76558 | CBIO | Hackathon

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

1 Motivation

Given a data consisting of the methylation sites in the DNA, it might be interesting to check for correlation between different methylation sites. An interesting question might arise concerning the meaning of these correlations, and one could think of many interpretations of such an analysis. In our work we will investigate the relation between physical distance of two methylation sites in the genome and the correlation of those sites.

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 Work Plan

Our work plan includes 4 chapters:

3.1 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:

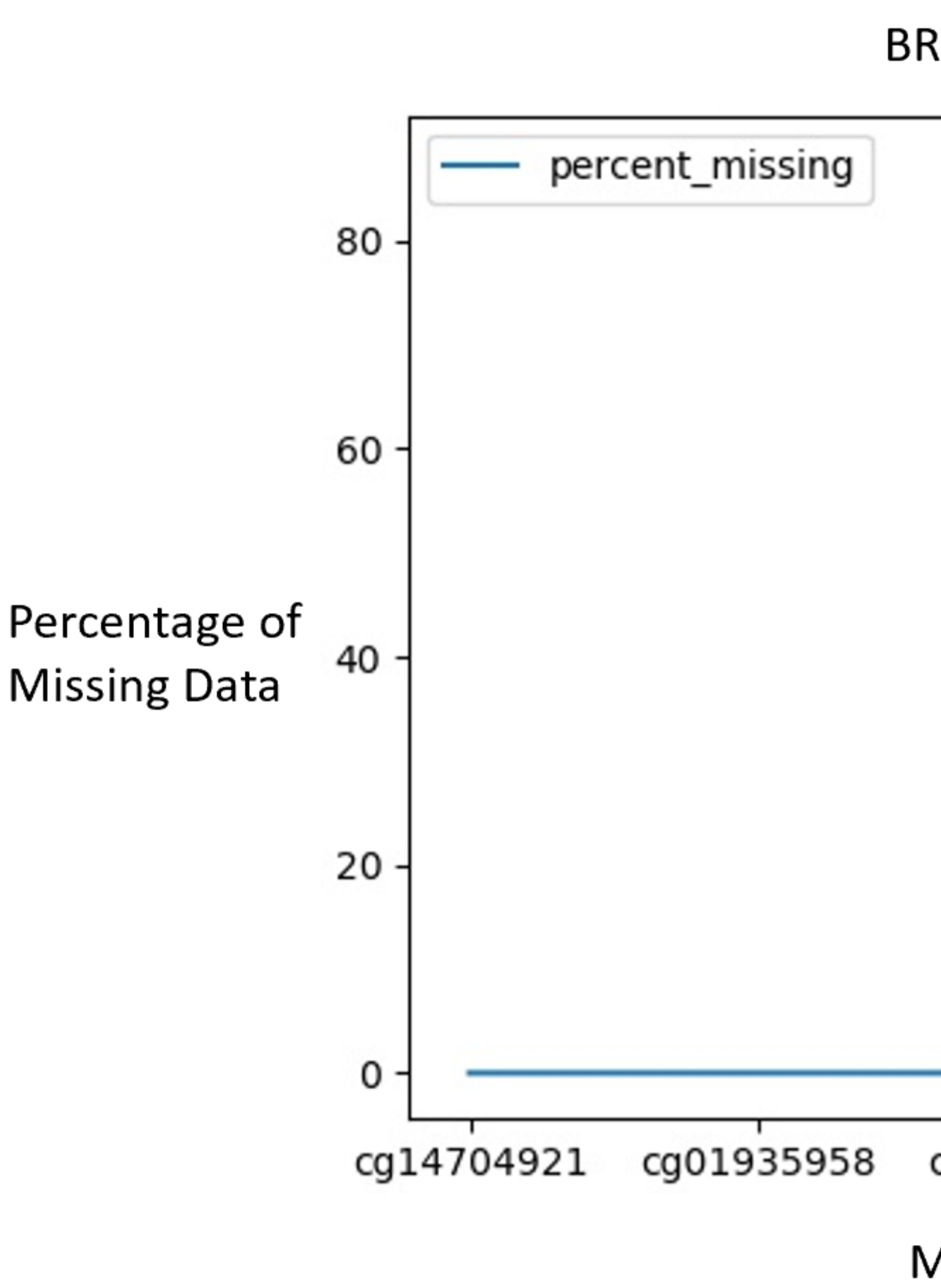
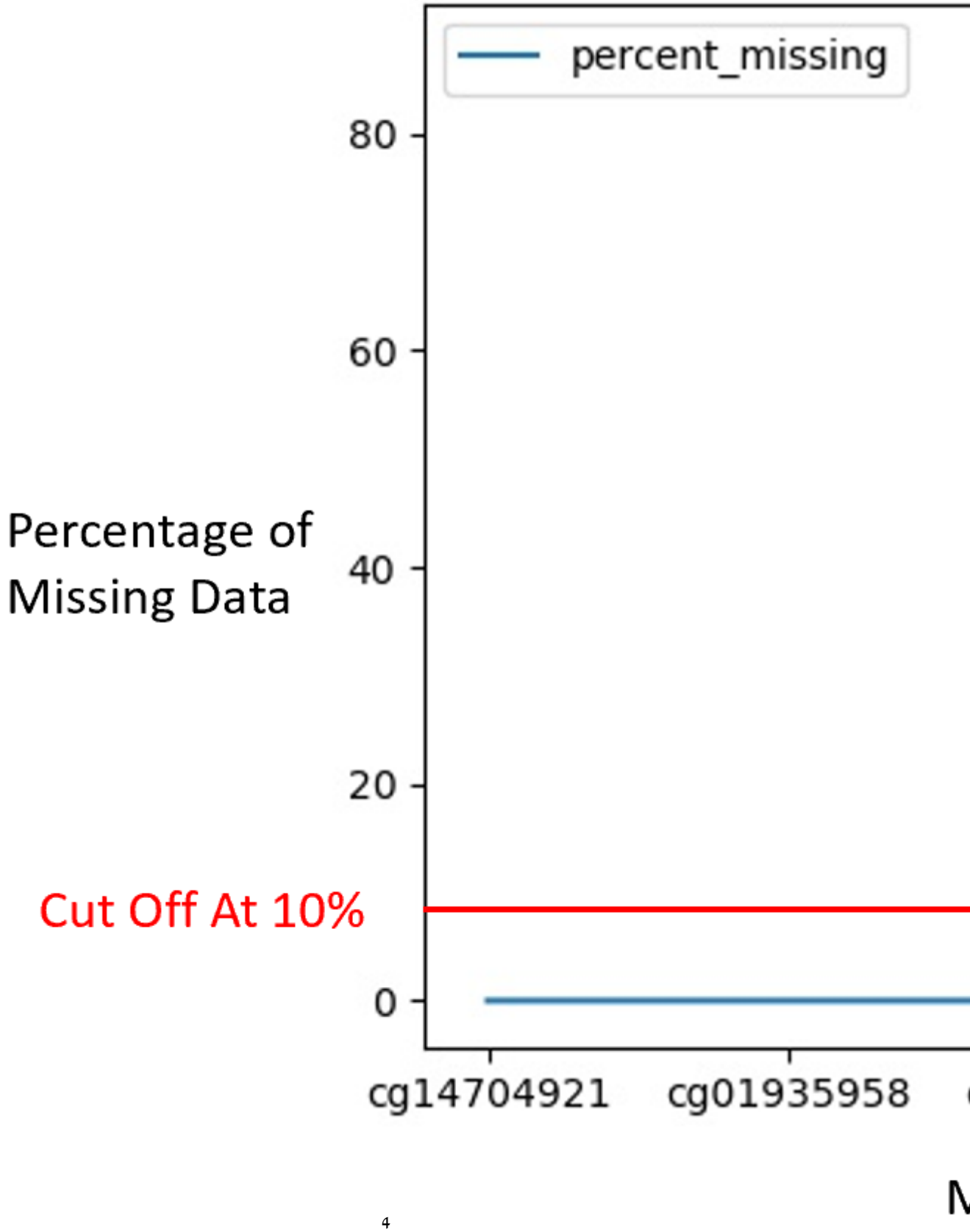


Figure 1: Example of missing data percentage in BRCA file

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:

BR


percent_missing
80
60
40
20
0
Percentage of
Missing Data
Cut Off At 10%
cg14704921
cg01935958

M

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 reserach on this question online **(link to paper)**, we decided that the best solution was to use

3.1.1 Within Tissue

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

3.1.2 Between Tissues

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

3.2 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

3.3 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

3.4 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

4 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:

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

5 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