

Biosequence Algorithms, Spring 2005
Exercise 1, February 1, 2005, at 12.15–14 in MT2

1. Using asymptotic notation, derive upper and lower bounds for the number of different
 - (a) prefixes and suffixes
 - (b) substrings, and
 - (c) subsequencesof a string $S = s_1 \dots s_n$.
2. (Gusfield, Ex. 1.2) A *circular* string of length n is a string in which character n is considered to precede character 1. (Bacterial and mitochondrial DNA is typically circular.) Design a linear-time algorithm to determine whether a linear string α is a substring of a circular string β . (Use the existence of a linear-time exact matching algorithm to solve this problem.)
3. (Gusfield, Ex. 1.3) **Suffix-prefix matching.** Give an algorithm that takes in two strings α and β , of lengths n and m , and finds the longest suffix of α that exactly matches a prefix of β . The algorithm should run in $O(n + m)$ time.
4. Show the character comparisons performed by
 - (a) the naive method and
 - (b) the Z algorithm

to search for occurrences of the pattern “AATAAT” in the target

ACAATAATAAT .

5. The Z algorithm works in linear time. This would seem to suggest that the algorithm examines any character at most a fixed number of times. Does this conjecture hold? Either justify a fixed upper bound to the number of times that any character is examined by the algorithm, or give a counter-example showing that some characters may be examined by the algorithm an unlimited number of times.