ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Epigenetic plasticity as a signature for memory allocation

Miniproject BIO-322

Guillaume Belissent and Camille Challier

23rd December 2022

1- Introduction

Epigenetic modifications, which refer to changes in DNA that do not alter the underlying genetic code, have been linked to a wide range of biological processes and diseases. Histone acetylation is an epigenetic process that removes the positive charge on the histones, weakening the attraction with negatively charged DNA. As a consequence, the condensed chromatin is transformed into a more relaxed structure that is associated with greater levels of gene transcription. In this project, we will be analyzing high-dimensional transcriptomics data in multiple cells of a mouse brain under three cell types that we call KAT5, CBP and eGFP. Our goal is to use machine learning techniques to predict, for each cell, the cell type based on its gene expression levels. We will begin by a review of the feature selection, followed by an interpretation of the data using visualization techniques. We will then describe the machine learning techniques used to predict the experimental condition and present our findings. Through this work, we hope to gain new insights into the markers of epigenetic modifications in different experimental conditions.

2- Features selection

High-dimensional data can reduce the performance and increase the run time of machine learning algorithms. With our 32285 genes, we need to perform some dimension reduction techniques. For that, we first used unsupervised feature selection methods. There were no missing values. We removed features that have approximately constant values (zero variance) that provide little to no information on the outcome. We kept only one feature within highly correlated feature groups (high multicollinearity). With these techniques, we obtained around 24 101 features that allow us to use linear machine learning methods.

To select the best genes and reduce dimensions, we tried different techniques. First we added a condition on the number of times that the gene expression is different from zero (call rates). We then focused on genes that are up- or down-regulated depending on the experimental condition. We tried to select the best features using an ANOVA F-test, a Mean Absolute Difference Classification and a Pearson's correlation factor. We did not manage to obtain adequate results with these methods. Selecting features by computing the difference of the arithmetic mean of each gene expression in each experimental condition, was one of the best dimension reduction techniques that we found. We finally choose to use a t-test technique, which shows the best accuracy/overfitting ratio. We can observe in *Figure-1* the different accuracies of a Multinomial model for different data sizes using different features selection techniques. We finally reduced the number of features to 6000. We then applied a PCA to reduce our features to 3000, which conserved 0.85 % of the explained variance as shown on *Figure-2*.

3- Visualisation

Using PCA we visualize the first 3 principal components. We can observe that eGFP is fully separated but CBP and KAT5 are not very distinguishable. We can again find this idea of differentiation of eGFP when looking at a confusion matrix, where the linear model has more difficulty classifying the 2 acetylated groups than the non-acetylated one.

We also used UMAP to visualize data and attempt dimension reduction but we did not manage to find hyperparameter settings that worked nicely. We also attempted to use

supervised UMAP which provided elegant figures but had a tendency to overfit (*Figure-3).* Also we only found a version implemented in python.

Using features selection techniques we can analyze the more important predictors.By using a t-test, we found that the 10 most representative genes are: Mid1, Polr1b, Hexb, Gm42418, Gm11867, Camk1d, Gm47283, AY036118, Sec61a1, Lrp1. Using mean difference comparison, we observed that the genes that show a higher difference of expression levels are:

- between CBP and KAT5 : Polr1b, Lrp1, Mid1, Gm26917, Camk1d
- between CBP and eGFP : Gm42418, Gm26917, Hexb, Snrnp70, Rsrp1
- between KAT5 and eGFP : Gm42418, Snrnp70, AY036118, Hexb, Gm26917

These predictors effectively classify but not sufficiently the experimental conditions, as demonstrated in *Figure-4*. We also made a heatmap (*Figure-5*) representing these genes grouped by samples. It shows the distribution of gene expression across each cell type. We can observe that gene expressions vary across the different cell types.

4- Linear and non-linear machine learning techniques

To predict the experimental condition, we first try to use some linear techniques. Using our 24 101 selected predictors, a simple logistic classifier obtained an accuracy of 0.896. By tuning a LassoClassifier on hyperparameter lambda in the search space of [`1e-8,1e-2`], we found a pretty good accuracy of 0.910 with cross-Validation with lambda = 8.5e-5. A Ridge Classifier has a high running time, so we haven't tried a lot of lambda values between 1e-2 and 1e-5. We found an accuracy of 0.891 with lamba 1e-3.

