Problem Set 9

Due 4pm, Wednesday, December 11th

This homework assignment is optional. If you choose to hand in this homework for credit, I will drop your two lowest Problem Set scores. In other words, your score on this assignment will replace your next-lowest homework score, if this score is higher. This is in addition to dropping your lowest score on PS 0 to PS 8. To obtain credit for this assignment, you must hand it in by Wednesday at 4pm. This is a hard deadline. Solutions will be posted on Wednesday, promptly at 4:00 pm, so students can use them to study for the final exam.

Your name:

Collaboration is allowed on this homework. You must hand in homework assignments individually. List the names of the people you worked with:

You may not use an alignment program to calculate the alignments on this problem set.

Homework must be submitted by 4pm in MI650 or electronically to Canvas or mstolzer@andrew.cmu.edu.
1. Profile alignment: A multiple sequence alignment of \( k \) sequences with symbols drawn from alphabet \( \Sigma \) can be treated as a single string over a larger alphabet of size \( |\Sigma|^k - 1 \). Durbin calls this a sequence profile. Progressive multiple alignment heuristics build a multiple alignment by repeatedly applying the global pairwise alignment algorithm to larger and larger profiles.

Pairwise profile alignment requires a modification of the recurrence relation in the global pairwise alignment algorithm. The goal in this problem is to derive the recurrence relations required for profiles \( s \) and \( t \), each of which contain two sequences. For columns \( i \) and \( j \) in profiles \( s \) and \( t \), \( s[i] = s[i_1] \in \Sigma_2 \) and \( t[j] = t[j_1] \in \Sigma_2 \), where \( \Sigma_2 = \Sigma' \times \Sigma' \setminus \{-\} \). \( \Sigma_2 \) is the expanded alphabet representing all possible combinations of symbols and gaps that can appear in the column of an alignment of two sequences. For example, when \( \Sigma = \{A, C, G, T\} \), this expanded alphabet contains 24 combinations of symbols, \( \Sigma_2 = \{AA, AC, AG, AT, A-, \ldots, TT\} \).

For this problem, assume similarity scoring. In each sub-problem, you are asked to write down the recurrence relation for a specific assignment of symbols and gaps to \( s[i] \) and \( t[j] \). In addition, show the last column of the alignment of \( s[1 \ldots i] \) and \( t[1 \ldots j] \) corresponding to each of the three entries in the recurrence.

The solution to the first problem is shown as an example.

(1) Given \( x \in \Sigma \) and

\[
\begin{align*}
s[i] &= x \\
t[j] &= x
\end{align*}
\]

i. complete the recurrence relation for the profile-profile alignment in terms of \( M \), \( m \), and \( g \)

\[
A[i, j] = \max \left\{ \begin{array}{c}
A[i - 1, j] + 4g \\
A[i - 1, j - 1] + 4M \\
A[i, j - 1] + 4g
\end{array} \right. 
\]

ii. for each of the entries in the recursion relation in part (i), show the last column of multiple alignment of \( s[1 \ldots i] \) and \( t[1 \ldots j] \) that would result if the corresponding entry in the recursion gave the maximum score.

\[
\begin{align*}
A[i - 1, j]: &\quad \frac{x}{x} \\
A[i - 1, j - 1]: &\quad \frac{x}{x} \\
A[i, j - 1]: &\quad \frac{-}{x} \\
\end{align*}
\]
(a) Given $x, y \in \Sigma$ and

\[
s[i] = \begin{cases} x & i = 1, \\ y & i > 1, \\ \end{cases} \quad t[j] = \begin{cases} x & j = 1, \\ y & j > 1. \\ \end{cases}
\]

i. complete the recurrence relation for the profile-profile alignment in terms of $M$, $m$, and $g$

ii. for each of the entries in the recursion relation in part (i), show the last column of multiple alignment of $s[1..i]$ and $t[1..j]$ that would result if the corresponding entry in the recursion gave the maximum score.
(b) Given no gaps in $s[i]$ or $t[j]$

i. complete the recurrence relation for the profile-profile alignment in terms of $p(\cdot, \cdot)$, $g$, $s[i_1]$, $s[i_2]$, $t[j_1]$, and $t[j_2]$.

