

BMI/CS 576 Fall 2009
Midterm Exam

Prof. Colin Dewey

Friday, October 30th, 2009 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 5).

Problem	Score	Max Score
1	_____	10
2	_____	5
3	_____	10
4	_____	10
5	_____	15
Total	_____	50

1. (10 pts) Suppose we are sequencing the string **GTACTTACC** via fragment assembly and by finding a Hamiltonian path through the overlap graph of the fragments. Give *four* distinct fragments (reads) of length *four* such that the following two conditions hold.
 - There exists an assembly of the reads that reconstructs the original string.
 - The *greedy* algorithm for finding the shortest superstring fails to give an optimal solution.

Show how the fragments you have chosen satisfy the two conditions.

The reads **GTAC** and **TACC** must be included as they contain the first and last characters, respectively, of the superstring. This leaves **TACT**, **ACTT**, **CTTA**, and **TTAC** as candidates for the other two reads. The key point in this problem is that unless you include **TACT** or **TTAC** as reads, **GTAC** will be joined with **TACC**, and you will get a suboptimal solution.

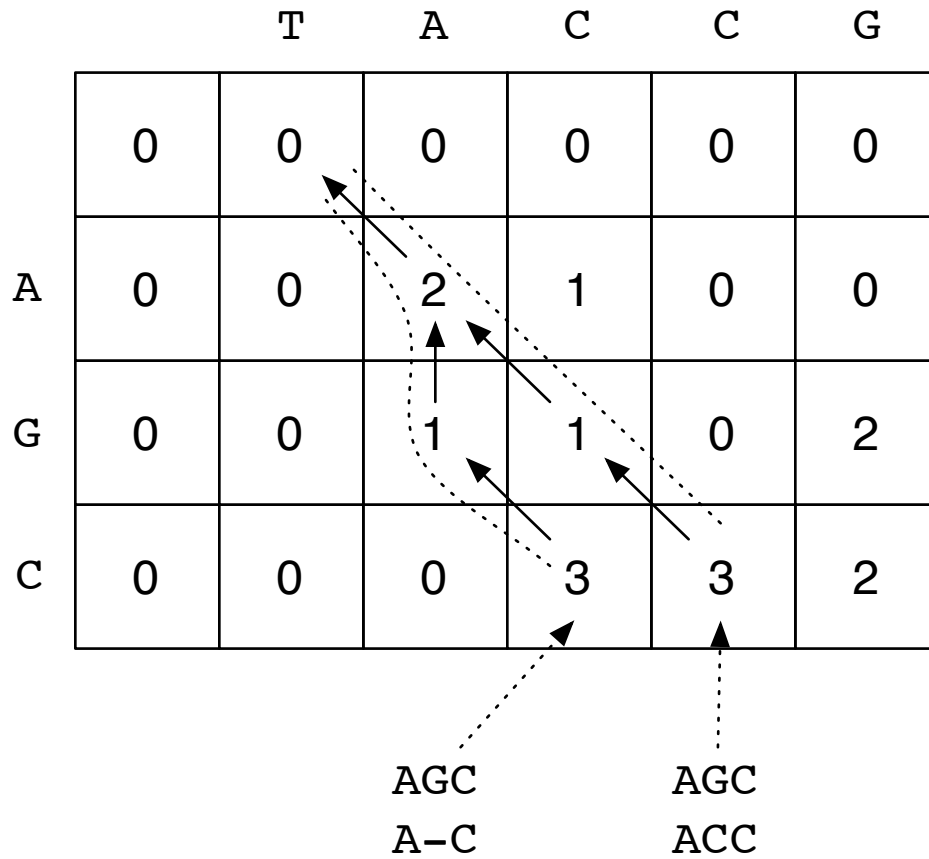
Thus, the best answer to this problem is the set $\{\text{GTAC}, \text{ACTT}, \text{CTTA}, \text{TACC}\}$. There exists an assembly of these reads that reconstructs the original string: **GTAC** \rightarrow **ACTT** \rightarrow **CTTA** \rightarrow **TACC**. However, the greedy algorithm will first pick the two length 3 overlaps between these reads: **GTAC** \rightarrow **TACC** and **ACTT** \rightarrow **CTTA**. Using these edges leaves us with two contigs **GTACC** and **ACTTA**, which have no overlap, and the greedy algorithm will either return **GTACCACTTA** or **ACTTAGTACC**. Both of these superstrings have length 10, which is greater than the optimal length of 9.

If a specific tiebreaker is assumed for the greedy algorithm (so that it is deterministic), it is also possible to include **TACT** or **TTAC** in the set of reads.

2. (5 pts) With sequencing-by-hybridization (SBH) data, a confused student attempts to determine a superstring by finding a Hamiltonian path (instead of an Eulerian path) through the SBH graph. Will such a path correspond to a valid superstring? Why or why not?

Not necessarily. In the SBH problem, a superstring is valid if it contains all k -mers in the spectrum. This means that a superstring's path through the SBH graph contains all edges (each of which represents one k -mer in the spectrum). A Hamiltonian path traverses each node exactly once, but may not traverse all edges. Thus, a Hamiltonian path may not give a valid superstring. For example, if the true superstring is **ATCAT** and $k = 3$, the spectrum is **ATC**, **TCA**, **CAT** which forms a length three cycle in the SBH graph. A Hamiltonian path will only include two of the edges in the cycle, producing an invalid superstring.

