Prediction of active regulatory regions using pattern recognition within signal shape of epigenetic datasets

Abstract

Regulatory regions are noncoding regions of the genome that contain a dense cluster of transcription factor (TF) binding sites and increase the expression of target gene(s). Regulatory regions play a crucial role in determining the phenotype of a cell and are selectively activated in a spatiotemporal fashion. Due to the paucity of experimentally validated regulatory regions, most methods predict active regulatory regions utilizing genomic properties associated with enhancers and promoters. Recently, a large number of massively parallel regulatory assays were developed making it now possible to train supervised models with a large number of regions that are experimentally shown to becapable of regulatory activity. Unlike previous regulatory region prediction methods that focus on the enrichment of different epqigenetic signals, herein, we develop a framework using the template recognition algorithm, matched filter, to also take into account the signal pattern within different epigenetic datasets associated with active regulatory regions. We were able to combine the matched filters from multiple epigenetic signals using highly simple transferrable machine learning models that can be used to predict active regulatory regions in a cell-type specific manner across diverse eukaryotic species. We utilized this method to predict active promoters and enhancers in ENCODE cell-lines and showed that the difference in TF binding patterns can also be utilized to differentiate gene-proximal and -distal regulatory regions. To conclude, we provide a framework to utilize the shape within individual epigenetic marks as well as a framework to combine information from the shape of different epigenetic marks to predict cell-type specific active promoters and enhancers and their genomic properties.

Introduction

Enhancers are gene regulatory elements that activate expression of target genes from a distance \cite{}. Enhancers are turned on in a space and time-dependent manner leading to the formation of a large assortment of cell-types with different morphologies and functions even though each cell in an organism contains nearly identical genome \cite{}. Moreover, changes in the sequences of regulatory elements is thought to play a significant role in the evolution of species \cite{}. Understanding enhancer function and evolution is currently an area of great interest because variants within distal regulatory elements are also associated with various traits and diseases during genome-wide association studies \cite{}. However, the vast majority of enhancers and their spatiotemporal activities remain unknown because it is not easy to predict their activity based on DNA sequence or chromatin state \cite{}.

Traditionally, the regulatory activity of enhancers and promoters were experimentally validated in a non-native context using low throughput heterologous reporter constructs leading to very few experimentally validated enhancers that function at the same time in a mammalian tissue \cite{}. In addition to the small numbers, the validated enhancers were typically biased towards conserved noncoding regions \cite{} with particular patterns of chromatin or transcription factor binding \cite{} making these validated enhancers inappropriate for training supervised machine learning models of enhancers. Instead, a majority of the methods for predicting enhancers are trained based on properties associated with enhancers such as clusters of TF-binding sites or TF-binding motifs, conservation, and chromatin features associated with active regulatory regions in the genome \cite{}. Active enhancers and promoters typically contain sites of accessible DNA in which TFs bind (peaks in DNase-I hypersensitivity or DHS) and are flanked by histone proteins that contain several post-translational modifications\cite{}. In particular, acetylation of Lys27 in the histone protein H3 (or H3K27ac modification) is enriched on the nucleosome flanking active regulatory regions while the level of methylation on Lys4 on the same protein is used to separate enhancers (H3K4me1-enriched) and promoters (H3K4me3-enriched). However, as very enhancers have been validated, it remains challenging to assess the performance of different enhancer prediction methods and the optimal method to combine information from multiple chromatin marks to make cell-type specific regulatory predictions remains largely unknown.

In recent times, due to the advent of next generation sequencing, a number of assays were developed to experimentally test the regulatory activity of up to a hundred thousand regions in a massively parallel fashion \cite{}. The validity of the regulatory regions identified using massively parallel regulatory assays (MPRA) for training machine learning models that predict enhancers remains controversial as the sensitivity and selectivity of these assays remains questionable. A majority of these MPRAs test the regulatory activity of different regions by assessing its ability to induce gene expression in a plasmid after transfecting it into a cell-type of interest \cite{}. Such assays may not recapitulate the native chromatin environment found in chromosomes, which may be necessary for assessing whether the regulatory region is active in its genomic environment \cite{}. However, the regulatory regions identified using MPRAs could provide a treasure trove of information on these regulatory elements if they could be utilized for predicting enhancers in a cell-type dependent fashion.

