2012 4th International Conference on Bioinformatics and Biomedical Technology
IPCBEE vol.29 (2012) © (2012) IACSIT Press, Singapore

Analysis of gene expression behaviour by nucleosome and protein binding profiles

BichHai Ho and TuBao Ho

Japan Advanced Institute of Science and Technology

Abstract. To achieve certain level of regulation over gene expression, cell may simultaneously take various pathways with participation of both *cis* and *trans*-regulatory factors in its four transcriptional phases of orchestration, initiation, access, elongation. To put it simply, gene regulation is exerted through the presence of those factors in regulatory as well as coding regions. However, the precise regulatory mechanisms remain elusive, as previous works on gene regulationusing nucleosome occupancy, binding of chromatin modifiers and regulators, etc., have come to diverse conclusions. To elucidate gene expression behaviour, we focus on nucleosome binding and protein binding information, which is based on the widely accepted hypothesis that they compete for access to*cis*-regulatory elements in regulatory processes. Investigation of 20 elongation factors and nucleosome occupancy in the downstream region from around TSS showed that the correlation of their binding profiles is predictive ofhigh, medium, and low expression levels, with nearly 70% accuracy.Furthermore, analysis of the most discriminative factors gives some insights into the effect of regulatory factors on expression behaviour during elongation phase.

Keywords: gene expression, nucleosome binding, protein binding, elongation factor

1. Introduction

Eukaryotic genomes are packaged inside cell nucleus under chromatin structure, which is formed like a bead-on-string fiber of nucleosomes. As a fundamental unit, each nucleosome contains a core of octamer histone proteins, H2A, H2B,H3, and H4, wrapped around by 147bp of DNA [1]. Increasing biological observations have shown that, more than DNA compacting, nucleosomesplay a role in various cellular processes such as transcription, DNA replication, DNA repair, etc., by occluding the access of biological machineries to *cis*-regulatory elements.In regulating transcription, one of the most important activities, nucleosome propertiesat different phases, i.e., orchestration, initiation, access, and elongation, have been associated with gene expression behaviour[2]. However, diverse conclusions drawn still leave the transcription process incompletely understood.

For instance, as pointed out in [3], the question of whether nucleosome occupancy is related to gene expression is not consistently explained, e.g., supported by [4,5]and refuted by [6,7]. From nucleosomal architecture at promoter, such information as positioning, spacing, and ± 1 nucleosome occupancies, it is possible to determine expression states (e.g., expressed and unexpressed), not expression levels, though[3,8].Hence, study on nucleosome binding properties only probably is not enough to deliver principles at global scale regarding gene expression. This drawback is understandable as in addition to nucleosome, chromatin modifiers and regulators also participate in transcription controlling. Evidently, by considering histone modifications,remodelling and transcription machinery factors, etc.,the derived dependence networks of those factors and gene expression managed to recover meaningful regulatory pathways[9,10].Taken together, we aim to explain the expression behaviour based on nucleosome andprotein binding, i.e., elongation regulatory factors.

More concretely, we utilized nucleosome and 20 elongation factor binding profiles in region from -350 bp to 850 bparound TSS to predict the 3 expression levels(low, medium, and high expression) in 4641

S.*cerevisiae*genes. Based on the biological observation that they compete with each other for binding to theunderlying DNA sequence and that some regulatory factors translocate or dismantle nucleosome during transcription[11], we quantified such interaction by calculating the correlation of their occupancy, called correlation profile. Our results from SVM classification model demonstrate that such information, to some extent, helpsdetermine expression levels, accuracyof nearly 70%. Furthermore, analysis of the most discriminative factors, such as Ctk1 and PolII (Rbp3),suggests some insights about their effects on the resulting gene expression during elongation phase.

2. Results and discussion

2.1. Nucleosome and elongation factor binding uncorrelated with gene expression

In S.*cerevisiae*, nucleosome disassembly is coupled directly with gene activation, i.e., nucleosome binding properties assumingly are related to regulation activities. However, as mentioned above, occupancy either genome-wide[6,7]or at promoter, e.g., theflanking ±1nucleosomes [3], is notindicative of gene expression levels. Hence, we first sought to see in our investigated regions of 1200 bpfrom around TSS toward downstreamwhether they show any correlation. Scatter plots of several factors against gene expression (Fig. 1) show minimal correlation. Other factors (plots not shown) exhibited the same tendency, with average correlation coefficient Spearman's $R^2 < 0.04$. We, therefore, speculate that nucleosome binding as well as elongation factor binding individually are not enough to explain gene expression behaviour.

