Bioinformatics and Biology Insights

ORIGINAL RESEARCH

Different Functional Gene Clusters in Yeast have Different Spatial Distributions of the Transcription Factor Binding Sites

Wei-Sheng Wu

Lab of Computational Systems Biology, Department of Electrical Engineering, National Cheng Kung University, Taiwan.
Corresponding author email: wessonwu@gmail.com

Abstract: Transcription factors control gene expression by binding to short specific DNA sequences, called transcription factor binding sites (TFBSs), in the promoter of a gene. Thus, studying the spatial distribution of TFBSs in the promoters may provide insights into the molecular mechanisms of gene regulation. I developed a method to construct the spatial distribution of TFBSs for any set of genes of interest. I found that different functional gene clusters have different spatial distributions of TFBSs, indicating that gene regulation mechanisms may be very different among different functional gene clusters. I also found that the binding sites for different transcription factors (TFs) may have different spatial distributions: a sharp peak, a plateau or no dominant single peak. The spatial distributions of binding sites for many TFs derived from my analyses are valuable prior information for TFBS prediction algorithm because different regions of a promoter can assign different possibilities for TFBS occurrence.

Keywords: yeast, promoter, TFBS, spatial distribution

Bioinformatics and Biology Insights 2011:5 1–11

doi: 10.4137/BBI.S6362

This article is available from http://www.la-press.com.

Introduction

Cells regularly face variable environments and can sense many different signals, including temperature, oxidative, and osmotic pressure, beneficial nutrients, and harmful chemicals.[1–5] Through signal transduction pathways, these signals can modulate the activities of transcription factors (TFs). The active TFs can regulate the gene expression of many targets genes to produce appropriate proteins, which work together to enable cells to adapt to new environmental conditions.[6–14] TFs regulate gene expression by binding to specific DNA sequences, called transcription factor binding sites (TFBSs), in the promoters of the target genes. Therefore, studying the spatial distribution of the TFBSs in the promoters may provide insights into the molecular mechanisms of gene regulation.

By using the TFBS data derived from the ChIP-chip data, Harbison et al[15] constructed the spatial distribution of the TFBSs in yeast promoters. They found that binding sites are not uniformly distributed over the promoter regions. Although their findings are interesting, their method has two problems. First, their spatial distribution of the TFBSs was inferred relative to the translation start codon. Since TFs control gene expression at the transcriptional level and transcription is initiated from the transcription start site (TSS), it is more biologically meaningful to infer the spatial distribution of the TFBSs relative to the TSS than to the translation start codon.[16,17] Second, their spatial distribution of the TFBSs was constructed by counting the number of TFBSs located at each site of a promoter. Because the number of promoter sequences that contain a site near the TSS (say 100 bp upstream from the TSS) is much larger than that of a site far upstream from the TSS (say 1000 bp upstream from the TSS), the number of TFBSs found at a site near the TSS will tend to be larger than that of a site far upstream from the TSS. Therefore, to avoid the bias caused by unequal numbers of promoter sequences that contain different sites, it is better to construct the spatial distribution of the TFBSs by counting the frequency of TFBSs than counting the number of TFBSs. The frequency of TFBSs located at a site is calculated by dividing the number of TFBSs located at that site by the number of promoter sequences which contain that site.

In this study, I developed a method to construct the spatial distribution of the TFBSs, which can solve the two problems of Harbison et al's method.[15] My method constructed the spatial distribution of the TFBSs relative to the TSS by counting the frequency of TFBSs located at each site of a promoter. Using my method, I can construct the spatial distribution of the TFBSs for any set of genes of interest. I found that different functional gene clusters have different spatial distributions of the TFBSs. I also found that binding sites of different TFs may have different spatial distributions.

Methods

Datasets

The TFBS locations in the yeast genome for 117 TFs were retrieved from the paper of MacIsaac et al.[18] They used two binding motif discovery algorithms, PhyloCon and Converge, to identify the genome-wide locations of the TFBSs. These two algorithms are based on TF-DNA binding evidence (determined by the *P*-value of the ChIP-chip data) and the phylogenetic conservation constraint (requiring the same binding sites present in the orthologous promoter regions of phylogenetically related yeast species). In this paper, I used four TFBS datasets from MacIsaac et al's paper,[18] which were derived from different ChIP-chip binding *P*-value and phylogenetic conservation constraints: I) ChIP-chip binding *P*-value <0.001 and conserved in at least three of the four yeast species: *S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus*; II) $P < 0.001$ and conserved in at least two yeast species; III) $P < 0.005$ and conserved in at least three yeast species, and IV) $P < 0.005$ and conserved in at least two yeast species.

The genomic coordinates of the TSS of 4560 yeast genes were retrieved from Nagalakshmi et al's paper,[19] in which a high-resolution transcriptome of the yeast genome was generated by a high-throughput RNA-seq method. The lists of 2803 singleton and 1501 duplicate genes were defined according to Ensembl gene family annotation.[20] The lists of 914 essential and 3387 non-essential genes were downloaded from Saccharomyces Genome Deletion Project Website.[21] The lists of 2079 stress and 2290 non-stress genes were defined according to Gasch et al's paper.[3] A gene is called a stress gene if it has a fold change larger than two under at least one of the five stress conditions: heat shock, oxidative shock,

osmotic shock, amino acid starvation, and nitrogen depletion.

