Higher-order partial least squares for predicting gene expression levels from chromatin states

Shiquan Sun  
School of Computer Science  
Northwestern Polytechnical University  
Xi’an 710072, Shaanxi, China  
Email: sqsun@nwpu.edu.cn

Xifang Sun  
School of Science  
Xi’an Shiyou University  
Xi’an 710065, Shaanxi, China  
Email: xfangsun@126.com

Yan Zheng  
School of Computer Science  
Northwestern Polytechnical University  
Xi’an 710072, Shaanxi, China  
Email: yan.zheng@nwpu-bioinformatics.com

Abstract—Multiple publications have indicated that gene expression levels are strongly affected by chromatin mark combinations via at least two mechanisms, i.e., activation or repression. But their combinatorial patterns remain unresolved. To further understand the relationship between histone modifications and gene expression levels, here in this paper, we introduce a purely geometric higher-order representation, tensor (also called multidimensional array), which might contain more hidden information from chromatin states to predicting gene expression levels. The prediction models were learned from regions around upstream 10k base pairs and downstream 10k base pairs of the transcriptional start sites (TSSs) over three species (i.e., Human, Rhesus Macaque, and Chimpanzee) with five histone modifications (i.e., H3K4me1, H3K4me3, H3K27ac, H3K27me3, and Pol II). Experimental results demonstrate that the proposed method is more powerful for predicting gene expression levels than several commonly used methods. Specifically, our method improves the performance on both criteria, R and RMSE as high as 1.7% and 11%, respectively.

Index Terms—Higher-order partial least squares; chromatin states; tensor decomposition; gene expression levels;

I. INTRODUCTION

In epigenetic level, gene expression process is highly regulated by histone proteins [1]. Each unit of chromatin like beads wrap around 147 base pairs of a DNA subsequence to affect gene expression levels. Specifically, histone modifications like methylation, acetylation, and phosphorylation play critical roles in transcriptional regulation process. It’s highly impact the process of gene expression by chemical modification of chromatin condensation and DNA accessibility when genetic information are converted into gene products [2]. These modifications are shown to regulate gene transcription with active or repressive manners [3]. For example, tri-methylation on K4 of histone H3 (i.e., H3K4me3) is primarily associated with transcriptional activation [4], [5], while tri-methylation on K27 of histone H3 (i.e., H3K27me3) are primarily associated with transcriptional repression [6], [7].

One of challenges of this study is to discover and characterize what chromatin mark combinatorial patterns can affect the process of gene expression, further revealing systematic de novo genome annotation. Extensive studies have been conducted on this issue [8]–[10], however, up to now, it is still limited knowledge to understand the degree of complexity of “histone code”. Recent studies have shown that machine learning-based methods can statistically offer higher prediction power to predict gene expression levels, and it can be considered as a promising way to reveal some interesting results in many cases [11]–[13]. For example, Chen et al., utilized support vector machine to train a prediction model for each bin. The results demonstrated that all bins are useful to predict gene expression levels, but they are not equally informative. In order to investigate the higher-order interactive relationship between chromatin features, they modeled an interaction model $y = \sum_i x_i + \sum_{i<j} x_i x_j$ to predict the gene expression levels, where the expression level $y$ as a linear combinations of the effects of individual histone modification features $x_i$ and their products $x_i x_j$ [14]; Dong et al., established a two-step model using linear regression model and random forest method to reveal the relationship between chromatin features and gene expression levels across various cellular contexts. The best bin was selected to represent the remaining signals for each histone modification. The predictor matrix was formed from the best bin for each histone and the whole gene expression levels [15]; Zhou et al., developed a linear mixed model to evaluate the association of each and joint contribution of the five marks with gene expression levels across three species. The marginal effects of each mark are the summation of all window size. The higher-order interactions between markers were also studied by considering them as the covariates in linear mixed model [16].

To naturally characterize higher-order interactions between different markers, tensor representation (also called multilinear or N-way) are frequently introduced to model higher-order interactions in different research fields [17], [18]. More recent study [19], [20] leveraged tensor representation to integrate different omics, environmental, and phenotypic data sets to uncover unclear biological problems; Also, our previous work [21] used tensor representation to identify transcription factor binding sites. All results from these applications are demonstrated that tensor representation enable to achieve a powerful performance.

