BMI/CS 576 Fall 2016
Midterm Exam

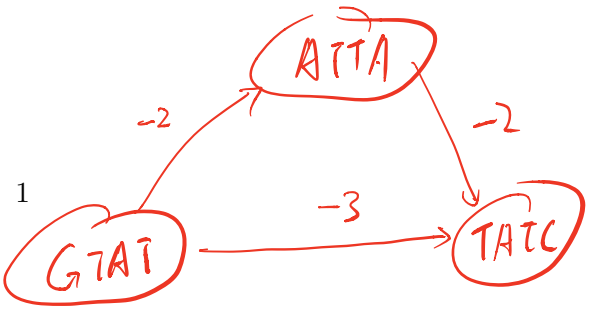
Prof. Colin Dewey

Thursday, October 27th, 2016 9:30am-10:45am

Name: KEY

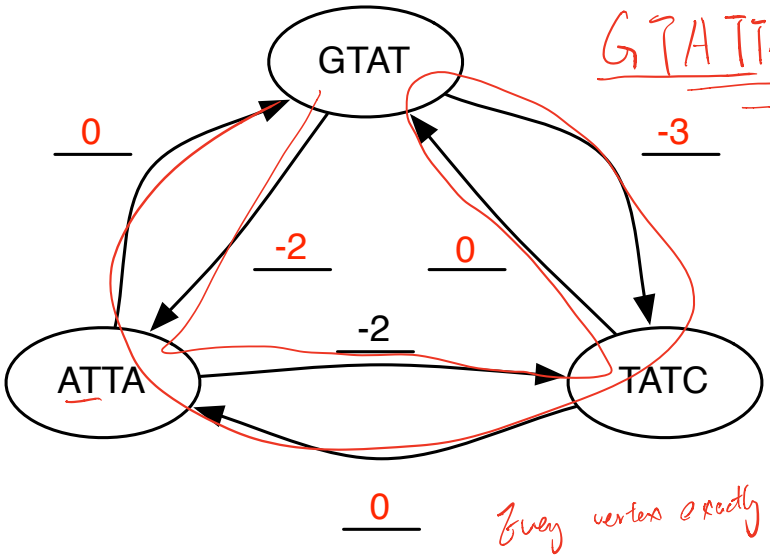
Write your answers on these pages and show your work. You may use the back sides of pages as necessary. Before starting, make sure your exam has every page (numbered 1 through 7). You are allowed 2 (double-sided) pages of notes. Calculators are *not* allowed.

Problem	Score	Max Score
1		25
2		25
3		25
Total		75



1. (25 points) Suppose that you are given the following three reads that are derived from some short genome: `ATTA`, `GTAT`, and `TATC`.

   (a) (5 points) Suppose you take the fragment assembly approach to assembling the genome given these reads and construct the overlap graph for the reads, with negated edge weights. Fill in the missing edge weights in the graph below.



   (b) (5 points) Give (i) one minimum weight Hamiltonian path through the graph in (a) *and* (ii) the superstring corresponding to that path.

   i. There is one minimum weight Hamiltonian path through the graph: `GTAT` → `ATTA` → `TATC`

   ii. This path corresponds to the superstring `GTATTATC`

   (c) (5 points) Would the greedy algorithm give a minimum length superstring in this case? Explain your reasoning.

   No. The greedy algorithm would first select the edge `GTAT` → `TATC`, and then either edge `TATC` → `ATTA` or `ATTA` → `GTAT`, resulting in the superstring `GTATCATTA` or `ATTAGTATC`, respectively, which are both nine bases long. The minimum length superstring is eight bases long.

GTATTATC

(d) (6 points) Suppose that the superstring you determined in part (b) was the true genome. Further, suppose that instead of being given reads, you are given the k-mer spectrum for that genome. What is the smallest that k can be such that the Eulerian path approach to reconstructing the genome will succeed? Explain your reasoning.

For the Eulerian path approach to succeed, each k-mer of the genome must be present in the genome exactly once. Thus, we seek the smallest value of k such that this property holds. For the minimum length superstring determined in part (b), `GTATTATC`, the smallest value of k is four, because all 4-mers of the superstring are unique, and there is one 3-mer that is not unique (`TAT`).

