Published online May 20, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 9 ***2889–2900***
DOI: 10.1093/nar/gkh614

Identification of sparsely distributed clusters of *cis*-regulatory elements in sets of co-expressed genes

Gabriel Kreiman*

Center for Biological and Computational Learning, McGovern Institute for Brain Research, Massachusetts Institute of Technology, 45 Carleton Street, MIT E25-201B, Cambridge, MA 02142, USA

Received January 14, 2004; Revised March 25, 2004; Accepted April 26, 2004

ABSTRACT

Sequence information and high-throughput methods to measure gene expression levels open the door to explore transcriptional regulation using computational tools. Combinatorial regulation and sparseness of regulatory elements throughout the genome allow organisms to control the spatial and temporal patterns of gene expression. Here we study the organization of *cis*-regulatory elements in sets of co-regulated genes. We build an algorithm to search for combinations of transcription factor binding sites that are enriched in a set of potentially co-regulated genes with respect to the whole genome. No knowledge is assumed about involvement of specific sets of transcription factors. Instead, the search is exhaustively conducted over combinations of up to four binding sites obtained from databases or motif search algorithms. We evaluate the performance on random sets of genes as a negative control and on three biologically validated sets of co-regulated genes in yeasts, flies and humans. We show that we can detect DNA regions that play a role in the control of transcription. These results shed light on the structure of transcription regulatory regions in eukaryotes and can be directly applied to clusters of co-expressed genes obtained in gene expression studies. Supplementary information is available at http://www.mit.edu/~kreiman/resources/cisregul/.

INTRODUCTION

Transcriptional regulation plays a fundamental role in many biological processes ranging from development to immunity to learning and memory. Recent sequencing efforts suggest that the total number of genes does not correlate well with the behavioral complexity of an organism. This complexity may arise, partly at least, from more intricate genetic regulatory mechanisms. Two recent sources of information promise to accelerate progress in our understanding of gene expression and its regulation. First, we now have sequence information from multiple species. Second, it is now possible to interrogate the expression levels of thousands of genes simultaneously. Combining these two types of data allows us to ask which sequence elements govern the levels of mRNA molecules.

Gene expression by RNA polymerase II is orchestrated by multiple protein transcription factors (TFs) that bind specific sequences in the DNA (1–4). We refer to the DNA sequences to which these factors bind to enhance or inhibit transcription as *cis* elements. Three specific aspects of the TF–DNA interaction complicate the computational search for regulatory elements. Binding sites for a given TF are quite variable and we generally lack accurate models of the binding energy between the TF and DNA (5). Furthermore, TFs can bind DNA near the transcriptional start site (TSS), usually called the promoter region, but they can also act at a distance of tens of thousands of base pairs away from the TSS (6,7). These long-distance interactions substantially increase the noise in any procedure to search for regulatory elements. Finally, groups of TFs may cluster along the DNA to form modules responsible for specific regulatory roles and therefore the binding affinity of a particular TF may also depend on the sequence surrounding its binding site (8,9).

Several computational techniques to search for individual TF binding sites, here called 'motifs', have been proposed (10–15). Some of these techniques have achieved considerable degrees of success, particularly when applied to sequences from prokaryotic organisms or yeasts, although in some cases the false positive rates still remain high. The extrapolation of these techniques to higher eukaryotes like mammals remains difficult for several reasons. Non-coding sequences are longer in humans or mice compared to yeast. Additionally, in several paradigmatic examples, transcriptional regulation has been shown to require the combinatorial interplay of multiple factors (3,8,9,16). The binding of a single TF in general cannot account for the complex spatial and temporal regulation of gene expression in higher eukaryotes. As an example, p65 was found to bind to 209 sites on human chromosome 22 alone by ChIP-chip analysis. Furthermore, it did not affect transcription at many of those sites (17).

In some cases, the investigators know or strongly suspect that a particular group of TFs plays a role in the transcriptional regulation of the set of genes under study. Several algorithms have been proposed recently to study this scenario (18–25). Here we address a different variant of the problem where the

