Predicting gene expression from five functionally distinct histone markers

Akshat Gupta
Carnegie Mellon University
Pittsburgh, PA 15213

Sanat Mishra
Carnegie Mellon University
Pittsburgh, PA 15213

Zhen Yang
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Epigenetic markers are known to regulate gene expression without altering underlying genomic sequences. Individual epigenetic modifications are classified by whether they upregulate or downregulate gene expression. In this project, three machine learning approaches have been compared in their ability to predict mammalian gene expression using five uncorrelated histone markers. These methods will be useful in making *in-silico* inferences on the activity of a gene.

1 Introduction

Histone modifications play a crucial role in modulating gene regulation. Some modifications can upregulate genes, while others downregulate them. Epigenetics reflects the environmental conditions an individual has been exposed to and can explain the overexpression or repression of almost all genes without altering the DNA sequences. Changes in epigenetic patterns on genes can explain the onset of diseases such as cancer and schizophrenia, among others, while also illustrating the molecular basis of aging.

At a high level, we can predict whether a given gene will be expressed or not by considering the effects of histone marks by moving a sliding window across the genome. The same coordinates on the genome can exhibit both epigenetic marks of activation and repression at different levels. Thus, predicting these modifications' net effect on gene expression forms an interesting and challenging machine-learning problem. In this project, we are considering the effect of the following modifications H3K4me3 *(activator)*, H3K4me1 *(repressor)*, H3K36me3 *(indicates activation)*, H3K27me3 *(repressor)* and H3K9me3 *(repressor)* on the expression of multiple genes. Marker types are associated with promoter regions, enhancer regions, transcribed regions, Polycomb repression, and heterochromatin regions.

1.1 Computational problem

This project aims to predict gene expression, which is a binary outcome based on five epigenetic modifications denoted by discrete values. The input data will be an array of epigenetic levels from 6601 genes. Each gene has 100 bins with five core histone modification marks. This makes our input data matrix have dimensions 660100 x 5. The final output will be an array of 0/1 target variables where a high gene expression level corresponds to target label = 1, and low gene expression corresponds to target label = 0. The dimensions of this vector will be 6601 x 1.

The dataset used here is on "E047" (Primary T CD8+ naive cells from peripheral blood) cell type from the Roadmap Epigenomics Mapping Consortium (REMC) database.

1.2 Data Structure

Aiming to capture both neighboring-range and long-range interactions among histone modifications and gene regulation, for each gene, a bin of length 100 basepairs (bp) are selected from 5,000 bp upstream of the transcription start site (TSS) to 5,000bp downstream of the TSS. This can be visualised in Figure 1.

The training data is a matrix where each row represents a bin for a given gene. In total, there are 6,600 genes and 100 bins per gene. This yields an input matrix with 660,000 rows, 6 columns which the first 5 columns correspond to the selected 5 histone markers. The last column represents the binarized gene expressions. A gene is labeled as 1 if it's turned on and labeled as 0 if it's turned off. Thus, the training data is a 660,000 x 6 matrix.

The testing data is also a matrix follows the same format as the training data, which includes 6,600 genes and yields a 660,000 x 6 matrix.

1.3 Existing models and machine learning approaches

Human genome reference sequences are widely studied to investigate the effect of genetic variance and its association with host diseases. Smith and Meissner (2013) and Sokolov et al. (2023) found that epigenetic mechanisms such as DNA methylation and accessibility play a major role in host phenotype as well. However, unlike human genetics, epigenetic references are less available and understudied. Reference Epigenome Mapping Centers (REMCs) established by NIH Roadmap Epigenomics Program Bernstein (2010) aims to elucidate the effect of epigenomic landscape across tissues and cell types on human health.

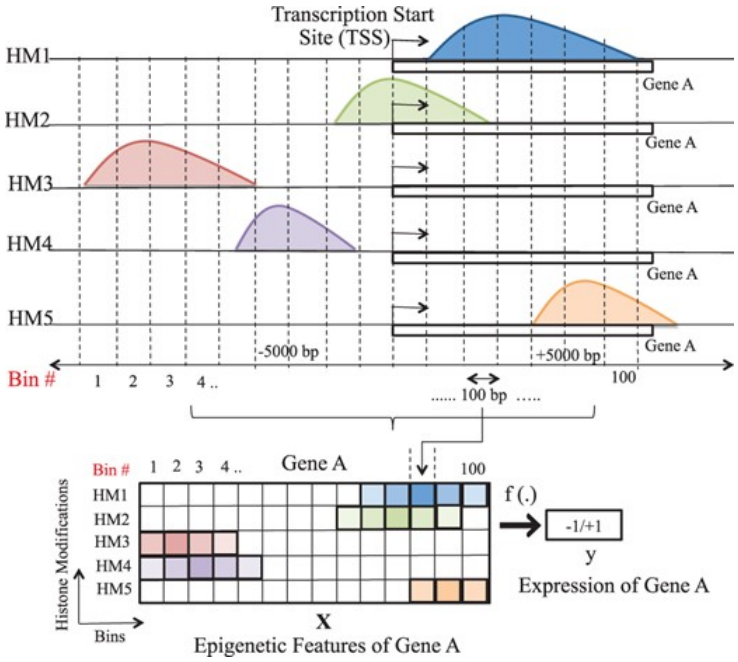

Figure 1: Bins of size 100 near the TSS, where presence of each histone mark is measured.(Ritambhara et al. (2016)

