Prediction of Enhancer RNA Activity Levels from ChIP-seq-derived Histone Modification Combinatorial Codes

Nawanol Theera-Ampornpunt
Department of Computer Science
Purdue University
West Lafayette, Indiana, USA
ntheeraa@purdue.edu

Somali Chaterji*
Department of Computer Science
Purdue University
West Lafayette, Indiana, USA
schaterj@purdue.edu

Abstract—**Motivation: Transcription-regulatory elements (TREs) are critical modulators of gene transcription, with promoters and enhancers being major hubs for gene regulation. Of these, enhancers are distal regulatory elements, typically short 50–150 basepairs DNA regions and are identified via the large numbers of genomic regions displaying enhancer-like chromatin features, such as histone modifications at regulatory sites. Given that enhancers operate at sites vastly removed from their cognate genes, their computational prediction is challenging. While ChIP-seq datasets have been used to map out parts of the epigenome, only a fraction of predicted enhancers have been found to be functionally active. Thus, recent methods, such as GRO-seq, have been used to elucidate functionally active enhancers. However, GRO-seq data is sparse, being experimentally complex, and often at insufficient resolution. Hence, in this paper, we evaluate the ability of ChIP-seq data to predict the eRNA activity, as given by GRO-seq, using as inputs the signatures of key histone modifications that are known to be predictive of enhancers. We implement two models with accurate performance for the entire range of eRNA activity, spanning 5 orders of magnitude, for the human embryonic stem cell type H1. We then perform a detailed comparison with a prior computational approach for (binary) enhancer activity classification.**

Results: Our best performing model is the Convolutional Neural Network (CNN)-based regression model, indicating that the epigenetic patterns are location invariant. We achieve an overall RMSE of 0.82 compared to the baseline's (ZeroR) 1.30. In the comparison with a prior approach, our model achieves AUC of 0.75 against the prior approach's 0.56.

Keywords-component; Enhancer prediction, GRO-seq, convolutional neural network, histone modification signatures

I. Introduction

Transcription-regulatory elements (TREs) are critical modulators of gene transcription, modulating the cell-type specific and condition-specific phenotypic states of different cell types under various physiological conditions (Maston, et al., 2006). The TREs, promoters and enhancers, are major hubs for gene regulation, and despite their conventional distinction, share many commonalities in their epigenetic signatures (Core, et al., 2014). The accurate mapping of TREs and their transcription levels is critical to the understanding of the gene regulatory landscape. Of TREs, enhancers are distal regulatory elements acting as cis-regulatory modules (CRMs), populating large parts of the mammalian genomes (Visel, et al., 2009). They are typically short DNA regions, 50–150 basepairs (bps) and are identified via the large numbers of genomic regions displaying enhancer-like chromatin features, such as histone modifications at regulatory sites. Stemming from the fact that the chromatin is folded and coiled in the nucleus, in spite of being located thousands of sequences away from the transcription start sites (TSSs), enhancers can come in close spatial proximity to the TSSs of the genes that they regulate. This availability of enhancers vastly removed from the genes, while affording efficiency to cellular regulation processes in the three-dimensional space, makes their computational prediction and annotation challenging. The dynamic modification of the TRE-associated chromatin elements, such as histones, alter DNA packaging and modify the gene regulatory landscape. These resulting chromatin states, constituting a part of the cellular epigenome, are distinct in different cell types, thus explaining how the different cell types exhibit distinct phenotypes, despite being hardwired by the same genomic DNA. Of the chromatin features pointing to the presence of enhancers, however, only a subset of these signatures point to the presence of functionally active enhancers, which are acted upon by trans-acting regulatory factors, such as transcription factors (TFs) (Natarajan, et al., 2012; Yip, et al., 2012), for effecting the enhancer effect. For example, the histone modification H3K27ac in combination with H3K4me1 is an important predictor of functionally active enhancers (Creyghton, et al., 2010). On the other hand, poised enhancers are enhancers that require some kind of activation stimulus to become functionally active. This could be enhancers that are involved in cell differentiation of embryonic stem cells, for example, and are recruited (or, made active) only when the cell undergoes differentiation to a cardiomyocyte phenotype.

Most recent computational techniques use machine learning (ML) algorithms to identify the characteristic chromatin features in experimentally determined enhancers and then use these features to predict new enhancers, in the same cell-type (cell-type specific) or across cell-types (general-purpose enhancers) (Kim, et al., 2015; Kim, et al.,

[*] To whom correspondence should be addressed

978-1-5090-3050-7/17/$31.00 ©2017 IEEE

2016; Kim, et al., 2016; Rajagopal, et al., 2013; Whitaker, et al., 2015). The most widely-used technique for mapping out enhancers genome-wide is ChIP-seq (Park, 2009), which is chromatin immunoprecipitation followed by deep sequencing. ChIP-seq datasets have been rapidly accumulating in data repositories, the largest repository being from the ENCODE (Encyclopedia of DNA elements) project (Consortium, 2012). While the data obtained from ChIP-seq experiments has been used to map out the epigenome, it has been found that only a fraction of these enhancers are actually functionally active. Thus, recent methods, such as global nuclear run-on sequencing, or GRO-seq (Core, et al., 2008), have been used to map out the functionally active enhancers. GRO-seq, as motivated by the pervasive genomic transcription paradigm (Jacquier, 2009; Kapranov, et al., 2007), makes use of the enhancer RNA (eRNA) transcription levels to indicate the activity level of enhancers (Natoli and Andrau, 2012). However, GRO-seq data is relatively sparse, being experimentally demanding to acquire, and is available for only a limited number of cell types, often with lower depths of sequencing. Thus, the resolution of GRO-seq data may not suffice for the identification of all active enhancers and in all cell types (Zhu, et al., 2013). Hence, in this paper, we wanted to evaluate the ability of ChIP-seq data to predict the eRNA activity levels, using as inputs the signatures of some key histone modifications that have been known to be good predictors of enhancers.

