Classification and Interactive Visualization of tumor types in Neurofibromatosis using RNA-seq and Drug Screening data

Pranay Methuku
Georgia Tech

Laura Mora
Georgia Tech

Reagan Kan
Georgia Tech

Swetha Singu
Georgia Tech

Introduction

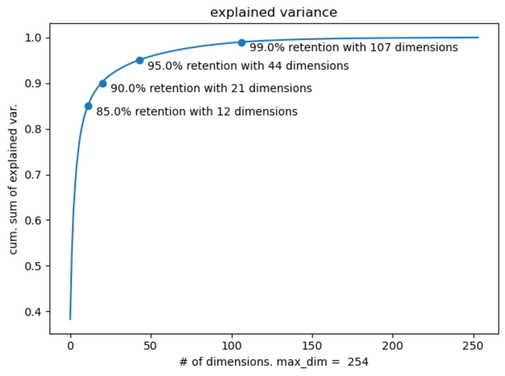
Neurofibromatosis (NF) is a rare genetic disorder of the nervous system affecting the development of nerve cell tissues that usually causes benign tumors with a nearly 10% chance of malignancy. It does not currently have a cure; thus, the most important elements of management are **early diagnosis** and treatment of the effects of the disease. The objective of this project is to identify molecules associated with different tumor types of NF1 and NF2 and possible therapeutic targets using **machine learning algorithms** to classify the RNA sequence data and identify targets using drug screening data.

Data

The data was obtained from **Synapse** as part of the **NFHackathon**. The combined files are about **3GB in size** and include about 65,000 genes involving 6,130,214 records, which were later transformed to a structured dataset with **255** tumor samples and **15k** gene symbols.

Approaches

After an initial structural transformation, the data will go through preprocessing. Three methods will be tested: **variance stabilizing** transformation of Deseq2, **minmax scaler**, and **log transformation normalization**. Next, the data dimensionality is reduced using PCA. We choose the number of PCA components that preserves 99% of the data variance. Finally, **Random Forests** are trained on the preprocessed data for subclassification of 8 types of classes which include normal samples and tumor samples.

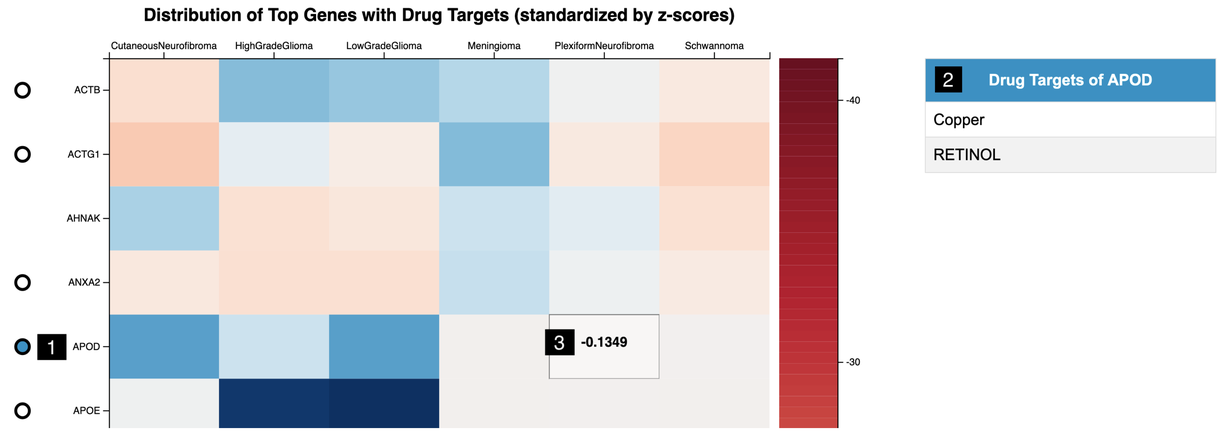


The above graph shows the explained variance as the number of PCA components increase. We can observe that 107 dimensions preserves 99% variance.

