

BMI/CS 576 Fall 2008
Midterm Exam

Prof. Colin Dewey

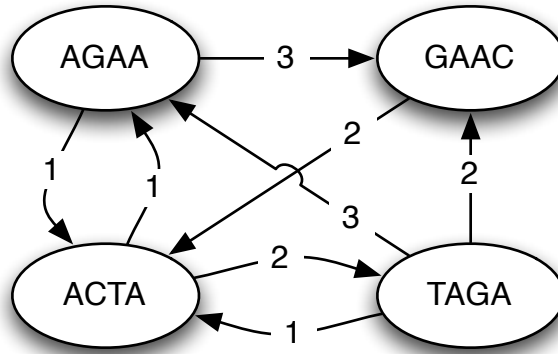
Friday, November 7th, 2008 9:55-10:45am

Name: _____ SOLUTIONS _____

Write your answers on these pages and show your work. You may use the back sides of pages as necessary. Before starting, make sure your exam has every page (numbered 1 through 7).

Problem	Score	Max Score
1	_____	25
2	_____	25
3	_____	25
4	_____	25
Total	_____	100

1. (25 pts) Let $S = \{AGAA, GAAC, TAGA, ACTA\}$ be a set of sequencing reads.
- (a) (15 pts) Give the *overlap graph* for this set of reads. You may omit weight zero edges.



- (b) (10 pts) For your overlap graph, give one *hamiltonian path* and the assembled superstring to which it corresponds. Any valid hamiltonian path is acceptable.

There are 7 possible hamiltonian paths if we do not include zero weight edges in the overlap graph:

Path	Superstring
AGAA → ACTA → TAGA → GAAC	AGAACTAGAAC
AGAA → GAAC → ACTA → TAGA	AGAACTAGA
ACTA → TAGA → AGAA → GAAC	ACTAGAAC
TAGA → AGAA → GAAC → ACTA	TAGAACTA
TAGA → ACTA → AGAA → GAAC	TAGACTAGAAC
TAGA → GAAC → ACTA → AGAA	TAGAACTAGAA
GAAC → ACTA → TAGA → AGAA	GAAC TAGAA

2. (25 points) Suppose we have discovered a specimen of a *hodag*, a fictional eukaryotic organism that is part of Wisconsin folklore. We sequence its genome and also acquire the sequence for many of its mRNAs. We wish to determine the positions of the exons in the genome by aligning each mRNA, x , against the genome, y . We will modify the *affine gap global alignment algorithm* such that (1) insertions are only allowed in y (these will correspond to introns) and, (2) insertions at the beginning and end of y have zero cost. An example alignment path for this algorithm is given in Figure 1. Let the lengths of x and y be n and m , respectively.

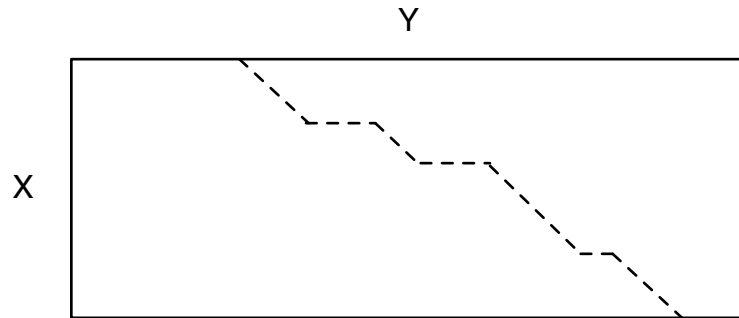


Figure 1: An example alignment path for this algorithm

- (a) (5 pts) Give the *initialization* formulas for this algorithm.