We are the first to perform such a fine-grained activity analysis of enhancers, in which we use multiple histone modification signatures organized spatially around the peak of the epigenetic element of choice, in our case enhancer-related histone modifications. In order to do this, we use a set of four predictive histone modifications at the sites of the binding of the histone acetyltransferase enzyme (HAT), coactivator p300. From the combinatorial signatures obtained therefrom, we use a suite of regression models to predict the eRNA activity levels, as would otherwise be obtained from GRO-seq datasets. Gauging the activity levels of enhancers can further prime the efforts toward epigenome editing for high-throughput screening of regulatory elements in the native genome in different cell types (Chaterji, et al., 2017).

Specifically, we set ourselves the problem of creating a regression curve and explore different forms of regression to solve the problem. This is a more challenging problem than the previously solved one (Zhu, et al., 2013) of classifying the eRNA activity into two clusters, one corresponding to the high expression and the other to the low expression. In this binary classification paper, the authors define a heuristic rule that a strong enhancer is one that has eRNA+ on both the sense and the antisense strands (note that eRNAs are bi-directionally transcribed) and a weak enhancer as one that has eRNA- on both the sense and the antisense strands. Here, we predict the activity level of the eRNA as a numeric value, which can then be thresholded to create classes by strength of activity, if desired (and which we do for a comparative evaluation with (Zhu, et al., 2013)). Further, knowing the precise levels of eRNA synthesis not only enables the further use of ChIP-seq data but also highlights the ability of epigenetic signatures to convey data pertaining to the pervasive nature of RNA transcription, which can also afford insights into the three-dimensional DNA packaging status (Danko, et al., 2015). This is because the ultimate RNA transcription levels will be dependent on the spatial conformation of the coiled DNA.

Overall, by comparing and contrasting different ML algorithms, some of which are more amenable to parallelization (Mahadik, et al., 2014; Mahadik, et al., 2017; Theera-Ampornpunt, et al., 2016), we also plan on contributing to the recent trend toward federating computational genomics algorithms (Chaterji, et al., 2017), abetting the creating of a domain specific language in genomics (Mahadik, et al., 2016). Not only is it important to come up with accurate ML algorithms to enable precision medicine but it is important to leverage appropriate parallelization techniques and pipeline some of these algorithms in an efficient manner, as done efficiently in related domains (e.g., metagenomics (Meyer, et al., 2017)) to speed up the closer-to-real time analysis of genomics data for delivering personalized (genomics) medicine.

Computational Challenges and Our Specific Contributions: There are several computational challenges that arise in solving our target problem, specifically that of using ChIP-seq data to obtain eRNA activity, which is in fact the most direct predictor of enhancer activity. First, the density of data is not uniform for enhancers at various activity levels. A vast majority of the enhancers lies within the range of 0.1–10, while the lower end of the range is 0.1 and the upper end of the range is 46064. This makes it difficult to fit a regression model that performs equally well for the entire range. Second, the range of eRNA activity levels spans 5 orders of magnitude, and therefore, using an absolute scale for the error in prediction will disproportionately inflate the metric for errors toward the upper end of the range if we use a traditional metric for prediction error such as Root Mean Square Error (RMSE). Third, the underlying data is inherently non-linear and hence a linear model is unable to predict the activity levels accurately.

In this paper, we solve each of the above mentioned problems and create a suite of regression models. We give the high level solution approach here and represented in Figure 1 and provide the details in Section 2. For the first problem of varying density of data points for the different eRNA levels, we adjust the distribution of the training samples by subsampling from the ranges with high density of samples, and duplicating samples in the sparser ranges. The resulting uniform distribution of the output feature lets the models focus on the entire range equally. For the second problem of distortion of the RMSE metric, we first put the output feature through a log transformation that reduces the spread of the data and then apply the traditional metrics of RMSE. For the third problem of non-linearity, we ML models that do not assume linearity of the data (such as, Deep Convolutional Neural Networks (CNN) for the regression problem).

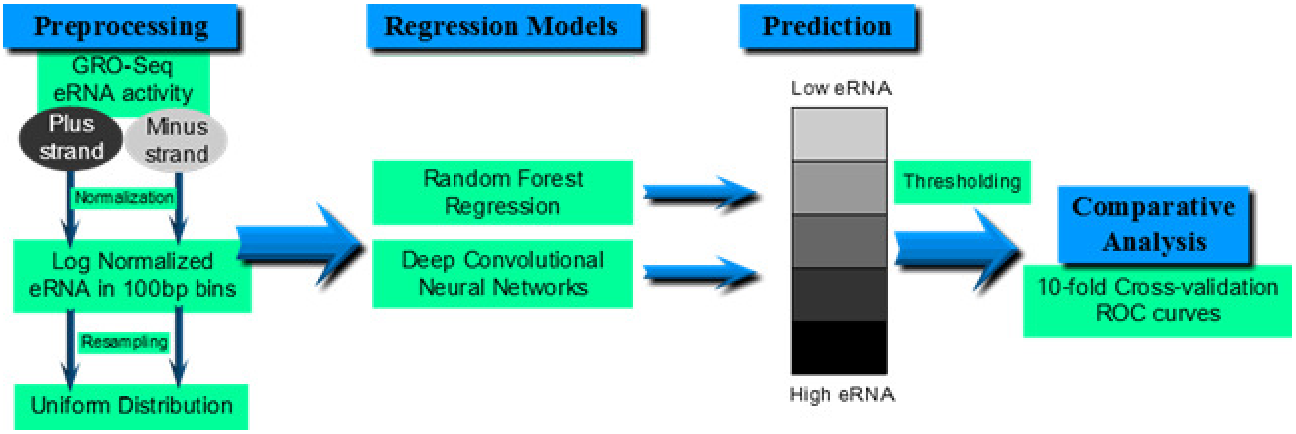

Fig. 1. Overall workflow. Preprocessing step adjusts the characteristics of the dataset to be suitable for the regression models. A suite of regression models is trained. Prediction errors are reported separately for each of five ranges of the eRNA activity level. In the comparative analysis, we transform the regression problem into the simpler classification problem in order to compare against a prior work that uses GRO-seq data for binary classification of eRNA activity.