*Tel: +1 617 253 0547; Fax: +1 617 253 2964; Email: kreiman@mit.edu

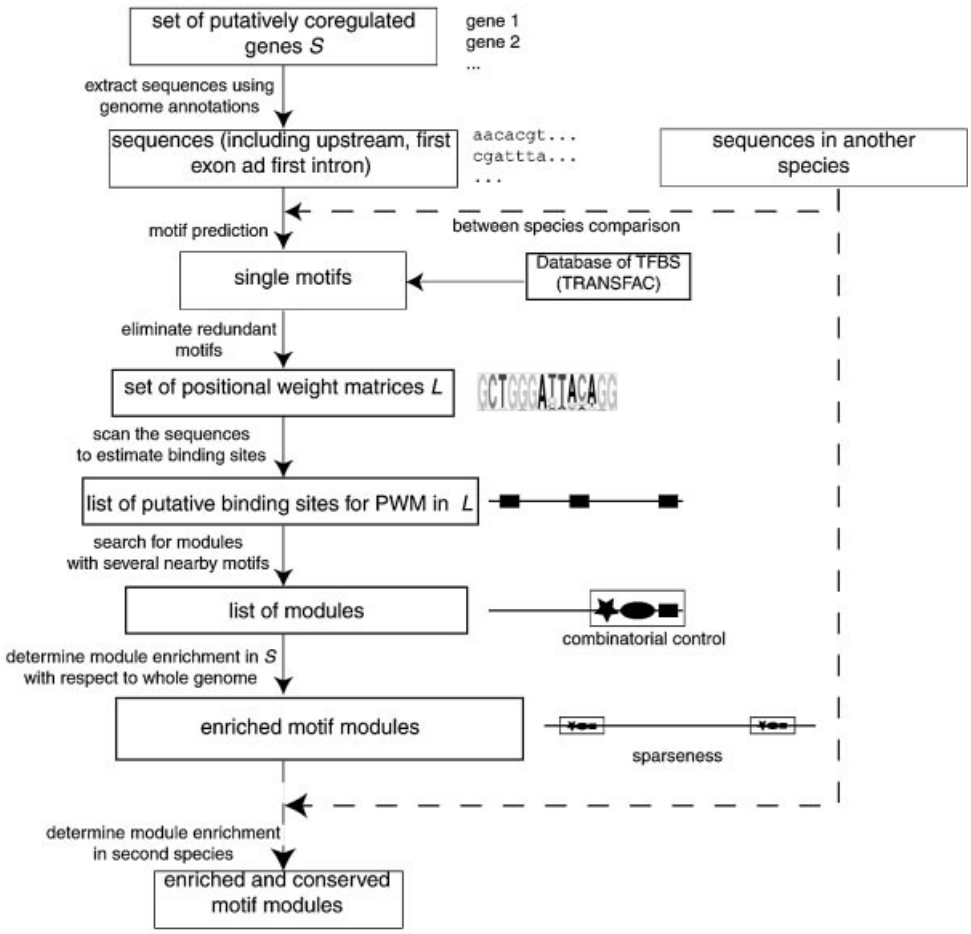

Figure 1. Overall scheme. Schematic description of our approach to find *cis* elements in eukaryotes based on combinatorial usage of transcription factors and sparseness of the regulatory modules. The approach involves searching for co-occurrences of motifs that are highly enriched in the set of potentially co-regulated genes (*S*) with respect to the set of all genes in the corresponding genome. The upstream region, first exon and first intron are retrieved for each gene in *S*. Non-conserved sequences can be masked to reduce the level of noise. A list of individual PWMs (*L*) is created by (i) searching for new motifs (using the motif finding programs alignACE, MEME and MotifSampler) on the upstream, first exon and first intron sequences of the genes in *S* and independently (ii) from motifs from the TRANSFAC database. Modules are defined by clusters of motifs within small DNA segments. Module enrichment is evaluated by comparing the occurrences of the module in the set *S* against occurrences in all the genes in the genome. The boxes indicate the output of the previous step and the arrows indicate the processs(es) involved in each step.

biologist is faced with a set of genes without any knowledge about the TFs involved in their regulation. This setting may arise, for example, in a DNA microarray experiment where one of the outputs is a cluster of genes that share a similar expression pattern.

Combinatorial regulation of transcription and sparseness of regulatory modules in the whole genome underlie the organization of *cis* elements in complex eukaryotic systems. Multiple TF binding sites are clustered together along the DNA forming modules that are required to control the expression of each gene (1,8,9). These modules of regulatory elements should occur infrequently throughout the whole genome. The rationale for this is that specificity requires a sparse code. Here we show that combinatorial regulation and sparseness can guide the search for regulatory elements in higher eukaryotes. The evolutionary conservation of important regulatory elements has been discussed and applied extensively (see for example 26–29). We therefore do not discuss comparisons across species in detail here but we incorporate conservation between species into our algorithm. We combine these three ideas into an algorithm to search for *cis* regulatory elements in sets of potentially co-regulated genes (Fig. 1). We illustrate the performance of the algorithm on random sets of genes as a negative control as well as on three separate sets of biologically validated co-regulated genes, ranging from yeast to humans. We show that we can correctly find many of the known regulatory regions in these sets of genes without any *a priori* knowledge about which TFs are involved or where to search.

