CS 175

Problem Set 4

Due: 20 November 2017 11:59pm

General Instructions

- The answer sheet for this problem set should be submitted as a PDF file. You may use any word processing software to create the answer sheet. The name of the PDF file to be submitted should follow the following format: [CS 175] < *Student Number* > − < *Last Name*, *First Name* > − Problem Set 4.pdf. For example: [CS 175] 190800001 - De la Cruz, Juan - Problem Set 4.pdf.
- If you have consulted references (books, journal articles, online materials, other people), cite them as footnotes to the specific item where you used the resource/s as reference.
- Submission of the problem set answers should be done via e-mail. Attach the PDF file, and write as the subject header of the e-mail: [CS 175] < *Student Number* > − < *Last Name*, *First Name* > − Problem Set 4. For example: [CS 175] 190800001 - De la Cruz, Juan - Problem Set 4. Send your answers to janmichaelyap@gmail.com.
- **You should receive a confirmation e-mail from me stating receipt of your deliverable within 24 hours upon your submission of the problem set**. If you have not received any, forward your previous submission using the same subject header once more.
- If you have any questions regarding an item (EXCEPT the answer and solution) in the problem set, do not hesitate to e-mail me to ask them. However, **questions regarding this problem set forwarded/received on or after 12:01am of 17 November 2017 will NOT be entertained**.

Questions

1. Consider the following set of reads from a short super sequence:

   - R1: TGACT
   - R2: GACTG
   - R3: TGGAC
   - R4: ACTGG
   - R5: CTGGA

   Assuming *perfect data*, you are to simulate de Bruijn graph-based de novo assembly given the reads above. Provide answers to the following instructions/questions as the steps in the simulation.

   (a) Generate the k-mers for each read where $k = 3$.

   (b) Draw the resulting de Bruijn graph based on the k-mers generated on the previous step.

   (c) What are the assembly/ies that can be derived from the generated de Bruijn graph?

2. Suppose that a gene set consisting of 100 genes were found to be significantly differentially expressed in an experiment. GO term enrichment analysis will be performed as qualitative validation as to the importance of the gene set, with the gene set to be compared with a database containing 5,000 genes. Below is a table showing the molecular function GO terms that were used to annotate each gene in the gene set, together with the number of genes with the annotation both in the differentially expressed gene set and the reference database:

Molecular Function GO Term	Sample Frequency	Background Frequency
catalytic activity	12	600
kinase inhibitor activity	30	46
chemokine receptor antagonist activity	24	1,188
oxidoreductase activity	11	1,217
identical protein binding	33	512
hijacked molecular function	1	613
transporter activity	10	489
GTP binding	2	2
MAP kinase kinase activity	78	3,844
protein tag	20	1,034

Assuming χ^2 distribution with 1 degree of freedom as the reference distribution for evaluating significance of gene set enrichment and using $\alpha = 0.05$ as threshold:

(a) Which GO terms were significantly enriched in the gene set? Give each significantly enriched terms their respective fold change.

(b) Which GO terms were significantly upregulated? Downregulated?