

## In Silico Prediction of Transcription Factors that Interact with the E2F Family of Transcription Factors

### Sheng Zhong

Department of Biostatistics  
Harvard University  
655 Huntington Ave,  
Boston, MA, 02115, USA  
Email: szhong@hsph.harvard.edu

### Qing Zhou

Department of Statistics  
Harvard University  
Cambridge, MA, 02138, USA  
Email: zhou@stat.harvard.edu

### Paloma Giangrande

Department of Molecular  
Genetics and Microbiology  
Duke University  
Durham, NC, 27710, USA  
Email: phg@duke.edu

### Joseph R. Nevins

Department of Molecular  
Genetics and Microbiology  
Duke University  
Durham, NC, 27710, USA  
Email: j.nevins@duke.edu

### Wing H. Wong

Department of Statistics and  
Department of Biostatistics  
Harvard University  
Cambridge, MA, 02138, USA  
Email: wwong@hsph.harvard.edu

### Abstract

We describe a computational approach to predict transcription factors that interact with a given transcription factor, or a given family of transcription factors. We first collect a set of upstream sequences, to which a particular transcription factor or a family of transcription factors may bind. This set of upstream sequences is regarded as our training set. We collect a set of a large number of randomly chosen upstream sequences as the control set. We define a random variable to represent the clustering information of any putative transcription factor binding sites (TFBSs) in the control set. We calibrate the observed clusters of TFBSs in the training set to the distribution of the random variable representing the clustering information in the control set. We select the significant clusters from the training set and report the putative transcription factors that can bind to the TFBSs in these clusters. These reported transcription factors are candidates of interactive partners of the transcription factor (family) we

started from. We applied this approach to discover transcription factors that may cooperate with E2F family proteins. We have identified 15 candidate interactive partners of E2F. Among them, 5 have been suggested or verified by previous biological studies.

### 1 Introduction

At sequences level, eukaryotic gene expressions are usually controlled by regulatory modules rather than isolated single transcription factor binding sites (TFBS) [1]. Berman et al [2] introduced an approach to utilize TFBS clustering information to discover cis-regulatory modules in *Drosophila* genome. However, according to our knowledge, there are few successful mammalian module discovery exercises being described. This is mainly due to the large number of mammalian transcription factors and the high false positive rates of TFBS prediction algorithms. In this paper we show that by starting from a promoter set of pre-collected responsive genes of a given TF, it is possible to computationally

predict its interacting partners with high sensitivity and specificity.

The E2F family of transcription factors are essential for the timely activation of genes involved in DNA replication and cell cycle control, exerting both positive and negative effects on gene expression [3]. However, it has been shown that the presence of an E2F site is not sufficient for G1/S phase transcriptional regulation. Some transcription factors can interact with E2F and activate promoters in a synergistic manner. Such interactions stabilize DNA binding and contribute to promoter-specificity [4,5]. In this paper we use E2F as an example to demonstrate a computational approach in identifying transcription factors that can interact with a given transcription factor to either activate or repress gene expression. Our algorithm reported 15 such transcription factors. In the predicted transcription factors, there is strong evidence for TFE3, YY1, “ccaat” binding protein, and RAR to interactively work with E2F on promoter sequences [4,6,7,8]. Besides, P53 has been shown to be able to interact with E2F physically [9].

## 2 Method

### 2.1 Preparing training set and control set

We collected 155 human and 41 mouse candidate E2F responsive genes from microarray experiments (microarray data not shown here). We identified their orthologous genes in the other species by reciprocal BLAST [10], and pooled these orthologous genes together with the original E2F responsive genes in our analyses. The upstream regions of these genes were extracted from Human Genome assembly in UCSC [11] and Celera’s mouse genome database [12]. We used 1000 nucleotide (nt) upstream from transcription start site (TSS) and 50 nt downstream to TSS as the upstream region. We were able to get 332 such upstream regions. They were used as the training set.

2000 randomly selected upstream regions,

covering 1000 nt upstream and 50 nt downstream to TSS, were used as the control set. We used Repeat Masker [13] to mask out simple repeats and ALUs in both the training set and the control set.

### 2.2 Position specific score matrix (PSSM) calls

We downloaded all (342) vertebrate position specific score matrices (PSSM) from Transfac [14]. We applied the Match [14] program to scan every upstream sequence for PSSM calls. Every PSSM call was a potential TFBS. The cut off was set to minimize the sum of false positive calls and false negative calls. Match gave the position of every PSSM call.

