BMI/CS 576 Fall 2019
Midterm Exam

Prof. Colin Dewey

Wednesday, October 30th, 2019 5:30-7:00pm

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 9). You are allowed 2 (double-sided) pages of notes. Calculators are *not* allowed.

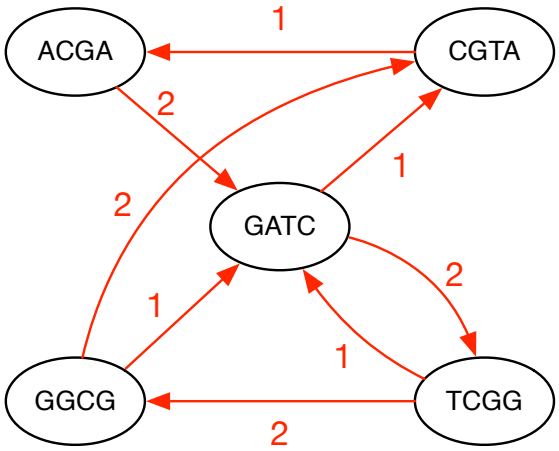
Problem	**Score**	**Max Score**
1		30
2		30
3		30
Total		90

1. (30 points) Suppose that we perform a shotgun sequencing experiment and obtain the following set of five error-free reads (listed in lexicographical order):

reads
ACGA
CGTA
GATC
GGCG
TCGG

(a) (15 points) Suppose that we use the fragment assembly approach.

i. (8 points) Restricting to overlaps with non-zero length, draw the **eight** edges *with weights* in the overlap graph below that would be considered in this approach.



ii. (4 points) There is one optimal assembly that could be output by this approach. Give the assembly and its corresponding path in the graph.

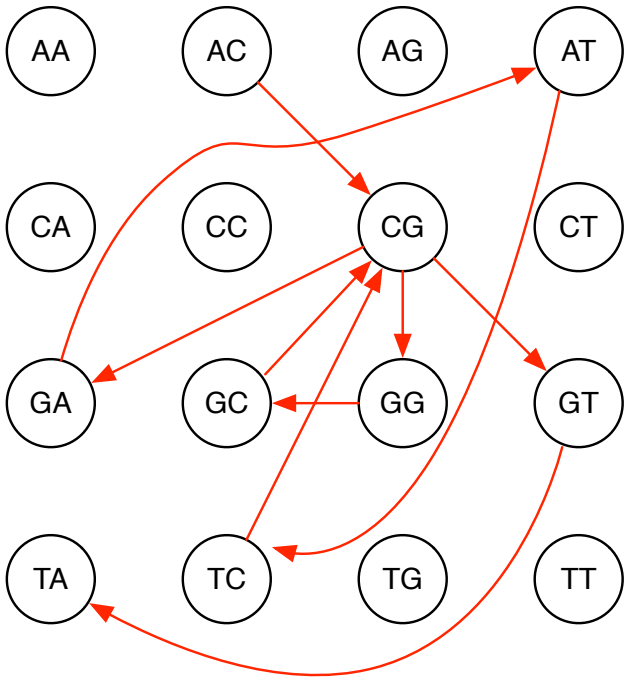
assembly = ACGATCGGCGTA
path = ACGA → GATC → TCGG → GGCG → CGTA

iii. (3 points) Will the greedy algorithm for fragment assembly succeed in identifying this optimal assembly? Explain your reasoning.

Yes. All edges that are part of the optimal assembly path have a weight of 2, which is larger than the weights of all edges that are not part of the optimal assembly. Therefore, during the greedy algorithm the edges that are part of the optimal assembly will be the first four edges that are selected and will form a Hamiltonian path.

(b) (15 points) Suppose that we instead use the *de Bruijn* approach to assembly (in which a k-mer spectrum is obtained from the reads) with $k = 3$.

i. (8 points) Draw the edges in the SBH graph below that would be used in this approach.



ii. (4 points) There is more than one possible assembly that could be output by this approach. List *all* possible assemblies. For each assembly, specify its corresponding path in the graph.

There are two possible assemblies:

A. assembly = ACGATCGGCGTA
path = AC → CG → GA → AT → TC → CG → GG → GC → CG → GT → TA

B. assembly = ACGGCGATCGTA
path = AC → CG → GG → GC → CG → GA → AT → TC → CG → GT → TA

iii. (3 points) Considering the original set of reads, are all of these assemblies equally valid? Explain your reasoning.