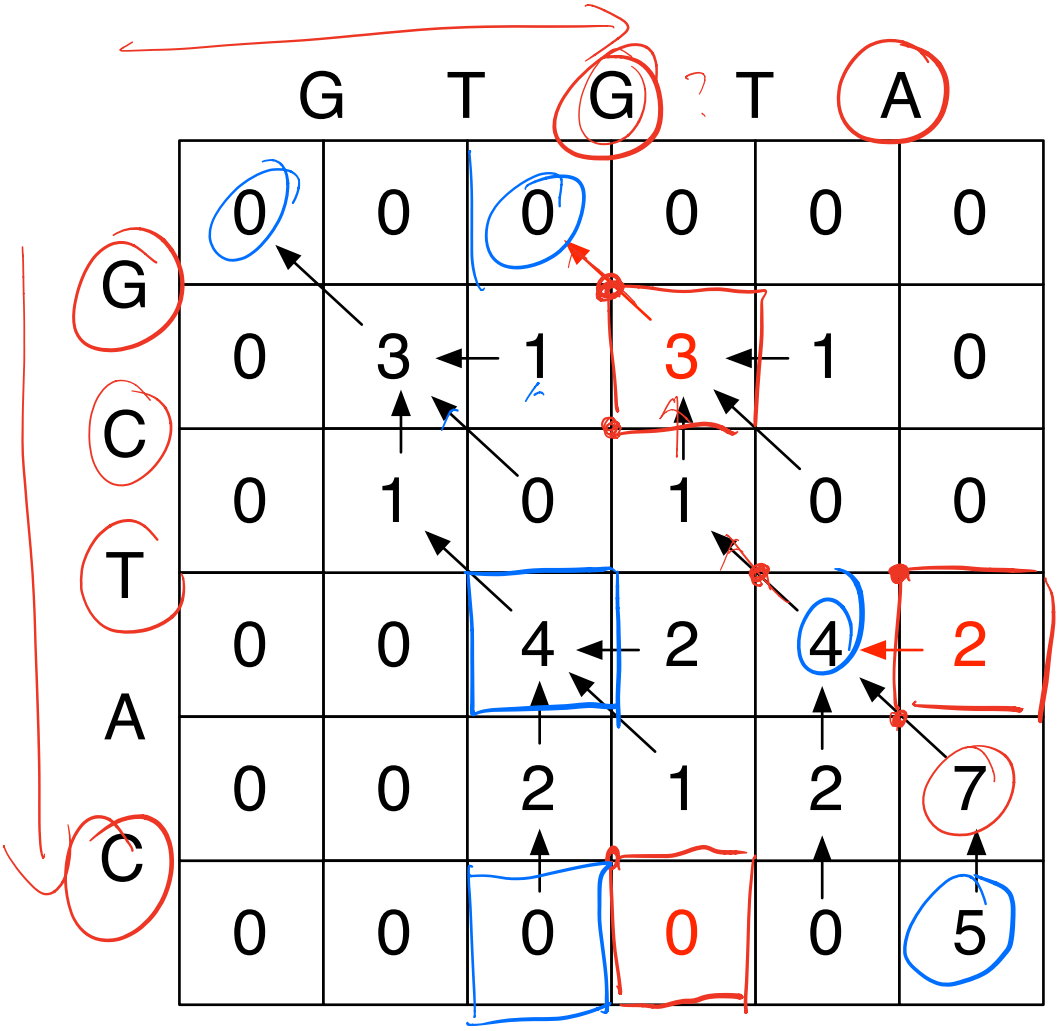
(e) (4 points) Considering only the structure of the graph in part (a) (i.e., ignoring the vertex labels and edge weights), does that graph have an Eulerian cycle? Explain your reasoning.

yes

This graph is complete, with each vertex having $indegree(v) = outdegree(v) = 2$, and is therefore balanced. Since this graph is balanced, it must have an Eulerian cycle.

2. (25 points) Consider the dynamic programming matrix for the *local* alignment of the sequences `GCTAC` and `GTGTA` with a linear gap penalty with parameters $match = 3$, $mismatch = -3$, and $space = -2$. Recall that the recurrence for this dynamic programming problem is

$$M[i,j] = \max \begin{cases} M[i-1,j-1] + S(x_i, y_j), \\ M[i-1,j] + space, \\ M[i,j-1] + space, \\ 0 \end{cases}$$



(a) (5 points) Fill in the (i) values *and* (ii) traceback pointers for the empty cells in the dynamic programming matrix above.

(b) (5 points) Give (i) the optimal local alignment score and (ii) a local alignment that achieves this score.

i. 7

ii. `GCTA`
`G-TA`

(c) (10 points) Using this matrix, determine the longest prefix of the first sequence (`GCTAC`) such that for some prefix of the second sequence (`GTGTA`), the optimal *local* alignment of the two prefixes is the same as the optimal *global* alignment of the same two prefixes. Briefly explain how you used the matrix to determine your answer.

