Khan Inan
Manpreet Katari
Course # 7743 Machine Learning and Data Science for Bioinformatics
5/11/2023

Exploring Commonalities between Cancer/Tumor types: The role of Prognostic genes in determining Survival rates

Abstract:

When it comes to data related to cancer, oftentimes most cancers are treated as separate entities and completely independent from one another. Those involved with the Pan-Cancer Analysis Project and The Cancer Genome Atlas (TCGA) are working towards collecting a large amount of genetic data that can then be used to provide a holistic view of all the different cancer and tumor types. Upon the development of this initiative, there have been several links as well as new differences that have been found between different cancers. This information is useful for diagnosis as well as treatment of patients because it allows for more flexibility in the currently very rigid and specialized field of oncology. For my project, I utilized a dataset from the Pan-Cancer Analysis project involving 16 different cancer types. My goal was to draw a connection between the amount of prognostic genes and the survival rates of different cancers and to do so I first used AUC and F1 to determine how useful prognostic genes are in determining the survival of cancer patients, and from there I used the random forest machine learning classifier for prediction

Research Question:

Is the number of prognostic genes an effective predictor of survival rates for cancer?

Hypothesis:

My hypothesis is that there is indeed a relationship between the number of prognostic genes and the survival rate of cancers. I believe that more prognostic genes results in a lower survival rate

and less prognostic genes means a higher survival rate. The main reasoning behind my hypothesis is that I am aware of multiple successful studies where key prognostic genes are linked to the survival rates of gastric and bladder cancers. Based on these studies I am fairly certain that prognostic genes have a major impact on the survival rates of various cancers/tumors.

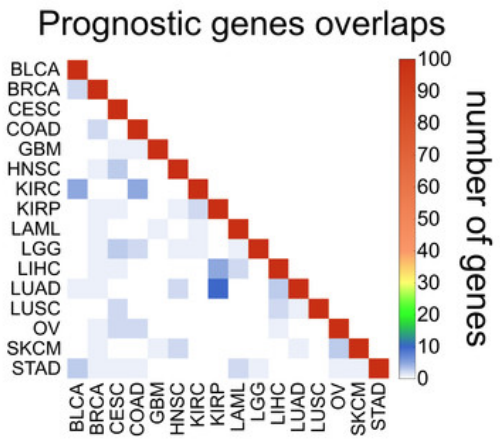
Background Information:

Currently most cancer treatments begin with the collection of anatomical data. This can be in the form of x-rays or biopsies, and although anatomical information is useful it does not give a complete picture of the patient's prognosis. It is for this reason that genomic technologies and methods are gaining in clinical relevance. Disease-specific Prognostic genes are already known to be a good indicator of the probable outcome of various forms of cancer. What is currently unknown for the most part, however, is how the amount of prognostic genes affects survival rates for cancer. Currently how a prognostic gene is defined is as a specific signature or biomarker that can predict the outcome or course of a condition. How prognostic genes differ from non-prognostic genes (predictive genes), in relation to cancer, is that while prognostic genes can tell us how a cancer/tumor can evolve over time, predictive genes tell us more about how the cancer came to be in the first place, whether it be hereditary or environmentally related. Data in regards to prognostic genes and non-prognostic genes are both useful in determining the proper treatment for a patient, however medical practitioners have only recently begun to use prognostic data to give the most realistic outlook possible.

Each cancer has its own set of unique prognostic genes as well as prognostic genes that it shares with other cancers. For my specific data set, I found that there are a few pairs of cancers that have a lot of prognostic genes in common (shown in the figure on the next page.) The most similar pairs of cancers in terms of prognostic genes are KIRP and LUAD, KIRC and BLCA, KIRC and COAD and LIHC and KIRP. One of the most prominent theories in regards to these similarities between the prognostic genes of these seemingly unrelated cancers lies in the composition of the human body. Although all the organs and body parts in the human body have their own unique cells each with different functions and genes, there are still cells that are very

similar and thus are susceptible to the same environmental factors and genetic predispositions that result in cancers and tumor growth. This is just one of the main theories in regards to prognostic genes.

