Enhancer Prediction and Feature Ranking of Enhancer Marks

Albert Kuo
University of Chicago
albertkuo@uchicago.edu

Abstract

Several epigenomic marks are available for predicting enhancers, but the integration of these marks continues to be a challenge. Three statistical models – a hidden Markov model, a support vector machine, and logistic regression – were implemented to integrate the data and better understand these enhancer marks, both in terms of their correlation with each other and their power to predict enhancers. The combined results provide further insight into the interrelations between enhancer marks and demonstrate that certain enhancer marks are more predictive than others.

I. Introduction

Enhancer prediction is critical for understanding the regulatory mechanisms behind gene expression. While different cells in the human body share the same genome sequence, different genes are expressed in different types of cells. The regulation of gene transcription involves both proximal and distal DNA sequences. The proximal element, located close to the gene, is the promoter while the distal regions are the enhancers. Enhancers interact with promoters through DNA looping and can be bound with proteins known as transcription factors to influence gene expression. Since faulty gene expression can result in many diseases such as cancer, this motivates the need to better understand enhancers and gene regulation.

Unlike promoters and gene-coding regions, which are well annotated, enhancer prediction remains a challenge because they are often located far away from their target genes. Several statistical models and experimental techniques have been proposed for enhancer prediction using epigenomic data such as enhancer sequence patterns and ChIP-seq data [11]. Certain histone modifications are understood to be associated with enhancers, although the exact nature and strength of the relationship is still unclear. In addition, the presence of histone acetyltransferase p300/CBP, a chromatin-modifying enzyme, is also associated with enhancers as it enables DNA looping for enhancer-promoter interactions [4]. Recently, enhancer RNAs (eRNAs) have been detected with high-throughput RNA sequencing (RNA-Seq), whose expression is correlated with enhancer regions [9]. However, these enhancer-identifying characteristics are not individually sufficient to describe or identify enhancers. For instance, because not all enhancers are transcribed, the presence of eRNA reflects only a fraction of enhancers in the genome. Moreover, it has been suggested that histone modifications or co-activators patterns do not adequately describe enhancers and are often characteristic of both enhancers and promoters, raising further challenges for identifying enhancers [2].

The work I did this summer consisted of two parts. First, I examined the relationships between histone marks, p300, and eRNA using a variety of graphical methods and explored different ways to integrate the data in order to better predict enhancers. Second, I implemented three statistical models, a hidden Markov model (HMM), a support vector machine (SVM), and logistic regression, to predict enhancers, with the goal of understanding which features are most important for enhancer prediction. Since my work in the second part builds upon earlier results in the first part, I will focus my discussion primarily on the methods and results of the statistical models.

II. Methods

I. Datasets

The lymphoblastoid cell line GM12878 under the ENCODE Project was studied. In particular, ChIP-seq data for eight histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H4K20me1) and the p300 transcriptional coactivator was used. The histone modifications data was provided in intervals of 200 base pairs. To integrate data for p300, enrichment of p300 binding was defined to be positive with any overlap in the intervals.

Data for eRNA, measured in transcripts per million (TPM) reads, was taken for CAGE based enhancers under the FANTOM project. For the purposes of our analyses, a binarized version of the eRNA data was used. Given an interval of 200 base pairs, eRNA expression was determined to be positive when there was a nonzero TPM value for any overlapping interval.

For the supervised learning models, SVM and logistic regression, a positive class set was constructed consisting of all intervals that were enriched in p300 binding ($n = 113,883$). A corresponding negative set of approximately equal size was constructed by taking a matching interval in the negative set for each interval in the positive set ($n = 107,085$). The average distance (m) between intervals in the positive set was calculated and each matching interval was located ($m/2$) base pairs downstream of the interval in the positive set. The negative set was constructed in this way to mimic the distribution of the positive set and avoid an over-abundance of intervals with no marks, which would provide little additional information about the relative importance of features and artificially enhance the classification score of the models.

II. Hidden Markov Model

A hidden Markov model (HMM) is an unsupervised learning model. It was used to model observed combinations of chromatin marks, eRNA, and p300 as a product of independent Bernoulli random variables, similar to studies done in the past [8]. Chromatin states, the hidden states, are then inferred from the observed combinations. Since a HMM assumes the system being modeled to be a Markov process, this corresponds well to the sequential correlation present in the genome, as chromatin states are likely to span multiple intervals.

Spectacle, which is based on the ChromHMM code, implements a spectral learning algorithm for hidden Markov models [10]. The number of states to be used was determined empirically to achieve a balance between capturing the potential complexity of the combinations and retaining interpretability of the states.

