76558 | CBIO | Hackathon

Guy Lutsker 207029448, Nivi Shenker 207227687, Eran Eben Chaime 308240597, Or Amar 311166169

1 Introduction

DNA methylation is an epigenetic mechanism in which a methyl group is added to cytosine. This mechanism causes silencing of gene expression and has an important role in multiple biological processes such as gene transcription regulation, aging, differentiation of stem cells, genomic imprinting and more. It has been shown that the patterns of DNA methylation are non-random, well regulated and tissue-specific **(link)**. These conclusions are consistent with the biological importance of DNA methylation. Given this notion, it is reasonable to think that the methylation status of close CpG sites is co-dependent. Indeed, several studies have demonstrated that there is a correlation in methylation status between neighboring CpG sites in healthy tissues **(link)**. In oncogenic tissues, a different scenario holds in the sense of tissue methylation sites status. On one hand, cancer silences tumor suppressor genes by methylating their promoter regions where on the other hand, hypomethylation has been recognized as a cause of oncogenesis **(link)**. This relationship raises a question concerning the co-methylation patterns in oncogenic tissues. It has been shown that the methylation status in these tissues was aberrant, at least for CpG islands **(link) (link2 - needs to be checked)**. In our work, we wish to delve into the data and get a better understanding of the phenomena of co-methylation between neighboring sites. Moreover, we aimed to see if co-methylation holds in tumorigenic tissues.

2 Research Questions

Is there a link between physical distance of methylation sites and the correlative distance? If such a link exists, does it hold for cancer tissues as well?

3 Preprocessing

Our data consists of 24 files of methylation patterns in 12 different tissues in both healthy and sick (Cancer) individuals. The first problem we encountered is the missing data in the files - many files had nan values. To handle this problem we chose to plot the percentage of missing values over the number of features (~450K methylation sites) to make an educated decision:

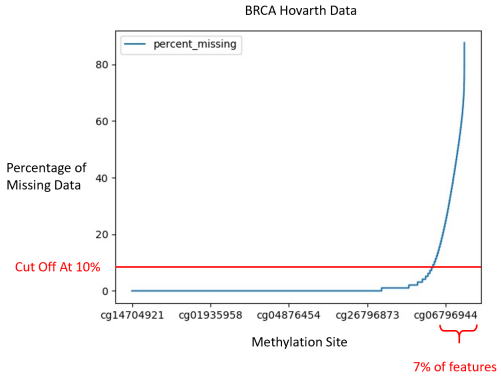

Figure 1: Example of missing data percentage in BRCA file

As we can see the features with most missing values do not take up much of the data itself and so we can intelligently remove some of the data and still keep most of it. In order to retain as much features as possible, while still getting rid of the "bad" features we chose to remove features with more than 10% nan values.

As we can see this method allowed up to retain 93% percent of the data, and now we have more "quality" data. But we still have the same problem! we still have missing values, and so data analysis is possible with missing values like this. There are many ways around this problem, and it is commonly called in statistics as data imputation. Common solutions involve replacing the missing values with the mean or the median. After some research on this question online **(link to paper)**, we decided that the best solution was to use KNN imputation. Our Intuition behind this step is that we assume that samples which are close in high dimensional space (~ 450K) are similar, and so we can use a weighted sum of the 3 most nearest neighbors to fill the missing values. After this step we have our preprocessed data!

4 Data Analysis &Visualization

Now that we have our data ready or work, its time to start working on it. But the data is incredibly sparse, with a relatively small amount of samples - for example in healthy tissue we have 538 people with ~450K features. To get a feel for the data we are working with, we ventured a bit into data science to help us understand what is going here.

4.1 Principle Component Analysis (PCA)

The first technique we chose to apply to our data is a linear dimensionality reduction named PCA. PCA is an unsupervised learning method which tries to perserve linear relationships in the data and is commonly used as a first step in analysis. Here we tried to see if PCA will be able to distinguish between tissues in a combined file of all the healthy tissues combined:

