Detecting Transcriptional *cis*-Regulation from Gene Expression Data

Yoshinobu Igarashi igarashi@kuicr.kyoto-u.ac.jp
Yasushi Okuno okuno@kuicr.kyoto-u.ac.jp
Jean-Philippe Vert vert@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Keywords: transcription, cis-regulation, gene expression, microarrays

1 Introduction

The regulation of gene expression in eukaryotes is a complex process still to be elucidated. It has been known for a long time that the coordinated actions of particular transcription factors play a fundamental role in this process, but discovering such cis-regulatory actions remains a challenging problem, both experimentally and computationally.

We have developed a systematic approach to discover combinations of regulatory motifs which potentially regulate certain genes under certain conditions, using gene expression data obtained from microarray experiments. Given a list of motifs and a table of microarray data, our method returns a ranked list of motif combinations which seem to play a particular role in the regulation of gene expression.

We applied this method to track cis-regulatory factors in the yeast *S. cerevisiae*, using a compilation of public microarray data in combination with a list of known or putative transcription factor binding site motifs.

2 Method and Results

2.1 Scoring Gene Clusters Using Microarray Data

We used a compilation of publicly available microarray data for the yeast resulting in 522 measures for 5050 genes. This table was mean-centered and scaled to unit variance first for each gene, then for each experiment.

For any given list of genes we consider its mean profile and variance profile. Due to the prior data normalization, the mean profile of the set of all genes is constantly equal to 0, while its variance profile is constantly equal to 1. As a result the profiles corresponding to a given subset of genes gives information concerning the possible co-regulation of these genes:

- if the mean profile is significantly positive (resp., negative) for some experiments, it means that these genes are on average over-expressed (resp., repressed) during these experiments;
- if the variance profile is significantly smaller than 1 for some experiments, it means that the genes are particularly co-expressed during these experiments.

From these observations we derived several *scores* (such as the average, the minimum or the maximum of these profiles) to quantify how co-regulated the genes are.

2.2 Tracking Motif Combinations

We used a list of 356 DNA motifs including 37 known transcription factor binding site motifs (used in [1]) and considered all possible combinations of several motifs. For each combination we computed the list of genes which contain all motifs in their 1000 bp upstream regions. We computed the scores corresponding to these gene lists, and ranked all combinations by decreasing scores.

For each combination we also compared the profiles corresponding to this combination with the profiles corresponding to the combinations where a single motif has been removed (for example, we compared the combination (A,B,C) with the three combinations (A,B), (A,C) and (B,C)). The profiles were compared by computing the difference of mean profiles and the ratio of variance profiles, and applying the previous scores. The combinations were also ranked by decreasing scores.

2.3 Results

For each score we obtained a ranked list of motif combinations. The analysis of these combinations reveals some known and some new combinations of motifs, and the analysis of the corresponding mean and variance profiles provides clues on why these combinations have high scores. As an example Table 1 contains motif combinations with particularly high scores[1]. Among these combinations some have

Table 1: Combinations of transcription factors with highest score

Score	Motif combination
2.296	ALPHA1 , MET31-32
2.267	GAL , putative G-proteins
2.069	PAC , putative RRSE10 , RRPE
1.980	SFF , ndt80
1.947	ALPHA1' , ALPHA1 , MET31-32
1.946	MCM1 , ALPHA1 , ALPHA2
1.901	GAL , putative lipid and fatty acid transport
1.843	SFF , SFF' , ndt80
1.838	MCM1 , MCM1' , ALPHA1 , ALPHA2
1.824	ndt80

been shown to potentially interact (e.g, the PAC, RRPE and RRSE combination has been highlighted in [1]) while others suggest new interactions between transcription factors (e.g., the genes containing the ALPHA1 and MET31-32 binding site motifs appear to be strongly over-expressed during an amino-acid starvation experiment).

3 Acknowledgements

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan and from the Japan Society for the Promotion of Science.

References

[1] Pilpel, Y., Sudarsanam, P., and Church, G.M., Identifying regulatory networks by combinatorial analysis of promoter elements, *Nature Genetics*, 29(2):153–159, 2001.

[1]The score considered in Table 1 is the maximum of the mean profile.