Genome-wide *In-silico* Identification of Transcriptional Regulators Controlling Cell Cycle in Human Cells

Ran Elkon[1#], Chaim Linhart[2#], Roded Sharan[2], Ron Shamir[2], and Yosef Shiloh[1]

[1] The David and Inez Myers Laboratory for Genetic Research, Department of Human Genetics and Molecular Medicine, Sackler School of Medicine

[2] School of Computer Science

Tel Aviv University, Tel Aviv 69978, Israel

[#] These authors contributed equally to this work.

Correspondence should be addressed to Y.S.

Telephone: 972-3-6409760

Fax: 972-3-6407471

e-mail: yossih@post.tau.ac.il

Running title: Transcriptional regulation of human cell cycle

Key words: Transcriptional regulation, cell cycle, functional genomics.

Abstract

Dissection of regulatory networks that control gene transcription is one of the greatest challenges of functional genomics. By utilizing human genomic sequences, models for binding sites of known transcription factors and gene expression data, we demonstrate that the *reverse engineering* approach, which infers regulatory mechanisms from gene expression patterns, can reveal transcriptional networks in human cells. To date, such methodologies were successfully demonstrated only in prokaryotes and low eukaryotes. We developed computational methods for identifying putative binding sites of transcription factors and for evaluating the statistical significance of their prevalence in a given set of promoters. Focusing on transcriptional mechanisms that control cell cycle progression, our computational analyses revealed eight transcription factors whose binding sites are significantly over-represented in promoters of genes whose expression is cell cycle dependent. The enrichment of some of these factors is specific to certain phases of the cell cycle. In addition, several pairs of these transcription factors show a significant co-occurrence rate in cell cycle-regulated promoters. Each such pair suggests functional cooperation between its members in regulating the transcriptional program associated with cell cycle progression. The methods presented here are general and can be applied to the analysis of transcriptional networks controlling any biological process.

Introduction

With completion of sequencing of the human genome, focus has shifted from sequencing and mapping genes to functional genomics. The goal of functional genomics is not merely to assign genes into functional categories, but also to provide a comprehensive understanding of genetic networks — to disclose how gene products interact and regulate each other to produce coherent and coordinated physiological processes and responses to homeostatic challenges (Lockhart and Winzeler 2000). A hallmark of functional genomics is the attempt to characterize biological pathways and processes in a holistic manner (Lander and Weinberg 2000). The holistic approach has become feasible in the study of biological systems thanks to the availability of genome sequences of many organisms, the maturation of high-throughput genome-scale technologies, and the development of computational tools to analyze the rapidly accumulating volume of biological data.

Regulation of transcription is a key component of physiological networks. Indeed, it is the endpoint of many signal transduction pathways emanating from either extracellular or intracellular triggers. Transcription of genes is controlled primarily via regulatory sequence elements that are recognized and bound by transcription factors (TFs). Transcriptional regulation in eukaryotes is combinatorial in nature. The expression pattern of any particular gene is determined by an interplay among several TFs that bind its promoter. Thus, a major task of deciphering transcriptional regulation networks is to identify combinations of TFs that cooperate in the regulation of genes and form a recurrent regulation motif, termed a *regulation module*. Recent works successfully undertook a computational approach for genome-wide mapping of transcriptional regulation modules involved in the regulation of Drosophila development (Berman et al. 2002; Halfon et al. 2002; Markstein et

al. 2002). Transcriptional modules in mammalian cells were defined and identified by several pioneering works (Frech et al. 1998; Kel et al. 1999; Wasserman and Fickett 1998).

The use of DNA microarrays to study global gene expression profiles is emerging as a pivotal technology in functional genomics. Comparison of gene expression profiles under different biological conditions reveals the corresponding modifications in the cellular transcriptional programs. Microarray measurements do not, however, directly reveal the regulatory networks that underlie the observed transcriptional modulation. Combining promoter analysis with microarray results can shed light on those networks. Recent studies integrated computational promoter analysis and microarray data to identify novel transcriptional regulatory networks in *S. cerevisiae* (Jelinsky et al. 2000; Pilpel et al. 2001; Tavazoie et al. 1999) These studies demonstrated that genes that are co-expressed over multiple biological conditions are often regulated via common mechanisms, and, hence, share common cis-regulatory elements in their promoters.

We developed novel computational approaches that utilize the human genome and data from high-throughput functional genomics technologies to dissect transcriptional regulation networks. Our methods identify TFs whose binding sites are significantly over-represented in specific sets of promoters, as well as pairs of TFs whose binding sites exhibit a significant co-occurrence rate. Applying these methods to the analysis of cell cycle regulation in human cells disclosed key regulators in the cell cycle transcriptional program and pointed to several possible inter-connections among these regulators.

Results

Extraction of putative promoters from the human genome data

As a first step in our analysis we constructed a set of putative promoter sequences of the known human genes. To this aim we downloaded the human genome data assembled into genomic contigs by the NCBI Reference Sequence project (Maglott et al.

2000)ftp://ftp.ncbi.nih.gov/genomes/H_sapiens, release of June 2001). We used the version in which human repetitive sequences are masked (mfa files). From these genomic contigs, putative promoter sequences of known human genes were extracted based on genes' start annotations provided by NCBI (gbs files provided at the same url). We determined the length of sequence around the putative TSS in which to search for transcriptional regulatory elements by examining the location distribution of 1075 empirically validated TF binding sites in human promoters (data from TRANSFAC database (Wingender et al. 2000)). Since 80% of these elements were located within 1,200 bases upstream of the genes' transcription start site (TSS) (data not shown), our analyses were confined to this region. Clearly, current knowledge is biased towards binding sites short distances from the TSS. Certain regulatory elements were demonstrated to act over very great distances, up to several kilobases from TSS, but it is clear that ample information resides in sequences in close proximity to the TSS. Our promoter set contains sequences for putative promoter regions of 12,981 known human genes, each 1,200 bp in length. This promoters set is referred to as the '13K set'. To estimate the accuracy of this promoters set, we compared it with experimentally validated human promoters taken from the EPD database (Praz et al. 2002). EPD contains validated promoter sequences for 247 distinct human genes. The 13K set contains promoter sequences for 180 of these genes. When the pairs of putative and validated promoters were alligned, the distance between the putative and true TSS was within 200 bp in 70% of cases (data not shown). The 13K set can be downloaded from http://www.cs.tau.ac.il/~rshamir/prima/PRIMA.htm.

In-silico identification of TFs that synergize with E2F

The aim of our first approach is to reveal, by *in-silico* analysis, TFs that cooperate with any particular TF of interest. The scheme of the analysis is as follows: A set of promoters of genes that are directly regulated by the TF of interest (termed *targets* of this TF) is constructed and scanned for over-represented binding sites corresponding to other TFs.