We developed a new supervised machine-learning method trained from large numbers of experimentally active regulatory regions in MPRAs to accurately predict active enhancers and promoters in a cell-type specific manner. Unlike previous enhancer prediction methods that focus on the enrichment (or signal) of different epigenetic datasets, we developed a method to also take into account the signal pattern within different epigenetic datasets associated with active regulatory regions. The epigenetic signal shapes associated with these regulatory regions are conserved across millions of years of evolution and these models can be used to predict enhancers and promoters in different cell-types and tissues and across diverse eukaryotic species. We further created simple to use transferrable machine learning models with six parameters that can be used to predict enhancers and promoters in several eukaryotic species like fly, mouse, and human. We applied these models to predict active enhancers and promoters in the H1-hESC, a highly studied human cell-line in the ENCODE datasets. These analyses show that the pattern of transcription factor binding (TF) and co-binding varies between enhancers and promoters. The pattern of TF and co-TF binding at active enhancers is much more heterogeneous than the corresponding patterns on promoters. The pattern of TF binding can be used to distinguish enhancers from promoters with high accuracy. Thus, our methods provide a framework that utilizes different epigenetic genomics datasets to predict active regulatory regions in a cell-type specific manner and then utilizes further functional genomics datasets to identify key TFs associated with active regulatory regions within these cell-types.

Results

Aggregation of epigenetic signal to create metaprofile:

We developed novel methodology to predict activating regulatory elements utilizing the epigenetic signal patterns associated with promoters and experimentally identified regulatory regions using MPRAs \cite{}. Active enhancers and promoters tend to be depleted of histone proteins and contain accessible DNA on which various transcription factors and cofactors bind \cite{}. These regulatory regions also tend to be flanked by nucleosomes that contain histone proteins with certain characteristic post-translational modifications \cite{}. These characteristics lead to an enriched “double peak” signal containing troughs on regulatory regions within different ChIP-Seq experiments for various histone modifications such as acetylation on H3K27 and methylations on H3K4 \cite{}. We created a metaprofile of these double peak signals that was then utilized in a pattern recognition algorithm for predicting active regulatory region in a cell-type specific manner.

To create these metaprofiles, we aggregated the modENCODE ChIP-chip signals for the histone modification H3K27ac at active STARR-seq peaks (see Figure 1 and Methods) identified in the S2 cell-line of fly. There is a large amount of variability in the distance between the two maxima of the double peak in the ChIP-chip signal (Figure S1). Even though the minimum tends to occur in the center of these two maxima on average, the distance between the two maxima in the double peaks can vary between 300 and 1100 base pairs. During aggregation, we aligned the two maxima in the H3K27ac signal across different STARR-seq peaks, followed by interpolation and smoothening the signal before calculating the average metaprofile. In addition, an optional flipping step was performed to maintain the asymmetry in the underlying H3K27ac double peak because it may be associated with the directionality of transcription \cite{}. Finally, we also calculated the dependent metaprofiles for thirty other histone marks and DHS signal by applying the same set of transformations to these datasets. The metaprofile for the histone marks associated with active regulatory regions were also double peak signals and the maxima across different histone modification signals tended to align with each other on average (Figure S2). In contrast, as expected, the DHS signal displayed a single peak at the center of the H3K27ac double peak (Figure 1). In addition, repressive marks such as H3K27me3 were depleted in these regions and the metaprofile for these regions did not contain a double peak signal (Figure S2).

Occurrence of metaprofile is predictive of regulatory activity:
We evaluated whether these metaprofiles can be utilized to predict active regulatory regions using matched filters, a well-established algorithm in template recognition. A matched filter is the optimal pattern recognition algorithm that uses a linear filter to recognize the occurrence of a template in the presence of stochastic noise \cite{}. We evaluated whether the occurrence of the regulatory metaprofiles identified for the histone marks and DHS can be used to predict active regulatory regions using receiver operating characteristic (ROC) and precision-recall (PR) curves. The PR curve is more sensitive to false positives in the presence of skewed datasets with larger fraction of negatives \cite{}. The matched filter score is higher in genomic regions where the template pattern occurs in the corresponding signal track while the matched filter score is low when only noise is present in the signal (Figure 1). Due to the aforementioned variability in the double peak pattern, the H3K27ac signal track is scanned with multiple matched filters with templates that vary in width between the two maxima in the double peak and the highest matched filter score with these matched filters is used to rate the regulatory potential of this region (see Methods). The dependent profiles are then used on the same region with the matched filter to score the corresponding genomics tracks.