)

By utilizing the dataset, Kundaje (2015) found disease-associated genetic variants are enriched in tissue-specific epigenomic marks, which stress the importance of regulatory elements and their impact on human health. Advanced algorithms such as DeepChrome developed by Ritambhara et al. (2016) also utilized the above gene regulatory factors to predict gene expressions. The same dataset will be used in our project, and we will train it with traditional clustering algorithms discussed in class. By doing so, this project aims to explore the predictive nature of histone modification signals on gene expression levels.

2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, aimed at gaining insights, understanding patterns, and identifying key features within a dataset. It involves a variety of techniques and statistical methods to summarize, visualize, and interpret the data before implementing any formal modeling or hypothesis testing.

The EDA directed out model choices as well as affirmed the inferences.

We made the following plots to understand the nature of our data

1. Correlation scatter plots
2. Heatmap based on correlation values

2.1 Correlation scatter plots

Correlation scatter plots are a powerful visualization tool used in exploratory data analysis to examine the relationship between two variables. By analyzing the pattern formed by the data, we can assess the strength, direction, and form of the relationship between the variables. A positive correlation is indicated by dots forming an upward trend, while a negative correlation shows a downward trend. The scatter plot also helps identify outliers and clusters within the data, which can provide valuable insights into the nature of the relationship. Additionally, correlation coefficients, such as Pearson's correlation coefficient, can be calculated from scatter plots to quantitatively measure the strength and direction of the relationship.

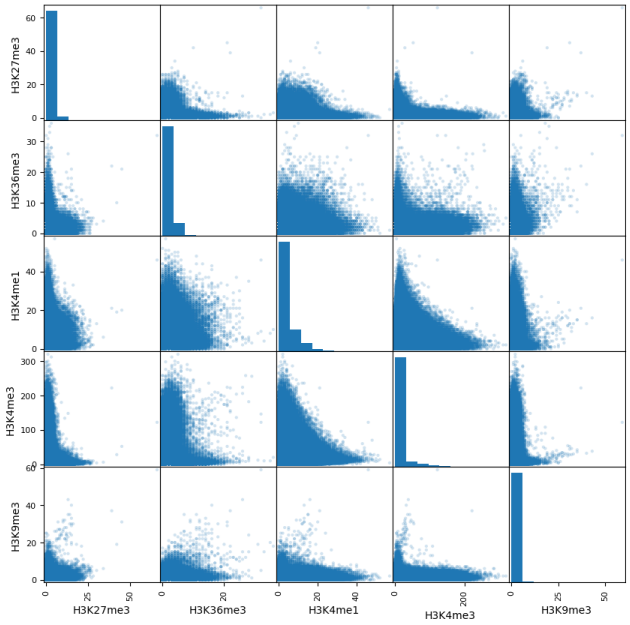

Figure 2: Bivariate plots between each of the five markers

The nature of the correlations between the histone markers are visible on Figure 1. For a given histone marker, it can be seen that the trend among data points is regular. No outliers can be detected.

2.2 Heatmap based on correlation values

To calculate the pairwise correlations between each pair of markers, a heatmap with the Pearson Correlation coefficients was plotted.

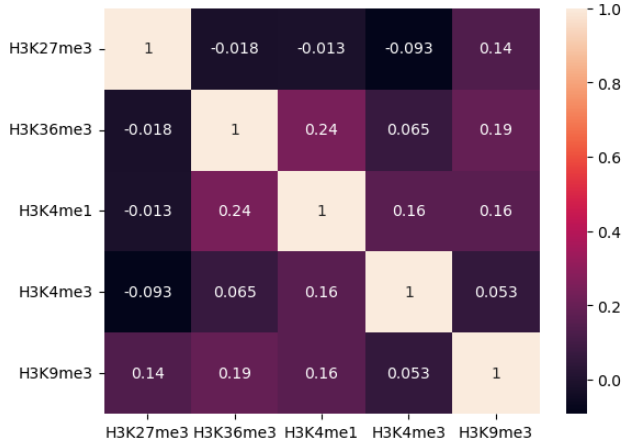

Figure 3: Heatmap of correlation coefficients for all histone markers

In Figure 2, we see that the correlation coefficients are small and are close to zero. This is in line with what was expected, owing to the fact that the histone markers are functionally and spatially independent of one another. While there appears to be low correlation between the markers, there can exist complex dependencies. There might exist higher level relationships that are not captured by expression of markers alone as suggested by Rothbart and Strahl (2014).

