Zeinab Navidi - Project Proposal - CSC2431 - 2022 Sep 28

**Title: Artificial Intelligence for Predicting Drug Effect on Genetics**

**Background:** Single-cell sequencing technology provides information on individual cells, which can be used to learn the pattern of cells' genetics, and obtain a better understanding of a sample containing heterogeneous cell types. This information cannot be obtained using bulk sequencing technology, as it reports average information across many cells. Single-cell gene expression is one type of information that can be extracted from cells that measure thousands of genes' activity (or expression) at the time of sequencing, which helps better characterization of cellular functions [resource1]. Applying drugs to diseased cells can change cell states (or in other words their gene expression profile), and their functionality. Understanding how drugs modify gene expression is essential for designing effective treatments to achieve healthy cell states. For instance, if perturbing the expression of a gene is found to affect cell viability, then a drug targeting that gene would have a significantly higher likelihood of success than one without such target. Finding the most effective treatment for each patient considering their genetic information requires observing a sample's behavior after applying different combination of drugs. The limitation is that in practice, possible combinations of many different drugs are too numerous to measure in laboratories and thus impractical. However, computational methods do not have this limitation and can effectively predict the effect of different treatments on one sequenced sample by deploying AI tools developed for predicting post-treatment gene expression for different drugs.

**Objective:** There remained challenges in developing effective and robust AI tools for the described application. Towards these goals, the objective of this project will be divided up into two subtasks to be done based on the interests and expertise of the students as course timing allows:

- Optimal dimensionality reduction of high dimensional pre-treatment and post-treatment gene expression profiles across different cell states (pre- vs. post-treatment) and cell types.
- Developing a robust Machine Learning model for predicting post-treatment cell state (gene expression profile) from pre-treatment cell state.

Students can implement different methods such as PCA, or neural networks like Variational Autoencoder and compare their performance. Prior experience in Python programming (specifically PyTorch) and understanding of different neural networks models would be helpful.

**Data:** The proposed data for this project includes pre-processed gene expression profile of 14811 cells treated with four drugs. Each cell contains the expression measurement of 58347 genes, and the data is provided in the form of cell by gene matrix, which each element in the matrix indicating the expression of a gene in a cell. This data can be downloaded from this page with 173.7M size. There are other pre-processed datasets available on this GitHub page, including the same format and information with different cell types and drugs, which can be used depending on students' interests.

**Other Resources:** Two review papers related to the drug effect prediction topic: [resource2, resource3]. Here is a link to a related method: [resource4].

**Recommended Computing Resources:** Based on my experience in training single-cell data on deep models, I estimate at 8 CPU cores, 1 GPU, and at least 32G memory will be required.