Solutions to Homework 2

Due: 1:30 pm Friday, September 21

Collaboration is allowed on this homework. You must hand in homework assignments individually. You may not use an alignment program to calculate the alignments on this problem set.

Your name:

List the names of the people you worked with.

*Homework must be submitted by 1:30 pm in MI646.*
1. Profile alignment: The goal in this problem is to derive the recurrence relations required for $s \in \Sigma_2$ and $t \in \Sigma_2$, profiles of two sequences each, where $\Sigma_2 = \Sigma' \times \Sigma' \setminus \{(.,.)\}$. 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.

(a) Given $x \in \Sigma$ and

\[ s[i] = \begin{cases} x & \text{if } i = 1 \\ x & \text{otherwise} \end{cases} \quad t[j] = \begin{cases} x & \text{if } j = 1 \\ x & \text{otherwise} \end{cases} \]

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

\[ A[i, j] = \max \begin{cases} A[i-1, j] + 4g \\ A[i-1, j-1] + 4M \\ A[i, j-1] + 4g \end{cases} \]

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.

\[ A[i-1, j]: \begin{array}{c} x \\ x \end{array} \]

\[ A[i-1, j-1]: \begin{array}{c} x \\ x \\ x \end{array} \]

\[ A[i, j-1]: \begin{array}{c} x \\ x \end{array} \]
(b) Given \( x, y \in \Sigma \) and
\[
\begin{align*}
s[i] &= \frac{x}{y} \\
t[j] &= \frac{x}{y}
\end{align*}
\]

i. complete the recurrence relation for profile-profile alignment in terms of \( M, m, \) and \( g \):
\[
A[i, j] = \max \begin{cases} 
A[i-1, j] + 4g \\
A[i-1, j-1] + 2M + 2m \\
A[i, j-1] + 4g
\end{cases}
\]

ii. for each of the entries in the recursion relation in part (i), show the last column in the alignment of \( s[1 \ldots i] \) and \( t[1 \ldots j] \) in terms of \( x, y, \) and \( \_ \) for each entry in the recurrence.
\[
\begin{align*}
A[i-1, j]: & \quad \frac{x}{y} \\
A[i-1, j-1]: & \quad \frac{x}{y} \\
A[i, j-1]: & \quad \frac{\_}{y}
\end{align*}
\]
(c) Given \( x, z \in \Sigma \) and
\[
\begin{align*}
  s[i] &= \frac{x}{x} && t[j] = \frac{z}{z}
\end{align*}
\]

i. complete the recurrence relation for profile-profile alignment in terms of \( M, m, \) and \( g \);
\[
\mathcal{A}[i, j] = \max \begin{cases}
  \mathcal{A}[i-1, j] + 4g \\
  \mathcal{A}[i-1, j-1] + 4m \\
  \mathcal{A}[i, j-1] + 4g
\end{cases}
\]

ii. for each of the entries in the recursion relation in part (i), show the last column in the alignment of \( s[1 \ldots i] \) and \( t[1 \ldots j] \) in terms of \( x, z, \) and \( \_ \) for each entry in the recurrence.
\[
\begin{align*}
  \mathcal{A}[i-1, j]: & \quad \frac{x}{x} \\
  \mathcal{A}[i-1, j-1]: & \quad \frac{x}{x} \\
  \mathcal{A}[i, j-1]: & \quad \frac{\_}{\_}
\end{align*}
\]
(d) Given $x, y \in \Sigma$ and
\[
s[i] = \frac{x}{y}, \quad t[j] = \frac{y}{y}
\]

i. complete the recurrence relation for profile-profile alignment in terms of $M, m,$ and $g$:
\[
A[i, j] = \max \begin{cases} 
A[i-1, j] + 2g \\
A[i-1, j-1] + 2m + 2g \\
A[i, j-1] + 4g
\end{cases}
\]

ii. for each of the entries in the recursion relation in part (i), show the last column in the alignment of $s[1 \ldots i]$ and $t[1 \ldots j]$ in terms of $x, y,$ and $\_\_$ for each entry in the recurrence.