The noise to signal ratio for the Match program’s PSSM calls is usually very high. Huang et al [15] reported the overall noise to signal ratio to be 8.9, i.e., on average, for every PSSM call being a real binding site, 9 false calls are reported at the same time.

### 2.3 Clusters of PSSM calls

To reduce false positive reports, we looked into clusters of PSSM calls. For a given PSSM, a cluster was defined by a significant grouping of its calls on any training upstream sequence. The significance of the grouping was addressed by comparing to all possible groupings of calls of the same PSSM on all the control sequences.

We introduced a sliding window scheme to find the groupings of PSSM calls on all the control sequences. The window size was set to be 300 nt. The window was shifted by 100 nt in each step of the scanning process. The initial window covered the first 300 nt on the first control upstream sequence. For a given PSSM, if it had  $n_1$  calls,  $n_1 = 0, 1, 2, \dots$ , within the initial window, we would say a grouping of  $n_1$  calls was found. At the next step, the window was shifted to cover from 100 nt to 400 nt of the first control sequence. In the second window we would find a grouping of  $n_2$  calls ( $n_2 = 0, 1, 2, \dots$ ). After finishing scanning the first control

sequence, the sliding window continued to cover the first 300 nt of the second control sequence. After the window slid through all control sequences, we added up the number of windows that covered the same number of PSSM calls. Given the sequence set, the number of windows that covered  $n$  PSSM calls was a random variable with respect to  $n$ . We named this random variable as the grouping random variable. By scanning the control sequences, we obtained a background distribution for the grouping random variable for one PSSM. Such a background distribution was calculated for every PSSM.

For every PSSM, setting the  $p$  value to be 0.0005, we calculated a critical value (CV) for its grouping random variable. The CV was so defined that less than 0.0005 of the sliding windows contained equal to or more than CV calls of this PSSM, in all the control sequences. Table 1 gives the background distributions and the CVs of the grouping random variables for all E2F PSSMs.

PSSM	0	1	2	3	4	5	6	7	8	9	10	11	12	CV
VSE2F1_Q3	10266	3120	1179	631	369	210	119	56	36	5	8	1	0	11
VSE2F1_Q4	15674	323	3	0	0	0	0	0	0	0	0	0	0	2
VSE2F1_Q6	14192	1405	331	58	12	1	1	0	0	0	0	0	0	5
VSE2F_Q1	15992	8	0	0	0	0	0	0	0	0	0	0	0	2
VSE2F_Q2	15926	72	1	0	1	0	0	0	0	0	0	0	0	2
VSE2F_Q3	15809	173	17	0	0	1	0	0	0	0	0	0	0	3
VSE2F_Q6	14833	1043	108	12	3	1	0	0	0	0	0	0	0	4
VSE2F_Q3	15922	71	6	1	0	0	0	0	0	0	0	0	0	2
VSE2F_Q4	15456	506	31	6	0	1	0	0	0	0	0	0	0	3

Table 1. The background distributions of the grouping random variables of E2F PSSMs. The first column contains the names of the PSSMs; the 2<sup>nd</sup> column contains the numbers of windows that covers 0 PSSM call; the 3<sup>rd</sup> column contains the numbers of windows that contain 1 call; and so on. The last column contains critical values when  $p$  value is set as 0.0005.

The CV attached to every PSSM can be regarded as a combined measure of two independent components. One is the biological meaning. The other is the specificity of a PSSM. When a PSSM has a very high CV, say, larger than 6, then this

PSSM is likely to be non-specific, because a grouping of so many binding sites can hardly have satisfactory biological interpretation. For this reason we discarded 62 PSSMs with CV bigger than 6. The following analysis was performed with the remaining 280 PSSMs.

We used the same sliding window system to scan the training sequences. Fixing a PSSM, if any sliding window contained more than or equal to CV calls, we would identify the calls within this window as a cluster of TFBSs. We identified all clusters for all PSSMs. Grouping together the PSSMs for the same transcription factor, we identified 40 transcription factors with 87 binding site clusters on 63 upstream regions in the training set. These data are available at <http://biosun1.harvard.edu/~szhong/E2F/macluster.xls>

#### 2.4 Rating clusters

Simple criteria were applied to rate the clusters. Table 2 gives the ratings and the criteria.

