FALL 2017 BCB5200 Final Exam (Dec 7, 2017)

Name: _____
Total final points = 100 pts

Part II. Python Programming Take Home Exam (30point) (Due: Dec 11, 2017 4:00pm)

Contaminant oligonucleotide sequences such as primers and adapters can occur in both ends of next-generation sequencing (NGS) reads. These adapter sequences have to be removed as they can hinder correct mapping of the reads and influence SNP calling and other downstream analyses.

Many software tools exist for adapters removal (https://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools#Trimming_and_adapters_removal). Now, you will make your own adapter removal software using Python (and/or BioPython). I strongly recommend you to use the BioPyhon as the previous homework.

Software arguments:

1. Get FASTQ file as input of sequence (first argument)
2. Get FASTQ file output name (second argument)
3. Get adapter sequence to check and remove if the read starts with the adapter (third argument)
4. Get adapter sequence to check and remove if the read ends with the adapter (fourth argument)
5. Get the minimum length to filter out if the length of trimmed reads is smaller than the minimum length (fifth argument)

Software workflow:

1. Input FASTQ format file and output FASTQ format file.
2. Get two adapters separately to compare start and end of reads.
3. Get the minimum read length after removal adapters (Be careful. Change the string argument to integer).
4. If the length of reads < minimum length, then "continue" (filter out the read).
5. If the reads have both adapters:
    a. Remove the adapters from the read
    b. If the adapter removal read >= minimum length, then save it.
    c. Else, then filter out the read.
6. If the reads have start adapter:
    a. Remove the adapter from the read
    b. If the adapter removal read >= minimum length, then save it.
    c. Else, then filter out the read.
7. If the reads have end adapter:
    a. Remove the adapters from the read

b. If the adapter removal read >= minimum length, then save it.
c. Else, then filter out the read.

Requirements using Test Dataset:

1. Download compressed FASTQ file and decompress it.
   ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR020/SRR020192/SRR020192.fastq.gz
2. Adapter for start of read
   `GATGACGGTGT`
3. Adapter for end of read
   `ATAACGCCCAT`
4. Minimum length should be 80
5. Run the program as below:
   ```
   $ python adapter_removal.py SRR020192.fastq SRR020192_filtered.fastq
   GATGACGGTGT ATAACGCCCAT 80
   ```
6. Check the number of reads of new filtered file
   ```
   Saved 40678 reads
   ```
7. Verify the case of both adapters exist
   Read ID: SRR020192.17
   ```
   * Original
   @SRR020192.17 E0LM4JH01BLTGJ/2
   GATGACGGTGTTTACATTGTTCCACCACTCATCTCCTCTGTCATGCCCAAAGTCTTCTCAAACCTTCGACAGTTTGG
   CCTGAAACCGACCCGGACTGACAAAACGGACGCTGAGATAACGCCCAT
   +DDDDDDDDDDDDDDDDDDIIIIIIIIIIIIIIIIIIIIIIIIIIIDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
   DDDDDDDDDDDDDDDDDBB>>550000555.:AAAAAA:4444<111<@
   * Filtered
   @SRR020192.17 E0LM4JH01BLTGJ/2
   TTACATTGTTCCACCACTCATCTCCTCTGTCATGCCCAAAGTCTTCTCAAACCTTCGACAGTTTGGCCTGAAACCGA
   CCCGGACTGACAAAACGGACGCTGAG
   +DDDDDDDIIIIIIIIIIIIIIIIIIIIIIIIIDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
   DDDDDDBB>>550000555.:AAAAAA
   ```
8. Verify the case of only start adapter exist
   Read ID: SRR020192.8
9. Verify the case of only end adapter exist
   Read ID: SRR020192.31
10. Compress your program and output file and return with yourname_final_part2.tar.gz