Title: Gene selection to predict the cancer type from genome-scale CRISPR–Cas9 screens
Life Sciences

Team members : HYUNG JUN YANG (006183230), HONG-PYO LEE (005877724)

Abstract

Cancer genome scale screens have provided tons of useful data about the functional role of numerous genes to drive cancer growth. However, it is very challenging to find the genetic mutations used for new drug targets from the huge amounts of genomic data, since drug targetable genes should be only functioning on the growth of a specific cancer type. Here, we applied machine learning algorithms to find the drug targetable set of genes which only classify a specific cancer type. Among diverse methods, Recursive feature elimination (RFE) -Random forest and Lasso-support vector machine (SVM) show 90.7% and 89.6% average accuracy for classification of carcinoma cancer type, respectively. Surprisingly, only 200 genes among ~17000 genes are selected as the optimal number of genes for the classification of carcinoma cancer with 92.3% prediction accuracy. Since the set of 200 genes for classification carcinoma means biologically functioning on only the carcinoma cancer type, we strongly expect to find new drug targets against the growth of carcinoma cancer from the 200 genes.

1. Introduction

Cancer is a complex disease derived from genetic and epigenetic mutations that develop uncontrollable cell proliferation. With sequencing and genetic techniques remarkably advancing, numerous mutations has been founded across diverse cancer types. DNA micro-arrays has been widely applied to observe thousands of mutational genes simultaneously in cancers and determine whether those genes are active, hyperactive or silent in normal or cancerous tissue. However, it is not clear whether such mutations and genes in cancers are functional cancer drivers. Therefore, a central challenge to the development of new cancer therapies is to systematically investigate the role of these mutational genes to derive cancer uncontrollable growth.

As the CRISPR-Cas9 system has emerged as a powerful tool for genome editing and transcriptional regulation of specific genes in a variety of cancer types, this technique has been remarkably improved the accuracy of testing functional roles of genes to derive cancer growth. Extraordinary efforts have characterized cancer growth dependencies of all genes using genome-scale in vitro CRISPR screens in hundreds of cancer cell lines. Since this new technique generates huge amounts of raw data, new analytical methods must be developed to sort out whether each cancer type have distinctive signatures of gene function on cancer growth over other types of cancer cells. Further, the distinctive information of genes possibly gives to deeply understand the molecular mechanisms about a growth of specific cancer type, which leads to new targets for cancer therapies.

In this work, we applied machine learning algorithms to find the optimal set to classify cancer type with raw data with hypothesis that each cancer type has a distinct set of genes which represent whole growth phenotype of the specific cancer. Our primary goal is to identify a minimal feature (gene) set that still achieves reasonable classification of cancer type, so that feature selection is the key problem in this project.

2. Dataset and Features

Two publicly available datasets were used for this work. Both set of data were obtained through the Dep Map project which have characterized the role of genes on cancer growth phenotype using genome-scale in vitro CRISPR screens. The first set of data characterized from 517 cancer cell lines was used for training and validation purposes. The second set of data characterized from 325 cancer cell lines was used for testing purposes. The each cell line has 16,183 gene scoring results and the total number of features we used is 16,183.

In both data sets, the raw data was in the form of a matrix containing positive or negative numbers indicating the impact (dependency) of a particular gene on cancer cell growth. If the positive number or negative number were represented in a particular gene (column) and a cell line (row), that indicated the role of the gene was respectively to inhibit (positive number) or promote (negative number) growth of the cancer cell.

3. Methods

a. Preprocessing data

We categorized cancer cell lines to 2 subtypes of cancer such as carcinoma and non-carcinoma. Total 517 cancer cell lines in training set were divided as 301 carcinoma and 216 non-carcinoma.

b. Principal component analysis (PCA)

Due to the vast number of features (~18000 genes), we used principal component analysis (PCA) to reduce dimensions and visualize the data, and examined if the first two or three principal components suggested distinct clusters that correspond to relevant characteristics of the cancer type. PCA is frequently used for dimension reduction, although the linear algorithm of PCA may not be able to capture non-linear relationships between data dimensions.

c. Filter method

To reduce the number of features and remove the redundant features, we applied two filter feature selection methods such as correlation based feature selection and feature selection by F-test statistics. The correlation based feature selection evaluates the quality of correlations among a feature, other features and the classification to select the feature subset which highly correlated to the classification, independently to other features. The feature selection by F-test statistics is a method to find the maximum relevance between a gene and a class label and then the correlation of the gene pair in that class is measured to minimize redundancy.

d. Wrapper methods

i. Recursive Feature Elimination (RFM with Support Vector Machine or Random forest)

We applied Recursive Feature Elimination to find the optimal set of genes for classification from the whole set of genes. As briefly explained the algorithm, we trained all samples with support vector machine (SVM) or random forest (RF) with the selected features and obtained all weight vectors, W_k (weight vector of k^{th} cancer type). Among all elements of W_k, remove the features with the smallest | W_{jk} | on the basis of the following hypothesis " the importance of a feature j should be related to the magnitude of its weight | W_{jk} |. Until getting the optimal set of genes with given number (m), iterate this process. The whole procedure was following:

1. Train SVM (or RF) on the active features.
2. Remove the j^{th} feature with the smallest $\Sigma_k w^2_{jk}$.
3. Go back to step 1 or stop if there are only m features

Linear SVM and RF was used in this process and when retraining the models in step 1, we used the weight vector from the previous step as a starting point. We applied the RFE algorithm in sklernn package to our data.

e. Embedded methods (selection from model)

Without retraining the model, we can estimate the feature importance of each features by the coefficients each features after one training. We used three different models for selection.

i. Least Absolute Shrinkage and Selection Operator (LASSO) method

LASSO is used as one of embedded methods to select the subset of genes. LASSO is a linear regression with L1 regularization. A model loss function is constructed with additional regularization term $\lambda||w||_1$ on the training data (x_i , y_i), i = 1...., N, where w is a k dimensional vector of weights corresponding to each feature dimension k. The L1 regularization term ($\lambda||w||1$) pushes the weights of correlated features to zero, thus prevents overfitting and improves model performance. Model interpretation ranks the features according to the LASSO feature weights.

ii. Random Forest

Random forests is an ensemble learning method for classification and regressions. The training algorithm for the random decision forest is bagging (bootstrap aggregating). The bagging repeatedly selects a random sample with replacement and fit the models to these samples. Then, prediction of new data is made by averaging the predictions from all the individual models. This bootstrapping method decreases the variance of the model without increasing the bias. The advantage of this method is that it can evaluate the importance of features. It can be done by permuting the feature of train data and computing the out-of-bag error before and after the permutation.

iii. Deep neural network (DNN)

DNN is an artificial neural network with multi-layers between the input and output layers. The main advantage of DNN is that it can model complex non-linear relationships. We use DNN for feature selection by estimating the feature importance using permutation. Using permutation importance, we replace the certain feature with random noise instead of removing a feature. Then, we compute the score after replacing and evaluating the feature importance. We used Keras package for deep neural network and Eli5 package for estimating the permutation importance.

4. Results

a. PCA

Projection along PCA plane was not able to distinctively recognize the difference between carcinoma cancer and others.

b. Performance comparison with respect to accuracy

The prediction accuracies of all methods are relatively high considering the fact that we used less than 1,000 features among 16,183 features. Table 1 shows that Select from Lasso SVM and RFE-RF show the highest performance. The prediction accuracy of them is about 0.9 and it is very high value considering the fact that we selects only less than 1,000 features among 16,183 features. Figure 2 also shows that the performance of each methods is dependent on the number of features to select. Especially, the performance of the selection from DNN is most sensitive to the number of features. The best prediction accuracy is obtained when the number of features are 200, 300, and 900. Our objective is to have minimum number of features for reducing cost for gene scoring. Therefore, we conclude 200 features using Lasso SVM as the optimal number of feature and feature selection method.

Table 1 Average prediction accuracy of each feature selection methods

Methods	**F test**	**Select from Lasso SVM**	**Select from RF**	**Select from DNN**	**RFE-SVM**	**RFE-RF**
Average Accuracy	0.874	0.907	0.892	0.862	0.881	0.896

a. Performance comparison in Deep leaning method with respect to the number of layers and nodes

From several experiments, we observe that feature selection from DNN is highly sensitive to the diverse hyper-parameters of DNN. Especially, the number of layers and nodes highly affect to the performance of DNN based feature selection. We performed sensitivity analysis for the performance of DNN based feature selection with respect to the number of layers and nodes. Figure 3 shows that the performance of DNN based feature selection method is highly sensitive to the number of layers and nodes.

