BioInformatics & Data Handling

TASK 1:

Write a program for each of the following:

- Recursively find all FASTQ files in a directory and report each file name and the percent of sequences in that file that are greater than 30 nucleotides long.
- Given a FASTA file with DNA sequences, find 10 most frequent sequences and return the sequence and their counts in the file.
- Given a chromosome and coordinates, write a program for looking up its annotation. Keep in mind you'll be doing this annotation millions of times. Output Annotated file of gene name that input position overlaps.

Input:

- fastq directory
- fasta directory(sample.fasta)
- coordinates_to_annotate.txt and hg19_annotations.gtf.

NOTE:

1. Keep in mind; we will use the results of these tasks to assess your ability. This is a chance for you to show off your programming skills and style.
2. A Python solution is ideal.
3. Sample input files have been provided for each task.
4. Make sure you understand the file formats (FASTQ, FASTA, GTF) to perform these tasks correctly.
5. Please make sure each task can run on the command line.
6. In the spirit of assessing your programming abilities, please avoid using 3rd-party tools to solve these problems (parsers and formatters excluded).

TASK 2 :

1. Parse the given Example.hs_intervals.txt file. The file contains information on covereage on exon level in a hybrid capture panel. The file is a tab-delimited text file. Report the mean target coverage for the intervals grouped by GC% bins. Bin in 10%GC intervals (e.g. >= 0 to < 10; >= 10 to < 20; etc). Note that in the file, GC values range from 0 to 1 rather than percentage.

   Notes: Relevant Columns in the file: %gc and mean_target_coverage.

TASK 3: (Optional)

1. Given a list of variant IDs, using Ensembl API retrieve information about alleles, locations, effects of variants in transcripts, and genes containing the transcripts.
2. Create a repository on GitHub and upload your code there. Make some minor changes to your code locally, and use a local Git installation to commit the changes to your GitHub repository.

   Notes:

   - (Link for information about API is here https://useast.ensembl.org/info/docs/tools/index.html).
   - You can use either PERL API or REST API.
   - Example variant id-: *rs56116432*

Cloud Computing:

1. How would you architect a framework for sharing large files (10Gb-25Gb) on the cloud with access controls at the file level? We want to share the same file with multiple users without making a copy. The users should be able to have access to the data on any cloud platform to run bioinformatics analysis pipelines. The users can run any cloud service, there is no restriction. The framework's responsibility is only to make data accessible with access controls.

2. Evaluate the benefits and limitations of using containerization and container orchestration technologies, such as Docker and Kubernetes, for deploying and managing bioinformatics HPC workloads in the cloud.

SQL:

1. For the following SQL statement, what is wrong with it and how would you fix it:

```
-- Question:
SELECT UserId, AVG(Total) AS AvgOrderTotal
FROM Invoices
HAVING COUNT(OrderId) >= 1
```