Understanding Gene Dependency Map and Machine Learning Classification of Cell Lineages

Group Member: Yiling Peng* Department of Mathematics, Northeastern University

April 20, 2023

1 Background information

Cancer is a leading cause of mortality worldwide, with an estimated 10 million cancer-related deaths in 2020 alone. In recent years, there has been a growing interest in using genomic data to better understand cancer vulnerabilities and develop more effective treatments. A global healthcare company based in New Jersey is committed to improving worldwide health by providing innovative health solutions. As part of this commitment, the company is interested in using publicly available genomic data sets to identify cancer vulnerabilities and develop predictive models.

The task at hand involves analyzing CRISPR KO gene dependencies, mutational and copy number profiles, and RNA expression data from over 1000 cancer cell lines. By integrating diverse data types and machine learning methods, the company aims to identify cancer vulnerabilities and differential cancer gene dependencies that can be targeted for effective cancer treatments.

One key aspect of this research is to classify the lineages that depend on essential genes for tumorigenesis. To achieve this, the company seeks suitable algorithms that can integrate the various data types and machine learning methods. The research seeks to answer the following questions: What are the cancer vulnerabilities and differential cancer gene dependencies that can be identified? How can predictive models be classified using CRISPR KO gene dependencies, mutational and copy number profiles, and RNA expression data from over 1000 cancer cell lines that share common genetic determinants?

The findings from this research are expected to contribute to the development of more effective cancer treatments that can improve the quality of life of cancer patients worldwide.

2 Research Questions

In this project we will adress the following questions:

1)Find the most appropriate machine learning algorithm for data classification and identification of significant clusters of lineages.

2)Evaluate the uniqueness of classification schemes by linking gene dependencies to other data sets like RNA expression, and copy number data.

3)Assess if the machine learning algorithm can identify predictors of general viability loss, not limited to specific lineage-specific genes only.

3 Related works

The previous study from [1] explored the use of computational intelligence methods to predict therapeutic targets in nine different cancer types. Simple models achieved reasonable performance for some cancers, but more complex models integrating gene-gene interaction data via network embedding features and a robust feature selection approach produced well-performing models for all nine cancer types. Limitations included biases in the training dataset and difficulty in interpreting important network features in terms of biological or graph-based properties. The study demonstrated the potential of computational methods to generate hypotheses for therapeutic targets in oncology.

3.1 Gene expression

The study in [2] indicates that individual gene expression profiles possess predictive value for various perturbations, which cannot be accurately captured by basic measures of gene-set level expression.

4 Dataset and Features

The Cancer Cell Line Encyclopedia (CCLE) is a comprehensive collection of cancer cell line genomic, transcriptomic, and pharmacological data, available for download on the DepMap portal. With information on 1,086 cell lines from 31 primary diseases, including mutational data, copy number profiles, RNA expression data, and CRISPR KO gene dependency data, the dataset has been extensively processed. In this paper, we use the DepMap CCLE data to investigate the gene essentiality of cancer cells across multiple lineages

and primary diseases, focusing on the gene effect scores derived from CRISPR knockout screens. We also leverage other genomic data, including mutational and copy number profiles, to deepen our understanding of cancer cell biology. Through this analysis, we aim to gain insights into the molecular mechanisms driving cancer progression and contribute to the existing knowledge base. Access the dataset at: https://depmap.org/portal/download/

4.1 Datatype and background knowledge

Geneko technology refers to the use of the CRISPR/Cas9 system to perform gene knockouts in a high-throughput manner. This approach allows researchers to systematically study the effect of knocking out each gene in a given cell type or organism, and to identify genes that are essential for survival or have other important functions.

Gene expression refers to the process by which the information encoded in a gene is used to create a functional protein or RNA molecule. This process involves the transcription of DNA into RNA, and the subsequent translation of RNA into protein. The level of gene expression can be measured using various techniques, such as RNA sequencing (RNA-seq), and provides information about which genes are being actively used by a cell at a given time.

Gene copy number refers to the number of copies of a particular gene that are present in an organism's genome. Changes in gene copy number can have a significant impact on gene expression and protein function, and are associated with a range of diseases, including cancer.

