ECE/CS 498 DS Spring 2019 - Mini Project 2
Unsupervised Single-Cell Analysis in Triple-Negative Breast Cancer

Disclaimer

As was the case of MP1, the data and the analytical pipeline (which was based on a real experiment [1]), has been sufficiently modified for you to learn and practice important concepts in data science. As such, the results of the analysis you perform do not accurately represent the reality of biological research. However, if you wish to learn more, representative papers and tutorials [1,2,3,4,5] are provided in the References section.

Broad Problem Statement

Triple-negative breast cancer is an aggressive type of cancer that is poorly understood both in terms of its chemistry and its treatment. Some years ago, women with triple-negative breast cancer who also had diabetes and were taking an anti-diabetic drug, metformin, saw their tumors shrinking. Why their tumors shrank was unclear. In this project, we're going to investigate how metformin impacts genes that might be responsible for the spreading of triple-negative breast cancer.

You'll be exploring various new data science methods in solving this problem and are free to use Python packages unless otherwise specified (see Useful Python libraries and APIs section for help).

Concepts you will learn and apply

- Data pre-processing and visualization
- Bayesian Networks
- Statistical analysis (Kolmogorov–Smirnov test, Multiple testing, Q-Q plot)
- Dimensionality reduction (Principal Component Analysis, t-SNE)
- Clustering (K-Means, Gaussian Mixture Model clustering, Hierarchical clustering)

Biology Primer

The human genome is a sequence of approximately 3 billion nucleic acids, and can be thought of as a large string from a 4 letter alphabet {A,C,G,T}. A *gene* is a substring in the genome that contains a sequence of nucleotides which can be used to construct proteins. *Proteins* are the workhorses of a cell and are responsible for carrying out important biomolecular functions. A gene is said to be "expressed" in a particular cell if it is actively being used to create its corresponding protein. Expression can be quantitative: A highly expressed gene will produce a large amount of its corresponding protein, and a repressed gene will produce little to none.

These proteins are involved in pathways (series of actions among molecules in a cell that leads to a certain product or a change in the cell).

Single-cell RNA sequencing (scRNA-seq) is a recently developed technique to examine the expression level for each gene contained in each of the cells under study. For simplicity, in this project we will assume that we have already generated expression levels for every gene in each cell. Though we are glossing over some of the technical details in how expression data is measured, we will consider some cases of how the quality of scRNA-seq data is susceptible to external factors like temperature, experiment duration and amount of reagent consumed.

As an example dataset, we will consider expression data from tumor cells. Specifically, triple-negative breast cancer: A subtype of breast cancer that does not have any standard targeted therapies [6]. Population studies have shown that an anti-diabetic drug, metformin, inhibits the growth of cancer cells in triple-negative breast cancer. Since cancer-targeting drugs usually work by altering the expression of genes involved in crucial pathways (for example, replicate immortality related pathways), **the biologists wish to identify genes with significantly altered expression induced by metformin as well as the pathways associated with these genes.** Identification of these genes can help biologists design laboratory experiments to establish the molecular mechanism of metformin in triple-negative breast cancer, and ultimately help clinicians provide better care to the patients.

Can you help the biologists?

Data

The biologists have obtained approximately 400 triple-negative breast cancer cells, half of which are treated with metformin (referred to as *metformin-treated cells*) and half that are not (referred to as *baseline cells*).The cells were sequenced using scRNA-seq, and the resulting data reflect ~20,000 genes and their associated gene expressions for baseline and metformin-treated cells.

Files

GeneExpression_Baseline.csv: Gene expression matrix for baseline cells.

Chromosome	Gene ID	…	Coding Length	Cell 1	…	Cell N
Gene 1 information*				104*	…	78
…				…	…	…
Gene M information				92	…	84

*expression level of Gene 1 in Cell 1 is 104.

GeneExpression_Metformin.csv: Gene expression matrix for metformin-treated cells, similar to **GeneExpression_Baseline.csv.**

Note: Gene ID is a unique identifier of genes. A given gene ID will appear and only appear once in **GeneExpression_Baseline.csv** and in **GeneExpression_Metformin.csv.** Furthermore, genes are sorted according to their positions in the genome in both files.
(Hint: Think about what index you want to use for your dataframe.)