Since insertions are not allowed in x , we eliminate the I_X values. We must initialize the first row and first column of the DP matrices.

$$\begin{aligned}
 M(0, j) &= -\infty, \forall j \\
 I_Y(0, j) &= 0, \forall j \\
 M(i, 0) &= -\infty, \forall i \geq 1 \\
 I_Y(i, 0) &= -\infty, \forall i \geq 1
 \end{aligned}$$

- (b) (15 pts) Give the *recursion* formulas for this algorithm.

$$\forall i \geq 1, j \geq 1$$

$$\begin{aligned}
 M(i, j) &= \max \begin{cases} M(i-1, j-1) + S(x_i, y_j) \\ I_Y(i-1, j-1) + S(x_i, y_j) \end{cases} \\
 I_Y(i, j) &= \max \begin{cases} M(i, j-1) + g + s \\ I_Y(i, j-1) + s \end{cases}
 \end{aligned}$$

- (c) (5 pts) Give the *traceback* step for this algorithm. Be sure to specify which entry of the dynamic programming matrix (or matrices) we must start from.

With the recursion formulas above, insertions at the end of y are penalized, so we do not start at entry (n, m) as in standard global alignment. Instead, we start at the maximum M entry in row n . That is, we start at $M(n, j')$, where $j' = \operatorname{argmax}_j M(n, j)$. Then we traceback as in the standard affine gap global alignment algorithm until we hit the first row, at which point the remainder of the alignment is an insertion at the beginning of y .

Alternatively, the recursion formula for I_Y in the last row ($i = n$) may be specially defined as:

$$I_Y(n, j) = \max \{I_Y(n, j - 1), M(n, j - 1)\}$$

so that insertions at the end of Y are not penalized and we start the traceback from the maximum of $M(n, m)$ and $I_Y(n, m)$, just like the standard algorithm.

3. (25 pts) In HW #3, you implemented the hidden Markov model (HMM) shown in Figure 2 for predicting the locations of exons and introns in pre-mRNA sequences. To improve this model we can add additional states to represent the *donor splice site*, which occurs at the boundary of an exon and the intron immediately downstream of it. We will model the donor splice site as a length six sequence that includes the last two nucleotides of an exon and the first four nucleotides of the downstream intron. The 3rd and 4th positions of the donor splice site are always *G* and *T*, respectively. The other positions in the splice site may have any nucleotide, but each position is characterized by a different distribution over the nucleotides. Exons must have length > 2 , introns must have length > 4 , and the last exon does not have a splice site at its end.

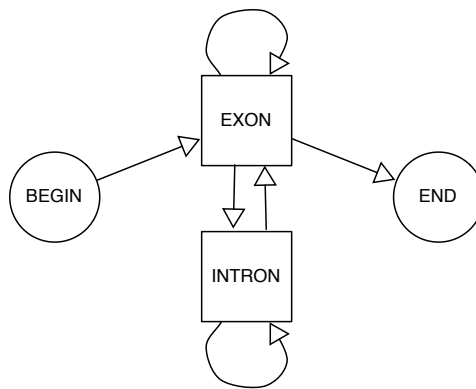
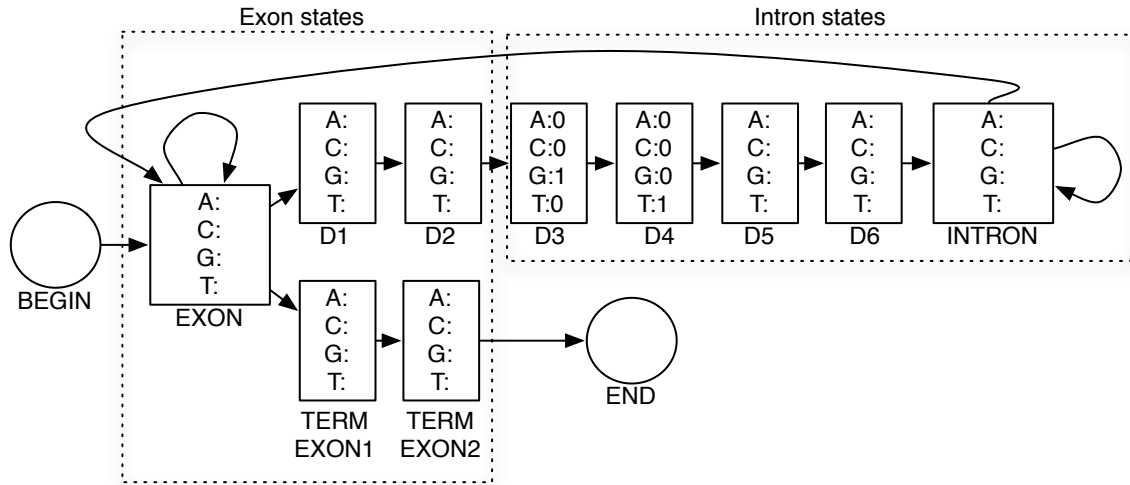