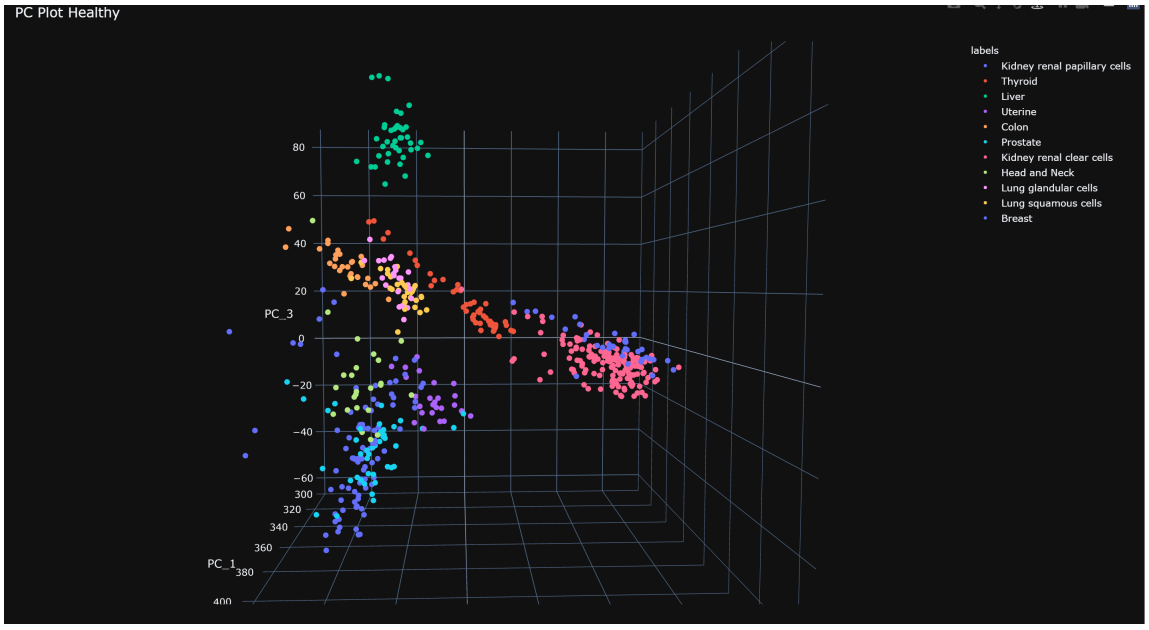

Figure 2: PCA of Healthy Tissue Combined

As we can see this unsupervised method was capable of distinguishing different tissues! Side note: the separation is even more pronounced in 3D : https://www.cs.huji.ac.il/~guy_lutsker/PCA_Plot_Healthy.html . This is very reasuring as it means that even though the data is very sparse, there is yet a strong signal in the data.

4.2 Uniform Manifold Approximation and Projection (UMAP)

The next step we took is to analyze the data using a non-linear approach. UMAP is non-linear dimensionality reduction method which is capable of perserving both global & local structures in data. This method could prove

useful to us since the manifold of our data lies in a very high dimension and so a method as robust as this could help us under stand the data better. UMAP analysis heavily relies on a neighbors graph generation, and we can learn quite a bit about the manifolds structure from the neighbors connection:

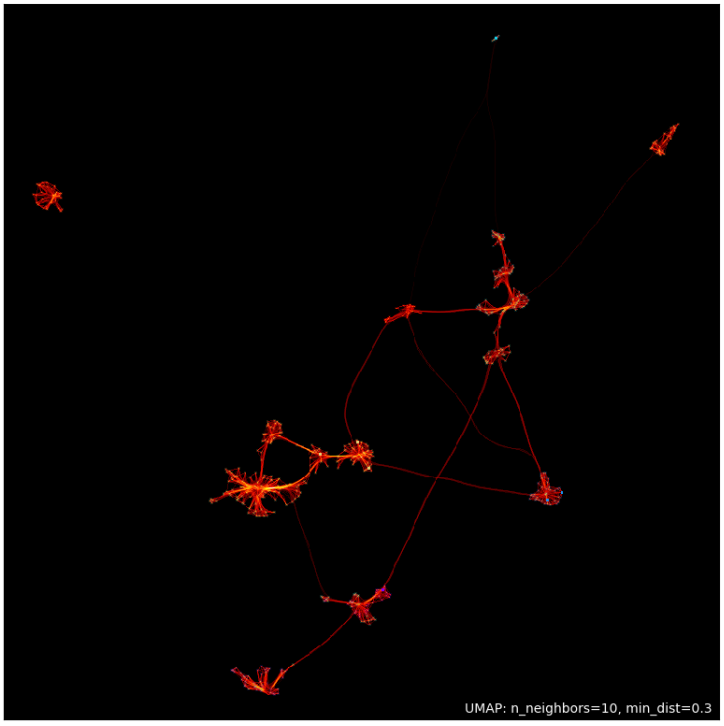

Figure 3: Neighbors graph on UMAP on Healthy Tissue Combined

And the UMAP graph itself:

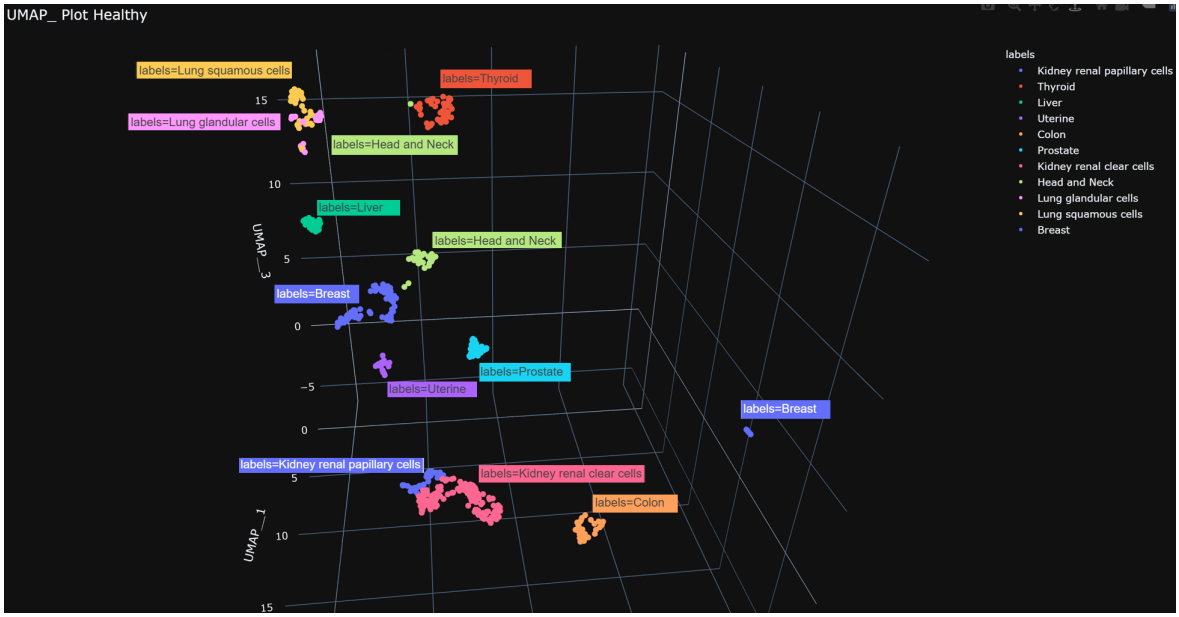

Figure 4: UMAP of Healthy Tissue Combined

Here the results are really good! UMAP with its local & global manifold preserving magic was able to sepa-

rate the tissues almost perfectly. In addition, it has also been able to capture the similarity of tissues, for example we can that both Kidney renal paplilary cells & Kidney renal are 2 clusters "smashed" together, while still being seperated into 2 distinct clusters. The same is true for Lung data, and all of these phenomena are more pleasing in 3D: https://www.cs.huji.ac.il/~guy_lutsker/UMAP_Plot_Healthy.html

4.2.1 Within Tissue

- Removing features with high percentage of missing data
- Use a data completion method (such as KNN)

4.2.2 Between Tissues

- Merging the data from several tissues
- Removing features with high percentage of missing data
- Use a data completion method (such as KNN)

4.3 Visualization

- Visualize the data to have a better interpretation of what we are working with, using dimensionality reduction techniques such as PCA, UMAP, and diffusion maps
- Cluster the data to see if we can recognize some intrinsic properties in the manifold of the high dimensional data, with techniques such as K-means, HDBSCAN or soft clustering methods such as NMF, LDA

4.4 Connection Between "Correlative Distance" & Physical Distance

- Calculating physical distance between methylation sites (in base pairs)
- Calculating "correlative distance" based on Pearson correlation
- Finding if we have a connection between "correlative distance" and physical distance

4.5 Connection Between Healthy Tissue & Cancer Tissue

- Same pipeline in section 3.3 will serve us for this section as well
- Find if the same correlations preserve in cancer tissues compared to healthy tissues

5 Model

Denoting k as the number of samples (people) and n as the number of features (methylation sites), we can write our data as follows:

$\{x\}_{i=1}^{k} \in \mathbb{R}^n$ e.g x_5^2 is the second methylation site of the fifth person in our data. In order to find the relation between physical distance of two sites in the genome (given in bp) and "correlative distance" we need to normalized the physical distance. To do so, we chose to use "Min-Max Normalization", this normalization preserves linear relationships in the data. We denote the normalized value of the physical distance with l_n. In order to calculate the "correlative distance" we use the Pearson Correlation coefficient denoted with $Corr$ and is given by the formula:

$Corr(x^i, x^j) = \frac{\sum_t (x_t^i - \bar{x^i}) \cdot (x_t^j - \bar{x^j})}{\sqrt{\sum_t (x_t^i - \bar{x^i})^2 \cdot (x_t^j - \bar{x^j})^2}}$. Notice that both l_n and $Corr$ have unit-less outputs and therefore we are allowed to compare them. We expect to find that two closely located sites will have similar methylation patterns and vice-versa. Therefore, we have constructed the following formula:

$$f(x^i, x^j) = \frac{1}{|Corr(x^i, x^j,)| - l_n(x^i, x^j)}$$

Our first intuition was to subtract the physical distance from the correlation in order to infer where the genome is "compacted" or "spreaded". However, we understand that both high positive and high negative correlation might

have some important biological insight (e.g both might suggest regulatory regions). Therefore, we have decided to take the absolute value of the correlation. Our last step was to invert the result because we expect to find the most interesting findings in this way. In order to preserve the information about the sign of the correlation, in the final step of the analysis we will use the following formula which incorporates the sign of the correlation:

$$g(x^i, x^j) = sign(Corr(x^i, x^j)) \cdot f(x^i, x^j)$$

6 Computational Tools

- KNN for missing data
- PCA for data visualization
- UMAP for data visualization
- Clustering algorithms such as K-means, HDBSCAN or soft clustering methods such as NMF, LDA
- Permutation Tests