CS 4630/5630 Kresman: Homework 1 short.

CS 4630: Do first two problems. CS 5630: Do all three problems.

Learning Outcome: Do a simple computation in a STEM domain.

Bioinformatics warm-up: Bioinformatics integrates biology and informatics in the collection and analysis of complex data such as genomics data. Messenger RNA (aka mRNA) is a family of RNA molecules that convey genetic information from DNA to the ribosome. Will learn about this topic a bit later in the semester, but for now couple of warm-up exercises that may perk up your interest😊

**Problem 1**:

Messenger RNA: Find all occurrences of a motif (for example, CAG) in an mRNA sequence - a string containing some combination of the letters (aka ‘bases’) A, C, T, G, etc. A sample mRNA sequence: CCACTGCACTCACCGCACCCGGCCAATTTTTGTGTTTTTAGTAGCAGACTAAATACCATATAGTGAACACCTAAGA

Write a python app that reads the mRNA sequence and the motif from an input file HW1shortData.txt (1st line is mRNA sequence, 2nd is motif) and spits out three lines:

* mRNA sequence
* motif
* starting positions the motif occurs, for example: 12, 345 - meaning the motif occurs only at starting positions 12 and 345.

**Problem 2**:

Bioinformaticians look for similarity between two genome sequences, i.e. how far apart the two sequences are (aka distance between two sequences). Two popular distance metrics: Hamming and Levenshtein. Google either one and explain one of them – how it works, limitations if any, and an example walk-through. No need to write code.

**Problem 3**:

Extend Problem 2, by writing code to compute the distance for the metric you chose. Assume line 3 and line 4 of the input file are the two sequences.

Notes:

* See HW1shortData.txt for a sample input file, though your code should work with any valid input.
* Coding - Use ONLY the primitives we covered through Unit 4.
* Canvas turn-in: hw1short.ipynb. ONE file, please – add a header like Problem 1, etc. so one can easily infer what your response relates to.