Such over-representations may point to a functional link between the over-represented TFs and the TF of interest. Here we employed this scheme in an attempt to ferret out TFs that cooperate with E2F. Since robust statistics require as large a set of E2F targets as possible, we used recent results published by Ren et al. (Ren et al. 2002), who combined ChIP (chromatin immunoprecipitation) and microarray technologies to identify 124 genes whose promoters bind either E2F1 or E2F4 *in-vivo*. Our 13K set contains promoter sequences for 103 of these genes. This set of E2F target promoters was scanned with experimentally-derived position weight matrices (*PWMs*) for 107 human TFs (PWMs are from TRANSFAC database (Wingender et al. 2000)). The occurrence frequency of each PWM in the E2F target set and in the 13K set, which served as a background set, was compared, and an analytical score computed for the significance of its observed abundance in the E2F target set (see Methods for details). For those PWMs that achieved a highly significant analytical score we applied an additional empirical test vs. random promoter sets. We determined the occurrence frequency of those high-scoring PWMs on 10,000 subsets of promoters that were randomly chosen from the 13K set and with the same size as the target set (103 promoters). We report only PWMs whose abundance on the E2F target set was significantly higher than on the random sets. The screening criterion we applied corresponded to $p<0.05$ after accounting for multiple testing (see Methods for details). We identified four significantly enriched PWMs in the E2F target set (Table 1). As expected, the PWM of E2F itself is highly enriched in this set. Since E2F is a true positive in this set, the identification of its PWM demonstrates the ability of our approach to detect true signals. PWMs of three TFs — NF-Y, CREB and NRF-1 — are also significantly enriched, pointing to possible functional links between these TFs and E2F.

Utilization of functional annotation in dissection of regulatory mechanisms

Hughes et al. (Hughes et al. 2000) demonstrated that groups of functionally related genes in *S. cerevisiae* often share common cis-regulatory elements in their promoters. Hence, analyzing promoters of genes with common function could reveal regulatory elements characteristic to specific functional categories. We examined whether this approach could be applied to human promoters, using the functional categorization of human genes provided by the LocusLink DB (Maglott et al. 2000), which employs the standard Gene Ontology vocabulary for description of biological processes (Ashburner et al. 2000). We focused on four cell cycle-related categories: cell cycle control, mitotic cell cycle, DNA metabolism, and M phase (some genes are assigned to several functional categories, hence the groups are not mutually exclusive). The methodology described above was applied to each category, again using the 13K set as the background set and scanning with all 107 PWMs. Significantly enriched PWMs were revealed in all functional categories (Table 2). The E2F PWM is enriched in all categories, reflecting its central role in regulating these processes. Notably, it is enriched in promoters of genes known to function in the M phase of the cell cycle. This is in accordance with recent studies (Ishida et al. 2001; Polager et al. 2002) showing that E2F's role in controlling the cell cycle goes beyond its previously documented control of the entry into the S phase. NF-Y and NRF-1 PWMs are enriched in three out of the four categories, Sp1 PWM is enriched in the cell cycle control and DNA metabolism categories, and ETF and ATF PWMs are enriched in the cell cycle control and the M phase categories, respectively.

Deciphering regulatory mechanisms using gene expression data

Next, we undertook the reverse engineering approach which infers transcriptional regulatory mechanisms from gene expression data. We analyzed the human cell cycle dataset published recently by Whitfield et al. (Whitfield et al. 2002). Their study recorded genome-

wide gene expression levels over multiple time points during the progression of cell cycle in HeLa human cell line; 874 genes showed periodic expression patterns over several cell cycles. Our 13K promoters set contains putative promoter sequences for 568 of these genes. Whitfield et al. (Whitfield et al. 2002) partitioned the cell cycle regulated genes according to their expression periodicity patterns into five clusters, corresponding to cell cycle phases G1/S, S, G2, G2/M and M/G1. We analyzed clusters of 103, 105, 122, 145 and 93 promoters, respectively.

We searched for significantly enriched PWMs in the entire set of the 568 cell cycle-regulated promoters using the 13K set as the background set. Six out of the 107 PWMs, corresponding to E2F, NF-Y, NRF-1, Sp1, ATF and CREB TFs, were significantly over-represented in this target set (Table 3*a*). We then searched for PWMs enriched only in specific phase clusters; Arnt and YY1 PWMs were specifically enriched in the G1/S and the M/G1 clusters, respectively (Table 3*b*). Caution must be exercised when examining whether PWMs that were enriched in the entire set favor any specific phase cluster. Given their significant over-representation in the entire set, random partitions of the dataset are also expected to yield clusters where these PWMs are enriched with respect to their genomic prevalence. So, what should be tested is whether these PWMs favor any specific phase cluster given their prevalence in this dataset rather than their genomic background prevalence. Hence, in this examination, the set of 568 cell cycle-regulated promoters was used as the background set. E2F PWM was found to be significantly over-represented in the G1/S and S phases ($p=3.2*10^{-7}$ for the observed prevalence in these 2 clusters together) and under-represented in the M/G1 cluster ($p=0.015$); NF-Y PWM was over-represented in the G2 and G2/M phases ($p=0.0096$ for the observed prevalence in these 2 clusters together); and Sp1 PWM slightly favored the G1/S cluster ($p=0.02$). NRF-1, ATF and CREB PWMs were more uniformly distributed and showed no bias for any particular phase (Fig. 1).

We examined the location distribution of the computationally identified binding sites of the enriched PWMs. The putative binding sites for E2F, NF-Y, NRF-1, Sp1, ATF and CREB tend to concentrate in the proximity of the TSS (Fig 2). This observation is in agreement with experimental data on the locations of *in-vivo* binding sites of E2F (Kel et al. 2001) and NF-Y (Mantovani 1998). In addition to the fact that the positions of the computationally identified hits are not uniformly distributed, but rather concentrated near the TSSs, we also observed that their occurrence rate declines sharply downstream the putative TSSs (data not shown). These observations provide an additional indication for the accuracy of the putative promoters we used.

Identification of co-occurring pairs of TFs

The approach described thus far identified TF PWMs that were enriched in target sets of promoters, with the tests performed separately on each PWM. Finding several enriched PWMs on the same target set may indirectly point to functional links between the corresponding TFs. We sought a direct method to test the associations between distinct PWMs. In an effort to identify pairs of PWMs that exhibit a significant tendency to appear together in the same promoters, we examined whether the prevalence of promoters containing hits for two PWMs was significantly higher than would be expected if the PWMs occurred independently. This analysis was applied to the set of 568 promoters of cell cycle-regulated genes. We examined all possible pairs formed by the 9 PWMs found to be enriched in any of the analyses reported above. Eight pairs showed a significant tendency to co-occur in this promoter set. Each such pair constitutes a hypothetical regulatory module, or a part thereof (Fig. 3). Figure 3 suggests that NRF-1, Sp1, ETF and E2F may constitute transcriptional modules of higher orders, i.e., recurrent motifs of three or four TFs.