3 Methods

3.1 Logistic Regression

The first method implemented for our analysis is logistic regression. Since we are performing binary classification, and logistic regression was inherently developed for classification tasks, logistic regression is a good choice for our study.

Secondly, the coefficients of the logistic regression model can be easily interpreted to understand the relationship between the input features (epigenetic modifications) and the target variable (gene expression). This provided useful insights into the biological mechanisms underlying gene expression. Additionally, the input data has *only* five features (epigenetic modifications), making it a low-dimensional problem. Logistic regression is well-suited for low-dimensional problems where the number of features is not too large.

3.2 Decision Tree

The second method we chose is the decision tree. A decision tree generates easily interpretable results by constructing trees illustrating the decision process. Since our data has relatively few features, the trees won't be overly complicated and will mostly be easy to understand. Also, the decision tree makes the decision by the majority of the data, meaning each class is weighted by the proportion of positive and negative outcomes in that class. Therefore, it considers the skewness of the data, which is beneficial for dealing with biological data.

3.3 Support Vector Machine

The third method we plan implemented is a support vector machine to classify the data. We performed this using the Scikit-learn package. SVM is a highly robust classification algorithm that can work very well for our dataset as it is less prone to overfitting than other classification algorithms. The relationship between our discretized data and the gene expression is unknown, complex, and needs to be inferred. For such a task, SVM is the ideal choice for an algorithm. Since SVM can incorporate non-linear decision boundaries, it can go one step beyond logistic regression and correctly classify points that have been previously misinterpreted. Building a robust model via SVMs also helps classify any data that has missing values so it can work well on sparse datasets.

4 Results

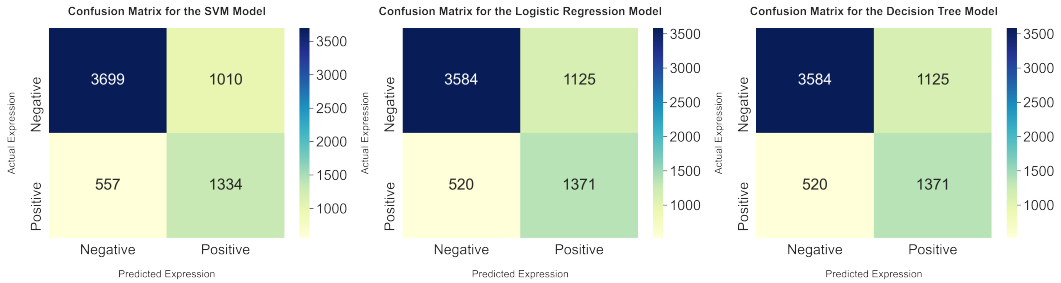

Figure 4: Confusion matrices summarising the model performances. All three models are able to capture gene expression correctly from the histone data.

The confusion matrices summarise the actual number of correctly and incorrectly labeled data in Figure 4. In particular, the SVM predicts a higher number of false negatives while the Decision Tree predicts a higher number of false positives. In the former case, genes thought to be inactive can inhibit the full understanding of say a particular disease that might use a gene but is not considered for analysis. While, in the latter case, more false positives will lead to erroneous genes that need not be examined in future steps. To overcome these problems, we can consider using a SVM model with a GO-term analysis downstream to filter out genes that are wrongly called.

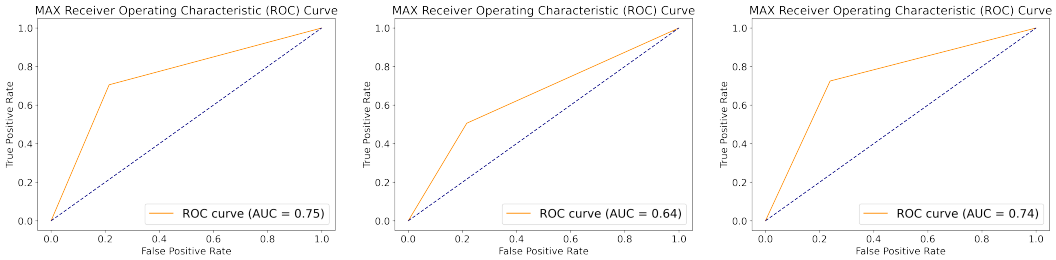

Figure 5: Comparing ROC curves among all the models. L-R: SVM, Logistic regression, and Decision Tree

Our problem is a binary classification problem for which a regular ROC curve is the most optimal way of understanding which model performs better. The DT appears to have a larger AUC (Area under the curve) and can be considered the best model among the three. This is evident from Figure 5. However, as discussed previously, the trade-off among false positives and false negatives might render the Decision Tree inferior to our particular biological problem.

