CISC-471 FALL 2019

HOMEWORK 3

Please work on these problems and be prepared to share your solutions with classmates in class on Thursday September 26. Assignments will **not** be collected for grading.

Programming

Write a program in the language of your choosing (I recommend Python) and verify that it works on the sample data. For each problem be prepared to tell us why you think your algorithm is correct (whether you program worked on the sample data or not). Also provide an estimate of the time and space complexity of your algorithm.

Problem 4.14: Implement and algorithm that counts the number of occurrences of each ℓ-mer in a string of length n. Run it over a bacterial genome and construct the distribution of ℓ-mer frequencies. Compare this distribution to that of a random string of the same length as the bacterial genome.

I found a genome for the bacteria Chlamydia here: ftp://ftp.sanger.ac.uk/pub/project/pathogens/Chlamydia/Ct_L2_UCH-1_454.dna

There are n= 1038869 symbols in the sequence. Find frequency counts for ℓ up to 9, and print stats of any frequencies that appear out of the ordinary. Use your own creativity to decide what "out of the ordinary" means.

To do this you would need to be able to determine the expected frequency of an ℓ-mer in string of length n.

Once you have stats for the Chlamydia bacteria repeat the experiment for a random 'cgta' string of length n.

BONUS: Can you explain the biological relevance of ℓ-mers that appear much more frequently in the bacterial genome than expected?

Problems

These questions come from *An Introduction to Bioinformatics Algorithms* by Neil C. Jones and Pavel A. Pevzner.

Problem 4.11: The search trees in the text are complete k-ary trees: each vertex that is not a leaf has exactly k children. It is also balanced: the number of edges in the path from the root to any leaf is the same (this is sometimes referred to as the height of the tree). Find a closed-form expression for the total number of vertices in a complete and balanced k-ary tree of height L. (Note: In mathematics, an expression is said to be a *closed-form expression* if it can be expressed analytically in terms of a finite number of elementary operations.)

Problem 4.13: Given a long text string T and one shorter pattern string s, and an integer k, find the first occurrence in T of a string (if any) s' such that $d_H(s, s') \leq k$. What is the complexity of your algorithm?