Discussion

The computational approaches presented here utilize the human genome sequence and data obtained by large-scale functional genomics technologies to determine putative regulatory mechanisms that control the transcriptional program of the cell cycle in human cells. Our analyses identified eight TFs whose regulatory sequences are significantly enriched in promoters of cell cycle-regulated genes. The enrichment of several of these TFs was shown to be specific for certain phases of the cell cycle.

The E2F family is well documented as a prime regulator of the mammalian cell cycle. Pathways that modulate the activity of E2F are frequently disrupted in human cancers, leading to misregulated cellular proliferation (Nevins 2001). The E2F PWM obtained highly significant enrichment scores in all our analyses, demonstrating the sensitivity of our methods to reveal true signals. The role of this family of TFs in the cell cycle was underscored by several recent studies showing that E2F regulates not only genes that function in the G1/S and S phases, but also many M phase genes (Ishida et al. 2001; Polager et al. 2002). Our analysis indicates that the E2F PWM is indeed enriched in promoters of genes that are expressed in G2, although its enrichment in promoters of genes that are expressed in G1/S and S phases is much more prominent (Fig. 1).

Published experimental data support our findings on most of the other TFs as well. NF-Y and Sp1 PWMs obtained highly significant enrichment scores. Though involved in many different aspects of cellular life, both TFs have an established role in the regulation of the cell cycle. NF-Y was demonstrated to control the expression of several key regulators of the cell cycle (Jung et al. 2001; Manni et al. 2001; Yun et al. 1999). The transcriptional activity of Sp1 is modulated in a cell cycle-dependent manner through its phosphorylation by Cyclin A-CDK complexes (Fojas de Borja et al. 2001). In addition, several cell cycle

regulators were reported to be controlled by Sp1 (Cram et al. 2001; Eto 2000; Martino et al. 2001; Paskind et al. 2000).

Our analysis shows that E2F and NF-Y binding sites, as well as E2F and Sp1 binding sites, significantly co-occur in promoters of cell cycle-regulated genes, suggesting functional cooperation between these TFs in the regulation of cell cycle progression. Experimental evidence supports the existence of such relations. Physical interactions were demonstrated between members of the E2F and Sp1 families (Rotheneder et al. 1999), and functional cooperation between E2F and Sp1 was reported in several cell cycle-related promoters (Chang et al. 2001; Huang et al. 2001; Nishikawa et al. 2001; Parisi et al. 2002; Rotheneder et al. 1999). As for E2F and NF-Y, co-occurrence of functional binding sites for both TFs was reported in several promoters, including Cdc2, TK, POLA, Cyclin A and several histone genes (Matuoka and Yu Chen 1999). Functional synergism between E2F and NF-Y was demonstrated in the regulation of the E2F-1 promoter (van Ginkel et al. 1997). Our findings substantially expand the generality of these functional links, pointing to possible synergism between these TFs on dozens of cell cycle-regulated promoters.

Other TFs that were significantly over-represented in cell cycle-related promoters in our analyses have not been established as prominent regulators of the cell cycle, but data suggest they are involved in regulation of cellular proliferation. ATF/CREB is a family of over a dozen TFs that bind a common regulatory element, the ATF/CRE (cAMP Response Element) motif. One member of the family, CREB, undergoes cell cycle-regulated phosphorylation (Saeki et al. 1999), and was recently reported to control the expression of multiple cell cycle regulatory genes (Klemm et al. 2001). Over-expression of another family member, ATF2, inhibits the G1/S phase transition in human cancer cell line (Crowe and Shemirani 2000), and is directly involved in the regulation of cyclin A (Djaborkhel et al. 2000) and cyclin D1 (Recio and Merlino 2002).

YY1 was reported to control several S-phase induced genes (Johansson et al. 1998; Wu and Lee 2001). Over-expression of YY1 was reported to induce DNA synthesis (Petkova et al. 2001). Furthermore, a cell cycle-regulated physical interaction between YY1 and pRb was reported in the same study. These findings link YY1 to induction of the S phase. In contrast, we found the YY1 PWM to be under-represented in the S phase, but significantly enriched in the M/G1 cluster.

Arnt forms a dimeric TF with the aryl hydrocarbon receptor (AhR). It is implicated in developmental processes and tissue homeostasis. Several studies linked the AhR-Arnt dimer to cell cycle regulation. Activation of AhR was reported to induce G1 arrest (Puga et al. 2000; Weiss et al. 1996). Recently, this negative regulation was shown to depend on physical interaction between AhR and pRb (Elferink et al. 2001). In agreement, we find the enrichment of the Arnt PWM in the G1/S cluster.

Transition of cells from quiescence to proliferation increases the cell demand for energy. One way of responding to the increased demand for ATP is to modulate the activity of the respiratory chain components. NRF-1 regulates the expression of many genes required for mitochondrial respiratory function (Evans and Scarpulla 1990). A recent study demonstrated that NRF-1 activity is enhanced by phosphorylation upon serum-induced proliferation, leading to transcriptional induction of cytochrome c, a major component of the respiratory apparatus (Herzig et al. 2000). The induction of cytochrome c was associated with enhanced energy production by the mitochondria in preparation for entry to the cell cycle. The induction of cytochrome c in response to serum was shown to be mediated by both NRF-1 and CREB (Herzig et al. 2000). Interestingly, this is one of the pairs we identified, and is possibly involved in the cellular metabolic transition to the proliferative phase. In addition, our analysis suggests that NRF-1, together with Sp1, ETF and E2F, form a recurrent motif of three or four TFs (Fig. 3).

By employing genome-wide *in-silico* computational analyses of promoters, we identified key regulators of the transcriptional program of the cell cycle in human cells. Several pairs of these TFs showed a significant co-occurrence rate on promoters of cell cycle-regulated genes. We expect that our findings will provide guidelines for experimental dissection of the regulatory mechanisms controlling the cell cycle in mammalian cells. Moreover, the methods demonstrated here are general and can be applied to the analysis of transcriptional networks controlling any biological process. We anticipate that this type of transcriptional regulation network dissection will become an integral part of the analysis of data obtained from gene expression microarrays and large scale chromatin immunoprecipitation studies, not only in low eukaryotes but also in mammals.

Methods

A set of known human TF position weight matrices. Binding sites that are recognized and bound by TFs are commonly modeled by consensus sequences or *position weight matrices* (*PWMs*). As the latter are more informative, we used this type of model in our promoter analysis. PWMs for known human TF binding sites were obtained from the TRANSFAC database (Wingender et al. 2000) (release 5.4, April 2002). A total of 107 PWMs that correspond to distinct TFs (according to the TF name's field in the PWM entry) were used in our analyses. Some TFs recognize similar binding sites so this PWM set might contain correlated matrices. All PWMs we used are based on at least 5 binding sites.

