Predicting Complete Remission in Leukemia Patients using Gene Expression

Ophir Gal, Noam Auslander, Daoud Meerzaman
Center for Biomedical Informatics & Information Technology (CBIIT), National Cancer Institute, National Institute of Health, Rockville, MD, 20850, USA

NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

Abstract

We wanted to see if we can use machine learning algorithms to predict complete remission in AML patients who have received induction therapy, based on their pre-treatment gene expression. We did a Principal Component Analysis to get a sense of the data, its outliers, and what methods may work well with the data. We used tests to get an estimate for the optimal hyperparameters to be used with the different classifiers. We performed statistical significance tests to determine genes which are differentially expressed between the samples of patients who achieved complete remission after 2 courses of treatment and those who did not. Embedded cross validations were conducted to both select our features (genes) and to get an AUC result using the training set for prediction. Lastly, using those genes a final ROC curve was plotted to determine the best performing model on the unseen test data set.

Background

Acute Myeloid Leukemia (AML) is a cancer of the blood and bone marrow. Childhood AML is a type of cancer in which the bone marrow makes a large number of abnormal blood cells. Cancers that are acute usually get worse quickly if they are not treated. Cancers that are chronic usually get worse slowly.

After receiving the RNA-seq (gene expression) data from the research group, we looked at the clinical data and searched for categories that would be interesting and perhaps useful to analyze or try to predict.

We decided to use the “Complete Remission” category in order to predict pre-treatment whether patients would go into complete remission. This could help decide on a more suitable course of treatment, be it more aggressive chemotherapy or a different therapy altogether. The bone marrow samples were taken at different years from patients ranging from infants to young adults, which suggested the data might vary by age or year of diagnosis.

Methods

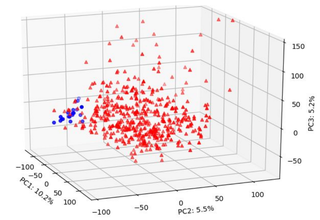
- We used RNA-seq data of m samples by n genes:
  Input: $X \in R^{m \times n}$, $y \in \{0,1\}^{m \times 1}$
- We tried to learn the mapping : $F : x \rightarrow \{CR, Not\ in\ CR\}$
- Several PCA plots were taken to examine the data, remove its outliers and get a sense of algorithms that could be used
- Different feature selection methods were then performed, crossed with different machine learning algorithms
- Cross validation was employed both in the feature selection process and when assessing algorithm performance
- Hyperparameters of classifiers were tuned to optimize performance
- Lastly, the final ROC results were produced using the test set which was set aside at the beginning

Results

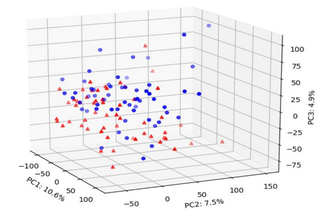
PCA plots produced to check for outliers and get a sense for the data ; Plot of AUCs for various K-values ; ROC curves with AUC results taken using different feature selection methods

a) PCA taken with all data (473 AML vs. 20 healthy bone marrow samples clustered in blue)
b) PCA taken with balanced subset (114 AML samples: Complete Remission vs. Not in Complete Remission in blue & red respectively)
c) PCA taken with all training data (colors vary by year of diagnosis)
d) PCA taken with all training data (colors vary by age at diagnosis)
e) AUCs from different Ks to estimate optimal K for K-NN classifier
f) ROC curves of K-NN and SVM algorithms using 3 feature selection methods: Hill Climbing, Recursive Feature Elimination, and Randomized Lasso. The curves were produced using cross validations done on the balanced training set (91 AML samples)

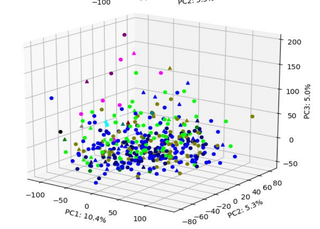
a)



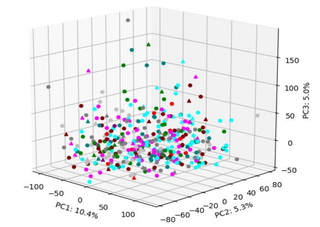
b)



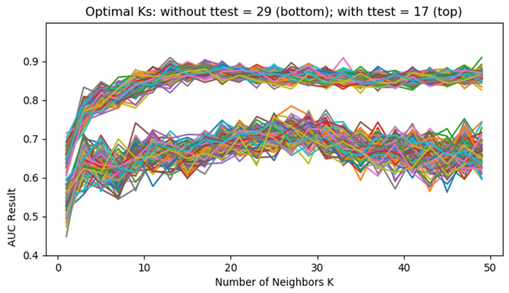
c)