Constructing the spatial distribution of the TFBSs for a functional gene cluster

Let A be a functional gene cluster, eg, essential genes, singleton genes, stress genes, etc. The procedure of constructing the spatial distribution of the TFBSs for A is as follows. Let x be the site relative to the TSS, ie, $x = -10$ stands for the site that is 10 bp upstream from the TSS. For each x in the promoter of a gene in A, I checked whether a TFBS is located at that site or not by using the TFBS data of 117 TFs in MacIsaac et al's paper.[18] The same process was applied to all genes in A. Then I counted the total number, $n(x)$, of TFBSs located at site x for all genes in A. Finally, the TFBS frequency $f(x)$ at site x was obtained by dividing $n(x)$ by the total number, $s(x)$, of promoter sequences in A that contain site x, ie, $f(x) = n(x)/s(x)$. For example, in the spatial distribution of the TFBSs for the 2803 singleton genes in yeast (see Fig. 1), I found $f(-116) = 0.073 = 196/2684$, which means that at the site 116 bp upstream from the TSS, 196 TFBSs were found to occupy that site and 2684 promoter sequences of the 2803 singleton genes contain that site. Note that all the spatial distributions shown in the figures are smoothed with a sliding window of size 41 bp.

Identifying the enriched TFBSs in a functional gene cluster

Let A be a functional gene cluster, eg, essential genes, singleton genes, stress genes, etc. The procedure of identifying the enriched TFBSs in A is as follows. A model based on hypergeometric distribution[10] is used to test whether the enrichment of a specific TFBS (eg, Abf1) in A is statistically higher than random expectation. The P-value for rejecting the null hypothesis that enrichment of the specific TFBS in A is by chance can be computed as the following formula:

$$p = P(x \geq m_a) = \sum_{x \geq m_a} \frac{\binom{n_a}{x}\binom{N-n_a}{M-x}}{\binom{N}{M}}$$

$$= 1 - \sum_{x=0}^{m_a-1} \frac{\binom{n_a}{x}\binom{N-n_a}{M-x}}{\binom{N}{M}}$$

where $N = 6576$ is the number of genes in the yeast genome, M is the number of genes in A, n_a is the number of genes (in the yeast genome) whose promoter

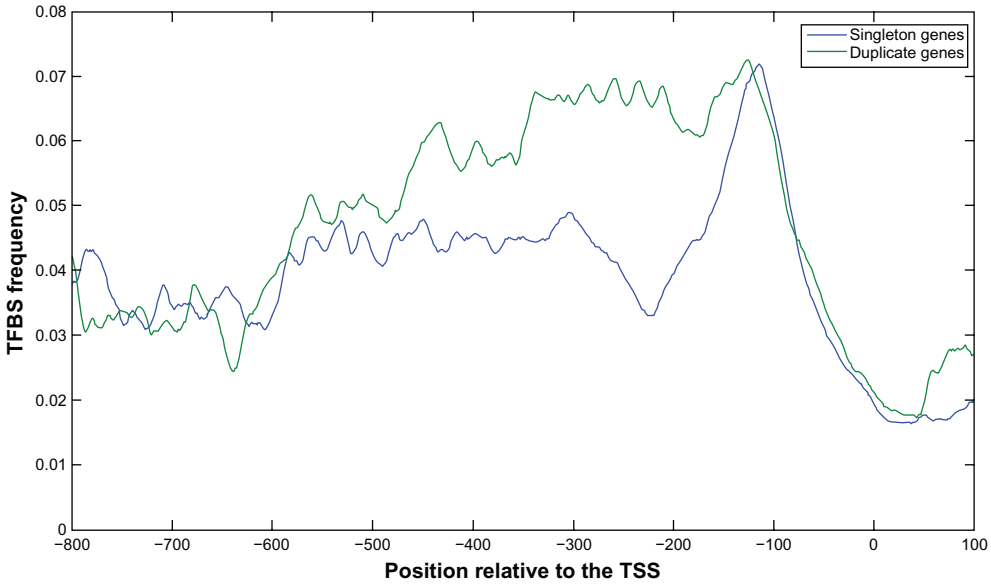

Figure 1. The spatial distributions of the TFBSs for singleton and duplicate genes. It can be seen that singleton genes have a sharply peaked distribution of the TFBSs, whereas duplicate genes have a dispersed distribution of the TFBSs. The difference between the spatial distributions for singleton and duplicate genes is statistically significant (K-S test, P-value $<10^{-4}$).

contain the specific TFBS, m_a is the number of genes (in A) whose promoter contain the specific TFBS, and $\binom{n_a}{m_a} \triangleq \frac{n_a!}{m_a!(n_a-m_a)!}$. This procedure is applied for each of the 117 TFBSs used in this study.

