BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 13 2010, pages 1579–1586
doi:10.1093/bioinformatics/btq248

Genome analysis Advance Access publication May 7, 2010

Discover regulatory DNA elements using chromatin signatures and artificial neural network

Hiram A. Firpi[1], Duygu Ucar[1] and Kai Tan[1,2,*]

[1]Department of Internal Medicine and [2]Department of Biomedical Engineering, University of Iowa, 2294 CBRB, 285 Newton Road, Iowa City, IA 52242, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Recent large-scale chromatin states mapping efforts have revealed characteristic chromatin modification signatures for various types of functional DNA elements. Given the important influence of chromatin states on gene regulation and the rapid accumulation of genome-wide chromatin modification data, there is a pressing need for computational methods to analyze these data in order to identify functional DNA elements. However, existing computational tools do not exploit data transformation and feature extraction as a means to achieve a more accurate prediction.

Results: We introduce a new computational framework for identifying functional DNA elements using chromatin signatures. The framework consists of a data transformation and a feature extraction step followed by a classification step using time-delay neural network. We implemented our framework in a software tool CSI-ANN (chromatin signature identification by artificial neural network). When applied to predict transcriptional enhancers in the ENCODE region, CSI-ANN achieved a 65.5% sensitivity and 66.3% positive predictive value, a 5.9% and 11.6% improvement, respectively, over the previously best approach.

Availability and Implementation: CSI-ANN is implemented in Matlab. The source code is freely available at http://www.medicine.uiowa.edu/Labs/tan/CSIANNsoft.zip

Contact: kai-tan@uiowa.edu

Supplementary Information: Supplementary Materials are available at *Bioinformatics* online.

Received on March 26, 2010; revised on April 30, 2010; accepted on May 5, 2010

1 INTRODUCTION

Cis-acting regulatory DNA elements, such as promoters, enhancers and insulators play an essential role in establishing precise temporal and tissue-specific gene expression patterns. Systematic and precise mapping of these regulatory DNA elements, especially enhancers, is a prerequisite for understanding gene expression programs in both healthy and diseased cells. Experimentally, enhancers can be mapped using the powerful technique, chromatin immunoprecipitation coupled with microarray chip (ChIP-Chip) (Kim and Ren, 2006) or short-read sequencing (ChIP-Seq) (Park, 2009). However, this approach is limited by the availability of a large number of chIP-grade antibodies specifically recognizing the transcription factors (TFs) of interest. On the other hand, enhancers can be computationally predicted based on the observation that they often contain dense clusters of TF binding sites (TFBS) in a short stretch of DNA (<1000 bp) and are often conserved. Methods relying on clustering of TFBS (Frith *et al.*, 2003; Pennacchio *et al.*, 2007; Sinha *et al.*, 2003) require prior knowledge of the binding specificities of the TFs involved which is still quite limited. Methods based on sequence conservation (Blanchette *et al.*, 2006; King *et al.*, 2005; Visel *et al.*, 2008) require precise alignment of regulatory DNA sequences from multiple species, which is not necessarily true for all elements.

Histone proteins in chromatin are subject to a number of covalent modifications, primarily at their N-terminal tails, including methylation, acetylation, phosphorylation, ubiquitylation and ADP-ribosylation. These chromatin modifications have profound influences on gene expression (Schones and Zhao, 2008). Numerous genome-wide ChIP-Chip/Seq studies have provided data on the distribution of histone modifications in various model organisms and cell types. A picture is now emerging in which distinct genomic regions such as enhancers, promoters and gene bodies (both protein coding and non-coding RNA genes) have distinct histone modification signatures (Heintzman and Ren, 2009; Schones and Zhao, 2008). For example, high levels of histone 3 lysine 4 methylation have been found at gene promoters and at many enhancers (Heintzman *et al.*, 2007, 2009; Wang *et al.*, 2008). In addition, it has been shown that many regulatory elements carry these epigenetic modifications only in specific cell/tissue types or according to environmental conditions, which cannot be determined by comparative genomics based on sequence alone. Collectively, these observations suggest that epigenetic signatures could be an alternative and powerful way to pinpoint regulatory DNA elements in the genome.

Given the rapid growth of genome-wide chromatin modification data from different species and cell types, there is now a pressing need for computational tools capable of integrating various histone modification maps to discover regulatory DNA elements. Recently, several groups have started to develop computational tools to address this need. Heintzman *et al*. (2007) were the first to develop a computational tool for predicting promoters and enhancers in HeLa cells using six histone modification maps covering 1% of the human genome. Their algorithm predicts promoters and enhancers based on correlation to the average histone modification profiles trained on known examples. In spite of the success of the profile-based method, it is limited in two aspects: (i) the contribution of each histone modification mark to the classification method

*To whom correspondence should be addressed.

and their interdependency was not examined; (ii) the window size of histone modification patterns (10 Kb) was chosen arbitrarily. To improve the profile-based approach, Won *et al.* introduced a Hidden–Markov Model (HMM) based method and used simulated annealing to optimize the window size and the choice of histone modification marks (Won *et al.*, 2008). Evaluated using a set of known enhancers, the HMM-based and profile-based method achieved a positive predictive value [PPV = TP/(TP + FP)] of 54.8 and 53.0%, respectively, and a sensitivity [Sn = TP/(TP + FN), where TP is true positives, FN is false negatives] of 74.1 and 68.9%, respectively. Besides these two supervised learning-based approaches, Hon *et al.* introduced an unsupervised approach to identify histone modification signatures by aligning segments of histone modification data (Hon *et al.*, 2008). Using the same set of known enhancers, they showed that the unsupervised method achieved a sensitivity of 53.5%. No PPV was reported for the unsupervised method.