Linear methods are not enough to perfectly classify the data and we thus try nonlinear methods. First of all, we tuned a RandomForest, with cross-validation hyperparameters tuning of n_trees between 10 and 1500 and max_depth between -1 and 10. We found an accuracy of 0.66504. We thus choose to focus on another machine learning model like XGBoost. We found an accuracy of 0.897 using the hyperparameters num_round [50,500], max_depth [2,10], eta [1e-2,0.2], min_child_weight [1,6], subsample [0.6,0.9], colsample_bytree [0.6,0.9] and gamma.

. By reducing our features to 1200 using a mean difference selection, we used SVM or SVC (for Support vector classifier). This method for selecting predictors overfit the training set before SVM hyperparameter tuning such that the cross-validation provided erroneous results.

Finally, concerning Neural Networks : Due to the high computation time we did not manage to tune them precisely or use large and deep neural nets. Using a method for dimension reduction using t-value (reducing the number of predictors down to 6000) or using mean difference techniques followed by a PCA, we obtained an average cross validation error above 0.1 exceeding the error for lasso classification. For better use of NNs it would be good to have better predictors selection methods and access to better computing power.

5- Conclusion

As a conclusion, we can see in *Figure-6* that many of the selected genes are relevant to neural activity (Hexb, Mid1). Others indicate higher levels of transcription (SNRNP70, Polr1b, Rsrp1). This project has enabled us to identify genes that play a role in determining the experimental condition. While better predictor selection methods and computing power could have improved analysis, we have still gained valuable insights into epigenetic plasticity signature.

6- Annexe : Click **HERE** to access the drive with all figures.

