Discovering Distal Regulatory Elements by Integrating Multiple Types of Chromatin State Maps

Li Teng
Department of Internal Medicine,
University of Iowa, 2289 CBRB,
Iowa City, IA 52242, USA
li-teng@uiowa.edu

Kai Tan
Department of Internal Medicine, Department of Biomedical Engineering,
University of Iowa, 3292 CBRB,
Iowa City, IA 52242, USA
kai-tan@uiowa.edu

***Abstract*— While the annotation of human protein-coding sequences is now fairly comprehensive, the identification of regulatory sequences remains difficult. With higher resolution, fewer artifacts and greater coverage, short-read-sequencing-based technologies have made striking impact on genome research. Given the rapid accumulation of genome-wide chromatin state data, there is a pressing need for computational methods to analyze these data. In this paper, we developed a Multiple Layer Perceptron (MLP) framework to predict transcriptional enhancers by integrating multiple types of chromatin state maps, including histone modifications, DNase I cleavage, and DNA methylation. Comparisons with previous work using known enhancers from three cell types suggest that our algorithm is more robust and has higher precision and sensitivity. We anticipate that the new method will be a valuable tool for genome-wide mapping of various DNA regulatory elements in a wide variety of cell types, tissues and growth conditions.**

Keywords; Next-Generation Sequencing; Gene Regulation; Enhancer; Epigenomics; Machine Learning, Artificial Neural Network.

I. INTRODUCTION

Now a major focus of genomics is to identify all gene regulatory elements within the genome [1]. Among these, transcriptional enhancers are distal-acting elements that orchestrate gene regulation critical for cell lineage specification [2]. Mapping enhancers experimentally is limited by the availability of high-quality antibodies. Recently, genome-wide studies have revealed the existence of striking correlations between the local chromatin modification state and the presence and, possibly, the functional state, of transcriptional cis-regulatory elements [3-5]. In the meanwhile, short-read-sequencing technologies expand previously focused readouts from a variety of DNA preparation protocols to a genome-wide scale and have fine-tuned their resolution to single base resolution [6]. Sequencing-based techniques are making striking impact on studies of genomics and genetics.

Epigenetic mechanisms constrain gene expression by adapting regions of the genome to maintain either gene silencing or gene activity [7]. Notably, the wide range of post-translational covalent modifications of histone tails may convey distinctive regulatory information and confer functional properties on specific genomic sites [8, 9]. Using histone modification information, Firpi. H. et al [10] proposed a computational framework, CSI-ANN, to discover DNA regulatory elements. In CSI-ANN signals of several histone modifications were linearly combined and used as the input to a time-delay neural network. However, combining multiple features into one usually results in loss of information.

Another major epigenetic regulatory mechanism is DNA methylation. It contributes to gene silencing and heterochromatin formation. Methods to locate the sites of DNA methylation at specific loci as well as on a global scale have been developed over the last three decades [11]. Using MethylC-Seq Lister R. et al [12] published the first genome-wide, single-base-resolution maps of methylated cytosines in the human genome recently.

Accessibility of DNA is critical in the control of gene expression and DNase I hypersensitive site (DHS) represents a region of DNA that is relatively accessible to macromolecules. Mapping of DHSs has been used to identify a variety of active cis-regulatory elements including enhancers. In the past it has been limited to analysis of single loci. Recently, progresses have been made in generating genome-wide maps of DNase HS sites using DNase-Seq that identifies sites of DNase I digestion at single base resolution[13].

Different types of chromatin state maps enable a more comprehensive characterization of active cis-elements with base-pair resolution. In this paper we developed a Multi-Layer Perceptron (MLP) framework to integrate multiple types of single-base-resolution chromatin state maps, including histone modifications, DNase I cleavage, and DNA methylation, to predict enhancers genome-wide. To the best of our knowledge, this is the first work that integrates multiple chromatin state maps to predict functional DNA elements. Comparison to previous methods using data on human B, K562 and ES cells suggests that our MLP-based model is more robust and has higher precision and sensitivity.

II. METHOD

A. Data source

1) Histone modification ChIP-Seq data: Raw ChIP-Seq sequencing reads for human ES cells were obtained from NCBI Short Read Archive from reference [14]. For B and K562 cells, we downloaded the data from ENCODE project website [1]. For histone modification

with multiple replicates, we used the record with the largest number of mapped reads.