Scanning a set of promoters for over-represented PWMs. We developed a program, called PRIMA (PRomoter Integration in Microarray Analysis), written in Perl and C, for scanning a given set of promoters for TF binding sites and identifying PWMs that are significantly over-represented in the examined set in comparison with a background set of promoters. Given a PWM P of length l, both strands of each promoter are scanned by sliding

a window of length *l* along the promoter. At each position of the window, a similarity score is computed between *P* and the corresponding subsequence of the promoter. Denote by *p(i,j)* the frequency of base *i* at position *j* in the PWM *P*. Given a promoter subsequence $s_1 s_2 \ldots s_l$, we define its similarity to *P* as follows:

$$sim(P, s_1 s_2 ... s_l) = \prod_{j=1}^{l} p(s_j, j)$$

In order to identify putative binding sites, or *hits*, of a TF, a threshold *T(P)* for the similarity score of the TF's PWM *P* is determined. Subsequences with a similarity score above *T(P)* are regarded as hits of *P*. The threshold *T(P)* is controlled by two parameters, α and β. The first parameter controls the rate of hits of *P* in random sequences as follows: A set of 400 random promoters of the same length as the real promoters is generated by an order-2 Markov model learnt from the background promoters. A threshold T_1 is computed, such that α percent of the random promoters contain one or more sites whose similarity score to *P* is above T_1. The second parameter, β, controls the rate of hits of *P* in a background set of promoters. A threshold T_2 is computed, such that β background promoters contain one or more sites whose similarity score to *P* is above T_2. The threshold *T(P)* is set as the minimum of T_1 and T_2. Unless otherwise stated in the text, in the reported experiments, the 13K set was used as the background set of promoters, α=10%, and β=1,000. Although the choice of these particular parameter values is somewhat arbitrary, the choice of other values gave similar results.

Once a similarity score threshold is set, the PWM *P* is used to scan the promoters. Given a set *B* of *n* background promoters, and a subset *T* of *m* target promoters, we compute an analytical score for the observed enrichment of PWM *P* in *T* with respect to its abundance in *B*. Suppose there are *h* hits of *P* in *T*, where at most three hits are counted per promoter. Let n_1, n_2 and n_3 denote the number of background promoters containing one, two, or at least

three hits, respectively. Assuming that T is randomly chosen out of B, the analytical score for the probability of observing at least h hits in T is:

$$p = \frac{\sum_{i+2j+3k \geq h} \binom{n_1}{i}\binom{n_2}{j}\binom{n_3}{k}\binom{n-n_1-n_2-n_3}{m-i-j-k}}{\binom{n}{m}}$$

We used the computed analytical score as a first filter. PWMs that achieved $p \leq 0.001$ were subjected to an empirical statistical test. We tested how often each of these PWMs received at least h hits on 10,000 random sets of promoters. Each set was generated by randomly choosing a subset of m background promoters from B. We report the PWMs whose observed abundance in T ranked among the top five within the 10,000 random sets. The implied significance level of this cut-off is 0.05, when applying Bonferroni correction for multiple testing of 107 distinct PWMs.

PRIMA software can be downloaded from http://www.cs.tau.ac.il/~rshamir/prima/PRIMA.htm

Identification of co-occurring pairs of PWMs. Given a set of m promoters, and a pair of PWMs, P_a and P_b, denote by f_a, f_b the number of promoters that contain a hit for P_a, P_b, respectively. Let f_{ab} be the number of promoters with a hit for both P_a and P_b. The p-value for observing f_{ab} or more promoters containing hits for both PWMs is:

$$p = \sum_{h=f_{ab}}^{\min\{f_a, f_b\}} \frac{\binom{f_a}{h}\binom{m-f_a}{f_b-h}}{\binom{m}{f_b}}$$

In this analysis we used α=20%, β=2,000. Overlapping hits of P_a and P_b were omitted from counting. We only report pairs that remain significant ($p<0.05$) after accounting for the multiple testing performed (36 pairs were tested).

Supplementary data. Full lists of genes whose promoters were found to contain high scoring sites for any of the enriched TFs reported in Table 1 and Table 3, are provided as supplementary data.

Accession numbers of reported PWMs. The accession numbers in TRANSFAC database (Wingender et al. 2000) of the reported TF's PWMs are: E2F - M00516, Sp1 - M00196, NF-Y - M00185, NRF-1 - M00652, ETF - M00695, ATF - M00338, CREB - M00113, Arnt - M00236, YY1 - M00069.

Acknowledgments

R. Elkon is a Joseph Sassoon Fellow. R. Sharan was supported by an Eshkol Fellowship from the Ministry of Science, Israel. This study was supported by a research grant from the Ministry of Science and Technology, Israel. This work was carried out in partial fulfillment of the requirements for Ph.D. degree of R. Elkon.

References

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25-29.

Berman, B.P., Y. Nibu, B.D. Pfeiffer, P. Tomancak, S.E. Celniker, M. Levine, G.M. Rubin, and M.B. Eisen. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A* 99: 757-762.

Chang, Y.C., S. Illenye, and N.H. Heintz. 2001. Cooperation of E2F-p130 and Sp1-pRb complexes in repression of the Chinese hamster dhfr gene. *Mol Cell Biol* 21: 1121-1131.

Cram, E.J., B.D. Liu, L.F. Bjeldanes, and G.L. Firestone. 2001. Indole-3-carbinol inhibits CDK6 expression in human MCF-7 breast cancer cells by disrupting Sp1 transcription factor interactions with a composite element in the CDK6 gene promoter. *J Biol Chem* 276: 22332-22340.

Crowe, D.L. and B. Shemirani. 2000. The transcription factor ATF-2 inhibits extracellular signal regulated kinase expression and proliferation of human cancer cells. *Anticancer Res* 20: 2945-2949.

Djaborkhel, R., D. Tvrdik, T. Eckschlager, I. Raska, and J. Muller. 2000. Cyclin A down-regulation in TGFbeta1-arrested follicular lymphoma cells. *Exp Cell Res* 261: 250-259.

Elferink, C.J., N.L. Ge, and A. Levine. 2001. Maximal aryl hydrocarbon receptor activity depends on an interaction with the retinoblastoma protein. *Mol Pharmacol* 59: 664-673.

Eto, I. 2000. Molecular cloning and sequence analysis of the promoter region of mouse cyclin D1 gene: implication in phorbol ester-induced tumour promotion. *Cell Prolif* 33: 167-187.

Evans, M.J. and R.C. Scarpulla. 1990. NRF-1: a trans-activator of nuclear-encoded respiratory genes in animal cells. *Genes Dev* 4: 1023-1034.

Fojas de Borja, P., N.K. Collins, P. Du, J. Azizkhan-Clifford, and M. Mudryj. 2001. Cyclin A-CDK phosphorylates Sp1 and enhances Sp1-mediated transcription. *Embo J* 20: 5737-5747.

Frech, K., K. Quandt, and T. Werner. 1998. Muscle actin genes: a first step towards computational classification of tissue specific promoters. *In Silico Biol* 1: 29-38.

