Homework 2

November 3, 2020

Wildcard KMP

Describe a modification of Knuth-Morris-Pratt (KMP) algorithm in which you only want to report the first occurrence and the pattern can contain any number of wildcard symbols '*', each of which matches an arbitrary substring. For instance, the pattern "al*thm*s" matches "ilovealgorithmscourse": the first wildcard matches *"gori"* and the second wildcard matches an empty substring. The algorithm should run in O(n + m), where n is the size of the text and m the size of the pattern.

You can use standard KMP implementation from any website referencing it of write your own.

Cyclic rotation alignment

Implement an algorithm that given strings s and t finds the best local alignment of any cyclic rotation of t to any cyclic rotation of s. You can assume that t is much shorter than s. The algorithm should run faster than $O(nm^2)$ (e.g. $O(nm)$ or $O(nmlog(m))$ are OK and $O(n^2m^2)$ or $O(nm^2)$ are not), where n is the size of s and m the size of t.

z

Read mapping

Implement an algorithm that given the reference file (long nucleotide sequence up to 1000 bp) in FASTA format and read file (short nucleotide sequences up to 1000 sequences up to 50 bases) in fasta format align reads to the reference (this process is usually called mapping and resulting alignments are called mappings). We don't allow mappings with more than one-base difference (one mismatch or one insertion/deletion of a single nucleotide). Output to console in a format that explicitly shows start and end positions on the reference and alignment itself. Make some memory/speed optimization given that we don't keep the mappings with more than one nucleotide difference.

Deadline - November 15 (midnight EST)