We used 10-fold cross validation to assess the performance of matched filters for individual histone marks to predict active regulatory regions identified in a STARR-seq experiment. In a single whole-genome STARR-seq experiment, several plasmids that contain a single core promoter upstream of a luciferase gene are transfected into cells \cite{}. The transfected plasmids are used to test the regulatory activity of different regions of the fly genome by placing them in the polyA-tail of the luciferase gene as differences in the gene's expression occur due to the differences in the regulatory activity of the tested region. In Figure 2, we observe that the H3K27ac matched filter is the single most accurate feature for predicting active regulatory regions (AUROC=0.92, AUPR=0.72) identified using STARR-seq. This is consistent with the literature as H3K27ac enriched peaks are often used to predict active promoters and enhancers \cite{}. In general, several histone acetylation (H3K27ac, H3K9ac, H4K12ac, H2BK5ac, H4K8ac, H4K5ac, H3K18ac) marks as well as the H1, H3K4me2, and DHS matched filters are the most accurate marks for predicting regulatory regions (see Figure 2 and Table S1) because the matched filter scores for active regulatory regions on these marks are higher than the matched filter scores for non-regulatory regions (Figure S3). The degree to which the matched filter scores for regulatory regions are higher than the matched filter scores for the rest of the genome is a measure of the signal to noise ratio for regulatory region prediction in the corresponding feature's genomic track and the larger the separation between positives and negatives, the greater the accuracy of the corresponding matched filter for predicting active regulatory regions. Interestingly, the distribution of matched filter scores for regulatory regions are Gaussian for each histone mark except for a bimodal distribution for the H3K4me1, H3K4me3, and H2Av matched filter scores (Figure S3). We also show that the matched filter scores are more accurate for predicting active regulatory regions than enrichment of signal alone as they outperform the histone peaks on ROC and PR curves (Figure S4).

While a single STARR-seq experiment identifies thousands of active regulatory regions, these regions display core-promoter specificity and different sets of enhancers are identified when different core promoters are used in the same cell-type \cite{}. As we wanted to create a framework to predict all the regulatory regions active in a particular cell-type, we combined the regulatory regions identified from multiple STARR-seq experiments in the S2 cell-type and reassessed the performance of the matched filters at predicting these regulatory regions. Merging the STARR-seq peaks from multiple core promoters in the S2 cell-type leads to higher AUROC and AUPR for the matched filters from most histone marks (Figure 2).

Machine learning can combine matched filter scores from different epigenetic features:

We combined the normalized matched filter scores (see Methods) from six different epigenetic marks associated with active regulatory regions by the Roadmap Epigenomics Mapping \cite{} and the ENCODE \cite{} Consortia using a linear SVM \cite{} and the SVM achieved a higher accuracy than the individual matched filters (Figure 2). These models are trained to learn the patterns in the matched filter scores for different epigenetic marks within experimentally verified regulatory regions and we chose these marks as we wanted to assess the applicability of these machine learning models to predict active regulatory regions across different cell-types and species. As expected, the linear SVM models outperformed the individual matched filter scores as they are able to leverage information from multiple epigenetic marks. In addition, the six-parameter SVM model displayed higher accuracy after combining the peaks identified using different core promoters. In a linear SVM model, the normalized matched filter score for each epigenetic feature in a particular region is scaled by its optimized weight and added together to form the discriminant function. The sign of the SVM discriminant function is then used to predict whether the region is regulatory. The features with large positive and negative weights are predicted to be important for discriminating regulatory regions from non-regulatory regions in such models. They can also be used to measure the amount of non-redundant information added by each feature in the SVM model. According to the SVM model, the acetylations (H3K27ac and H3K9ac) are the most important feature for predicting active regulatory regions from inactive regions. While the DHS matched filter performed the second best as an individual feature (AUPR in Figure 2), the information in DHS is highly redundant with the information in the histone marks as indicated by the fact that it has the lowest weight among the six features in the SVM model. We utilized several other machine learning algorithms to combine the machine learning models and found that they all displayed nearly similar accuracy and similar features were more important across these different models (Figure S5).

