CJSJ

Genetic Association Testing and Predictive Modeling for Non-Small Cell Lung Carcinoma RNA Sequencing

Lyon Kim

Abstract— **We developed Python code to analyze a bulk RNA sequencing dataset consisting of lung tissues from healthy and non-small cell lung carcinoma (NSCLC) patients. Our preliminary goal was to find genes that were positively or negatively associated with cancer. To do so, we tested the null hypothesis, which stated that the average gene expression between cancer and non-cancer patients is equal. We rejected the null hypothesis since 86.8% of the genes showed a difference in expression. After finding genes associated with cancer, we built a machine learning logistic regression model from the gene expression data. To efficiently measure the performance of our method's ability to predict cancer, we randomly split the data into training (80% of data) and testing datasets (20%) and used five-fold cross validation. By adjusting the probability threshold for classifying cancer, we created an ROC curve, representing the trade-off between the fpr (false positive rate) and the tpr (true positive rate). Ultimately, we hope that these mechanisms can help increase the chance for an early diagnosis of NSCLC, which is crucial to controlling and even preventing it.**

I. INTRODUCTION

X = gene expression (matrix - $N \times M$)
N = number of patients (1118)
M = number of genes (10077)
y = cancer labels (vector of length N) (0 is non-cancer, 1 is cancer)
μ^n = vector of means of each gene in non-cancer patients (has length of M)
μ^c = vector of means of each gene in cancer patients (has length of M)
N_n = number of non-cancer patients
N_c = number of cancer patients
σ_n = vector of the standard deviation of the gene expression of non-cancer patients
σ_c = vector of the standard deviation of the gene expression of cancer patients
X^n_{ji} = an arbitrary data point from a non-cancer patient that is drawn from a normal distribution in which the mean is μ^n and the variance is σ^2_n
X^c_{ji} = an arbitrary data point from a cancer patient is drawn from a normal distribution in which the mean is μ^c and the variance is σ^2_c
Z = Z values: represent how many standard deviations are in the difference of empirical means (has length of M)
$\bar{Z}$ = normalized Z values: Z values scaled to standard normal distribution under the null hypothesis
ℓ_j = the quantitative measure of how much each patient corresponds to the likelihood of cancer (has length of N)
p_j = predicted probability that patient j has cancer
β = vector of coefficients for every gene that directly relates to getting the probability of cancer
D^c = a measure of the amount of cancer genes expressed by the patient
D^n = a measure of the amount of non-cancer genes expressed by the patient

We assumed that each data point (X^n_{ji} and X^c_{ji}) is normally distributed:

$$X^n_{ji} \sim \mathcal{N}(\mu^n_i, (\sigma^n_i)^2) \quad X^c_{ji} \sim \mathcal{N}(\mu^c_i, (\sigma^c_i)^2) \tag{1}$$

First, we calculate the gene expression means ($\hat{\mu}^n$ and $\hat{\mu}^c$) for each gene i for $1 \le i \le M$

$$\hat{\mu}^n_i = \frac{\sum_{j=1}^{N_n} X^n_{ji}}{N_n}, \quad \hat{\mu}^c_i = \frac{\sum_{j=1}^{N_c} X^c_{ji}}{N_c} \tag{2}$$

We calculate σ_n and σ_c

$$(\sigma^n_i)^2 = \frac{\sum_{j=1}^{N_n}(X^n_{ji} - \mu^n_i)^2}{N_n}, \quad (\sigma^c_i)^2 = \frac{\sum_{j=1}^{N_c}(X^c_{ji} - \mu^c_i)^2}{N_c} \tag{3}$$

Figure 1. Defining Notation

II. METHOD

Discovering Genes Associated with NSCLC

We aimed to detect genes with a significant difference in gene expression between cancer and non-cancer patients. To do so, we tested the null hypothesis that the average gene expression in both types of patients is equal. The data we used was from a normalized, merged lung cancer transcriptome dataset that listed for each patient whether each gene was expressed or not through methods of differential expression analysis, batch effect correction, and filtering of genes with low variance.

