Module_3: *(Template)*

Team Members:

Jack Hancock and Shreesh Kalagi

Project Title:

Expression of p53 and Bcl-2 in the Apoptosis Pathway Across Cancers of Varying Rates of Metastasis

Project Goal:

This project seeks to identify any correlation between the expression of TP53 and BCL-2 and the pathologic spread of cancer from the primary tumor site in both Lung Adenocarcinoma and Thyroid Carcinoma.

Disease Background:

- Cancer hallmark focus:

Evasion of Apoptosis

- Overview of hallmark:

Apoptosis, or programmed cell death, is a protective process that removes damaged or unnecessary cells. In many cancers, this system becomes disrupted, allowing abnormal cells to survive and multiply. The evasion of apoptosis is one of cancer's defining hallmarks, as it enables tumors to grow unchecked and resist treatment. Cancer cells often achieve this by altering key regulatory genes such as p53 and Bcl-2, which normally balance cell survival and death

- Genes associated with hallmark to be studied (describe the role of each gene, signaling pathway, or gene set you are going to investigate):

p53: A tumor suppressor gene that detects DNA damage and triggers apoptosis when repair is not possible. Loss or mutation of p53, which occurs in over half of human cancers, prevents the cell from initiating this death response, leading to unchecked survival

Bcl-2: A gene family that regulates mitochondrial signals in apoptosis. Some members, like Bcl-2 and Bcl-XL, prevent cell death, while others, such as like Bax and Bak, promote it. Overexpression of Bcl-2 allows cancer cells to resist apoptosis, while reduced p53 activity further enhances this effect

- Prevalence & incidence Lung adenocarcinoma is among the leading causes of cancer-related deaths globally. With an estimated 230,000 new cases annually in the U.S. Thyroid carcinoma, while far less deadly, is the most common endocrine malignancy, affecting roughly 44,000 Americans each year.

- Risk factors (genetic, lifestyle) & Societal determinants LUAD is strongly linked to tobacco use, occupational carcinogen exposure, and air pollution. THCA risk

increases with prior radiation exposure and female sex. Socioeconomic status and access to screening play roles in detection and treatment outcomes for both cancers.

- Standard of care treatments (& reimbursement) LUAD treatment includes surgery, chemotherapy, radiation, and increasingly, targeted therapies like EGFR or ALK inhibitors. Despite these options, resistance due to apoptosis evasion remains a major problem. THCA treatment typically involves surgical resection followed by radioactive iodine therapy and thyroid hormone suppression, with generally excellent outcomes.
- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology) Biological Mechanisms

Lung Adenocarcinoma (LUAD): Lung adenocarcinoma arises from the epithelial cells of the distal airways and alveoli, where gas exchange occurs. These cells are normally responsible for producing surfactant and maintaining the thin barrier between air and blood. Chronic exposure to carcinogens such as tobacco smoke, pollution, or asbestos damages the bronchial and alveolar epithelium, leading to inflammation and abnormal cell repair. Over time, this causes metaplasia and dysplasia of glandular epithelial cells within the lung parenchyma. At the cellular level, cancerous transformation results in uncontrolled proliferation, loss of polarity, and invasion through the basement membrane into surrounding tissues. On a molecular level, abnormal activation of signaling pathways controlling cell division, angiogenesis, and DNA repair allows these cells to grow independently and metastasize through lymphatic and vascular routes to other organs such as the brain, liver, and bones.

Thyroid Carcinoma (THCA): Thyroid carcinoma originates from the follicular cells of the thyroid gland, located in the anterior neck. These cells normally absorb iodine from the bloodstream and synthesize thyroid hormones (T3 and T4), which regulate metabolism throughout the body. Structural or mutational damage to follicular cells disrupts normal thyroid follicle organization and hormone synthesis. The cancer often begins as a small nodule that can progress to invade nearby lymph nodes or tissues in more advanced stages. On a cellular level, transformed follicular cells lose regulated growth control but often retain some differentiation, allowing continued hormone production in early disease. Molecularly, alterations in intracellular signaling pathways that regulate growth and differentiation lead to increased proliferation and decreased apoptosis. Despite these changes, thyroid carcinoma typically maintains partial glandular function, which contributes to its slower progression and generally favorable prognosis.

Data-Set:

We were provided a large clincial dataset developed by The Cancer Genome Atlas which includes patient data collected over a decade. The patients in the data set expressed over 11,000 tumors across 33 variants of cancer. Genomic data was collected and summarized for each patient amongst several relavant genomic markers. In addition, patient metadata, which includes demographics, tumor staging, and patient history data is supplied adjuntively.

Our study focuses on the prevalance of two genes - TP53 AND BCL-2 - both of which are included in the TCGA genomic data sheet (the sheet includes genomic data produced by sequencing and molecular profiling platforms, normalized computationally). We chose to focus on the prevalance of these genes only in patients with either Lung Adenocarcinoma (LUAD) or

Thyroid Carcinoma (THCA) in order to determine any correlation between TP53/BCL-2 expression and rate of metastasis.