$A[i-1, j]: \quad \frac{x}{\_}$

$A[i-1, j-1]: \quad \frac{x}{y}$

$A[i, j-1]: \quad \_ \quad \frac{y}{y}$
(e) Given \( x, y \in \Sigma \) and
\[
    s[i] = x \quad t[j] = y
\]

i. complete the recurrence relation for profile-profile alignment in terms of \( M, m, \) and \( g \):
\[
    A[i, j] = \max \begin{cases} 
        A[i-1, j] + 2g \\
        A[i-1, j-1] + m + 2g \\
        A[i, j-1] + 2g 
    \end{cases}
\]

ii. for each of the entries in the recursion relation in part (i), show the last column in the alignment of \( s[1 \ldots i] \) and \( t[1 \ldots j] \) in terms of \( x, y, \) and \( \_ \) for each entry in the recurrence.

\[
    A[i-1, j]: \quad \_ \quad \_ \\
    A[i-1, j-1]: \quad \_ \quad y \\
    A[i, j-1]: \quad \_ \quad y
\]
(f) Given \(x, y \in \Sigma\) and
\[
s[i] = \frac{x}{-} \quad t[j] = \frac{x}{y}
\]

i. complete the recurrence relation for profile-profile alignment in terms of \(M, m,\) and \(g;\)
\[
A[i, j] = \max \begin{cases} 
A[i-1, j] + 2g \\
A[i-1, j-1] + M + m + 2g \\
A[i, j-1] + 4g
\end{cases}
\]

ii. for each of the entries in the recursion relation in part (i), show the last column in
the alignment of \(s[1 \ldots i]\) and \(t[1 \ldots j]\) in terms of \(x, y,\) and \(_\_\) for each entry in the
recurrence.

\(A[i-1, j]:\)
\[
\frac{x}{-}
\]

\(A[i-1, j-1]:\)
\[
\frac{x}{x} \frac{x}{y}
\]

\(A[i, j-1]:\)
\[
\frac{y}{y}
\]

\(A[i, j]:\)
\[
\frac{x}{y}
\]
2. Multiple sequence alignment: 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 preprocessing 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 three sequences:

\[
\begin{align*}
X: \text{CORD} & \quad Y: \text{CARED} & \quad Z: \text{BOARD}
\end{align*}
\]

Given a scoring function where \( M = 3 \), \( m = -3 \), and \( g = -2 \), the optimal alignments for the three possible word pairs are

- \text{COR-D}
- \text{CARED}
- \text{CO-RD}
- \text{BOARD}
- \text{C-ARED}
- \text{BOAR-D}

(a) Give the number of matches, mismatches, and gaps associated with each of the three pairwise alignments. What is the total alignment score for each alignment?

- \( Y \text{ and } Z : 3M + m + 2g = 2 \)
- \( X \text{ and } Z : 3M + m + g = 4 \)
- \( X \text{ and } Y : 3M + m + g = 4 \)
(b) These three sequences were aligned with the scoring function $M = 3$, $m = -3$, and $g = -2$, using progressive alignment according to this guide tree:

![Guide Tree Diagram]

The resulting sequence-profile alignment is shown in the alignment matrix below. Highlight or circle the cells in the matrix that correspond to the traceback from the optimal solution to the start of the alignment.

