Cancer Type Classification and Mortality Prediction

Jing Lin, MEng[1]; Lawrence Wong, SB[1]; Maggie Wu, MEng[1], Yun Boyer, MEng[1]

[1]Department of EECS, Massachusetts Institute of Technology, Cambridge, MA

Abstract

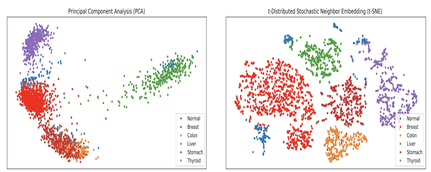
Abnormal gene expression due to mutations in gene regulation mechanisms results in uncontrolled proliferation in cancer cells. Recent advances in machine learning as applied to gene expression profiling data have allowed for greater accuracy in the classification of cancer cell types and have greatly aid in predicting clinical outcomes of cancer patients. However, these prediction methods often deal with gene expression microarray datasets in high-dimensional space. Moreover, the time complexity to generate these datasets scales proportionally to the number of probed genes. We demonstrate with a logistic regression model and a 2D convolutional neural network that the classification accuracy will not suffer when using a lower-dimensional subset of the dataset. This subset of genes is carefully selected and verified through multiple feature selection techniques and extensive literature research on known driver genes. Subsequently, this subset combined with a linear regression model and a dense neural network is utilized to predict mortality of cancer patients. These results combined will help shed light on the mechanisms of gene regulation in cancer cells and set the stage better for informed expression data analysis studies.

Introduction

The classification of different cancer types is one of the most important works in cancer diagnosis and has been applied to multiple common cancers. With recent advances in technology, cancer classification using gene expression level of the entire genome (20,000 - 25,000 genes) has become more feasible and has gained importance in genetic testing procedures. Numerous research studies have shown that cancers and mortality are reflective of changes in the expression levels of cancer-related genes. Given that there are 291 reported cancer-related genes, approximately 1.5% of all the genes in the human genome, the majority of the gene expression data consisting of non-cancerous genes are unnecessary for the prediction of cancer. Also, the time to measure gene expression using microarray technologies increases proportionally to the number of probed genes. Through this study, we would like to know if we can, as accurately as using all genes in the genome, classify cancer types with only a subset of genes. With this subset, we would like to further examine our accuracy in predicting the mortality of cancer patients.

Data

The Cancer Genome Atlas (TCGA) datasets provided by the National Cancer Institute catalogs genetic mutations responsible for cancer. This project focuses on the gene expression RNA-seq dataset generated by the Illumina HiSeq 2500 protocol for five cancer types: *Breast (BRCA), Colon (COAD)*, *Liver (LIHC)*, *Stomach (STAD)*, and *Thyroid (THCA)*. Each dataset set is a matrix with rows as identifier genes, columns as patient sample IDs, and entries as mean-normalized gene expression across all TCGA cohorts. Patient samples consist of both **solid tissue normals** taken from normal tissues near the site of the tumor and **primary tumor** taken directly from the site of the tumor. Dimensionality reduction technique is used to visualize structural patterns and variations within the dataset.



Methods

Gene Selection per Cancer Type

- *Reported Driver Genes.*
- *LASSO Regression*. Pick genes of greatest coefficient.
- *Hypothesis Testing*. Compare the T-statistics of gene expressions between normal and cancer cells. Use Holm & Benjamini for False Discovery Rate < 0.05.
- *Network Centrality Measure*. Create Network structure with nodes as genes and edges as correlation with respect to driver genes (threshold>0.8). Determine additional potential driver genes by choosing genes with high betweenness centrality.

Classification of Gene Expression Patterns

- All Gene Expressions vs. Selected Gene Expressions
- Methods for Cancer Type Classification
  - *Baseline: Multi-class Logistic Regression (LoR)*
  - *Improved: Hybrid Neural Network (hNN/dNN-C)*

Prediction of Mortality for Cancer Patients

- Methods for Mortality Prediction with Selected Genes
  - *Baseline: Linear Regression (LiR)*
  - *Improved: Dense Neural Network (dNN-P)*

