Prediction of Nucleosome Positioning Based on Support Vector Machine

Jihua Feng, Jianping Xiao, Ying Lu, and Qiufu Shan

***Abstract*—When we used DNA sequences of Saccharomyces cerevisiae whose nucleosome positioning data have been experimentally determined to train a Support Vector Machine (SVM) to predict the nucleosome formation potential of any given sequence of DNA, we observed that chromatin structure has an impact on the evolution of genomic DNA molecules. We have found, on average, 15% lower predictive accuracy rates in nucleosomal DNA than in linker DNA. This widespread local rates heterogeneity represents an evolutionary footprint of nucleosome positions and reveals that nucleosome organization is a genomic feature conserved over evolutionary timescales.**

***Index Terms*—Evolutionary footprints, nucleosome, predictive accuracy rates, transcriptional start sit (TSS).**

I. Introduction

The combination of DNA and some associated proteins in eukaryotic is called chromatin. In all eukaryotes, the basic subunit of chromatin structure called nucleosomes, are all same [1]. The genomes of all eukaryotic organisms are packaged into nucleosomes, comprising 147-bp segments of DNA wrapped around a histone octamer. Nucleosomes are separated by linker DNA and form the basis of higher order packaging of genomes into chromatin [2]. But in both euchromatin, heterochromatin of the interphase cell nucleus, and mitotic chromosomes, the structure of nucleosome remains constant throughout cell life cycle.

Because of the interaction between the protein –protein, protein-DNA, and the formation of some complex macromolecular complexes, transcriptional regulation of eukaryotes becomes a multi-stage complex process. As an important stage of the gene regulation in eukaryotes, nuclesome is an important part of the genetic mechanisms. So the research about the statistical location of nucleosome in the genome is a prerequisite for understanding how nucleosome affect genome DNA's evolution through its own position [3].

In this paper, with the DNA sequence of yeast genome from NCBI database, the physical characteristics of DNA and transcription factor binding sites, we took yeast nucleosome positioning data as a control group to train the support vector machines(SVM) to predict nucleosome packaging capacity of yeast genomic DNA. When we analyzed predictive accuracy rates of SVM model, we gained the interesting conclusion on the structure of chromatin which represents DNA's evolution.

II. Experiment Methods

A. Data Preparation

Our study includes the three parts of data: the first comes from nucleosome positioning experimental data of Lee *et al.* [4], which also includes data of transcription factor binding sites and of DNA structure in yeast; the second, the yeast's 16 chromosomes DNA sequence, is derived from NCBI database; the third part is derived from the experimental data of 4792 yeast gene with high degree of confidence in the David‘s research [5], including proven transcription start sites (TSS) and transcription termination sites (TTS). Due to the heterogeneity of the above data, we reconstructed the some data according to our research purpose.

B. Data Processing

1) Nucleosome positioning data processing

The primitive nucleosomal positioning experimental data provided by Lee *et al.* come form the platter form of Affymetrix tiling microarray with 4-bp resolution. Thus, we must first get, using interpolation method, nucleosome occupancy ratio data which should cover each site of the yeast genome. Further, for the nucleosome positioning data obtained by the hidden Markov model, we set the nucleosome occupancy DNA sequence to 1, and DNA without nucleosome to 0. As a result, a set of binary data corresponding to the yeast‘s 16 chromosome with 0's and 1's are structured which represent the positions of nuclesome.

2) Data alignment

The data were aligned with Transcription Start Sites (TSS) of 4792 genes which were obtained through David's work and we selected 800bp data from both upstream and downstream of TSS. After averaged them, we obtained the data distribution map of genome-wide aligned with gene's TSS.

3) Transcription factor binding sites (TFBS) data processing

Next, we applied a method similar to the nucleosome position data processing. We processed 126 transcription factor binding sites data in which 0's and 1's mean transcription factor binding sites' positions [6]. In order to extract and analyze data conveniently, transcription factor binding sites on 16 chromosomes are integrated to a single set of data.