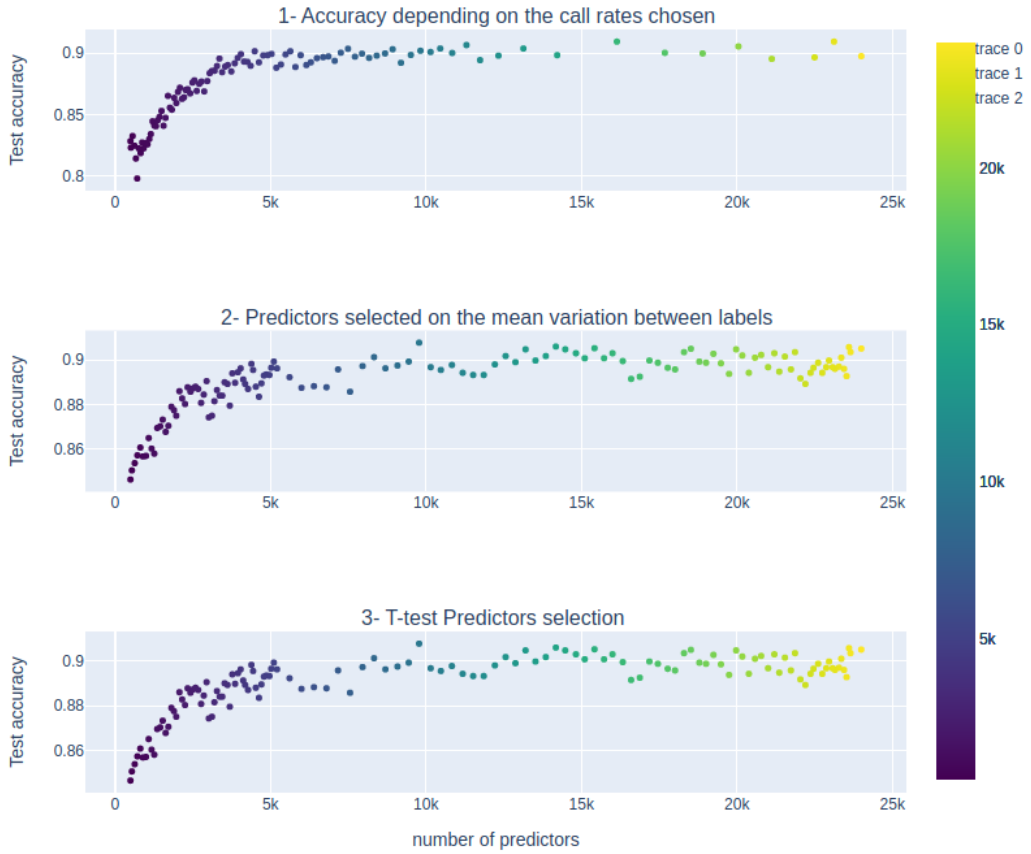

Figure 1 : Accuracy variation with different features selection techniques

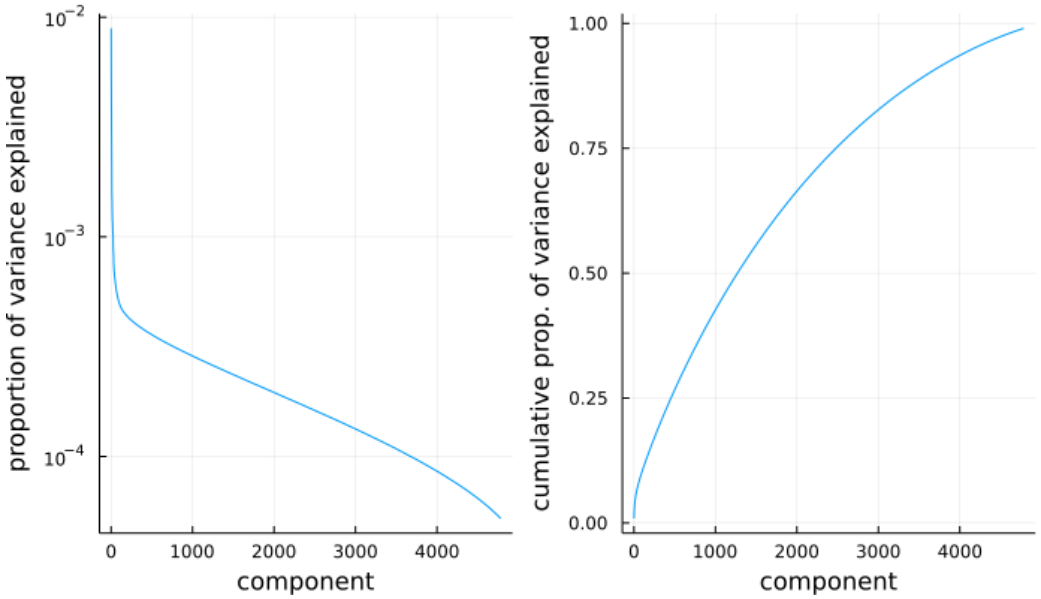

Figure 2 : PCA explained variance plot

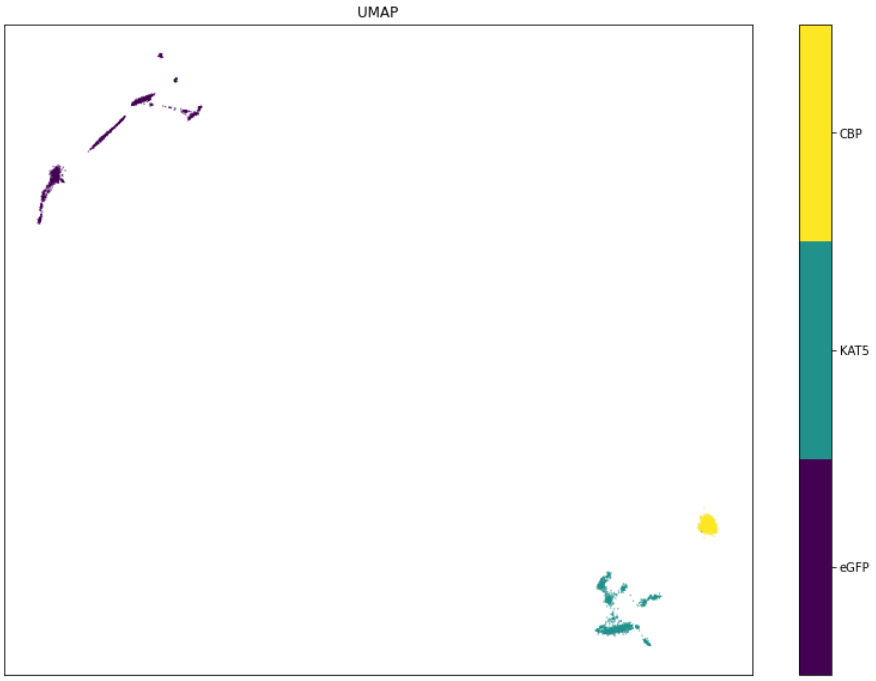

Figure 3 : Supervised UMAP 2 components. (See Supervised_UMAP.ipynb)

Figure 4 : Spatial representation of the labels repartition using the estimated 4 more important genes.