Testing each Gene's Significance

Next, we attempted to test the null hypothesis that the two means are equal ($\mu^n = \mu^c$). For genes that we reject the null hypothesis, it is unlikely that differences between the two conditions are solely due to experimental noise. The empirical average of the sample for non-cancer and cancer patients can be calculated as:

$$\hat{\mu_i}^n \sim \mathcal{N}(\mu^n_i, \frac{(\sigma^n_i)^2}{N_n}), \quad \hat{\mu_i}^c \sim \mathcal{N}(\mu^c_i, \frac{(\sigma^c_i)^2}{N_c}) \tag{4}$$

Next, we define σ^2 to be a measure of the variability of each gene, Z, and $\bar{Z}$. The Z value represents the number of standard deviations between the difference of cancer and non-cancer means for a given gene.

$$\sigma^2_i = \frac{\frac{(\sigma^n_i)^2}{N_n} + \frac{(\sigma^c_i)^2}{N_c}}{1/N_n + 1/N_c} \quad Z_i = \frac{\hat{\mu^c_i} - \hat{\mu^c_i}}{\sigma_i} \quad \bar{Z}_i = \frac{\hat{\mu^c_i} - \hat{\mu^c_i}}{\sigma_i \sqrt{\frac{1}{N_n} + \frac{1}{N_c}}} \tag{5}$$

We can then calculate the null distribution of Z.

$$Z_i \sim \mathcal{N}(0, \frac{1}{N_c} + \frac{1}{N_n}), \quad \bar{Z}_i \sim \mathcal{N}(0, 1) \tag{6}$$

Now, in order to test the null hypothesis, we tested whether $|\bar{Z}| > 3.33$ and found 8747 genes to be significant. Next, we prove that genes passing the null hypothesis test have a p-value of less than 0.001.

$$P(|\bar{Z}|_i > 3.33) = 2(P(\bar{Z}_i > 3.33)) \approx 0.00086 < 0.001 \tag{7}$$

Figure 2. Method for Testing Gene Significance

Creating a Logistic Regression Model for the Prediction of NSCLC

In the second part of our project, we used the package scikit-learn to fit a logistic regression model to our data. By five-fold cross validating our model, we were able to show that our model predicts cancer with high accuracy from gene expression.
In the logistic regression model, we predict the probability of having cancer (p_j) as:

$$p_j = \frac{1}{1 + e^{-\ell_j}} \tag{8}$$

Now we define ℓ as written below with β_0 being the intercept:

$$\ell_j = \sum_{i=1}^{M} (\beta_i * X_{ji}) + \beta_0 \tag{9}$$

We define the likelihood (L) as the probability of observing y given X, which is the product of the probabilities of observing each patient's condition. β is chosen to maximize the likelihood of observing all the data points.

$$L = P(y|X) = \prod_{j=1}^{N} (p_j^{y_j} * (1 - p_j)^{1-y_j}) \tag{10}$$

Next, we randomly split our data into training (80%) and testing data (20%), with a different 20% for each of the five partitions. This method is used to measure our ability to predict future data based on past data. In order to classify patient j as cancer or non-cancer, we thresholded the predicted probabilities at α. By varying α from 0.00 to 1.00, we generated an ROC curve of the true positive rate (tpr) vs. the false positive rate (fpr). The tpr is defined as proportion of cancer patients correctly classified. The fpr, on the other hand, is the proportion of non-cancer patients incorrectly classified.

III. RESULTS

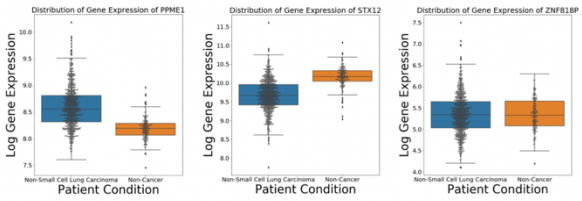

Figure 3. Genes Associated with NSCLC

These three graphs are boxplots of gene expressions in both cancer and non-cancer patients.

- Left: This boxplot represents the gene, PPME1, with the largest decrease in average gene expression from cancer to non-cancer patients.
- Middle: This second graph represents the gene, STX12, with the largest increase in average gene expression from cancer to non-cancer patients.
- Right: This gene, ZNF818P, was found to have no change in average gene expression from cancer to non-cancer patients. This graph serves as a control gene.