4) Yeast's DNA data processing (physical characteristics)

The DNA structure data include GC content, Melting

Manuscript received March 1, 2013; revised May 30, 2013. This work was supported by the National Natural Science Foundation of China (Grant No. 31160234) and The Foundation Research Project of Yunnan Province Application Project (Grant No. 2011FB082).

The authors are with Yunnan University of Nationalities, China (e-mail: fengjihua@21cn.com, 275939685@qq.com).

DOI: 10.7763/IJBBB.2013.V3.253

temperature, Enthalpy change, Free energy and so on. We obtained physical characteristics of DNA data which covered the whole genome with 1-bp resolution by using interpolating method, and then normalized those primitive data.

C. Support Vector Machine (SVM) Prediction

1) Genome-wide nucleosome position prediction

In our study, Support Vector Machine's training data include the experimental nucleosome positioning data, yeast DNA structure data and integrated Transcription Factor Binding Sits data.

Then, about 10,000 DNA fragments with the length of 4000bp are randomly selected in genome-wide (predict window length is set to 50bp, the algorithm iteration number is 1000) to predict the DNA-nucleosome affinities. These DNA fragments were randomly distributed in 16 chromosomes. We observed that the difference of average prediction accuracy rates between fragments changed little. But there are significant differences of average prediction accuracy rates between the nucleosomal packaging DNA and link DNA. After statistical analysis, we found that the prediction accuracy rates of the nucleosomal DNA are generally lower than that link DNA ($p < 10^{-7}$). For a better understanding of the relationship between nucleosome prediction accuracy rates and regulation, we made a further study on both upstream and downstream of gene's TSS.

2) Prediction around the TSS

We used the characteristics of the DNA structure such as Clash strength, the Entropy, Rise, Tip and the data of the Transcription Factor Binding Site as SVM model's input, and used the experimental nucleosome positioning data as training data to predict DNA-nucleosome affinity around TSS. During the prediction, in order to verify the robustness of the algorithm, we selected different predictive windows (length of 30, 40, 50, of 60bp) with 1bp step, and then aligned and average the predictive accuracy rates. As a result, the average distribution map of predictive accuracy rates of TSS is shown as red line in Fig. 1.

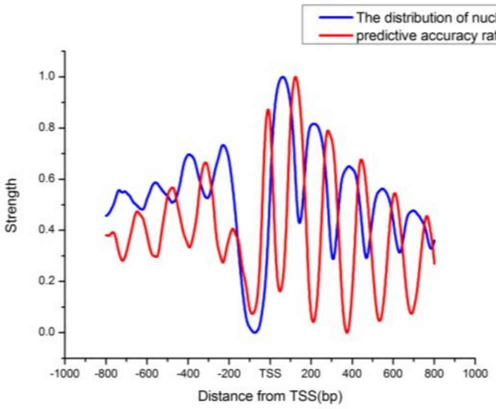

Fig. 1. The distribution of nucleosome and predictive accuracy rates around TSS.

In the figure, the blue line represents the distribution of nucleosome obtained from experimentally, red line show predictive accuracy rates of Support Vector Machine.

In order to verify the reproducibility of the experimental results, we have chosen a different structure of the yeast DNA data to do a crossover experiment. The results are similar to the Fig. 1. It demonstrates that our results are universal.

III. Results

A. Relevance between the Predicting Accuracy and the Nucleosome Positioning around TSS

From Fig. 1, there has an intense relevance, namely negative correlation, between occupancy rates and predict accuracy: high prediction rates exist in the nucleosome linker region, and low rates exist in packing region. And it's consistent with the result of experiment from random genome.