Rating	Criterion	# TF	# cluster
0	Clusters of putative E2F binding sites	1	10
High	Showing up on more than one promoters, including at least one human promoter	15	40
Middle	Showing up on one promoter, but locating within the core promoter region	3	3
Low	Otherwise	21	24

Table 2. Criteria for rating clusters. The core promoter region is defined as within 150 nt of the

TSS.

### 3 Result

15 transcription factors with 40 clusters were rated “High”, 3 transcription factors with 3 clusters were rated “Middle”, 21 transcription factors with 24 clusters were rated “Low”. The higher the rating of a cluster was, the more likely it corresponded to real a cluster of TFBSs. We reported the 15 high rating transcription factors as the transcription factors that were likely to interact with E2F family of transcription factors. Our algorithm gave the cluster and its nearest putative E2F binding site as a module. Appendix 1 and 2 show the high and middle rating clusters, the genes on which the clusters were found, and the positions of putative binding sites.

### 4 Discussion

#### 4.1 Checking the results

By searching literature, we found 6 transcription factors that were verified at various levels to interact with E2F family factors. As discussed in the introduction section, 5 of them showed up in our 15 reported transcription factors. Only one transcription factor, DMP1 [16], was missed by our algorithm. This was due to the lack of a PSSM for DMP1 in Transfac.

This result strongly argues that our algorithm can serve as a good guidance in looking for E2F’s cooperative partners. It would be interesting if in vitro or in vivo experiments can be done to test the modules given by our algorithm. We also see that our approach is limited by the fact that not all transcription factors have a PSSM compiled for their binding sites. Future work could first apply motif finding software like Bioprospector [17] to harvest de novo PSSMs and combine those PSSMs with PSSMs provided by Transfac. Then feed all those PSSMs into our algorithm.

#### 4.2 E2F’s clusters

The transcription factors that generated the largest number of clusters were the E2F family factors, which had 10 clusters on 10 different promoters. This was expected because the training set was selected to be promoters of E2F responsive genes. 10 by itself was not a large number, which indicates our selection criterion was very stringent. We may have missed some real binding sites or binding site clusters, but we would expect real binding sites to be significantly enriched in the putative clusters that we identified.

#### 4.3 A cluster of two putative ATF6 binding sites on MCM2’s promoter

There is a cluster of two putative ATF6 binding sites on the MCM2 promoter (Figure 1). ATF6’s PSSM is a very specific one (CV=2), so that any two calls happening within 300 nt would be identified as a cluster by our algorithm. The two putative binding sites on MCM2’s promoter are connected to each other and show a palindromic pattern. Such properties make this cluster statistically extremely significant. Moreover, this cluster occupies the core promoter region, and not far from it there are two putative E2F binding sites. The evidence above leads us to two hypotheses. First, ATF6 or a transcription factor that binds to ATF6 consensus binding site is important to activate or repress MCM2’s promoter. Secondly, ATF6 may cooperate with E2F family transcription factors in activating or repressing MCM2’s promoter.

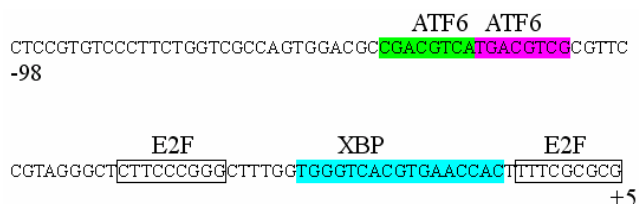


Figure 1. Putative E2F binding site and ATF6 binding sites on the MCM2 promoter.

#### 4.4 A cluster of putative P53 binding sites on the promoter of Apoptotic Protease Activating Factor 1 (Apaf1)

Our algorithm identified two putative p53 binding sites on the Apaf1 promoter. Hereafter we call the site closer to TSS BS1 and the site further to TSS BS2. The PSSM that generated this cluster is called P53\_02, which describes half of p53's tetramer binding site. It has 5 calls (2 at the same position: -603, on different strands) within the -800 nt to -500 nt region. Our algorithm did not require these calls to be very close to each other. However, a real p53 binding site usually has its two half-tetramer sites within 15 nt from each other. Interestingly, the putative half binding sites grouped by themselves into two pairs, and the distances within any pairs are both smaller than 15 nt. Figure 2 shows this cluster.



Figure 2. Putative E2F binding site and P53 binding sites on Apaf1 promoter.