Constructing the random expectation of the spatial distribution of the TFBSs

The random expectation of the spatial distribution of the TFBSs is calculated by averaging the distributions constructed using 1,000 "randomized" genomes in which the binding sites in each promoter region were redistributed randomly and independently in each promoter region.[15]

Statistical testing for the difference between two spatial distributions of the TFBSs

Kolmogorov-Smirnov (K-S) test was used to test if two spatial distributions of the TFBSs are statistically different in location or shape.[22] Assume that $F_n(x)$ is the empirical distribution constructed by the n samples collected from the first distribution and $F_m(x)$ is the empirical distribution constructed by the m samples collected from the second distribution, then the Kolmogorov–Smirnov statistic $D_{n,m} = \sup_x |F_n(x) - F_m(x)|$ quantifies a distance between $F_n(x)$ and $F_m(x)$. The null hypothesis of K-S test is that these two distributions are the same and the null hypothesis is rejected with a P-value equals to $P\left(K > \sqrt{nm/n+m} \cdot D_{n,m}\right)$ where K denotes the Kolmogorov distribution and the cumulative distribution function of K is given by $P(K \leq x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1} e^{-2i^2x^2}$.

Results

The spatial distributions of the TFBSs for singleton genes and duplicate genes are significantly different

Gene duplication plays an important role in evolution because it is the primary source of new genes. Many studies showed that gene duplicability varies considerably among genes.[23–25] Some genes only have a single copy whereas the other genes have multiple copies in an organism. It would be interesting to know whether singleton and duplicate genes have different gene regulation mechanisms. I studied this issue by constructing the spatial distributions of the TFBSs for singleton and duplicate genes. I found that singleton genes have a sharply peaked distribution of the TFBSs, whereas duplicate genes have a dispersed distribution of the TFBSs (see Fig. 1). The difference between the spatial distributions for singleton and duplicate genes is statistically significant (K-S test, P-value $<10^{-14}$). Moreover, the binding sites of Abf1 were found to be enriched in singleton genes but not in duplicate genes, while the binding sites of Rap1, Fhl1, Sfp1, Yap5 and Msn2 were found to be enriched in duplicate genes but not in singleton genes (see Table 1). This suggests that the gene regulation mechanisms for these two kinds of genes may be different.

The spatial distributions of the TFBSs for essential genes and non-essential genes are significantly different

Essential genes in yeast are those genes required for laboratory growth on rich media.[21] The deletion of any one of these genes is sufficient to confer a lethal phenotype. Such genes make excellent potential drug targets.[26] It is estimated that 17.8% of the

Table 1. The enriched TFBSs (P-value <0.001) in the functional gene cluster under study (see Supplementary Table 2 for details).

Singleton genes	Abf1	**Duplicate genes**	Rap1, Fhl1, Sfp1, Yap5, Msn2
Essential genes	Abf1, Rpn4, Reb1	**Non-essential genes**	Skn7, Phd1, Nrg1, Sut1, Sok2, Cin5
Non-stress genes	None	**Stress genes**	Fhl1, Gcn4, Cin5, Skn7, Rap1, Gln3, Sfp1, Xbp1, Ume6, Phd1, Yap7, Hsf1, Sok2, Pho2, Msn2, Yap6, Yap1, Swi4, Bas1,Sut1, Rox1, Yap5, Nrg1, Met32, Ino4

yeast genome is essential.[27,28] On the other hand, non-essential genes are those genes when deleted may have some fitness effects but the yeast still can survive. It would be interesting to know whether essential and non-essential genes have different gene regulation mechanisms. I studied this issue by constructing the spatial distributions of the TFBSs for essential and non-essential genes. I found that essential genes have a sharply peaked distribution of the TFBSs, whereas non-essential genes have a dispersed distribution of the TFBSs (see Fig. 2). The difference between the spatial distributions for essential and non-essential genes is statistically significant (K-S test, P-value $<10^{-5}$). Moreover, the binding sites of Abf1, Rpn4 and Reb1 were found to be enriched in essential genes but not in non-essential genes, while the binding sites of Skn7, Phd1, Nrg1, Sut1, Sok2 and Cin5 were found to be enriched in non-essential genes but not in essential genes (see Table 1). This suggests that the gene regulation mechanisms for these two kinds of genes may be different.

The spatial distributions of the TFBSs for stress genes and non-stress genes are significantly different

Stress genes are those genes whose protein products can help cells fight against the deleterious effects induced by environmental stresses such as high temperature, high acidity, nutrient depletion, etc.[1–5] On the other hand, non-stress genes are those genes whose functions are not related to the cell's complex stress adaptation mechanism. It would be interesting to know whether stress and non-stress genes have different gene regulation mechanisms. I studied this issue by constructing the spatial distributions of the TFBSs for stress and non-stress genes. I found that non-stress genes have a sharply peaked distribution of the TFBSs, whereas stress genes have a dispersed distribution of the TFBSs (see Fig. 3). The difference between the spatial distributions for stress and non-stress genes is statistically significant (K-S test, P-value $<10^{-24}$). Moreover, no TFBSs were enriched in non-stress genes but the binding sites of 25 TFs were enriched in stress genes (see Table1). This suggests that the gene regulation mechanisms for these two kinds of genes may be different.