III. Support Vector Machine

SVMs are a supervised learning model for classification. A SVM finds a non-probabilistic binary linear classifier, optimized for separating observations in the two classes with a gap that is as wide as possible [3]. The Python library scikit-learn was used to implement a support vector machine with a linear kernel. A linear kernel was chosen since it was more efficient computationally, its feature weights are more easily interpretable, and there was no apparent gain in classification score using other kernels.

To prevent overfitting, a soft margin SVM was used and the penalty paremeter was tuned using cross-validation [3]. To prevent multicollinearity, the data was pre-processed using principal components analysis (PCA). As the principal components are orthogonal to each other, there is no correlation between the components. These components are then used as features for the SVM.

IV. Logistic Regression

Logistic regression is a well-known probabilistic classification model. The application of this model to our data used p300 enrichment as the binary dependent variable (y) and the eight histone marks and eRNA expression as the features (x).

To decorrelate the features, logistic ridge regression was used with scikit-learn. Ridge regression applies a L2 penalty to the log likelihood of the weights (w), i.e.

$$l(w) = \ln[\prod_{i}^{N} P(y_i|x_i, w)] - \lambda||w||_2^2$$

where $\lambda||w||_2^2 = \sum_j w_j^2$ is the L2 regularization term and λ is the penalty parameter, which was tuned with cross-validation.

III. Results

I. Hidden Markov Model

Spectacle was first run with eight histone modifications and eRNA as features. It outputs both the model parameters and the chromatin state assignments for each interval. Chromatin states with similar emission parameters are automatically grouped together (Fig. 1). Based on the enrichment of different features, which have specific biological functions, each chromatin state can be assigned to a possible role. Although the primary interest in our analysis is on enhancers, the applicability of the HMM thus extends to other chromatin states.

Possible enhancer states were determined by the levels of H3K4me1 and TPM (or eRNA expression). P300 enrichment for each of these states was subsequently inspected (Fig. 2). In general, the highest percentages of p300 overlap corresponded with high levels of TPM, H3K4me1, and other histone marks.

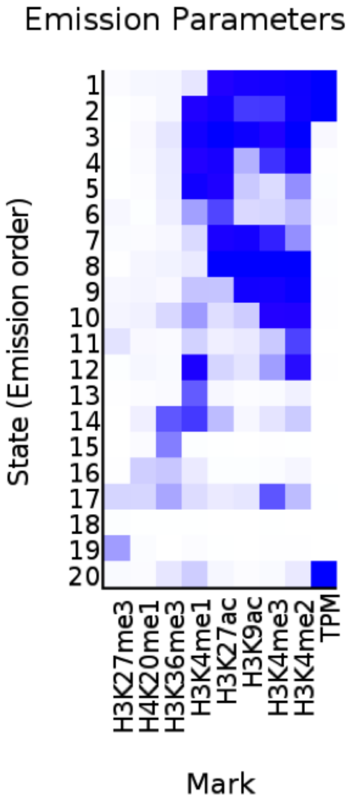

Figure 1: *Heat map of emission parameters. The saturation of each cell reflects the probabilities for each feature, marginalized over all other features.*

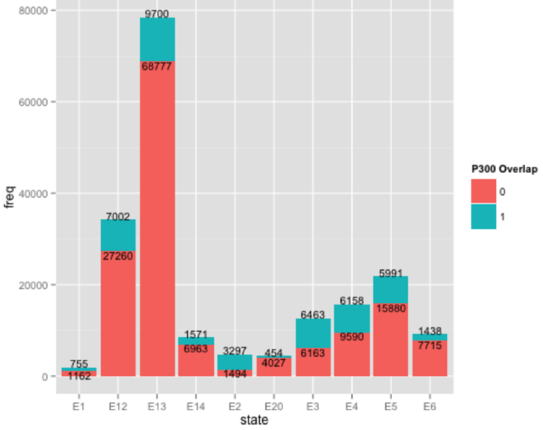

Figure 2: *Stacked bar graph of p300 overlap for each enhancer state. 0 indicates no overlap and 1 indicates overlap.*

As another way to validate the results, p300 overlap was used as an additional feature in the HMM. With an increasing number of features, a high number of states becomes more challenging to interpret. Since Spectacle uses a spectral learning algorithm, the appropriate number of states to be used can be deduced from the rank of the tensor. However, because determining the rank of a tensor is not a trivial problem, the number of states was determined empirically instead.

With ten states, the TPM feature is mainly reduced to state 4 (Fig. 3) instead of spanning three different states (Fig. 1). Moreover, state 4 aligns with similarly high levels of p300 and H3K4me1. The number of intervals in each state are also more balanced with a lower number of states. State 9, which consists of sections with no marks, has the highest number of intervals, with 104,352 intervals, and state 3 has the lowest number of intervals, with 11,383 intervals.