To assess the information contained in other epigenetic marks, we combined the matched filters from all 30 measured histone marks along with the DHS matched filter in a separate SVM model (Figure S6) and this model displayed higher accuracy than the 6 feature model presented in Figure 2. The feature weights in the SVM model indicated that H3K27ac contains the most information regarding the activity of regulatory regions. However, we found that a few other acetylations such as H2BK5ac, H4acTetra, and H4K12ac contain additional non-redundant information regarding the activity of these regulatory regions and might improve the accuracy of regulatory region prediction from machine learning models (Figure S7).

Distinct epigenetic signals associated with promoters and enhancers:

We proceeded to create individual metaprofiles and machine learning models for the two classes of regulatory activators – promoters (or proximal) and enhancers (or distal). We divided all the active STARR-seq peaks into promoters or enhancers based on their distance to the closest transcription start site (or TSS). Due to the conservative distance metric used in this study (1kb upstream and downstream of TSS), the enhancers are regulatory elements are not close to any known TSS even though a few promoters may actually function as enhancers. We then created metaprofiles of the different epigenetic marks on the promoters and enhancers and assessed the performance of the matched filters for predicting active regulatory regions within each category (Figure 3). The highest matched filter scores are typically observed on promoters and the matched filters for each of the six marks tended to perform better for promoter prediction. The H3K27ac matched filter continues to outperform other epigenetic marks for predicting active promoters and enhancers (Figure 3). In addition, the DHS, H3K9ac, and H3K4me2 matched filters also performed reasonably for promoter and enhancer prediction. Similar to previous studies \cite{}, we observed that the H3K4me1 metaprofile peforms better for predicting enhancers while it is close to random for predicting promoters. Similarly, the H3K4me3 metaprofile can be utilized to predict promoters and not enhancers. The histogram for matched filter scores show that H3K4me1 matched filter score is higher near enhancers while the H3K4me3 matched filter score is higher near promoters (Figure S8). The mixture of these two populations lead to bimodal distributions for H3K4me1 and H3K4me3 matched filter scores when calculated over all regulatory regions.

We created two different six-parameter SVM models to learn the combination of features associated with promoters and enhancers. These SVM models outperformed the individual matched filters at predicting active enhancers and promoters. In addition, the weights of the individual features identified the difference in roles of the H3K4me1 and H3K4me3 matched filter scores at discriminating active promoters and enhancers from inactive regions in the genome. The promoter-based (enhancer-based) SVM model performed much more poorly at predicting enhancers (promoters) indicating that the unique properties of these regions (Figure S8). We also created two SVM models utilizing matched filter scores for all thirty histone marks as features for predicting enhancers and promoters. The additional histone marks provided independent information regarding the activity of promoters and enhancers as these features increased the accuracy of these models (Figure S9). The weights of different features indicate that H2BK5ac again displays the most independent information for accurately predicting active enhancers and promoters (Figures S10 and S11). We observe similar trends and accuracy with several different machine learning models (Figures S9-S11).

The epigenetic underpinnings of active regulatory regions are highly conserved in evolution:

In order to assess the transferability of these metaprofiles and machine learning models for predicting regulatory regions in other tissues and cell-types, we assessed the accuracy of these models for predicting regulatory elements identified using the FIREWACh assay in mouse embryonic stem cells (mESC) \cite{}. During the FIREWACh assay, random nucleosome-free regions in mESC were captured and assayed for regulatory activity of the GFP gene by utilizing a lentiviral plasmid vector and inserted (or transduced) these vectors into the chromosome in mESC cells. The active regulatory elements were sequenced after fluorescence-activated cell sorting was used to identify cells in which GFP was expressed. As the FIREWACh assay tests the regulatory activity of nucleosome free regions after transducing it into the chromosome, the regulatory activity of this region is typically tested in its native chromatin environment and transduction-based assays form a more stringent test for regulatory activity. However, the shorter length of the tested region could lead to a different set of biases in the regulatory regions identified using the FIREWACh assay. In addition, as these regulatory regions were identified using a single core promoter in FIREWACh, the performance of the different models are probably underestimated similar to Figure 2.