The spatial distributions of the binding sites for different TFs are significantly different

In the previous results, the spatial distribution of the TFBSs for a functional gene cluster as constructed by considering the binding site data of all TFs. It would be interesting to know whether the spatial distributions of the binding sites for different TFs are different. Therefore, I constructed the spatial distribution of the

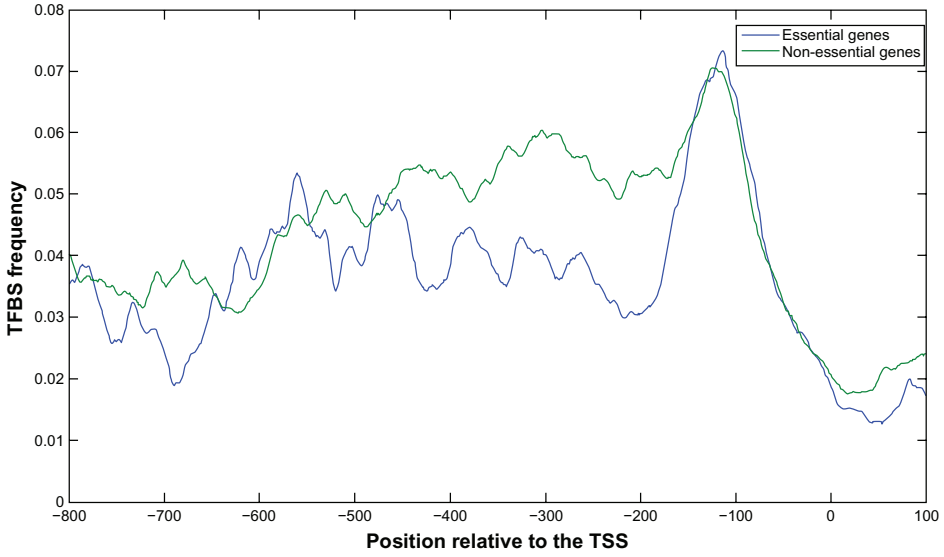

Figure 2. The spatial distributions of the TFBSs for essential and non-essential genes. It can be seen that essential genes have a sharply peaked distribution of the TFBSs, whereas non-essential genes have a dispersed distribution of the TFBSs. The difference between the spatial distributions for essential and non-essential genes is statistically significant (K-S test, P-value $<10^{-5}$).

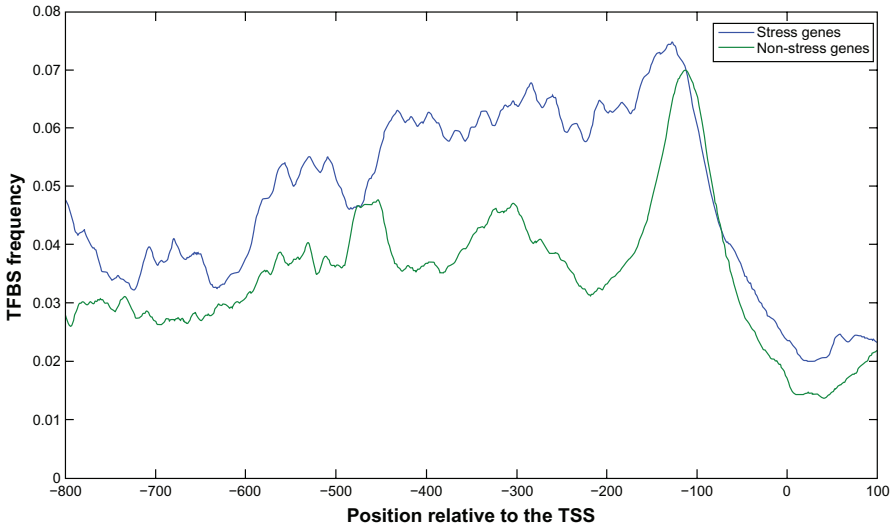

Figure 3. The spatial distributions of the TFBSs for stress and non-stress genes. It can be seen that non-stress genes have a sharply peaked distribution of the TFBSs, whereas stress genes have a dispersed distribution of the TFBSs. The difference between the spatial distributions for stress and non-stress genes is statistically significant (K-S test, *P*-value $<10^{-24}$).

binding sites for each available TF.[18] For example, 356 Abf1's binding sites in the promoter regions of 308 Abf1's target genes in the yeast genome are used to construct the spatial distribution of the binding sites for Abf1. After constructing the spatial distributions of the binding sites for various TFs, I found that these spatial distributions could be divided into three categories (see Fig. 4): 49 TFs in the first category, 27 TFs in the second category and 41 TFs in the third category (see Supplementary Table 1 for the detailed list of the TFs in each category). The spatial distributions of the binding sites for the first group of TFs, eg, Abf1, Reb1, and Hap4, have a sharp peak and can be fitted by a Power-law distribution ($p(x) \sim (x + 150)^{-0.2}$), suggesting that a specific distance between the binding sites and the TSS is required for proper functioning of these TFs. The distributions for the second group of TFs, eg, Rap1, Fhl1, and Skn7, have a plateau and can be fitted by a Gaussian distribution ($p(x) \sim N(\mu = -300, \sigma = 200)$), suggesting that a range of distances between

