Detection of Tissue Specific Genes by Putative Regulatory Motifs in Human Promoter Sequences

Katsuhiko Murakami katsu@gsc.riken.jp
Toshio Kojima tkojima@gsc.riken.jp
Yoshiyuki Sakaki sakaki@gsc.riken.jp
RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa, Japan

Keywords: promoter, tissue-specific gene expression, position weight matrix, regulatory motif, CpG islands

1 Introduction

Gene expression of multi-cellular organisms is regulated by transcription factors (TFs) that interact with regulatory *cis*-elements on DNA sequences. To find the functional regulatory elements, computer searching can predict TF binding sites (TFBS) using position weight matrices (PWMs) that represent positional base frequencies of collected experimentally determined TFBS. However, it is still difficult to tell authentic sites from false positives. Reports have shown that particular TFBS are concentrated in promoters, though a general tendency is uncertain. Computational approaches to reveal structure of promoter as combination of TFBS are required. Here we have examined the correlation between predicted TFBS and promoters, and identified two PWM groups, 1) PWMs whose TFBS are clustered in promoters mainly by the existence of CpG islands (CGI), 2) PWMs whose TFBS are clustered in promoter independent of CGI. As an application of the groups, we show that tissue specific genes can be extracted by finding clusters of predicted TFBS of selected PWMs in promoters.

2 Method and Results

2.1 Scoring PWMs by Partial Correlation Coefficients

DNA sequences of promoters and introns in chromosomes 20, 21 and 22 were collected from the genes in RefSeq by the annotation of UCSC genome DB [2]. The annotation of 5' end of RefSeq were modified if 5' UTR is extended by the entries in DBTSS [5]. Promoter sequences were taken from genome sequence according to the modified annotation. Intron sequences, without the first introns, were used as non-promoter sequence. Overlapping genes were discarded by the annotation of UCSC genome DB. Repeat sequences were partially removed so that the ratio of repeat regions to the total sequence in promoter is almost the same as that in non-promoter sequences. We prepared 768 promoter and 15,444 non-promoter subsequences, each of which is 600 bp in length.

Every subsequence was analyzed as follows. TFBS were predicted by MATCH program [1] with TRANSFAC matrix 6.3 with the options of 'high quality matrix' and 'minimize false negatives'. We defined accumulated score (*AS*) of predicted TFBS as an index of the density of a cluster of TFBS for a PWM. A CGI score of *start_p* was calculated by CpGProD program [4]. The score indicates the probability that the region is a "start CpG island" (a CpG island which is located over a transcription start site). Then partial correlation coefficients between the two variables among promoter, CGI, and AS, were computed controlling the other variable. We use P to denote promoter, namely $P = \{0, 1\}$,

and I to denote CGI score, and C to denote accumulated score. A partial correlation coefficient $\boldsymbol{r}_{PI.C}$ is the correlation between P and I while controlling for C. Fig. 1 shows a plot of $\boldsymbol{r}_{IC.P}$ against $\boldsymbol{r}_{PC.I}$ for various PWMs. A partial correlation coefficient is useful when we want to know pure correlation between two variables and there might be other causal factors. Suppose that correlation between P and C depends entirely on the common cause I. Either I is constant or I varies, the partial correlation coefficient between P and C should be zero (or small), because there is no reason P and C are correlated, while correlation coefficient is not expected to be zero. Using two partial correlation coefficients, we identified two PWM groups, 1) CGI-related PWMs, whose TFBS are clustered in promoters mainly by the CGI, and 2) CGI-independent PWMs, whose TFBS are clustered in promoters independent of CGI. Larsen *et al.* found that all housekeeping genes have CGI covering transcription start site [3]. Since the second PWM group are independent of CGI, it is suggested that transcription factors corresponded to the second PWM group are associated with tissue-specific regulation.

2.2 Detection of Tissue Specific Genes

Given a promoter sequence, we searched clustered predicted TFBS by each PWM with significant Z-score above 3, compared with random sequences of the first order Markov model. Let n be the number of PWMs whose TFBS clusters were considered as significant for the promoter sequence. We define differential correlation coefficient score DCC by

$$DCC = \sum_{i=1}^{n} (\boldsymbol{r}^i_{PC.I} - \boldsymbol{r}^i_{IC.P})$$

which indicate how much the TFBS clusters are correlated with promoters independent of CGI. Here $\boldsymbol{r}^i_{PC.I}$ and $\boldsymbol{r}^i_{IC.P}$ represents the partial correlation coefficients of i th PWM. From the genes whose expression in tissues are described in the human gene expression index (HuGE Index) database, we collected genes on the three chromosomes: (72 genes including 51 tissue specific (TS) genes). TFBS clusters were found in 40 genes of them. If we take genes with higher DCC score with a cutoff value, we could extract 20 genes; 18 of which (90%) were TS genes. The event was significant (p-value < 0.04) under cumulative hypergeometric distribution. It is also interesting that DCC is not related with CGI scores. These results suggest that clusters of TFBS associated with the CGI-independent PWMs, which we identified, really play important roles in expression of TS genes.

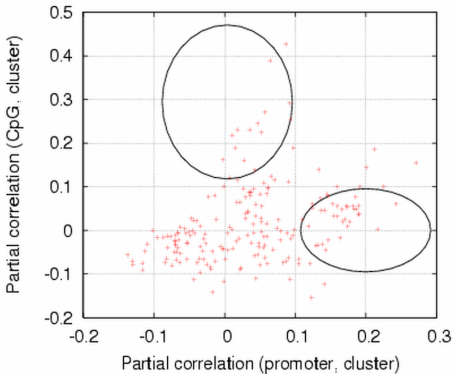

Figure 1: A plot of $\boldsymbol{r}_{IC.P}$ against $\boldsymbol{r}_{PC.I}$ for various PWMs. For PWMs around the top circle, clusters of predicted TFBS are due to mainly CpG islands. For PWMs around the right circle, the cluster of predicted TFBS is little correlated with CGI and more correlated with promoter. The two circles were drawn manually to show the idea.

References

[1] Kel, A.E. *et al.*, MATCH: a tool for searching transcription factor binding sites in DNA sequences, *Nucleic Acids Res.*, 31(13):3576–3579, 2003.

[2] Kent, W.J. *et al.*, The human genome browser at UCSC, *Genome Res.*, 12(6):996–1006, 2002.

[3] Larsen, F. *et al.*, CpG islands as gene markers in the human genome, *Genomics*, 13(4):1095–1107, 1992.

[4] Ponger, L. and D. Mouchiroud, CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences, *Bioinformatics*, 18(4):631–633, 2002.

[5] Suzuki, Y. *et al.*, DBTSS: database of human transcriptional start sites and full-length cDNAs, *Nucleic Acids Res.*, 30(1):328–331, 2002.