SBES375 - Bionformatics.
Spring 2024.
Final Project.
Deadline: Sunday – May 19th, 2024, at 2:00pm.
Presentation: Sunday – May 19th, 2024. (Presentation schedule will be sent later).

- You can submit before the deadline. Please arrange with me for that.
- No late submissions.
- Only one member from each team should submit the project files to Blackboard.

Teams: 3 or 4 students per group.

Here are three 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) and its variant Omicron; or between the variant Delta and the variant Omicron.

- GISAID (Global Initiative on Sharing All Influenza Data) is a public database for SARS-Cov-2 sequences and its variants.
  - It is a collaborative work between the German government and a non-profit organization (Friends of GISAID).
  - To access the GISAID data you need to register with an educational institution email address (e.g. Cairo University). Then, wait sometime until your registration gets approved.

- Pick up only one country from any continent. This country must have sequences for both SARS-Cov-2 and its Omicron variant; or for both variants Delta and Omicron.

- Download 10 SARS-Cov-2 (or Delta variant) sequences from the selected country.
  - We will call them the reference sequences.
  - Construct a consensus sequence from these sequences. At each sequence location, the nucleotide/amino acid of the consensus sequence will be the most dominant one across all the sequences at that location.
  - We will use this consensus sequence as a single representation for the reference sequences.

- Download 10 sequences for the SARS-Cov-2 Omicron variant.
  - We will call them the case sequences.
  - Apply any multiple sequence alignment technique on these sequences.

- Ways of comparison:
  - Construct a phylogenetic tree between all the above 20 sequencs.
  - The average percentage of the chemical constituents (C, G, T, and A) and the CG content, if applicable, between the reference sequences and the case sequences.
  - Extract the dissimilar regions/columns between the alignment of the case sequences and the consensus sequence (the representative reference).

Project Idea #2:

You are required to conduct a study to analyze gene expression (GE) data for the Lung Squamous Cell Carcinoma (LUSC). The GE data (sent with this statement) are described as:

- Two GE files:
  1. "lusc-rsem-fpkm-tcga-t_paired.txt", GE data for tissues with cancer,
  2. "lusc-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:

- Hypothesis Testing. 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 (volcano plot).
  - For the hypothesis testing method, apply the appropriate test statistic for the following two cases:
    1. Samples are paired,
    2. Samples are independent.
       - Report the set of DEGs in the above two pairing cases, and report how different these two sets of genes.
  - For the volcano plot method, use the set of DEGs obtained by the hypothesis that data are paired and perform Gene Set Enrichment Analysis (GSEA) on this set of genes.
    - Suggestion: you can use this GSEA Software.

Project Idea #3:

You are required to analyze a biological network of protein-protein interactions (PPIs) using the NetworkX python package. The package website contains helpful information about networks (graphs) and how to use the package.

- You can get your own PPIs network or you can use the interactome titled "PathLinker_2018_human-ppi-weighted-cap0_75.txt". You can download the interactome file from here.
  - This file represents a directed interactome: each interaction starts from the tail node to the head node.
  - This file represents each PPI in one line of four pieces of information (columns).
    - First: the tail/source/start protein node.
    - Second: the head/destination/end protein node.
    - Third: the interaction confidence. It has a range from 0 to 1. The higher this value, the more probable this interaction to happen.
    - Fourth: the method used to identify this interaction.
  - Each protein is represented by its UniProt ID.
- Your analysis can include, but not limited to, the following:
  - Construct a graph (biological network) from the above interactome.

- Given two proteins, list the acyclic shortest path(s) between these two nodes in a text file.
  - Provide the total path score.
  - Provide the weight of each interaction in the path(s).
  - If more than one path, report all the paths.
  - Use NetworkX and matplotlib to draw the sub-network formed by these shortest paths.
- Given one protein, list all the directly connected proteins to it in a text file.
  - Report the degree (number of connections) of this protein in a separate line.
  - Provide each connected protein in a line with its corresponding interaction weight.
- Given a set of proteins:
  - Draw a histogram for the proteins degree.
  - Rank these proteins from the highly connected to the least in a text file, where each line is a protein and its corresponding degree.
- Provide a conversion map between the protein UniProt ID and its gene name.
  - You can be provided by one protein ID or a set of protein IDs, and then you need to get their corresponding gene names.
- Convert the above graph as an unweighted graph and save it using the adjacency matrix method. You need to search for it.

Submission for all 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.

Good luck!