Therefore, the gene PPME1, which is highly prevalent in cancer patients, is associated with NSCLC, whereas the gene STX12 is negatively associated with NSCLC. Studying these genes may yield insight into the biological mechanism of NSCLC. In fact, inhibition of PPME1 has already been proven to prevent cell proliferation and induce apoptosis in certain cancer cells [2]. On the other hand, the gene STX12, which is a part of the SNARE complex, helps to control tumorigenesis by regulating cancer cell invasion [3].

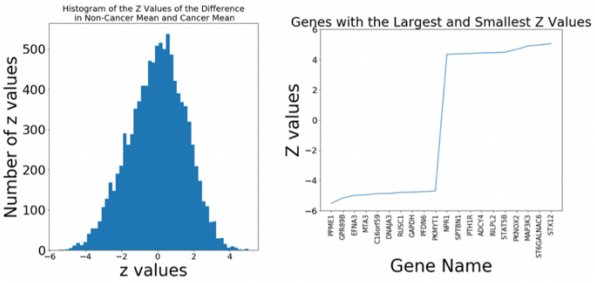

Figure 4. Genes with the Largest and Smallest Z Values are most associated with NSCLC

By finding the Z values, we were able to quantify the difference of the average gene expression between cancer and non-cancer patients. We used the Z values to find candidate genes that could be related to cancer. Next, we tested for each gene for significance, and we found 8747 significant genes, including the twenty genes in the figure on the left.

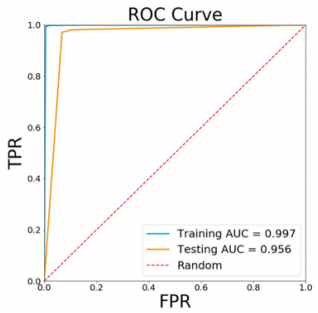

Figure 6. Training and Testing Accuracy

We randomly split our data into training (80%) and testing (20%) data in five fold cross-validation. This method is used to measure our ability to predict future data based on past data. We use the training data to create the model and use the testing set as if it were new data to see how well our model works. The ROC curve model performed significantly better on the training data (Fig. 6). On the other hand, the testing data represents a more realistic measure of an ability to predict on new patients. The overall testing accuracy was around 96%, indicating that this model can be used for early detection of NSCLC. With methods for early diagnosis, it could lead to an improved prognosis. This model can be further improved in the future by training the model with more data and more careful gene selection.

IV. CONCLUSION

In this project, we successfully detected more than 20 genes that were both positively and negatively associated with cancer, specifically non-small cell lung carcinoma (NSCLC), and fit a logistic regression model that predicts a patient's risk of cancer based on their RNA sequencing data. In the future, with the provided merged lung cancer transcriptome dataset, we would modify our predictive logistic regression model to take in only the genes that we found to be significant, or associated with cancer, hence combining my two methods. We would also test for significant genes at different stages of cancer. From this, we could predict a patient's risk of a certain stage of NSCLC and their survival rate. We believe that by finding genes associated with cancer, we can shed light on the biological mechanism of cancer and raise awareness about these genes. Our predictive logistic regression model can be used for early detection of NSCLC. In addition, we believe that it can also be used to detect other types of cancer by training a new logistic regression model on another dataset.

V. REFERENCES

[1] Lim, Su Bin, Swee Jin Tan, Wan-Teck Lim, and Chwee Teck Lim. "A merged lung cancer transcriptome dataset for clinical predictive modeling." Scientific data 5 (2018): 180136.
[2] Li, Jing, Sufang Han, Ziliang Qian, Xinying Su, Shuqiong Fan, JiangangFu, Yuanjie Liu et al. "Genetic amplification of PPME1 in gastric and lung cancer and its potential as a novel therapeutic target." Cancer biology & therapy 15, no. 1 (2014): 128-134.
[3] Meng, Jianghui, and Jiafu Wang. "Role of SNARE proteins in tumourigenesis and their potential as targets for novel anti-cancer therapeutics." Biochimica et Biophysica Acta (BBA)-Reviews on Cancer 1856, no. 1 (2015): 1-12.