Although the success by these previous methods is encouraging, there is still a large room for improvement judging by the PPV and sensitivity values reported above. Part of the reason for the limited success of previous methods is that they do not fully employ the signal from the ChIP-Chip/Seq data. From a pattern recognition point of view, by improving the feature extraction step in these methods, one can ensure that potentially important signals are not missed. For instance, besides amplitude, the shape of the signal peaks (broad versus narrow, symmetric versus asymmetric, etc.) could also be very informative for distinguishing enhancers from other types of functional DNA elements and among different types of enhancers. With this in mind, we hypothesize that by introducing efficient data transformation and feature extraction procedures before classification, we can increase the overall accuracy of a classifier for predicting regulatory DNA elements using genome-wide chromatin signatures.

2 METHODS

2.1 Data source and processing

2.1.1 HeLa cell ENCODE data Six histone modification ChIP-Chip data (H3, H3Ac, H4Ac, H3K4me, H3K4me2 and H3K4me3) were obtained from Heintzman *et al.* (2007). Genomic coordinates of 74 training enhancers were obtained from the same study.

2.1.2 Human CD4$^+$ T cell data T-cell histone modification ChIP-Seq data were obtained from Wang *et al.* (2008). There are 39 histone modifications. To create a high confidence training set of enhancers, we first selected distal p300 binding peaks (2.5 Kb away from known RefSeq TSS) mapped using ChIP-Seq in Wang *et al.* (2009). We then chose those p300 binding peaks whose length is <1 Kb to better narrow down the binding site centers. From this set of distal and narrow p300 peaks, we chose those that overlapped with computationally predicted enhancers from the PreMod database. The resultant set contains 213 enhancers.

2.1.3 Background sequences We chose a set of random genomic loci as the background sequences. For both HeLa and CD4$^+$ T cells, the number of random sequences is 10 times that of training enhancer sequences.

2.1.4 Histone modification data preprocessing The preprocessing of HeLa cell ChIP-chip data was implemented as described in Won *et al.* (2008). Briefly, normalized ChIP-chip probe log ratios within a 100-bp window were first averaged. For T cell data, total ChIP-Seq tag counts in a 200-bp window were computed. To normalize across different chIP-Seq experiments, total tag counts in each 100 bp window were divided by the total tag counts in a given histone modification dataset. For both HeLa cell ChIP-Chip and T cell ChIP-Seq data, a linear interpolation was then applied to account for missing data.

2.1.5 Feature value calculation The resolution of preprocessed histone modification data is 100 bp. For prediction, we use an optimal input window size of 2000 bp according to our experiments as well as Won *et al.* (2008). Preprocessed data are converted to feature values using two functions: mean and energy,

$$y_m^j = \frac{1}{n}\sum_i^n x_j(i) \tag{1}$$

$$y_e^j = \sum_i^n x_j^2(i) \tag{2}$$

where $x_j(i)$ is the preprocessed signal for histone modification type j at position i in the input window and n is the number of data points in each window ($n = 20$).

2.1.6 Bhattacharyya distance Bhattacharyya distance (BD) measures the separation of two probability distributions. It is defined as

$$D_B(p,q) = -\ln(\mathrm{BC}(p,q)) \tag{3}$$

where p and q are probability distributions on the same domain X,

$$\mathrm{BC}(p,q) = \sum_{x\in X}\sqrt{p(x)*q(x)} \tag{4}$$

is the Bhattacharyya coefficient. The range of D_B is between 0 and positive infinity and larger values indicate better separation of two distributions.

2.2 Fisher discriminant analysis

Fisher discriminant analysis (FDA) is a feature extraction/reduction technique for finding a linear combination of features that optimally separate two classes. The measure for separation between two classes (i.e. Fisher discriminant ratio) is defined to be the ratio of the variance between the two classes and variance within the classes:

$$S = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} \tag{5}$$

To normalize variations across different histone modification data, feature values calculated from individual histone modification data are first Z-score transformed before they are used as input to FDA. The output of the FDA is a one dimensional feature value.

2.3 Time-delay neural network

For classification, we implemented a time-delay neural network (TDNN). Figure 2B shows the general architecture of the TDNN used in this study. It consists of one input layer, one hidden layer and one output layer. Unlike feed-forward neural networks, nodes (termed time-copies) in the hidden layer of a TDNN receives partial signal of the input. The number of data points (termed receptive field) per hidden node time-copy is controlled by the number of delays plus one. During training, weights in each time-copy of the hidden layer are constrained to be the same. We used the Particle Swarm Optimization (PSO) technique to train the neural network (see Supplementary Materials for detail). We tested different number of hidden layer nodes and delays to find the optimal TDNN structure. A sigmoid activation function was used at each node.

$$y = \frac{1}{1+e^{-x}} \tag{6}$$

Therefore, the output of the network is a bounded value between 0 and 1.