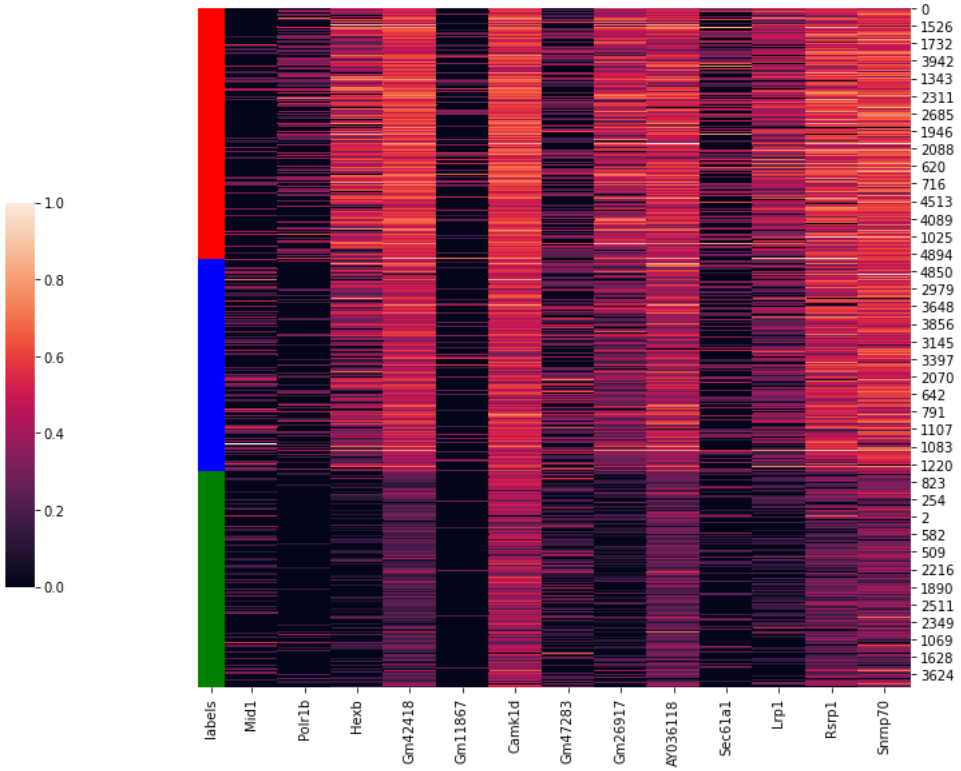

Figure 5 : Heatmap of gene expressions for most relevant genes. Made with seaborn (see clustermap.ipynb)

We can observe the function of the genes using databases like genecards.org, ncbi.nlm.nih.gov …	**Rsrp1** is a gene that encodes for components of ribosomes.
Genes starting in **Gm** are predicted genes. They are believed to exist based on computational analyses of genomic data, but have not yet been experimentally verified. There is little known about their function.	**MID1** is a gene that provides instructions for making a protein called midline 1 (MID1). This protein plays a role in the development of the central nervous system.
SNRNP70 (also known as U1-70K) is a protein that is involved in the process of splicing.	**Polr1b** is a gene that encodes for RNA Polymerase 1 subunit b.
Hexb is a protein encoding gene that encodes for a subunit of beta-hexosaminidase (A & B). Beta-hexosaminidases play a critical role in neurons. These enzymes are found in lysosomes, which are structures in cells that break down toxic substances and act as recycling centers.	**LRP1** is a gene that encodes for LRP1 protein. This plays a crucial role in the metabolism of cholesterol and other lipids in the body. Defects of this gene have been linked to Alzheimer's disease.

Figure 6 : Functions of the most important genes.