Andrew Smith
Final Project Proposal

Cancer Proteome Exploration

The National Cancer Institute Genomic Data Commons stores a variety of data from cancer research programs, including a large collection of cancer proteomes from the MD Anderson Cancer Center and The Cancer Proteome Atlas. Since we know that protein dysregulation can be both cause and effect in cancer cases, identifying anomalous trends in cancer proteomes can be helpful in diagnostics, understanding the mechanisms of cancer, and identifying potential drug targets.

For this project, I propose to make a data display tool to showcase this proteome data. I would retrieve cancer proteomes from NCI GDC for a particular cancer, for example esophageal, and create a SQL database that stores both this protein expression data and related annotations about the proteins from Entrez in a relational fashion. This involves identifying and downloading relevant data, designing a database schema, and converting/inputting the data according to that schema.

Then, I will create a website using CGI and Python so that users can search for specific proteins and retrieve information about that protein in that type of cancer. I would use a Python script to make a database connection with mysql.connector and retrieve and format the relevant data. At the top of the webpage will be a summary of the protein and key information about it. Below will be visualizations representing that protein's expression. This could include a heatmap showing the expression level for that protein in each of the samples, as well as a chart showing the proteins whose expression levels are most correlated with the target protein. Because these correlations might be expensive to repeatedly compute, I could precompute them and store the results in an ExpressionCorrelations table in my database. The visualizations can be done in Matplotlib and served with CGI and Jinja templating.

I will use CSS to style the page, and AJAX in JavaScript to retrieve results when a search is made, as well as jQuery to add an autocomplete feature to the search and validate the search form. It would be possible to just send the data in JSON form from the CGI script and build the visualizations client-side using a JS library like canvasJS or Chart.js, rather than doing it with Python server-side. This might stretch further from what we covered in this class but would allow the visualizations to be more interactive.

The goal of this site is to display data about cancer proteomes in a useful way to help researchers more quickly and easily identify important protein markers and correlations.