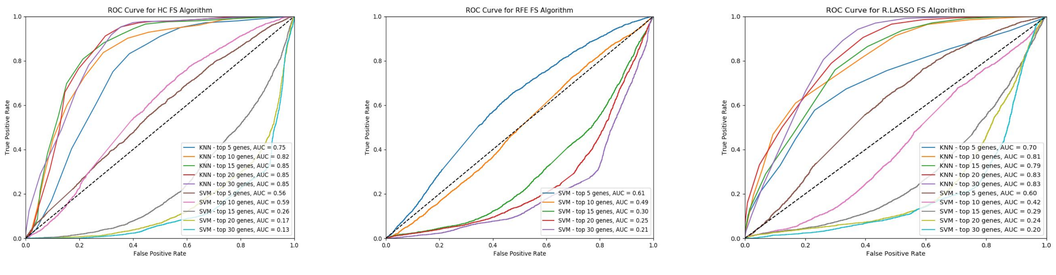
d)



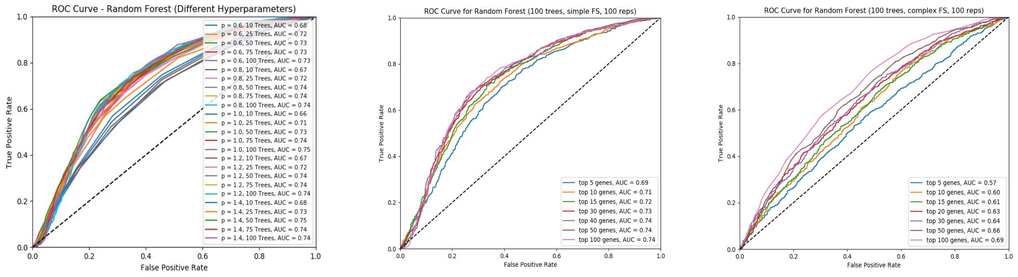
e)



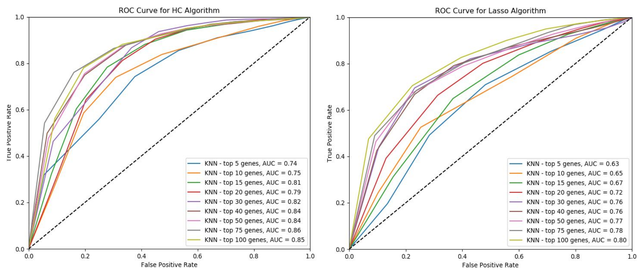
f)



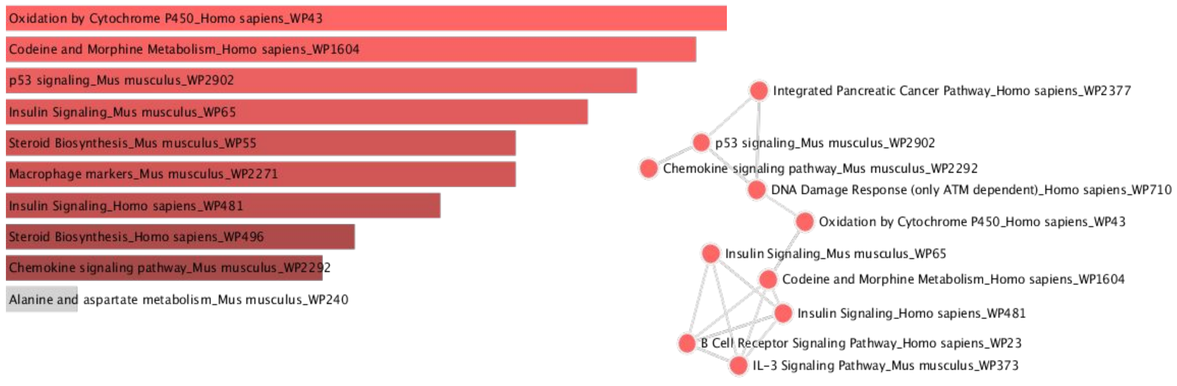
Tuning Random Forest Classifier (done on training data set)



K-NN Model Performance on all Training Data (331 samples)



Pathway Enrichment Analysis



Final K-NN Model Performance (83 samples)