More information cocnering the dataset and collection protocol is included in the article "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics"

Data Analyis:

Methods

The machine learning technique I am using is: Principal Component Analysis, followed by Classification

In a Principal Component Analysis (PCA), we will reduce the dimensionality of our data in order to identify clusters of data that may inform us on the patterns in the dataset. Specifically, we will take data from TP53 and Bcl-2 gene expression and "unlabel" the data in order to create a scatter plot using PCA as the machine learning technique. Visualizing the data, if we notice a dilineation of the data into distinct clusters, we will then move forward with the Logistic Regression supervised machine learning model. Taking the same gene expression data, but this time labelled with certain cancer types, we can try to classify different gene expression levels into different forms of cancer by creating a decision boundary between the data clusters. **

Analysis

Our working version of code includes a mix of the code provided to us in class and code formatted from prompting with ChatGPT. The code starts with a PCA analysis to identify clustering and then later labels the dataset in order ot identify classification of gene expression as a correlate of cancer type between LUAD and THCA. I consulted ChatGPT in trying a few different methods of supervised learning models, including Logistic Regression and a Decision Tree. We decided to move forward with a Logistic Regression of the data in order to draw a decision boundary, which the code implements, ultimately producing a graph with a deicison boundary. We included code for a graph of the logistic regression and its efficacy as a decision boundary model as well.

```
# %%
from sklearn.linear_model import LogisticRegression
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, roc_auc_score
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import os

# %% --- LOAD DATA ---
```

```
data_folder = r"C:/Users/ugj3eb/Desktop/Third Sem/Computational
BME/Comp BME/Module 3"
expr_file = os.path.join(data_folder, "Gene_subset_data.csv")
meta_file = os.path.join(data_folder, "patient_cancer_type.csv")

expr = pd.read_csv(expr_file, index_col=0, low_memory=False)
meta = pd.read_csv(meta_file)

# Keep only LUAD and THCA
meta = meta[meta["cancer_type"].isin(["LUAD", "THCA"])]

# Align intersection of samples
common_samples = expr.columns.intersection(meta["sample"])
expr = expr[common_samples]
meta = meta[meta["sample"].isin(common_samples)]

# Transpose to samples x genes
X = expr.T
X = X.dropna(axis=1, how='all')
X.columns = X.columns.astype(str)

# Labels
y = meta.set_index("sample").loc[X.index, "cancer_type"]

print(f"Data shape: {X.shape}")
print(y.value_counts())

# %% --- IMPUTE MISSING VALUES ---
imputer = SimpleImputer(strategy='mean')
X_imputed = pd.DataFrame(imputer.fit_transform(X), index=X.index,
columns=X.columns)

# %% --- STANDARDIZE ---
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X_imputed)

# %% --- TRAIN/TEST SPLIT ---
X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.5, random_state=42, stratify=y
)

# %% --- PCA (fit ONLY on training data) ---
pca = PCA(n_components=2)
X_train_pca = pca.fit_transform(X_train)
X_test_pca = pca.transform(X_test)

# %% --- LOGISTIC REGRESSION TRAINED ONLY ON TRAIN DATA ---
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)
```

```
# Evaluate
y_pred = model.predict(X_test)
acc = accuracy_score(y_test, y_pred)

# Convert labels to numeric for AUC
y_test_num = (y_test == "LUAD").astype(int)
y_proba = model.predict_proba(X_test)[:, 1]

print(f"Test accuracy: {acc:.3f}")

# %% --- LOGISTIC REGRESSION IN PCA SPACE (for visualization only) ---
model_2d = LogisticRegression(max_iter=1000)
model_2d.fit(X_train_pca, y_train)

# Create meshgrid for decision boundary
x_min, x_max = X_train_pca[:, 0].min() - 1, X_train_pca[:, 0].max() +
1
y_min, y_max = X_train_pca[:, 1].min() - 1, X_train_pca[:, 1].max() +
1

xx, yy = np.meshgrid(np.linspace(x_min, x_max, 300),
                     np.linspace(y_min, y_max, 300))

Z = model_2d.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)

# %% --- PLOT TRAINING DATA DECISION BOUNDARY ---
plt.contourf(xx, yy, Z, levels=50, cmap="RdBu", alpha=0.6)
plt.contour(xx, yy, Z, levels=[0], colors='black', linewidths=2)
sns.scatterplot(x=X_train_pca[:, 0], y=X_train_pca[:, 1], hue=y_train,
                edgecolor='k', palette="Set1")
plt.title("Decision Boundary (Training Data, PCA space)")
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.show()

# %% --- PLOT TEST DATA DECISION BOUNDARY ---
plt.contourf(xx, yy, Z, levels=50, cmap="RdBu", alpha=0.6)
plt.contour(xx, yy, Z, levels=[0], colors='black', linewidths=2)
sns.scatterplot(x=X_test_pca[:, 0], y=X_test_pca[:, 1], hue=y_test,
                edgecolor='k', palette="Set1")
plt.title("Decision Boundary (Test Data, PCA space)")
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.show()

# %% --- PLOT: Predicted probability vs true label (Model fit
visualization) ---
```

```
plt.figure(figsize=(7,5))

# Sort by predicted probability for a clean curve
sorted_idx = np.argsort(y_proba)
probs_sorted = y_proba[sorted_idx]
truth_sorted = y_test_num.values[sorted_idx]

plt.plot(probs_sorted, label="Predicted probability of LUAD")
plt.scatter(range(len(truth_sorted)), truth_sorted,
            color="black", s=20, label="True label (0=THCA, 1=LUAD)")

plt.title("Logistic Regression Fit: Predicted Probability vs True
Labels")
plt.ylabel("Probability / True Label")
plt.xlabel("Sorted Samples")
plt.legend()
plt.tight_layout()
plt.show()
```