In this paper, we leverage tensor representation, which intuitively involves more interaction information for chromatin features, to predict gene expression levels. The predictors for each gene were represented by a matrix as input (rather than a vector), in which each row indicates histone markers while
We also normalized the real data set with two steps for each factors and catalyzes the transcription of mRNA. The actual regions; Pol II directly interacts with chromatin remodeling enhancers and promoters; H3K27me3 mark represses genomic (H3K27me3), and Pol II mark. The reason we choose these consistent with the previous work [16]. Five histone marks not only because their molecular functions have been relatively well studied, but also because they represent five marks not only because their molecular functions have been relatively well studied, but also because they represent a wide variety of transcription initiation regulators. H3K4me1 mark presents at both active and poised enhancers; H3K4me3 mark actives transcription start sites; H3K27ac mark actives enhancers and promoters; H3K27me3 mark represses genomic regions; Pol II directly interacts with chromatin remodeling factors and catalyzes the transcription of mRNA. The actual gene expression levels are measured by RNASeq and quantified as RPKM (reads per kilobases per million mapped reads). We also normalized the real data set with two steps for each species:

- We used COVERAGEBED tool [24] to convert the reads into the given window for each mark and each individual, and then we normalized the peak read counts for each individual of each mark by subtracting the number of mapped reads divided by total number of mapped reads and input reads divided by total number of input reads (the detailed procedure see the reference [16]);
- We used logarithmic transformation \( \log_2(x + \theta) \) to normalize the data. In order to obtain the optimal parameter \( \theta \) in prediction phrase, we divided the whole data set into two parts. One-third of data set was used to optimize the parameters \( \theta \), then the optimized \( \theta^* \) was added to the same modification of the remaining two-thirds of data set to train the prediction model and test their performance.

### B. High-order representation

In this section, we first give the detailed description how the original data were represented by a higher-order representation. The first step is to divide the each gene body, which flanking TSSs both sides with 10k base pairs, into different bins for each individual and each mark (the first two steps of Fig 2). For each gene, the data for each bin and each histone marker was reformulated into a three-dimension data structure (genes × marks × bins), which each gene was represented as matrix. The gene expression levels were used the averaged values across 8 individuals. The detailed process to form tensor data is showed in Fig 1.

As shown in Fig. 1, the first step is to show the different signal patterns near the TSSs over different marks. The histone marks H3K4me3 and Pol II show more informative, while H3K27me3 and H3K3me1 show weaker informative. Each mark was represented by multiple bins (e.g., 41 bins). Therefore, we first combined the five marks into a matrix for each gene. The third step in Fig. 1, we used the contour of distribution of each gene to represent its signals. Assume we have 26115 genes. Finally, we collected all genes to form a tensor data \( \mathcal{X} \).

#### C. High-order model and algorithm

Higher-order partial least squares (or \( N \)-way partial least squares, NPLS) was proposed by Bro et al. [25]. It is adapted to high-order data \( \mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N} \). Here, \( N \) is the number of order for high-order data \( \mathcal{X} \) (in our case, \( N = 3 \)), and the variable \( I_i \) represents the dimensionality of the mode \( i \). The response variable \( \mathcal{Y} \in \mathbb{R}^{I_1} \) is the averaged gene expression levels across eight individuals for all marks.

The optimization model of NPLS is easily reformulated from standard PLS model as:

\[
\begin{align*}
\max_{\{\mathbf{P}^{(n)}\}_{q=1}^Q} & \left[ \text{cov} \left( \mathcal{X} \times (2) \mathbf{P}_{(2)}^{(1)}T \times (3) \cdots \times (N) \mathbf{P}_{(N)}^{(N-1)}, \mathcal{Y} \right) \right]^2 \\
\text{s.t.} \quad & \mathbf{P}_{(n)}T \mathbf{P}_{(n)} = \mathbf{I}, \mathbf{q}^T \mathbf{q} = 1. 
\end{align*}
\]

To solve this model, we should find the optimal \( \mathbf{p}_1 \) and \( \mathbf{p}_2 \) such that:

\[
\mathcal{X} = \mathbf{t}_1 \otimes \mathbf{p}_1 \otimes \mathbf{p}_2 + \mathbf{E}_1
\]

where the operation \( \otimes \) is the outer product. The first latent variable \( \mathbf{t}_1 \in \mathbb{R}^{I_1} \) is extracted from the tensor data \( \mathcal{X} \) to provide the maximum of covariance between \( \mathbf{t}_1 \) and the response variable \( \mathcal{Y} \). The \( \mathbf{p}_1 \in \mathbb{R}^{I_2} \) and \( \mathbf{p}_2 \in \mathbb{R}^{I_3} \) is the loading vector for mode 2 and 3, respectively and \( \mathbf{E} \) is the residual of data \( \mathcal{X} \) after the first extraction. For given the number of factors \( f \), the predicted variable can be estimated by the equation,

\[
\hat{\mathcal{Y}} = \mathbf{Tb}
\]

where \( \mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_f) \), \( \mathbf{b} \) is the regression coefficient with respect to \( \mathbf{T} \). The MATLAB code of \( N \)-way partial least squares is freely available at: http://www.models.life.ku.dk/source/nwaytoolbox.

