

Algorithms in Bioinformatics II, SS2004  
**Assignment sheet # 8**

Daniel Huson

June 9, 2004

**1 Longest prefix repeat (3 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 Building a suffix array (4 points)**

By hand, show in detail how the algorithm of Manber and Myers builds the suffix array for the text **barbara**. Your description should include the variables *pos*, *inv*, *count*, *bh* and *b2h*.

**3 Worst case text for binary search (3 points)**

In the first speed-up of the binary search algorithm for suffix trees, we introduced the *mlr* variable to save comparisons. However, the worst-case runtime requirement remained  $O(h \log n)$ , where  $h$  is the length of the query and  $n$  is the length of the text. In this context, please construct a query and an infinite family of texts that require the worst-case runtime and explain why it works.

Assignments due: **Monday, June 14, 10am**