Verify and validate your analysis:

We decided to test our model by splitting our data randomly into two halves: "training" data and "testing" data. We trained our model using the model.fit method with half of the data, and then developed a prediction of what type of cancer (LUAD or THCA) other gene expression levels may be classified as. Then, we ran the "testing" half of the data to assess the accuracy of our model. Our "Test Accuracy," as measured by the rate of correct predictions, was calcuated to be 0.988 (highly accurate):

```
# %% --- TRAIN/TEST SPLIT ---
X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.5, random_state=42, stratify=y
)

# %% --- LOGISTIC REGRESSION TRAINED ONLY ON TRAIN DATA ---
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

# Evaluate
y_pred = model.predict(X_test)
acc = accuracy_score(y_test, y_pred)

# Convert labels to numeric for AUC
y_test_num = (y_test == "LUAD").astype(int)
y_proba = model.predict_proba(X_test)[:, 1]

print(f"Test accuracy: {acc:.3f}")
```

In order to further validate our results, we also looked at published litearture. In lung adenocarcinoma, the tumour suppressor gene TP53 is mutated in approximately 40-50% of

cases, making it one of the most frequently altered genes in this disease (https://pubmed.ncbi.nlm.nih.gov/37072703/). In an article by Nature, it is stated,"Mutations in TP53 are common in lung adenocarcinoma but rare in differentiated thyroid carcinoma: for example, TP53 mutations are described as frequent in lung cancers but occur at very low rates in differentiated thyroid tumours". (https://www.nature.com/articles/s41419-022-05408-1). Looking at BCL2, anti-apoptotic BCL2 expression tends to be highest in thyroid carcinomas while being among the lowest in lung adenocarcinomas, supporting the idea that thyroid carcinomas may rely less on pro-survival BCL2 overexpression for apoptosis evasion than lung adenocarcinomas

Conclusions and Ethical Implications:

The PCA plot and decision boundary developed by our code based on the testing dataset showed that there is a distinct clustering of datapoints with regards to gene expression of BCL-2 and TP53 (and associated genomic markers of apoptotic evasion) and cancer type between LUAD and THCA. The "Test Accuracy" metric, which relies on accuracy of prediction with the trained model, being 0.988 represents a good fit of our model to the test data. With regards to out project goal, this conclusion showed us that there is, in fact, a correlation between BCL-2/TP53 expression and cancer type, where LUAD and THCA are highly studied and have been observed to exhibit different rates of metastasis.

There are a number of ethical implications of our findings. For one, physicians may be able to use gene expression data, which may be less invansive and less expensive than taking a patient to the operating room, in order to classify cancer type. However, it is imperative that physicians do not rely solely on gene expression data in order to calssify cancer types, seeing as this method has not been sufficiently studied and is not a definitive method of cancer identification.

Another important ethical implication includes false identification of cancer based on gene expression alone. It is possible that patients have abnormal gene expression levels that may represent baseline metrics for them, but might be suggestive of cancer by our model. False identification of cancer poses risks of increase health insurance rates, emotional burden, and liability by scientists and providers for unecessary procedures. Thus, it is important to complete a full analysis of the patient, including any necessary screenings or scans, in order to ensure an accurate diagnosis.

Limitations and Future Work:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

NOTES FROM YOUR TEAM:

What cancer hallmark do we want to study?

- Evasion of apoptosis
  - Involves the cell signaling pathways induced by p53 supressor gene damage, as well as Bcl-2.

    - Is there a significant difference in quanity of these two genes in cancers which metastasize faster than others? (Is there a correlation between precesne of these genes and spread of cancer).

How do we want to analyze the data?

- If we want to identify "clusters" in the data between gene expression for different cancer types, we can do a Principal Component Analysis which will unlabel the data and plot the points. If we then want to re-label the data and run a supervised machine learning model (i.e. Logistic Regression), we can draw a decision boundary to help us see if there is an effective way to identify cancer type based on gene expression.

How do we test accuracy of our model?

- Consulting with ChatGPT: One way to test accuracy of a PCA model is to use the .fit method to create a model, and then the .predict method to predict how a piece of data would be classified. Then, running the testing dataset through the model, if we calculate a ration of correct predictions to total guesses, we can develop a test accuracy metric.

QUESTIONS FOR YOUR TA:

We have no questions for our TA at this time.