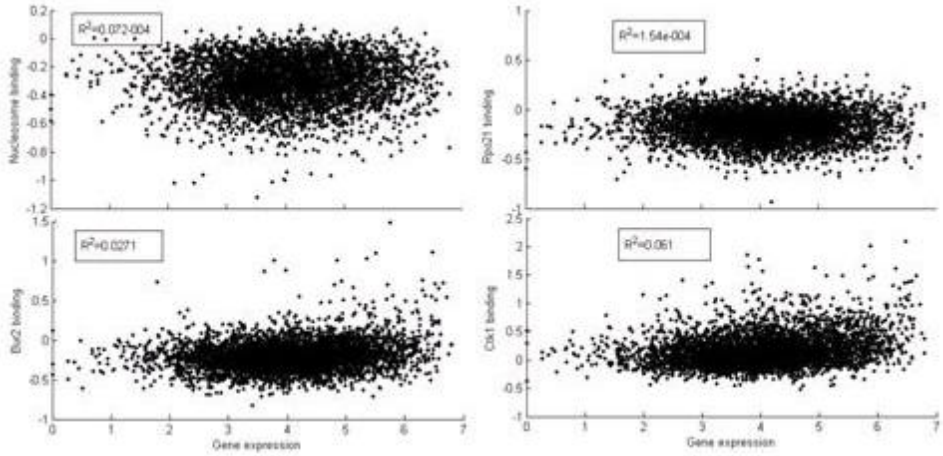

Fig. 1: No correlation between nucleosome/regulatory factor binding and gene expression levels

2.2. Prediction of gene expression levels by binding correlation

Although, genomic studies have shown that it is not possible to determine gene expression levels by nucleosome occupancy information, the controlling role of nucleosome as a repressive factor to transcription still stands. Works considering such information along with other factors such asvarious chromatin modifiers and regulators [12], etc., have attained some insights into gene expression behaviour. Binding of regulatory factors to the underlying DNA may translocate or evict nucleosomes[11]; hence, the resulting nucleosomal organization can be characterized to help explaining gene expression behaviour in general. In our work, we proposed to use binding correlation between nucleosome and elongation factors and derived a multiclass classification model to illustrate its usefulness to predicting gene expression levels.

We computed correlation profiles as input for SVM classifier training, as described in Section 4.2 and 4.3 respectively. For 3 classes of high, medium, and low expression levels, the best parameters for RBF kernel function ($C = 8, \gamma = 0.125$) gave 10-fold cross-validation average accuracy of≈ 69.7%, t-test p-value ≈5e-11(Table 4). This result clearly demonstrates that the interaction information between nucleosome and regulatory binding contributes to distinguish gene expression levels.

2.3. Impact of elongation factors on gene expression levels

Further analysis by feature selection, as described in Section 4.4, gives the best discriminant correlations between nucleosome and elongation factor binding, hereinafter called factors, in distinguishing genes oflow, medium, and high expression levels from the rest (Table 1, factors of order lower than 4 not shown).We examine the best 3 factors for each level (5 distinctive ones in total), with F-score considerably higher than

the others (almost double), which show consistency in their roles during elongation with previous reports in literature.

The elongation complex is composed of RNA polymerase II (PolII) complex and all the elongation factors, which interact with each other. PolII is phosphorylated at its C-terminal repeat domain (CTD) during elongation. Our data contain 3 PollII factors Rbp3, Rbp7, and Rpo21; however, only Rpb3 is discriminative. This is consistent with report by Mayer et al.[13]that Rbp3 reveals PolII enrichment through the transcribed region.Ctk1, a CDT kinase for Ser2 phosphorylationpeaking at about 600bp to 1kb downstreamof TSS, may determine how productive the elongation complex is [13]. CCR4-NOT complex, containing Dhh1 and Not5), was shown by Kruk et al.[14] to directly regulate transcription elongation by promoting the resumption of elongation of arrested PolII once it encounters transcriptional blocks *in vivo*. Lastly, FACT (Pob3) complex, working mostly at 5' end with flaking nucleosomes, was recently found to have a role in turning genes from repressed to active state [15]. Taken together, we may reasonably establish the following:

- Low expression level may be determined mostly by how PolII is active (by Ctk1), how PolII is promoted (by Not5), and if gene is turned to active state (by Pob3).
- Medium and high expression levels are distinguishedby how PolII is enriched downstream (by Rbp3), how PolII is active (by Ctk1), and how PolII is promoted (by Dhh1).

