

BCB 567/CprE 548 Bioinformatics I
Fall 2007
Homework 2
Due Tuesday, September 25

1. (5 points) Given two sequences, which score is larger: their local alignment score or global alignment score? Why? How does their semi-global alignment score compare with the other two scores?
2. (5 points) Consider affine gap alignment with parameters α, β, g , and h . Write an equation for the score of the best possible path that travels outside of the k -band for two strings of equal length n .
3. (5 points) Sequences generated by shotgun genome sequencing projects, called *shotgun reads*, can be thought of as 500-700 base pair sequences read from a random location on the genome of interest. Sometimes, foreign DNA sequence from a bacteria, called *vector sequence*, can contaminate the end of a shotgun read. Consider two shotgun reads \mathcal{A} and \mathcal{B} that have been read from overlapping locations on the genome. If it wasn't for the presence of the contaminating vector sequence, we would be able to find a high scoring suffix-prefix alignment between them.

We wish to define a modified version of semi-global alignment, to possibly allow for ignoring the ends of the reads. We call an ignored region on the end of a read a *tail*. We will use a parameter t to bound maximum size of the tails. Describe how you would modify semi-global alignment using parameter t to allow for tails.

4. (5 points) Consider the following formulation of spliced alignment. You wish to only penalize small blocks of gaps. A block of gaps longer than a certain length X is considered to be excised from one of the sequences during splicing and is therefore not penalized during alignment. You are given 4 parameters for alignment: α for matches, β for mismatches, γ for gaps, and X : the gap threshold. Any block of gaps with total length $g_L < X$ is penalized $g_L \times \gamma$. Any block of gaps with length $g_L \geq X$ is penalized 0.
 - (a) (3 points) Describe an algorithm that finds the best spliced alignment between two strings that runs in at most $O(n^3)$ time. Writing out the formulas for filling in the DP table is sufficient.
 - (b) (2 points) How would you generalize your solution such that you score a block of gaps with $g_L \geq X$ with the parameter σ instead of 0?

(Note: There exists a way to solve this problem in $O(n^2)$ time. While the solution requires no advanced mathematics, finding this solution is difficult. Therefore, full credit will be awarded for a correct $O(n^3)$ solution.)

5. (5 points) Bacteria have circular DNA molecules. Consider the full sequence alignment problem for two bacteria molecules \mathcal{A} and \mathcal{B} . We can construct linearization of a circular molecule by choosing a random point to cut the circle. There are n possible linearizations for molecule \mathcal{A} , which we will label $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n$. There are m possible linearizations of \mathcal{B} , which we will label $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_m$. One way to solve this problem is to consider all mn possible pairs $(\mathcal{A}_i, \mathcal{B}_j)$ and then solve the full-sequence comparison problem for each pair. This will require $O(m^2n^2)$ total time.

Find an algorithm that solves the full sequence alignment problem for bacteria molecules that is asymptotically faster than $O(n^2m^2)$. Analyze its runtime and space requirements.