UC Berkeley, CS 176: Algorithms for Computational Biology (Fall 2014)

Problem Set 1

Instructor: Nir Yosef

Out: Sep 11, 2014 Due: Sep 25, 2014

1. (**Tandem arrays)** A substring α contained in string S is called a *tandem array* of β (called the base or the repeat unit) if α consists of more than one consecutive copy of β. For example, if $S = xyzabcabcabcabcpq$, then $\alpha = abcabcabcabc$ is a tandem array of $\beta = abc$. Note that α also is a tandem array of a longer base, namely *abcabc*. A *maximal* tandem array is a tandem array that cannot be extended either left or right. Given the base β, a tandem array of β in S can be described by a pair of numbers (s, r), where s denotes the starting location in S and r denotes the number of times that β is repeated.

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

   (b) Assume $|S| = n$. Give an $O(n)$-time algorithm that takes S and β as input, finds all maximal tandem arrays of β with odd number of repeats (r odd), and outputs the pair (s, r) for each occurrence. (Since maximal tandem arrays of a given base can overlap, a naive algorithm would establish only an $O(n^2)$-time bound.) *Hint*: Use the Z-algorithm.

   Biological relevance: Tandem arrays of DNA sequences are common in eukaryotic genomes. For example, microsatellites, also called short tandem repeats (STR), are tandem arrays of 1–6 bp repeat units that are widely used as genetic markers in forensic science and in linkage analysis (a method of locating disease-influencing genes). Also, telomeres, the regions at the ends of chromosomes, contain long (10–15 kb) tandem arrays of the hexanucleotide repeat $TTAGGG$. There are other types of tandem arrays, including minisatellites and satellites, which differ in the total size and the length of the repeat unit.

2. **(Minimal unique substring)** A minimal unique substring U of some text string T is defined as a substring that satisfies the following properties:

   i. (Uniqueness) U occurs exactly once in T.

   ii. (Minimality) all proper (is neither the empty string nor the entire string) prefixes of U occur at least twice in T.

   iii. U is of length at least ℓ for some given fixed parameter ℓ.

   Give a solution to the problem of finding all minimal unique substrings that runs in $O(m)$ time, where $m = |T|$.

   Biological relevance: This problem has applications in PCR *primer* design in DNA sequencing. (See Wikipedia entries for PCR and primer, if you are not familiar with those terms.)

3. **(k-mer frequency count)** The term k-mer refers to all possible substrings of length k that are contained in a string. For a string S, construct a data structure in $O(N)$, efficient for determining how many times a given k-mer appears. That is, for any k-mer you can count the number of its occurrences in S in time at most $O(k \log N)$ in the worst case, where k is the length of the k-mer and N the length of the string. Is $O(k \log N)$ optimal or the bound can be improved? How about the case when we want to be as memory efficient as possible.

4. **(All Pairs LCP)** You are given k strings, each of length n. Give an algorithm to find the longest common prefix for each pair of strings in time $O(kn + p)$, where p is the number of pairs with a common prefix of non-zero length.