

Combining Genome-Wide ChIP and Mutant Expression Data to Decipher the Regulatory Logic of MBF and SBF during the Yeast Cell Cycle

Feng Gao⁽¹⁾ and Harmen J. Bussemaker^(1,2)

⁽¹⁾ Department of Biological Sciences and ⁽²⁾ Center for Computational Biology and Bioinformatics, Columbia University, 1212 Amsterdam Ave MC 2441, New York, NY 10027, USA
Email: fg2037@columbia.edu, hjb2004@columbia.edu

ABSTRACT

Different transcription factors can form a complex and work together to induce or repress gene expression. Here we show how regression analysis can be used to uncover associations between transcription factors. MBF, a heterodimer of Mbp1 and Swi4, and SBF, a heterodimer of Swi4 and Swi6, are two well-studied cell-cycle regulators in yeast. Genome wide ChIP log ratio for Mbp1, Swi4 and Swi6 were converted into promoter occupancies. Multivariate regression analysis was subsequently used to explain the transcriptional response to deletions of Mbp1 or Swi4 in terms of a cell combination of these occupancies. This approach naturally uncovers association between transcription factors. The results showed that MBF has significant correlation with Mbp1 deletion gene expression data, while SBF showed a significant correlation with Swi4 deletion gene expression data. This technique is very general and has potential for detection of context dependent transcription factor association.

Key words: Transcription factor, Combinatorial regulation, Microarray analysis, ChIP, Promoter occupancy, Regression analysis

Introduction

Transcription factors (TF) have the ability to bind certain DNA sequence elements and regulate mRNA transcription. In yeast, 141 transcription factors have been reported to have DNA binding and transcription regulation activity [1]. Chromatin immunoprecipitation (ChIP) can be used to analyze the genome-wide binding profile of each transcription factor. In a recent report, Lee et al determined binding profiles for 131 transcription factors for 6270 genes in yeast [2]. When regulating gene expression, transcription factors can form a complex that either induces or represses gene expression. MBF, a heterodimer of transcription factor Mbp1 and Swi4, and SBF, a heterodimer of Swi4 and Swi6, are two well-studied combinations regulating gene expression during the G1/S phase in yeast cell cycle [3, 4]. We previously used regression analysis to explain expression data in terms of counts of cis-regulatory elements in the promoter of each gene, and developed an algorithm called REDUCE to identify DNA binding motifs based on a single microarray experiments [5]. Here we use regression analysis to explain expression data in terms of transcription factor occupancies of each promoter, as measured in ChIP experiments. Combining ChIP data for Mbp1, Swi4 and Swi6 with expression data for Mbp1 and Swi4 deletion strains, we could predict that transcription regulation by these factors is conferred only through the heterodimers MBF and SBF.

Methods

Transcription factor binding and gene expression data: We downloaded the microarray log-ratios for the ChIP experiments on the transcription factors Mbp1, Swi4 and Swi6 performed by Lee et al. [2] and the gene expression data for the Mbp1 and Swi4 deletion experiments published by Hughes et al. [6].

Converting ChIP log-ratio to promoter occupancy: For each combination of transcription factor f and gene g , we converted the ChIP log-ratio L to a promoter occupancy O using the following transformation:

$$O_{f,g} = f((L_{f,g} - \alpha_f) / \beta_f) \quad (1)$$

Here $f(x)$ is the sigmoidal function given by $f(x) = e^x / (1 + e^x)$. The parameters α and β determine the position and slope of the distribution respectively. Since occupancies are always positive and insensitive to outliers in the log-ratio, it makes sense to multiply them to analyze the effect of transcription factor combinations on expression. To fix α and β , we first identified the ten most significant motifs found by the REDUCE algorithm among all oligonucleotides up to heptamers from the ChIP data for Mbp1, Swi4 and Swi6 [5]. Values for α and β were subsequently determined that maximize the sum of R^2 cover the top ten motifs.

Regression Analysis: After converting binding log ratio to occupancy, we fit the following multivariate model to the expression log-ratios E_g for gene g between a wild type strain and a mutant for one of the two proteins Mbp1 and Swi4:

$$E_g = C + F_1 O_{1,g} + F_2 O_{2,g} + F_3 O_{3,g} + F_{1,2} O_{1,g} O_{2,g} + F_{1,3} O_{1,g} O_{3,g} + F_{2,3} O_{2,g} O_{3,g} + F_{1,2,3} O_{1,g} O_{2,g} O_{3,g} \quad (2)$$

The intercept C is close to zero and F_i etc. are regression coefficients or slopes. Terms are included for all possible combinations of the three transcription factors. Statistical parameters (t -value and p -value) were computed for each coefficient using the R software package.

