Prediction of Biological Functions by Histone Modification Patterns Profiling

Yiou Xiao\(^1\), Kishan G. Mehrotra\(^1\), Chilukuri K. Mohan\(^1\), Pratibha Choudhary\(^2\) and Ramesh Raina\(^2\)

\(^1\)Department of Electrical Engineering and Computer Science
\(^2\)Department of Biology
Syracuse University, Syracuse, NY, 13210, USA
(yixiao, mehrotra, ckmohan, pchoud01, raraina)@syr.edu

Abstract

Histone modifications provide an important layer of gene regulation in eukaryotes. In this paper, we propose an approach that identifies the histone modification patterns most relevant for specific biological functions, such as flowering in plants. We first propose a new pattern scoring method, which evaluates the importance of each combinatorial pattern of histone modifications; this is used along with logistic regression, Support Vector Machines, and naïve Bayesian classifier algorithms to predict gene functions. This approach is shown to be successful in inferring significant patterns verified by independent gene function data, outperforming other pattern scores used in current histone modification analysis research.

**keywords:** Histone modification, gene expression, predicting biological function

1. Introduction

Feeding the world’s population requires designing robust crops, with improved yield and enhanced resistance to diseases. In plants, histone modifications have been associated with many biological processes, including development [3], flowering time [14], and pathogen defense [2]. To understand plant fitness, we need to understand how the histone modifications regulate development, flowering time and pathogen defense but no prior studies have explored the same. In this study, we attempt to understand the relationship between certain combinatorial patterns of histone methylation and acetylation in regulating plant development, flowering time, and pathogen defense in *Arabidopsis thaliana* (*A. thaliana*). This is a novel step in the early stages of the epi-genomics era, and we believe that there is a tremendous potential for the use of similar computational methods to predict how these patterns regulate plant development, flowering time, stress and defense responses.

Gene expression in eukaryotes is regulated at several levels, including transcription, post-transcription, translation, and post-translation. In higher organisms, genomic DNA is packaged with the help of histone proteins such as H3, H4, H2A, and H2B. Each unit of DNA and histone proteins assembly is known as the nucleosome. *N*-terminal tails of the histone proteins are subjected to various modifications such as acetylation, methylation, ubiquitination, and sumoylation, which regulates open or close state of the chromatin.

Histone modification is a post-translational mechanism, which allows eukaryotes to have an additional important layer of gene regulation, by opening up the space within neighboring nucleosomes or packaging them more tightly. Relaxed nucleosomes allow access to the transcription factors, hence facilitating gene activation, whereas condensed nucleosomes restrict the access of transcription factors, resulting in gene repression. These changes may be transient or can be inherited into future generations, possibly affecting the fitness of an organism in response to various environmental stimuli.

Histone modifications such as acetylation and methylation have been shown to regulate development of plants, and recent work has shown that they also regulate stress tolerance in plants [15, 9]. Histone acetylation is mainly associated with gene activation, whereas methylation is associated with both gene activation and repression. Gene expression can be turned on and off based on the presence of active or repressive methylation marks on genes; in *A. thaliana*, these marks occur mostly on lysine (K) and/or and arginine (R) residues of H3 and H4 histone proteins. H3K4me, H3K36me marks are associated with gene activation, however, H3K9me, H3K27me are associated with gene repression.

Histone acetylation and methylation marks occur in various combinatorial patterns in the promoters and/or in the coding sequences of genes, leading to different outcomes of gene expression. These combinatorial patterns of the histone marks may function cooperatively or antagonistically to regulate gene expression, and have been studied mainly in humans [26], but to a much smaller extent in *A. thaliana* [5, 7].

This paper reports the results of our analysis of the Chromatin ImmunoPrecipitation sequencing (ChIP-seq) dataset from Luo, et al., [19], which was produced to analyze histone modifications patterns for the regulation of natural antisense transcripts. We selected nine abundant marks from the ChIP-seq dataset, i.e., H3K4me2/3, H3K9me2, H3K27me1/3, H3K36me2/3, H3K9ac, H3K18ac, and total H3 occupancy. We have succeeded in discovering patterns which are unique to plant development, flowering time, stress response and pathogen defense. This study is useful to understand the...
regulation of gene expression related to these biological processes and might be helpful in designing better crops.

