffixes*?

5 Worst-case runtime complexity of WOTD algorithm (1 points)

What is the worst-case runtime complexity for building the complete WOTD suffix tree, and why?

Assignments due: **Monday, May 21, 10am**