Figure 2: The HMM from HW#3

- (a) (20 pts) Draw the structure of an updated HMM that includes states for modeling the donor splice site. Indicate any emission probabilities that are known before training the model.



- (b) (5 pts) In HW #3, you were given a set of training sequences for which the positions of exons and introns were known. If you were not given any sequences for which the positions of exons and introns were known, how would you train your model?

The Baum–Welch (EM) algorithm can be used to train the parameters of an HMM with unlabeled sequences.

4. (25 pts) You are given a set of homologous nucleotides related by the tree in Figure 3.

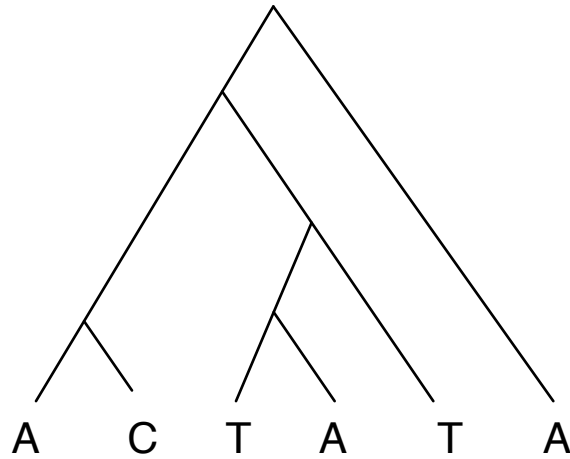
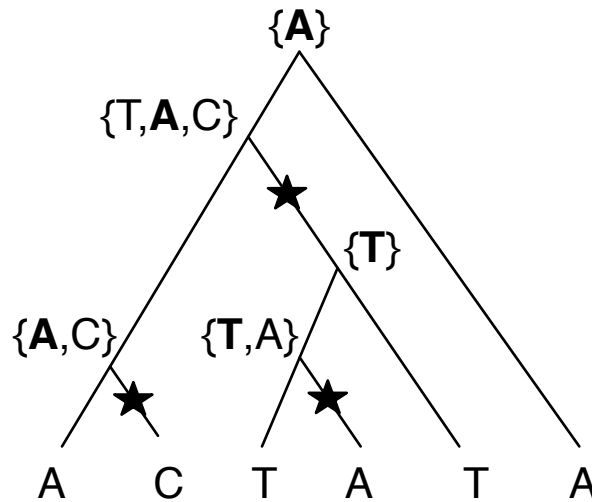


Figure 3: The tree

- (a) (20 pts) Use Fitch's (unweighted) parsimony algorithm to compute the minimum number of substitutions needed to explain the data at the leaves. Be sure to give (1) the sets of possible states at the internal nodes calculated during the leaves \rightarrow root stage, (2) the ancestral states at the internal nodes picked during the root \rightarrow leaves stage, and (3) the number of substitutions calculated.



Ancestral states are in **bold** and substitution events are indicated by stars on the branches. *Three* substitutions are required for this data and tree.

- (b) (5 pts) Which node in this tree is the outgroup? Briefly explain why the node you have selected is the outgroup.

The rightmost leaf is the outgroup in this tree. This leaf is the outgroup because all of the other leaves share a more recent common ancestor than they do with this leaf.