ii. for each of the entries in the recursion relation in part (i), show the last column of multiple alignment of $s[1..i]$ and $t[1..j]$ that would result if the corresponding entry in the recursion gave the maximum score.
(c) Given that the second symbol of \( s[i] \) is a gap

i. complete the recurrence relation for the profile-profile alignment in terms of \( p(\cdot, \cdot) \), \( g \), \( s[i_1] \), \( t[j_1] \), and \( t[j_2] \).

ii. for each of the entries in the recursion relation in part (i), show the last column of multiple alignment of \( s[1..i] \) and \( t[1..j] \) that would result if the corresponding entry in the recursion gave the maximum score.
(d) Given that the second symbol of \( t[j] \) is a gap

i. complete the recurrence relation for the profile-profile alignment in terms of \( p(\cdot, \cdot) \), \( g \), \( s[i_1] \), \( s[i_2] \), and \( t[j_1] \).

ii. for each of the entries in the recursion relation in part (i), show the last column of multiple alignment of \( s[1..i] \) and \( t[1..j] \) that would result if the corresponding entry in the recursion gave the maximum score.
(e) Given that the second symbols of both \( s[i] \) and \( t[j] \) are gaps

i. complete the recurrence relation for the profile-profile alignment in terms of \( p(\cdot, \cdot) \), 
\( g \), \( s[i_1] \), and \( t[j_1] \).

ii. for each of the entries in the recursion relation in part (i), show the last column of 
multiple alignment of \( s[1..i] \) and \( t[1..j] \) that would result if the corresponding entry 
in the recursion gave the maximum score.
2. Progressive alignment is a multiple alignment strategy in which the pairwise alignments between all pairs of sequences in the data set are calculated in a pre-processing step. The resulting pairwise alignment scores are used to construct a “guide tree”. The multiple alignment is then obtained by repeatedly merging partial multiple alignments in an order that is determined by the guide tree. As this problem demonstrates, the final result can be quite sensitive to the order in which the profiles are merged.

In this problem you are given four sequences:

X: ARIL  
Y: ARID  
Z: CHIDE  
W: CHILD

There are six possible combinations of four words. Given a scoring function where $M = 1$, $m = -1$, and $g = -1$, the optimal alignments for those six pairings are

ARIL  ARIL-  ARIL-  ARID-  ARI-D  CHI-DE  
ARID  CHIDE  CHILD  CHIDE  CHILD  CHILD-

(a) Suppose that this is the guide tree for the progressive alignment of these four sequences:

```
X
 /   \
Y   Z
 / \  /
W   
```

Of the six profiles shown on the previous page, which two will be aligned to give the final alignment of four sequences?

(b) Align these profiles using the recurrence relations from the previous question. Show your alignment matrix with scores and traceback on the attached template. Highlight or circle the cells in the matrix that correspond to the optimal solution.
(c) Write down the resulting multiple alignment of four sequences and compute its sum-of-pairs score.

(d) Does this give you the same score as the entry in the lower right hand corner of your alignment matrix? If not, why not?
(e) Suppose that this is the guide tree for the progressive alignment of these four sequences:

```
X
 W
 Y
 Z
```

Of the six profiles shown on page 8, which two will be aligned to give the final alignment of four sequences?

(f) Align these profiles using the recurrence relations from the previous question. Show your alignment matrix with scores and traceback on the attached template.

(g) Write down the resulting multiple alignment of four sequences and compute its sum-of-pairs score.
(h) Is the MSA you obtained in (g) the same as the MSA you obtained in (c)? Are the sum-of-pairs scores the same?

(i) What is the pairwise alignment of ARID and ARIL induced by this multiple sequence alignment (in (g))? What is the score of the induced alignment? How does it compare to the score of the optimal pairwise alignment of ARID and ARIL shown at the start of question 2?