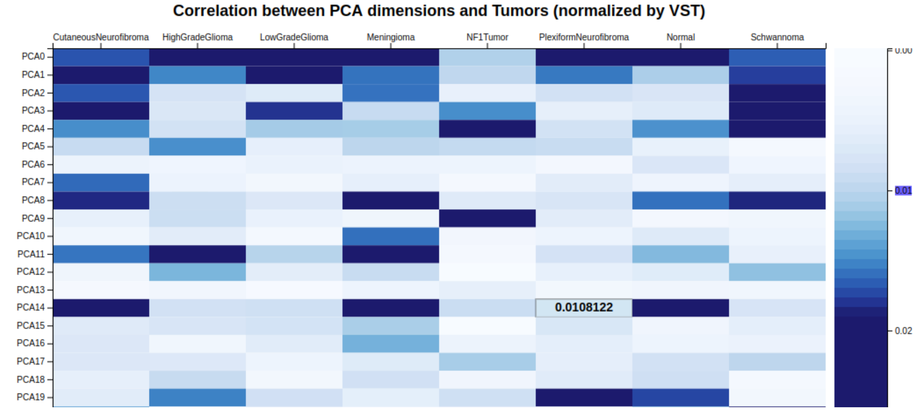
Therapeutic targets are identified using **differential gene analysis**. The z-scores relative to the normal tumor type are analyzed and the smallest and largest values are used to pick the top 10 differentially expressed genes. These genes are then used to get the common name of their known associations.

Experiments and Results

Of the three preprocessing methods, variance stabilizing transformation of Deseq2 produced random forests with the best test metrics. We also experimented with different random forest hyper parameters like **max depth**, **n estimators**, **random state** using **GridSearch** with a **5-fold** cross-validation scheme. The best classifier had **16 estimators** and a **max depth of 100**. Different evaluation metrics were generated to test the performance of the classifier and heat maps are generated for PCA component/tumor-type correlation **(Heatmap 2)** and the top genes and drug targets for drug screening data **(Heatmap 1)**.



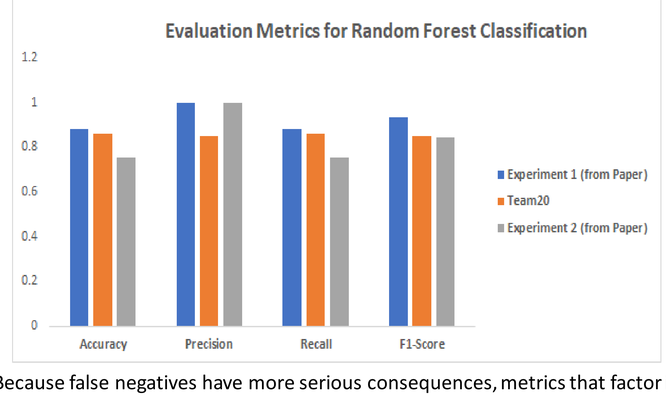
Heatmap 1: Distribution of Top Genes with Drug Targets. The rows represent specific human gene symbols, and the columns correspond to a tumor type. Thus, each grid in the heatmap is a gene/tumor pair. There are two interactive components to the heatmap. Selecting the radio buttons (1) on the left column will display the drug targets associated with the gene in a menu on the right-hand side (2). Hovering over the grids (3) will display the z-score of that gene/tumor combination relative to a normal tumor type on the linear color scale.



Heatmap 2: Correlation between PCA dimensions and Tumors. The rows represent PCA dimensions and the columns correspond to tumor types. Thus, each grid in the heatmap is PCA dimension/tumor pair. Hovering over the grids will display the normalized VST score of that gene/tumor combination relative to a normal tumor type. The color scale is non-linear in this case; it is exponential. We can observe that as the number of dimensions increase, the correlation values decrease, signifying less important dimensions.

Evaluation

The metrics like **Accuracy, Precision, Recall, F1-score** are compared for the classifier tested with the evaluation metrics of another study. In that paper, Experiment 1 includes all Latent variables and Experiment 2 includes top 40 latent Variables. Although our dataset has twice as many tumor types, thus more complex, our classifier can still achieve comparable performance.



Because false negatives have more serious consequences, metrics that factor in false negatives, like recall and accuracy, are more insightful.