Halfon, M.S., Y. Grad, G.M. Church, and A.M. Michelson. 2002. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res* 12: 1019-1028.

Herzig, R.P., S. Scacco, and R.C. Scarpulla. 2000. Sequential serum-dependent activation of CREB and NRF-1 leads to enhanced mitochondrial respiration through the induction of cytochrome c. *J Biol Chem* 275: 13134-13141.

Huang, D., M. Jokela, J. Tuusa, S. Skog, K. Poikonen, and J.E. Syvaoja. 2001. E2F mediates induction of the Sp1-controlled promoter of the human DNA polymerase epsilon B-subunit gene POLE2. *Nucleic Acids Res* 29: 2810-2821.

Hughes, J.D., P.W. Estep, S. Tavazoie, and G.M. Church. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 296: 1205-1214.

Ishida, S., E. Huang, H. Zuzan, R. Spang, G. Leone, M. West, and J.R. Nevins. 2001. Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis. *Mol Cell Biol* 21: 4684-4699.

Jelinsky, S.A., P. Estep, G.M. Church, and L.D. Samson. 2000. Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes. *Mol Cell Biol* 20: 8157-8167.

Johansson, E., K. Hjortsberg, and L. Thelander. 1998. Two YY-1-binding proximal elements regulate the promoter strength of the TATA-less mouse ribonucleotide reductase R1 gene. *J Biol Chem* 273: 29816-29821.

Jung, M.S., J. Yun, H.D. Chae, J.M. Kim, S.C. Kim, T.S. Choi, and D.Y. Shin. 2001. p53 and its homologues, p63 and p73, induce a replicative senescence through inactivation of NF-Y transcription factor. *Oncogene* 20: 5818-5825.

Kel, A., O. Kel-Margoulis, V. Babenko, and E. Wingender. 1999. Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J Mol Biol* 288: 353-376.

Kel, A.E., O.V. Kel-Margoulis, P.J. Farnham, S.M. Bartley, E. Wingender, and M.Q. Zhang. 2001. Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. *J Mol Biol* 309: 99-120.

Klemm, D.J., P.A. Watson, M.G. Frid, E.C. Dempsey, J. Schaack, L.A. Colton, A. Nesterova, K.R. Stenmark, and J.E. Reusch. 2001. cAMP response element-binding protein content is a molecular determinant of smooth muscle cell proliferation and migration. *J Biol Chem* 276: 46132-46141.

Lander, E.S. and R.A. Weinberg. 2000. Genomics: journey to the center of biology. *Science* 287: 1777-1782.

Lockhart, D.J. and E.A. Winzeler. 2000. Genomics, gene expression and DNA arrays. *Nature* 405: 827-836.

Maglott, D.R., K.S. Katz, H. Sicotte, and K.D. Pruitt. 2000. NCBI's LocusLink and RefSeq. *Nucleic Acids Res* 28: 126-128.

Manni, I., G. Mazzaro, A. Gurtner, R. Mantovani, U. Haugwitz, K. Krause, K. Engeland, A. Sacchi, S. Soddu, and G. Piaggio. 2001. NF-Y mediates the transcriptional inhibition of the cyclin B1, cyclin B2, and cdc25C promoters upon induced G2 arrest. *J Biol Chem* 276: 5570-5576.

Mantovani, R. 1998. A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Res* 26: 1135-1143.

Markstein, M., P. Markstein, V. Markstein, and M.S. Levine. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc Natl Acad Sci U S A* 99: 763-768.

Martino, A., J.H.t. Holmes, J.D. Lord, J.J. Moon, and B.H. Nelson. 2001. Stat5 and Sp1 regulate transcription of the cyclin D2 gene in response to IL-2. *J Immunol* 166: 1723-1729.

Matuoka, K. and K. Yu Chen. 1999. Nuclear factor Y (NF-Y) and cellular senescence. *Exp Cell Res* 253: 365-371.

Nevins, J.R. 2001. The Rb/E2F pathway and cancer. *Hum Mol Genet* 10: 699-703.

Nishikawa, N., M. Izumi, M. Yokoi, H. Miyazawa, and F. Hanaoka. 2001. E2F regulates growth-dependent transcription of genes encoding both catalytic and regulatory subunits of mouse primase. *Genes Cells* 6: 57-70.

Parisi, T., A. Pollice, A. Di Cristofano, V. Calabro, and G. La Mantia. 2002. Transcriptional regulation of the human tumor suppressor p14(ARF) by E2F1, E2F2, E2F3, and Sp1-like factors. *Biochem Biophys Res Commun* 291: 1138-1145.

Paskind, M., C. Johnston, P.M. Epstein, J. Timm, D. Wickramasinghe, E. Belanger, L. Rodman, D. Magada, and J. Voss. 2000. Structure and promoter activity of the mouse CDC25A gene. *Mamm Genome* 11: 1063-1069.

Petkova, V., M.J. Romanowski, I. Sulijoadikusumo, D. Rohne, P. Kang, T. Shenk, and A. Usheva. 2001. Interaction between YY1 and the retinoblastoma protein. Regulation of cell cycle progression in differentiated cells. *J Biol Chem* 276: 7932-7936.

Pilpel, Y., P. Sudarsanam, and G.M. Church. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* 29: 153-159.

Polager, S., Y. Kalma, E. Berkovich, and D. Ginsberg. 2002. E2Fs up-regulate expression of genes involved in DNA replication, DNA repair and mitosis. *Oncogene* 21: 437-446.

Praz, V., R. Perier, C. Bonnard, and P. Bucher. 2002. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res* 30: 322-324.

Puga, A., S.J. Barnes, T.P. Dalton, C. Chang, E.S. Knudsen, and M.A. Maier. 2000. Aromatic hydrocarbon receptor interaction with the retinoblastoma protein potentiates repression of E2F-dependent transcription and cell cycle arrest. *J Biol Chem* 275: 2943-2950.

Recio, J.A. and G. Merlino. 2002. Hepatocyte growth factor/scatter factor activates proliferation in melanoma cells through p38 MAPK, ATF-2 and cyclin D1. *Oncogene* 21: 1000-1008.

Ren, B., H. Cam, Y. Takahashi, T. Volkert, J. Terragni, R.A. Young, and B.D. Dynlacht. 2002. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev* 16: 245-256.

Rotheneder, H., S. Geymayer, and E. Haidweger. 1999. Transcription factors of the Sp1 family: interaction with E2F and regulation of the murine thymidine kinase promoter. *J Mol Biol* 293: 1005-1015.

Saeki, K., A. Yuo, and F. Takaku. 1999. Cell-cycle-regulated phosphorylation of cAMP response element-binding protein: identification of novel phosphorylation sites. *Biochem J* 338 (Pt 1): 49-54.