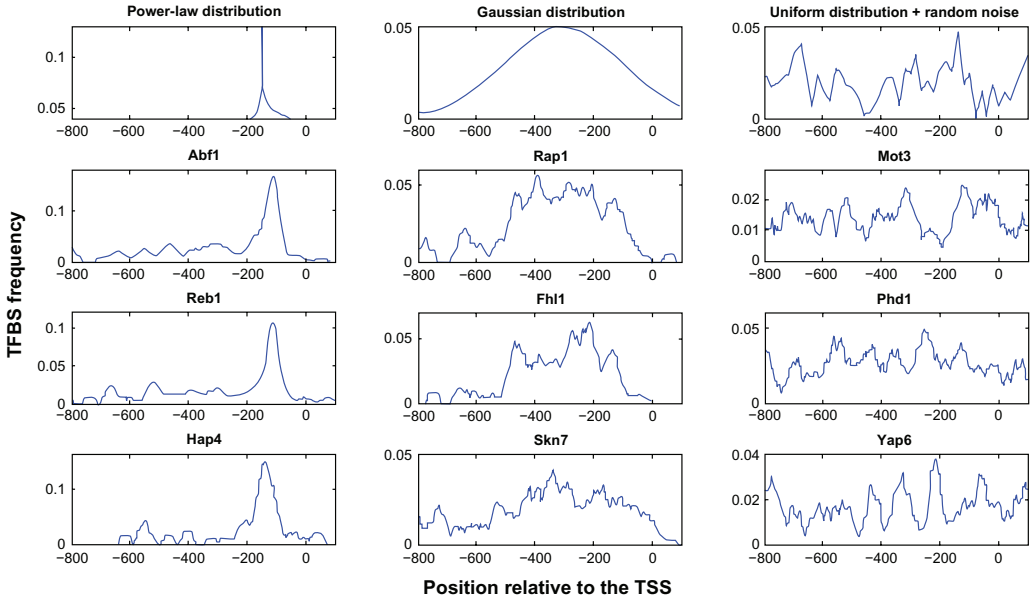

Figure 4. The spatial distributions of the binding sites for different TFs. It can be seen that spatial distributions of the binding sites for different TFs are different and can be roughly divided into three categories: I) the distributions have a sharp peak, II) the distributions have a plateau, and III) the distributions do not have any dominant peak.

the binding sites and the TSS is important for proper functioning of these TFs. However, the distributions for the third group of TFs, eg, Mot3, Phd1, and Yap6, do not have any dominant peak and can be fitted as an uniform distribution plus a random noise, suggesting that the distance between the binding sites and the TSS is not important for proper functioning of these TFs.

Discussions

The proposed method can extract biologically meaningful results

To show that my method can extract more biological insights than Harbison et al's approach,[15] I constructed the spatial distribution of the TFBSs for all genes in the yeast genome by using the TSS (in this study) and the translation start codon (in Harbison et al's study) as the reference point, respectively. When the TSS is used as the reference point, I found that the TFBSs are highly enriched in a ~100 bp region (ranging from 80 to 180 bp upstream from the TSS) with a sharp peak at −115 bp. The sharp peak is significantly higher than random expectation, supporting a strong positioning bias of the TFBSs relative to the TSS (see Fig. 5). In contrast, no sharp peak can be observed when the start codon was used as the reference point (see Fig. 5). This observation suggests my method is more powerful than Harbison et al's approach to construct biologically meaningful spatial distributions of the TFBSs.