2.4 Cross validation study

For the cross validation study, we used 74 enhancers and 106 promoters obtained from Heinztman *et al*. (2007). We implemented a 5-fold cross validation and ran the experiment 30 times to generate results presented in Table 1. During each cross validation run, information of the test data was not used in the training step. This cross validation experiment was designed to be the same as in Won *et al*. for comparison purposes.

2.5 Genome-wide enhancer predictions

To make genome-wide predictions, we use a sliding window of 2.5 Kb. Consecutive windows are overlapped by 1.25 Kb. For each 2.5 Kb window, we first identify the TDNN output at the center position of the window. A prediction is made if the TDNN output value is above the cutoff value of 0.49. This cutoff is chosen based on achieving a PPV close to 66% which was the PPV obtained using the smaller but better characterized HeLa cell ENCODE data.

3 RESULTS

3.1 Data transformation enables better separation of enhancers from background

In machine learning, data transformation is a powerful means to reveal hidden patterns in complex and high-dimensional data that might not be apparent in the untransformed data. It has been widely used to analyze a variety of biomedical data, for instance, Fourier transformation of nuclear magnetic resonance (NMR) data (King and Kuchel, 1994), principle component analysis of gene expression microarray data (Raychaudhuri *et al.*, 2000), wavelet transformation of DNA sequences to identify functional elements and nucleosome positions (Thurman *et al.*, 2007; Zhang *et al.*, 2008). To the best of our knowledge, the utility of data transformation has not been adequately examined in the case of histone modification data. All previous approaches only used the average signal (arithmetic mean) within an input window. Although the moving-average approach is useful for finding chromatin signatures around functional DNA elements, it misses many aspects of the signal such as the shape of the signal distribution which could be used to distinguish different modification patterns from the same epigenetic mark.

In this study, in addition to average signal, we use a second mathematical function, energy, to transform raw histone modification data (see Section 2). Energy is a commonly used data transformation function in signal processing (D'Alessandro *et al.*, 2003; Smart *et al.*, 2007). It highlights the amplitude difference in the data. Using mean and energy functions, each histone mark data is then transformed into two sets of feature values. To evaluate whether data transformation is useful for distinguishing enhancers from background sequences, we compared the feature value distributions at the two classes of sequences. We used 74 known enhancers and 740 randomly selected background sequences. Both are from HeLa cell ENCODE region (Birney *et al.*, 2007). We then used BD (Kailath, 1967) to measure the separation between the two distributions (see Section 2). As shown in Figure 1, for some histone marks, such as H3K4me3, H4Ac and H3, the transformed data using energy function is better at separating enhancers from background sequences than arithmetic mean. Therefore, by data transformation, additional signal could be extracted to help the classification task.

3.2 Feature extraction from histone modification data

Depending on the number of histone marks and feature functions used in data transformation, we could have a large feature space.

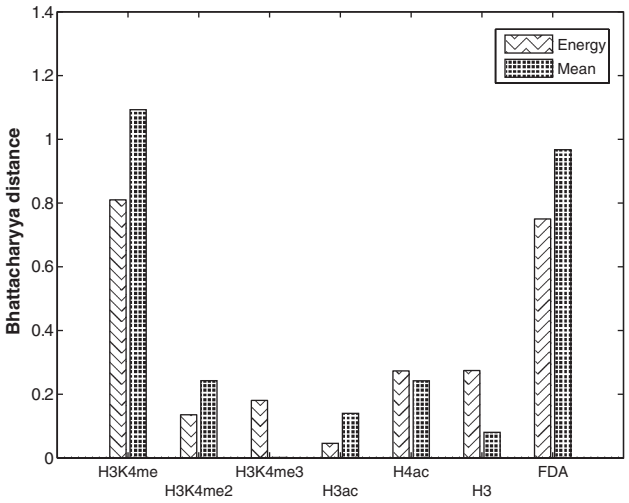

Fig. 1. Discriminative power of data transformation and feature extraction. BD between enhancers and random sequences are calculated using mean (stripe) and energy (square) signals for each histone mark as well as an artificial feature processed through Fisher Discriminative Analysis.

Some of the features may be irrelevant. For instance, several recent reports have shown that various functional DNA elements are associated with different *subsets* of histone modifications (Heintzman *et al.*, 2009; Khalil *et al.*, 2009; Wang *et al.*, 2008). In a pattern recognition system, feature extraction is arguably the most important step (Duda, 2000). It is the stage where the input data is analyzed with regard to a set of features and dimensionality of the feature space is reduced, i.e. non-relevant features are discarded and relevant ones are kept for further analysis. Given these considerations, a feature extraction step will be beneficial for pinpointing relevant histone modifications.