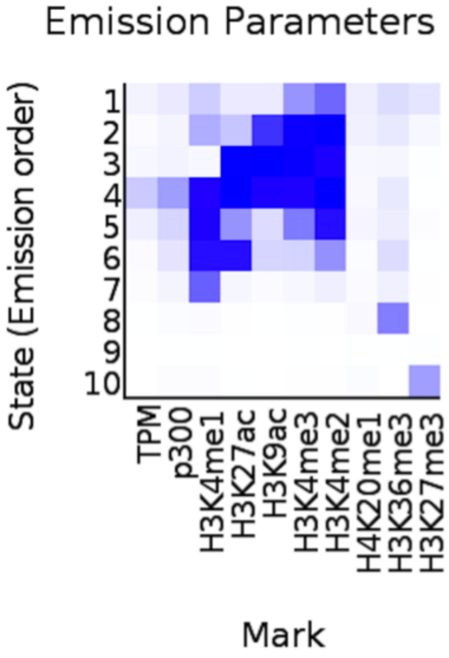

Figure 3: *Heat map of emission parameters with ten states.*

II. Support Vector Machine

Since a linear kernel was used in the SVM, the relative importance of the features can be determined using the absolute value of the feature coefficients (Fig. 4). The three features with the highest coefficients are H3K4me1, H3K4me2, and H3K4me3. The penalty parameter (C) was tuned with cross-validation and the mean cross-validation (CV) score peaked at 0.7117.

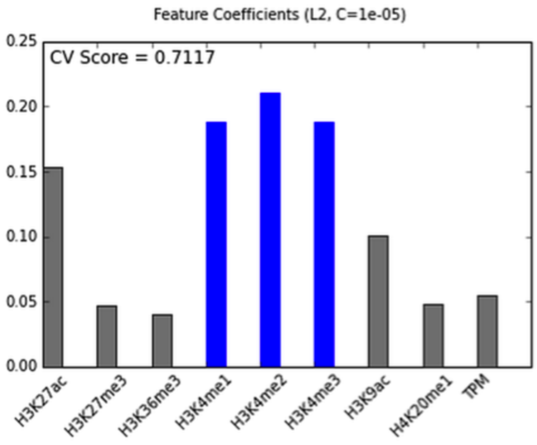

Figure 4: *Bar graph of the absolute value of the feature weights in the SVM. The three greatest weights are highlighted in blue.*

Principal components analysis was used to decorrelate the features. While the components are harder to interpret, the feature weights will not be influenced by correlation among the features. The top three components accounted for 76% of the variance in the observations and the top five components for 87% of the variance. Therefore, only the top five components were used as features in the SVM.

The greatest feature weights are for the first component and the third component (Fig. 5). The first component is a combination of several characteristic marks of enhancers, including H3K4me1 and H3K27ac, while the third component has a combination of marks that more closely reflects a gene-coding region, given the high level of H3K36me3 (Fig. 6).

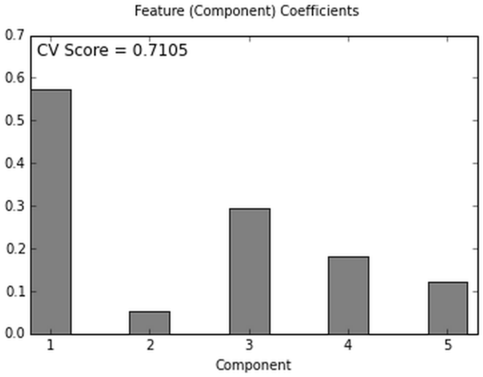

Figure 5: *Bar graph of the absolute value of the feature weights for each component in the SVM.*

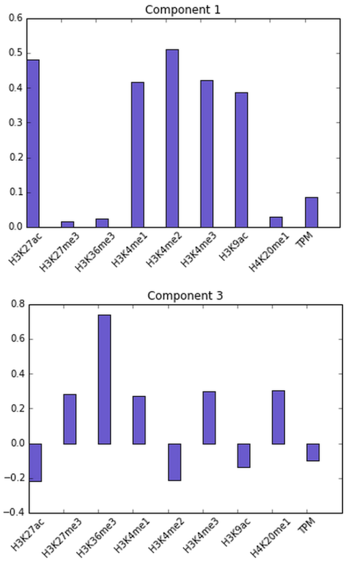

Figure 6: *Bar graph of the linear combination coefficients for principal components 1 and 3.*

III. Logistic Regression

The relative importance of the features in the logistic regression model was determined by the absolute value of their coefficients. Since multicollinearity may influence the feature coefficients, correlation between the features was reduced using a L2 penalty.