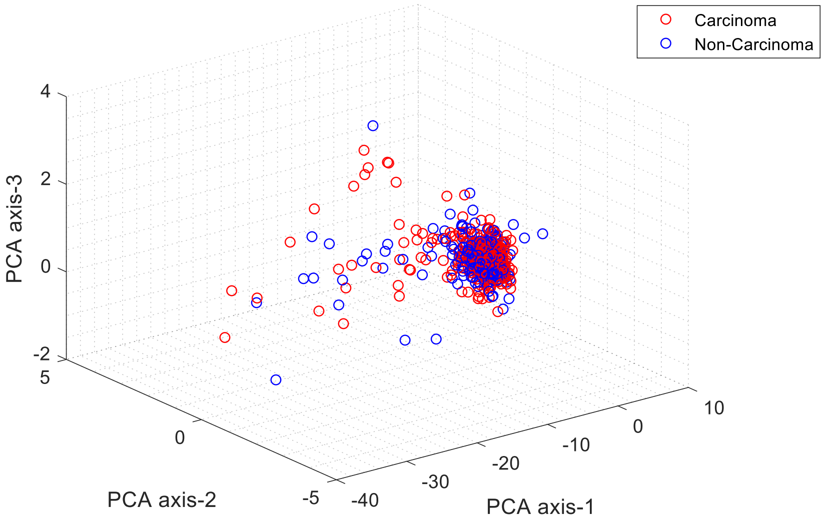

Figure 1 Projection of data to PCA plane

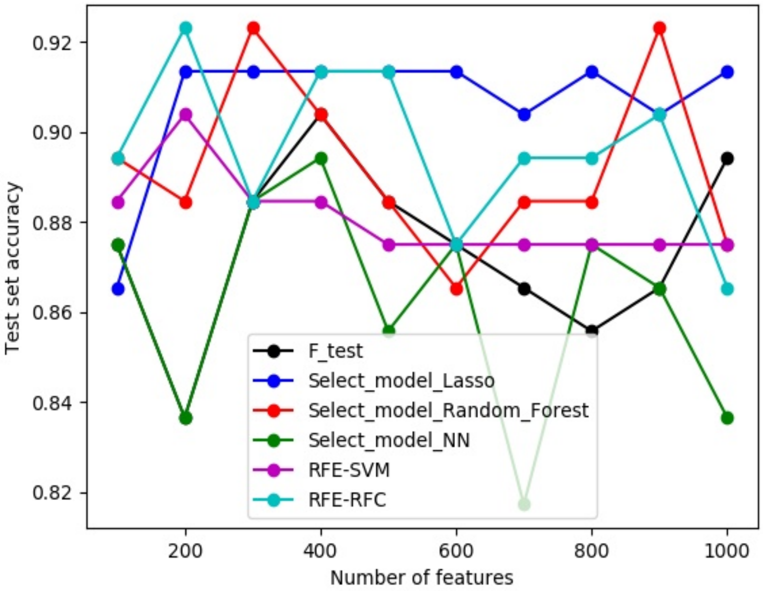

Figure 2 Comparing prediction accuracy with respect to methods and the number of features

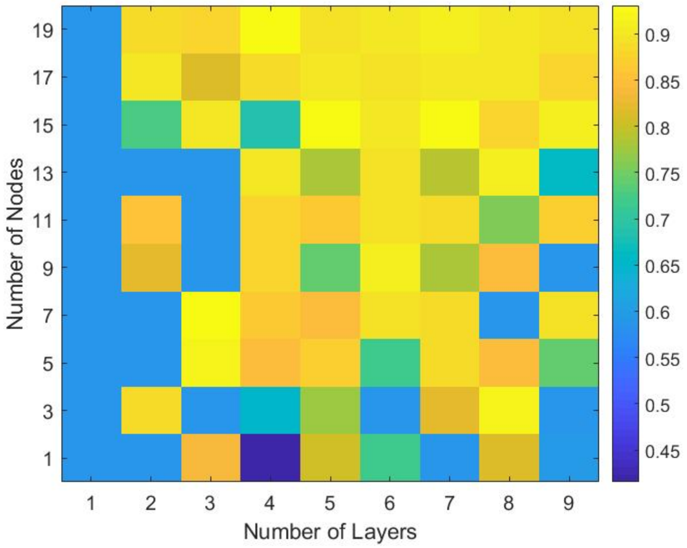

Figure 3 Prediction accuracy of deep learning based feature selection with respect to the number of layers and nodes.

5. Discussion

In summary, through machine learning methods, we are able to classify carcinoma cancer types from the given information of genome scale screens and identify the major set of genes related with carcinoma cancers in fairly good accuracy. Among feature selection diverse methods, RFE-Random forest and select from Lasso SVM show the best average performance. The optimal number of feature for cancer type classification is 200, and its prediction accuracy is 0.923. Moreover, the performance of deep learning-based feature selection is highly sensitive to the number of nodes and layers. Thus, Hyper-parameter optimization is necessary for efficient deep learning-based feature selection. Eventually, we hope to personalize treatment plans based on their predicted disease outcomes, thereby improving the quality of care and reducing the cost of cancer management.

6. Future works

We would like to extend our methodology to explore more specified types of cancers (4 types of cancers such as blood cancer, carcinoma, sarcoma, and others). Furthermore, it is worthy to examine the biological functions represented from the 400 selected genes to find the new target for carcinoma cancer treatment, since the selected features for the classification of carcinoma cancer could be specifically involved in the growth of carcinoma cancer.

7. Contribution

H.Y. and H.L. designed the general strategies for this machine learning tasks. H.L. collected and modified raw data. H.Y. created and revised codes for machine learning algorithms of our strategy. H.L. and H.Y. wrote the manuscript.

8. Reference

1. Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.

2. Li, Y., Chen, C. Y., & Wasserman, W. W. (2015, April). Deep feature selection: Theory and application to identify enhancers and promoters. In International Conference on Research in Computational Molecular Biology (pp. 205-217). Springer, Cham.