The metaprofiles for individual histone marks learned using active enhancers identified with the STARR-seq assay in the S2 cell-line were used with matched filters to predict the regulatory activity of different regions in mESC based on the epigenetic marks in mESC (Figure 4). The matched filters for individual histone marks displayed similar accuracy for predicting regulatory regions in mESC as in the original S2 cell-line. We also show that the matched filter learned from S2 cell-line can be utilized to predict active promoters and enhancers in the BG3 cell-line of fly (Figure S12). In addition, the 6-parameter SVM models learned using STARR-seq data in S2 cell-line were also highly accurate at predicting active enhancers and promoters in mouse (Figure 4) and the BG3 cell-line (Figure S12). This indicates that the epigenetic profiles associated with active enhancers and promoters are conserved over 600 million years of evolution and they can even be used to predict regulatory regions in higher eukaryotes.

Different Transcription Related Factors (TRFs) bind to enhancers and promoters

We utilized the 6 parameter SVM model to predict active regulatory regions in the H1-human embryonic stem cell (hESC) based on the epigenetic datasets measured by the ENCODE consortium. Using these models, we predicted 43463 active regulatory regions, of which 22828 are within 2kb of the TSS and are labeled as promoters. A large proportion of the predicted enhancers are found in the introns and intergenetic regions (Figure S13). The predicted promoters and enhancers are significantly closer to active genes than might be expected randomly (Figure S14).

As the ENCODE consortium has measured binding data for 60 transcription related factors in the H1-hESC cell-line using ChIP-seq, we further studied the differences in TRF binding at promoters and enhancers (Figure 5). A number of histone acetylation and methylation related factors (SAP30, KDM4A, PHF8, and members from the CHD family) have binding sites on enhancers and promoters. However, a larger fraction of the promoters are bound by general transcription factors such as TATA-associated factor (TAF) 1, TAF7, GTF2F1, and TATA-box binding protein (TBP) than enhancers. SP1, SP2, and SP4 are promoter-associated transcription factors \cite{} and tend to have a large fraction of their ChIP-seq peaks on active promoters. On the other hand, a larger proportion of the NANOG, POU5F1, and BCL11 binding sites are found on enhancers than on promoters. These transcription factors are known to play key roles in stem-cell pluripotency and are required for the propagation of undifferentiated embryonic stem-cells in culture. As expected, repressors such as SUZ12, ZNF274, and FOSL1 have very few ChIP-seq peaks on active promoters and enhancers. Overall, TRF-binding at enhancers is more heterogeneous than TRF-binding at promoters and this is consistent with the absence of a sequence code (or grammar) that can be easily utilized to identify active enhancers on a genome-wide fashion.

We then analyzed the co-association of pairs of TRFs by analyzing the overlap between peaks of same TRF on enhancers and promoters separately. To do this, we calculated the proportion of enhancers (or promoters) that contain a ChIP-seq peak for a particular

TRF also contain a ChIP-seq peak for a second TRF. While a number of general trends regarding TRF co-association is captured by these analyses, we observe that the co-association patterns of TRFs at promoters are distinct from its co-association patterns at enhancers. These TRF co-associations could lead to mechanistic insights of cooperativity between TFs. For example, similar to a previous study \cite{}, CTCF and ZNF143 may function cooperatively as they are observed to co-occur frequently at distal regulatory regions in this study. Similarly, SP4 and RXRA co-occur quite often on enhancers and promoters.

In Figure 5, we show that the patterns of TF binding within regulatory regions can be utilized in a logistic regression model to distinguish active enhancers from promoters with high accuracy (AUPR = 0.89, AUROC = 0.87). We were also able to identify the most important features that distinguish promoters from enhancers. In addition to TATA-box associated factors such as TAF1, TAF7, and TBP, the RNA polymerase-II binding patterns as well as chromatin remodelers such as KDM4A and PHF8 are some of the most important factors that distinguish promoters from enhancers in H1-hESC. This provides a framework that can be utilized to identify the most important TFs associated with active enhancers and promoters in each cell-type.