Results

We first validated the transformation to occupancies according to equation (1) by performing regression of Mbp1 deletion expression data on the binding data for Mbp1. A similar analysis was performed for Swi4. The value of R^2 increased from 1.77% (log-ratio) to 2.63% (occupancy) for Mbp1 ($\alpha = 1.7$ and $\beta = 0.5$), and from 0.38% (log-ratio) to 0.94% (occupancy) for Swi4 ($\alpha = 2.0$ and $\beta = 0.7$). Next, we fit the model of equation (2) to the Mbp1 and Swi4 deletion data. The results are listed in table 1 and 2. Only the product of Mbp1 and Swi6 occupancy showed a significant regression coefficient in the Mbp1 deletion experiment, while only the product of Swi4 and Swi6 was significant in the Swi4 deletion experiment.

Table 1. Coefficient result of regression analysis in Mbp1 deletion experiment

| Potential Elements | t value | Pr(> t) |
|--------------------|---------|----------|
| Mbp1 | -0.681 | 0.49571 |
| Swi4 | 0.765 | 0.44450 |
| Swi6 | -0.999 | 0.31774 |
| Mbp1:Swi4 | -1.148 | 0.25110 |
| Mbp1:Swi6 | 2.847 | 0.00443* |
| Swi4:Swi6 | 0.258 | 0.79675 |
| Mbp1:Swi4:Swi6 | 0.811 | 0.41728 |

* Significant statistical difference, $p < 0.05$.

Table 2. Coefficient result of regression analysis in Swi4 deletion experiment

| Potential Elements | t value | Pr(> t) |
|--------------------|---------|----------|
| Mbp1 | 0.902 | 0.3672 |
| Swi4 | 1.060 | 0.2892 |
| Swi6 | -0.181 | 0.8563 |
| Mbp1:Swi4 | -1.466 | 0.1426 |
| Mbp1:Swi6 | -0.973 | 0.3305 |
| Swi4:Swi6 | -2.314 | 0.0207* |
| Mbp1:Swi4:Swi6 | 1.494 | 0.1352 |

* Significant statistical difference, $p < 0.05$.

Discussion

Our proposed sigmoidal transformation of ChIP microarray data log-ratios to promoter occupancies reduces noise for small binding log-ratios and in addition suppresses the effect of outliers. The inherent merit of using inferred occupancies instead of log-ratios is clear from the fact that the value of R^2 for univariate regression on the ChIP data of the transcriptional response to the deletion of a transcription factor significantly improves for both the Mbp1 and the Swi4 deletion experiments.

Different transcription factors can form a complex and work together to regulate gene expression. As we have shown, the occupancy representation for transcription factor binding data naturally allows for multivariate regression of gene expression on transcription factor combinations to be used as a methodology for detecting transcription factor associations using only binding and gene expression data as input. Our unbiased analysis showed that it is necessary for both Mbp1 and Swi4 to form a complex with Swi6, in agreement with the previous experimental observations [3,4]. We expect our approach to be generally useful for deciphering genome-wide regulatory network in yeast, and possibly higher eukaryotes as well.

Acknowledgements. We thank Marcel van Batenburg, Barrett Foat, and Crispin Roven for assistance and helpful suggestions. F.G. was supported by a grant from the Netherlands organization of Scientific Research (NWO).

References

- [1] M.C. Costanzo, J.D. Hogan, M.E. Cusick, B.P. Davis, A.M. Fancher, P.E. Hodges, P. Kondu, C. Lengieza, J.E. Lew-Smith, C. Lingner, K.J. Roberg-Perez, M. Tillberg, J.E. Brooks and J.I. Garrels, The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Research*, 28(1):73-76, 2000
- [2] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford and R.A. Young, Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799-804, 2002
- [3] C. Koch, T. Moll, M. Neuberg, H. Ahorn and K. Nasmyth, A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science*, 261(5128):1551-7, 1993
- [4] B.J. Andrews and I. Herskowitz, The yeast SWI4 protein contains a motif present in developmental regulators and is part of a complex involved in cell-cycle-dependent transcription. *Nature*, 342(6251):830-3, 1989
- [5] H.J. Bussemaker, H. Li and E.D. Siggia, Regulatory element detection using correlation with expression. *Nature Genetics*, 27(2):167-71, 2001
- [6] T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburty, J. Simon, M. Bard and S.H. Friend, Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109-26, 2000