SBE459 – Bioinformatics.
Final Project.
Deadline: Sunday – July 11th, 2021.

- You can submit before the deadline. Please arrange with me for that.
- No late submissions.

Group members: Two or three students per group.

Here are two project ideas. Please pick up only *ONE* of them.

Project Idea #1:

You are required to conduct a comparative study related to SARS-Cov-2 (COVID-19).

- Comparison between different COVID-19 sequences.
  - Use the NCBI database to download the sequences below.
  - Compare the DNA sequences of SARS-Cov-2 between:
    - Five sequences from Egypt.
    - Other sequences from USA, Europe, China, and the gulf area.
      - At least two sequences from each geographic region.

- The comparison should be including, but not limited to, the following:
  - Sequence alignments,
  - Phylogenetic trees,
  - The percentage of the chemical constituents (C, G, T, and A) and the CG content,
  - Find the functional products/interpretations of the dissimilar regions among the aligned sequences with reference to the Wuhan city reference sequence.

Project Idea #2:

You are required to conduct a study to analyze gene expression (GE) data for two types of cancer. These types are:

1. Lung Squamous Cell Carcinoma (LUSC),
2. Kidney Renal Clear Cell Carcinoma (KIRC).

The GE data (sent with this statement) for each cancer type:

- Two GE files for each cancer type:
  1. "type-rsem-fpkm-tcga-t_paired.txt", where type $\in$ {lusc, kirc}: GE data for tissues with cancer,
  2. "type-rsem-fpkm-tcga_paired.txt": GE data for tissues in a healthy case.
- Data are paired: each GE file will have the same number of cases (patients) and in the same order.
- Files are tab-separated.

Requirements:

- For each cancer type, infer the differentially expressed genes (DEGs):
  - Use the following methods to identify DEGs:
    1. Hypothesis testing,
    2. Fold change,
    3. Both of them.
  - Report the set of DEGs from each of the above methods, and report how similar/different the DEGs from the first two methods are.
  - Use the set of DEGs from the third method to perform Gene Set Enrichment Analysis (GSEA) on this set of genes.
    - Suggestion: you can use this GSEA Software.

Submission for both ideas:

- Support your findings/results/conclusions with figures.

- You have to deliver the following:
  - All the code scripts you used for your analysis,
    - Comments are a must.
  - Project report:
    - It should look like a research paper. It should have the following sections:
      - Introduction,
      - Methods: describe all the steps carefully and include all the used software packages,
      - Results and Discussion: report your results in details and discuss them,
        - You can augment your results with textual files or spreadsheets.
      - Conclusion: list the overall findings of your analysis.
      - **Members Contribution: list in details what each member in your group did in this project. Each member in the group may receive a different grade based on the contribution weight.**

  - Presentation:
    - You will be given a few minutes to represent your work online.
    - Prepare yourself for discussing your analysis and findings.

Notes:

- The report and the code scripts have to be delivered at least one day before the presentation day.

Good luck!