The Achilles and SCORE projects are two large-scale efforts to perform CRISPR knockout screens in a variety of cell types. By measuring the effect of gene knockouts on cell growth and survival, these projects have generated a large dataset of gene effect scores, which can be used to identify genes that are essential for cell viability or have other important functions.

Dataset	**Data Type**	**Dimension**
CRISPR gene effect	Numerical	17386x1086
CRISPR gene dependency	Numerical	17386x1086
CCLE gene count number	Numerical	25368x1766
CCLE expression	Numerical	19221x 1406
sample info	Categorical	1840

Table 1: Description of the data used in our analysis.

5 Machine Learning Methods

5.1 Correlation analysis

Pearson correlation analysis is a statistical method used to measure the strength and direction of the linear relationship between two continuous variables. It calculates the correlation coefficient, which ranges from -1 to +1, where -1 represents a perfect negative correlation, +1 represents a perfect positive correlation, and 0 represents no correlation. In various fields, including biology, medicine, and finance, Pearson correlation analysis has numerous application. In my research paper, I used Pearson correlation analysis to identify the genes that are most essential to the viability of cancer cells.

5.2 PCA and TSNE analysis

Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are commonly used techniques for dimensionality reduction and data clustering. PCA is a statistical method that reduces the dimensionality of large datasets while retaining as much information as possible. It does so by identifying the underlying patterns in the data and projecting them onto a lower-dimensional space. t-SNE, on the other hand, is a nonlinear dimensionality reduction technique that is particularly useful for visualizing high-dimensional datasets. It aims to preserve the pairwise distances between points in the original high-dimensional space by creating a low-dimensional representation of the data. Both PCA and t-SNE are powerful tools for data analysis and visualization, and are frequently used in various fields, including machine learning, biology, and social sciences, among others.

5.3 logistic regression and Linear regression

Logistic regression and linear regression are commonly used statistical techniques in machine learning and data analysis. Linear regression models the relationship between a dependent variable and one or more independent variables to predict future observations. In contrast, logistic regression models the relationship between a binary dependent variable and one or more independent variables, and is useful when predicting the probability of a binary outcome. Logistic regression provides a probability estimation, making it useful for decision-making tasks. Linear regression is versatile and can handle a variety of predictors, making it useful for both prediction and interpretation tasks. Together, these techniques can be used to analyze data and make accurate predictions.

5.4 Neural Network

Neural networks classify complex data by identifying patterns and relationships through interconnected layers. Each layer produces a transformed output, with the final layer predicting the class or label. The network adjusts its connections during training to improve accuracy. Neural networks are a proven effective method for classification.

5.5 SVM

SVM is a popular machine learning method for classification and regression analysis that constructs a hyperplane or multiple hyperplanes to separate different classes in a dataset. It aims to find the hyperplane with the largest margin between classes and can handle both linearly and non-linearly separable datasets using various kernel functions. SVM is useful for high-dimensional data, particularly when the number of features exceeds the number of samples.

5.6 Random Forset

Random forest is a versatile machine learning technique for classification and regression analysis. It consists of a collection of decision trees, where each tree makes a prediction for the target variable. The random forest algorithm randomly selects a subset of features and samples to build each decision tree, which helps to reduce overfitting and improve accuracy. During prediction, the algorithm aggregates the predictions of all trees to make a final decision. Random forest can handle both categorical and continuous data and is effective in handling high-dimensional datasets.

5.7 KNN

K-Nearest Neighbors (KNN) is a machine learning algorithm used for classification and regression analysis. KNN works by finding the K nearest data points in a training dataset to a new input data point, and then classifying the new data point based on the majority class among its K nearest neighbors. The value of K is a hyperparameter that can be tuned to optimize the model's performance. KNN is a simple and effective method for handling both linear and non-linear datasets, and can work well with noisy data. However, KNN can be computationally expensive and requires a large amount of memory to store the training data.

5.8 KMean

K-means is a widely used unsupervised machine learning technique for clustering analysis. The algorithm aims to partition a dataset into k distinct clusters, where each cluster represents a group of similar data points. K-means works by randomly selecting k initial centroids, assigning each data point to the closest centroid, and then recalculating the centroids based on the mean of the data points in each cluster. This process is repeated until the centroids no longer change or a maximum number of iterations is reached. K-means is useful for finding structure in large datasets .