In order to generate a combination of features that provides the best discrimination between enhancers and background while reducing noise, we perform an FDA (see Section 2) on the set of 12 features [6 histone marks $\times$ 2 functions (mean and energy)] for the HeLa ENCODE data. FDA linearly combines multiple features and produces a one-dimensional feature that is optimized for pattern classification purposes (Duda, 2000). It reduces the dimensionality of the data, which helps to improve classification accuracy because classifiers perform poorly when the size of the feature space is hyper-dimensional (curse of dimensionality). Because only one feature will be used in a classifier, FDA also serves to reduce the number of parameters used in the classifier. As shown in Figure 1, the FDA feature is equally capable of discriminating enhancers from background sequences as the most discriminative single feature H3K4me. The FDA feature is then used as the input to an artificial neural network (ANN) classifier.

3.3 A TDNN classifier

There are many types of statistical classifiers one can use for prediction purpose. We used an ANN in this study since it can be trained with a discriminative criterion and can handle non-linearity in the data (Haykin, 1998). The following features have to be learned in order to accurately extract chromatin signatures by an ANN: (i) the neural network has to recognize peaks that may occur at non-fixed

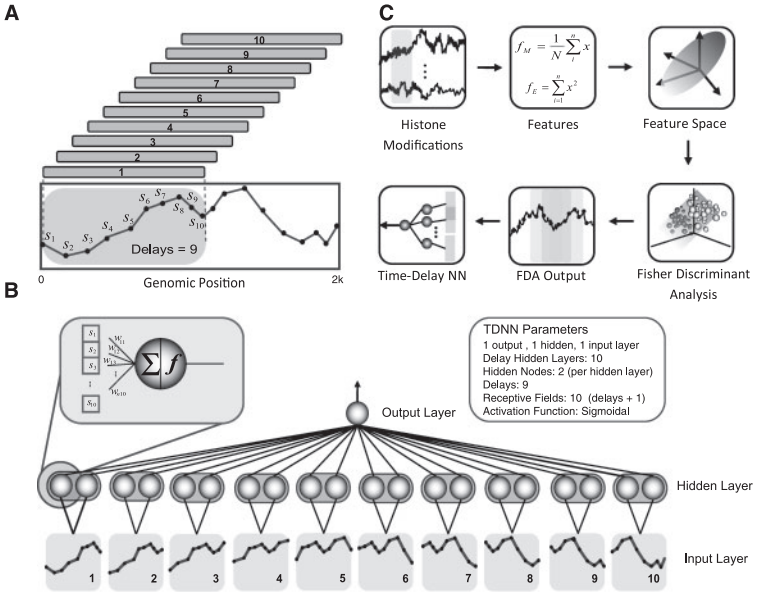

Fig. 2. Overview of the CSI-ANN framework. (**A**) An input data window with shifted sub-windows. (**B**) TDNN. A TDNN with one hidden layer and one output layer. Each time-copy layer (from 1 to 10) receive a portion (10 points or 1000 bp segment) of the entire 2 Kb window (20 points in total); each sliding window is denoted by a numbered light-gray bar. Each bar was slid one point for each time-copy (set of weights is the same for all time copies). (**C**) Flow chart of the CSI-ANN framework.

positions in the input window, i.e. the network has to learn that the peak is a feature independent of shifts in its position; (ii) the network has to recognize features even when they appear at different relative positions. This situation arises in cases where different features (e.g. histone marks) occur in the input window with different relative distances. Classical feed-forward ANNs have problems in solving these spatial dependencies inherent in epigenomic data. To address these two issues, namely the shifting signal and the varying relative positioning of patterns in the signal, we propose to use the TDNN (Fig. 2A and B). This architecture was originally designed for processing speech signal with local time shifts (Waibel *et al.*, 1989). It addresses both problems by imposing certain restrictions on the network topology and by the way in which weights are updated (see Section 2).

The classic training algorithm for neural networks, the back propagation algorithm, has a very slow convergence rate and can get stuck in local minima. In this study, we implemented the PSO algorithm for training TDNN. PSO uses a population of potential solutions, called a swarm of particles, that explores a search space of fitness values for an optimal value (Eberhart, 2001). The algorithm's evaluation continues until either a stop condition is met, denoting convergence to a globally optimal particle, or the maximum number of generation is reached (see Supplementary Methods for details). Because of its stochastic and population nature, PSO is faster and less likely to be trapped in local minima compared to back propagation.

To determine the optimal architecture for the TDNN, we tested a range of delay values and a number of hidden layer nodes using ENCODE data from HeLa cell (see Supplementary Materials). We also tested different input window lengths: 1, 2, 3, 6 and 10 Kb (Supplementary Tables S1, 2 and 4). As shown in Supplementary Table S4, using a 2 Kb window, a TDNN with a delay of 9, 2 hidden layer nodes and 1 output node achieved the best performance when evaluated using the HeLa cell ENCODE data. Therefore, we used this TDNN architecture for the rest of our study.