The workflow of our experiments was given in Fig. 2. We first extended upstream region and downstream region to 10k base pairs around TSSs, and then divided into multiple bins ( e.g., 41 bins if 500 base pairs for each bin). Finally, gene expression levels were measured by a 3-order tensor rather than a matrix, i.e., 26115 genes × 5 marks × 41 bins. The histone density in each bin was logarithm-transformed by \( \log_2(x + \beta_k) \) with respect to the parameter \( \beta_k \) (the parameter
for kth bin). In order to determine the optimal $\theta^*_k$ for each bin, we divided the whole data into two parts: one-third of dataset was used for finding the optimal parameter $\theta^*_k$ and then the same $\theta^*_k$ was added to the corresponding bins in the remaining data set. The gene expression levels $Y$ was also logarithm-scaled using the equation $\log 2(Y)$. A high-order multivariate regression model was developed using the logarithm-scaled training data set and the 10-fold cross validations was used to avoid the over-fitting in training model phrase. Finally, the performance of method were measured by the Pearson’s correlation coefficient (R) and root mean square error (RMSE).

III. RESULTS

In our experiments, to avoid the risk of over-fitting to training prediction model and obtain the reliable results, we used 10-fold cross-validation (with 10 random splitting replicates) in which nine parts are used for training the prediction model, while the remaining part for testing the performance of learned model.

We also incorporated three popular machine learning-based methods, including simple linear regression (denoted as LR), random forest (denoted as RF), and support vector machine regression (denoted as SVR) to compare with the proposed method. All these methods were implemented in R, and the parameter of random forest (the number of trees) was set as 500, and support vector regression used default parameters in R. The performance of each methods were evaluated by two criteria, namely Pearson’s correlation coefficient (R) and root mean square error (RMSE) which are formulated by the following equations:

$$R = \frac{\sum_{i=1}^{I_1} (Y_i - \mu_Y) \left( \hat{Y}_i - \mu_{\hat{Y}} \right)}{\sqrt{\sum_{i=1}^{I_1} (Y_i - \mu_Y)^2 \sum_{i=1}^{I_1} (\hat{Y}_i - \mu_{\hat{Y}})^2}}$$

$$RMSE = \sqrt{\frac{\sum_{i} (Y_i - \hat{Y}_i)^2}{I_1}}$$

where $Y$ is the real gene expression levels, while $\hat{Y}$ is the predicted the expression levels; $I_1$ is the number of genes; $\mu_Y$ is the mean of gene expression levels $Y$.

A. Experiments on simulation data sets

Our first experiments were conducted on simulation data sets. Interestingly, we found that the distribution of each mark is similar to the normal distribution but does not exactly the same as it. Therefore, in this simulation experiments, we simulated the five different histone modifications according to their expression levels, which are described the height $h$ of their distributions, $h = 3, 2, 1, 0.5, 0.4$. 10,000 genes were simulated and the distributions of expression levels for each gene were divided into 100 bins to represent the expression levels of corresponding simulated histone. Similar to the PLS
model, we also introduced a latent variable $Z$ to simulate the data in simulation experiments.

Suppose the latent variable $Z \sim N(0, \beta)$, then

$$y = Z + \varepsilon^y \quad \text{and} \quad \log (\sigma_i^2) = Z \times h_i + \varepsilon_i^\sigma$$

where $\varepsilon^y \sim N(0, 0.2)$ and $\varepsilon_i^\sigma \sim N(0, \gamma_i)$, $i = 1, 2, 3, 4, 5$.

The simulation data $\mathcal{X}$ with 100 bins can be obtained by partitioning the density function $N(0, \sigma_i^2)$ into 100 intervals and calculating their area of corresponding intervals for each mark.

Fig. 3 shows the prediction of four methods with respect to varies parameters $\beta$. As shown in Fig. 3, our method (npls) steadily outperforms others on both criteria, R (left) and RMSE (right). The second-best performance is achieved by random forest regression. Compared with other three methods, linear regression method linearly increases on RMSE.

B. Experiments on real data sets

Our second experiments were conducted on real data sets. In this section, we investigated the relationship between gene expression level and chromatin features based on three species (Humans, Chimpanzees, and Rhesus Macaques). The results from the previous work [16] have shown that five marks are significantly enrich near TSSs regardless of species, and the enrichments pattern is robust with respect to the choice of the size of the TSS regions.
Fig. 3. Comparison of four algorithms for predicting gene expression level. The left figure demonstrates the averaged correlation coefficients with varies parameter $\beta$ over 10 random splitting replicates while the right figure shows the averaged root mean square error with varies parameters $\beta$ over 10 replicates. lr: linear regression; rf: random forest; svr: support vector regression; npls: $N$-way partial least squares.

