Homework 3

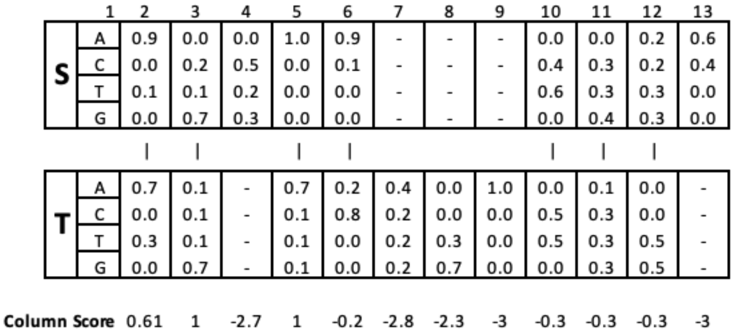
CS 4364/5364
Spring 2022

Due: 23 February 2022

1. **(25 points)** *Profile Alignment Problem* Given two sequence profiles T and S, both of size $\sigma \times n$ (that is each represents a sequence of length n, but with probabilities of each character from the alphabet at each position), determine the optimal alignment (i.e. which columns of S align with which columns of T) under the scoring scheme δ.

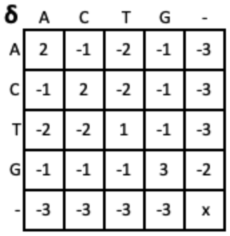
   Your task: **modify** the Needlman-Wunch global alignment algorithm to consider these profiles rather than sequences. You can assume that the replacement costs are defined in a function $\delta(a,b) \to \mathbb{Z}, \forall a, b \in \Sigma \cup \{'-'\}$. Give the algorithm, an explanation of correctness, and analysis of it's running time.

   An example alignment is shown below over the alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{T}, \mathtt{G}\}$, as well as it's alignment score. Note that the score for a column is now no longer the value of δ for the two characters being aligned, but the weighted sum of these values.



	1	2	3	4	5	6	7	8	9	10	11	12	13
S	A	0.9	0.0	0.0	1.0	0.9	-	-	-	0.0	0.0	0.2	0.6
	C	0.0	0.2	0.5	0.0	0.1	-	-	-	0.4	0.3	0.2	0.4
	T	0.1	0.1	0.2	0.0	0.0	-	-	-	0.6	0.3	0.3	0.0
	G	0.0	0.7	0.3	0.0	0.0	-	-	-	0.0	0.4	0.3	0.0
		|	|		|	|				|	|	|	
T	A	0.7	0.1	-	0.7	0.2	0.4	0.0	1.0	0.0	0.1	0.0	-
	C	0.0	0.1	-	0.1	0.8	0.2	0.0	0.0	0.5	0.3	0.0	-
	T	0.3	0.1	-	0.1	0.0	0.2	0.3	0.0	0.5	0.3	0.5	-
	G	0.0	0.7	-	0.1	0.0	0.2	0.7	0.0	0.0	0.3	0.5	-
Column Score		0.61	1	-2.7	1	-0.2	-2.8	-2.3	-3	-0.3	-0.3	-0.3	-3

Figure 1: Alignment of two profiles



δ	A	C	T	G	-
A	2	-1	-2	-1	-3
C	-1	2	-2	-1	-3
T	-2	-2	1	-1	-3
G	-1	-1	-1	3	-2
-	-3	-3	-3	-3	x

Figure 2: Scoring scheme