For the ith prefix of the first sequence and the jth prefix of the second sequence, the score of an optimal local alignment of that pair of prefixes can be found by

$$\max_{\substack{0 \le i' \le i \\ 0 \le j' \le j}} M[i', j']$$

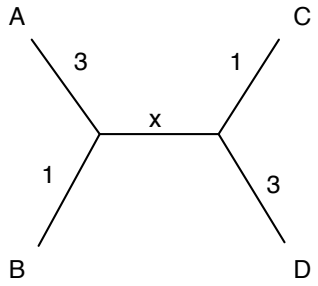
i.e., finding the maximum entry (i^*, j^*) in the submatrix with top-left entry $(0, 0)$ and bottom-right entry (i, j). Note that every global alignment of a pair of sequences is also a valid local alignment, and thus the score of an optimal local alignment is always greater than or equal to the score of an optimal global alignment. For the optimal local alignment of the two prefixes to also be an optimal global alignment, we must have that $(i^*, j^*) = (i, j)$, and that the traceback from entry (i, j) ends at entry $(0, 0)$. Thus, we seek the entry in the matrix (i, j) with largest i such that the criteria in the previous sentence are met. For this matrix, that entry is $(3, 2)$, which has score 4 and corresponds to the prefix `GCT` of the first sequence being aligned to the prefix `GT` of the second sequence.

(d) (5 points) Suppose that we modify the local alignment algorithm such that we start our traceback of an alignment from the lower-right entry of the dynamic programming matrix (as in global alignment) and traceback from that value until we hit an entry with value zero. Briefly describe what task this algorithm solves *in general* (e.g., not specifically for the sequences in this problem).

An entry $M[i, j]$ in matrix represents the optimal score of an alignment of some *suffix* of the ith *prefix* of the first sequence with some *suffix* of the jth *prefix* of the second sequence. Thus, the entry $M[m, n]$ (the lower-right entry) represents the optimal score of a (global) alignment of some *suffix* of the first sequence with some *suffix* of the second sequence. And an alignment achieving that score can be obtained by tracing back from the lower-right entry until a zero is reached.

3. (25 points) Phylogenetic trees

(a) (5 points) Given the true unrooted tree below, for what *values of the branch length* x will the UPGMA algorithm correctly reconstruct the structure of the tree (ignoring the estimated branch lengths), given pairwise distances between the four leaves? Briefly justify your answer. You should consider the rooted tree produced by UPGMA equivalent to the unrooted tree that one would obtain by removing the root node.



For UPGMA to reconstruct the correct tree structure, it must either pick (A,B) or (C,D) as the first pair to merge. If it picks either of those pairs first, any ordering of the last two merges will result in a rooted tree that is equivalent to the true unrooted tree given above. For (A,B) or (C,D) to be picked first by UPGMA we would need $d_{AB} = d_{CD} = 3 + 1 = 4 < d_{BC} = 1 + x + 1 = 2 + x$. Thus, for $x > 2$, UPGMA will correctly reconstruct the structure of the tree.

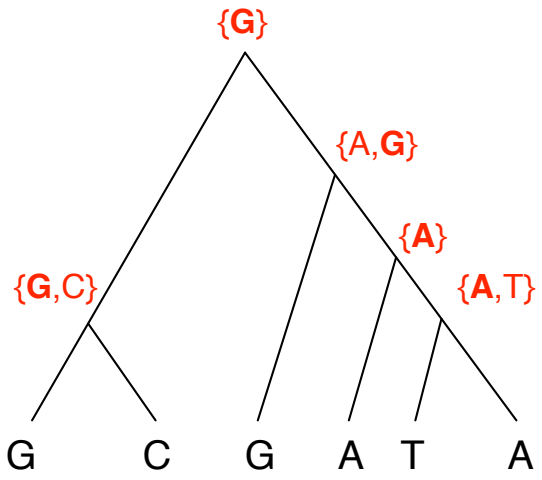
(b) (5 points) Given the same unrooted tree as in part (a), for what *values of the branch length* x, if any, do the pairwise distances between the four leaves obey the properties of an ultrametric? Briefly justify your answer.

There are no values of the branch length x such that the pairwise distances between the four leaves are an ultrametric. To see this, note the pairwise distances between the three leaves, A, C, and D. These distances are

$$\begin{aligned} d_{AC} &= 3 + x + 1 = 4 + x \\ d_{AD} &= 3 + x + 3 = 6 + x \\ d_{CD} &= 3 + 1 = 4 \end{aligned}$$

For $x = 0$, two of these distances are equal and smaller than the third. For $x > 0$, all three distances are different. For an ultrametric, two of these distances need to be equal and greater than or equal to the third. Thus, no value of x results in these distances satisfying the ultrametric property.

(c) (10 points) For the rooted tree below with extant characters at the leaves, give (i) the minimum unweighted parsimony cost of this tree and (ii) an assignment of characters to the ancestral nodes that achieves this cost.



i. 3. Three union operations were required in computing the R_i sets at the ancestral nodes, and thus the minimum cost (# of changes) is three.

ii. See bold characters in tree above.

(d) (5 points) Briefly explain why one might prefer to use weighted parsimony to compute the cost of a tree instead of unweighted parsimony.

One would prefer to use weighted parsimony if certain changes between characters are more or less likely than other changes. In this case one would want the cost of a lower probability change to be higher than that of a higher probability change.