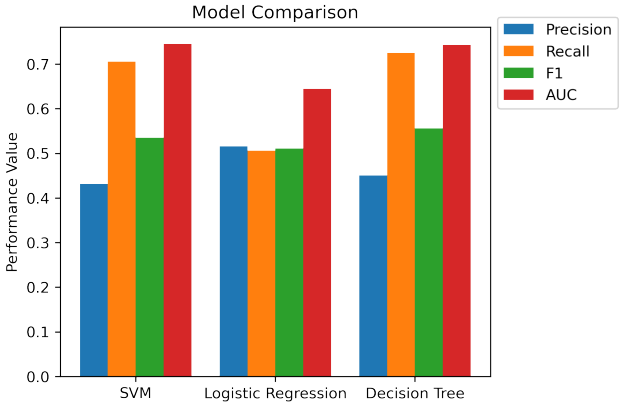

Figure 6: Model comparison between SVM, Logistic regression, and Decision Tree. SVM and Decision Tree perform comparably.

The three models can be best looked at from the plot in Figure 6. Although, the AUC for the DT appears to be larger than that for the SVM as is seen in Figure 5, here we can explicitly compare the two. The methods we used matched the performance of the models used in literature. (Ritambhara et al. (2016)). The Decision Tree and SVM performances were comparable.

In all three figures (Figure 4, Figure 5, and Figure 6) the logistic regression performs the worst. This is possibly because it fails to capture non-linear relationships among the histone marks.

5 Discussion and future directions

Linear SVMs cannot model complex feature interactions. To capture this aspect of the data, we looked at more complex decision boundaries using a RBF (Radial Basis Function) kernel and a polynomial kernel. In case the data is such that it requires a non-linear decision boundary to classify gene expression, a RBF kernel is better suited. In the data that was used in this project, the RBF kernel performed the best mamong different kernels.

It is known that decision Trees (DT) are a good choice for biological data as there are hierarchical relationships among different proteins that belong to multiple pathways. Thus, it was surprising that the DT and SVM had comparable performances. However, as was explained earlier when looking at the correlation plots, it could be that this is because of higher level relationships between the histone proteins that the DTs are able to learn. The only concern for a decision tree in handling non-linearity

is that they require deeper trees to do so, increasing the risk of overfitting. Lastly, SVMs might outperform DTs on a higher dimensional dataset in terms of speed and that can be another reason for choosing a SVM classifier.

For the reasons that SVM and DT are well suited for this classification problem, logistic regression suffers i.e. non-linearity of data. Logistic Regression may not capture complex feature interactions as well as Decision Trees, which can naturally model interactions between features by splitting on multiple features in a hierarchical manner. While it is possible to incorporate feature interactions in Logistic Regression by manually creating interaction terms, this can quickly become cumbersome for high-dimensional datasets.

To improve model implementation and generalisation, the following additions will prove useful.

5.1 Exploratory Data Analysis

The methods used in this project to detect outliers are not based on a statistical approach. Since the dimension of the input data (i.e. the number of features) were few, it was possible to detect the presence of outliers by eye. However, calculating the Median Absolute Deviation (MAD) or the Tukey's Fences to discover outliers can be an important task in generalising the model to work with more input features.

5.2 Dimensionality reduction

Although not applicable to the data used in this project implementation, given more number of histone markers a dimensionality reduction methods such as PCA would help pick out the most important features. Adding this can help make the model more generalisable.

5.3 Neural netowrks

Newer and more rigorous methods such as neural networks can outperform the models presented here. A neural network (ResNets or Convolutional Neural Network) can significantly improve the task of classification as their architecture is well suited to handle non-linear data.

References

B. E. Bernstein. The nih roadmap epigenomics mapping consortium. *Nature Biotechnology*, 28:1045–1048, 2010.

A. Kundaje. Integrative analysis of 111 reference human epigenomes. *Nature*, 518:317–330, 2015.

S. Ritambhara, L. Jack, R. Gabriel, and Q. Yanjun. Deepchrome: deep-learning for predicting gene expression from histone modifications. *Journal of Bioinformatics*, 32:1639–1648, 2016.

S. B. Rothbart and B. D. Strahl. Interpreting the language of histone and dna modifications. *Biochimica et Biophysica Acta*, 1839:627–643, 2014.

Z. Smith and A. Meissner. Dna methylation: roles in mammalian development. *Nature Reviews Genetics*, 14: 204–220, 2013.

A.V. Sokolov, DM. Manu, and D.O.T. Nordberg. Methylation in mad1l1 is associated with the severity of suicide attempt and phenotypes of depression. *Clinical Epigenetics*, 15:1, 2023.