Gene selection to predict the cancer type from genome-scale CRISPR–Cas9 screens

HYUNG JUN YANG 1 and HONG-PYO LEE2

1Department of Energy Resources Engineering, Stanford University
2Department of Mechanical Engineering, Stanford University

BACKGROUND AND SIGNIFICANCE

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 found 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 dependences 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. 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.

APPLIED METHODS FOR GENE SELECTION

CONCLUSIONS AND FUTURE WORK

Conclusions

- Among diverse methods, RFE-Random forest and select from Lasso SVM show the best average performance.

<table>
<thead>
<tr>
<th>Methods</th>
<th>F_test</th>
<th>Select from Lasso SVM</th>
<th>Select from Random forest</th>
<th>Select from DNN</th>
<th>RFE-SVM</th>
<th>RFE-Random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Accuracy</td>
<td>0.874</td>
<td>0.907</td>
<td>0.892</td>
<td>0.862</td>
<td>0.881</td>
<td>0.896</td>
</tr>
</tbody>
</table>

- The optimal number of feature for cancer type classification is 200, and its prediction accuracy is 0.923.
- The performance of deep learning-based feature selection is highly sensitive to the number of nodes and layers.
- Hyper-parameter optimization is necessary for efficient deep learning-based feature selection.

Future work

- Explore more specified types of cancers (4 types of cancers such as blood cancer, carcinoma, sarcoma, and others)
- Examine the biological functions represented from the selected genes to find the new target for specific cancer treatment

REFERENCES


Two publicly available datasets were used for this work. Both set of data were obtained through the DepMap 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 25 cancer cell lines was used for testing purposes. Total number of features is 16,183, and feature selection is performed among these features.

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 growth. If the positive number or negative number were 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 cells.