Homework 9

CS 4364/5364
Spring 2022

Due: 9 May 2022

Because of the reliance of the particular assignments in this class on mathematical notation, and the fact that all assignments will be submitted electronically, students are encouraged to use LaTeX to formalize their responses. **For those enrolled in the graduate section the use of latex is *required*.** This assignment (like all others) will be posted on the course `github`[1] as source code as well as in PDF form on the course website. Graduate students will need to include the `.tex` files as well as a PDF, this is optional but encouraged for undergraduates.

Question (25 points): Given a segment of the genome G and an RNA sequence R, describe an algorithm (along with proofs of correctness and running-time) to align the RNA to a segment of the genome (i.e. all of the RNA, substring of the Genome) allowing for gaps left by the removal of introns.

Use the following penalty/reward structure:

- Matches, mismatches, and single character indels will be scored using a function $\delta(a, b)$, defined for $a, b \in \Sigma \cup \{\text{-}\}$.
- Intron (represented as long segments of deletions from the genome) will be scored not based on length but on the two characters in the RNA bordering the intron. Let s_1, s_2 and p_1, p_2 be the characters before and after the intron gap respectively (in the RNA of course). The penalty of inserting the intron between them (a substring in R is $....s_1s_2p_1p_2...$) is scored by $\gamma(s_1, s_2, p_1, p_2)$. Notice that γ does not take into account the intron length (deletion in the genome).

Your algorithm should have a running time of $O(mn^2)$ (assuming $|G| = m, |R| = n$) and use $O(mn)$ space.

[1] `github.com/deblasiolab/CS4364-documents`