5.9 Feature Selectiont

Feature selection is a technique used in machine learning to reduce the number of features in a dataset while still retaining the most relevant ones. One popular method for feature selection is SelectKBest, which ranks the features based on their statistical significance and selects the top k features. This approach can help to reduce overfitting and improve the accuracy of the model. SelectKBest can be used with various machine learning algorithms, including SVM, to identify the most important features for classification or regression analysis.

6 Research and Discussion

To start with, we loaded four datasets: gene expression data, knockout effect data, copy number data, and sample information data. We then extracted the intersection of these datasets, which allowed us to focus our analysis on the shared subset of data across all four datasets. This approach enabled us to identify potential relationships between gene expression, knockout effect, and copy number variation in the specific subset of samples represented in all three datasets.

6.1 Identification of Essential Gene

Pearson correlation is a statistical measure that evaluates the linear relationship between two continuous variables. It produces a correlation coefficient, which ranges from -1 to 1, indicating the strength and direction of the association between the variables. In our study, we aimed to identify the genes that are most essential to the viability of cancer cells. To achieve this goal, we employed Pearson correlation analysis.Specifically, we calculated the correlation coefficient between gene expression levels and cancer cell viability. The intuition is that if a gene is highly correlated with cancer cell viability, then its expression levels might play a crucial role in the survival of cancer cells. To identify the genes that are most strongly correlated with cancer cell viability, we sorted the correlation coefficients in descending order and selected the top 100 genes. Moreover, to filter out the genes with low expression levels, we further selected the genes with expression levels above the 90th percentile of the distribution. This approach allowed us to identify a subset of genes that are highly correlated with cancer cell viability and might be potential targets for cancer therapy.

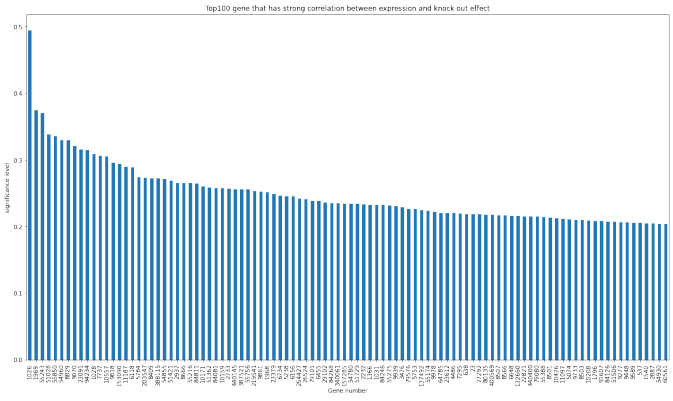

Figure 1: Top100 Gene with strong Expression-Knockout Effect correlation

A few top genes with high correlation coefficients and high expression levels in descending order are: (RPA2, UXT, GSS, EIF3G, ATP6AP2, PSMD6, EIF2S3, SRPRA,). They are good targets for cancer treatment.

6.2 Lineage Classification

After running principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) analyses, I plotted the results of each dataset separately and then combined the expression and copy number datasets to identify any patterns or clusters. The graphs revealed significant clusters within the data, indicating that there may be distinct groupings of cell lineages. These findings suggest that further investigation into the genetic basis of these clusters may lead to the identification of potential therapeutic targets for cancer treatment.

Figure 2: PCA Graph of Labeled by Lineages

To further explore the potential of machine learning algorithms in predicting cell lineage based on multi-omics data, I developed a neural network model that incorporated expression, KO effect, and copy number data. The ANOVA

Figure 3: TSNE Graph of Labeled by Lineages

F-value method was used for feature selection, and the data was scaled using standard scaling. The neural network model had an input layer with 64 nodes, two hidden layers with 128 nodes each, and an output layer with a softmax activation function for classification. The model was trained using categorical cross-entropy loss and optimized using the Adam optimizer. The default metric used to evaluate the model was accuracy, and the model was trained for 20 epochs with a batch size of 32. However, the accuracy of the neural network model was found to be 0.75, which was lower than the logistic regression model with an accuracy of 0.8 and random forest. These results suggest that while neural networks have the potential to predict cell lineage, logistic regression can be an effective tool for this task when using multi-omics data, with important implications for cancer research and identifying potential therapeutic targets.