3.4 CSI-ANN framework for predicting enhancers

The overall workflow of our computational framework, termed Chromatin Signature Identification by Artificial Neural Network (CSI-ANN), is shown in Figure 2C. It contains three components: (i) data transformation of preprocessed histone modification data; (ii) feature reduction by FDA; (iii) classification by TDNN. Figure 3A depicts the neural network output for the set of 74 known enhancers and 740 background sequences in HeLa cell. It shows a clear separation of most enhancers from background sequences. Figure 3B shows the input histone mark signals for a selected enhancer and the average signals for the set of enhancers and background sequences.

We compared the performance of CSI-ANN with two existing approaches for predicting enhancers using histone modification data: the profile-matching based method by Heintzman *et al*. (2007) and the HMM-based method by Won *et al*. (2008). Following Won *et al*., we first conducted 30 independent 5-fold cross validations using the same set of 74 enhancers and 6 histone modification datasets as in Won *et al*. Table 1 shows the results of the comparison. Using two histone modifications, CSI-ANN achieved a 2% improvement in PPV over the HMM method. Using six histone modifications, CSI-ANN achieved a slightly better improvement in PPV (2.24%) compared to the HMM method.

To further characterize the performance of CSI-ANN on larger genomic regions, we trained CSI-ANN using the same 74 enhancers as in the cross validation study and made predictions across the entire ENCODE region in HeLa cell. To verify the computationally

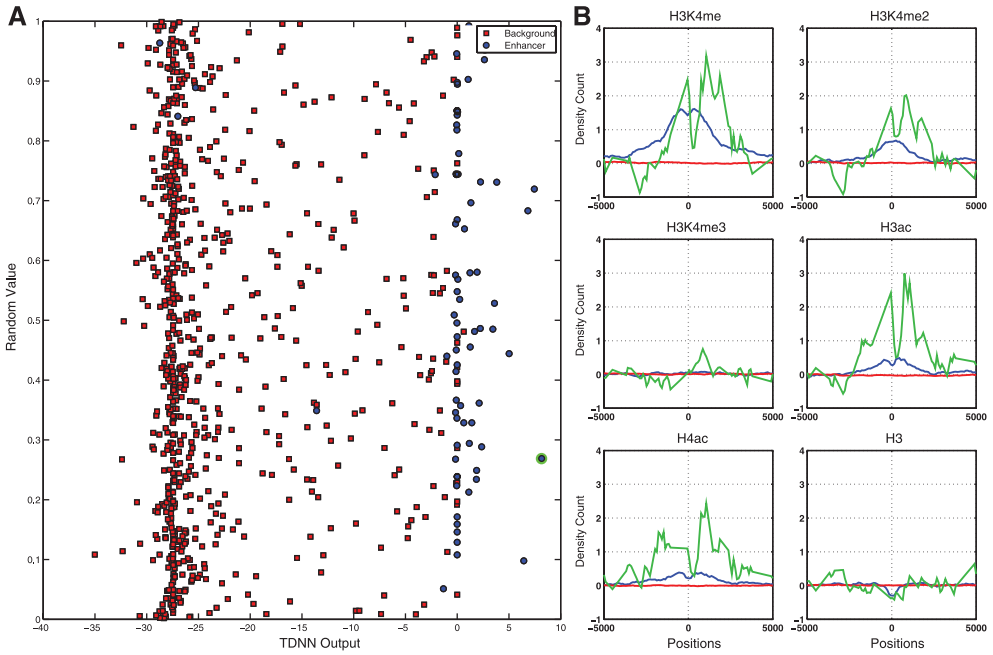

Fig. 3. CSI-ANN outputs for enhancers and background sequences. (**A**) CSI-ANN outputs of 740 background sequences (red squares) and 74 training enhancers (blue circles). (**B**) Histone modification signals (green line) at an example enhancer (green circled on left). Also shown are average signal for each histone mark at enhancers (blue line) and background sequences (red line).

Table 1. Enhancer prediction cross validation results

Method (input data)	PPV (±SD)
PM (6 histone marks)	78%
PM (2 histone marks)	85%
HMM (6 histone marks)	93.52% (±1.83%)
HMM (2 histone marks)	94.06% (±0.89%)
CSI-ANN (6 histone marks)	95.76% (±1.77%)
CSI-ANN (2 histone marks)	96.22% (±2.14%)

PM, profile-matching method by Heintzman *et al.*; HMM based method by Won *et al.*; CSI-ANN, Chromatin Signature Identification by ANN.

