Sun *et al. BMC Bioinformatics* 2018, **19**(Suppl 5):113
https://doi.org/10.1186/s12859-018-2100-y

BMC Bioinformatics

RESEARCH Open Access

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

Shiquan Sun[1,3*†], Xifang Sun[2†] and Yan Zheng[1]

From Biological Ontologies and Knowledge bases workshop at IEEE BIBM 2017
Kansas City, MO, USA. 14 November 2017

Abstract

Background: Extensive studies have shown 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 are still unclear. 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 borrow more unknown interactions in chromatin states to predicting gene expression levels.

Results: The prediction models were learned from regions around upstream 10k base pairs and downstream 10k base pairs of the transcriptional start sites (TSSs) on 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 to predicting gene expression levels than several other popular methods. Specifically, our method enable to get more powerful performance on both commonly used criteria, R and RMSE, as high as 1.7% and 11%, respectively.

Conclusions: The overall aim of this work is to show that the higher-order representation is able to include more unknown interaction information between histone modifications across different species.

Keywords: Higher-order partial least squares, Chromatin states, Tensor decomposition, Gene expression levels, Histone modification

Background

In epigenetics, histone modifications like methylation, acetylation, and phosphorylation play critical roles in transcriptional regulation process [1]. Specifically, during gene expression process, each unit of chromatin like beads wrapping around DNA subsequences (about 147 base pairs) is 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 in this study is to discover or characterize what chromatin mark combinatorial patterns can affect the process of gene expression, further revealing complex gene expression mechanisms in downstream analysis [8–13]. This topic have attracted extensive attentions [14–16], 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

*Correspondence: sqsun@nwpu.edu.cn
†Equal contributors
[1]School of Computer Science, Northwestern Polytechnical University, 710072 Xi'an, Shaanxi, People's Republic of China
[3]Department of Biostatistics, University of Michigan, 48109 Ann Arbor, MI, USA
Full list of author information is available at the end of the article

BioMed Central

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

prediction power to predict gene expression levels, and it can be considered as a promising way to reveal some interesting results in many cases ([17, 18], Devadas L, Yen A, Kellis M. Various localized epigenetic marks.predict expression across 54 samples and reveal underlying chromatin state enrichments. 2015; bioRxiv 030478, unpublished). 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 $y \sim \sum_i x_i + \sum_{i<j} x_i x_j$ to predict gene expression levels, where the expression level y as a linear combinations of the interactions between individual histone modification features x_i and their products $x_i x_j$ [19]; 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 [20]; 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. 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 [21].

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 [22, 23]. More recent studies ([24], Khan SA, Ammad-ud-din M. tensorBF: an R package for Bayesian tensor factorization. 2016; bioRxiv 097048, unpublished) leveraged tensor representation to integrate different omics, environmental, and phenotypic data sets to uncover unclear biological problems; Also, our previous work [25] 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 each column represents the bin we combined (see Fig. 1). To make the proposed method scalable, three popular machine learning-based methods, including linear regression, random forest [26] and support vector regression [27], were conducted on a series of simulation and real data sets. The results demonstrate that the proposed method gave a statistically significant improvement compared with other prediction models.

Methods

Data sets and pre-processing

In this study, we used the real data sets which are from lymphoblastoid cell lines (LCLs) over three species, namely Human (GSE47991 and GSE19480), Rhesus Macaque (GSE60269), and Chimpanzee (GSE60269), and these data set are all available in Gene Expression Omnibus (GEO). For each species, eight individuals were considered, and for each individual, 26,115 genes were considered in our experiments.

The preprocessing workflow of real data was completely consistent with the previous work [21]. Five histone marks were queried to contribution in gene expression levels: promoter marks (H3K4me1, H3K4me3, H3K27ac), repressor mark (H3K27me3), and Pol II mark. The reason we choose these 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:

(i) We used COVERAGEBED tool [28] 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 [21]);
(ii) We used logarithmic transformation $\log 2\,(x + \theta)$ to normalize the data. In order to obtain the optimal parameter θ in prediction phrase, we divided the whole data set into two parts. One-third of data set was used to optimize the parameters θ, then the optimized θ^* was added to the same modification of the remaining two-thirds of data set to train the prediction model and test their performance.

