CM122/CM222: Algorithms in Bioinformatics **Spring 2021**

Midterm Exam

- This exam is take-home. It is due Tuesday, April 27th, at 2 P.M. Pacific Time.
- The exam is open notes, but communication between students about exam questions is prohibited. If you have any questions, please email the instructors.
- Questions are free response. Solutions with no justification or work shown will receive no credit, regardless of correctness.
- If you have access to a printer or scanner, please print the exam and fill it out, then scan it into a PDF and upload it to CCLE. If you do not have access to a printer or scanner, you can hand-write the exam, take pictures with your phones, and upload to CCLE.
- If you cannot access CCLE, contact the instructors to arrange a way to submit it.

Name and ID:

1. Multiple choice, and true/false questions. (15 points)

   a (5 points) Suppose a coverage of λ and that 25% of all your reads have sequencing errors. What is the probability that X or fewer reads with errors span a certain position? We use the same notation as class lecture 3. Shortly explain your answer in the box below. Some partial credit might be given.

   B
   - ☐ 0.25ppois(X, λ)
   - ☒ ppois(X, 0.25λ)
   - ☐ ppois($0.25X$, λ)
   - ☐ None of the above.

   b (2 points each) Select all that are true (no explanations needed for these questions).

   F ☐ As sequencing coverage decreases, the consensus algorithm for identifying single nucleotide polymorphisms becomes more accurate.

   T ☐ For the indexing/hashing read alignment algorithm, splitting reads into D substrings allows a mismatch tolerance of $D - 1$.

   F ☐ Doubling the interval between checkpoints in the FM Index (used by Bowtie) speeds up the algorithm at the cost of additional memory usage.

   F ☐ Suppose there is no sequencing error. When you use k-mers to build a De Bruijn graph, shorter k-mers generally produce less branching (or less tangling) in the graph than longer k-mers do.

   T ☐ Given a k-mer spectrum, an Euler path on a De Bruijn graph provides the same solution as a Hamiltonian path on an overlap graph with edges for overlaps of $k - 1$.

2. Re-sequencing (20 points total)

For problems (a-b) below, assume a genome length of $N = 3 * 10^9$ bases, and that $M = 9 * 10^8$ reads have been sequenced, each of which are $L = 100$ bases long.

a) (5 points) What is the average coverage?

$\frac{M*L}{N} = 30X$

b) (5 points) What is the percentage of bases expected to have exactly 20 reads covering the position?

dpois(20, 30) $\approx$ 0.0134

c) (5 points) How long will it take to align reads to a genome, given a genome length N, M reads of length L, where each nucleotide comparison takes t seconds, using the naive approach of "sliding" a read across the genome?

$N*M*L*t$

d) (5 points) Given a genome length N, reads of length L, and a mismatch tolerance of 1, how many rows do we expect a table to have using an index/hashing approach? How many genomic positions do we expect within each row of the table?

$4^{1/2}$ rows, $\frac{N}{4^{1/2}}$ entries / row

3. Burrows Wheeler Transform (30 points)

a. (20 points) Given that the Burrows Wheeler transform of a string is "CGGTC$CA", find the original string.

GATCCGC$

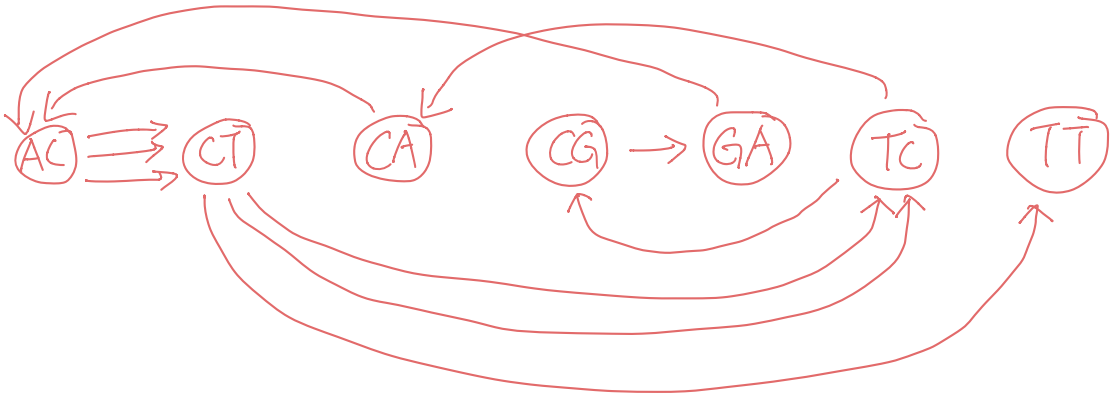
b. (10 points) Write the suffix array for the BWT matrix.

$GATCCGC	7	8
$G	1	2
$GATCCG	6	7
$GAT	3	4
$GATC	4	5
$	0	1
$GATCC	5	6
$GA	2	3

4. Assembly (35 points total)

a) (25 points) Given the genome "ACTCACTCGACTT", identify the 3-mers and construct a De Bruijn graph from them.

[ACT, CTC, TCA, CAC, ACT, CTC, TCG, CGA, GAC, ACT, CTT]



b) (10 points) Is this De Bruijn graph unique to this genome? If not, provide another genome that can be generated from the graph.

No, ACTCGACTCACTT