Data report on the methylization of cell-free DNA

Daniel, Thomas & Hans-Henrik

March 8, 2024

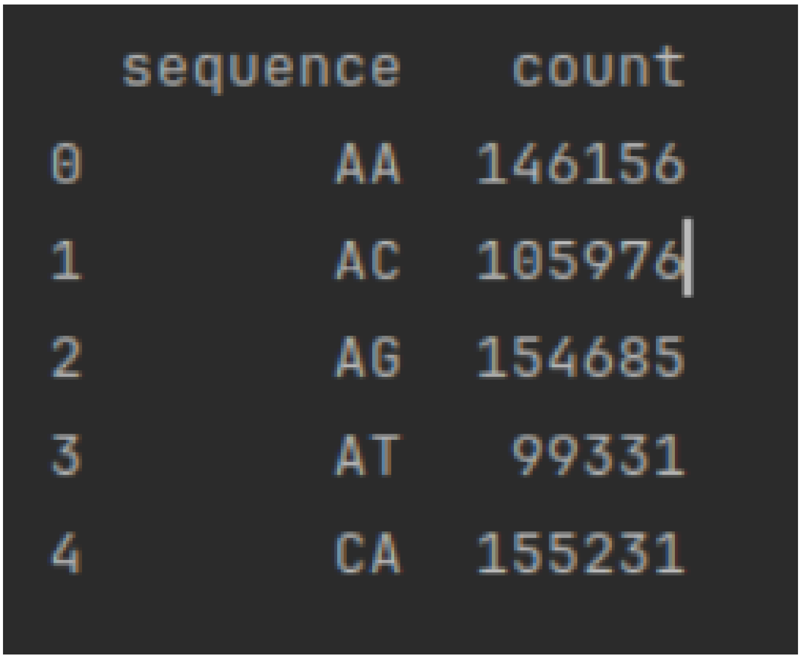
Contents

1 Introduction

In this project we will be looking at attempting to implement a model to predict, based on data from blood analysis, whether a given person potentially has cancer. The method in which we will be doing this, is by looking at methylized/unmethylized cell-free dna in peoples blood, specifically the fragmentation patterns. This is preferable to other methods, since it is both less invasive, and less expensive.

The data we will be looking at in this project, has been provided to us by Søren Bessenbacher, and contains information regarding fragmentation patterns for approximately 230 people diagnosed with cancer, and 230 people as a control group.

2 The data



```
  sequence   count
0       AA  146156
1       AC  105976
2       AG  154685
3       AT   99331
4       CA  155231
```

Figure 1: An example of the data

In the first column we have the k-mer, in this specific figure it is a 2-mer, and in the second column we have the count of that specific sequence. In addition to the control and the test samples, we also have a background file, which details the total number of combinations of sequences.

3 Normalizing the data

The raw data with just the counts, would likely not work, thus we want to normalize the data. The way in which we do this is by taking the sum of all the counts in a given file, and then dividing each count cell by this sum, thus giving us a ratio of the data. The way in which we've programmed this can be found in the *scripts/* folder of this repository, called *normalize_data.sh*. This script does as previosly mentioned and writes the ratios into a new file, which can be found in the folder *processed_data/normalized_data*. We then combine all these matrices into a single matrix with a R-script, which can also be found in the *scripts/*. The new file can be found in the the *processed_data/combined_data/*.

3.1 Normalizing with the background

In order to see the ratios of the samples compared to the potential ratios of the region in question, one can also normalize with the background file found in each k-mer folder. The way in which we approached this, was as previous by taking the sum of each count column, and then in addition we divide each cell in each sample with their respective background ratio. We did this with the *normalize_data_with_background.sh* file, which can be found in the *scripts/* folder.

3.2 PCA

To make the data approachable, it would be desirable to transform the data into a smaller dimension, we do this by using *Principal Component Analysis*. The way in which do this is by using the function from *sklearn* called PCA. An example could be:

```python
from sklearn.decomposition import PCA
...
def pca_fit(data):
    pca = PCA(n_components=2)
    pca.fit(data)
    data_pca = pca.transform(data)

    return data_pca
...
```

4 Data exploration

Now that the data has been normalized, we thought it natural to explore the data, and see if we could find something. The first thing we thought of was using *clustering*. We thought of using *k-means clustering*, and again we used the function *sklearn*. One can see an example of how we implemented this:

```
from sklearn.cluster import KMeans
...
def kmeans_cluster(data, n_clusters):
    kmeans = KMeans(n_clusters=n_clusters)
    kmeans.fit(data)
    y_kmeans = kmeans.predict(data)
    plt.scatter(data[:, 0], data[:, 1], c=y_kmeans, s=50, cmap='viridis')
    centers = kmeans.cluster_centers_
    plt.scatter(centers[:, 0], centers[:, 1], c='black', s=200, alpha=0.5)
...
```

One can also see an example of the resulting plot from taking the *combined_ 2mers* from the folder *processed_ data/combined_ data/with_ background* here:

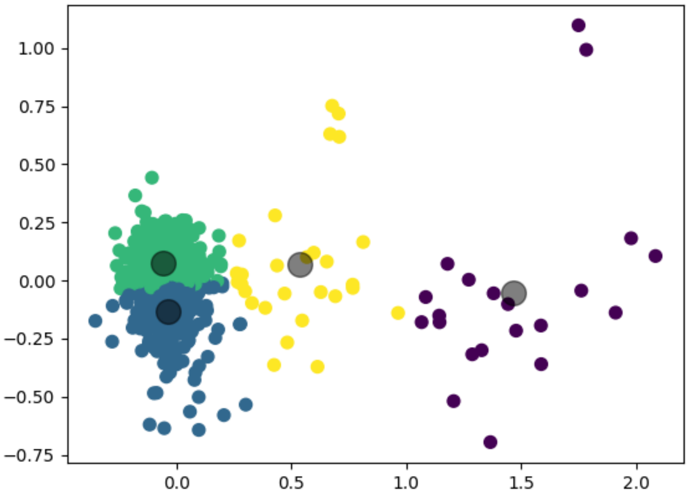


Figure 2: A k-means clustering of the combined 2-mers file

Since we have; methylated/unmethylated, cancer/healthy. Which would result in four groupings of the data. We chose to make 4 clusters of the data.