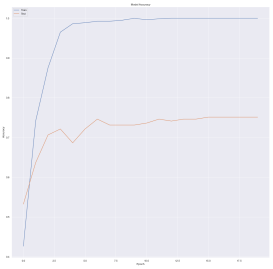

Figure 4: Learning curve of nerual network classification

The study revealed that the impact of feature selection on accuracy varied depending on the machine learning algorithm and the dataset. While feature selection did not significantly improve accuracy in the case of the random forest model, it did lead to a higher accuracy in the logistic regression model. These results suggest that the effectiveness of feature selection is algorithm-dependent and dataset-specific. Therefore, it is crucial to weigh the potential advantages and disadvantages of feature selection carefully before employing it in a machine learning task.

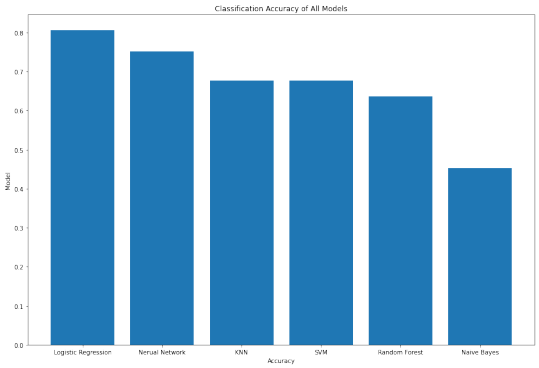

Figure 5: Accuracy Ranking of different Machine Learning method

6.3 Compare KMeans

We employed the K-means clustering algorithm to cluster different datasets, without using any predefined labels. However, we observed that the resulting graphs of the various datasets were diverse and dissimilar from one another. To assess the effectiveness of the clustering algorithm, we compared the Rand score of each classification scheme. The analysis revealed that the classification schemes were distinctive, and the Rand score differed across them. This underscores the significance of selecting the appropriate clustering algorithm and optimal parameters for each dataset, as the performance of the algorithm can fluctuate significantly depending on the data.

6.4 CERES Scores

CRISPR-Cas9 knockout screens are commonly used to infer gene essentiality, and [2]CERES (CRISPR Enrichment through Reference Alignment and Synthetic Evaluation) is a computational method that generates CERES scores for this purpose. These scores are obtained by normalizing the observed fold-changes of guide RNA (gRNA) abundance to account for differences in the effectiveness of gRNA targeting and cell growth rates. Positive CERES scores indicate essential genes, whereas negative scores indicate non-essential or toxic genes. . In summary, CERES scores provide a way to generate cell viability data that can be used to summarize lineage features and to train general viability models.

The formula for CERES (CRISPR Enrichment through Reference Alignment and Synthetic Evaluation) score is:

$$\text{CERES} = \log_2\left[\frac{(1-w)(\text{shRNA}_\text{c}\text{ount}+0.5)}{(\text{total}_\text{c}\text{ount}+1)} + \frac{w\text{median}(\text{shRNA}_\text{c}\text{ount}+0.5)}{(\text{total}_\text{c}\text{ount}+1)}\right] - \text{gene}_\text{e}\text{ffect}_\text{m}\text{edian} \quad (1)$$

where: w is a weight used to balance individual shRNA effects with the median shRNA effect. $\text{shRNA}_\text{c}\text{ount}$ is the number of reads for a specific shRNA. $\text{total}_\text{c}\text{ount}$ is the total number of reads for all shRNAs targeting a gene. $\text{gene}_\text{e}\text{ffect}_\text{m}\text{edian}$ is the median gene effect across all cell lines.

6.5 Lineage Cluster

In this study, CERES scores were utilized to analyze and cluster cell lineages. Specifically, the author computed the average CERES score from cells in a specific lineage and applied K-mean and Agglomerative Clustering algorithms to cluster the lineages. To evaluate the similarity between the clusters generated by both algorithms, the author measured the Rand Index between the labels generated by each algorithm. The findings indicate that the clusters generated by both algorithms are similar. This work highlights the potential of CERES scores and clustering algorithms for analyzing cell lineages, providing important implications for future research in this field.