For our design, we create two regression models—Random Forest Regressor (RFR) and Deep Convolutional Neural Networks (CNN). Through our evaluation, we bring out which classifier performs well for which part of the dataset and then hypothesize the reason behind the performance of each. We then bring out the relative strengths and weaknesses of each regression model toward the use of enhancer-related chromatin features (ChIP-seq) for enhancer activity prediction, as relating to GRO-seq data. It turns out that the best performing regression model is the Random Forest, followed closely by the CNN, which highlights a hitherto unknown fact about the data, namely that it is invariant to location. The property of CNN that it does not overfit even with relatively sparse input data turns out to be useful for our domain. To the best of our knowledge, we are the first to apply CNN-based regression to a computational genomics problem.

In this paper, we make the following contributions:

1. In our study, we use genome-wide locations of enhancers. Thus, our dataset comprises of both intergenic and intragenic (e.g., intronic) enhancers[1]. Ours is thus a prediction model on the most comprehensive set of locations.
2. Our prediction model performs most accurately for the highest eRNA activity levels, and outperforms the baseline by 35%–65% depending on the cell type and the sense of the strand.
3. We perform a detailed comparison with the most relevant prior computational approach for enhancer classification and bring out a detailed ROC characterization by our approach and the prior approach.
4. We also find out what percentage of the active enhancers are bound by some of the important H1 cell type-associated TFs and by the co-activator p300.

The rest of our paper is organized as follows. Section 2 describes the dataset and experimental strategy. Section 3 presents prediction results and insights from them. We follow this with related work and then conclude the paper.

II. Methods

In this section, we describe the datasets, data preprocessing steps, and the machine learning models used. We set out to assay the relationship between chromatin modifications and eRNA activity, as obtained from the more readily available ChIP-seq datasets and the sparser GRO-seq datasets. So, in our regression models, the histone modifications are the input variables and the GRO-seq levels are the output variables. To train, we used ChIP-seq and GRO-seq datasets for the human embryonic stem cell, H1 and a mature human cell type, IMR90. We attempt same-cell prediction (i.e., train on H1, predict on H1, etc.) as well as cross-cell prediction (i.e., train on H1, predict on IMR90, etc.). We consciously choose two very different human cell lines to observe the generalizability of our prediction models. Figure 1 shows the overall workflow of our approach. First, the dataset undergoes a preprocessing step that makes it suitable for our problem. Then, the same dataset is used to train our regression models. Prediction performance is measured in terms of RMSE and median absolute error. Given that the eRNA activity level spans multiple orders of magnitude, we divide the entire range into three bins, and then report RMSE and median error for each bin separately. To provide a comparison with the most relevant prior work, we turn to the easier binary classification problem, given that we are the first to implement a regression suite. We train one of our regression models, CNN, which can be used as a classifier as well, by thresholding the predicted value. The ROC curves for both approaches are built and the AUC scores are compared. Matthew's correlation coefficient (MCC) is also reported.

[1] The ability to distinguish between promoters and enhancers is fuzzy in the dataset because of their similar signatures, and therefore, we do not attempt to distinguish between promoters and enhancers in our predictions. Instead, we focus on overall nascent eRNA predictions from GRO-seq.

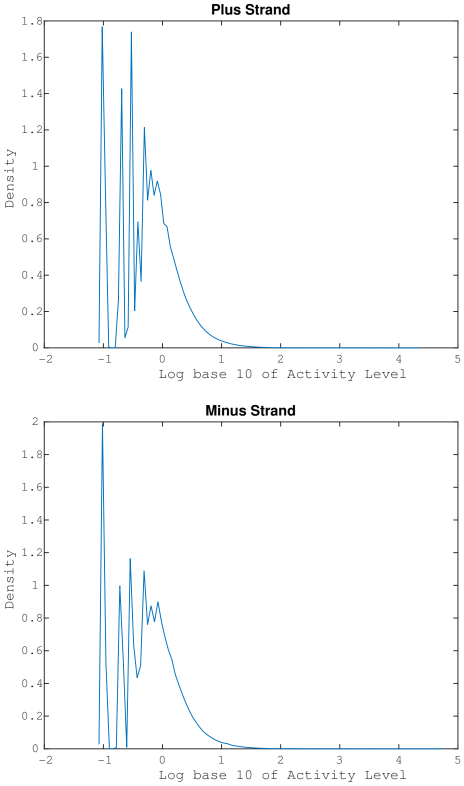

Fig. 2. Distribution of activity level for each strand.

A. *Datasets*

We use four histone modifications as the input features, H3K4me1, H3K4me3, H3K4me2, and H3K27ac. This choice follows recent analysis that shows that these modifications are top predictors for enhancers (Kim, et al., 2015; Whitaker, et al., 2015), also constituting the combinatorial histone code for enhancer activity (Zhu, et al., 2013). The ChIP-seq reads of these histone modifications give us the enhancement level of the modification. We used the RPKM measure (Hawkins RD et. al., 2010) to normalize the replicates, then binned the locations into 100 base pair (bp) intervals.