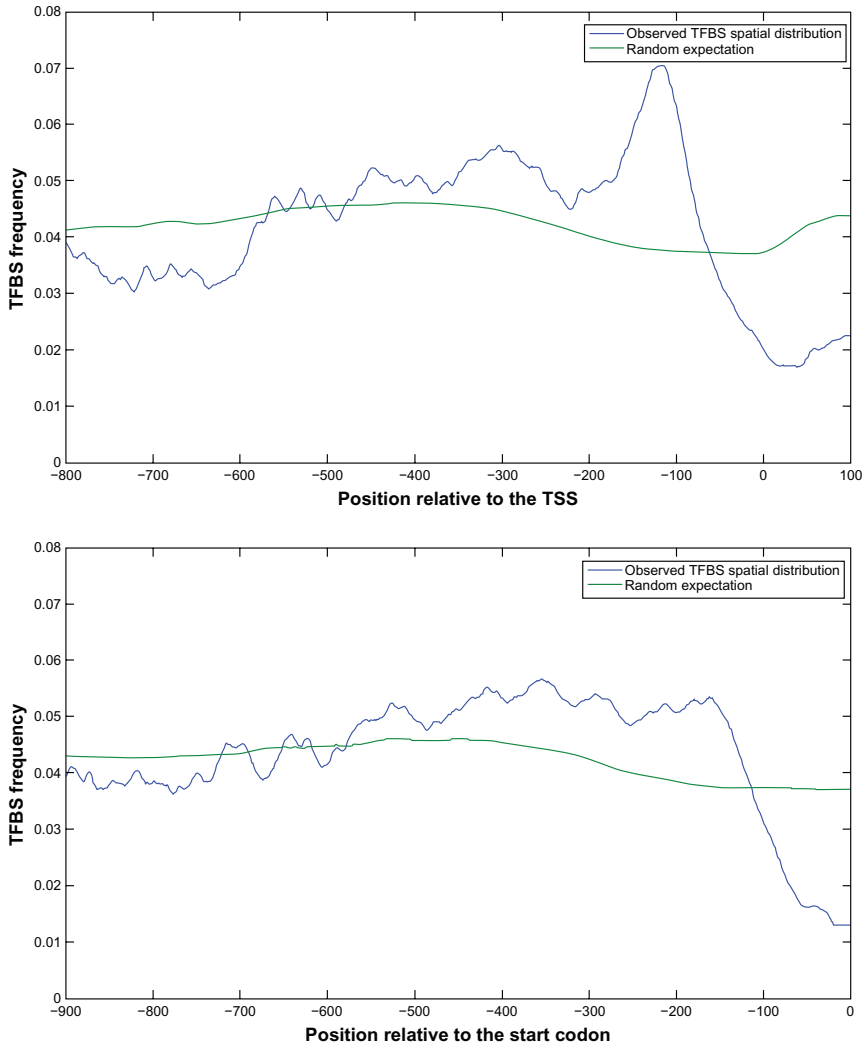

Figure 5. The spatial distributions of the TFBSs relative to the TSS v.s. the start codon. I constructed the spatial distribution of the TFBSs for all genes in the yeast genome by using the TSS (in this study) and the translation start codon (in Harbison et al's study) as the reference point, respectively. A sharp peak can be seen in the spatial distribution of the TFBSs relative to the TSS and the sharp peak is significantly higher than random expectation, supporting a strong positioning bias of the TFBSs relative to the TSS. However, such sharp peak cannot be seen in the spatial distribution of the TFBSs relative to the start codon, suggesting that my method is more powerful than Harbison et al's approach to construct biologically meaningful spatial distributions of the TFBSs.

The proposed method is robust

Since the total number of TFBSs that could be identified strongly depends on the parameter settings in TFBS prediction algorithms, I performed my analyses using four different TFBS datasets to ensure the robustness of my findings. I found that these four distributions of the TFBSs have very similar patterns, showing the robustness of my results (see Fig. 6a). Moreover, since TFBSs in divergent promoters (a common promoter region shared by two divergently transcribed adjacent genes) cannot be assigned to one of the two adjacent genes unambiguously, I excluded these TFBSs from my analyses to avoid the potential bias. I found that the distributions of the TFBSs are almost the same whether I included divergent promoters in my analyses or not, showing that my method is very robust (see Fig. 6b). In addition, I used three different window sizes (31 bp, 41 bp, and 51 bp) of the sliding window for smoothing the spatial distributions of the TFBSs. I found that these three distributions of the TFBSs have very similar patterns, showing the robustness of my results (see Fig. 7).

The spatial distributions of the binding sites for various TFs are useful

The spatial distributions of the binding sites for different TFs derived from my analyses can be used as valuable prior information for TFBS prediction.