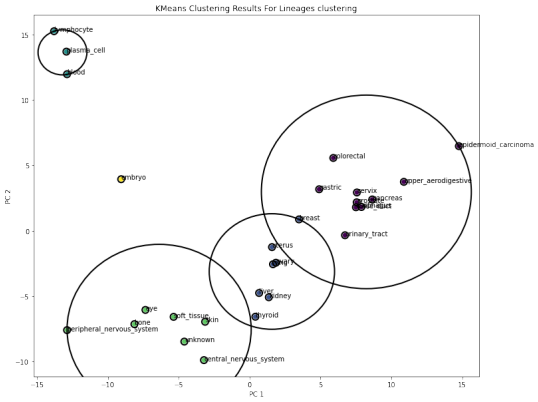

Figure 6: KMeans Clustering Results For Lineages clustering

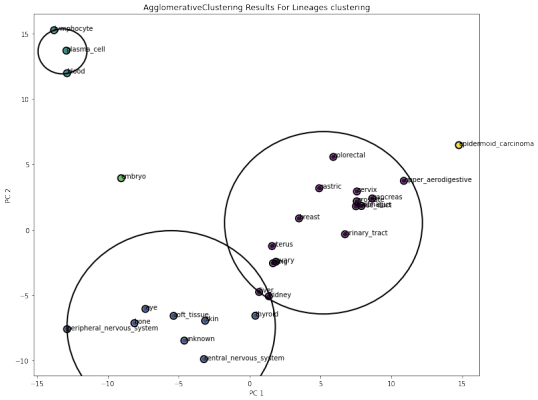

Figure 7: AgglomerativeClustering Results For Lineages clustering

6.6 Predicator of General Viability

To find out the third question, I used CERES scores and linear regression to train models that predict general viability

loss based on expression data. Specifically, I used linear regression to train the models, with expression data as the input and CERES scores as the output. The results indicate that the models achieved a high level of accuracy, with a 99 % accuracy rate. This work provides important insights into the use of CERES scores and linear regression for predicting general viability loss and has implications for future research in this area.

7 Project Conclusion

In conclusion, this study investigated the gene essentiality of cancer cells across multiple lineages and primary diseases using the DepMap CCLE data. The study employed a combination of techniques, including Pearson correlation analysis, PCA, t-SNE, and machine learning algorithms, to identify potential relationships between gene expression, knockout effect, and copy number variation in the specific subset of samples represented in all three datasets. The study revealed that machine learning algorithms can be effective in predicting cell lineage based on multi-omics data, with important implications for cancer research and identifying potential therapeutic targets. However, the effectiveness of feature selection is algorithm-dependent and dataset-specific, and it is crucial to weigh the potential advantages and disadvantages of feature selection carefully before employing it in a machine learning task. Overall, the findings of this study contribute to the existing knowledge base of cancer cell biology and provide insights into the molecular mechanisms driving cancer progression.

References

[1] Bazaga, A., Leggate, D., Weisser, H. Genome-wide investigation of gene-cancer associations for the prediction of novel therapeutic targets in oncology. Scientific Reports, 10(1), 10787. https://doi.org/10.1038/s41598-020-67846-1 1

[2] Dempster, J. M., Krill-Burger, J. M., McFarland, J. M., Warren, A., Boehm, J. S., Vazquez, F., Hahn, W. C., Golub, T. R., Tsherniak, A. (2020). Gene expression has more power for predicting in vitro cancer cell vulnerabilities than genomics. bioRxiv. https://doi.org/10.1101/2020.02.21.959627 1, 5

[3] Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., Beroukhim, R. (2017). Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature genetics*, 49(12), 1779-1784.

[4] DepMap. (2023). Cancer Cell Line Encyclopedia (CCLE). Retrieved from https://depmap.org/portal/download/.

[5] Gupta, R. (2019). Machine learning for cancer diagnosis: Challenges and opportunities. Journal of Oncology Practice, 15(7), 371-375.

[6] Hsu, C., Chang, Y. (2020). Feature selection in machine learning for cancer diagnosis: A survey. Pattern Recognition, 106, 107406.