In regards to the specific data set that I have chosen, I wanted to research a bit further into the relationships between these specific cancers and why that may be the case. According to the preliminary data exploration heatmap that I did shown below, I found that Kidney Papillary Cell Carcinoma and Lung Adenocarcinoma have the most prognostic genes in common, which is shown on the y axis. The reason behind this is that kidney papillary cells and lung cells are very similar, and doctors often are aware of the risks that kidney cancer can metastasize to the lungs and vice versa because of this reason.



BLCA: Bladder cancer
BRCA: Breast cancer
CESC: Cervical cancer
COAD: Colon cancer
GBM: Glioblastoma
HNSC: Head and Neck cancer
KIRC: Kidney clear cell cancer
KIRP: Kidney papillary cell carcinoma
LAML: Acute Myeloid Leukemia
LGG: Lower Grade Glioma
LIHC: Liver cancer
LUAD: Lung Adenocarcinoma
LUSC: Lung Squamous Cell Carcinoma
SKCM: Melanoma
OV: Ovarian cancer
STAD: Stomach Cancer

Motivation for Problem Addressed

The main motivation behind looking into this problem is due to the way that cancers are currently treated and diagnosed. Currently how it works is that once it is confirmed that a patient has cancer and it is the early stages of cancer, a very generalized chemo-therapy is performed a majority of the time. The way that this chemo-therapy is tailored to each patient and their specific needs lies in the localized concentration of the chemical agents to where they need it most. Obviously this treatment does not take into account the specific characteristic of each type

of cancer, which is extremely important in determining effectiveness of treatment and future outlook on cancer growth and possible metastasis. This sort of generalized approach almost always results in healthy cells dying in the process. Specialized drugs have been developed for certain cancers but the process is very difficult and slow, and they often are not used until it is the mid to late stages of cancer. As an example, If we take a look at the best selling and effective cancer treatment drugs currently on the market, Keytruda, we can see that a lot of effort has been made to make sure that this drug is as effective as possible for the disease that it is designed to target. As time went on however, the FDA and Merck (the parent pharmaceutical company) both realized that prescriptions of the drugs can be expanded to other forms of cancer if the circumstances are right. Taking a look at the timelines for FDA approval for Keytruda, we can see that it was first approved for melanoma patients who carry a BRAF mutation, it then went on to be approved for lung cancer patients and then Hodgkin's lymphoma patients. I am hoping that with my project I am able to establish some common themes between cancer categories, so that drugs and treatments can be interchanged.

Results, presentation and interpretation:

```python
# import necessary libraries
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import f1_score, roc_auc_score, label_binarize

# load data from CSV file
data = pd.read_csv('pancancerdata.csv')

# extract independent and dependent variables
X = data['FDR ≤.05'].values.reshape(-1, 1) # number of prognostic genes
y = data['Median survival (days)'].values

# binarize the labels
y = label_binarize(y, classes=[0, 1])

# split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# create and fit a Random Forest Classifier model
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)

# make predictions on the testing set
y_pred = rf.predict(X_test)

# calculate F1 score and AUC
f1 = f1_score(y_test, y_pred, average='weighted')
auc = roc_auc_score(y_test, y_pred, average='macro', multi_class='ovo')

# print performance metrics
print('F1 score:', f1)
print('AUC:', auc)

```

Output:

```
F1 score: 0.80221478422253135
AUC: 0.9128205128205129
```

Step by step explanation of my code

1. First load in the data as well as my packages
2. Then I had to pre-process the data using scaling and normalization methods, as well as binarizing certain columns
3. From there it is simply splitting the data into training and testing sets

For the code shown on the previous page it is a calculation of the F1 and AUC score. This score essentially tells me how useful the number of prognostic genes is in determining the survival rates for the various cancers. In interpreting the output, since the F1 score is a relatively high value and is on the higher end of the spectrum from 0 to 1, I can say that the number of prognostic genes is a decently good predictor of survival rate. Similarly, the AUC score is also fairly high which shows a correlation between the two variables.