Table 2. Enhancer predictions in ENCODE region of untreated HeLa cell

Marker	PM (%)	HMM (%)	CSI-ANN (%)
p300 ($n=94$)	77 (81.91)	82 (87.23)	79 (84.04)
DHS ($n=587$)	165 (28.11)	179 (30.49)	243 (41.40)
TRAP220 ($n=76$)	43 (55.84)	47 (61.04)	54 (71.05)
DHS or TRAP220 or p300	206 (52.96)	213 (54.76)	258 (66.32)

For rows p300, DHS and TRAP220, values in parenthesis indicate sensitivity. For the last row, values in parenthesis indicate PPV.

predicted enhancers, we used location data of three markers across the ENCODE region: p300, DnaseI Hypersensitivity Sites (DHS), and TRAP220 as described in Won *et al*. p300 is a transcriptional co-activator that binds to enhancers. TRAP220 is a component of the mediator complex that has been shown to bind to enhancers as well as promoters. DHSs are nucleosome-free regions that are often occupied by enhancers. We only considered p300, TRAP220 and DHS sites that are distal (>2.5 kb) from any Transcription Start Site (TSS) to avoid confusion with promoters. For comparison purpose, we first set the output threshold of our neural network to make 389 predictions, the same number of predictions as reported in Heintzman *et al.* and Won *et al*. Table 2 shows the sensitivity and PPV for each of the three methods. CSI-ANN obtained a large sensitivity gain based on DHS and TRAP220 marker (10.9 and 10.0%, respectively) compared to the HMM method with a small loss of sensitivity based on p300 marker (3.2%). Overall, CSI-ANN achieved a sensitivity gain of 5.9% over the HMM method when averaged over all three markers. In addition, CSI-ANN achieved a PPV of 66.32% (258/389), 11.6% more than that of the HMM method (54.76%, 213/389).

Figure 4 plots the number of predictions supported by at least one enhancer marker against the total number of predictions at various classifier thresholds. As can be seen from this figure, across the range of 161–420 total predictions, CSI-ANN has a higher rate of true positives than the other two existing methods. In addition, CSI-ANN using only 6 histone marks outperformed the HMM method using 10 histone marks.

3.5 Genome-wide enhancer prediction in human CD4$^+$ T cell

Next we applied CSI-ANN to identify enhancers in the human CD4$^+$ T cell genome. We used a set of 39 histone modification ChIP-Seq data published by Wang *et al*. (2008), including 18 histone acetylation marks, 19 histone methylation marks and the histone variant H2A.Z (see Section 2). For the TDNN, we used the same architecture as in the HeLa cell ENCODE region study (delay = 9, 2

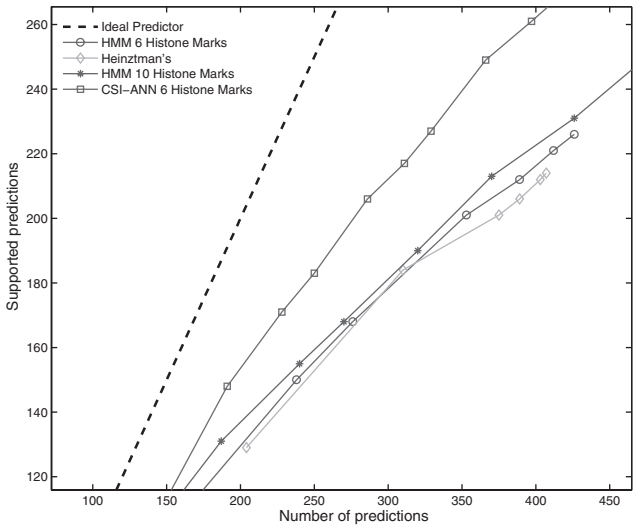

Fig. 4. Number of true positives as a function of total predictions. Dashed line, ideal predictor; diamond, template matching method in Heintzman *et al*. using two histone marks; circle, HMM-based method of Won *et al*. using six histone marks; star, HMM method using 10 histone marks; square, CSI-ANN using six histone marks.

hidden layer nodes per time-copy) since it gave the best performance (Supplementary Table S3). To train the neural network, we used a set of 213 high-confidence enhancers (see Section 2) and 2130 background sequences. Input data windows in the training set are 2 kb long. In total, CSI-ANN made 36 769 predictions using the set of 39 histone modification data. CSI-ANN was able to identify known T-cell specific enhancers. For instance, it correctly located several previous characterized enhancers around the *Ifng* (interferon gamma) gene locus that exhibit T-cell specific expression (Fig. 5A).

To further validate the set of predicted enhancers, we conducted a series of computational analyses as described below in details. Overall, these analyses are designed to examine whether complementary lines of evidence exist to corroborate the predicted enhancers by taking advantage of a diverse set of published data.

We first evaluated the predicted enhancers using genome-wide p300 location (4826 sites) (Wang *et al.*, 2009) and DHS data (73 898 sites) (Boyle *et al.*, 2008) in human CD4$^+$ T cell. Overall, 13.1% of our predictions are supported by both p300 and DHS markers and 69.2% by at least one of the two markers (Fig. 5B).