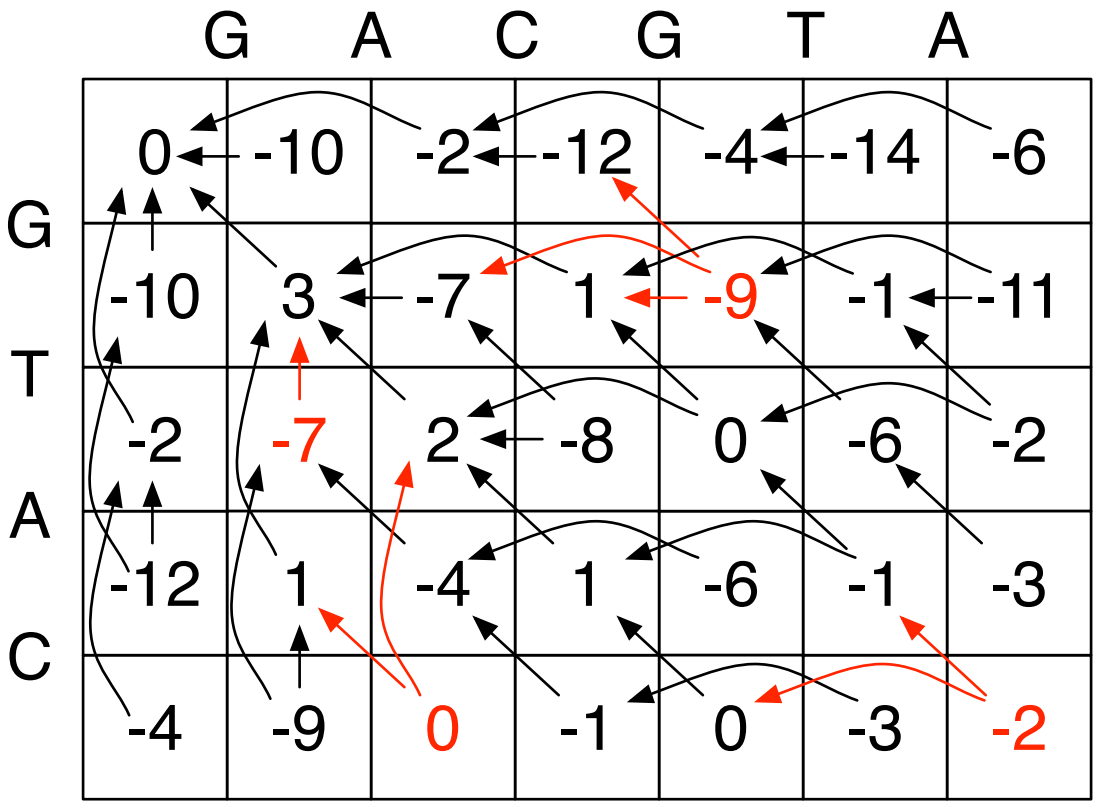
No. Assembly (B) (ACGGCGATCGTA) does not contain two of the reads (TCGG and ACGA) as substrings and there are no reads that cover its first two bases. In contrast, assembly (A) (ACGATCGGCGTA) contains all reads as substrings and is the solution to the shortest superstring problem (part a.ii.).

2. (30 points) Suppose we wish to model a form of molecular evolution in which insertion and deletion events involving two bases are much more common than insertion and deletion events involving individual bases. To perform a *global* alignment of two sequences with such a model, we modify the dynamic programming recurrences to be:

$$M[i,j] = \max \begin{cases} M[i-1,j-1] + S(x_i, y_j), \\ M[i-1,j] + space, \\ M[i,j-1] + space, \\ M[i-2,j] + double\_space, \\ M[i,j-2] + double\_space \end{cases}$$

where the fourth case in the max is only considered for $i \geq 2$ and the fifth case in the max is only considered for $j \geq 2$.

Using these modified recurrences, we wish to align the the sequences `GTAC` and `GACGTA` with parameters $match = 3$, $mismatch = -1$, and $space = -10$ and $double\_space = -2$. Below is the dynamic programming matrix with a few entries missing.