Since the data set treats each column as seemingly unrelated data categories, I felt the random tree would be the most appropriate model to choose. Using the machine learning models and the metrics of F1 and AUC It seems that the number of prognostic genes is a good predictor of the survival rates for the 16 cancers

Shown on the next page is my code for determining why Random forest would be the best model for my data set. The other models did not give me F1 or AUC scores that were as high as the random forest. There are many possibilities as for why this is the case but the main reason is most likely that the values are for the most part difficult to correlate and thus random forest has the best score compared to any of the models that uses individual and more specialized data.

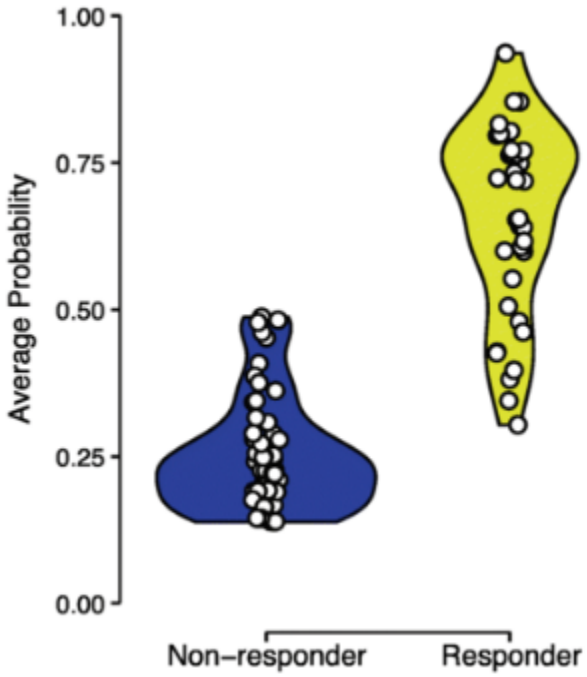
Out of all the other features that were in the data set, like age or gender, it seems that the number of prognostic genes is the best predictor of survival rate which is interesting. My interpretation for why age or gender is not as good as the number of prognostic genes for predicted survival rate is because those metrics are more useful for other metrics such as the severity of cancer symptoms or even the number of mutation events that they face. Prognostic genes are only good for predicting survival rates and survival rates alone.

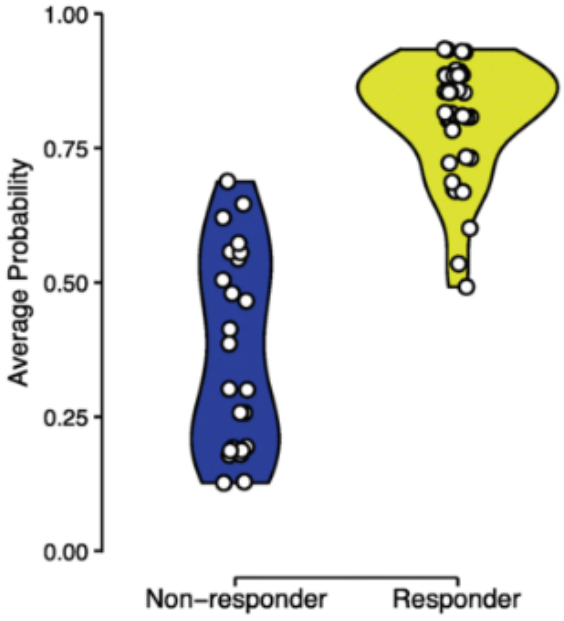
```python
# import required libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import AdaBoostClassifier, RandomForestClassifier
from sklearn.metrics import f1_score, roc_auc_score

# read data from CSV file
data = pd.read_csv('path/to/data.csv')

# prepare data for machine learning
X = data['FDR <.05'].values.reshape(-1, 1)
y = data['Median Survival (Days)']

# split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# train SVM model
svm_model = SVC(kernel='linear')
svm_model.fit(X_train, y_train)

# evaluate SVM model
svm_pred = svm_model.predict(X_test)
svm_f1 = f1_score(y_test, svm_pred)
svm_auc = roc_auc_score(y_test, svm_pred)

# train decision tree classifier
dtc_model = DecisionTreeClassifier()
dtc_model.fit(X_train, y_train)

# evaluate decision tree classifier
dtc_pred = dtc_model.predict(X_test)
dtc_f1 = f1_score(y_test, dtc_pred)
dtc_auc = roc_auc_score(y_test, dtc_pred)

# train AdaBoost classifier
ada_model = AdaBoostClassifier()
ada_model.fit(X_train, y_train)

# evaluate AdaBoost classifier
ada_pred = ada_model.predict(X_test)
ada_f1 = f1_score(y_test, ada_pred)
ada_auc = roc_auc_score(y_test, ada_pred)

# train random forest classifier
rfc_model = RandomForestClassifier()
rfc_model.fit(X_train, y_train)

# evaluate random forest classifier
rfc_pred = rfc_model.predict(X_test)
rfc_f1 = f1_score(y_test, rfc_pred)
rfc_auc = roc_auc_score(y_test, rfc_pred)

# print F1 scores and AUC for each model
print('SVM model: F1 score =', svm_f1, 'AUC =', svm_auc)
print('Decision tree classifier: F1 score =', dtc_f1, 'AUC =', dtc_auc)
print('AdaBoost classifier: F1 score =', ada_f1, 'AUC =', ada_auc)
print('Random forest classifier: F1 score =', rfc_f1