Tavazoie, S., J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. 1999. Systematic determination of genetic network architecture. *Nat Genet* 22: 281-285.

van Ginkel, P.R., K.M. Hsiao, H. Schjerven, and P.J. Farnham. 1997. E2F-mediated growth regulation requires transcription factor cooperation. *J Biol Chem* 272: 18367-18374.

Wasserman, W.W. and J.W. Fickett. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* 278: 167-181.

Weiss, C., S.K. Kolluri, F. Kiefer, and M. Gottlicher. 1996. Complementation of Ah receptor deficiency in hepatoma cells: negative feedback regulation and cell cycle control by the Ah receptor. *Exp Cell Res* 226: 154-163.

Whitfield, M.L., G. Sherlock, A.J. Saldanha, J.I. Murray, C.A. Ball, K.E. Alexander, J.C. Matese, C.M. Perou, M.M. Hurt, P.O. Brown, and D. Botstein. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* 13: 1977-2000.

Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter, and F. Schacherer. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 28: 316-319.

Wu, F. and A.S. Lee. 2001. YY1 as a regulator of replication-dependent hamster histone H3.2 promoter and an interactive partner of AP-2. *J Biol Chem* 276: 28-34.

Yun, J., H.D. Chae, H.E. Choy, J. Chung, H.S. Yoo, M.H. Han, and D.Y. Shin. 1999. p53 negatively regulates cdc2 transcription via the CCAAT-binding NF-Y transcription factor. *J Biol Chem* 274: 29677-29682.

Web Site References

ftp://ftp.ncbi.nih.gov/ genomes/H_sapiens; Human Genome data at NCBI.

http://www.gene-regulation.de; TRANSFAC database.

http://www.ncbi.nlm.nih.gov/LocusLink; LocusLink database.

http://genome-www.stanford.edu/Human-CellCycle/Hela/data.shtml; Human cell cycle microarray dataset.

Figure legends

Fig. 1 Representation of TF PWMs in the cell cycle phase clusters. The eight circles correspond to the PWMs that were highly enriched in promoters of cell cycle-regulated genes (Table 3). Each circle is divided into 5 zones, corresponding to the phase clusters. The number adjacent to the zone represents the ratio of its prevalence in promoters contained in each of the cell cycle phase clusters to its prevalence in the set of 13K background promoters. Note that several TFs show a tendency towards specific cell cycle phases: e.g., over-representation of the E2F PWM in promoters of the G1/S and S clusters, and its under-representation in promoters of the M/G1 cluster.

Fig. 2 Distribution of locations of TFs putative binding sites found in 568 cell cycle-regulated promoters. Promoters were divided into six intervals, 200 bp each. For each of the PWMs listed in Table 3, the number of times its computationally identified binding sites appeared in each interval was counted (after accounting for the actual number of bps scanned in each interval. This number changes as the masked sequences are not uniformly distributed among the six intervals). Locations of NRF-1, CREB, NF-Y, Sp1, ATF and E2F binding sites tend to concentrate in the vicinity of the TSSs (chi-square test, $p<0.01$).

Fig. 3 Pairs of PWMs that co-occur significantly in promoters of genes regulated in a cell cycle manner. We examined whether the nine PWMs reported in Tables 1-3 can be organized into regulatory modules. For each possible pair formed by these PWMs, we tested whether the prevalence of cell cycle-regulated promoters that contain hits for both PWMs is significantly higher than would be expected if the PWMs occurred independently. Eight significant pairs were identified, each connected by an edge. The corresponding p-value is indicated next to the edge. The edge connecting the E2F-NRF1 pair is dashed to indicate that its significance is borderline.

Table legends

Table 1. A set of 103 promoters corresponding to E2F target genes reported by Ren et al. (Ren et al. 2002) was scanned for over-represented binding sites corresponding to 107 human TF PWMs. Four significantly enriched PWMs were found. Indicated for each one are: the number of promoters with hits of the PWM and the total number of hits of the PWM (some promoters have multiple hits of a PWM), the analytical score for observing such enrichment, and the rank of the PWM's abundance in the E2F target set relative to its abundance in 10,000 sets of randomly selected promoters of the same size as that of the E2F target set. Similarity score thresholds for declaring hits were stringently determined in order to enable identification of real enrichments in the examined set. Therefore, the number of promoters having E2F binding sites in this E2F target set is underestimated. Nevertheless, the observed occurrence rate of E2F is highly significant. Notably, the enrichment of the NF-Y PWM in this set is even more significant than the enrichment of the E2F PWM. Full lists of genes whose promoters were found to contain high scoring sites for the enriched TFs are provided in supplemental Tables A1-A4.

Table 2. Promoters in the 13K set were assigned to functional categories. Functional annotations of genes were extracted from LocusLink DB, which utilizes the GO vocabulary (Maglott et al. 2000). Four categories related to cell cycle, containing a total of 672 distinct genes, were analyzed (certain genes are assigned to several categories; hence the categories are not mutually exclusive). The number of promoters and the TF PWMs significantly enriched in each category are indicated. Indicated for each over-represented PWM are the analytical score for observing such enrichment, and the rank of the PWM's abundance in the functional category relative to its

abundance in 10,000 sets of randomly selected promoters of the same size as that of the functional category set are. Numbers in parentheses represent the number of random sets in which the PWM was equally abundant as in the functional category set.

Table 3. *a*. A set of 568 promoters of cell cycle-regulated genes scanned for over-represented TF PWMs, disclosing six significantly enriched PWMs. Information for each PWM is as in Table 1. *b.* Whitfield et al. (Whitfield et al. 2002) partitioned the cell cycle-regulated genes according to their expression periodicity patterns into five clusters corresponding to different phases of the cell cycle. When the promoter sequences of these clusters were scanned for enriched PWMs, two PWMs were enriched in a specific phase cluster, but not in the 568 set as a whole. Full lists of genes whose promoters were found to contain high scoring sites for the enriched TFs are provided in supplemental Tables B1-B8.

Supplementary Tables. Tables A1-A4 and B1-B8 list the genes whose promoters were found to contain high scoring sites for the TFs reported in Table 1 and 3, respectively. For each gene, a header line specifies its ID and symbol in NCBI's LocusLink DB (http://www.ncbi.nlm.nih.gov/LocusLink). The putative binding sites contained in the gene's promoter follow the header line. For each site the table specifies its sequence, strand, position (relative to putative TSS) and its similarity score to the TF's PWM. In these lists genes are sorted such that those with the highest scoring sites are at the top.

Tables

Table 1. Enriched TF PWMs in promoters of E2F target genes.

TF	Number of promoters with hits	Number of hits	Analytical score	Rank relative to abundance in random sets
E2F	28	35	1.9×10^{-10}	1
NF-Y	44	64	1.7×10^{-14}	1
CREB	28	41	2.5×10^{-5}	1
NRF-1	32	77	3.1×10^{-4}	3

Table 2. Enriched TF PWMs in promoters of genes that function in the cell cycle.