We further attempted to investigate how the bindings of nucleosome and those factors interact by interpreting the Spearman correlation coefficient for each gene, i.e.,> 0 (+) if two bindings increase/decrease together and <0 (-) if two bindings vary inversely. Table 2 shows the dominating direction of correlation for each factor, based on the percentages of genes with that tendency. Only Ctk1 and Rpb3 exhibit clear directions. Rbp3-containing PolII competes with nucleosome for access to DNA, while Ctk1 only interact with PolII complex in the sense that the more repressive nucleosome binding is, the more Ctk1 is needed for PolII to maintain the above medium expression level.

Table 1: Best discriminant nucleosome-elongation factor correlations

Order	Low		Medium		High	
	Factor	**F-score**	**Factor**	**F-score**	**Factor**	**F-score**
1	Ctk1	0.01876	Rpb3	0.01012	Ctk1	0.01238
2	Not5	0.01861	Dhh1	0.00921	Rpb3	0.01179
3	Pob3	0.01336	Ctk1	0.00735	Dhh1	0.00735
4	Cdc39	0.00635	Ctr9	0.00658	Pcf11	0.00569

Table 2: Direction of correlation

Complex	Factor	Direction
CTK	Ctk1	+ (69%)
PolII	Rbp3	- (72%)
CCR4-NOT	Dhh1	+ (56%)
CCR4-NOT	Not5	+ (59%)
FACT	Pob3	- (51%)

3. Conclusion

Nucleosome has been implicated in various cellular processes, including transcription. While its controlling role is widely accepted, how nucleosome individually and cooperatively with other factors takes part in regulatory activities remains elusive. Here, we examinethe interaction of nucleosome and elongation factor binding around TSS to explain gene expression behaviour by quantifying by their correlation value. Our classification model has shown that correlation profiles can distinguish expression levels, to some extent. Further analysis of the most discriminative factors suggests that to attain precise regulation, cell may employ more of differentelongation factorsin each expression level.

4. Methods

4.1. Data preparation

The experimental data were all prepared on *Saccharomyces cerevisiae*grown in YPD medium. Datasets were gathered as follows:

- *Nucleosome binding*: High resolution map of nucleosome occupancies at 4bp resolution across the whole yeast genome was taken from Lee et al. [16].
- *Protein binding*: ChIP-Seq data of 20 regulatory factors (Table 3) were obtained from Venters et al. [2],including RNA polymerase (Pol) II and elongation regulators, as classified by Gene Ontology. The data were distributed in 25 bp windows of $\log_2$ occupancy signal.
- *Gene expression*: Gene expression data were obtained from Xu et al. [17], in which expression values for each gene are tiling-array hybridizations normalized to genomic DNA. For further analysis, we divided values into 3 levels by interval discritization, corresponding to low, medium, and high.
- Genomic annotations of *Saccharomyces cerevisiae* (SGD/sacCer2 assembly) were extracted from the tables provided by the UCSC Genome Browser [18].

We chose to analyse the region of 1200 bp length around TSS (from -350 to 850 bp) as it covers the most informative part of promoter architecture, including ±1 nucleosomes and nucleosome free region (NFR)[3], and the most concentrated binding of elongation regulatory factors [2]. After preprocessing, our data include 4641 genes, belonging to 3 classes (221 low, 2911 medium, and 1509 high) with their corresponding nucleosome-elongation factorcorrelation profiles.