Neural Network Architectures

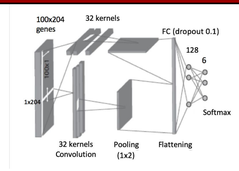


Fig. 1a. hNN

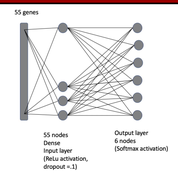


Fig. 1b. dNN-C

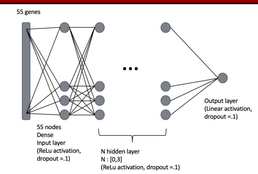


Fig. 1c. dNN-P

Results

Gene Selection Visualization

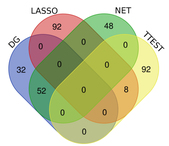


Fig. 2a. BRCA

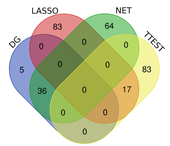


Fig. 2b. COAD

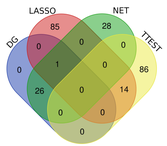


Fig. 2c. LIHC

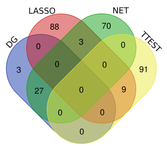


Fig. 2d. STAD

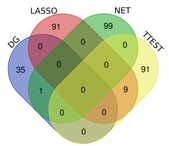


Fig. 2e. THCA

Fig 2. Per Cancer Gene Selection.
The network is built on the correlation of driver genes with other genes. Thus, many driver genes are selected as important genes. However, both T-test and LASSO regression find none of the driver genes as the top 100 significant genes. That is, mutated driver genes can cause other genes to overly express and therefore lead to cancer.

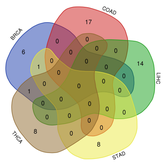
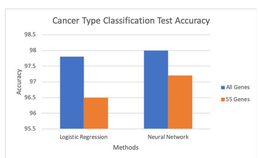


Fig. 3. Combined

Fig 3. Combined Gene Selection.
T-test and LASSO regression are unbiased toward driver genes, providing informative genes that are overexpressed. The union of these genes for all cancers is shown (total=55). There is almost no overlap between any two cancers, implying that genes associated with cancer are very different, which is beneficial for classification.

Cancer Classification & Mortality Prediction



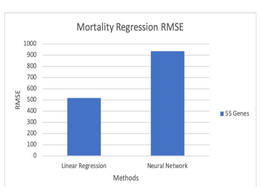


Fig. 4. Method Comparison.

Gene Information Distribution

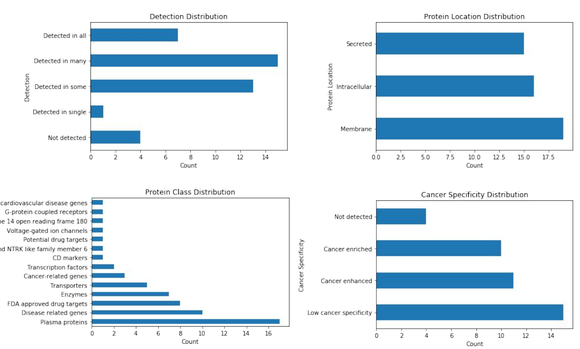


Fig. 5. Gene Detection, Protein Location and Class, and Cancer Specificity

Discussion

Cancer Classification: Based on t-SNE result and feature selection analysis, cancers are distinctly differentiable, thus, LoR can classify well (97.8% with all genes and 96.5% with selected genes). hNN/dNN-C performs slightly better (98.0% with all genes and 97.2% with selected genes). The performance with selected genes only drops around 1% for both models which confirms that these selected genes have good predictive value in determining the cancer type.

Days-to-live Prediction: To avoid small dataset size and knowing each cancer has different significant genes, we predict days-to-live for all cancers combined (total=512 samples). The RMSE is high for both LiR and dNN-P. This is due to a lack of training data and different death rates for different cancers.

Selected-Gene-Information: 1. The majority of selected genes have a differential expression *detectable* among cancers, which confirms that they are indeed important. 2. The *locations* of the protein associated with genes are evenly distributed amongst three categories. 3. The genes are mainly associated with three *classes*: FDA approved drug targets, disease-related genes and plasma protein (helps protect against apoptosis). 4. A majority of the genes' normalized expression value is significantly elevated (>1x in particular tissue), enriched (>4x in particular tissue), or enhanced (>4x in groups of tissues) among cancer cells.

Limitations: We ignore still-alive-patients (~80%) when predicting mortality which assumes all patients with cancer will die within a certain time frame. A better model should consider both mortality and survival rate.

Conclusions

Gene Selection per Cancer Type

- Informative genes, selected from T-test and LASSO regression, vary greatly across cancer types.

Classification of Gene Expression Patterns

- The data is mostly separable so both LoR and hNN perform well. Accuracy remains high with 55 genes.

Prediction of Mortality for Cancer Patients

- Better model required to predict mortality well.

Next Steps

- Perform mortality prediction differentiated by cancer.
- Explore the underlying biological relationship between selected genes and driver genes.