Problem 1: Computing products of probabilities (50 points)

A typical subroutine in several of the methods discussed in class require the computation of the product of the probabilities of the symbols in a string. Assume that we are dealing with a Bernoulli source, and therefore that symbols are independently generated. Suppose we are given a text \( x \) of size \( n \) over the alphabet \( \Sigma = \{A, T, C, G\} \) and that we know \( p_A, p_T, p_C, p_G \) in advance.

**Q (20pts):** Design an algorithm to compute

\[
p(y) = \prod_{i=1}^{\left|y\right|} p_{y[i]}
\]

in \( O(1) \) for any of the substrings \( y \) of the text \( x \). A substring \( y \) is identified by a pair \((i, j)\) where \( i \) is the start of the substring, and \( j \) is the end of the substring. You are allowed to spend \( O(n) \) time in a preprocessing phase, just once. After that, your algorithm should be able to return \( p(y) \) in constant time for any substring \( y \) in the text.

**Q (10pts):** How would you avoid cancellation problems induced by multiplying long sequences of very small real numbers?

**Q (20pts):** Discuss how to extend your algorithm for a Markov model of order \( h > 0 \). What is the time complexity of your algorithm? Is it possible to obtain the same time complexity as for the Bernoulli case, that is \( O(n) \) preprocessing once and then constant time for any substring \( y \) of the text \( x \)?

Problem 2: Maximum likelihood (30 points)

We have seen in class (see slides “Intro to Probability and Statistics”) that the maximum likelihood estimate for the probabilities of the Bernoulli model corresponds to the intuitive idea of computing the proportion of the occurrences in the text for each symbol of the alphabet.

Formally, the ML estimate for a Bernoulli model is

\[
p_{A}^{ML} = \frac{f(A)}{f(A) + f(C) + f(G) + f(T)}
\]

and similarly for \( p_{C}^{ML} \), \( p_{G}^{ML} \) and \( p_{T}^{ML} \). Recall we denote by \( f(y) \) the number of occurrence of \( y \).

**Q:** Prove that \( P(x|p_{\text{ML}}^{ML}) > P(x|p) \) for any other choice of \( p \neq p_{\text{ML}}^{ML} \).

**Hint:** show that \( \log(P(x|p_{\text{ML}}^{ML})/P(x|p)) > 0 \) and use the fact that the relative entropy is always positive.
Problem 3: Profile HMM (20 points)

Build a profile HMM on the alignment below. Use one state for columns 1,2,3,7,8,9 and one insert state for columns 4,5,6.

   1 2 3 4 5 6 7 8 9
1 ACAA--ACC
2 AGAAGTCTC
3 ACA---ATG
4 TCAC--ATC
5 ACCG--ATC

Note: You are expected to work on this assignment on your own. Anything that you submit that comes from “external sources” (a friend, a web page, a book, etc.) must be acknowledged and will be graded accordingly.