Table 3: List of elongation factors and their complex

Complex	**Factor**
CCR4-NOT	Not5, Dhh1, Cdc39, Not3, Pop2, Ccr4
Pol II	Rpb3, Rbp7, Rpo21
ESS1	Ess1, Bye1
BUR1,2	Bur2
CPF	Pta1
CTK	Ctk1
FACT	Pob3
FCP1	Fcp1
PAF1	Ctr9
PCF11	Pcf11
Iws1	SPT6
Dst1	TFIIS

Table 4: 10-fold cross-validation performance

Turn	**Accuracy**
0	65.9483
1	66.5948
2	65.9483
3	70.3017
4	66.3621
6	69.0086
7	70.5172
8	67.7328
9	67.7086
10	68.6552
Average	67.87776

4.2. Derivation of correlation profile

To derive correlation profiles, we first create the corresponding nucleosome and protein binding profiles in the above-mentioned region. Each region was divided into non-overlapping bins of 150 bp length. As a resulted, each profile is represented as an 8-dimension vector. Then, the correlation values between paired profiles are calculated using non-parametric Spearman's rank correlation. This measurement well reflects the

dependency nature of the interaction between nucleosome and protein binding. Correlation profile for each gene is, then, represented as a 20-dimension vector, showing how the nucleosome occupancy is related to the binding of 20 elongation factors.

4.3. Support vector machine classifier

Given a training set containing instance-class pairs $(x_i, y_i), i = 1, 2, \ldots, l$ where $x_i \in R^l$ and $y_i \in \{-1, 1\}$ is a class label, an SVM classifier is a hyperplane $x^T \varphi(x_i) + b$, where $\varphi(x_i)$ is a function mapping x_i into a higher (maybe infinite) dimensional space, that best separates the two classes. The hyperplane is obtained by solving the following primal optimization problem (1) and its dual quadratic optimization problem (2):

$$\begin{aligned} &\text{Minimize: } \frac{\mathrm{w}^{\mathrm{T}}\mathrm{w}}{2} + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{l}} \xi_{\mathrm{i}} \\ &\text{Subject to: } \mathrm{y}_{\mathrm{i}}\left(\mathrm{w}^{\mathrm{T}}\varphi(\mathrm{x}_{\mathrm{i}}) + \mathrm{b}\right) \geq 1 - \xi_{\mathrm{i}} \\ &\qquad \xi_{\mathrm{i}} \geq 0 \ \mathrm{i} = 1, 2, \cdots, \mathrm{l} \end{aligned} \tag{1}$$

$$\begin{aligned} &\text{Minimize: } \frac{\alpha^{\mathrm{T}}\mathrm{Q}\alpha}{2} - \mathrm{e}^{\mathrm{T}}\alpha \\ &\text{Subject to: } \mathrm{C} \geq \alpha_{\mathrm{i}} \geq 0 \ \mathrm{i} = 1, 2, \cdots, \mathrm{l} \\ &\qquad \mathrm{y}^{\mathrm{T}}\alpha = 0 \end{aligned} \tag{2}$$

where e is an unit vector, $C > 0$ is an error penalty parameter, $Q_{ij} = y_i y_j K(x_i, x_j)$, $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ is a kernel function. In our work, we employed *Radial Basis Function (RBF)* kernel $K(x_i, x_j) = \exp\left(-(x_i - x_i)^2\right)$ to build SVM classifiers.

Our problem, prediction of 3 gene expression levels, is a multiclass classification one. We used LibSVM[19] for parameter learning, training and testing classifier. Performance was evaluated by the average accuracy of 10-fold cross-validation.

4.4. Feature selection with Fisher criterion

Feature selection is a process of selecting a subset of features that contribute the most to distinguishing instances from different classes. We used Fisher score (F-score)to rank the discriminative strength of each feature. This measure is simple, effective and independent of the choice of classification method. The score is defined as follows. Given a dataset X with two classes, denote instances in class 1 as X^1, and those in class 2 as X^2. Assume $\overline{X}_j^k$ is the average of the *jth* feature in X^k, the F-score of the *jth* feature is:

$$\mathrm{F(j)} = \frac{\left(\bar{\mathrm{x}}_{\mathrm{j}}^{1} + \bar{\mathrm{x}}_{\mathrm{j}}^{2}\right)^2}{\left(\mathrm{s}_{\mathrm{j}}^{1}\right)^2 + \left(\mathrm{s}_{\mathrm{j}}^{2}\right)^2} \quad \text{where} \quad \left(\mathrm{S}_{\mathrm{j}}^{\mathrm{k}}\right)^2 = \sum_{\mathrm{x} \in \mathrm{X}^{\mathrm{k}}} \left(\mathrm{x}_{\mathrm{j}} - \bar{\mathrm{x}}_{\mathrm{j}}^{\mathrm{k}}\right)^2 \tag{3}$$

The numerator indicates the discrimination between two classes, and the denominator indicates the scatter within each class. The larger the F-score indicates a more discriminative feature. We took one-vs-all scheme, i.e., converting the multiclass problem into 3 binary classification ones, and calculated F-score of feature sets for each classifier.Gist software package[20] was used.