High-order representation

In this section, we give more detailed description how the original data were represented by a higher-order

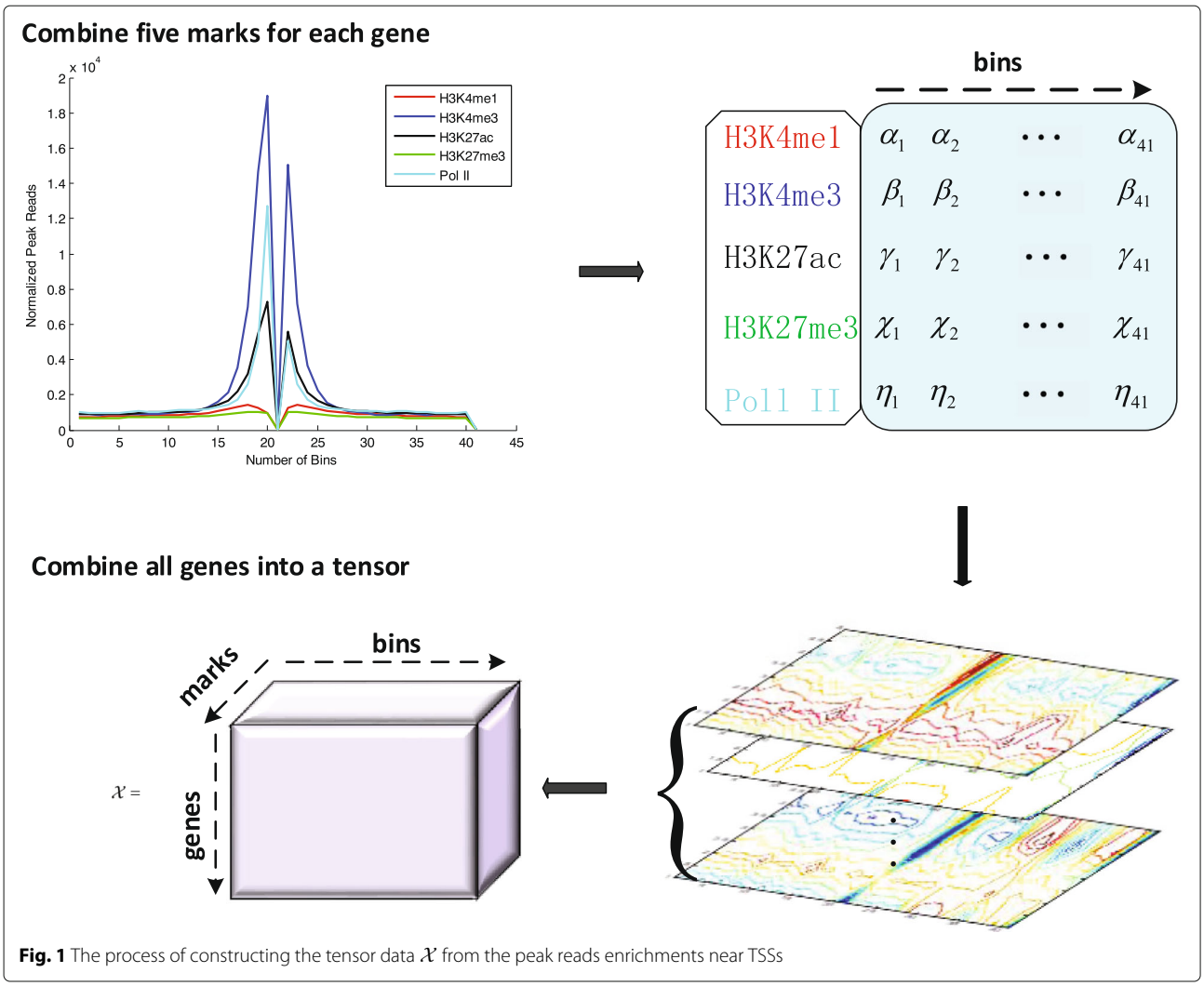

Fig. 1 The process of constructing the tensor data $\mathcal{X}$ from the peak reads enrichments near TSSs

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. 1). For each gene, the data for each bin and each histone marker was reformulated into a three-dimension data structure (genes × marks × bins), therefore, each gene was represented by a matrix instead of a vector. 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. In third step (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 $\mathcal{X}$.