2) DNase I hypersensitivity data: Raw DNase-Seq sequencing reads for the three cell types were obtained from the ENCODE project website [1].

3) DNA methylation data: Single-base-resolution methylome map of human ES cells was published in [12]. The same type of data is not available for B and K562 cells.

4) Distal p300 peaks used for training: We first selected short distal p300 binding peaks (at least 2.5 Kb away from the closest RefSeq transcription start site (TSS), and peak length < 1Kb) mapped using ChIP-Seq in [12, 15] and by the ENCODE Consortium, respectively. Then we chose those p300 peaks that overlap with computationally prediced enhancers from the PReMod [16] database. Table 1 shows the number of training enhancers for each cell type, respectively.

TABLE I. NUMBER OF DISTAL P300 USED FOR TRAINING.

#	*B*	*K562*	*ES*
p300	717	576	580

5) Background loci: Random genomic loci were used as the background. For each cell type, a set of random loci 10 times the number of the training enhancers was generated. MLP model for each cell type was then trained, respectively.

B. Ehancer Prediction Using Multilayer Perceptron

Figure 1 shows our MLP-based algorithm for predicting enhancers by integrating multiple types of chromatin state maps. There are three main steps, data preprocessing, genome wide enhancer prediction, and post-processing. Each type of chromatin modification data (different histone modifications, DNase I cleavage and DNA methylation) is used as an input feature to the MLP.

1) Data preprocessing: The raw sequencing data for each feature were aligned to the human genome (build hg18) using BOWTIE [17]. The genome was partitioned into 200bp bins, total sequence tag count in each 200bp window was computed for each feature, respectively. For the single-base-resolution methylome map of ES cells, we first calculated a methylation level for each CpG cytosine with at least 10 sequencing reads as the ratio of methylated reads to the total number of reads. For each bin a DNA methylation level is then calculated as the average methylation level of CpG cytosines within the window. Each feature was normalized by a Z-score transformation. Feature value for each genomic locus was represented by the summation of signal values within a 1Kb region. This process makes the algorithm more robust.

2) Multilayer Perceptron: A three-layer feedforward neural network with 20 hidden neurons and one output neuron was implemented. Each MLP model was trained for each combination of features and cell type using 5-fold cross-validation, respectively. The results presented below are based on 10 independent runs of cross-validations.

3) Post-processing: A 2.5Kb filter window moving along the genome was used to make genome-wide predictions. An enhancer prediction is made when the center position of the filter window has an MLP output value above the cutoff (0.5) and represents the highest output value within the window.

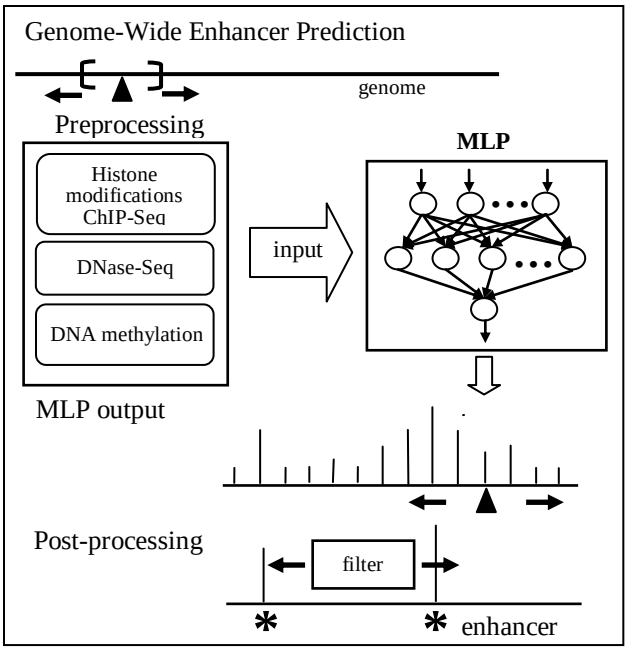

Figure 1. Overview of the MLP-based algorithm for enhancer prediction.