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>C</th>
<th>A</th>
<th>R</th>
<th>E</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>-</td>
<td>C</td>
<td>O</td>
<td>R</td>
<td>-</td>
<td>D</td>
</tr>
<tr>
<td>Z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>← -4</td>
<td>← -8</td>
<td>← -12</td>
<td>← -14</td>
<td>← -18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>↑</td>
<td>↖</td>
<td>↖</td>
<td>↖</td>
<td>↖</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-4</td>
<td>-6</td>
<td>← -10</td>
<td>← -14</td>
<td>← -16</td>
<td>← -20</td>
</tr>
<tr>
<td></td>
<td>↑</td>
<td>↖</td>
<td>↖</td>
<td>↖</td>
<td>↖</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>-8</td>
<td>-10</td>
<td>-6</td>
<td>← -10</td>
<td>← -12</td>
<td>← -16</td>
</tr>
<tr>
<td></td>
<td>↑</td>
<td>↖</td>
<td>↑</td>
<td>↖</td>
<td>↖</td>
<td>↖</td>
</tr>
<tr>
<td>A</td>
<td>-12</td>
<td>-14</td>
<td>-10</td>
<td>← -12</td>
<td>← -14</td>
<td>← -18</td>
</tr>
<tr>
<td></td>
<td>↑</td>
<td>↖</td>
<td>↑</td>
<td>↖</td>
<td>↖</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>-16</td>
<td>-18</td>
<td>-14</td>
<td>← -4</td>
<td>← -6</td>
<td>← -10</td>
</tr>
<tr>
<td></td>
<td>↑</td>
<td>↖</td>
<td>↑</td>
<td>↑</td>
<td>↖</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-20</td>
<td>-22</td>
<td>-18</td>
<td>-8</td>
<td>-9</td>
<td>0</td>
</tr>
</tbody>
</table>

(c) How many optimal multiple alignments are there?

*Three*
(d) Write down the resulting optimal multiple alignment(s) of the three sequences.

```
| B O A R - D |
| C O - R - D |
| C A - R E D |

| B O A R - D |
| - C O R - D |
| - C A R E D |

| B O A R - D |
| C - O R - D |
| C - A R E D |
```

(e) Compute the sum-of-pairs score of the optimal multiple sequence alignment. (If there is more than one optimal multiple alignment, just calculate the sum-of-pairs score for one of them.) Show the number of matches \((M)\), mismatches \((m)\) and gaps \((g)\) in each column. Give the total number of matches, mismatches and gaps and the total numerical score.

```
| B O A R - D |
| C O - R - D |
| C A - R E D |

\[
\begin{align*}
\text{m} & \quad \text{M} & \quad \text{g} & \quad \text{M} & \quad \text{0} & \quad \text{M} \\
\text{m} & \quad \text{m} & \quad \text{g} & \quad \text{M} & \quad \text{g} & \quad \text{M} \\
\text{M} & \quad \text{m} & \quad \text{0} & \quad \text{M} & \quad \text{g} & \quad \text{M}
\end{align*}
\]

\[8M + 4m + 4g = 4\]
(f) Does this give you the same score as the entry in the lower right hand corner of your alignment matrix? If not, why not?

No, the scores are different. This is because the scores in the alignment matrix reflect comparisons across profiles, not within profiles; i.e., each number in the alignment matrix is the sum of two pairwise scores. The final SP scoring of the alignment in part (e) is the sum of the scores comparing all three pairs of symbols.

(g) These three sequences were aligned with the scoring function $M = 3$, $m = -3$, and $g = -2$, using progressive alignment according to this guide tree:

The sequence-profile alignment is shown in the alignment matrix below. Highlight or circle the cells in the matrix that correspond to the traceback from the optimal solution to the start of the alignment.
(h) How many optimal multiple alignments are there?

One

(i) Write down the resulting optimal multiple alignment(s) of the three sequences.

\[
\begin{align*}
& BOAR - D \\
& CORD \\
& CARED \\
& m \ M \ g \ M \ 0 \ M \\
& m \ g \ M \ M \ g \ M \\
& M \ g \ g \ M \ g \ M \\
9M + 2m + 6g = 9
\end{align*}
\]

From a quantitative perspective, this alignment is better. The sum-of-pairs similarity score of this alignment is 9, whereas the alignment in (b) has a score of 4. From a qualitative perspective, this alignment is better because it correctly aligns O’s with O’s and A’s with A’s.

(j) What is the pairwise alignment of CORD and CARED induced by this multiple sequence alignment? What is the score of the induced alignment? How does it compare to the score of the optimal pairwise alignment of CORD and CARED?

\[
\begin{align*}
& CO - R - D \\
& CARED \\
\end{align*}
\]

The score is 3. This is a suboptimal alignment. The score is lower than the optimal alignment score, which is 4.