Gene Expression Network and Prediction for Breast Cancer

Initial Proposal

Group Member: Yidan Ma, Jinglun Li, Zhengkuan Tang, Ziming Wei

Data

The dataset we use is *Breast cancer gene expression - CuMiDa*, which is publicly available on Kaggle (https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida). Which contains 151 samples and 54676 genes.

Aim

The aim of our project is to use gene networks and machine learning algorithms to verify the biological significance of genes and find out other possible genes that are less studied or could be targeted by drugs in order to treat Breast Cancer.

Novelty

Some of the researchers have published their prediction models on Kaggle, but their aim is to predict the result more than to find out the gene importance and roles in the gene network. In our final project, we would focus more on the biological gene network and Gene ontology part and combine the data from other Gene libraries like NCBI. We attest that none of us have used the same dataset or applied the same pipeline in other projects.

Part1.Exploratory Data Analysis (Validation) (Yidan)

First, we would like to use Python to preprocess the data we downloaded from Kaggle. We will perform integrity checks using Pandas, coding features manually and standardize the data using packages from scikit-learn. Then, Quality control will be performed on the raw dataset. In addition, the expression of genes will be scaled to prevent highly expressed genes from dominating results. Finally, exploratory data analysis will be performed using visualization tools like Weka and t-SNE to get the distribution and the characteristics of the dataset.

As the genes in the raw dataset are in Affymetrix format, we will convert the Affymetrix format to a commonly used gene symbol.

Part2. Biological Background Research (Jinglun)

We will identify validated breast cancer related genes in the Open Targets Platform (https://platform.opentargets.org/), and select genes with a high association score with breast cancer using a proper threshold and generate our target genes list. Similarly, genes associated with each subtype can be extracted from the website.

Then we would like to find out whether these biologically significant genes' expression levels can be differentiated under normal conditions and abnormal conditions (Patients with breast cancer or patients with each subtype of cancer) using Cuffdiff based on the dataset. After that, we can identify the differentially expressed genes that are specific to each type or are common

to several types. This can be regarded as an initial result to test whether the information we get from the dataset matches what has already been published.

Part3. Machine Learning Based Prediction Model (Zhengkuan)
Feature selection methods like Backward Feature Elimination or Random Forest would be used to identify important genes for the outcome. Then clustering methods such as Hierarchical Agglomerative Clustering would be applied to determine if the samples from the dataset are clearly clustered into several categories. We would check the consistency between the clustering result and the “Type” column which stands for type of breast cancer. Based on the result from the last two steps, the ML-based prediction model would be built on one or more clusters to see the relationship between the gene and (different type of) breast cancer.

Part4. Gene Network and Pathway analysis (Ziming, Jinglun)
After getting the result from the differential expression and clusters using machine learning algorithms, a gene network showing the up-regulation or down-regulation of genes in the network would be constructed. Hubs and differentially expressed gene locations would be recorded for further analysis. Biological significant nodes would be highlighted to see if they are differently expressed.

The genes included in the constructed network are mainly from two attributes. The first way is explained in Part2. The second way to filter out candidate genes is based on the differential expression gene analysis, where we find out the genes that are highly differentiated in gene expression between the control group and breast cancer group using Cuffdiff. After getting the candidate genes, we use the BioGRID to find out the correlation between the candidate gene and download the gene network data. We decided to use NetworkX as the package to build our network. The correlation and differentially expressed genes could be easily discovered from the gene network we build. The biologically significant genes identified in part two will be emphasized in networks.

We will also implement pathway analysis in published breast cancer related genes, core genes found in the network, and genes with significantly different expression levels through GO (using GO-Elite and ToppGene) and KEGG. The enriched pathways that differ between published genes and core nodes genes will let us better understand the mechanisms.

We expect that after this part of the analysis, we will be able to gain a deeper understanding of some of the critical genes, based on their differential expression levels, position in the network, biological significance, and involved pathways. Some of them could be explored as a potential drug target in future studies.