Transcription factor binding: does order matter?

Gabriel Kreiman[1], Nambi Nallasamy [2]

Keywords: gene expression regulation, transcription factor binding sites, *cis* elements

1 Introduction

Gene expression in eukaryotes is orchestrated by the binding of transcription factors (TFs) to *cis* regulatory elements in the DNA [1]. Regulatory regions in eukaryotes show a hierarchical structure of separate modules with multiple TF binding sites within each module. In many cases, the position, order and orientation of separate modules can be changed considerably without affecting transcriptional rate. However, the detailed internal structure within modules is only poorly understood. Further understanding the structure of transcriptional control regions could help improve current algorithms to detect regulatory elements [2]. Here we asked whether the order of TF binding sites within a module matters or not. We studied the relative orientation and order for all possible pairs from a large collection of yeast and mammalian TF binding sites. Throughout non-coding regions upstream of genes, we observed a large number of examples where one particular ordering was much more prevalent than the reverse order. Our observations suggest that the internal structure of regulatory modules, in particular the spatial order of TF binding to DNA, may play an important role in the specificity of gene expression control.

2 Results

A brief schematic description of the methodology is shown in Figure 1. In order to computationally assess whether the order of the binding of two TFs matters or not, we explored the frequencies in which different possible ordering arrangements occur throughout non-coding regions of the genome upstream of genes. For a pair of TF binding sites *i* and *j*, n_{ij} indicates the number of genes where the binding site for *i* was 5' of *j* and n_{ji} indicates the number of genes where the binding site for *j* was 5' of *i*. Under the null hypothesis, the two orders should be equally prevalent and the value of n_{ij} should be close to n_{ji}. We defined the ordering *p* value as the probability of observing a given difference between n_{ij} and n_{ji} by chance using a binomial distribution and computing the cumulative probability given by:

$$p_{order}(i,j) = \sum_{x=n_{ij}+1}^{n} \binom{n}{x} \alpha^x (1-\alpha)^{n-x}$$

where $n=n_{ij}+n_{ji}$ and $\alpha=0.5$ under the null hypothesis. Given that there were multiple PWM pairs that passed the interaction criteria, we used a Bonferroni correction using the total number of TF pairs as the total number of hypotheses to evaluate. We observed many examples in yeasts, humans and mouse where one order of PWMs was much more prevalent than the reverse order. A summary of the results for the mouse case is shown in Table 1.

3 Figures and tables

[1] Center for Computational and Biological Learning, McGovern Institute for Brain Research, Massachusetts Institute of Technology, Boston, North America, E-mail: kreiman@mit.edu

[2] Engineering and Applied Sciences, Harvard University E-mail: nallasam@fas.harvard.edu

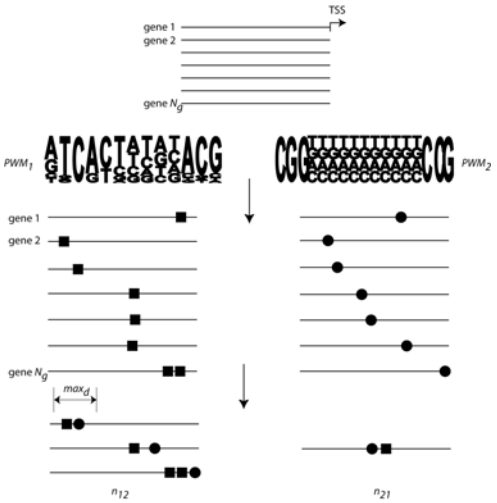

Figure 1: Schematic illustration of the methodology

Comparison of the frequency of different possible arrangements of two TFs. The TF binding models are given by the weight matrices (PWM_1 and PWM_2). The PWMs are used to scan the sequences upstream of the transcription start site of a set of N_g genes, determining the putative binding sites for PWM_1 and PWM_2 (squares and circles respectively). Subsequently, we consider those cases where PWM1 occurs within a given distance max_d 5' of PWM2 (n_{12}, left) and those cases where PWM2 occurs within max_d bp 5' of PWM1 (n_{21}, right). Finally, we computed whether the difference between n_{12} and n_{21} could arise by chance.

PWM_1 name	PWM_2 name	n_{12}	n_{21}	log(p)	r	PWM_1 name	PWM_2 name	n_{12}	n_{21}	log(p)	r
P300	ER	796	19	-208	41.89	ARP1	MEF2	323	16	-76	20.19
LYF1	RORA1	708	28	-172	25.29	AREB6	NRSF	321	19	-73	16.89
PAX4	NMYC	636	15	-167	42.4	GKLF	TAXCR	370	39	-69	9.49
E2F	CHOP	795	62	-164	12.82	1-Oct	PAX4	296	16	-68	18.5
PAX4	USF	614	15	-161	40.93	CDPCR	PAX4	288	16	-66	18
ARP1	PAX4	694	41	-155	16.93	3HD XBP1	MEF2	446	77	-65	5.79
ER	COUP	569	22	-139	25.86	AML1	1-Oct	499	103	-64	4.84
GATA1	PAX5	618	43	-132	14.37	FOXJ2	EVI1	378	52	-63	7.27
AHR	AHRAR	645	63	-123	10.24	TAXCR	NRSF	410	69	-61	5.94
CHOP	LYF1	522	26	-122	20.08	GR	BRACH	257	14	-60	18.36
AP1FJ	FOXJ2	520	31	-116	16.77	HAND1	ZID	306	35	-56	8.74
RORA1	TCF11	468	19	-114	24.63	AHR	AP1FJ	262	25	-52	10.48

Table 1: List of pairs of PWMs with non-uniform distribution of order arrangement

Pairs of *PWMs* (TRANSFAC release 6.0, *Hs* and *Mm*) where one order was significantly more frequent than the reverse order ($p<10^{-6}$). n_{12} = number of genes where the binding sites of the first *PWM* were 5' of those for the second *PWM*, *log(p)* = log of the order *p* value, *r* = ratio of n_{12} and n_{21}. A subset of *PWM* pairs is shown here. Shaded entries yielded $p < 10^{-6}$ in mouse.

4 References

[1] Davidson, E. et al. (2002). A genomic regulatory network for development. *Science* **295**, 1669-1678

[2] Kreiman, G. (2004). Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucl. Acids Res.* **32**, 2889-2900