QualityControl.csv: Experiment conditions and corresponding quality records collected from previous scRNA-seq experiments.

Temperature	Time	Viability	Quality
cool*	short*	high*	good*
…	…	…	…
cool	long	low	bad

*A scRNA-seq experiment conducted under cool temperature, in a short period of time duration led to high cell viability and good data.

BayesInferenceBase.csv: Experiment conditions collected for baseline cells in **GeneExpression_Baseline.csv.**

Cell Name	Time	Temperature
Cell 1*	short*	cool*
…	…	…
Cell N	long	warm

*BaselineCell 1 went through scRNA-seq under cool temperature, in a short period of time duration.

BayesInferenceMetf.csv: Experiment conditions collected for baseline cells in **GeneExpression_Metformin.csv**, similar to **BayesInferenceBase.csv.**

PathwayDictionary.txt

*2-arachidonoylglycerol biosynthesis:ABI3,...,ZNF658B

…
Xanthine and guanine salvage pathway:ACE2,...,ZYG11A

*Gene ‘ABI3’, …, and ‘ZNF658B’ are involved in pathway “2-arachidonoylglycerol biosynthesis”.

Useful Python libraries and APIs

- numpy
- pandas
- matplotlib.pyplot
- seaborn.heatmap
- scipy.stats.ks_2samp
- sklearn.decomposition.PCA
- sklearn.manifold.TSNE
- sklearn.cluster.KMeans
- sklearn.mixture.GaussianMixture
- sklearn.cluster.AgglomerativeClustering

References

1. Athreya, Arjun P., et al. "Unsupervised single-cell analysis in triple-negative breast cancer: A case study." *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 2016.
2. Hunter, Lawrence. "Molecular biology for computer scientists." *Artificial intelligence and molecular biology* (1993): 1-46.
3. http://hemberg-lab.github.io/scRNA.seq.course/
4. http://data-science-sequencing.github.io/
5. http://genomicsclass.github.io/book/
6. https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies/targeted-therapies-fact-sheet
7. Hautaniemi, Sampsa, et al. "A novel strategy for microarray quality control using Bayesian networks." *Bioinformatics* 19.16 (2003): 2031-2038.

Task 0: Getting Started

Answer the following questions based on **GeneExpression_Baseline.csv**:

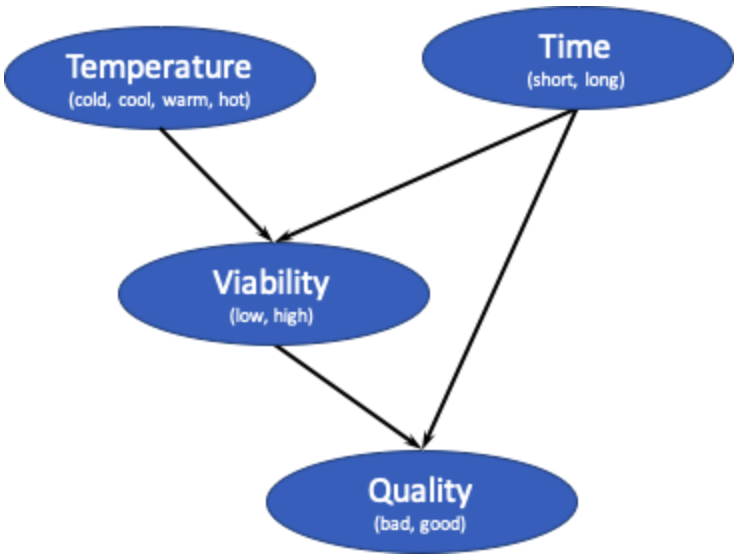
1. In the context of statistical analysis, why do biologists need multiple cells to identify genes with significantly altered expression?
2. How many cells were sequenced?
3. How many genes were sequenced?
4. Are genes equally distributed across chromosomes?
5. Plot the distribution of coding length.

Task 1: Data Cleaning and Visual Inspection

1. Bayesian Network for Quality Control

Not all cells used in the experiment yield high-quality data. Identifying and removing potentially problematic cells is critical for downstream analysis. In scRNA-seq, factors such as machine failures, batch effects might impact the quality of the data. For the purpose of this project, we will limit the factors to the ones described below.