III. RESULTS

A. Comparison of the MLP-based algorithm to CSI-ANN

In CSI-ANN all chromatin modification features were combined into a one dimensional feature using linear combination achieved by Fisher Discriminant Analysis (FDA). Time-delay neural network is then used to make predictions. Combining features usually results in loss of information. And the performance of time-delay neural network is sensitive to the settings.

Table II shows a comparison between the MLP-based method and CSI-ANN. The experiments were carried out using B cell data. CSI-ANN was trained with three histone modifications (H3K4me1, H3K4me3 and H3K27ac) because these histone modifications are known to be associated with enhancers. These three histone modifications were also used to train the MLP model. The MLP-based model was trained with two sets of features, histone modifications only, histone modifications plus DNase I cleavage map, respectively. The performance was evaluated in terms of the average positive prediction value (PPV), specificity, and sensitivity generated from independent cross-validations during training and testing.

Neural network cutoff was set to 0.5 for the experiments reported in the table. As can be seen, the MLP-based method was able to increase the sensitivity by more than 13% when trained with the same set of histone modifications. When DNase I cleavage map was included in the training data the sensitivity was increased by an additional 6%. In addition, the MLP-based algorithm is not sensitive to the cutoff, while CSI-ANN is sensitive to the cutoff (result not shown).

TABLE II. COMPARISION OF PERFORMANCE OF THE MLP-BASED ALGORITHM AND CSI-ANN TRAINED WITH DIFFERENT FEATURES.

Model	**CSI-ANN**	**MLP**	
Features	*Three histone modifications*	*Three histone modifications*	*Three histone modifications +DNase-Seq*
PPV	82.19% (5.24%)	82.23% (5.25%)	82.84% (4.69%)
Specificity	96.32% (3.30%)	97.87% (0.58%)	98.13% (0.60%)
Sensitivity	61.22% (33.4%)	74.67% (6.08%)	80.47% (6.04%)

B. Performance Comparison of Integrating Different Features by the MLP

1) Integrating DNase I cleavage map: We plotted the cumulative distributions of DNase I cut count in 1Kb enhancer regions and 1Kb random loci. Figure 2(a,c) show that B cell and K562 cell enhancers have signifiantly more DNase I cuts than random loci (Kolmogorov-Smirnov test (KS test) p-values are 2.85e-289, 2.39e-254, respectively). For ES cell the difference is less significant with a KS test pvalue of 2.55e-15, as shown in Figure 2(e).

Figure 2(b,d,f) show the performance comparison of the MLP-based algorithm using histone modification alone and using both histone modification and DNase I cleavage map. Experiments were carried out on the three cell types, respectively. Ten rounds of 5-fold cross-validation for each model were performed. Cutoff was set to values from 0.1 to 0.9 with a step size of 0.1. A F1 score (harmonic mean of precision and recall) was calculated for each run of cross-validation. Each point in Figure 2(b,d,f) shows the average of 50 runs. As shown in the figure, at a cutoff of 0.5 the F1 score was increased by 6.35% for B cell, 4.05% for K562 cell, and 1.37% ES cell, respectively. An explanation for the smaller increase in ES cell may be the unique biology of pluripotent cells in which the poised promoters are more abundant but strong enhancers are depleted [17].

2) Integrating DNA methylation map: CpG cytosine methylation is another major epigenetic modification that has essential roles in genome regulation, development and disease. Stadler MB et al [18] have generated base-pair-resolution methylomes for mouse embryonic stem cells and they identified low-methylated regions (LMRs) in which the CpGs present intermediate, yet low levels of methylation in the range of 10%-50%. They indicated that several lines of evidence including genomic position, conservation, chromatin state, regulatory activity and transcription factor occupancy support the hypothesis that LMRs are indeed active distal regulatory regions. In human DNA methylomes [12] similar characteristics about CpG methylation were observed. We expected that enhancers in ES cell also show a lower cytosine methylation level comparing to background.

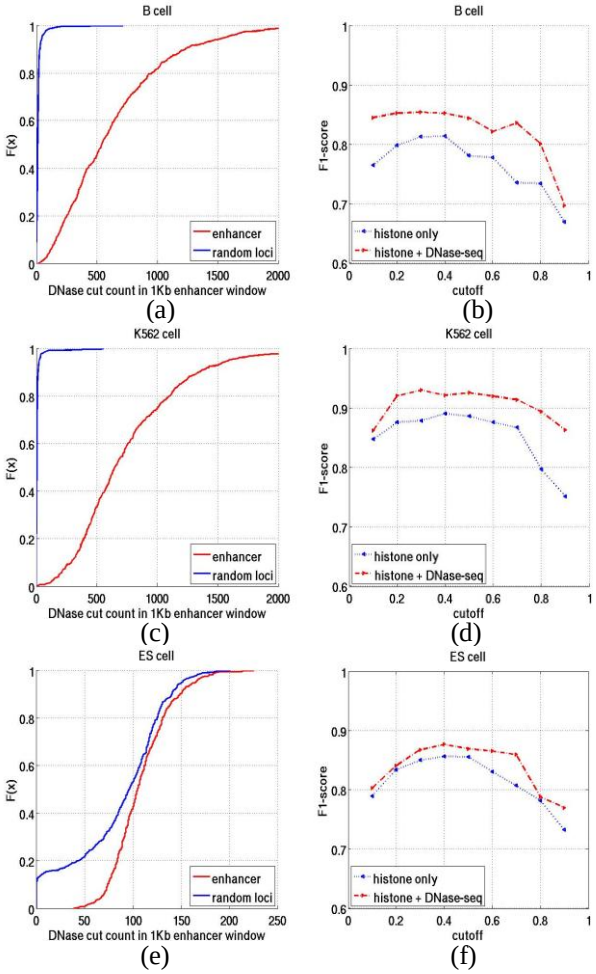

Figure 2. (a,c,e) Cumulative DNase I cut count distributions of enhancers comparing to that of random loci, (b,d,f) Performance comparison of the MLP-based algorithm with histone modifications only and with both histone modifications and DNase I cleavage map for B, K562 and ES cells, respectively.

Using a threshold of 15% and 65% methylation, we determined that 5.72% of the methylome of human ES cell are LMRs, compared to 4.1% of the methylome of mouse ES cell. Most of the methylomes are fully methylated regions (FMR) (94% in human ES cell and 89.4% in mouse ES cell). We found that all enhancers used for training are in LMRs. Figure 3 (a) shows the average DNA methylation level of all enhancers used for training and that of the same number of random loci. DNA methylation level shows a clear dip around enhancer center as the figure shows. Figure 3(b) shows the cumulative DNA methylation levels of enhancers and background loci. The two distributions are significantly different with a KS test p-value of 1.85e-12.

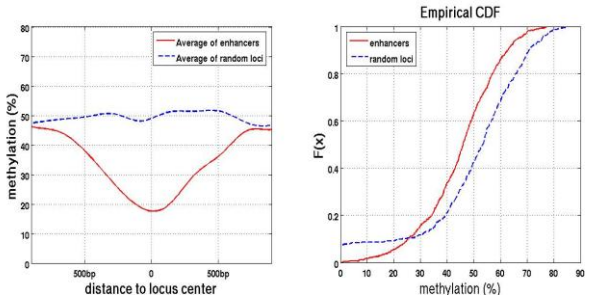

Figure 3. Comparison of DNA methylation levels between enhancers and random loci.

For ES cell we trained the MLP-based model with four different feature combinations, histone modifications only, histone modifications with DNase I cleavage map, histone modifications with DNA methylation map, and all four features. Experiments and F1 score calculation were carried out as described above.

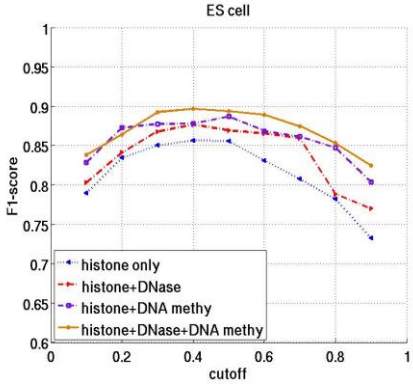

Figure 4. Performance comparisons among MLP models trained with four different sets of features.

As shown in Figure 4 integrating DNA methylation map with other chromatin features significantly improved the performance by an average of 4.29% compared to the model trained with only histone modification maps. This means that DNA methylation level is a useful feature to distinguish enhancers from background. By using all three types of chromatin state maps as input features, the performance was significantly increased by an average of 5.43%. The MLP model that integrates different chromatin state maps not only increases enhancer prediction accuracy but also provides a more robust model. The standard deviation of performance over cross-validations is the lowest among all trained models.

IV. CONCLUSIONS

In this report, we used MLP to integrate multiple types of chromatin state maps to predict enhancer elements in three human cell types. To the best of our knowledge, this is the first work to predict enhancers by integrating multiple types of chromatin state maps. Our result shows that enhancers show different characteristic in each chromatin state map (e.g. higher DNase cut count, lower DNA methylation level) compared to background. These features contribute to enhancer prediction when integrated in the MLP model. Integrating multiple chromatin state maps enables a more comprehensive and accurate characterization of active cis-regulatory elements. We anticipate that the new method will be a valuable tool for genome-wide mapping of DNA regulatory elements in a wide variety of cell types or tissues under diverse conditions.

ACKNOWLEDGMENT

We thank members of the Tan lab for valuable discussions. This study was supported by the National Institutes of Health grant HL073015.

REFERENCES

[1] E. Birney, et al., "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," Nature, vol. 447, pp. 799-816, Jun 14 2007.
[2] M. Bulger and M. Groudine, "Enhancers: the abundance and function of regulatory sequences beyond promoters," Dev Biol, vol. 339, pp. 250-7, Mar 15 2010.
[3] G. E. Zentner, et al., "Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions," Genome Res, vol. 21, pp. 1273-83, Aug 2011.
[4] N.D. Heintzman, G.C. Hon, R.D. Hawkins, P. Kheradpour, A. Stark, et al., "Histone modifications at human enhancers reflect global cell-type-specific gene expression," Nature, vol. 459, pp. 108-112, 2009.
[5] C. M. Koch, et al., "The landscape of histone modifications across 1% of the human genome in five human cell lines," Genome Res, vol. 17, pp. 691-707, Jun 2007.
[6] E. R. Mardis, "Next-generation DNA sequencing methods," Annu Rev Genomics Hum Genet, vol. 9, pp. 387-402, 2008.
[7] R. Jaenisch and A. Bird, "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals," Nat Genet, vol. 33 Suppl, pp. 245-54, Mar 2003.
[8] T. Jenuwein and C. D. Allis, "Translating the histone code," Science, vol. 293, pp. 1074-80, Aug 10 2001.
[9] B. Li, et al., "The role of chromatin during transcription," Cell, vol. 128, pp. 707-19, Feb 23 2007.
[10] H. A. Firpi, et al., "Discover regulatory DNA elements using chromatin signatures and artificial neural network," Bioinformatics (Oxford, England), vol. 26, pp. 1579-86, 2010.
[11] S. J. Clark, et al., "DNA methylation: bisulphite modification and analysis," Nat Protoc, vol. 1, pp. 2353-64, 2006.
[12] R. Lister, M. Pelizzola, R.H. Dowen et al., "Human DNA methylomes at base resolution show widespread epigenomic differences," Nature, vol. 462, pp. 315-321, 2009.
[13] A. Boyle et al, "High-Resolution Mapping and Characterization of Open Chromatin across the Genome," Cell, vol. 132, pp. 311-322, 2008.
[14] R.D. Hawkins, G.C. Hon, L.K. Lee, Q. Ngo et al., "Distinct Epigenomic Landscapes of Pluripotent and Lineage-Committed Human Cells," Cell Stem Cell, vol. 6, pp. 479-491, May 7 2010.
[15] Z. Wang, et al., "Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes," Cell, vol. 138, pp. 1019-31, Sep 4 2009.
[16] V. Ferretti, et al., "PReMod: a database of genome-wide mammalian cis-regulatory module predictions," Nucleic Acids Res, vol. 35, pp. D122-6, Jan 2007.
[17] J. Ernst, et al., "Mapping and analysis of chromatin state dynamics in nine human cell types," Nature, vol. 473, pp. 43-49, 2011.
[18] M. B. Stadler, et al., "DNA-binding factors shape the mouse methylome at distal regulatory regions," Nature, vol. 480, pp. 490-5, 2011.