High-order model and algorithm

Higher-order partial least squares (or N-way partial least squares, NPLS) was proposed by Bro et al. [29]. It is adapted to high-order data $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$. 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 $\mathbf{Y} \in \mathbb{R}^{I_1}$ 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:

$$\max_{\{\mathbf{P^{(n)}}\},\mathbf{q}} \left[\text{cov}\left(\mathcal{X} \times_{(2)} \mathbf{P}^{(1)T} \times_{(3)} \cdots \times_{(N)} \mathbf{P}^{(N-1)T}, \mathbf{Yq}\right)\right]^2 \tag{1}$$

$$\text{s.t.} \quad \mathbf{P}^{(n)T}\mathbf{P}^{(n)} = \mathbf{I},\ \mathbf{q}^T\mathbf{q} = 1. \tag{2}$$

To solve this model, we want to find the optimal $\mathbf{p}_1$ and $\mathbf{p}_2$ such that:

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

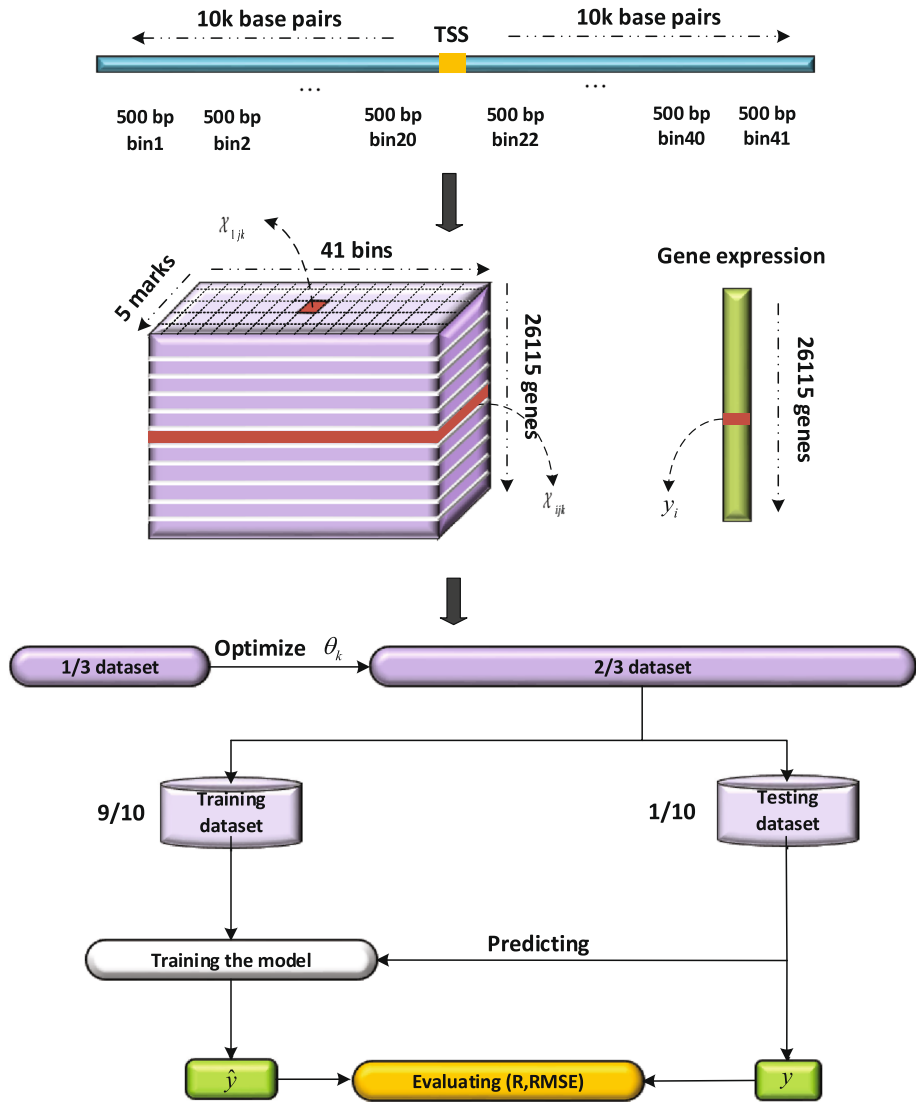

Fig. 2 Schematic diagram of the process of data analysis in our paper

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 $\mathbf{Y}$. 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 $\mathcal{E}$ is the residual of data $\mathcal{X}$ after the first extraction. For the given number of factors f, the predicted variable can be estimated by the equation,

$$\hat{\mathbf{Y}} = \mathbf{Tb}$$

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 $\times 5$ marks $\times 41$ bins. The histone density in each bin was logarithm-transformed by $\log 2\,(x + \theta_k)$ with respect to the parameter θ_k (the parameter for kth bin). In order to determine the optimal θ_k^* for each bin, we divided the whole data into two parts: one-third of dataset was used for finding the optimal parameter θ_k^* and then the same θ_k^* was added to the corresponding bins in the remaining data set. The gene expression levels $\mathbf{Y}$ was also logarithm-scaled using the equation $\log 2(\mathbf{Y})$. A