Section 2 presents the methodology we used, including discussion of related work. Section 3 describes the experimental simulations and results obtained using our approach. Concluding remarks are given in Section 4.

2. Methods

In this work, we focus on ChIP-Seq dataset where intervals of genes are associated with a signal strength score. The strength score is a simplified data representation of histone modification activity. The high values represent reliable and active histone modification regulation and the low values represent absence of histone modification or weak affinity. They are quantified by the number of reads (read is the unit in ChIP-Seq for counting matching antibodies sequences), where each interval is associated with a value. Because the interval lengths are not uniform, the common approach [24, 17] is to apply a fixed-length sliding window across the entire genome and calculate the accumulated read counts in each segment for simplicity. We devise the raw ChIP-Seq dataset into fixed-width segments of 100bp length using bedtools [22]; in our work, we consider 20 segments in the upstream 1000bp to downstream 1000bp range. We use the empirical values for segments size, up/down stream range from [19, 17, 25]

2.1 Related work

Although researchers have not fully understood the underlying mechanism of how histone acetylations and methylations control gene expression, several approaches have been proposed to help speed up the process of hypothesis...
generation and experiment design. Subsets of genes have been used to discover significant patterns, along with genome-wide pattern discovery, as summarized below.

Polling of individual histone modifications: For a subset of genes \( G_f \), the presence ratios for modifications are defined as a vector \( R^{(f)} = [R_1^{(f)}, R_2^{(f)}, \ldots, R_H^{(f)}] \), summarizing the histone modifications distribution for all genes in \( G_f \), where \( R_j^{(f)} = \sum_{g \in G_f} x_{g,j} / n_f \). Most works [16, 2] adopted direct polling to obtain the overview distributions of different histone modifications on genes in genome wide scale. The individual histone modification ratios are not used as the final results of their analysis because the importance of combinations rather than individual modification are well recognized in recent epigenetic studies [20]. In this paper, we use polling as our baseline against which other space transformation methods are compared.

Pairwise correlation analysis: Co-occurrence strength between histone modifications \( j \) and \( k \) can be measured by the cosine similarity \( \cos_{j,k} = (x_{g,j} \cdot x_{g,k}) / |x_{g,j}||x_{g,k}| \) where \( g \in G_f \) (see [13, 10]). We conducted pairwise similarity analysis for two functions \( f_1 = \text{flowering} \) and \( f_2 = \text{stress} \), observing small differences, e.g., H3 and H3K9me2 co-occur slightly more often in flowering than stress related genes.

Market basket analysis: Frequencies (number of occurrences) of item sets have been used [18, 23, 26] to obtain combinations of important histone modifications patterns. The pattern ratio \( r_{j,f} = n_{j,f} / n_f \) is used to measure the importance of pattern \( P_f \) for label \( f \).

Clustering: The self-organizing map (SOM) approach was proposed in [20] to infer clusters of modifications, based on the raw histone modification enrichment levels \( m_{j,f}^{(f)} \). But this unsupervised approach ignores the biological functions of genes, instead focusing on cluster visualization.

2.2. HiPSiS: Histone Profiling by Significance Score

We propose HiPSiS, an innovative method for modification pattern inference which focuses on evaluating patterns by evaluating a significance score. First, for each pattern \( P_f \) we compute the global ratio \( r_t \), as well as \( r_{t,f} \) for each \( G_f \). We adopt the FP-growth algorithm for efficient enumeration and indexing [12]. In this step, we only consider the patterns that exist in the data set so that the number of combinations is bounded by \( \min\{N, 2^H\} \).