Discussion

Our ability to accurately predict active regulatory regions in a cell-type specific manner using transferable supervised machine learning models that were trained based on regulatory regions identified using new NGS-enabled MPRAs distinguishes our method from previous works that were trained with regions that had various features associated with promoters and enhancers. Only a small number of these regions were typically tested experimentally and the precision/recall of these different features for regulatory region prediction remained unknown. These MPRAs were able to firmly establish that certain histone modifications occur on nucleosomes flanking active regulatory regions leading to the formation characteristic double peak pattern within the ChIP-signal \cite{}. We created matched filter models that were able to identify these patterns within the shape of the ChIP-signal in the presence of stochastic noise with the highest signal to noise ratio. Furthermore, we were able to combine the matched filter scores from different epigenetic features using simple transferrable linear SVM models and learned the most informative epigenetic features for regulatory region predictions.

While different acetylation modifications are associated with active regions of the genome, we were able to compare close to 30 histone marks for enhancer and promoter predictions. The H3K27ac matched filter remains the single most important feature for predicting active regulatory regions while H3K4me1 and H3K4me3 are known to distinguish different promoters from enhancers. However, our analysis shows for the first time the amount of redundancy in information in different epigenetic features for predicting active regulatory regions and shows that ChIP-experiments of H2BK5ac, H4acTetra, and H2A variants could also produce independent information that can improve the accuracy of promoter and enhancer predictions. In addition to these 30-feature SVM models, we also provide a simple to use six-parameter linear SVM model for combining H3K27ac, H3K9ac, H3K4me1, H3K4me2, H3K4me3, and DHS to predict active regulatory regions in a cell-type specific manner. The metaprofiles and the combination of epigenetic marks associated with active regulatory regions are highly conserved in evolution making these models highly transferable as shown in this work.

These six histone marks have been measured for a number of different tissues and cell-types by the Roadmap Epigenomics Mapping Consortium \cite{}, the ENCODE \cite{}, and the modENCODE Consortium \cite{}.

One aspect that is discussed less frequently is the effect of core promoter on enhancer and promoter prediction. MPRAs show that the regulatory activity of enhancers and promoters in a regulatory assay depends on the core promoter utilized during the experiment \cite{}. As the transcription factors that bind to each regulatory region are thought to play a key role in core-promoter specificity \cite{}, we think that machine learning models that contain sequence or motif-based features could be biased towards certain transcription factor binding sites when they are trained with regulatory regions identified experimentally using a single-core promoter. On the other hand, the performance of machine learning models that are trained with epigenetic features and contain no sequence-based information may be underestimated when utilizing data from a single core promoter as shown here in Figure 2. On comparing the predictions from such models with experiments using a single core promoter, some of the strongest predictions could be mislabeled as negatives even though they contain some regulatory activity leading to a lower accuracy estimate.

We also analyzed the differences in the patterns of TF binding at proximal and distal regulatory regions. The TRF binding and co-binding patterns at distal regulatory regions is much more heterogeneous than that at proximal regulatory regions. We think that this heterogeneity in TRF binding patterns makes it much more difficult to predict distal regulatory regions due to the absence of obvious sequence patterns in distal regulatory regions. We were able to create highly accurate machine learning models that are able to distinguish proximal promoter regions from distal enhancers based on the patterns of TF ChIP-seq peaks within these regulatory regions.

Methods

Creation of Metaprofile

- Identification of double peaks in H3K27ac around active STARR-Seq peaks
- Alignment, interpolation, and smoothing to create metaprofile
- Same set of operations on other chromatin associated marks to create dependent profile.

Matched Filter Algorithm

- Convolution of metaprofile with signal
- Noise leads negatives to have Gaussian distribution
- Positives have noise but also have pattern at varying intensities – hence Gaussian with more spread

AUROC/AUPR Curves

- Positives are peaks in MPRA
- Negatives are random regions of same width with H3K27ac signal that do not intersect with positives

Machine Learning Models (all implemented in scikit-learn)

- Random Forest

- Linear SVM
- Gaussian Naïve Bayes
- Ridge Regression

H1-hESC whole genome prediction

- Scan over chromatin with matched filter
- Used SVM to set cutoff

H1-hESC regions and gene expression

- Gencode v19 on hg19
- ENCODE2 data for gene expression (polyA)

H1-hESC TRF binding

- ENCODE2 TRF datasets (60 ChIP-Seq experiments)
- Intersectbed (-f 0.25)