Biological process category	Number of genes	TF	Analytical score	Rank relative to abundance in random sets
Cell cycle control (GO:000074)	223	ETF	1.5×10^{-7}	1
		E2F	1.5×10^{-6}	1
		NRF-1	2.5×10^{-5}	1
		Sp1	2.5×10^{-4}	4 (2)
Mitotic cell cycle (GO:0000278)	175	E2F	1.4×10^{-9}	1
		NF-Y	1.3×10^{-4}	1 (2)
		NRF-1	1.6×10^{-4}	1
DNA metabolism (GO:0006259)	240	E2F	6.7×10^{-5}	1
		NF-Y	4.6×10^{-4}	4 (2)
		Sp1	6.8×10^{-4}	5 (5)
M phase (GO:0000279)	100	NRF-1	5.9×10^{-6}	1
		NF-Y	2.5×10^{-4}	2 (2)
		ATF	3.4×10^{-4}	4 (5)
		E2F	3.8×10^{-4}	1

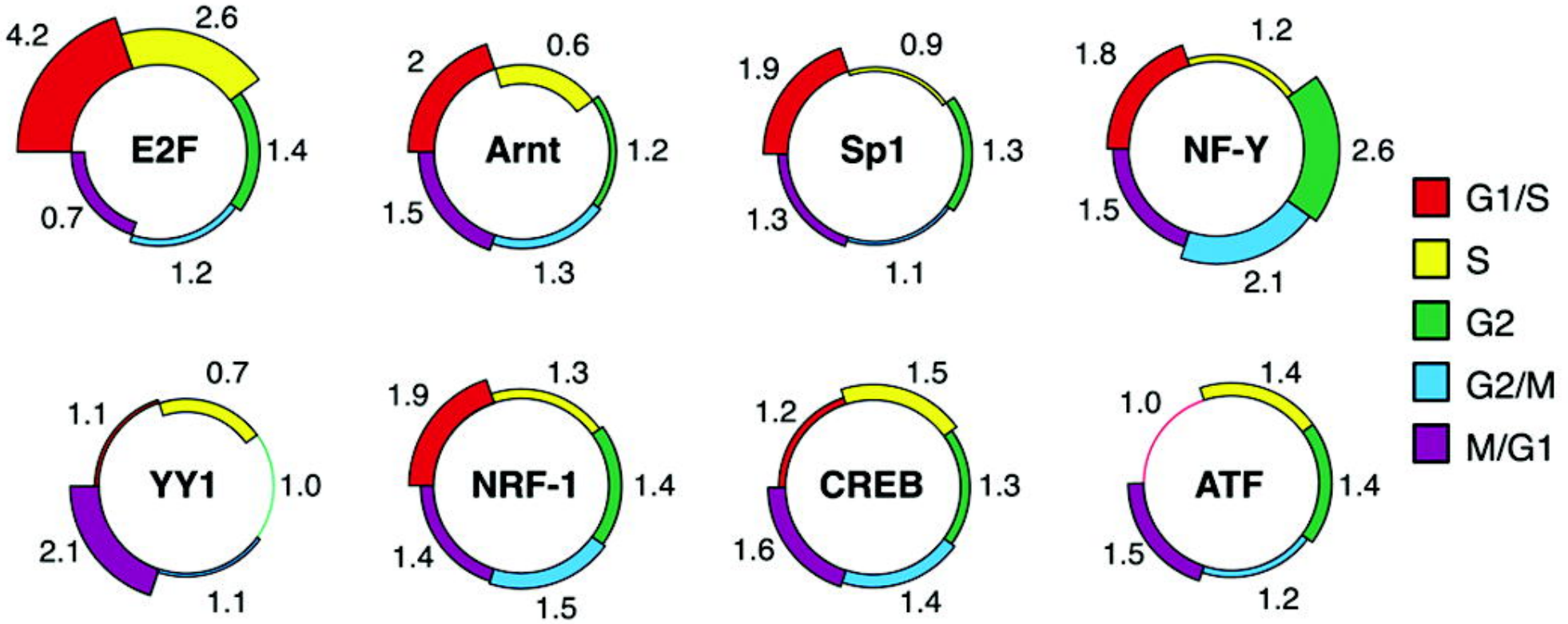
Table 3. Enriched TF PWMs in promoters of cell cycle regulated genes.