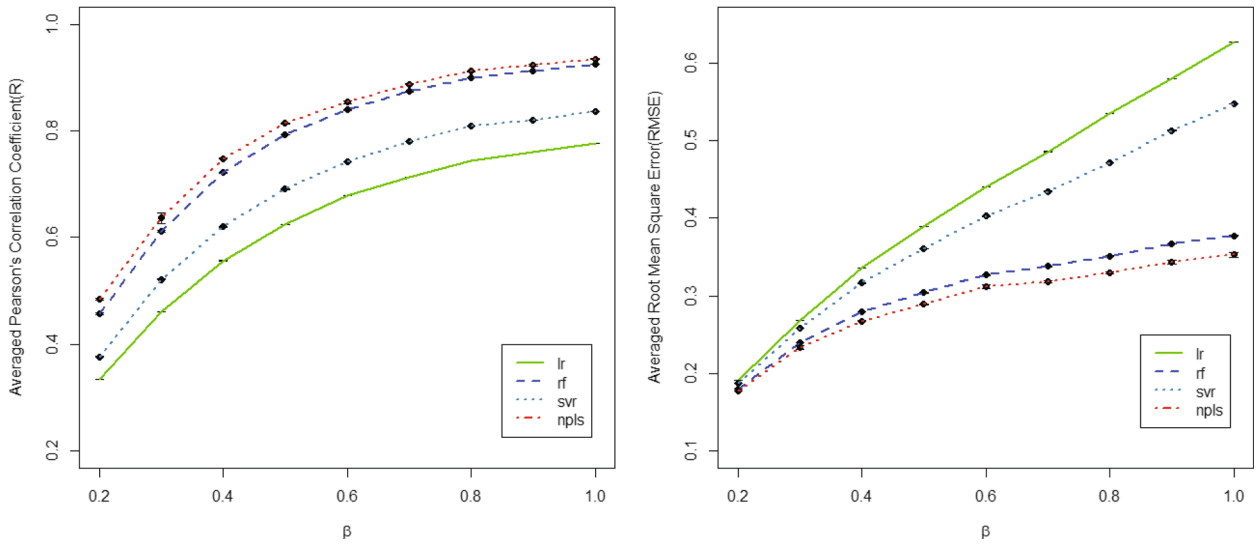

Fig. 3 Comparison of four algorithms for predicting gene expression level. The left figure demonstrates the averaged correlation coefficients with varies parameter β over 10 random splitting replicates while the right figure shows the averaged root mean square error with varies parameters β over 10 replicates. lr: linear regression; rf: random forest; svr: support vector regression; npls: *N*-way partial least squares

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 methods were measured by the Pearson's correlation coefficient (R) and root mean square error (RMSE).

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 (similar to previous study [30]).

We also compared with other three popular machine learning-based methods, i.e., simple linear regression (denoted as LR), random forest (denoted as RF), and support vector machine regression (denoted as SVR) to make the proposed method scalable. 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,

