CS 175

Problem Set 3

Due: 05 November 2018, 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] < *Last Name*, *First Name* > − Problem Set 3.pdf. For example: [CS 175] De la Cruz, Juan - Problem Set 3.pdf. Include the answer sheet in the .zip file detailed in the 3rd bullet point.
- 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.
- This problem set will require you to submit a .zip file containing several files (including the answer sheet) as detailed below. The name of the .zip file to be submitted should follow the following format: [CS 175] < *Last Name*, *First Name* > − Problem Set 3.zip. For example: [CS 175] De la Cruz, Juan - Problem Set 3.zip
- Submission of the problem set answers should be done via e-mail. Attach the .zip file, and write as the subject header of the e-mail: [CS 175] < *Last Name*, *First Name* > − Problem Set 3. For example: [CS 175] De la Cruz, Juan - Problem Set 3. Send your answers to jcyap@dcs.upd.edu.ph.
- **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 02 November 2018 will NOT be entertained**.

Questions

For this problem set, you will perform an assembly of an artificially generated set of reads from an *E. coli* using SPAdes and perform assembly quality analysis on the result using QUAST. Because of the sheer bulk of the reads, use only 1 thread and a maximum of 4GB of memory to perform the assembly using SPAdes. Be warned that the assembly will take at least half a day to finish, so do this activity as early as possible.

1. How many reads are there in the dataset?
2. What is the length of the assembly?
3. How many contigs did the assembly produce?
4. How many contigs are *below the threshold* QUAST considers to be relevant contigs?
5. What is the least amount of contigs needed to cover half of the length of the assembly?

Include in the .zip file the *scaffolds.fasta* and the *spades.log* files generated by SPAdes and the *report.pdf* generated by QUAST.