CM122/CM222: Algorithms in Bioinformatics **Spring 2020**

Midterm Exam

- This exam is take-home. It is due Tuesday, April 28th, 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 Gradescope. 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 Gradescope.
- If you cannot access Gradescope, contact the instructors to arrange a way to submit it.

Name and ID: Answer Key

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

a (5 points) Suppose 10% of all your reads have sequencing errors. What is the probability that more than X correct (error-less) reads 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.

□ $\sum_{i=X}^{\infty}$ dpois(i, 0.9λ)
□ $\sum_{i=X}^{\infty}$ dpois(0.9i, λ)
□ $\sum_{i=X}^{\infty}$ 0.9dpois(i, λ)
□ None of the above.

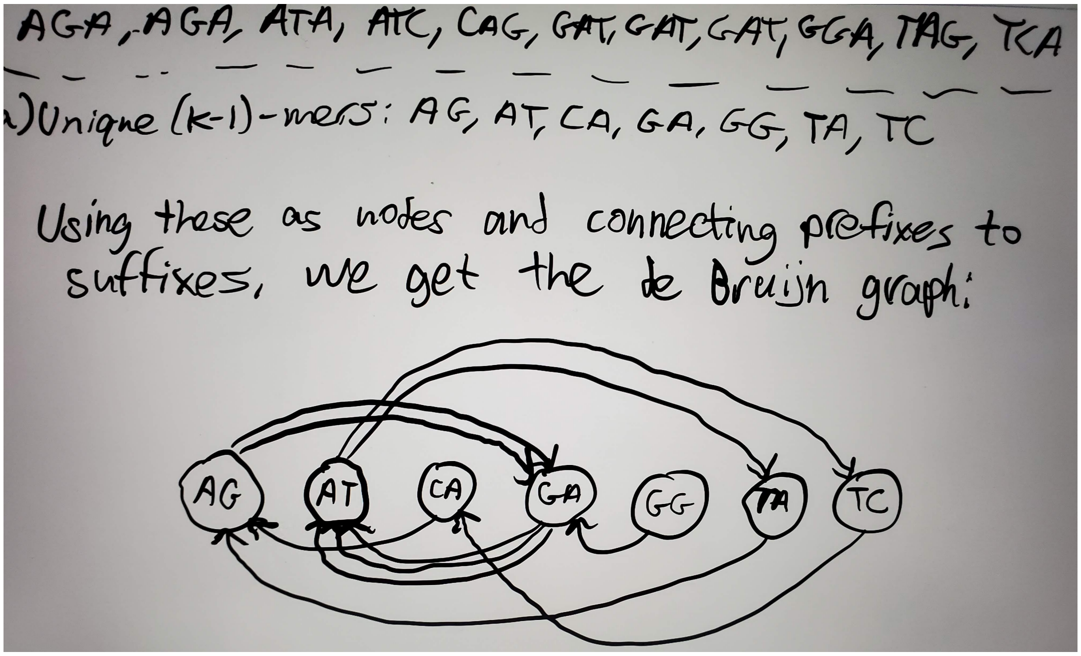
Answer: None of the above/A (with sufficient reasoning provided). dpois(i, λ) looks at the probability that exactly i reads span a certain position in a Poisson distribution defined by the parameter λ. Because only 90% of our reads do not have errors, we need to adjust the lambda parameter to reflect the distribution of usable reads, by scaling our λ parameter. We then sum over all positions greater than X.

Note: Though A is close, it reflects the probability of **at least** X error-less reads, not more than X error-less reads. It was not our intention to trick people on this part of the question, so we will give full credit to those who answered A with a sufficient explanation.

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

□ Suppose there is no sequencing error. When you overlap k-mers to assemble the genome, shorter k-mers generally produce better accuracy than longer k-mers do.
✓ There can be multiple solutions for finding an Eulerian path in a De Bruijn graph.
□ De Bruijn graphs are always Eulerian.
□ The length of a repeated segment in the genome can be inferred based on the average sequencing coverage at each position in the genome.
✓ The Velvet method uses Tour Bus error correction to remove bubbles in the De Bruijn graph.

Answer: I: False II: True III: False IV: We are tossing this question out and giving everyone points V: True



AGA, AGA, ATA, ATC, CAG, GAT, GAT, GAT, GGA, TAG, TCA

a)Unique (k-1)-mers: AG, AT, CA, GA, GG, TA, TC

Using these as nodes and connecting prefixes to suffixes, we get the de Bruijn graph:

2. Assembly (45 points total)

a) (25 points) Given the following k-mers, create a de bruijn graph via one of the methods discussed in the Compeau and Pevzner textbook chapter 3. Explain the steps being performed.
K-mers: AGA, AGA, ATA, ATC, CAG, GAT, GAT, GAT, GGA, TAG, TCA

Answer: See image above. Also accepted: the gluing method, as long is it generates the same graph.

b) (5 points) What are the start and end nodes for the de Bruijn graph? Is the graph Eulerian? If not, what edge would need to be added to make the graph Eulerian?

Answer: GG is the start node because one edge leaves but none enter. AT is the end node since 3 edges enter but only 2 leave. Thus, creating an edge from AT to GG would make the graph Eulerian.

c) (10 points) Traverse the graph and show the final assembled genome, listing the nodes traversed in order. If there are multiple possible paths, say so and pick one arbitrarily. Is your solution unique?

Answer: There are two possibilities, e.g. the solution is not unique. The possibilities are GGATAGATCAGAT or GGATCAGATAGAT (see below).

Possibility 1: GG → GA → AT → TA → AG → GA → AT → TC → CA → AG → GA → AT → builds GGATAGATCAGAT
Possibility 2: GG → GA → AT → TC → CA → AG → GA → AT → TA → AG → GA → AT → builds GGATCAGATAGAT

3. Burrows Wheeler Transform (10 points)

(10 points) Construct the Burrows Wheeler transform for the string "ELEAZAR".

Stage 1: Circular rotations
ELEAZAR$
$ELEAZAR
R$ELEAZA
AR$ELEAZ
ZAR$ELEA
AZAR$ELE
EAZAR$EL
LEAZAR$E

Stage 2: Lexicographic ordering
$ELEAZAR
AR$ELEAZ
AZAR$ELE
EAZAR$EL
ELEAZAR$
LEAZAR$E
R$ELEAZA
ZAR$ELEA

Stage 3: BWT is last column above. Thus:
Answer: RZEL$EAA

4. Re-sequencing (30 points total)

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

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

Answer: $M * L/N = (3 * 10^8) * 100/(3 * 10^9) = 10$
Also accepted if you left answer in terms of N, M, and L.

b) (10 points) What is the percentage of bases expected to have coverage 5x or greater?

Answer: 1 - ppois(4,10) = 97.075%. Also accepted if you left the answer in terms of λ and k.

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?

Answer: $N * M * L * t$ seconds

d) (10 points) How long will it take to align reads to a genome, given a genome length N, M reads of length L, index of kmers length L/3, where each nucleotide comparison takes t seconds, using an index/hashing approach? (When comparing the read to the genome, we compare the whole read, not just the two-thirds that aren't necessarily a perfect match)

Answer: $(N * 3 * M * L * t)/(4^{(L/3)})$ seconds