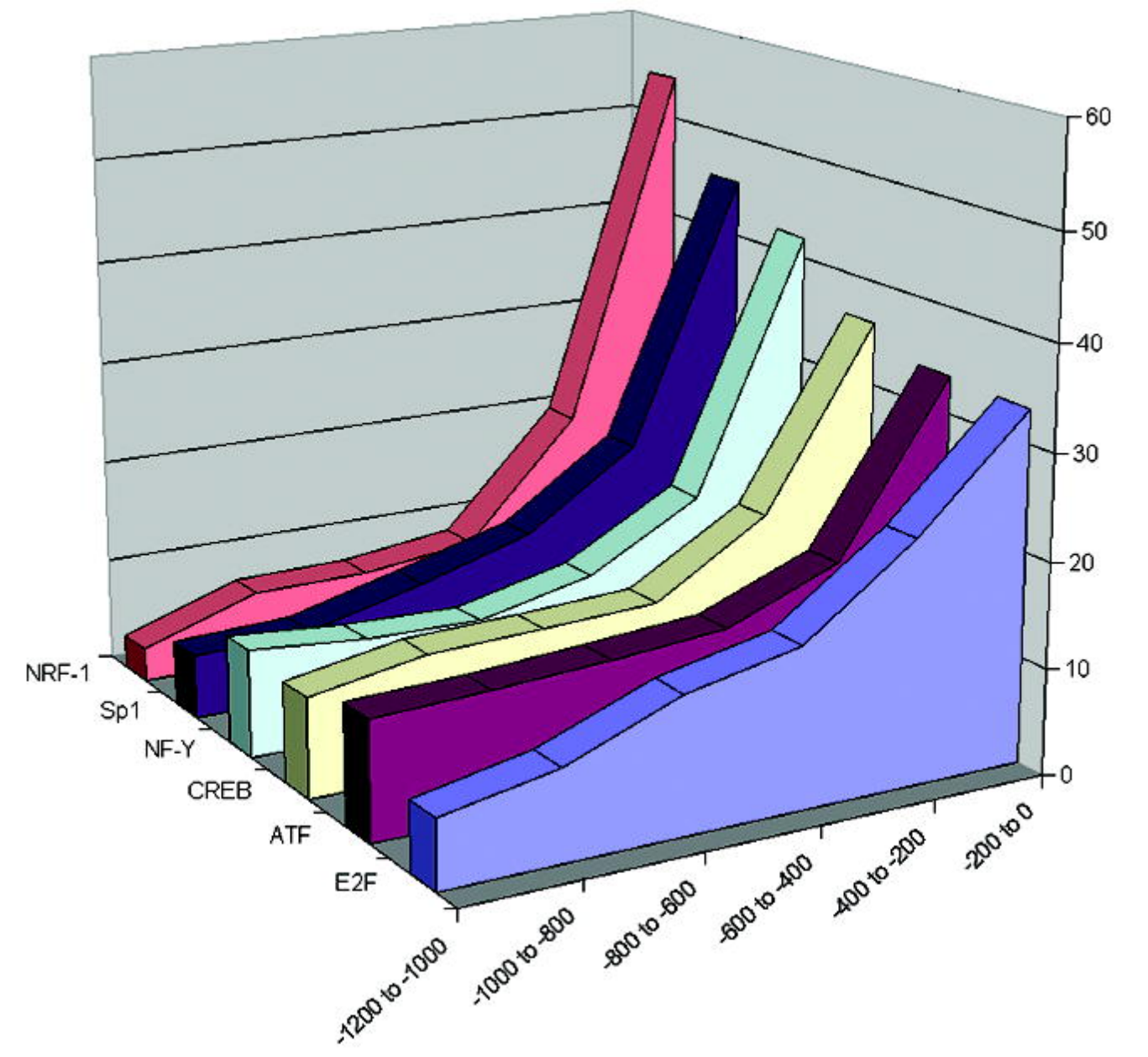
a

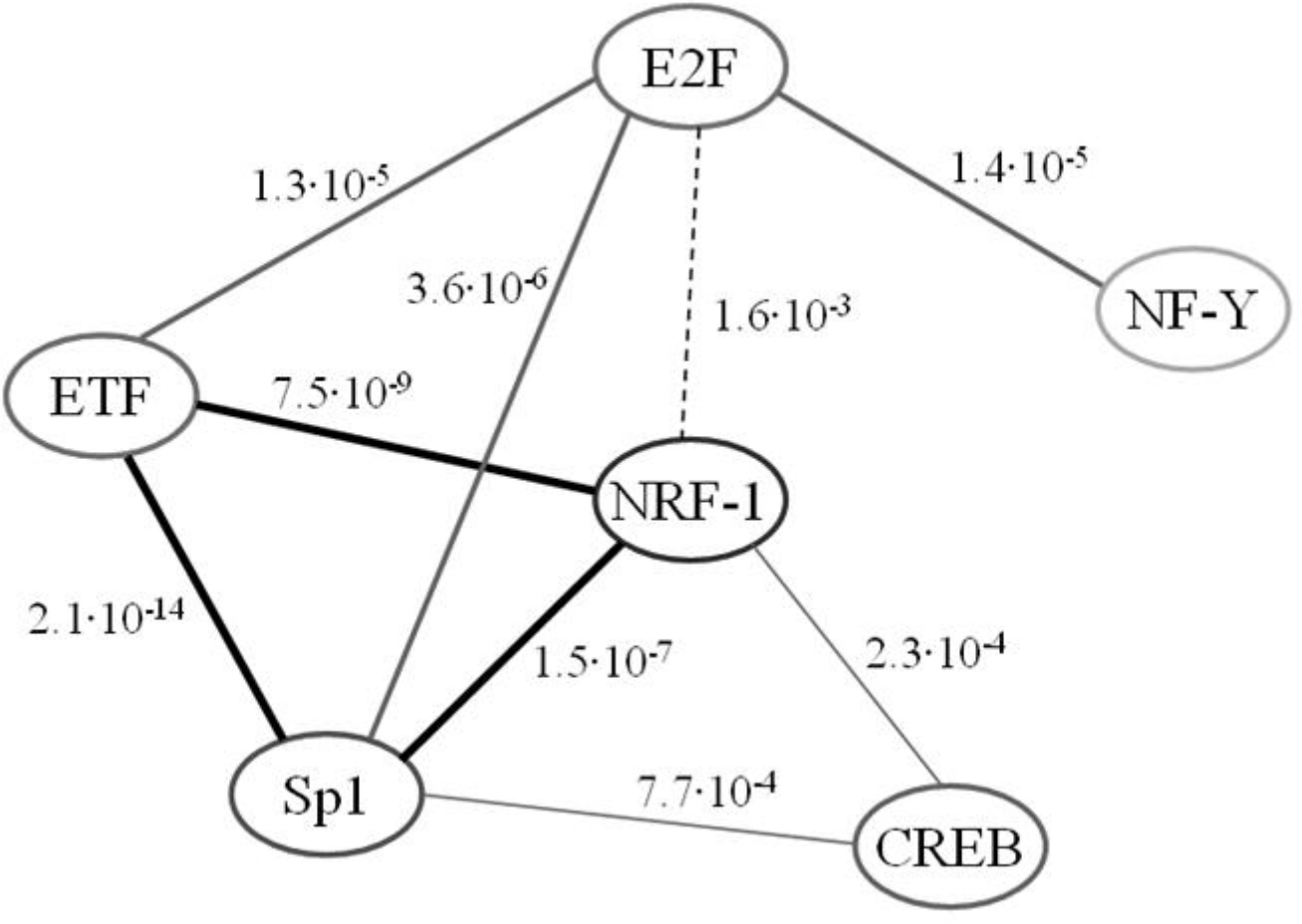
TF	Number of promoters with hits	Number of hits	Analytical score	Rank relative to abundance in random sets
NF-Y	152	203	$1.2x10^{-11}$	1
E2F	78	92	$1.2x10^{-8}$	1
NRF-1	127	234	$3.3x10^{-6}$	1
Sp1	223	365	$1.3x10^{-4}$	1
ATF	113	162	$5.3x10^{-4}$	2
CREB	91	117	$9.3x10^{-4}$	2 (1)

b

TF	Number of promoters with hits	Number of hits	Cell cycle phase	Analytical score	Rank relative to abundance in random sets
Arnt	33	37	G1/S	$5.1x10^{-4}$	5 (4)
YY1	20	25	M/G1	$8.1x10^{-4}$	5 (3)


4.2
2.6
E2F
1.4
0.7
1.2
2
0.6
Arnt
1.2
1.5
1.3
1.9
0.9
Sp1
1.3
1.3
1.1
1.8
1.2
NF-Y
2.6
1.5
2.1
0.7
1.1
YY1
1.0
2.1
1.1
1.9
1.3
NRF-1
1.4
1.4
1.5
1.2
1.5
CREB
1.3
1.6
1.4
1.0
1.4
ATF
1.4
1.5
1.2
G1/S
S
G2
G2/M
M/G1


60
50
40
30
20
10
0
NRF-1
Sp1
NF-Y
CREB
ATF
E2F
-1200 to -1000
-1000 to -800
-800 to -600
-600 to -400
-400 to -200
-200 to 0


E2F
NF-Y
ETF
NRF-1
Sp1
CREB
1.3·10-5
1.4·10-5
3.6·10-6
1.6·10-3
7.5·10-9
2.1·10-14
1.5·10-7
2.3·10-4
7.7·10-4