

CMSC858E: Algorithms in Biosequence Analysis

Homework 1

Due: In class, September 27, 2005

1. Give an algorithm that takes 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.
2. A substring α contained in a string S is called a tandem array of β (called the base) if α consists of more than one consecutive copy of β . For example, if $S = xyzabcabcabcabcqpq$, then $\alpha = abcabcabcabc$ is a tandem array of $\beta = abc$ or $\beta = abcabc$. A maximal tandem array is a tandem array that cannot be extended either left or right.

Give an example to show that two maximal tandem arrays of a given base β can overlap.

Suppose S has length n . Give an $O(n)$ -time algorithm that takes S and β as input, and finds every maximal tandem array of β .

3. When searching for the site where a particular oligonucleotide (short DNA sequence) might hybridize in a genome, we must check both the forward and reverse complement strands of each DNA sequence. Denote the oligonucleotide sequence by P and the input genomic sequence by T . Further, denote the reverse complement of sequence S by S' . Let $\$$ be a character not present in P or T . Discuss the strengths and weaknesses of the following algorithms for finding all the oligonucleotide binding sites:
 - (a) Exact string search for P in $T\$T'$.
 - (b) Exact string search for P in T , then P' in T .
 - (c) Exact string search for P in T , then exact string search for P in T accessed from last character to first character with the characters of T mapped to their reverse complements on the fly.
 - (d) Any other strategy you care to dream up...

Be sure to consider running time, memory required, memory locality, and ease of coding.

4. (a) Plot the expected number of occurrences e of a pattern P of length n in a string of length m whose characters are sampled independently and uniformly from an alphabet Σ

$$e = (m - n + 1) \left(\frac{1}{|\Sigma|} \right)^n$$

for various values of m , n , and $|\Sigma|$.

(b) Determine how many bases long P should be to ensure that occurrences of P are unlikely to be chance events in genomes of the following sizes:

- i. 1.8Mb (Haemophilus influenzae - the flu)
- ii. 4.7Mb (Escherichia coli - intestinal bacteria)
- iii. 100Mb (Caenorhabditis elegans - roundworm)
- iv. 180Mb (Drosophila melanogaster - fruit fly)
- v. 3.1Gb (Homo sapiens - human)

(c) Implement the naive algorithm (or the Z-algorithm, Knuth-Morris-Pratt, or use strcmp, etc) in the language (Perl, Python, C, C++, lisp, etc) of your choice and run it on some or all of the DNA sequence at

`www.umiacs.umd.edu/~nedwards/teaching/CMSC858E_Fall_2005/data/chr22.fasta`

Discuss the degree to which the empirical number of matches for various patterns is consistent with the theoretical model of part (a). Point out, in particular, any particularly significant deviations from the theoretical model.