

Inferring Combinatorial Regulation of Binding Sites and Transcription Factors on Gene Expression

Mamoru Kato^{1,2} Naoya Hata²
kato@cshl.edu hata@cshl.edu

Nila Banerjee² Michael Q. Zhang²
banerjee@cshl.edu mzhang@cshl.edu

- ¹ Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan
² Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, U.S.A.

Keywords: combinatorial regulation, gene expression, DNA microarray, chromatin immunoprecipitation

1 Introduction

The combinatorial regulation of transcription factors (TFs) and their binding sites is quite important for gene expression, in which a limited number of TFs can contribute to complicated differentiation and respond to a great number of changeable environments. However, only a few studies have directly tackled with this hard problem, to uncover the combinatorial regulation [2]. Here we propose a novel integrated system to elucidate combinatorial regulations of TFs and binding site motifs effective on gene expression. This system includes filtering data, finding interactions between TFs and motifs from usual DNA microarray data [4] and chromatin immunoprecipitation (IP) microarray data [3], finding significant combinations of TFs and their motifs using a new algorithm, and finding the combination effects on time series of gene expression. This system was applied to Yeast cell cycle, in which a variety of data are accumulated.

2 Method

Our system has four procedures. 1) Filtering both ORF pairs with the same promoter and ORFs with extremely homologous upstream regions by BLAST into one ORF, so as not to overestimate statistical significance. 2) Finding significant single motifs of given TFs. (a) Data setting. Foreground: genes of cell-cycle [4] and IP+ [3] (ex. G1 and Mbp1+ etc.), background: genes of non cell-cycle and IP-. (b) Searching 6- to 9-mer over-represented motifs by counting the number of motifs and using the test of 2*2 contingency table. (c) Merging over-represented motifs in order to simplify redundant information. For instance, when CGCGAAA is supposed to be a true motif, also CGCGAA and GCGAAA are usually found in the one less-mer searching. The algorithm executes exact alignment to select an associated pair if the outer gap is within 2 (CGCGAA- and -GCGAAA), and deletes an associated pair if the p -value of the extended motif (CGCGAAA) is less than both p -values of two motifs of the pair in the contingency table. After these sub-procedures, it can be said that a found motif should interact with a given TF from IP data.

3) Finding motif combinations. (a) Data setting. Foreground: G1, S, S/G2, G2/M, and M/G1 genes classified in [4], background: genes of non cell-cycle and with constant expression (from std-dev over time) and of not IP+. (b) Searching over-represented combinations of the single motifs in the previous result, by counting the number of genes with a combination and by using the test of 2*2 contingency table. We executed full search up to 3-combinations. (c) Merging over-represented combinations in order to simplify redundant information. For instance, when M1-M2-M3 is supposed to be a true motif combination, also M1-M2, M1-M3, and M2-M3 are usually found in the one less-combination searching. The algorithm executes the test of independence for a pair (ex. M1-M2 and

M2-M3) among a found combination set to select an associated pair, and deletes an associated pair if the p -value of the extended combination (M1-M2-M3) is less than both p -values of two combinations of the pair in the contingency table. 4) Finding combinations effective on gene expression. The system calculates the degree of expression coherence of ORFs with a combination based on the average Euclidean distance over all of the ORF pairs, and also outputs its p -value from computer simulation. According to the p -value, the system screens for combinations significantly contributing to time series of gene expression in [1] data.

3 Results

We reconstructed synergistic regulations (Fig. 1). The left result is experimentally validated, and the right result is a new statistically significant synergism of MBF (Mbp1 and Swi6), SBF (Swi4 and Swi6), and Mcm1.

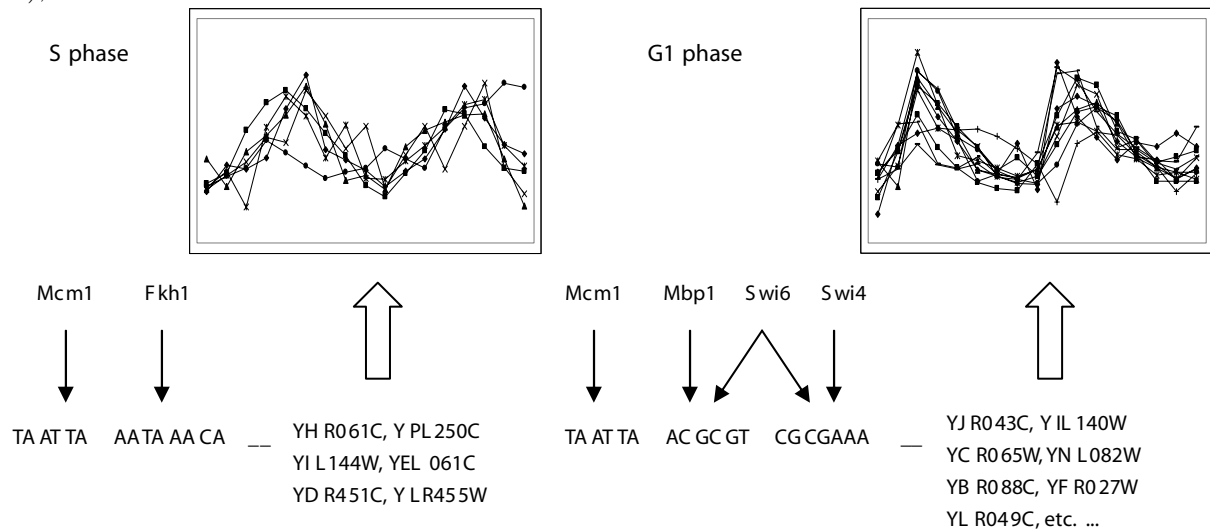


Figure 1: A part of reconstructed combinatorial regulations from usual microarray data, IP microarray data, and upstream sequence data. Prior knowledge on regulation is not used. Mcm1, Fkh1, Mbp1, Swi6, and Swi4 are TFs. An arrow from a TF to a motif shows a direct or indirect influence of the TF on the motif. A motif-bar-motif representation shows a significant motif combination. A motif combination exists on the upstream sequence of the following ORFs.

4 Acknowledgments

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

References

- [1] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., and Davis, R.W., A genome-wide transcription analysis of the mitotic cell cycle, *Molecular Cell*, 2:65–73, 1998.
- [2] Pipel, Y., Sudarsanam, P., and Church, G.M., Identifying regulatory networks by combinatorial analysis of promoter elements, *Nature Genetics*, 2:153–159, 2001.
- [3] Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S., and Young, R.A., Serial regulation of transcriptional regulators in the yeast cell cycle, *Cell*, 106:697–708, 2001.
- [4] 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, *Molecular Biology of the Cell*, 9:3273–3297, 1998.