We download GRO-Seq data for the H1 and the IMR90 cell lines from Gene Expression Omnibus (GEO: GSM1006728). The wiggle file contains normalized read counts mapped into 10 bp intervals. Only intervals with at least 0.1 reads per million are included. We use liftOver to lift the dataset from the hg18 to the hg19 human genome version. Then, we map normalized read counts into 100 bp intervals.

B. *Data Preprocessing*

Figure 2 shows the distribution of activity level in both strands, generated using kernel density estimation (KDE). Due to the values spanning 5 orders of magnitude, using the original scale will result in lower values being drowned out by the few extremely high values. To solve this, we transform the original activity levels into the log base 10 scale.

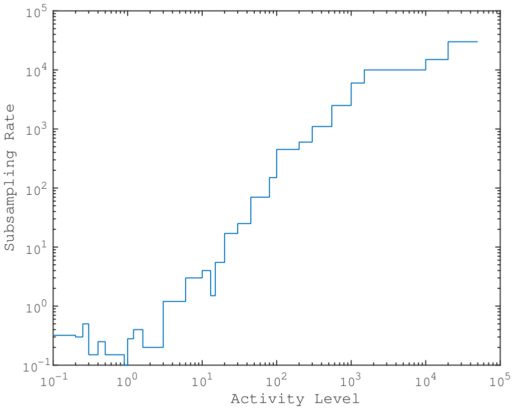

Fig. 3. Subsampling rate for different eRNA activity levels. Subsampling rate of > 1 indicates that samples are duplicated. This is done to compensate for the vastly different densities of data in the different activity levels.

After the log transformation, we are still left with the distribution skewed toward lower values. Models learned using such skewed dataset will focus mostly on the lower ranges, ignoring the higher ranges. This is undesirable. While locations with high activity level are rare, they are biologically important hubs, binding to master TFs, for example, and contributing to key cell identity-controlling genes (Whyte, et al., 2013). Thus, because the entire range is important for us, we need to modify the distribution in the training dataset so that the model focuses on each portion of the range of activity levels, uniformly. This is done by subsampling from the denser ranges, and duplicating samples from the sparser ranges. Figure 3 shows the subsampling rate we use. The reader will observe that the sampling rate is lower in the denser regions and progressively goes higher in the sparser regions, which are the ones with high eRNA activity levels. Finally, we end up with dataset that is uniform across the whole range. Note that this preprocessing step only applies to the training dataset, not to the test dataset. The test dataset only undergoes the log (base 10) transformation. All of the above is done after dividing the dataset into training set and test set, to ensure that there is no overlap between the two.

C. *Machine Learning Models*

The problem of predicting output in the form of real number is referred to as regression. There are many regression models available, ranging from simple curve fitting to complex models such as Deep Convolutional Neural Network (CNN). We selected two models that have been shown to perform well for biological datasets—Random Forest Regression (RFR) and Convolutional Neural Networks (CNN). RFR works by building multiple decision tree regression models, each trained using a slightly different dataset generated using a method called bootstrap aggregating. This helps reduce overfitting which is a common problem when using decision trees. CNNs have been shown to work well in various domains, such as, in image and voice recognition. CNNs are DNNs (Deep Neural Networks) with an additional constraint on the neural network. Instead of individual neurons picking up their own pattern in the data, neurons are grouped together and tiled in a way to cover different parts of the input features in order to capture a single pattern. This leads to better generalization when the relevant patterns are location invariant, which we believe is the case in this problem of predicting eRNA activity levels. A pattern being location invariant means that whether the pattern (of histone modifications in our case) "A-B-C" appears in genic locations 1–3 or some other locations 101–103, they are equally predictive of the eRNA activity levels.

We use the machine learning library scikit-learn for RFR. We use Keras as the frontend and Theano as the backend for CNN.

Hyperparameter optimization: We evaluate the prediction performance of the models using 10-fold cross-validation. Within each fold, the hyperparameter optimization method differs slightly for each model.

For RFR, the number of trees is fixed at 100, and each tree has max depth of 3.

For CNN, the architecture comprises of 32 convolutional filters of size 2 across each mod vector (20 inputs) treating each vector as a channel (for 4 channels total). The next layer uses 32 filters of size 1 row 2 columns, then max pooling is applied using size 1 row 2 columns as well. The output is then flattened and fed to a standard DNN with architecture 128-128 followed by a linear activated output layers. All convolutional and hidden layers use ReLU as activation and are initialized by the Gaussian fan in/fan out method. Dropout rate is .25 for the maxpooling layer and .5 for the two DNN hidden layers with 128 nodes each. The model was trained using Adadelta with a batch size of 128 that optimized RMSE as the loss function. Only one epoch is used for each fold.

III. Results

For our evaluation, we use the dataset with log transformation without resampling as the test set. First, we evaluate the ability of our models to predict the level of eRNA activity based on the histone modification signatures. Second, we convert the regression problem into a classification problem and comparatively evaluate the best of our models—Convolutional Neural Network (CNN) against logistic regression used by Zhu et al. (Zhu, et al., 2013). Third, we do an analysis to answer the question: what fraction of the transcription factor (TF) binding sites correspond to high eRNA activity.

A. *Predicting enhancer activity level*

Here we use CNN to predict the eRNA expression levels, formulated as a regression problem. To present the results in a more fine-grained manner, we bin the expression levels into 3 bins considering the range and different number of samples in each bin. Separate models are built for the sense and the antisense strands of the RNA. The results are shown in Figure 4 and 5. The RMSE is reported separately for each range of activity level. We focus our discussion on the rightmost bucket, corresponding to the strong eRNA expression levels, which in turn corresponds to the most active enhancers. Because of the log transformation, the error is reported in the transformed log (base 10) scale. This means that RMSE of 1 corresponds to the predicted value being 10 times higher or lower than the actual value, while RMSE of 0.5 corresponds to being off by 100.5 = 3.16 times. For comparison purposes, we include a baseline regressor, denoted as ZeroR, which always predicts a value right in the middle of the range of output values. Expectedly, it performs well in the bin that is in the middle of the range, but very poorly for the important class of strong enhancers, i.e., those corresponding to high eRNA expression levels.