Next, we corroborated our predictions using sequence conservation and TFBS information. First, we observed that 8124 (22.1%, $P < 10^{-4}$) of our predictions are conserved over a set of 17 vertebrate genomes (see Supplementary Methods). Next, we examined if our predictions are enriched for known TFBS. PreMod is a database of computationally predicted enhancers in human genome using a set of more than 500 TF DNA binding motifs (Ferretti *et al.*, 2007). We computed the overlap between

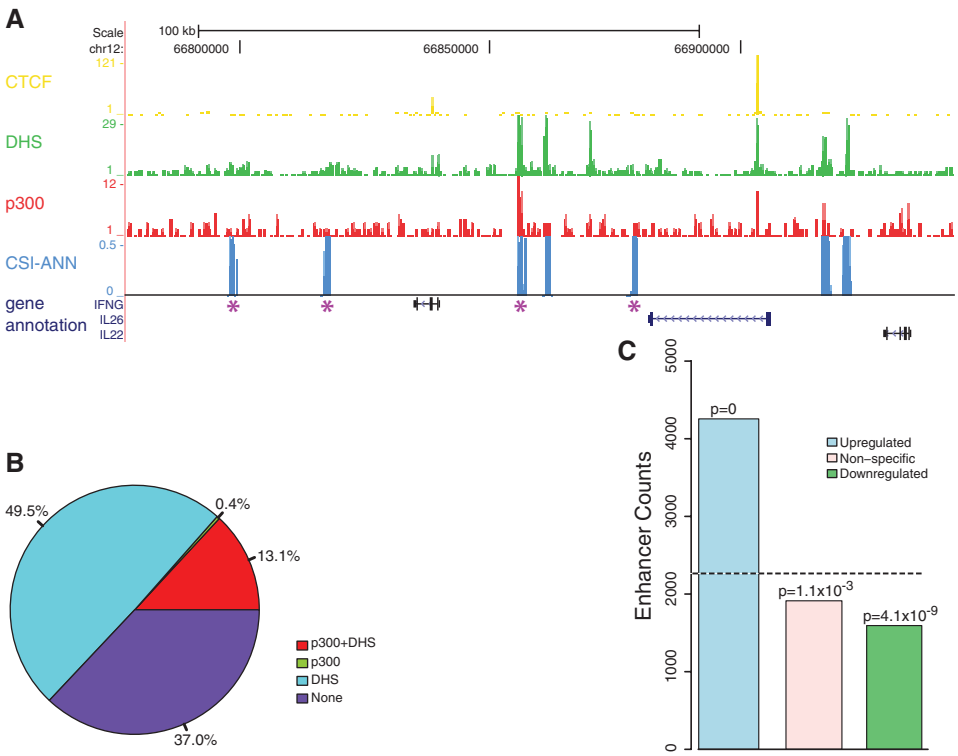

Fig. 5. Enhancer predictions in human CD4$^+$ T cell. (**A**) Histone modification profiles at a known T-cell specific locus, the interferon gamma gene locus, *Ifng*. Known enhancers recovered by prediction are indicated with a purple star. (**B**) Overlap of predicted enhancers with additional supporting genomic markers, p300 and DHS. (**C**) Enrichment of predicted enhancers associated with T-cell specific upregulated genes (blue) compared to downregulated genes (green), and genes with no expression change (pink) in T cell. Dashed line, average number of enhancers associated with randomly selected set of genes.

PreMod enhancer predictions and our predictions. We found 9037 (24.6%, $P < 10^{-4}$) overlapped predictions. In total, 26 662 (72.5%) of our predicted enhancers are supported by at least one of the four lines of evidence (i.e. p300, DHS, PreMod and sequence conservation) (Supplementary Table S5).

A hallmark of enhancers is their tissue-specific activity. In other words, if the predicted enhancers are functional in T cell, we would expect they are more likely to be associated with genes having T cell specific expression compared to non-specific genes. To identify genes specifically expressed in T cells, we used a compendium of microarray expression profiles covering 19 cell types in human (Su *et al.*, 2004). We then assigned a T cell expression specificity score for each gene based on entropy (see Supplementary Methods). Using this measure, it is possible to rank genes with respect to their expression specificity to a given tissue. We identified top 1000 genes that are specifically upregulated, exhibit no expression change and are specifically downregulated in $CD4^+$ T cell. We then examined if predicted enhancers are enriched within domains of T-cell specific genes (see Section 2). As shown in Figure 5C, we observed a 1.87-fold enrichment of predicted enhancers around T-cell specific genes compared to a set of random genes ($P = 0$). Additionally, for genes that show no expression change or are repressed in T cell, the number of enhancers is significantly depleted compared to random genes ($P = 1.1 \times 10^{-3}$ and $P = 4.1 \times 10^{-9}$, respectively, Fig. 5C).

4 DISCUSSION

We introduced a novel computational framework to identify enhancers using chromatin histone modification signatures. An important component of this novel approach is the introduction of data transformation and feature extraction to epigenomic data analysis. Although these concepts are not new in the field of pattern recognition, to the best of our knowledge, they have not been applied to epigenomic data. Unlike genomic sequences that are sequences of discrete alphabets (four nucleotides), epigenomic signals could be described as a continuous real-valued signal (analog nature). It bears a striking similarity to physiological signals such as electroencephalographic signals and NMR spectra. Thus, it could be beneficial if we use data transformation techniques to reveal hidden patterns in complex and high-dimensional epigenomic data. For instance, raw histone modification signal in an input window can be described using its mean, skewness and kurtosis that are different mathematical functions. These functions allow us to capture different aspects of chromatin modification signals many of which may be missed by previous approaches that only use arithmetic mean. For example, different shapes of histone modification peaks, e.g. symmetric versus asymmetric, could be captured by the skewness measure. These differences in peak shape might reflect yet undiscovered differences in the underlying molecular mechanisms that create the modifications in the first place.