For each specific function \( f \) and pattern \( P_f \), we assume that \( x_f = (r_{t,f} - r_t)^k / \sigma \sim N(0,1) \), where \( \sigma = r_t(1 - r_t)(1/n_f + 1/n_j) \). We quantify the importance of \( r_{t,f} \) using the cumulative probability from two tails of the normal distribution, and the final score assigned to pattern \( P_f \) is

\[
s_{t,f} = \log(Pr[x_{f} \leq r_{t,f} \cdot n_f]) - \log(Pr[x_{f} > r_{t,f} \cdot n_f]).
\]

The score of a pattern is essentially the log value of ratios of odds; and the ratio represents the inclination towards a certain choice. If \( s_{t,f} > 0 \), pattern \( P_f \) is considered important in function \( f \); in contrast, patterns with negative scores are considered to interfere with function \( f \).

As the output of HiPSiS , the matrix \( s_{t,f} \) contains the results of function specific histone modification patterns. Namely, it’s a mapping function from any combinatorial pattern \( \ell \) to a real value to quantify its importance with respective to function \( f \). The higher the score is the more confident and important pattern \( \ell \) is for biological function \( f \). As a result, HiPSiS can be used as a function specific pattern recognition method to select top \( k \) patterns in each function.

2.3. Gene function prediction

As a result, for each function, we have a large sparse matrix \( Z^{(f)} \) to represent the ownership and importance score of different patterns for each gene. The new input data is then fed into normal classification algorithms to evaluate the HiPSiS scoring system.

We map the original input data \( X \) on to new spaces using various pattern scoring methods to evaluate the latter. Since pairwise correlation and SOM clustering cannot quantify the importance of a particular combination of histone modifications, we performed the transformation using only the following methods.

1. Original space: We use the original binary matrix \( X \) without any modification.
2. Polling ratio weighted space: For each \( f \)-vs-rest classification problem, we transformed the original input data \( W^{(f)} \) by:

\[
W^{(f)} = X^{N \times H} R^{(f)}.
\]

3. Simple basket weighted space for function \( f \): For each record \( X_i \), we create a new record based on the matching results of all observed patterns. The new input data is:

\[
Y^{(f)} = A \begin{bmatrix} r_{1,f} \\ r_{2,f} \\ \vdots \\ r_{t,f} \end{bmatrix}.
\]

4. HiPSiS weighted space:

\[
Z^{(f)} = A \begin{bmatrix} s_{1,f} \\ s_{2,f} \\ \vdots \\ s_{t,f} \end{bmatrix}.
\]

In our experiments, we used the 4 different binary matrix \((X, W^{(f)}, Y^{(f)}, Z^{(f)})\) to train multiple learning algorithms.

We expect better performance by conducting this additional data feature transformation compared with original input.

We inferred the significant combinatorial patterns for each different biological function, and categorized genes into predicted functional groups by applying multiple classification
learning algorithms (Logistic linear[4], Naive Bayesian[21] and Support Vector Machine with linear or Gaussian kernels [6]) to the pattern scores discussed above.

3. Experiments and Results

3.1. Pattern recognition performance

![Figure 2: Emission probability of HMM model learned using ChromHMM with 5 hidden states. (a) With α = 0.9. States 2-4 clearly captured the planted pattern [2,7,8]. (b) When α = 0.5. The planted pattern is not obvious anymore.](image)

In order to evaluate the specificity and sensitivity of combinatorial patterns recognition, we compare HiPSiS with ChromHMM [10]. We use the simulated binary data $X_{N \times M}$ where $N = 10,000$, $M = 10$ and $P(X_{i,j} = 1) = 0.3$, $P(X_{i,j} = 0) = 0.7$\(^1\). Then we plant a pattern [2, 7, 8] with probability $\alpha$ a randomly selected subset $G_{test}$. The $\alpha$ parameter controls the confident level of the planted level and the binary values of $X_i$ is randomly toggled with probability $1 - \alpha$. Then we train the hidden markov model using ChromHMM with input data $G_{test}$. In the meanwhile, we calculate the pattern score using HiPSiS and find the top 5 patterns based on their scores $S_{test}$. It’s obvious that ChromHMM is capable of capturing pattern [2,7,8] when the confidence level is reasonably high Figure 2a but when the confidence is very low the pattern is not clear Figure 2b. Whereas HiPSiS is able to assign high scores for our planted patterns even with low confidence level. In Table 2, we managed to capture the planted pattern even with high noise level.