Between the three models, RFR (abbreviated as "RF" in the figures) performs best overall. RFR excels in extracting implicit features from the data and further, its training time is minimal because the constituent trees can be trained and executed in parallel. The best-in-class performance of RFR points to the possibility that there is spatial correlation in the data—different histone modification signatures laid out spatially on the "genome line" and close-by epigenetic modifications, such as histone modifications, work in concert to determine the gene expression level.

When a large dataset is being used, such as in our case, a practical concern is how much time is needed to train a model. To this end, we measure the training time of each model presented. For all models we use, prediction time is insignificant compared to training time. The time taken to train a single model is 872 seconds for CNN and 4886 seconds for RFR. For RFR this is the cumulative time and this is reduced by up to a factor of 100 (corresponding to 100 decision trees) through the simple parallelization strategy.

B. *Binary Classification of eRNA Activity Levels*

To provide a comparison to the most relevant prior work (Zhu, et al., 2013), which we refer to as "logistic regression" (LR in the plot), we use our regression model to perform binary classification of the eRNA activity levels. For the prior work, we do not have access to Zhu et al.'s implementation and we therefore create our own implementation of a logistic regression model following the steps outlined in the paper. We use as inputs the same 4 histone modification signatures as for our technique. The output of logistic regression is a probability value that gives the probability of a data point belongs to the positive class. We threshold this probability and sweep through the threshold values to generate a Receiver Operating Characteristic (ROC) for the logistic regression approach. For our approach, we take the CNN regression model and threshold the output to generate a binary classification. We sweep through the values of this threshold to generate the ROC for our approach. A higher value for the Area Under the Curve (AUC) is better and the range is from 0 to 1. A random classifier will achieve AUC of 0.5. Our training and test data both have an equal proportion of data points with zero eRNA activity levels (actually values < 0.1, which are mapped to a zero value due to the quantization effect) and non-zero activity levels. In the real dataset the proportion of zero eRNA activity levels is higher. However, this does not

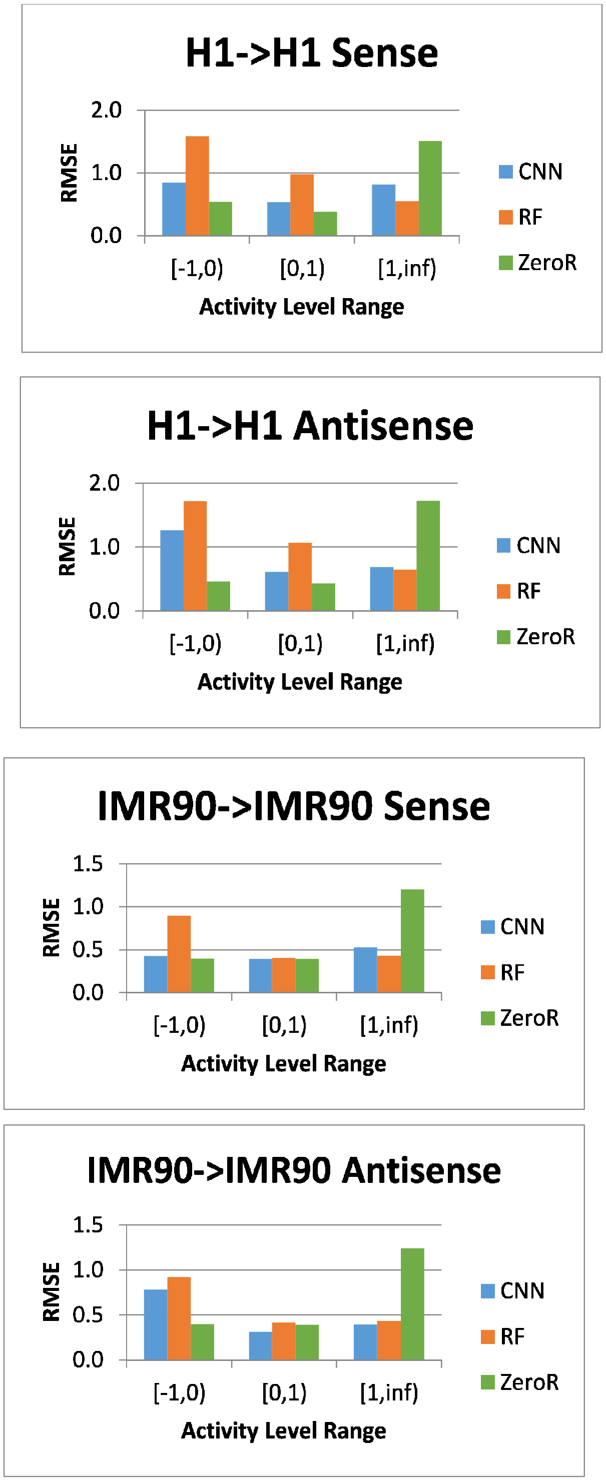

Fig. 4. RMSE of the same-cell prediction using our various regression models. We do this separately for the sense and antisense strands of the RNA. We observe that CNN gives lower RMSE overall, especially in the highest bin which is the most important.