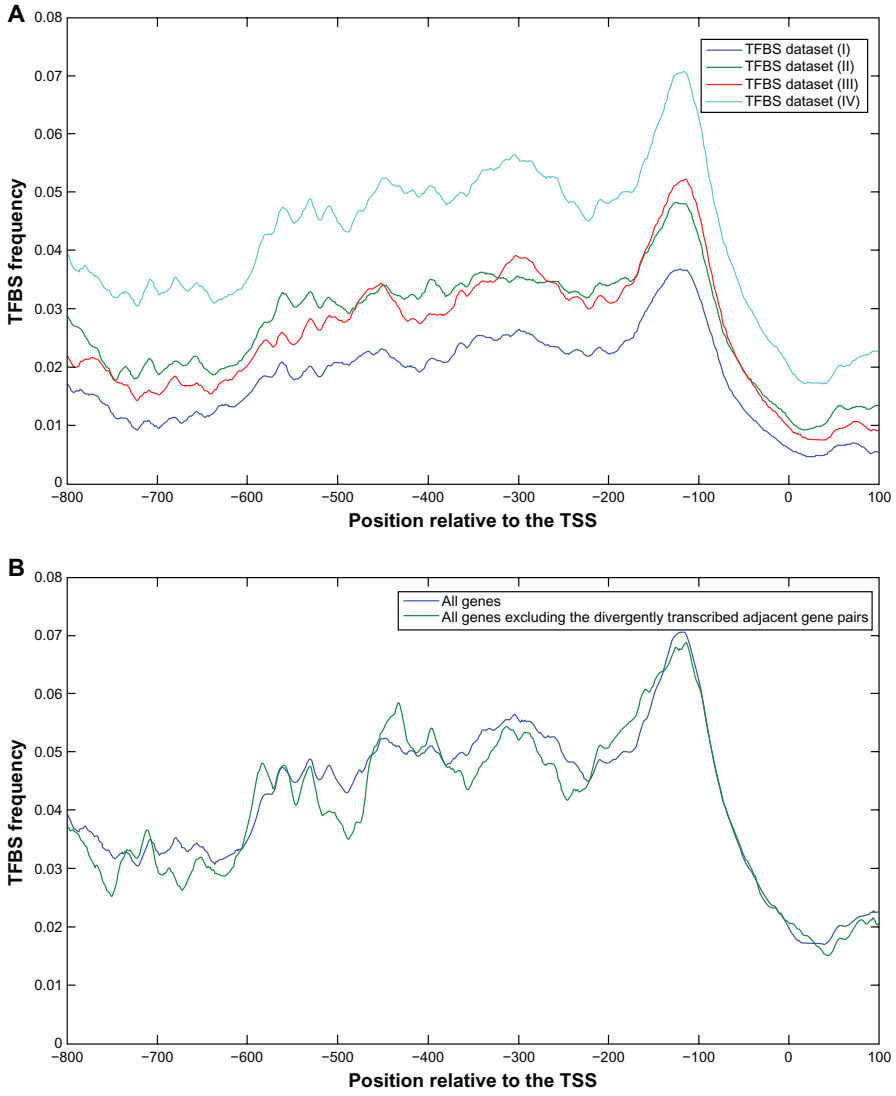

Figure 6. The robustness of my method. **A**) The four spatial distributions of the TFBSs constructed by using four different TFBS datasets have very similar patterns, showing the robustness of my method. **B**) The two spatial distributions of the TFBSs constructed by including or excluding the divergently transcribed adjacent gene pairs are almost the same, showing that my method is very robust.

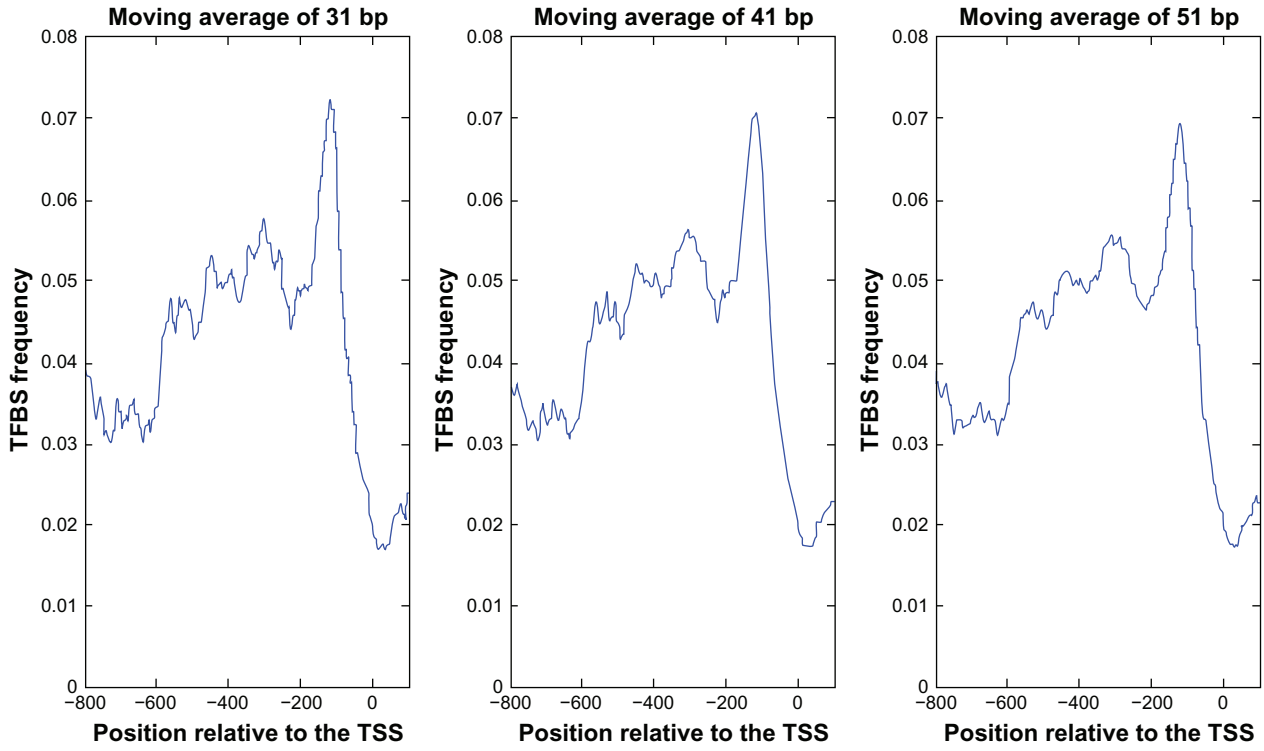

Figure 7. Using three different window sizes for smoothing the spatial distribution of the TFBSs. The three spatial distributions of the TFBSs constructed by using three different window sizes (31 bp, 41 bp, and 51 bp) of the sliding window have very similar patterns, showing my result is not sensitive to the chosen window size.

For example, when a TFBS prediction algorithm is performed to search Abf1's binding sites in a promoter of interest, it should focus on the DNA sequences from 100 bp to 200 bp upstream from the TSS because most of Abf1's binding sites prefer to locate in this region (see Fig. 4). In contrast, when search Yap6's binding sites in a promoter of interest, every region of the promoter should have equal attention because Yap6's binding sites have no preference to locate in a particular region of the promoter (see Fig. 4).