3. (10 pts) Find all optimal local alignments of the sequences AGC and TACCG. Assume a linear gap penalty with parameters $match = 2$, $mismatch = -1$, and $space = -1$. Give all values in the dynamic programming matrix and the traceback path for each optimal local alignment.



4. (10 pts) Suppose we wish to perform global alignments with a gap function

$$w(k) = \max(g_{short} + s_{short} \cdot k, g_{long} + s_{long} \cdot k)$$

where, $g_{short} > g_{long}$ and $s_{short} < s_{long}$ (all parameters are negative). In other words, each gap will be penalized with either a “short” gap function or a “long” gap function (with longer gaps tending to be penalized by the “long” gap function because $s_{short} < s_{long}$). We can implement this scoring function by introducing two additional dynamic programming matrices to the standard affine gap penalty algorithm. The matrix entries will be defined as:

- $M(i, j)$: best score of aligning $x_{1\dots i}$ and $y_{1\dots j}$ with x_i aligned to y_j
- $I_x^{short}(i, j)$: best score of aligning $x_{1\dots i}$ and $y_{1\dots j}$ with x_i aligned to a “short” gap.
- $I_y^{short}(i, j)$: best score of aligning $x_{1\dots i}$ and $y_{1\dots j}$ with y_j aligned to a “short” gap.
- $I_x^{long}(i, j)$: best score of aligning $x_{1\dots i}$ and $y_{1\dots j}$ with x_i aligned to a “long” gap.
- $I_y^{long}(i, j)$: best score of aligning $x_{1\dots i}$ and $y_{1\dots j}$ with y_j aligned to a “long” gap.

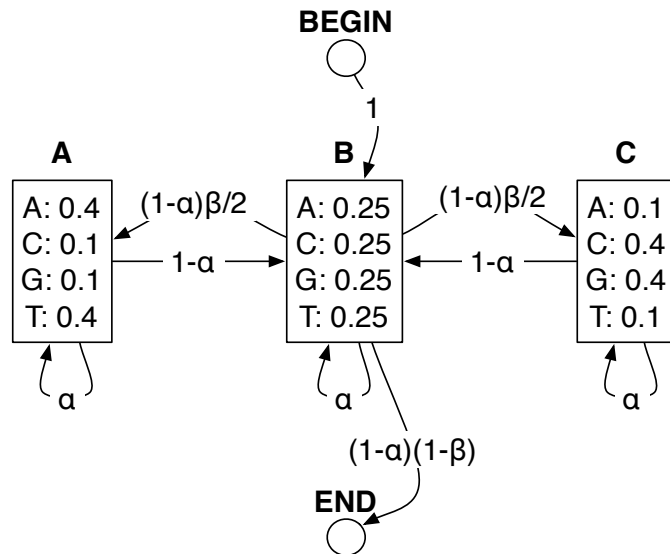
Give the dynamic programming recurrences for computing the values in these matrices. As before, assume that a gap in one sequence may not be followed by a gap in the other sequence.

$$\begin{aligned}
 M(i, j) &= S(x_i, y_j) + \max \begin{cases} M(i-1, j-1) \\ I_x^{short}(i-1, j-1) \\ I_y^{short}(i-1, j-1) \\ I_x^{long}(i-1, j-1) \\ I_y^{long}(i-1, j-1) \end{cases} \\
 I_x^{short}(i, j) &= s_{short} + \max \begin{cases} M(i-1, j) + g_{short} \\ I_x^{short}(i-1, j) \end{cases} \\
 I_y^{short}(i, j) &= s_{short} + \max \begin{cases} M(i, j-1) + g_{short} \\ I_y^{short}(i, j-1) \end{cases} \\
 I_x^{long}(i, j) &= s_{long} + \max \begin{cases} M(i-1, j) + g_{long} \\ I_x^{long}(i-1, j) \end{cases} \\
 I_y^{long}(i, j) &= s_{long} + \max \begin{cases} M(i, j-1) + g_{long} \\ I_y^{long}(i, j-1) \end{cases}
 \end{aligned}$$

5. (15 pts) A crazy biologist hypothesizes that genome sequences are composed of three types of segments: A, B, and C. Type A segments are composed of approximately 40% A, 10% C, 10% G, and 40% T. Type B segments are composed of roughly equal frequencies of the four bases. Type C segments are composed of approximately 10% A, 40% C, 40% G, and 10% T. Type A segments are never adjacent to Type C segments. Only Type B segments may be at the beginning or end of a sequence. The distributions of the lengths of the three types of segments are the same, $P(\ell) = \alpha^{\ell-1}(1 - \alpha)$. Type A segments occur at the same frequency as Type C segments.

- (a) (10 pts) Give the state transition diagram for a hidden Markov model that models genome sequences according to this hypothesis. Include emission and transition probabilities in the diagram.

To fully parametrize the model, we assume that the distribution of the *number* of B segments, n_B , (not the length of a single B segment) in a sequence is given by $P(n_B) = \beta^{n_B-1}(1 - \beta)$.



- (b) (5 pts) Given a new sequence, how would you test to see if it is best explained by this hypothesis, or by a simpler model that assumes that sequences are only composed of Type B segments? Name any algorithms that should be used for this task.

Let M_{crazy} and M_{null} denote the crazy and simple models, respectively. Given a new sequence, x , we can determine which model best explains x by computing $P(x|M_{crazy})$ and $P(x|M_{null})$ and choosing the model that gives the higher likelihood. These probabilities are computed for a HMM via the Forward algorithm.