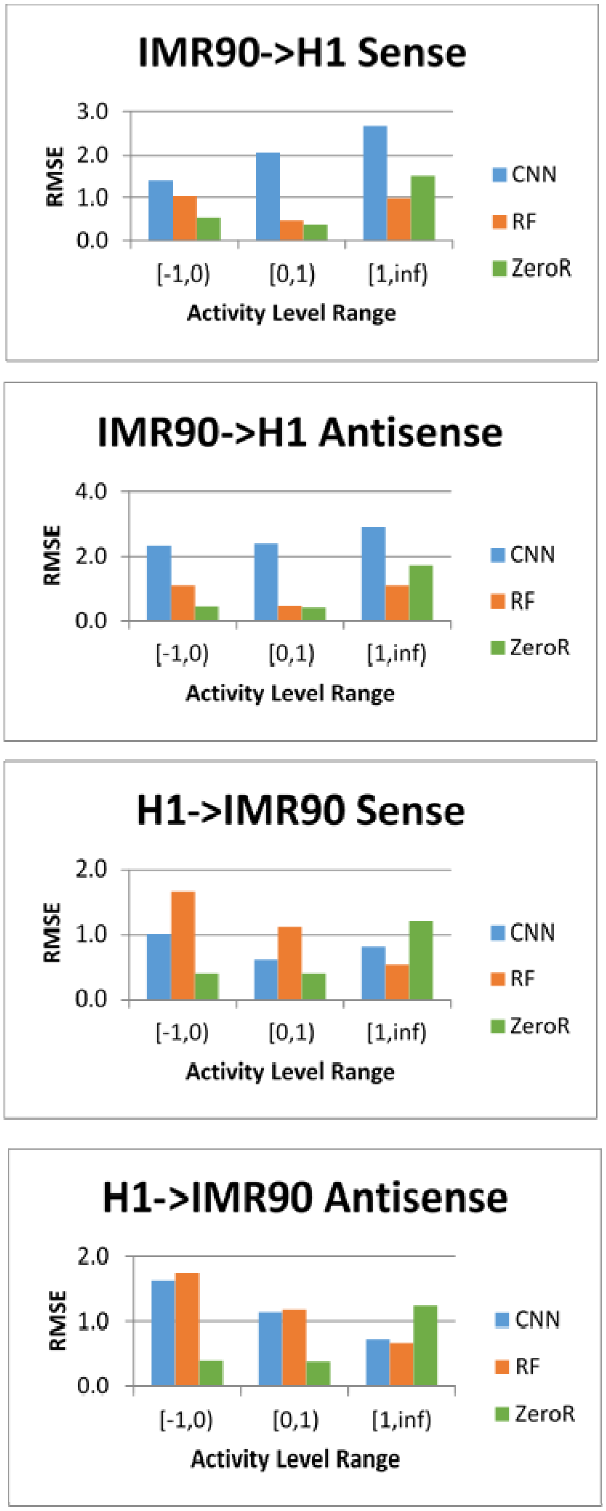

Fig. 5. RMSE of the cross-cell prediction using our various regression models. We do this separately for the sense and antisense strands of the RNA. We observe that CNN gives lower RMSE overall, especially in the highest bin which is the most important.

affect the true positive rate and false negative rate, and thus the ROC curve. In addition to the AUC score, we also report Matthew's correlation coefficient (MCC). MCC ranges from -1 to 1, 1 being the perfect classifier. A random classifier will achieve MCC of 0. The resultant ROC is shown in Figure 6. The prediction performance for both logistic

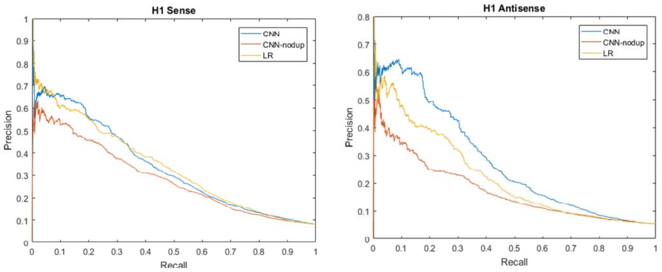

Fig. 6. Precision-recall curve of CNN and LR for classification problem. Models are built separately for the sense and antisense strands of the RNA. We observe that CNN gives comparable accuracy to LR for the sense strand, and outperforms LR for the antisense strand.

regression and CNN is almost identical for the plus and the minus strands. For logistic regression (Zhu, et al., 2013), the AUC is 0.56 and the MCC is 0.19 for both strands, while for our approach (CNN), AUC is 0.75 and 0.76, and MCC is 0.37 and 0.39 for the plus and the minus strands respectively. Thus, based on the AUC, we see an improvement of 34.8% in the classification performance with our approach. Note that the AUC for logistic regression that we obtained is much lower than that reported by (Zhu, et al., 2013), which was 0.935. The difference can be attributed to the fact that we are using a different dataset and we focus more on the high eRNA activity levels, since we are more interested in predicting the high activity enhancers.

C. *Enrichment of Transcription Factor binding in functionally active TREs*

We want to find out the prevalence of transcription factor binding sites, including the co-activator P300, among Transcriptional Regulatory Elements (TREs) classified by their activity. Specifically, we divided our GRO-seq data into 4 clusters by doing k-means on the eRNA activity level and classified them as inactive/not TRE, weakly active, moderately active, and strongly active based on the strength of GRO-seq signal. The plus and minus trands were clustered separately and the TREs were identified as intersection of both sets. Since the signal can be distorted by gene transcription, we only considered intergenic regions.

