The Estimations of Motif Effects with Longitudinal Mixed Model in Temporal Gene Expression Analysis

Song Jiuzhou, Jaime Bjarnason and Michael G. Surette

Keywords: Regulatory analysis, Motif discovery, Longitudinal mixed model, Gene expression analysis

1 Introduction

The identification and testing of relevant transcription factor (TF) binding sites is one of the most important and greatest challenges in a functional genomics era. Traditionally, TF binding sites have been characterized by different experimental methods, a very slow and inefficient process. Although the similarity comparisons and multiple alignments of upstream sequences can find many significant repeats or conserved sequences upstream of the coding region, the statistically significant meaning of the putative motifs is based only on the frequencies of the nucleotides or patterns against the genome species[1-5]. It doesn't indicate the probability that the putative motifs are TF binding sites or that they have biological relevance for gene expression. Real TF binding sites must be confirmed by wet-bench genetic analysis. How to screen the motif candidates is becoming a critical issue. The motif effect indicates the regulatory extent of the motif for a given gene expression, so the estimation of the individual and combinational motif effects on gene expression will provide alternative support for the motif candidates, and improve the quality and efficiency of the screening process. We propose a longitudinal mixed model to estimate motif effects in temporal gene expression analysis,

2 Material and Methods

In a temporal gene expression experiment of iron responsive genes in *S. typhimurium*,[6] we clustered genes on the basis of their expression profile across four conditions and time points via cluster analysis. We adopted the Mismatch Tree Algorithm (MITRA) approach to obtain composite regulatory patterns [7]. We then assume that Y_{ij} satisfies $Y_{ij} = b_{1j} + b_{2i}t_{ij} + b_{3i}t_{ij}^2 + e_{ij}$, where n_i is the number of longitudinal measurements available for the ith gene, and where all error components e_{ij} are assumed to be independently normally distribution with mean zero and variance s^2, Y_i equals $(Y_{i1}, Y_{i2}, ..., Y_{in_i})'$, e_i equals $(e_{i1}, e_{i2}, ..., e_{in_i})'$, b_i equals to $(b_{1i}, b_{2i}, ..., b_{3i})'$, and Z_i is the $(n_i \times 3)$ matrix, the columns of which contain only ones, all time points t_{ij} and all squared time points t_{ij}^2. Ones have believed that $\hat{b}_{i,OLS}$ are good approximations to the real subject-specific regression parameters b_i, that is $b_i = B_i + b_i$. Then a linear mixed-effects model can be defined as any model which satisfies mixed model equations $Y_i = X_i b + Z_i b_i + e_i$.

3 Result and Analysis

We found three motif candidates in the analysis, A1: GATAATAATTATT, A2: TAATGATAATCATT and A3: ATAATTATTATCA, B: GCGT_ACGC and motif C: GCCGGA. In the mixed model analysis, the Table 1 shows the significant tests of the fixed effects, the motif candidates and the interactions among motifs, and time and quadratic time are very significant. Then the maximum likelihood (ML) and restricted maximum likelihood (REML) are used to estimating for all parameters in the longitudinal mixed model as shown in Table 2. The motif candidates A1, A2 and A3 are similar to the Fur binding site, GATAATGATAATCATTATC [8]. The

Department of Biochemistry and Molecular Biology, Health Sciences Centre. University of Calgary, Calgary, AB, Canada, T2N 4N1
Email: songj@ucalgary.ca, bjarnaso@ucalgary.ca, surette@ucalgary.ca,

estimates for the parameters shows that significant effects seem to be present among the motif candidates A and B, although they have opposite effects. The motif candidate C has the weakest effects (0.0438). There are significant positive interactions between motif candidate B and time effects, and weaker interactions between motifs A and C and time effects. The table also indicates that the interaction of all motif candidates and quadratic time effects are negative and weak. Those results suggest that motif effects are strongly influenced by gene expression level over time. The results indicate the evaluation of motif candidates is possible via longitudinal analysis.

Table 1 Type 3 tests of fixed effects

Effects	DF	Den DF	F value	Pr>F
Motif	3	262	9.99	<0.0001
Time*Motif	3	262	23.67	<0.0001
$Time^2$ *Motif	3	262	18.00	<0.0001

Table 2. The estimations of main effects and interaction of the motif candidates

Effects	ML(s.e.)	REML(s.e.)
Motif A	0.3278(0.0854)	0.3278(0.0869)
Motif B	-0.3569(0.0887)	-0.3569(0.0901)
Motif C	0.0438(0.1482)	0.0438(0.1507)
Time * Motif A	0.0636(0.0436)	0.0636(0.0443)
Time * Motif B	0.2965(0.0452)	0.2965(0.0460)
Time * Motif C	0.6006(0.1129)	0.6006(0.1148)
$Time^2$ * Motif A	-0.0036(0.0047)	-0.0036(0.0048)
$Time^2$ * Motif B	-0.0166(0.0049)	-0.0166(0.0050)
$Time^2$ * Motif C	-0.1222(0.0185)	-0.1222(0.0188)

4 Reference

[1] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat Genet*, vol. 22, pp. 281-5, 1999.

[2] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church, "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae," *J Mol Biol*, vol. 296, pp. 1205-14, 2000.

[3] J. D. H. Frederick P. Roth, Preston W. Estep, and George M. Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation," *Nature Biotechnology*, vol. 16, pp. 939-945, 1998.

[4] V. R. Hao Li, Carol Gross, and Eric D. Siggia, "Identification of the binding sites of regulatory proteins in bacterial genomes," *PNAS*, vol. vol.99, pp. 11772-11777, 2002.

[5] G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics*, vol. 15, pp. 563-77, 1999.

[6] J. Bjarnason, C. M. Southward, and M. G. Surette, "Genomic profiling of iron-responsive genes in Salmonella enterica serovar typhimurium by high-throughput screening of a random promoter library," *J Bacteriol*, vol. 185, pp. 4973-82, 2003.

[7] E. Eskin and P. A. Pevzner, "Finding composite regulatory patterns in DNA sequences," *Bioinformatics*, vol. 18 Suppl 1, pp. S354-63, 2002.

[8] C. F.Earhart, "Uptake and Metabolism of Iron and Molybdenum," *Frederick C.Neidhardt, Editor in Chief. Escherichia coli and Salmonella Cellular and Molecular Biology*, vol. 1, pp. 1075-1090, 1996.