(a) (20 points) Fill in the (i) values *and* (ii) traceback pointers for the empty cells in the dynamic programming matrix above.

(b) (2 points) Give the optimal global alignment score.

The optimal global alignment score is -2, which is the lower-right entry in the matrix.

(c) (8 points) Give all global alignments that achieve the optimal score.

There are four optimal global alignments:

```
G--TAC
GACGTA

GT--AC
GACGTA

GTA--C
GACGTA

GTAC--
GACGTA
```

3. (30 points) Given below are three DNA sequences and all pairwise optimal global alignments between them with a linear gap penalty and parameters $match = +2$, $mismatch = -1$, and $space = -1$.

sequences

```
s1: ATGC
s2: CTGA
s3: TGCTG
```

optimal pairwise alignments
(match = +2, mismatch = -1, space = -1)

```
s1: ATGC
s2: CTGA
```

```
s1: ATGC--
s3: -TGCTG
```

```
s2: --CTGA
s3: TGCTG-
```

(a) (6 points) Compute the alignment score for each of the three given pairwise alignments.

$score(\text{s1} - s2) = -1 + 2 + 2 + -1 = 2$
$score(\text{s1} - s3) = -1 + 2 + 2 + 2 + -1 + -1 = 3$
$score(\text{s2} - s3) = -1 + -1 + 2 + 2 + 2 + -1 = 3$

(b) (4 points) Suppose we wish to compute a multiple alignment of the three sequences using a sum-of-pairs objective function and the same parameters as for the pairwise alignments. Using only the scores you computed in part (a), give an **upper bound** for the score of the optimal multiple alignment of the three sequences. Explain your reasoning.

Using only the scores from (a), an upper bound for the score of the optimal multiple alignment of the three sequences is the sum of the three optimal pairwise alignments, which is $2 + 3 + 3 = 8$. This is because the sum-of-pairs score of a multiple alignment with a linear gap penalty is equivalent to the sum of all pairs of pairwise alignments that it implies. This sum can thus be no larger than the sum of the scores of the optimal pairwise alignments.

(c) (12 points) We will use the *star alignment* approach to construct multiple alignments of the three sequences using the given pairwise alignments. Recall that star alignment constructs a multiple alignment by combining the optimal pairwise alignments of a selected "center" sequence to all other sequences. We will construct three multiple alignments, one for each possible selection of the center sequence. Given to you is the solution for the alignment using $s2$ as the center sequence. Compute the other two multiple alignments as well as their sum-of-pairs scores.

s1 as center	s2 as center	s3 as center
s1: ATGC-- s2: CTGA-- s3: -TGCTG	s1: --ATGC s2: --CTGA s3: TGCTG-	s1: ATGC--- s2: ---CTGA s3: -TGCTG-
alignment	alignment	alignment
score: 5	score: 5	score: 2

(d) (4 points) Suppose that we wish to use the technique of *iterative refinement* to improve upon a multiple alignment that was originally constructed by the star alignment algorithm (for example, the alignments you constructed in part (c)). If we remove the sequence used as the "center" for the star alignment, and realign it to the remaining multiple alignment, is it possible for the score of the entire multiple alignment to increase? Explain your reasoning.

No. When the center sequence is removed and realigned to the remaining multiple alignment, the only components of the sum-of-pairs score that can change are the scores of the implied pairwise alignments between the center sequence and each other sequence. However, the original star alignment already implies optimal pairwise alignments for all pairwise alignments involving the center. Thus the sum-of-pairs score cannot increase.

(e) (4 points) Consider the scenario in part (d), but with one of the "non-center" sequences removed and realigned to the remainder of the alignment. Is it possible for the score of the entire multiple alignment to increase? Explain your reasoning.

Yes. Within the original star alignment, the alignment of each "non-center" sequence was only guided by its optimal pairwise alignment with the center sequence and did not take into account any of the other sequences. If a "non-center" sequence is removed and realigned to the remaining multiple alignment, the realignment of the sequence takes into account all other sequences, not just the center.

For example, in part (c), consider the star alignment using s3 as the center, which has a score of 2. If we remove and realign s1 (a non-center sequence), we can obtain a new multiple alignment that is the same as the star alignment using s2, which has a higher score (5).