Extracting Regulatory Signals from the Upstream Region of Co-expressed Genes Derived from DNA Microarray Experiments

Shuichi Kawashima
shuichi@kuicr.kyoto-u.ac.jp

Hiroyuki Ogata
ogata@kuicr.kyoto-u.ac.jp

Minoru Kanehisa
kanehisa@kuicr.kyoto-u.ac.jp

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

1 Introduction

Since the various genome sequencing projects are rapidly producing new sequence data, it is now possible to analyze information of sequence data at the genomic scale. In addition, a large amount of data for gene expression profiles produced by the DNA microarray technology will be available in near future. Since the co-expression of genes strongly indicates co-regulation by complex genetic networks, the data of families of co-expressed genes derived from these profiles are useful for understanding the mechanisms of genetic networks in living organisms. Thus computational method for analyzing regulatory sequences of co-expressed genes is required.

Here, we present an analysis of upstream regions of the sets of genes grouped according to a large scale expression profile experiment.

2 Data and Method

van Helden *et al.* developed a method that isolates DNA binding sites for transcription factors from families of co-regulated genes [2]. The method is based on a regulatory paradigm in yeast that the upstream regions of the genes contain multiple copies of regulatory sequences. We employed the method to detect regulatory sites for the groups of genes determined by a microarray experiment on *Saccharomyces cerevisiae* [1]. The method counts the number of occurrences of all oligonucleotides of the selected size (six in the present analysis), and estimates their statistical significance. The probability of observing exactly n occurrences of oligonucleotide b is estimated according to the binomial formula:

$$\Pr(occ\{b\} = n \mid F\{b\}) = \frac{T!}{(T-n)! \times n!} \times (F\{b\})^n \times (1 - F\{b\})^{T-n}$$

where T is the total number of possible matching position, and $F\{b\}$ is the frequency of b observed in T. To estimate the significance of the observed occurrence n against background probability, we employed the significance index given by Tamames *et al* [3].

$$sig = \log \frac{\Pr(occ\{b\} = n \mid F_{ob}\{b\})}{\Pr(occ\{b\} = n \mid F_{bg}\{b\})}$$

where $F_{bg}\{b\}$ is the background frequency of b observed in all the upstream sequences (800 bases upstream of the initiation codon) in the genome, and $F_{ob}\{b\}$ is the frequency of b observed in the upstream sequences in the group of genes.

3 Result and Discussion

We analyzed three groups of genes reported in DNA microarray experiments [1]. In Table 1, oligonucleotides occurring in the upstream region of one group are shown with their significance indices. This group include seven genes, namely GYS2, CTT1, HSP42,HSP26, HSP12, YKL026 and CYGR043C, which are known to have stress response elements (STRE; AGGGG or CCCCT) in their non-coding upstream regions. Interestingly, oligonucleotide patterns that match to STRE sequences have highly significant indices as shown in the Table 1. Also for the second group, which consists of MLS1, IDP2, ICL1, ACS1, ACR1, FBP1 and PPC1, oligonucleotide patterns that partially match to the known regulatory element (CCRTYCRTCCG) are detected with highly significant indices. For the last group, which consists of YPL012W, YNL141W, YMR290C, SAM1, GPP1, YGR160W and YDR398W, oligonucleotide, TTTTTT, observed extremely high significance ($sig = 31.80$). Although the function of this pattern, as long as we know, is not characterized, the sequences may have relation to the regulation of this group.

Sequence	Occurrence	significance	Sequence	Occurrence	significance
tggctc	8	4.64	tagggg	4	2.25
cagggg	5	3.54	cgacgt	4	2.20
ataggg	6	3.22	ccctaa	5	2.20
cctgga	6	3.09	ggggga	4	2.18
ggctct	6	2.88	ctggct	5	2.14
ggggcg	4	2.86	aggggа	5	2.10
gggggg	4	2.64	aggggc	4	1.93
atgggg	5	2.54	ggcggc	4	1.90
taaaac	10	2.51	tgccag	5	1.87
aggggg	4	2.31	gcgacc	3	1.78

Table1. oligonucleotides with high significance.

In this study, we have shown the effectiveness of the method to characterize regulatory sequences of a set of genes defined by expression profile data. We plan to detect other functionally related genes that are co-regulated with the groups we analyzed. Correlation analysis of expression patterns could be used to enlarge the grouping of genes.

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas 'Genome Science' from The Ministry of Education, Science, Sports and Culture in Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] DeRisi, J.L., Iyer, V.R. and Brown, P.O., Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, 278:680–686, 1997.

[2] van Helden, J., André, B. and Collado-Vides, J., Extractiong regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies, *J. Mol. Biol.*, 281:827–842, 1998.

[3] Tamames, J., Casari, G., Ouzounis, C. and Valennncia, A., Conserved clusters of functionally related genes in two bacterial genomes, *J. Mol. Evol.*, 44:66–73, 1997.