If one observes that data quality is related to temperature, time, etc., one can identify the poor quality cells using a Bayesian approach. Consider the following Bayesian Network for quality control:



a. Give the factorization of the joint probability distribution.
b. Count the number of parameters needed to define the conditional probability distribution of the Bayesian Network for quality control.

c. Show the conditional probability tables P(Quality|Viability, Time), P(Viability|Temperature, Time), P(Temperature), P(Time) for the above network. Training data is provided in **QualityControl.csv**.
d. Calculate P(Quality|Temperature, Time) for all possible values of Quality, Temperature and Time. Show your calculation.
e. Use the calculated conditional probabilities and the collected data **BayesInferenceBase.csv, BayesInferenceMetf.csv** to determine the quality of the sequenced cells given Temperature and Time during the experiments. Report bad quality cells. Drop bad quality cells for the following analyses.

In the real world, poor quality cell filtering is usually done through other approaches [3]. Nonetheless, Bayesian Networks have been proven to be a reliable and useful model for quality control in other domains [7].

2. Data Standardization

A well-known bias in scRNA-seq experiments is the amount of reagent consumed. For each cell, the expression level from all genes are scaled by the same constant corresponding to the amount of reagent consumed for that cell. The constant varies across cells. Therefore, to compare the expression level across cells, we normalize the data.

a. Calculate the sum of expression level (summing up all genes) of each cell. Plot the histogram of the sums. You're expected to plot two histograms - one for baseline cells, and one for metformin-treated cells. Briefly summarize your observations.
b. Assume the sum of the expression level of all genes should be the same across cells. Normalize the gene expression matrices to get rid of the bias induced by the amount of reagent consumed so that the normalized sum of expression level of a cell equals to 1. Plot the histograms specified in Task 1.2.a. with the normalized data. (Hint: All cells will fall into one bin in the histogram.)

Use the normalized gene expression matrices for the following analyses.

Other normalization techniques include subtracting the mean, taking the z-score, etc. One should choose the normalize technique carefully based on one's understanding of the domain and the nature of the analyses.

3. Visual Inspection

a. Heatmap is a visual representation where individual values contained in a matrix are represented as colors. Plot heatmaps of the gene expression matrices. You're expected to plot two heatmaps - one for baseline cells, and one for metformin-treated cells. The heatmaps should have genes as rows and cells as columns. Briefly summarize your observations. (Hint: Make use of the *heatmap* API in the *seaborn* package; save your plot to a local file because plotting in Jupyter Notebook is sometimes inaccurate.)

b. Plot the distribution of gene expression across all cells for KCND2, TMEM239, and LINC00336, separately. Baseline cells and metformin-treated cells should be considered separately. (In total you're expected to plot 6 distributions which can be plotted in the same graph if you choose.) Briefly summarize your observations.
c. Plot the distribution of gene expression across all genes for BaselineCell_1, BaselineCell_2, MetforminCell_3, and MetforminCell_6, separately. (In total you're expected to draw 4 distributions which can be plotted in the same graph if you choose.) By visual inspection, do these distributions fit a Gaussian distribution?

Task 2: Statistical Analysis

Recall that the biologists wish to identify genes with significantly altered expression induced by metformin. A gene's expression is declared **altered** if the difference observed in its expression level between baseline cells and metformin-treated cells is statistically significant.

1. Kolmogorov–Smirnov (KS) Test

a. Is KS test a parametric test or a non-parametric test? When does one want to use non-parametric tests?
b. For each gene, find the p-value of a two-sample KS test on its expression across baseline cells vs. metformin-treated cells. In total you're expected to find ~20,000 p-values. (Hint: Make use of the *stats.ks_2samp* API in the *scipy* package.)
c. What is the null hypothesis of the KS test in our context? Take one gene as an example.
d. Count the number of genes with significantly altered expression at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Summarize your answers in a table.