Accordingly, we consider that due to nucleosome-linker DNA encoded lager number of transcription factor binding sites. So it is an important region to assemble transcription factor and has extraordinary significance to living systems [7]. These DNA fragment are relative conservation, and are not easily changing during evolution. Otherwise, it will bring grave consequences. Those DNA physical characteristics are against forming of nucleosome during evolution. Thus, it is easier to be captured by SVM. To the contrary, nucleosome-packing region were restricted fewer than linker-DNA during evolution, and expressed relatively active. The DNA changing rapidly in these areas cause that the SVM difficult to capture information of nucleosome's position. This conclusion coincides with Washietl et al. who found the substitution rates of nucleosome-linker DNA lower than nucleosome–DNA [8].

B. Nuclesome-Free Region Has Low Predicting Accuracy

Interestingly, in the nucleosome-free region (NFR) around TSS of gene, we find that lower occupancy rates region also has lower prediction accuracy relatively (shown in Fig. 2).

We observed that the prediction accuracy is obviously low in NFR areas, and different from other areas. This area is the place that Poly II and other transcription factor to form transcription machine together. Although those areas encode genome information such as promoter, TTS and enhancer, we suppose that those areas disperse nucleosome mainly by recruiting protein or the mechanism related to gene regulation. Hence, due to complicated cause, the performance of SVM is not good enough. Therefore, in NFR our model only acquired the low predict accuracy.

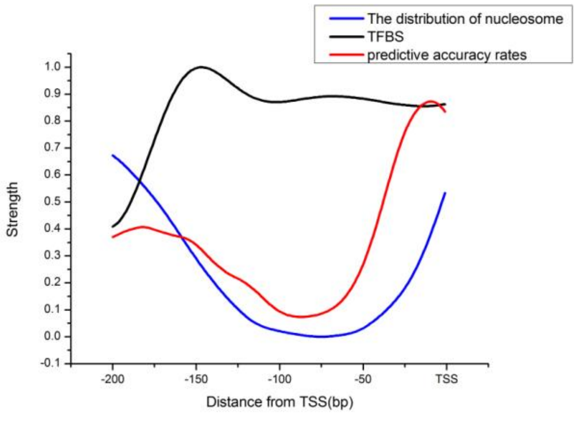

Fig. 2. The distributed of predictive accuracy rates and the distribution of nucleosome in NFR region.

C. The Relationship between Distribution of the Transcription Factor Binding Sites and Prediction Accuracy Rates

In the Fig. 2, the distribution of nucleosome and predictive accuracy rates are lower in the nucleosome deletion region, while transcription factor binding rates' density is higher. This region is a place where the transcription machinery is formed. The start of transcription in Eukaryotes requires correspondence of various protein factors. So, the lack of nucleosome many mainly caused by the binding of transcription factor in this region. Due to the particularity of this area, the role of DNA on the capability of nucleosome packaging becomes a secondary factor. In this condition, Support Vector Machine is hard to obtain predictive information.

D. Deferent Models' Comparison

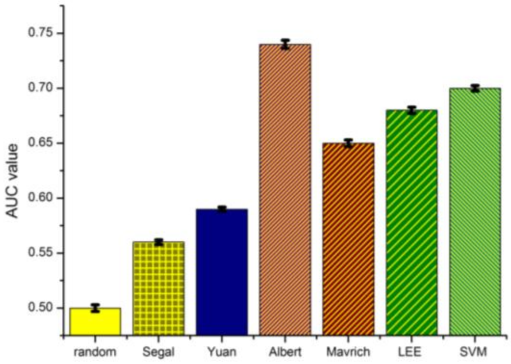

Fig. 3. Comparison of seven models.

We choose the cumulative error curve (AUC) as a performance measure (normalized) for seven models around TSS area [9]. Here, an AUC value of 1 indicates perfect prediction, i.e. all predicted nucleosomes are predicted with zero positional error, while an AUC value close to 0.5 corresponds to random prediction. The results are summarized in Fig. 3. In fact, the four models show only a modest predictive power with maximal AUC values of 0.57(Segal [10]), 0.59 (Yuan [11]), 0.64 (Mavrich [12]) and 0.66 (Lee [4]). This suggests that these models scores per se are a poor predictor of nucleosome positions. However, the AUC values are, in all cases, higher than expected by chance confirming our previous notion that in vivo nucleosomes are also positioned by sequence feature and that our SVM model(AUC 0.71) and Albert (AUC 0.74) [13] have a good performance to capture aspects of the sequence-dependent affinity of the nucleosome.