Fig. 4. The relationship between the numbers of factor and the performance on three species.

Herein, we considered the DNA regions around (10k) at the upstream and downstream regions of the TSSs in current study. In our model, the number of factors is an important parameter to affect the performance. To investigate how parameters affect the performance of proposed method, we check it under two criteria (R and RMSE) with the varies of factors (see Fig 4). We can see that the performances are robust when the number of factors is larger than 4.

The comparison of the results of four regression models over three species were summarized in Table I. As shown in Table I, our method is steadily better than others with respect to both averaged R and averaged RMSE on three species. For the performance on R, the proposed method improved roughly 1.2%, 1.7%, and 1.3% on Hum, Chi, and Rhe data sets, respectively, while RMSE was improved roughly 11%, 8%, and 8% on Hum, Chi, and Rhe data sets, respectively. For other methods, random forest regression outperforms other two methods on Hum and Chi data sets while support vector machine performs better on Rhe data set.

**IV. CONCLUSION**

In this paper, we proposed a higher-order representation method to predict gene expression levels from chromatin state enrichments. The effectiveness of proposed method was validated by simulation and real data sets. Our method can outperform others, most likely because higher-order representation method can integrate more information than standard representation method. These results again demonstrate
that the gene expression levels are strongly correlated with chromatin mark combinations.

Each chromatin state shows specific functional and annotation. Our work provides a way to study genomic annotation via chromatin mark combinations, which can extend the epigenetic functional interpretation of the human genome. Based on what we have done in previous studies [26]–[28], our further work is to try to incorporate epigenetic factors into the model.

ACKNOWLEDGMENT

We would like to acknowledge Dr. Xiang Zhou at University of Michigan, Ann Arbor for very helpful guide to pre-process the data set. This work is supported by the Fundamental Research Funds for the Central Universities (Grant No. 3102017QD098).

REFERENCES

[1] A. J. Bannister, T. Kouzarides, “Regulation of chromatin by histone modifications,” Cell research, vol. 21, pp. 381-395, Mar 2011.
[2] T. Kouzarides, "Chromatin modifications and their function," Cell, vol. 128, pp. 693-705, Feb 2007.
[3] R. Karlic, H. R. Chung, J. Lasserre, K. Vlahovic, and M. Vingron, "Histone modification levels are predictive for gene expression," Proceedings Of the National Academy Of Sciences Of The United States Of America, vol. 107, pp. 2926-2931, Feb 2010.
[4] H. Santos-Rosa, R. Schneider, A. J. Bannister, J. Sherriff, B. E. Bernstein, N. C. T. Emre, et al., "Active genes are tri-methylated at K4 of histone H3," Nature, vol. 419, pp. 407-411, Sep 2002.
[5] A. J. Ruthenburg, C. D. Allis, and J. Wysocka, "Methylation of lysine 4 on histone H3: Intricacy of writing and reading a single epigenetic mark," Molecular Cell, vol. 25, pp. 15-30, Jan 2007.
[6] T. S. Mikkelson, M. C. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, et al., "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," Nature, vol. 448, pp. 553-560, Aug 2007.
[7] A. Barski, S. Cuddapah, K. R. Cui, T. Y. Roh, D. E. Schones, Z. B. Wang, et al., "High-resolution profiling of histone methylations in the human genome," Cell, vol. 129, pp. 823-837, May 2007.
[8] B. D. Strahl and C. D. Allis, "The language of covalent histone modifications," Nature, vol. 403, pp. 41-45, Jan 2000.
[9] Ernst, Jason, and Manolis Kellis. "Discovery and Characterization of Chromatin States for Systematic Annotation of the Human Genome," Nature biotechnology vol. 28, pp. 693-694, 2010.
[10] S. H., Wu, et al. "Independent regulation of gene expression level and noise by histone modifications," Plos Computational Biology, vol. 13, pp. e1005585, Jun 2017.
[11] A. Daemen, O. L. Griffith, L. M. Heiser, N. J. Wang, O. M. Enache, Z. Sanborn, et al., "Modeling precision treatment of breast cancer," Genome Biology, vol. 14, pp. R110, Dec 2013.
[12] K. Kim, E. Bolotin, E. Thuesch, H. Y. Huang, M. W. Medina, and R. M. Krauss, "Prediction of LDL cholesterol response to statin using transcriptomic and genetic variation," Genome Biology, vol. 15, pp. 460, Sep 2014.
[13] L. Devadas, A. Yen, and M. Kellis, "Various localized epigenetic marks predict expression across 54 samples and reveal underlying chromatin state enrichments," bioRxiv, 2017.