2. Multiple Testing (Hint: Take a look at Lecture 15 and 17 in [4] for an introduction on multiple testing.)

a. P-value of 0.05 is generally considered a good threshold for significant discovery. What does a p-value of 0.05 represent in our context?
b. Based on the definition of p-value, if the null hypothesis is true, what distribution will the p-values follow? (Hint: Google the definition of p-value.)
c. If no gene's expression was altered, how many significant p-values does one expect to see at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Compare your answers with your results in Task 2.1.d. Show the comparison in a table.
d. Q-Q (quantile-quantile) plot is used to compare two probability distributions by plotting their quantiles against each other. Say you've performed N KS tests in Task 2.1.b. Following the procedure below, plot a Q-Q plot to compare the distribution of p-values of your statistical tests (Task 2.1.b., referred to as observed p-values) with the distribution of p-values when the null hypothesis is true (Task 2.2.b.):
    i. Sample N p-values from the expected distribution in Task 2.2.b (referred to as expected p-values).

ii. Take the -log10() of observed p-values and expected p-values.
iii. Rank observed p-values and expected p-values in ascending order separately.
iv. Take the pair of smallest p-values (one from observed p-values, one from expected p-values) and plot a point on an x-y plot with the observed p-value on the Y-axis and the expected p-value on the X-axis.
v. Repeat (iv) for the next smallest pair, for the next smallest, and so on until you have plotted all N pairs in order.
vi. Add the x=y line to your plot.

e. Answer the following questions:
   i. How does taking the -log10() of the p-values help you visualize the "tail" of the p-value distribution?
   ii. What can you conclude from the Q-Q plot? (Hint: Think about what it means if the Q-Q plot approximately aligns with the x=y line and what it implies about the null hypothesis.)

Task 3: Dimensionality Reduction and Clustering

In this task, you will apply clustering techniques to find out subpopulations of cells.

The results in Task 2.2.e. is related to the heterogeneity of cells. For example, the ~200 baseline cells might comprise multiple subpopulations and the response to metformin might differ between or even within such subpopulations (after all, cell responses are stochastic processes.) Which is to say, metformin might have only induced alterations in the expression levels of crucial genes in a subpopulation of cells instead of all cells. **Thus, it is essential to first identify such subpopulations before running statistical tests.**

Identifying subpopulations based on cells' gene expression profiles is essentially an unsupervised clustering problem. That is to identify clusters of cells based on the similarities of their gene expression profiles. This is a difficult problem because: 1) The number of clusters is not known a priori, 2) There is usually high level of noise in the data (both technical and biological), and 3) The number of dimensions (i.e. genes) is large.

When working with high-dimensional datasets such as gene expression matrices, it can often be beneficial to apply some sort of dimensionality reduction method. Projecting the data onto a lower-dimensional subspace could substantially reduce the amount of noise. An additional benefit is that it is typically much easier to visualize the data in a 2 or 3-dimensional subspace. Therefore, we will be performing Principal Component Analysis to reduce the dimensionalities of the gene expression matrices.

1. Principal Component Analysis (PCA)

The easiest way to visualize the data is by transforming it using PCA and then visualizing the first two principal components. **Note:** PCA, plotting, and calculation should be done separately for baseline cells and metformin-treated cells.

a. Treating cells as samples and genes as features (dimensions), perform PCA on the gene expression data. (Hint: make use of the *decomposition.PCA* API in the *sklearn* package. Select "full" for *svd_solver*.)
b. Order the principal components by decreasing contribution to total variance. Plot a scree plot to show the fraction of total variance in the data as explained by each principal component. How many principal components are needed in order to explain 30% of the total variance?
c. Plot a scatter plot of the gene expression with the first two components. Briefly summarize your observations.

2. t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a popular dimensionality reduction technique for visualization. t-SNE models each high-dimensional object by a two- or three-dimensional point such that similar projects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability. We are not expecting you to know t-SNE in detail, but it is useful to know how to run t-SNE with a package and interpret the results.

a. Treating cells as samples and genes as features (dimensions), perform t-SNE to visualize the gene expression data in 2D. (Hint: make use of the *manifold.TSNE* API in the *scikit-learn* package) **t-SNE and plotting should be done for baseline cells and metformin-treated cells separately.** Briefly summarize your observations.
b. Compare the results of PCA and t-SNE.

3. Clustering

We now performing clustering to identify the subpopulations of cells. For the purpose of this project, **clustering should be performed after projecting the data into 2D using PCA** (Task 3.1.a.). **Note**: In reality, various factors go into determining the number of dimensions to retain after PCA.

