Identification of Regulatory Controls for Sets of Co-expressed Genes

Shannan J. Ho Sui[1], James Mortimer[2], Brian P. Kennedy[2], Chris J. Walsh[1], Wyeth W. Wasserman[1]

Keywords: promoter, transcription factor binding sites, comparative sequence analysis, gene expression, microarray, regulatory network

1 Introduction.

The use of large-scale gene expression profiling experiments to decipher underlying transcriptional networks is an intriguing and challenging area of research in bioinformatics. Creative computational algorithms are required to elucidate the transcription factors (TFs) that give rise to observed co-expression patterns by searching for shared *cis*-regulatory motifs in the regulatory regions of co-expressed genes. We have developed an integrated approach that combines cross-species comparisons with promoter motif identification tools for the automated detection of significantly over-represented transcription factor binding sites (TFBS) in sets of coordinately expressed genes.

The use of position-specific scoring matrices (PSSMs) to detect known TFBS is well-established (reviewed in [1]). However, these methods typically yield a large number of false positive predictions due to the short, variable nature of TFBS. Dramatic improvements in the specificity of TFBS prediction are attained by limiting the search space to regions of conserved, non-coding DNA using a comparative genomics approach known as phylogenetic footprinting [2].

2 Methods.

The promoter regions of human genes (defined as 5kb upstream and 1kb downstream of the annotated transcription start sites), were aligned to the corresponding promoter regions of their mouse orthologs (as defined by EnsEMBL). Regions of the alignments with greater than 75% sequence conservation were searched for matches to 75 vertebrate-specific TF binding profiles present in the JASPAR database [3]. Computational methods utilized the TFBS suite of regulatory analysis Perl modules [4].

Two statistical measures were calculated to determine which, if any, TFBS were over-represented in the set of promoters for co-expressed genes. These represent two distinct models for counting the occurrences of binding sites.

The z-score uses a simple binomial distribution model to compare the *frequency of occurrence of a TFBS* in the set of co-expressed genes to the expected frequency estimated from a background set containing all genes on the microarray chip. For a given TFBS, let the random variable X denote

[1] Centre for Molecular Medicine and Therapeutics, Vancouver, BC, Canada. E-mail: (shosui, cjwalsh, wyeth)@cmmt.ubc.ca

[2] Department of Biochemistry and Molecular Biology, Merck Frosst Centre for Therapeutic Research, Point-Claire, Dorval, Quebec H9R 4P, Canada. E-mail: (james_mortimer, brian_kennedy)@merck.com

the number of predicted binding site nucleotides in the conserved non-coding regions of the co-expressed genes. Let p be the rate of occurrence of predicted binding site nucleotides in the background sequences. Using a binomial model with n events, where n is the total number of nucleotides examined from the co-expressed genes, and p is the probability of success, the expected value of X is $\mu = np$, with standard deviation $\sigma = \sqrt{np(1-p)}$. Let x be the observed number of binding site nucleotides in the conserved non-coding regions of the co-expressed genes. By applying the Central Limit Theorem and using the normal approximation to the binomial distribution with a continuity correction, the z-score is calculated as $z = \frac{x - \mu - 0.5}{\sigma}$. Then, the probability of observing x or more binding site nucleotides in the conserved non-coding regions of the co-expressed genes is given by $\Pr(X \geq x) \cong \Pr(Z \geq z)$.

In contrast, the one-tailed Fisher exact probability compares the *proportion of co-expressed genes* containing a particular TFBS to the proportion of the background set that contains the site to determine the probability of a non-random association between the co-expressed gene set and the TFBS of interest. It is calculated using the hypergeometric probability distribution that describes sampling without replacement from a finite population consisting of two types of elements [5]. Therefore, the number of times a TFBS occurs in the promoter of an individual gene is disregarded, and instead, the TFBS is considered as either present or absent.

3 Results.

The method was validated on a number of reference sets, and then applied to genes significantly down-regulated in cells treated with a compound known to inhibit the NF-κB signaling pathway. TFBS that were significantly over-represented in the down-regulated set relative to the background set are shown in Table 1.

	TFBS	TF Class	z-score p-value	Fisher p-value		TFBS	TF Class	z-score p-value	Fisher p-value
1	NF-κB	Rel/ NF-κB	0.0e+00	3.2e-09	7	SPI-B	ETS	1.7e-17	2.2e-03
2	p65	Rel/ NF-κB	0.0e+00	4.0e-08	8	HFH-2	Forkhead	8.5e-17	2.0e-03
3	c-Rel	Rel/ NF-κB	0.0e+00	1.5e-04	9	FREAC-4	Forkhead	3.8e-16	5.9e-04
4	p50	Rel/ NF-κB	0.0e+00	5.5e-04	10	Max	bHLH-ZIP	2.4e-07	8.6e-03
5	Pbx	Homeo	3.3e-32	2.1e-03	11	SRY	HMG	1.8e-04	8.8e-03
6	Sox-5	HMG	1.5e-18	4.1e-03					

Table 1: Significant TFBS detected in genes down-regulated by treatment with the inhibitor (p-values less than 0.01 for both the z-score and Fisher measures).

4 References and bibliography.

[1] Wasserman, W.W. and Krivan, W. 2003. In silico identification of metazoan transcriptional regulatory regions. *Naturwissenschaften* 90:156-66.

[2] Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N. and Wasserman, W.W. 2003. Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* 2:13.

[3] Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. and Lenhard, B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32:D91-4.

[4] Lenhard, B. and Wasserman, W.W. 2002. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* 18:1135-6.

[5] Fleiss, J.L. 1981. *Statistical methods for Rates and Proportions*. New York: John Wiley.