However, ChromHMM is more versatile in terms of representing patterns where each individual modification has different probability. For example, ChromHMM is capable of capturing patterns like [0,2,7/8/9] where 7, 8 and 9 are interchangeable. Whereas HiPSiS is not suitable for recognizing such patterns because it consider [0, 2, 7], [0, 2, 8] and [0, 2, 9] as different patterns. As a result, the scores for such patterns are not significantly high.

3.2. Histone Modification Dataset

In this study, we used ChIP-seq dataset from Luo et al, 2013 which was produced using aerial parts of two-week-old A. thaliana. We accessed this dataset through the National center for Biotechnology Information (NCBI, accession number SRA010097). This dataset contains global distribution of nine histone modifications (H3K4me2/3, H3K9me2, H3K27me1/3, H3K36me2/3, H3K9Ac, H3K18Ac, and total H3 occupancy).

3.3. Gene Labeling

Using TAIR gene ontology annotation [1], we created subsets of genes with specific corresponding functions (i.e. stress, stimulus and etc)\(^2\). The numbers of different annotations are summarized in Table 3.

![Table 3: Overview of TAIR gene GO annotation dataset](image)

3.4. Verification of Predicted Candidates for Specific Functions

We assigned a score to each gene in $G_f$ for each function $f$, using the following normalized pattern score:

$$S_f(g) = \frac{\sum_{\ell \in g} s_{\ell,f}}{||X_g||^1}.$$ (4)

Then, genes with high scores were selected as the potential candidates for label $f$. In this experiment, we only focused on $f$ = flowering because our domain experts created the gene list manually, whereas the labels for other $f$s were obtained through a keyword matching method using GO description, which is less reliable.

Table 4 shows that 14 out of 15 predicted strong candidates from $G_f$ were verified to be correct by domain experts using an independent data source, suggesting the effectiveness of the HiPSiS pattern inference approach. We verified the roles of candidate genes by manual checking.

\(^1\) In our experiment, we also tested multiple different choices for $P$ values

\(^2\) The sum of the numbers of genes for each function $f$ is not equal to the total number 28523, and some genes have multiple function annotations
Table 4: Verification of 15 genes with high scores $S$flowering time, predicted to have the functionality of “flowering time”

<table>
<thead>
<tr>
<th>GeneID</th>
<th>Name</th>
<th>Verification</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT3G48430</td>
<td>JMJ12</td>
<td>flowering activator</td>
</tr>
<tr>
<td>AT3G48590</td>
<td>NF-YC1</td>
<td>flowering activator</td>
</tr>
<tr>
<td>AT3G63010</td>
<td>GID1B</td>
<td>flowering activator</td>
</tr>
<tr>
<td>AT4G00650</td>
<td>FRIGIDA</td>
<td>flowering suppressor</td>
</tr>
<tr>
<td>AT4G08920</td>
<td>CRY1</td>
<td>flowering suppressor</td>
</tr>
<tr>
<td>AT4G15880</td>
<td>ESD4</td>
<td>flowering suppressor</td>
</tr>
<tr>
<td>AT4G24210</td>
<td>SLEEPY1</td>
<td>flowering activator</td>
</tr>
<tr>
<td>AT4G29830</td>
<td>VIP3</td>
<td>flowering suppressor</td>
</tr>
<tr>
<td>AT4G34530</td>
<td>CIB1</td>
<td>flowering activator</td>
</tr>
<tr>
<td>AT5G12840</td>
<td>NF-YA1</td>
<td>flowering suppressor</td>
</tr>
<tr>
<td>AT5G13790</td>
<td>AGL15</td>
<td>flowering activator</td>
</tr>
<tr>
<td>AT5G16320</td>
<td>FRL1</td>
<td>flowering suppressor</td>
</tr>
<tr>
<td>AT5G23150</td>
<td>HUA1</td>
<td>flower development</td>
</tr>
<tr>
<td>AT5G24470</td>
<td>PRR5</td>
<td>flowering activator</td>
</tr>
<tr>
<td>AT5G35840</td>
<td>PHYC</td>
<td>flowering repressor</td>
</tr>
</tbody>
</table>