IV. Conclusion

We use SVM models to predict DNA-nucleosome affinity. Our results establish that nucleosome organization is encoded in eukaryotic genomes. This newly characterized genetic information mainly occurs around gene's TSS areas and may indicate DNA evolution. The consistency of the predictions on the yeast genome using SVM models derived independently from yeast DNA structure data implies that the genomic signals for nucleosome positioning at TSS areas are strong.

References

[1] Z. Zhang and B.F. Pugh, "High-resolution genome-wide mapping of the primary structure of chromatin," *Cell*, vol. 144, no. 2, pp. 175-86, 2011.

[2] S. Sasaki *et al.*, "Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites," *Science*, vol. 323, no. 5912, pp. 401-4, 2009.

[3] A. Jansen and K. J. Verstrepen, "Nucleosome positioning in Saccharomyces cerevisiae," *Microbiol Mol Biol*, vol. 75, no. 2, pp. 301-20, Rev, 2011.

[4] W. Lee *et al*., *A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet*, vol. 39, no. 10, pp. 1235-44, 2007.

[5] L. David *et al.*, "A high-resolution map of transcription in the yeast genome," in *Proc Natl Acad Sci USA*, vol. 103, no. 14, pp. 5320-5, 2006.

[6] S. Veerla, M. Ringner, and M. Hoglund, "Genome-wide transcription factor binding site/promoter databases for the analysis of gene sets and co-occurrence of transcription factor binding motifs," *BMC Genomics*, pp. 145, Nov. 2010.

[7] R. Sadeh and C. D. Allis, "Genome-wide 're'-modeling of nucleosome positions," *Cell*, vol. 147, no. 2, pp. 263-6, 2011.

[8] S. Washietl, R. Machne, and N. Goldman, "Evolutionary footprints of nucleosome positions in yeast," *Trends Genet*, vol. 24, no. 12, pp. 583-7, 2008.

[9] H. R. Chung and M. Vingron, "Sequence-dependent nucleosome positioning," *J Mol Biol*, vol. 386, no. 5, pp. 1411-22, 2009.

[10] E .Segal *et al.*, "A genomic code for nucleosome positioning," *Nature*, vol. 442, pp. 772-778, 2006.

[11] G. C. Yuan and J. S. Liu, "Genomic sequence is highly predictive of local nucleosome depletion," *PLoS Comput. Biol.*, vol. 4, pp. e13, 2008.

[12] T. N. Mavrich *et al*., "A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome," *Genome Res.*, vol. 18, pp. 1073-1083, 2008.

[13] I. Albert, T. N. Mavrich, L. P. Tomsho, J. Qi, S. J. Zanton, S. C. Schuster, and B. F. Pugh, "Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome," *Nature*, vol. 446, pp. 572-576, 2007.

Jihua Feng received the B.S. degree in Computer Science and M.S. degree in Communication and Information System from Yunnan University, Kunming, China, in 2001 and 2006, and Ph.D. degrees in Bioinformatics from Sun Yat-Sen University, Guangzhou, China, in 2010, respectively.

He is currently Associate Professor and Director of Department Information Engineering, School of Electrical and Information Technology, Yunnan University of Nationalities. His research interests include gene regulation networks, data mining, nonlinear intelligence signal processing, and communication.

Jianping Xiao received the B.S. degree in Electronic and Information Engineering from Jilin Teacher's Institute of Engineering &Technology, Changchun, China, in 2010.

He is currently pursuing Master degree at School of Electrical and Information Technology, Yunnan University of Nationalities, Yunnan, China. His research interests have been in the areas of bioinformatics, machine learning, and data mining and signal processing.