Integrative Method for Identifying Combinatorial Regulation of Transcription Factors

Mamoru Kato[1,2] kato@src.riken.jp
Naoya Hata[2] hata@cshl.edu
Nila Banerjee[2] banerjee@cshl.edu
Michael Q. Zhang[2] mzhang@cshl.edu

[1] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, U.S.A.
[2] Laboratory for Medical Informatics, SNP Research Center, RIKEN, 1-7-22 Suehiro, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

Keywords: integrative genomics, combinatorial regulation, DNA microarray, ChIP microarray

1 Introduction

To identify combinatorial regulation of transcription factors (TFs) and their binding motifs is important for understanding gene expression. However, the customary approach [2, 3] in computational microarray analysis is to cluster gene expression patterns and to identify individual sequence motifs specific to each gene cluster. The limitations of this approach are: 1) it does not directly address combinatorial regulation by TFs; 2) Even if a sequence motif is found, it cannot identify the relevant TF(s) to the motif. To overcome these limitations, we propose a novel method that integrates chromatin immunoprecipitation (ChIP) microarray data, DNA microarray data, and combinatorial motif analysis to identify combinatorial regulation of TFs and their binding motifs. We applied this method to yeast cell cycle, and searched up to 3-order combinations of motifs with 6-9 mer.

2 Methods

Our method consisted of four steps. 1) We identified over-represented single motifs. For target genes (promoters) of a TF from each of 113 ChIP datasets [1], we took the intersection of these targets with each of the cell cycle-regulated gene set and the sub-classified gene sets of Spellman *et al* [2]. We defined each of these intersection sets as a foreground set. As a background set, we chose the intersection of non-target genes of a given TF and non-cell cycle genes. For each foreground-background pair, and for all possible motif words (6-9 bps), we calculated the statistic of the 2×2 contingency table test:

$$T = \frac{N\left(|ad - bc| - N/2\right)^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)}, \tag{1}$$

where N is the sum of a, b, c, and d; a and b are the occurrences of a given motif and other motifs in a foreground set, respectively; c and d are the same in a background set. We then calculated the p-value and selected significantly over-represented motifs. 2) We identified over-represented motif combinations. We took each of the phase-specific gene sets as a foreground set. As the background set, we took the intersection of the non-cell cycle genes and genes with constant expression profiles. We then searched all possible order-1, order-2, and order-3 combinations of motifs found in the previous step. For each of the combinations, we calculated the statistic of Eq. 1, where a and b are the numbers of upstream sequences with and without a given combination in a foreground set, respectively; c and d are the same in a background set. Then we calculated the p-value and selected significantly over-represented motif combinations. 3) We checked the coherence of expression profiles over time. We

calculated the average standard deviation (ASD) score for genes having a motif combination in the upstream sequences, and selected solely motif combinations whose genes had coherent expression patterns. The ASD score is defined as:

$$\text{Score} = E_i\left(\sigma_g\left(X_{i,g}\right)\right),$$

where $X_{i,g}$ is the normalized expression level of gene g at time i, σ_g is the standard deviation over genes, and E_i is the average over time. 4) We assigned particular TF combinations to the respective motif combinations by matching over-represented single motifs with over-represented TFs. Over-represented TFs are TFs that bind to a significant number of the promoters with a motif combination. To define the significant number, we used the p-value of the hypergeometric model:

$$P(t) = \frac{{}_TC_t \times {}_{K-T}C_{k-t}}{{}_KC_k} \text{ and } p = 1 - \sum_{i=0}^{t-1} P(i),$$

where K is the number of all promoters used and T is the number of promoters that are bound by a TF among the K promoters, and k is the number of promoters with a motif combination from a phase-specific gene set and t is the number of promoters that are bound by the TF among the k promoters, and p is the p-value.

3 Results and Discussion

We have succeeded in finding both known and new combinations of TFs and motifs. For example, our result was consistent with a combination in which Fkh1/2 and Ndd1 bind to GTAAACA, and Mcm1 binds to TTCCTAA at G2/M phase. One of new examples is the combination of Swi5 (a cell cycle regulator) and Ste12/Dig1 (mating regulators) at M/G1 phase. From these results, we found that many combinations are classified into three types of combinatorial mode: 1) combinations to wait for signals and then activate transcription; 2) combinations of TFs as a main player in the previous and the next cell cycle phase; 3) combinations of TFs that connect two different biological processes. Our big picture of combinatorial regulations have shifted the classical view of the mono-regulation of a transcription factor, to the view that most genes may be regulated by multiple transcription factors.

4 Acknowledgments

We would like to thank Akira Suyama, Tatsuhiko Tsunoda, and Toshihisa Takagi for their helps. This work was supported by a grant from the Japan Society for the Promotion of Science.

References

[1] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, Ch.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.-B., Volkert, Th.L., Fraenkel, E., Gifford, D.K., and Young, R.A., Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*, 298:799–804, 2002.

[2] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell*, 9:3273–3297, 1998.

[3] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M., Systematic determination of genetic network architecture, *Nature Gen.*, 22:281–285, 1999.