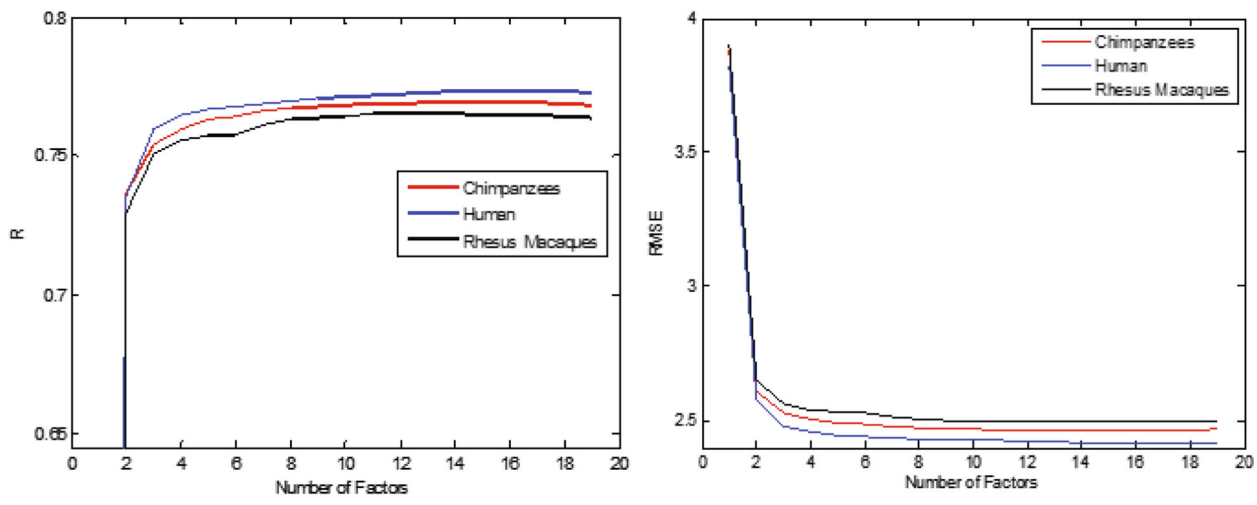

Fig. 4 The relationship between the numbers of factor and the performance on three species

namely Pearson correlation coefficient (R) and root mean square error (RMSE) which are formulated by the following equations:

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

$$RMSE = \sqrt{\frac{\sum_i \left(\mathbf{Y}_i - \hat{\mathbf{Y}}_i\right)^2}{I_1}}$$

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

Experiments on simulation data sets

Our first experiments were conducted on a series of 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 marks 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

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

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 $N\left(0, \sigma_i^2\right)$ into 100 intervals and calculating their area of corresponding intervals for each mark.

Figure 3 shows the prediction of four methods with respect to varies parameters β. As shown in Fig. 3, our method (npls) steadily outperforms others on both criteria, i.e., 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.

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 [21] 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.

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 this parameter 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 1. As shown in Table 1, 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 outperforms other two methods on Rhe data set.

Discussion and conclusion

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

Table 1 The performance of different models on three species data sets

	Linear model	Random forest	Support vector machine	NPLS(41bins)	NPLS(21bins)
Hum	0.769(2.43)	0.775(2.46)	0.774(2.46)	0.784(2.37)	0.787(2.35)
Chi	0.756(2.52)	0.767(2.47)	0.765(2.53)	0.780(2.41)	0.784(2.39)
Rhe	0.760(2.52)	0.761(2.51)	0.765(2.54)	0.774(2.46)	0.778(2.43)

Note: The number in bracket following the average R represents averaged RMSE over 10-flod cross validation (with 10 random splitting replicates). Hum: Human data set, Chi: Chimpanzee data set, and Rhe: Rhesus Macaque data set

Each chromatin state shows specific functional and annotation. Our study provides a way to study genomic annotation via chromatin mark combinations, which can extend the epigenetic functional interpretation of the human genome. Therefore, our further work is to incorporate epigenetic factors into the downstream analysis, such as gene expression analysis [9, 31], GO ontologies [32, 33], and disease-related ncRNAs [34].

Acknowledgements
We would like to acknowledge Dr. Xiang Zhou at University of Michigan, Ann Arbor for very helpful guide to pre-process the real data sets and simulation data sets.

Funding
Publication costs were funded by the Fundamental Research Funds for the Central Universities (Grant No. 3102017OQD098).

Availability of data and materials
All data set used in this work are available in Gene Expression Omnibus (GEO): Human (GSE47991;https://doi.org/https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47991 and GSE19480;https://doi.org/https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19480), Rhesus and Chimpanzee Macaque (GSE60269;https://doi.org/https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60269). The MATLAB code of higher-order partial least squares is freely available at: https://doi.org/http://www.models.life.ku.dk/source/nwaytoolbox.

About this supplement
This article has been published as part of *BMC Bioinformatics* Volume 19 Supplement 5, 2018: Selected articles from the Biological Ontologies and Knowledge bases workshop 2017. The full contents of the supplement are available online at https://doi.org/https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-5.

Authors' contributions
SS and XF conceived and wrote the manuscript. SS implemented the software and analyzed the data, and YZ analyzed the data. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]School of Computer Science, Northwestern Polytechnical University, 710072 Xi'an, Shaanxi, People's Republic of China. [2]School of Science, Xi'an Shiyou University, 710065 Xi'an, Shaanxi, People's Republic of China. [3]Department of Biostatistics, University of Michigan, 48109 Ann Arbor, MI, USA.

