FALL 2017 BCB5200 Final Exam (Dec 7, 2017)

Name: _____
Total final points = 100 pts

Part I. Take-= Home Exam (70point) (Due: Dec 8, 2017 4:00pm)

1. **Suffix tree and overlap problem. (15 points)**
   a. **Consider the text T = ACGACA. Show the suffix tree for T with terminator $.**

   Sol>

   b. **Consider the text T1 = ACGACA and T2 = GACACG. Show how to find the longest common substring of the two strings. Explain how you search the longest common substring.**

   Sol>

2. Genome assembly by de Buijn graph problem. (20 points)
Let's suppose you have 6 bp reads as below:
TCGACA, GTCGAC, AAGACG, CGACAT, AGTCGA, GACATA, ACATCG, AAGTCG, GACATC, AAAGTC

a. Build a de Bruijn graph with k=5 (you need to draw the paths). You need to draw all paths from the sequences.

Sol>

b. Do you see any tip in the graph? If yes, then remove the tip in the graph. If no, report "No tip".

Sol>

c. Do you see any bubble in the graph? If yes, then remove the bubble in the graph (by removing any node in this exam). If no, report "No bubble".

Sol>

d. Do you see any sequencing error(s)? If yes, then report the error(s) and correct the errors in the reads.

Sol>

e. Finally, report the longest sequence after the error corrections.

Sol>

3. Gene finding problem. (15 points)

a. Describe an open reading frame (ORF). (1 sentence answer expected)

Sol>

b. Describe the purpose of performing a homology search with a DNA sequences? (1 sentence answer expected)

Sol>

c. Why is it relatively easy to identify ORFs in prokaryotic genomes then eukaryotic genomes by computer analysis? (1-3 sentences answer expected)

Sol>

d. How can gene annotation or prediction tools determine the difference between the stop codon in the intron and the actual stop codon at the end of the exon? (1-3 sentences answer expected)

Sol>

4. **Mutations and Variant Calling. (10 points)**
   a. **What is the difference between SNP and SNV?**
   Sol>

   b. **Why coverage (depth of coverage) is important in variant calling?**
   Sol>

5. **RNA-Seq Analysis. (10 points)**
   a. **What are the benefits of RNA-Seq compared to microarray?**
   Sol>

   b. **A test for a disease has a sensitivity of 95% and a specificity of 91%. You plan to screen a population in which the prevalence of the disease is 0.2% (Prevalence is a statistical concept referring to the number of cases of a disease that are present in a particular population at a given time, whereas incidence refers to the number of new cases that develop in a given period of time). What is the positive prediction value (PPV) that is a proportion of positive test that are true positives and represent the presence of disease?**
   Sol>