Besides data transformation, feature extraction is arguably the most important component of a pattern recognition system. By combining all histone modifications features, we are exploiting a data space where irrelevant features are discarded and noise is decreased. To demonstrate the utility of data transformation and feature extraction for analyzing epigenomic data, we showed that CSI-ANN outperformed two existing methods in the task of identifying functional DNA elements using histone modification data.

In this study, we compared the performance of HMM with that of ANN. Although a powerful framework for biological sequence analysis, HMMs assume that states are independent, which may not hold true for histone modification data given the combinatorial nature of histone tail post-translational modification (Strahl and Allis, 2000). Also, the maximum likelihood criterion of the training algorithm leads to poor discrimination, i.e. it maximizes the probability of a given model generating the observed sequence, but does not minimize the probability of the other models generating the same sequence. Because ANNs can be trained with a discriminative criterion, these systems could perform better than those based on HMMs. In addition, given its nature, ANNs offer a potential of providing massive parallelism, adaptation and efficient algorithms for solving classification problems, which could be very beneficial when analyzing multi-dimensional and huge datasets such as epigenomic data.

Nonetheless, there still is room from improvement. First, in this study, we processed histone modification data using only two mathematical functions (mean and energy). Additional functions from domains such as statistics and information theory could be added to create additional features using raw histone modification data. Second, although a linear feature extraction method such as FDA produced satisfactory results, we would expect a non-linear combination method to provide better results than that from linear transformations. Third, for this study we only used chromatin histone modification data. Other data sources such as nucleosome positioning, TFBS and gene expression profiling can also be included. We could input all those different data sources into a feature extraction step to create even better informative feature(s) for the classifier.

ACKNOWLEDGEMENTS

We thank three anonymous reviewers for their helpful comments.

Funding: American Cancer Society (77-004-31 to K.T.); Pharmaceutical Research and Manufacturers of America Foundation (to K.T.); National Institute of Health (HL0007344-31 to H.A.F.); National Science Foundation and Computing Research Association (0937060 subaward CIF-239 to D.U.).

Conflict of Interest: none declared.

REFERENCES

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Blanchette,M. *et al.* (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, **16**, 656–668.

Boyle,A.P. *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.

D'Alessandro,M. *et al.* (2003) Epileptic seizure prediction using hybrid feature selection over multiple intracranial EEG electrode contacts: a report of four patients. *IEEE Trans. Biomed. Eng.*, **50**, 603–615.

Duda,R.D. *et al.* (2000) *Pattern Classification.* Wiley-Interscience, New York.

Eberhart,R. C. *et al.* (2001) *Swarm Intelligence.* Morgan Kaufman, San Francisco.

Ferretti,V. *et al.* (2007) PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res.*, **35**, D122–D126.

Frith,M.C. *et al.* (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.

Haykin,S. (1998) *Neural Networks: A comprehensive Foundation.* Prentice Hall, Upper Saddle River.

Heintzman,N.D. and Ren,B. (2009) Finding distal regulatory elements in the human genome. *Curr. Opin. Genet. Dev.*, **19**, 541–549.

Heintzman,N.D. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.

Heintzman,N.D. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.

Hon,G. *et al.* (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput. Biol.*, **4**, e1000201.

Kailath,T. (1967) The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans Comm. Tech.*, **15**, 52–60.

Khalil,A.M. *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA*, **106**, 11667–11672.

Kim,T.H. and Ren,B. (2006) Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genomics. Hum. Genet.*, **7**, 81–102.

King,D.C. *et al.* (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res.*, **15**, 1051–1060.

King,G.F. and Kuchel,P.W. (1994) Theoretical and practical aspects of NMR studies of cells. *Immunomethods*, **4**, 85–97.

Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.

Pennacchio,L.A. *et al.* (2007) Predicting tissue-specific enhancers in the human genome. *Genome Res.*, **17**, 201–211.

Raychaudhuri,S. *et al.* (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*, 455–466.

Schones,D.E. and Zhao,K. (2008) Genome-wide approaches to studying chromatin modifications. *Nat. Rev. Genet.*, **9**, 179–191.

Sinha,S. *et al.* (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19**(Suppl. 1), i292–i301.

Smart,O. *et al.* (2007) Genetic programming of conventional features to detect seizure precursors. *Eng. Appl. Artif Intell.*, **20**, 1070–1085.

Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Thurman,R.E. *et al.* (2007) Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.*, **17**, 917–927.

Visel,A. *et al.* (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.*, **40**, 158–160.

Waibel,A. *et al.* (1989) Phoneme recognition using time-delay neural networks. *IEEE Trans. Acousc. Speech Sig. Proc.*, **37**, 328–339.

Wang,Z. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

Wang,Z. *et al.* (2009) Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, **138**, 1019–1031.

Won,K.J. *et al.* (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, **9**, 547.

Zhang,Y. *et al.* (2008) Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics*, **9**, 537.