Recently the P53MH algorithm, an algorithm specific to identify p53 binding sites, has been introduced [18]. P53MH was supposed to have high sensitivity and specificity in identifying p53 binding sites, for the reason that it was specifically designed for p53 only, and quite a few p53 specific properties were tuned into the algorithm. P53MH reported 10 putative p53 binding sites on the Apaf1 gene and its 10kb flanking regions on both sides. This region was larger than what we used in the training set. One of the 10 reported binding sites was BS1, but BS2 was not reported. To compare with P53MH, we applied our algorithm on the whole Apaf1 gene and its 10kb flankings. No other putative clusters were identified except for the BS1\_BS2 cluster.

Fortin et al [19] and Moroni et al [20] reported Apaf1 as a transcriptional target for p53 in the

regulation of neuronal cell death. They demonstrated from electrophoretic mobility shift assays (EMSAs) that neuronal extracts exhibited p53-DNA binding activity at both BS1 and BS2. These results suggest that in the case of Apaf1, our algorithm for general binding site discovery beats an algorithm specific to one transcription factor.

We have also found a putative E2F binding site between these two p53 binding sites. It should be interesting to interrogate the interaction between p53 and E2F on Apaf1's promoter.

## Reference

- [1] Davidson, E.H. Genomic Regulatory Systems. Development and Evolution. Academic Press, San Diego. (2001)
- [2] Berman, B. P. et al. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci USA* **99**, 757-62 (2002).
- [3] Dyson, N. The regulation of E2F by pRB-family proteins. *Genes Dev* **12**, 2245-62 (1998).
- [4] van Ginkel, P. R., Hsiao, K. M., Schjerven, H. & Farnham, P. J. E2F-mediated growth regulation requires transcription factor cooperation. *J Biol Chem* **272**, 18367-74 (1997).
- [5] Rotheneder, H., Geymayer, S. & Haidweiger, E. Transcription factors of the Sp1 family: interaction with E2F and regulation of the murine thymidine kinase promoter. *J Mol Biol* **293**, 1005-15 (1999).
- [6] Giangrande, P. H., Hallstrom, T. C., Tunyaplin, C., Calame, K. & Nevins, J. R. Identification of E-box factor TFE3 as a functional partner for the E2F3 transcription factor. *Mol Cell Biol* **23**, 3707-20 (2003).
- [7] Schlisio, S., Halperin, T., Vidal, M. & Nevins, J. R. Interaction of YY1 with E2Fs, mediated by RYBP, provides a mechanism for specificity of E2F function. *Embo J* **21**, 5775-86 (2002).

- [8] Lee, H. Y. et al. All-trans retinoic acid converts E2F into a transcriptional suppressor and inhibits the growth of normal human bronchial epithelial cells through a retinoic acid receptor-dependent signaling pathway. *J Clin Invest* 101, 1012-9 (1998).
- [9] O'Connor, D. J. et al. Physical and functional interactions between p53 and cell cycle co-operating transcription factors, E2F1 and DP1. *Embo J* 14, 6184-92 (1995).
- [10] BLAST software.  
<http://www.ncbi.nlm.nih.gov/BLAST/>
- [10] Kent, W. J. et al. The human genome browser at UCSC. *Genome Res* 12, 996-1006 (2002).
- [12] <http://www.celera.com/>
- [13]  
<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>
- [14] Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 31, 374-8 (2003)
- [15] Haiyan Huang, Ming-Chih J. Kao, Xianghong Zhou, Jun S. Liu, Wing H. Wong. Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. *Journal of Computational Biology*, In press. (2004)
- [16] Hirai, H. & Sherr, C. J. Interaction of D-type cyclins with a novel myb-like transcription factor, DMP1. *Mol Cell Biol* 16, 6457-67 (1996).
- [17] Liu, X., Brutlag, D.L. & Liu, J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 127-38 (2001).
- [18] Hoh, J. et al. The p53MH algorithm and its application in detecting p53-responsive genes. *Proc Natl Acad Sci U S A* 99, 8467-72 (2002).
- [19] Fortin, A. et al. APAF1 is a key transcriptional target for p53 in the regulation of neuronal cell death. *J Cell Biol* 155, 207-16 (2001).
- [20] Moroni, M.C. et al. Apaf-1 is a transcriptional target for E2F and p53. *Nat Cell Biol* 3, 552-8 (2001).