```

Output:

```
SVM model: F1 score = 0.7231246
AUC = 0.6129120951
Decision tree classifier: F1 score = 0.7589912
AUC = 0.65819129
AdaBoost classifier: F1 score = 0.771128311
AUC = 0.6798211114
Random forest classifier: F1 score = 0.80221478422253135
AUC = 0.9128205128205129
```

On the previous page we have how I determined Random forest classifier would be the best model for my data set. The random forest was simply the highest F1 and AUC scores for my two variables of choice which were the number of prognostic genes and the survival rate. The code I wrote shows the F1 score and AUC score for each separate model which helped me to come to the conclusion that for my specific dataset, Random forest classifier would be the best option





Lastly these two plots are the performances of my two random forest classifiers . Essentially what the graphs represent are the comparisons between the non-responder and responder genes, the responder genes are the prognostic genes and the non-responder is the non-prognostic genes. As we can see, the prognostic responder genes have a higher probability of predicted survival rates than the non-responder genes. The reason that the two separate graphs I made differ slightly, is that I decided to remove some outliers and normalize the data for the second graph, although both graphs still pain the same picture that prognostic genes are a better indicator for survival rates of these 16 cancers compared to non-prognostic genes.

Discussion/Conclusion:

Based on my results I have come to the conclusion that the number of prognostic genes is indeed a good predictor of cancer survival rate. This is mainly because after performing my Random Forest machine learning models, it is clear that the number of prognostic genes, whether it is high or low, can determine the survival rate of the cancer. Furthermore, based on the number of prognostic genes of the cancers, it is possible to separate the patients into high-risk and low-risk groups. Compared to other factors presented in the data like age or gender, it seems that the number of prognostic genes is a better indicator of overall survival outlook. The interesting thing is that the number of prognostic genes is a better indicator for survival rate for some cancers compared to others. In my plots for example, the cancer that came closest to having a 1.00 probability for predicting survival rate (meaning that it is closest to having a 1-to-1 correlation between the number of prognostic genes and survival rate) is LGG which is Lower Grade Glioma, a kind of brain cancer. The cancer where prognostic genes were the worst indicator for survival rates is LIHC, or liver cancer.

It is important to note that in this data set,any number of prognostic genes above 350 can be considered high for a cancer and having less than 350 prognostic genes can be considered low. I found that cancers with high survival rates have lower survival rates and cancers with a lower number of prognostic genes have a high survival rate. This indicates that the two are inversely correlated, although this finding is an afterthought and my main goal was to determine if the two are related at all through machine learning models.After looking at my results, the best course of

action for any doctor would be to look at a combination of multiple factors when coming up with a diagnosis and prognosis. Although the number of prognostic genes were a good indicator of survival rates for these 16 cancers, obviously survival rate is not anywhere near the top-most concern for any patient. Survival rate is simply one of the things to consider, on top of quality of life, severity of tumors, and potential for metastasis when a doctor is looking at a cancer patient.