Conclusions

I developed a method to construct the spatial distribution of the TFBSs for any set of genes of interest. Unlike Harbison et al's approach, I used the transcription start site (TSS) as the reference point and counted the frequency of TFBSs for each site in the promoter of a gene. These two features make my method more biologically meaningful than their approach. Besides, I constructed the spatial distributions of the TFBSs using four different TFBS datasets and found highly consistent results, showing the robustness of my method. Moreover, I used my method to construct the spatial distributions of the TFBSs for different functional gene clusters. I found that singleton genes, essential genes and non-stress genes have sharply peaked spatial distributions of the TFBSs, whereas duplicate genes, non-essential genes, and stress genes have dispersed spatial distributions of the TFBSs. In addition, I found that binding sites for different TFs may have different spatial distributions. For example, the spatial distributions of the binding sites of Abf1, Reb1 and Hap4 all have a sharp peak; the distributions of the binding sites of Rap1, Fhl1 and Skn7 all have a plateau, but the distributions of the binding sites of Mot3, Phd1, and Yap6's do not have any dominant peak. In summary, my results show that different functional gene clusters have different spatial distributions of the TFBSs, indicating that gene regulation mechanisms may be very different among different functional gene clusters.

Disclosures

This manuscript has been read and approved by the author. This paper is unique and not under consideration by any other publication and has not been published elsewhere. The author and peer reviewers report no conflicts of interest. The author confirms

that they have permission to reproduce any copyrighted material.

Acknowledgements

I thank Drs. Wen-Hsiung Li, Zhenguo Lin, and Han Liang for helpful discussion. This study was supported by the Taiwan National Science Council NSC 99-2628-B-006-015-MY3.

References

1. Hohmann S, Mager WH. *Yeast Stress Responses*. Berlin: Springer-Verlag; 2003.
2. Wu WS, Li WH. Identifying gene regulatory modules of heat shock response in yeast. *BMC Genomics*. 2008;9:439.
3. Gasch AP, Spellman PT, Kao CM, et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*. 2000;11:4241–57.
4. Causton HC, Ren B, Koh SS, et al. Remodeling of Yeast Genome Expression in Response to Environmental Changes. *Mol Biol Cell*. 2001;12:323–37.
5. Gasch AP, Werner-Washburne M. The genomics of yeast responses to environmental stress and starvation. *Funct Integr Genomics*. 2002;1:181–92.
6. Wu WS, Li WH, Chen BS. Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. *BMC Bioinformatics*. 2006;7:421.
7. Alberghina L, Westerhoff HV. *Systems Biology: Definition and Perspectives*. Berlin: Springer-Verlag; 2005.
8. Palsson B. *Systems Biology: Properties of Reconstructed Networks*. New York: Cambridge University Press; 2006.
9. Alon U. *An Introduction to Systems Biology*. Florida: CRC press; 2007.
10. Wu WS, Li WH, Chen BS. Identifying regulatory targets of cell cycle transcription factors using gene expression and ChIP-chip data. *BMC Bioinformatics*. 2007;8:188.
11. Lin Z, Wu WS, Liang H, et al. The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcription regulation. *BMC Genomics*. 2010;11:581.
12. Wu WS. Identifying highly confident TF-gene regulatory relationships in yeast. *International Journal of Systems and Synthetic Biology*. 2010;1(1):121–33.
13. Tsai HK, Lu HH, Li WH. Statistical methods for identifying yeast cell cycle transcription factors. *Proc Natl Acad Sci U S A*. 2005;102:13532–13537.
14. Wu WS, Li WH. Systematic identification of yeast cell cycle transcription factors using multiple data sources. *BMC Bioinformatics*. 2008;9:522.
15. Harbison CT, Gordon DB, Lee TI, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*. 2004;431:99–104.
16. Kleinsmith LJ, Kish VM. *Principles of cell and molecular biology*, 2nd ed. New York: HarperCollins; 1995.
17. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *Molecular biology of the cell*, 4th ed. New York: Garland Science; 2002.
18. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*. 2006;7:113.
19. Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008;320:1344–9.
20. Ensembl. http://www.ensembl.org.
21. Saccharomyces Genome Deletion Project Website. http://www-sequence.stanford.edu/group/yeast_deletion_project/Essential_ORFs.txt.
22. Mendenhall W, Sincich T. *Statistics for Engineering and the Sciences,* 4th ed. Englewood Cliffs: Prentice-Hall; 1995.
23. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. Role of duplicate genes in genetic robustness against null mutations. *Nature*. 2003;421:63–6.
24. Zhang J. Evolution by gene duplication: an update. *Trends Ecol Evol*. 2003;18:292–8.
25. Taylor JS, Raes J. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*. 2004;38:615–43.
26. Seringhaus M, Paccanaro A, Borneman A, Snyder M, Gerstein M. Predicting essential genes in fungal genomes. *Genome Res*. 2006;16:1126–35.
27. Cole ST. Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur Respir J Suppl*. 2002;36:78s–86s.
28. Winzeler EA, Shoemaker DD, Astromoff A, et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*. 1999;285:901–6.

Supplementary Data

Supplementary Table 1. Contains the detailed lists of the TFs for the three different categories.

Supplementary Table 2. Contains the enriched TFBSs for each functional gene cluster under study.