Baseline cells and metformin-treated cells should be clustered separately. Visualize your results by plotting a 2D scatter plot just like you did in Task 3.1.c., but with each point colored by the clusters they belong to (use different colors for different clusters). For all clustering algorithms below, you will need to decide the optimal number of clusters by yourself and reason about it. Provide numbers, tables and/or graphs to support your reasoning.

a. K-Means Clustering

b. Gaussian Mixture Model Clustering (Hint: You might run into numerical issues if you use the GMM implementation in *scikit-learn*. You might want to try scaling up the data 100 or 1000 times if your results look weird.)
c. Single Linkage Hierarchical Clustering
d. Compare your results for different clustering methods and interpret them. Pick the results from your favorite one for the following analyses.

Task 4: Interpret Results

In this task, you are going to identify the set of genes with significantly altered expression and their associated pathways.

1. Identify altered genes

For each subpopulation (cluster) M_i in the metformin-treated cells, there are two possibilities:
1) The cells in M_i are not affected by metformin. Thus, they have the same gene expression profile as one of the subpopulations in the baseline cells. Consequently, M_i will be identical to one of the baseline subpopulations, while being distinctively different from the other subpopulations.
2) For cells in M_i, the expression levels of ~50-400 genes were altered by metformin. Thus, they have similar gene expression profiles as one of the subpopulations in the baseline cells, but not entirely the same. Consequently, M_i will be more similar to one of the baseline subpopulations compared to the other subpopulations, but not identical to any of them.

Based on the above information, answer the following questions:

a. For each metformin-treated subpopulations, determine whether or not it is affected by metformin. Show and explain your decision process in detail. Provide numbers, tables and/or graphs where necessary.
b. For each affected subpopulations M_i, identify the baseline subpopulation that M_i is most similar to. Show and explain your decision process in detail. Provide numbers, tables and/or graphs where necessary.
c. Identify genes with significantly altered expression by comparing each affected subpopulation with its corresponding baseline subpopulation. Use KS test with alpha level=0.0000025. This alpha level was chosen to account for multiple testing caveats implied in Task 2.

2. Gene set characterization

Pathway analysis is one technique which is a helpful tool to characterize sets of genes. To give you a flavor of how biological discoveries can be made from your analysis above, you will perform a “baby” version of pathway analysis with synthesized data instead of going through the actual software used for such analysis. Characterize the composition of your identified gene set by answering the following question:

a. What are the most common pathways these genes are associated with? Make use of **PathwayDictionary.txt**.
b. Are there any novel genes in your identified gene set ,i.e., genes that have not been implied in any pathways contained in **PathwayDictionary.txt**?

Deadlines

Checkpoint #	Deadline	Tasks	Requirements
1	2/25 11:59:59PM	Task 0	.ipynb with Task 0 completed
2	3/8 11:59:59PM	Task 1, 2	.ipynb with Task 0, 1 and 2 completed
3	3/27 11:59:59PM	Task 3, 4	● .ipynb with all tasks completed ● A Powerpoint presentation (.pdf)

Grading (Total 90 points)

- Task 0 - 6 points
- Task 1 - 17 points
- Task 2 - 15 points
- Task 3 - 20 points
- Task 4 - 14 points
- ipynb formatting - 9 points
- Powerpoint - 9 points

Submission Requirements

Please provide a single .ipynb file. Please label each section, task, and subsection accordingly.

- Write your names and NetIDs of group members in the beginning
- Explain all your work (include the code with comments)
- Write down the equations that are being used (for partial credit)
- All the charts should be appropriately formatted by showing the legend, axes labels, and chart title
- Each question answered should include the code you used to achieve the needed charts and/or tables and an explanation/interpretation

Similar to MP1, please also prepare a powerpoint with your results.
Recommendation on slides: one slide each for Task 0, 1.2, 1.3, 2.1, 2.2, 3.1-2, 4.2, two slides each for Task 1.1, 3.3, 4.1, and one slide specifying the contribution of each group member.

All submissions are done on Compass2G. Please don't zip your files for checkpoint 3. Upload them as separate files. Late submission policy is applicable. One submission per group.