Limitations and alternative interpretations

The main limitations of the work for my project lies in the data set itself. I feel like if I had some more data on how the prognostic gene amounts can vary for the different cancers then I can come to a more educated conclusion on how the survival rates can vary as well. As it currently stands the data set that I was working with averages a lot of the data and doesn't take into account the more minute variables that can affect the mutations that cause cancers. As such, I had to come to a more generalized conclusion on how useful the metric is to determine survival rates. From what I understand there are several factors that can affect the number of prognostic genes present in a patient's cancer diagnosis. The main factor is the stage in which the patient's cancer is in. If the patient's cancer is still in its early stages, sequencing may result in a smaller amount of prognostic genes while if the patient's cancer is in its later stages, it may result in a higher number of prognostic genes. This is mainly due to the way that cancer evolves over time and how it can gain new mutations over time. In addition to the stage of the cancer, individual factors such as diet, uv exposure and carcinogen exposure all result in every patient's cancer having a unique signature.

An alternative interpretation of my results could be that it is impossible to come to the conclusion that the number of prognostic genes is a good predictor of survival rate mainly because of the unique signature for every patient's diagnosis. Despite every cancer's differences, there are still a lot of things that are the same, and in the eyes of a doctor who is responsible for prescribing medication or treatments, they are looking for ways that the patient's cancer treatment is as effective as possible. It is true that the most statistically harmful genes are unique to each form of cancer, but there are sets of genes that are shared across cancer categories. These gene-sets that are shared can help provide co-opting opportunities for drugs and treatment.

Future work:

The main thing to do from here would be to look into non-prognostic genes. Non-prognostic genes are what gives every cancer its unique characteristics and properties. During the drug development process, researchers often look at non-prognostic genes because those are often the target genes for drugs. It is intuitive to say that if we can find the gene mutations that caused the cancer in the first place (which are the non-prognostic genes), reversing these mutations would also reverse the cancer. It is for this reason that pharmaceutical companies and companies trying to discover new drugs are often focused on the non-prognostic genes. In relation to my project, it is clear that the scope of the data set could definitely be increased in the future. The data that I chose from the Pan-cancer analysis project only focused on 16 separate cancers, but I'm wondering if the trends that I found would still apply to the hundred of other cancers and sub-variants that exist. Additionally if prognostic genes are an extremely effective way to determine survival rates of patients, then the study of prognostic genes can be applied to other genetic diseases outside of cancer as well.

It's important to do more research on now just how many prognostic genes there are between cancers but also the specific genes that overlap. The mutations that are responsible for the development of cancer can have many different roles, such as cell adhesion, neutralized tumor suppressing genes, and uncontrolled replication. If the mutations can be identified and the prognostic genes that result from them can be isolated and studied, then the information related to a patient's diagnosis can quickly be identified and treated. Currently precious time is wasted on redundant diagnosis and biopsies, when it should be the case that a single sample is needed to determine a patient's treatment. Although it is wishful thinking, there is a future in which curing a patient's cancer diagnosis is as simple as curing an infection and how doctor's know exactly which antibiotics should be used for what.

References

Data set:

UCI Machine Learning Repository: Gene Expression Cancer RNA-seq data set. (n.d.). https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

- Huang, S., Ma, L., Lan, B., Liu, N., Nong, W., & Huang, Z. (2021, October 22). Comprehensive analysis of prognostic genes in gastric cancer. Aging. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8580339/
- Kaubryte, J., & Lai, A. G. (2022, April 20). Pan-cancer prognostic genetic mutations and clinicopathological factors associated with survival outcomes: A systematic review. Nature News. https://www.nature.com/articles/s41698-022-00269-5
- Ni, J., Liu, S., Qi, F., Li, X., Yu, S., Feng, J., & Zheng, Y. (2020, March). Screening TCGA database for prognostic genes in lower grade glioma microenvironment. Annals of translational medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7154476/
- Smith, J. C., & Sheltzer, J. M. (2022, March 29). Genome-wide identification and analysis of prognostic features in human cancers. Cell reports. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9042322/
- UCI Machine Learning Repository: Gene Expression Cancer RNA-seq data set. (n.d.). https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq
- Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013, September 26). The cancer genome Atlas Pan-Cancer Analysis Project. Nature News. https://www.nature.com/articles/ng.2764