63. A Computational Approach to Discover Differential Cooperation of Regulatory Sites in Functionally Related Genes in Yeast Genome

Hsien-Da Huang[1], Jorng-Tzong Horng[2], Jing-Yue Hong[3], and Baw-Jhiune Liu[3]

Keywords: Regulatory Elements, Gene Expression, Data Mining, Promoter

1 Introduction

The availability of genome-wide gene expression data provides a unique set of genes from which to decipher the mechanisms underlying the common transcriptional response. A set of transcription factors which bind to target sites regulate the gene transcription cooperatively. The functional-specific combinations are discovered by a statistical approach and the over-represented repetitive elements involving in the combinations are possible to be transcription factor binding sites. The combinations facilitate to predict functional-specific putative regulatory elements and to identify genes potentially co-regulated by the putative regulatory elements. Our proposed approach is applied to *Saccharomyces cerevisiae* and the promoter regions of Yeast ORFs. The results of the study are available at http://bioinfo.csie.ncu.edu.tw/REDB/.

This study initially identifies the combinations of known regulatory sites extracted from TRANSFAC [1] and over-represented repetitive oligonucleotides statistically retrieved from RSDB [2] in the promoter regions of a particular set of genes selected. The data mining approach, mining association rules, is then applied to mine the associations from the combinations of over-represented repetitive elements and known regulatory sites. The combinations found in each functional gene group are then statistically analyzed in all other groups. Chi-square test is applied to determine the dependence of the sites of the combinations, as well as the R-value of each combination is computed among groups of genes to find its differential occurrences in each group of functionally related genes. Those target sites in highly dependent combinations with large enough R-values, i.e., the small enough p-values, in a functional gene group are candidates of putative functional-specific regulatory sites in the group because of their specificities in that functional gene group.

2 Methods

We first preprocess the target sites and gene promoter regions to find the combinations of known sites and over-represented repetitive oligonucleotides located in the promoter regions of the groups of functionally related genes. Next, AprioriAll algorithm [3] is applied to mine the association rules by combining the known sites and over-represented repeats. Chi-square test is then used to select certain interesting and significant rules. The R-value of each site combination is computed to investigate the differences of transcriptional regulation in different functional categories of genes. Finally, the over-represented repeats within the significant and differential combinations, which are mapping to the items in the association rules, are selected as putative regulatory sites [4-5].

[1] Department of Computer Science and Information Engineering, National Central University, Taiwan.
E-mail: damay@db.csie.ncu.edu.tw
[2] Department of Computer Science and Information Engineering, National Central University, Taiwan.
E-mail: horng@db.csie.ncu.edu.tw
[3] Department of Computer Science and Engineering, Yuan-Ze University, Taiwan.

In order to investigate the occurrence differences of the site combinations mined from different groups, we propose a R-statistic to compute the hypothesis that each occurrence time is consistent with the others. Computing R-statistic can extract the combinations whose the occurrence most varied across different functional categories. The statistic is denoted as R_j for each combination j given by formula 1, where m is the number of functional categories, $X_{i,j}$ is the occurrence times of the combination j in the functional category i, and N_i is the total number of genes, i.e., ORFs, in the ith functional category. The frequency f_j of the combination j in all of the functional categories is given by formula 2.

$$R_j = \sum_{i=1}^{m} X_{i,j} \log\left(\frac{X_{i,j}}{N_i f_j}\right) \quad (1) \qquad f_j = \frac{\sum_{i=1}^{m} x_{i,j}}{\sum_{i=1}^{m} N_i} \quad (2)$$

3 Results

As shown in Table 1, the number of genes in each functional categories are shown in the third row, as well as the numeric of MIPS functional categories. (a: Deoxyribonucleotide metabolism, b: Amino-acid transport, and c: Homeostasis of protons) For example, in the first row the combination "GATAA,aacgc" occur in the gene upstreams of functional categories, 010307, 010107, and 130102, with Chi-square values 7.20, 4.48, and 0.04, respectively. The Chi-square value greater than 3.84 is shown with parentheses. Similarly, the support values are 0.75, 0.00 and 0.00. The support value greater than 0.4 is shown with parentheses.

		Functional Categories (Numeric in MIPS)						
		010307[a]	010107[b]	130102[c]	010307[a]	010107[b]	130102[c]	
Number of Genes		12	23	32	12	23	32	
		χ^2 values			**Support**			***R***
010307	GATAA,aacgc	(7.20)	(4.48)	0.04	(0.75)	0.00	0.00	3.19
	GAGGA,cgcgtc	(6.12)	none	none	(0.42)	0.00	0.00	8.60
	GAGGA,acgcgtc	(6.12)	none	none	(0.42)	0.00	0.00	8.60
	ACGCGT,GAGGA	(5.18)	none	none	(0.50)	0.00	0.00	10.32
010107	cgtgc,gcgcc	0.01	(4.71)	0.30	0.17	(0.48)	0.00	6.76
	cgtgc,gccgc	0.01	(7.08)	0.18	0.17	(0.48)	0.00	4.12
	cgccg,cgtgc	0.69	(5.32)	0.03	0.08	(0.43)	0.00	4.96
	TTATC,cgccg	(8.00)	(5.28)	0.73	0.08	(0.52)	0.00	2.79
	ATATAA,cggcaa	1.50	(4.54)	0.73	0.08	(0.43)	0.00	2.82
130102	TTATC,atataat	3.27	0.61	(5.74)	0.00	0.09	(0.44)	6.46
	ATATAA,attataa	1.09	0.40	(6.73)	0.00	0.04	(0.41)	7.07

Table 1: The differential combinations of regulatory sites in five different functional categories (R > 2.0 and support > 0.4).

References

[1] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhäuser, M. Prüß, F. Schacherer, S. Thiele, and S. Urbach. "The TRANSFAC system on gene expression regulation." *Nucleic Acids Res.* 2001, 29, pp. 281-283.

[2] J.T. Horng, J.H. Lin, and C.Y. Kao, "RSDB-A Database of Repetitive Elements in Complete Genomes", *Proceedings of the Atlantic Symposium on Computational Biology and Genome Information Systems & Technology*, Durham, NC, USA, 2001, pp. 220-223.

[3] R. Agrawal, T. Imielinski, and A. Swami, "Mining Associations between Sets of Items in Large Databases", *Proc. of the ACM SIGMOD Int'l Conference on Management of Data*, Washington D.C., 1993, pp. 207-216.

[4] J.T. Horng and H.D. Huang. "Mining Putative Regulatory Elements in Gene Promoter Regions." 2002, *In Silico Biology*, 2, 0025.

[5] J.T. Horng, and H.D. Huang, C.C. Huang, and Y.K. Cheng. "Mining Putative Regulatory Elements in Gene Promoter Regions." Proceedings of the German Conference on Bioinformatics. 2001, pp. 90-95.