Assignment-5&6: Clustering Techniques and RNA Seq

IIIT-DELHI

Due Date: 10th November 12.00 pm

Total (50 points)

Instructor: Debarka Sengupta

Plagiarism: All submitted codes are expected to be the result of your individual effort. You should never misrepresent someone else's work as your own. In case any plagiarism case is detected you will get one grade reduction in final examination. Cite the resource wherever using other's code.

Instructions:

1) Allowed programing language are python and C .Allowed to use any external python libraries. If you see an ambiguity or inconsistency in a question, please seek a clarification from the teaching staff.

2) You must submit your working solution on backpack on the deadlines page from where you have downloaded this assignment instructions sheet. **No extensions on deadline. If you fail to submit within the time limit then your solution will not be evaluated.**

3) **Mention your enrollment no and name at starting of the each file . Write a clean code with proper comments at appropriate places as it will be checked.Your code should not give error and should not break .**

4) Generate good labeled diagrams and name then according to the problem and upload a zip folder with **rollno_name_assignmentX.zip** which should have your codes.figures and latex report.**Codes won't be checked if they don't follow the guidelines. Also provide the dataset.If the dataset is huge mention google drive link to download the data from.**

Hardware and Software Resources:

It is mandatory that you should do version controlling of all your homeworks. We will be using https://bitbucket.org/ for this purpose. Get a login id for free on bitbucket and create the homework repository. All homeworks should be saved inside the repository named hw_1 with your roll number as suffix. If your roll number is 1234 then the repository name will be hw1_1234. Every homework repository should have access level set as "private". You have to use the "share" option for your homework repository to share it with both of your TAs.

Problem 1: RNA Sequencing [25 points]

Task List

1. Select any **RNA-sequencing based study** and download **FASTQ** files and explain the rationale behind selecting it.
2. Perform **QC and generate bam files**. [External Tools Allowed]
3. Generate **Raw count data** using the above generated **bam files**.
4. Perform **normalisation and clustering techniques** on the data generated in step 3. Compare the clustering results with results generated using **PCA.**
5. List out **differentially expressed genes with their p-values and log fold change**. Mention the threshold values of **p-value and log fold change and their biological significance**. **Plot these differentially expressed genes corresponding to the cluster.**

Write a **LATEX Report** mentioning details about the protocol followed to generate the result for each task ,interpresentation of your result and mention in brief about each tool, file formats used.Submit all your script,plots and report in pdf file format.

Problem 2: Clustering on Single Cell Data [25 points]

Task List

1. Download any **single cell study** from **Recount database** and extract read count to perform **log transform and normalisation** of the data.
2. Perform at least **3 clustering technique** and compare the results with a set of parameters.Mention the rationale of choosing those **3 clustering techniques.**
3. Validate the results of the above techniques by applying **PCA or tSNE on the data and share your views on accuracy and ease of use of the 3 clustering techniques used**.
4. Discuss **biological significance of the data** used and the **generated results**.

Write a **LATEX Report** mentioning details about the protocol followed to generate the result for each task ,interpresentation of your result and mention in brief about each tool, clustering algorithm and file formats used.Submit all your script,plots and report in pdf file format.