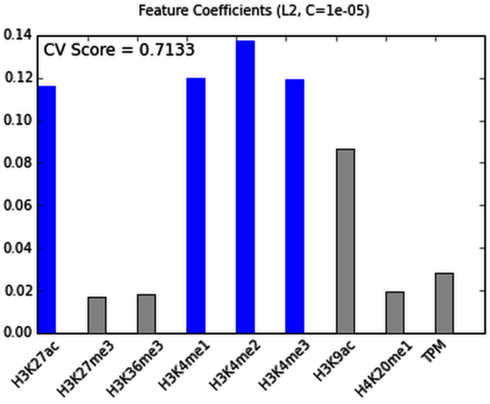

Figure 7: *Bar graph of the absolute value of the feature coefficients in logistic regression. The four greatest coefficients are highlighted in blue.*

Similar to the results from the SVM, the four most important features are H3K4me1, H3K4me2, H3K4me3, and H3K27ac. In addition, logistic regression performed similarly to the SVM, with a ten-fold cross-validation score of 0.7133.

IV. Discussion

The application of different statistical models reveals different aspects of the relationships between enhancer marks. In the unsupervised learning model of a HMM, the different features of histone modifications, eRNA expression, and p300 overlap are shown to be correlated, but are not guaranteed to co-occur. In part, this is due to the incomplete nature of the data for each feature, but it may also be a result of confounding signals from each feature. Therefore, it is likely that each enhancer mark only provides part of the picture and more accurate predictions can be made when the information provided by the enhancer marks are taken cumulatively.

The supervised learning methods predicted enhancers by learning p300-enriched enhancer regions, a technique commonly used in previous studies ([5],[12]). However, this technique is inherently limited, since the coactivator p300 only targets a subset of enhancers [6]. Thus, our analysis focused on determining the relative importance of different features rather than the classification score of our models. In settings where collecting a large number of epigenomic marks may be infeasible, knowing which features are most indicative of enhancers becomes especially important.

The results of both the SVM and logistic regression indicate that the mono-methylation of lysine 4 of histone H3 (H3K4Me1), the di-methylation of lysine 4 of histone H3 (H3K4Me2) and the tri-methylation of lysine 4 of histone H3 (H3K4Me3) are the most important features, followed by the acetylation of H3K27 (H3K27ac). The first principal component is a weighted combination of features (Fig. 6) that follows a pattern that is remarkably similar to the feature weights in both the SVM (Fig. 4) and logistic regression (Fig. 7). The importance of these features confirms prior studies that have linked these features with enhancers ([1],[7]). The relatively lower importance of eRNA expression may be explained by the small number of intervals that have nonzero TPM or may reflect an inherently lower level of overlap between eRNA expression and p300 enrichment.

To investigate these relationships further, more analysis can be done to determine a more robust comparison of the predictive power of different features. Other datasets, such as sequence motifs or binding sites conserved across different species, may also be incorporated into the analysis. Understanding the role of eRNA expression and why it has relatively lower importance in the supervised learning models is of particular interest. Lastly, exploring different training sets beyond p300 enrichment may also be beneficial for predicting a larger subset of enhancers.

V. Acknowledgments

I thank my mentor, Dr. Kevin Chen, and Jimin Song for their help and guidance. I also thank Dr. Eugene Fiorini and the DIMACS program for their support.

References

[1] Chepelev, Iouri, et al. "Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization." Cell research 22.3 (2012): 490-503.

[2] Core, Leighton J., et al. "Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers." Nature genetics 46.12 (2014): 1311-1320.

[3] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

[4] Eckner, Richard, et al. "Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor." Genes & development 8.8 (1994): 869-884.

[5] Fernández, Michael, and Diego Miranda-Saavedra. "Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines." Nucleic acids research 40.10 (2012): e77-e77.

[6] Ghisletti, Serena, et al. "Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages." Immunity 32.3 (2010): 317-328.

[7] Heintzman, Nathaniel D., et al. "Histone modifications at human enhancers reflect global cell-type-specific gene expression." Nature 459.7243 (2009): 108-112.

[8] Hoffman, Michael M., et al. "Integrative annotation of chromatin elements from ENCODE data." Nucleic acids research (2012): gks1284.

[9] Kim, Tae-Kyung, et al. "Widespread transcription at neuronal activity-regulated enhancers." Nature 465.7295 (2010): 182-187.

[10] Song, Jimin, and Kevin C. Chen. "Spectacle: Faster and more accurate chromatin state annotation using spectral learning." bioRxiv (2014): 002725.

[11] Whitaker, John W., et al. "Computational schemes for the prediction and annotation of enhancers from epigenomic assays." Methods 72 (2015): 86-94.

[12] Won, Kyoung-Jae, et al. "Prediction of regulatory elements in mammalian genomes using chromatin signatures." BMC bioinformatics 9.1 (2008): 547.