Table 5: The correlation of scores obtained using HiPSiS and simple basket analysis with background ratios. The correlation values indicate the similarity between patterns from entire dataset versus specific subset(function) of genes. The low correlation is desired because if the function-specific pattern scores are highly correlated with global patterns then they are not considered useful measures to distinguish functions. In this table, we have shown that HiPSiS can distinguish patterns from global dataset and function specific dataset.

<table>
<thead>
<tr>
<th>Function</th>
<th>HiPSiS</th>
<th>Market Basket</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stress</td>
<td>0.11</td>
<td>0.99</td>
</tr>
<tr>
<td>Stimulus</td>
<td>0.36</td>
<td>0.99</td>
</tr>
<tr>
<td>Defense</td>
<td>0.38</td>
<td>0.98</td>
</tr>
<tr>
<td>Development</td>
<td>0.08</td>
<td>0.98</td>
</tr>
<tr>
<td>Flowering</td>
<td>0.61</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Patterns with high scores are proposed as $f$-specific histone modification patterns, whereas others are considered to be irrelevant for function $f$. The proposed patterns of interest are the top five and bottom five patterns, respectively. We observed that although the histone modification H3K9ac is considered strong in multiple functions, it collaborates with different modifications in different functions. For example, H3K4me3 and H3K36me3 are the most important collaborators in function stimulus, whereas H3K18ac is the most important collaborator for defense.

3Descriptions come from definition lines supplied with TIGR gene annotation records (description is generally based on sequence similarity), as well as definition lines from GenBank records (written by the submitter). Other descriptions may be written by a curator and based upon information obtained from the available literature.
3.7. Evaluation of HiPSiS by gene function classification

We evaluated different pattern scoring systems by projecting the original binary histone modification data \( X \) on different feature spaces. We performed a stratified 5-fold cross validation for testing with 10 repeated randomly shuffled sequences of input data \( X \). For each function \( f \), we evaluated the performance of Logistic Regression Classifier with transformed binary data; we also experimented with Naïve Bayesian and SVM (with linear kernel as well as Gaussian kernel) classifiers, obtaining similar results.

We use the mean AUC of ROC curves to evaluate the classification performance for each function \( f \). Figure 3a shows that HiPSiS performs better than other pattern scoring scores. It is noticeable that most of the binary classifiers are very close to random classifier. The main reason is that overlap between \( G_f \) and \( G_f \) in binary feature space. The second cause stems the nature of function labels: we adopted the GO annotation as the function label where incomplete annotation are expected.

4. Concluding Remarks

Histone modifications play an important role in gene regulation. In this paper we propose an approach to predict combinations of histone modifications that are most relevant to each biological function. We propose a new pattern scoring method (HiPSiS ), which evaluates the importance of each combinatorial pattern of histone modifications by comparing with the background ratio. Compared with other pattern scores proposed in previous work, HiPSiS was shown to be capable of inferring significant patterns which were verified by independent gene function data. We also examined the combination of different pattern scoring methods with well-known classifier algorithms to predict gene functions, and observed that HiPSiS performed the best.

We were able to predict new function-specific histone modification patterns, which need to be experimentally verified in future studies. Future studies should also consider the locations and distributions of histone modifications across gene segments, which may be relevant, as implied in [13]. We also plan to work directly with raw enrichment values, and incorporate the locations of modifications into the feature set.

References


[12] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 1–12.