5. References

[1] K Luger, AW Mader, RK Richmond, DF Sargent, and TJ Richmond, Crystal structure of the nucleosome core particle at 2.8 A resolution, *Nature*, vol. 389, pp. 251-60, 1997.

[2] BJ Venters et al., A Comprehensive Genomic Binding Map of Gene and Chromatin Regulatory Proteins in Saccharomyces, *Molecular Cell*, vol. 41(4), pp. 480-92, 2011.

[3] JB Zaugg and NM Luscombe, A genomic model of condition-specific nucleosome behaviour explains transcriptional activity in yeast, *Genome Research*, vol. Advance access, pp. , 2011.

[4] S Shivaswamy et al., Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation, *PLOS Biology*, vol. 6(3):e65, pp. , 2008.

[5] I Tirosh and N Barkai, Two strategies for gene regulation by promoter nucleosomes, *Genome Research*, vol. 18(7), pp. 1084-91, 2008.

[6] N Kaplan et al., The DNA-encoded nucleosome organization of a eukaryotic genome, *Nature*, vol. 458(7236), pp. 362-6, 2009.

[7] KA Zawadzki, AV Morozov, and JR Broach, Chromatin-dependent transcription factor accessibility rather than

nucleosome remodeling predominates during global transcriptional restructuring in Saccharomyces cerevisiae, *Mol Biol Cell*, vol. 20(15), pp. 3503-13, 2009.

[8] J Wan, J Lin, D Zack, and J Qian, Relating periodicity of nucleosome organization and gene regulation, *Bioinformatics*, vol. 25(41), pp. 1782-8, 2009.

[9] H Yu, S Z S, B Zhou, H Xue, and J Han, Inferring causal relationships among different histone modifications and gene expression, *Genome Research*, vol. 18(8), pp. 1314-24, 2008.

[10] B v Steensel et al., Bayesian network analysis of targeting interactions in chromatin, *Genome Research*, vol. 20(2), pp. 190-200, 2009.

[11] E Segal and J Widom, What controls nucleosome positions?, *Trends Genet.*, vol. 25(8), pp. 335-43, 2009.

[12] DA Gilchrist et al., Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation, *Cell*, vol. 143(4), pp. 540-51, 2010.

[13] A Mayer et al., Uniform transitions of the general RNA polymerase II transcription complex, *Nat Struct Mol Biol.*, vol. 17(10), pp. 1272-8, 2010.

[14] JA Kruk, A Dutta, J Fu, DS Gilmour, and JC Reese, The multifunctional Ccr4-Not complex directly promotes transcription elongation, *Genes Dev.*, vol. 25(6), pp. 581-93, 2011.

[15] S Mahapatra, PS Dewari, A Bhardwaj, and P Bhargava, Yeast H2A.Z, FACT complex and RSC regulate transcription of tRNA gene through differential dynamics of flanking nucleosomes, *Nucleic Acids Res.*, vol. 39(10), pp. 4023-34, 2011.

[16] W Lee et al., A high-resolution atlas of nucleosome occupancy in yeast, *Nat Genet.*, vol. 39(10), pp. 1235-44, 2007.

[17] Z Xu et al., Bidirectional promoters generate pervasive transcription in yeast, *Nature*, vol. 457(7232), pp. 1033-7, 2009.

[18] D Karolchik et al., The UCSC Table Browser data retrieval tool, *Nucleic Acids Res.*, vol. 32(Database issue), pp. D493-6, 2004.

[19] C and Lin, C Chang, LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1--27:27, 2011.

[20] P Pavlidis, I Wapinski, and WS N WS, Support vector machine classification on the web, *Bioinformatics*, vol. 20, pp. 586-587, 2004.