The ChIP-seq data for 10 transcription factor (TF) binding proteins was downloaded from ENCODE (https://genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html) and peaks were called on these datasets using MACS2. For each binding protein, we estimated proportion of TREs overlapping with ChIP-seq peaks. Figure 7 illustrates these proportions for strong, moderate, and weak TREs as classified by their activity. From this, we find that the proportion varies significantly across these binding proteins. Some of these, such as POL2, TBP, and SIN3A, show a noticeably higher enrichment in most functionally active (strong) TREs than weak or moderate ones. The variation also explains the fact that different proteins bind to different classes of TREs.

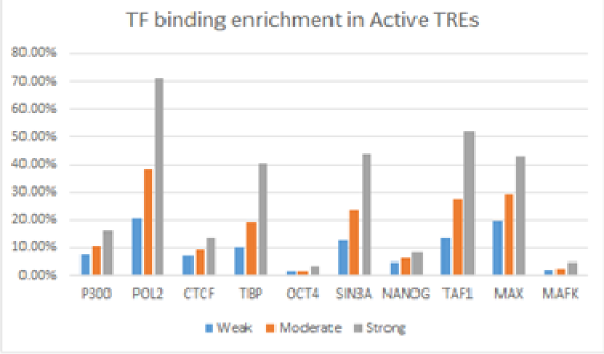

Fig. 7. Enrichment level of Transcription Factors (TF) in functionally active Transcriptional Regulatory Elements (TREs).

IV. Related Work

The computational identification of enhancers has proven challenging because of the large search space, scattered across the non-coding genome (98% of the genome) and the absence of locational signals relative to their target genes (Visel, et al., 2013). Several computational methods that use histone modification signatures to identify enhancer regions have been developed (Kim, et al., 2015; Kim, et al., 2016; Kim, et al., 2016; Rajagopal, et al., 2013; Visel, et al., 2009; Whitaker, et al., 2015), including a range of ML algorithms, such as Hidden Markov Models (HMMs) (Won, et al., 2008), Time-Delayed Neural Networks (TDNNs) (Firpi, et al., 2010), Support Vector Machines (Fernández and Miranda-Saavedra, 2012), and Random Forests (RFs) (Rajagopal, et al., 2013). Use of RFs in RFECS (Random Forest based Enhancer identification from Chromatin States) improved upon the limited number of training samples in previous approaches in order to determine the optimal set of histone modifications to predict enhancers. We recently used a deep neural network (EP-DNN) and made the DNN interpretable by identifying which histone modifications are important to enhancer prediction (Kim, et al., 2015). However, the enhancers predicted as a result of classification via these ML algorithms are not always functionally active (Bulger and Groudine, 2011). Functionally active enhancers are the ones that are actually exposed (spatially) to regulatory factors such that they can be acted upon by the trans-acting genomic regulators, such as TFs and microRNA (Ghoshal, et al., 2015; Ghoshal, et al., 2015), which often act in a coordinated manner (Martinez and Walhout, 2009). Given that recent high-throughput technologies, such as GRO-seq (Core, et al., 2008), afford us the ability to tease apart some of the biochemical indicators of enhancer functionality, in this paper, we wanted to leverage the huge influx of ChIP-seq datasets (Barski, et al., 2007; Consortium, 2012; Wang, et al., 2008) to get a fine-grained indication of the quantitative levels of eRNA activity, the most direct measure of enhancer activity. In doing so, using the classification variant of our ML algorithm (CNN), we beat the only binary classification algorithm that uses logistic regression (Zhu, et al., 2013) and go on to show that ChIP-seq signals can offer unprecedented insights into genomic regulation, enabling the reuse of the exhaustive and high resolution ChIP-seq datasets using sophisticated data analysis techniques.

V. CONCLUSION

Chromosomal DNA is organized into three-dimensional structures that modulate gene expression by bringing together distant, seemingly non-interacting gene regulatory elements, such as, enhancers situated at great distances from their cognate promoters. This three-dimensional conformation can thus orchestrate the fine workings of cellular events in different cell types and conditions. Capturing the effect of the three-dimensional conformation of the DNA without the need to run elaborate experiments can vastly improve the functional annotation of the genome and at lower costs of annotation. While the genome-wide mapping of enhancers itself is a challenging task, add to that the desire to map out the activity levels of enhancers, and the task becomes increasingly complex, yet more rewarding. The synthesis of eRNA points to the actual extent of interaction of enhancers with promoters and the quantitative levels of eRNA synthesis are obtained by GRO-seq. Thus, it seems worthwhile to try to predict the GRO-seq datasets, which are expensive to generate and have uneven sequencing depth, using more available ChIP-seq data. The ability to predict the precise activity levels of enhancers can enable the engineering of synthetic gene circuits with step-like dose response regulatory outputs and the ability to predict more exact response of CRISPR-Cas9 epigenome editing (Chaterji, et al., 2017; Dominguez, et al., 2015; Sander and Joung, 2014).

In this paper, we decipher the relationship between combinatorial histone modification codes and eRNA transcription levels, as given by GRO-seq data. The histone epigenetic signatures are derived from the abundantly available ChIP-seq datasets. We achieve this using regression models—random forests and convolutional deep neural networks. In addition, since such fine-grained estimation of eRNA activity levels has not been done in the past, we used a previous binary classification approach, classifying enhancers into active and inactive enhancers, as our baseline. In doing so, we got AUC of 0.75–0.76. Importantly, however, our primary problem in this paper was not this coarse-grained classification but rather, for the first time, to develop regression models to be able to estimate precise eRNA activity levels, covering the entire genomic range of enhancers and the entire spectrum of enhancer activity levels, ranging from inactive to poised to active enhancers. Specifically, we wanted to use ChIP-seq data as a surrogate for the lower quality GRO-seq data. Our results indicate that we were able to achieve this goal using a combination of data preprocessing to uniformly cover the entire range of enhancer activity levels and sophisticated regression models resulting in an overall RMSE improvements for the most active enhancer class, compaed to the ZeroR model, range from 35%–65% depending on the cell type and same-cell or cross-cell prediction. Such fine-grained prediction of eRNA activity levels may rule out the need to generate extensive amounts of new datasets by extracting the predictive capabilities of existing datasets to the maximum possible extent. Thus, integrating diverse high-throughput sequencing technologies will offer a more panoramic view of the mammalian epigenome and its sophisticated layers of regulation.

REFERENCES

[1] Barski, A., et al. High-resolution profiling of histone methylations in the human genome. Cell 2007;129(4):823-837.

[2] Bulger, M. and Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. Cell 2011;144(3):327-339.

[3] Chaterji, S., Ahn, E.H. and Kim, D.-H. CRISPR Genome Engineering for Human Pluripotent Stem Cell Research. Theranostics 2017;7(18):4445-4469.

[4] Chaterji, S., et al. Federation in genomics pipelines: techniques and challenges. Briefings in Bioinformatics 2017;102.

[5] Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57-74.

[6] Core, L.J., et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics 2014;46(12):1311-1320.

[7] Core, L.J., Waterfall, J.J. and Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 2008;322(5909):1845-1848.

[8] Creyghton, M.P., et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences 2010;107(50):21931-21936.

[9] Danko, C.G., et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nature methods 2015;12(5):433-438.

[10] Dominguez, A.A., Lim, W.A. and Qi, L.S. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. Nature Reviews Molecular Cell Biology 2015.

[11] Fernández, M. and Miranda-Saavedra, D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. Nucleic acids research 2012;40(10):e77-e77.

[12] Firpi, H.A., Ucar, D. and Tan, K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics 2010;26(13):1579-1586.

[13] Ghoshal, A., et al. An Ensemble SVM Model for the Accurate Prediction of Non-Canonical MicroRNA Targets. In, ACM-BCB Best Paper Award. ACM; 2015. p. pp. 403-412.

[14] Ghoshal, A., et al. MicroRNA Target Prediction using Thermodynamic and Sequence Curves. BMC Genomics 2015.

[15] Jacquier, A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nature Reviews Genetics 2009;10(12):833-844.

[16] Kapranov, P., et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 2007;316(5830):1484-1488.

[17] Kim, S., et al. Interpretable Deep Neural Networks for Enhancer Prediction. In, IEEE-BIBM. IEEE; 2015. p. pp. 1-8.

[18] Kim, S., Harwani, M. and Grama, A., Chaterji, S. EP-DNN: A Deep Neural Network-Based Global Enhancer Prediction Algorithm. Nature Scientific Reports 2016;6:1-13.

[19] Kim, S.G., et al. Opening up the blackbox: An interpretable deep neural network-based classifier for cell-type specific enhancer predictions. BMC Systems Biology 2016.

[20] Mahadik, K., et al. Orion: Scaling Genomic Sequence Matching with Fine-Grained Parallelization. In, Supercomputing 2014 (The International Conference for High Peformance Computing, Networking, Storage and Analysis). IEEE; 2014. p. pp. 1-11.

[21] Mahadik, K., et al. Scalable Genomic Assembly through Parallel de Bruijn Graph Construction for Multiple K-mers. In, Proceedings of the 8th ACM International Conference on Bioinformatics,

Computational Biology, and Health Informatics. ACM; 2017. p. 425-431.

[22] Mahadik, K., et al. SARVAVID: A Domain Specific Language for Developing Scalable Computational Genomics Applications. In, Proceedings of the 2016 International Conference on Supercomputing. ACM; 2016. p. 34.

[23] Martinez, N.J. and Walhout, A.J.M. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. BioEssays : news and reviews in molecular, cellular and developmental biology 2009;31(4):435-445.

[24] Maston, G.A., Evans, S.K. and Green, M.R. Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. 2006;7:29-59.

[25] Meyer, F., et al. MG-RAST version 4—lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. Briefings in Bioinformatics 2017;105.

[26] Natarajan, A., et al. Predicting cell-type–specific gene expression from regions of open chromatin. Genome research 2012;22(9):1711-1722.

[27] Natoli, G. and Andrau, J.-C. Noncoding transcription at enhancers: general principles and functional models. Annual review of genetics 2012;46:1-19.

[28] Park, P.J. ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 2009;10(10):669-680.

[29] Rajagopal, N., et al. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput. Biol 2013;9(3):e1002968.

[30] Sander, J.D. and Joung, J.K. CRISPR-Cas systems for editing, regulating and targeting genomes. Nature biotechnology 2014;32(4):347-355.

[31] Theera-Ampornpunt, N., et al. Fast training on large genomics data using distributed support vector machines. In, Communication Systems and Networks (COMSNETS), 2016 8th International Conference on. IEEE; 2016. p. 1-8.

[32] Visel, A., Rubin, E.M. and Pennacchio, L.A. Genomic views of distant-acting enhancers. Nature 2009;461(7261):199-205.

[33] Visel, A., et al. A high-resolution enhancer atlas of the developing telencephalon. Cell 2013;152(4):895-908.

[34] Wang, Z., et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics 2008;40(7):897-903.

[35] Whitaker, J.W., et al. Computational schemes for the prediction and annotation of enhancers from epigenomic assays. Methods 2015;72:86-94.

[36] Whyte, Warren A., et al. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell 2013;153(2):307-319.

[37] Won, K.-J., et al. Prediction of regulatory elements in mammalian genomes using chromatin signatures. BMC bioinformatics 2008;9(1):547.

[38] Yip, K.Y., et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol 2012;13(9):R48.

[39] Zhu, Y., et al. Predicting enhancer transcription and activity from chromatin modifications. Nucleic acids research 2013;41(22):10032-10043.