BIO 364, Section 1, Bioinformatics Algorithms, Dr. Perry Ridge

Name:

Complete the following questions. I have deliberately printed the exam on only one side of the pages so you can use the backs of pages as scratch paper, but I will only grade what is written on the front of each page and have provided adequate space for each question. To grade your exam, I must be able to read your writing/figures. Please take the necessary time to organize your work and write legibly. Include all your scratch paper stapled to the end. Be sure you read the question and follow the instructions. Many students lost points for being incomplete in their work. I've attempted in this redo to tell you more explicitly what I expect. Please write your name on the top of each page. Since this is a redo, it can only help your grade, so skipping questions you don't want to answer is reasonable. Additionally, question 1 on your redo can only be used to recover points lost on question 1 of the original exam, question 2 on your redo can only be used to recover points lost on question 2 of the original, etc. Meaning, if you've mastered one concept and got, for example, 100% on that topic on the first exam, you can't improve your grade by getting points on the same concept you already demonstrated mastery of on the first exam. You are welcome to use your book, any notes you've already taken, and your code. You may **not** use any other resources. This includes other students, people not in the class, the Internet, etc. I'm trusting each of you not to cheat, so please don't prove my trust incorrectly placed. Finally, you must sign this page in two places below. The first you indicate you followed all rules. The second you indicate that you have destroyed all physical and electronic copies of the exam. You may not keep a copy to prepare for the final, for your records, etc.

I promise I have followed all the rules of the exam outlined above. I have not discussed this exam with other students in the class, even if I'm already done, and will not until after the due date. I have not used Internet resources other than the Stepik textbook and have not received help from any other person or resource.

_____

Signature

I have deleted my electronic copy of this exam and emptied my trash (meaning, I have no way of retrieving a copy). Furthermore, I have not shared it with anyone. Finally, I have only made one physical copy of the exam, which I am turning in, and I have included any scratch paper I used.

_____

Signature

BIO 364, Section 1, Bioinformatics Algorithms, Dr. Perry Ridge

1. You have been given the whole genome sequence from a newly discovered bacterium and need to identify the origin of replication (*ori*). Unfortunately, this bacterium's genome is very long and is in fact the longest ever observed (~300B nucleotides)! In this somewhat bizarre bacterium, by divine intervention it was revealed to you that 5-mers with one mismatch could be helpful in your search, the *ori* region is only 350 base pairs in length, and that all the approaches we used for finding *ori* in E. *coli* in the textbook will work here. Please describe in detail how you would go about finding *ori* in this new genome. You need to justify/explain your choices and provide enough detail/examples that I am convinced you know what you are describing and are not superficially describing an approach with key words and not real understanding, and what you propose must be computationally feasible. For example, if you were to tell me that first you need to find the reverse complement of the genome, I would expect that you include something like "to find the reverse complement of a sequence, first I find the complementary bases and then reverse their order, for example, the reverse complement of ATTGC is GCAAT." This would demonstrate that you know how to find the reverse complement and that you know what a reverse complement is. (10 points)

BIO 364, Section 1, Bioinformatics Algorithms, Dr. Perry Ridge

2. I have provided you a set of homologous sequences from five different primates for a gene that is believed to be responsible for the higher intelligence observed in primates. We are interested in understanding where transcription factors bind upstream of the gene so that we can attempt to modulate expression of the gene in individuals with decreased cognition in a highly experimental treatment (i.e., this would never work in reality). We discussed several different approaches for finding motifs in sequences. You will need to use a GreedyMotif search and a GibbsSampler to find motifs for these sequences. Assume you are looking for 4-mers and make sure you work your algorithms in such a way that there's never a 4-mer with zero probability in your matrix. If there is a random component to an algorithm, make sure you clearly tell me what the "random" choice is and make sure that it is obvious to me that you understand what random means in the context of one, or both, of the algorithms. Make sure you report the score for each set of motifs. If you don't provide sufficient detail for me to see that you understand what to do, your writing is too messy for me to read, or the different steps are scattered across the page, giving you credit will be challenging. Finally, compare and contrast GreedyMotif search, Randomized Motif search, and the Gibbs Sampler. This includes, for example, relative strengths and weaknesses of each, which you would expect to get the most accurate results and why, and did the results of your work here match what you'd expect. GreedyMotif search starting on this page. (30 points, roughly 10 for each algorithm and 10 for your explanation)

```
GTTCAG
AATCAG
TATTCG
```

Complete your GibbsSampler starting here. Use two random starts and two iterations for each. I have included the same three sequences on this page so you don't have to flip back and forth between the two pages of the exam.

```
GTTCAG
AATCAG
TATTCG
```

3. Given the following set of **paired** kmers, reconstruct a genome using a de Bruijn graph. Show **all** your work, including your de Bruijn graph and the path you traveled in the graph. All nodes should be clearly labeled. Assume an insert size (the textbook called this *d*) of two and kmer length of three. Assembling a genome from these reads **is** possible. The assembled genome, even if correct, without the graph will not earn any points. (30 points)

```
AAT   ATT
ATT   TTC
ATT   TTC
ATT   TTG
CAA   GAT
CAT   CTT
GAT   ATT
TCA   GCT
TCA   TGA
TGA   CAT
TTC   TGC
TTG   TCA
TTG   TCA
```

4. In our work to algorithmically assemble genomes we made (at least) two assumptions about our data that are completely inconsistent with reality. **First**, we assumed there are no errors in our sequencing reads. However, Illumina sequencers, which is the most used next-generation sequencing technology, have an average error rate of ~1%. To further complicate things, the error rate changes across the sequence read. Typically, sequence reads are 100-250 base pairs in length. Close to the beginnings of the reads there are many fewer errors than near the end of the reads. **Second**, we assumed that the insert size (the textbook called this *d*) between the reads is identical for every pair of reads. In even the most stringent of protocols the insert size is never identical for all reads. Considering our algorithm for assembling read pairs above, answer the following two questions. **1)** If we remove these assumptions, what complications will occur with the algorithm as we learned it? **2)** If we remove these assumptions, would we need to modify our algorithm above? If yes, how? (For the record, De Bruijn graphs are the most popular approach for genome assembly and we do not have the luxury of making the assumptions above, but still manage to accurately assemble new genomes.) (30 points)