Published: 11 April 2018

References

1. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res. 2011;21:381–95.
2. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128:693–705.
3. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci USA. 2010;107:2926–31.
4. Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NCT, et al. Active genes are tri-methylated at K4 of histone H3. Nature. 2002;419:407–11.
5. Ruthenburg AJ, Allis CD, Wysocka J. Methylation of lysine 4 on histone H3: Intricacy of writing and reading a single epigenetic mark. Mol Cell. 2007;25:15–30.
6. Mikkelsen TS, Ku MC, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007;448:553–60.
7. Barski A, Cuddapah S, Cui KR, Roh TY, Schones DE, Wang ZB, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129:823–37.
8. Sun SQ, Peng QK, Shakoor A, Vol. 9. A Kernel-Based Multivariate Feature Selection Method for Microarray Data Classification; 2014, p. e102541.
9. Sun SQ, Hood M, Scott L, Peng QK, Mukherjee S, Tung J, Zhou X. Differential expression analysis for RNAseq using Poisson mixed models. Nucleic Acids Res. 2017;45:e106.
10. Peng J, Xue H, Shao Y, et al. A novel method to measure the semantic similarity of HPO terms. Int J Data Min Bioinforma. 2017;17:173.
11. Chen L, Jiang Y, et al. DisSim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs. Sci Rep. 2016;5:30024.
12. Peng J, Lu J, Shang X, et al. Identifying consistent disease subnetworks using DNet. Methods. 2017;131:104–10.
13. Hu Y, Zhou M, et al. DisSetSim: an online system for calculating similarity between disease sets. J Biomed Semant. 2017;28:71.
14. Strahl BD, Allis CD. The language of covalent histone modifications. Nature. 2000;403:41–5.
15. Ernst Jason, Kellis Manolis. Discovery and Characterization of Chromatin States for Systematic Annotation of the Human Genome. Nat Biotechnol. 2010;28:817–25.
16. Wu SH, et al. Independent regulation of gene expression level and noise by histone modifications. Plos Comput Biol. 2017;13:e1005585.
17. Daemen A, Griffith OL, Heiser LM, Wang NJ, Enache OM, Sanborn Z, et al. Modeling precision treatment of breast cancer. Genome Biol. 2013;14:R110.
18. Kim K, Bolotin E, Theusch E, Huang HY, Medina MW, Krauss RM. Prediction of LDL cholesterol response to statin using transcriptomic and genetic variation. Genome Biol. 2014;15:460.
19. Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biol. 2011;12:R15.
20. Dong XJ, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene expression using chromatin features in various cellular contexts. Genome Biol. 2012;13:R53.
21. Zhou X, Cain CE, Myrthil M, Lewellen N, Michelini K, Davenport ER, et al. Epigenetic modifications are associated with inter-species gene expression variation in primates. Genome Biol. 2014;15:547.
22. Freitas MP, da Cunha EFF, Ramalho TC, Goodarzi M. Multimode Methods Applied on MIA Descriptors in QSAR. Curr Comput Aided Drug Des. 2008;4:273–82.
23. Guzman E, Baeten V, Pierna JAF, Garcia-Mesa JA. Evaluation of the overall quality of olive oil using fluorescence spectroscopy. Food Chem. 2015;173:927–34.
24. Hore V, Vinuela A, Buil A, Knight J, McCarthy MI, Small K, et al. Tensor decomposition for multiple-tissue gene expression experiments. Nat Genet. 2016;48:1094–100.
25. Sun SQ, Zhang XP, Peng QK. A high-order representation and classification method for transcription factor binding sites recognition in Escherichia coli. Artif Intell Med. 2017;75:16–23.
26. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2:18–22.
27. Lu ZQJ, Vol. 173. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition; 2010, pp. 693–4.
28. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
29. Bro R. Multiway calibration. Multilinear PLS. J Chemometr. 1996;10:47–61.

30. Sun SQ, Peng QK, Zhang XK. Global feature selection from microarray data using Lagrange multipliers. Knowl-Based Syst. 2016;110:267–74.
31. Sun SQ, Peng QK. A hybrid PSO-GSA strategy for high-dimensional optimization and microarray data clustering. In: IEEE International Conference on Information and Automation, vol. 105. Hailar: IEEE; 2014. p. 41–6.
32. Cheng L, Sun J, Xu W, et al. OAHG: an integrated resource for annotating human genes with multi-level ontologies. Sci Rep. 2016;6:34820.
33. Peng J, Wang H, Lu J, et al. Identifying term relations cross different gene ontology categories. BMC Bioinformatics. 2017;18:573.
34. Hu Y, Zhou M, et al. Measuring disease similarity and predicting disease-related ncRNAs by a novel method. BMC Med Genomics. 2017;10:71.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit