Problem Set 2  

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

For Problems 2b and 2f, hand in your alignment matrix with scores and arrows on the alignment template on the syllabus page. Show the traceback on the matrix by highlighting or circling the cells in the matrix that correspond to the optimal solution.

Homework must be submitted by 4pm in MI646 or electronically to comp-bio@cs.cmu.edu.

1. Profile alignment: A multiple 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 up 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 \( s \in \Sigma_2 \) and \( t \in \Sigma_2 \), profiles of two sequences each, where \( \Sigma_2 = \Sigma' \times \Sigma' \setminus \{(\_\_\)\}. For profiles of two sequences, the score for aligning \( s[i] \) with both symbols in \( t[j] \) and the second symbol in \( s[i] \) with both symbols in \( t[j] \). Assume similarity scoring.

(a) Give the recurrence relation for calculating the alignment matrix \( a[i, j] \), for the case where \( s[i] = s_1 \) and \( t[i] = t_1 \), in terms of \( p(\cdot, \cdot), g, s_1, s_2, t_1, \) and \( t_2 \).
(b) Give the recurrence relation for calculating the alignment matrix $a[i, j]$, for the case where $s[i] = s_1$ and $t[i] = t_2$, in terms of $p(\cdot, \cdot), g, s_1, t_1, t_2$.

(c) Give the recurrence relation for calculating the alignment matrix $a[i, j]$, for the case where $s[i] = s_1 s_2$ and $t[i] = t_1$, in terms of $p(\cdot, \cdot), g, s_1, s_2$ and $t_1$. 
(d) Give the recurrence relation for calculating the alignment matrix $a[i, j]$, for the case where $s[i] = s_1$ and $t[i] = t_1$, in terms of $p(\cdot, \cdot), g, s_1,$ and $t_1$.

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 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:

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 righthand 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 3, 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. Give the resulting MSA here:

(g) What is the pairwise alignment of ARID and ARIL 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 ARID and ARIL?
3. The robotics department has created a robot that plays Rock-Paper-Scissors. In each round of this game, two players simultaneously present a hand signal representing a rock (closed fist), paper (open palm) or scissors (ring and index fingers extended). The winner of the round depends on which of the two signals is dominant:

<table>
<thead>
<tr>
<th>Hand Signal</th>
<th>Hand Signal</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper</td>
<td>Scissors</td>
<td>Scissors cut Paper: Scissors wins</td>
</tr>
<tr>
<td>Scissors</td>
<td>Rock</td>
<td>Rock blunts Scissors: Rock wins</td>
</tr>
</tbody>
</table>

If both players present the same signal, then the round is a tie.

The people programming the robot have installed a basic move-set where the robot will always play Scissors on the first round. The robot never presents the same signal twice in a row, and presents each of the other two moves with equal probability. Your goal is to construct a Markov model of this robot’s sequence of moves.

(a) Draw the topology of the Markov chain. Use as few states as possible. Clearly label your nodes and transition probabilities.

(b) What is the initial state probability distribution?
(c) What is the transition matrix?

(d) You plan to play the robot for five rounds. For each of the five rounds, what is the probability that the robot will present each of the possible signals? In other words, give the state probability distribution for the first five time steps.

(e) What set of moves should you play during the first five rounds in order to lose as infrequently as possible. In other words, what sequence of five moves maximizes the expected number of rounds in which you win or tie?
(f) Someone left the robot playing over a long weekend. When you return after the holiday, the robot has generated a very long sequence of hand signals. Using this data, you calculate the frequency of each of the hand signals. These frequencies correspond to the stationary distribution of the model.
Write down the system of equations that specify the stationary distribution for this model.

(g) To find the distribution, solve your equations.