MATERIALS AND METHODS

General overview

Given a set S of n_S potentially co-regulated genes, our algorithm searches for common sequence patterns that occur more frequently than expected by chance (see scheme in Fig. 1). The search is based on locating co-occurrences of

putative binding sites for TFs within short segments of DNA and ensuring that the co-occurring motifs are sparsely distributed throughout the genome. We tested the algorithm on different sequence sources including (i) artificial sequences with implanted modules, (ii) random sets of genes (negative control) and (iii) experimentally validated systems encompassing a wide range of sequence characteristics (positive controls): (iiia) the CLB2 cluster in yeasts (30), (iiib) a set of genes involved in pattern formation in flies (24) and (iiic) a set of genes co-expressed in human skeletal muscle (31). The list of genes in each set is shown in Table 2. Further characterization of the algorithm, performance details and a list of results for the different data sets are available as supplementary information at http://www.mit.edu/~kreiman/resources/cisregul. All the code is available upon request from the authors.

Sequences

Sequences were retrieved from the following sources: http://www.yeastgenome.org/ for the *Saccharomyces cerevisiae* sequences (release 01-21-2003); http://www.ncbi.nlm.nih.gov/ for the human and mice RefSeq sequences (release 06-21-2003); http://www.ensembl.org/ for the *Drosophila melanogaster* sequences (release 01-07-2003). The corresponding annotations were used to retrieve the best current estimation of the TSS for each gene. In cases of multiple alternative start sites, we used the one farthest upstream. We included in our search the upstream sequences, the first exon and the first intron. We restricted the search to the 5000 bp upstream of the TSS (see Discussion). If the first exon or intron was longer than 5000 bp they were trimmed to retain the 5000 bp closest to the TSS. In the CLB2 set in yeast the search was restricted to 1000 bp upstream of the TSS (extensive evidence suggests that this is the most important region for yeasts; 13,26,32,33). For the human and mouse genes, we restricted our analysis to the RefSeq set of genes (34). For the comparison between mouse and human genes, the orthology information was retrieved from the NCBI HomoloGene list (http://www.ncbi.nlm.nih.gov/HomoloGene/).

Motif models and scanning

We used position weight matrices (PWM) to model the binding specificity of each motif. PWMs take into account the frequency of each nucleotide (A, C, G and T) at each position (5,35). PWMs assume independence between different nucleotide positions and attempt to provide a first level approximation of the interaction energy of a TF with its binding site (5,36). We used two sources of PWMs: (i) from a database of known transcription factor binding sites, TRANSFAC public release 6.0 (37); (ii) putative novel motifs. The novel motifs were obtained by using three motif-finding algorithms: alignACE (13), MEME (12) and MotifSampler (11). The input to the motif-finding algorithms was the set of sequences for the genes in S (the boundaries for the sequences were defined in the previous section). For the human muscle set, we ran the motif-finding algorithms on both the raw sequences and the sequences after masking those segments not conserved in the mouse orthologs. Sequence conservation was determined using BLAST. The parameters for alignACE were -numcols 10 and the GC frequency. The parameters for MEME were -minw 6, -maxw 20, -dna, -mod tcm, -nmotifs 100, -evt 1, -minsites 3, -maxsites 500, 6th order background model and -revcomp. The parameters for MotifSampler were 6th order background model and -n 10. The background models and GC frequency were computed from the upstream sequences of all genes in the corresponding genome (5000 bp upstream of the TSS for mouse, human and flies and 1000 bp for yeast). The output of the motif search algorithms depends on the random initial conditions. We therefore ran 10 iterations of each motif search algorithm on the same sequences (the number of novel non-redundant PWMs reported by the motif-finding programs decreases with each successive iteration; see supplementary information.

The PWMs from the motif-finding algorithms and the motifs from TRANSFAC were merged. Redundancies (between PWMs obtained from different iterations of a motif-finding algorithm or different motif-finding algorithms or the motif-finding output and TRANSFAC) were removed by considering the similarity of the weight matrices. Similarity between two PWMs was assessed by the Spearman correlation coefficient between linearized weight matrices (in the best alignment). We used a threshold correlation coefficient of 0.70 to consider two motifs redundant (13). Furthermore, we only considered motifs with an information content (5) larger than 0.2331 bits/nt (this value corresponded to the lowest 5th percentile from the TRANSFAC database), a minimum length of 6 nt and a minimum of five sequences used to define the PWM. The resulting set of T non-redundant motifs, $L = \{m_1,\ldots,m_T\}$, was then used to search for modules (see below).

Given a PWM, we scanned all the sequences and assigned a score to each sequence segment. In the human muscle set, scanning was performed on the masked sequences (similar results but with higher levels of noise were obtained when using the raw sequences; see supplementary information). The score was given by

$$\theta = \sum_{i=1}^{i=w} \log\left(\frac{f_{ni}}{b_n}\right)$$

where w is the motif length, f_{ni} indicates the frequency of nucleotide n at position i ($n \in \{A,C,G,T\}$ and $i = 1,\ldots,w$; a pseudocount of one was added at each position to account for small sample bias; see 5) and b_n is the overall frequency of nucleotide n (5). Binding of a TF to DNA is likely to be a continuum whereby the protein spends more time bound to higher affinity sequences. However, for the present implementation we chose a binary threshold that classified each PWM as present or absent at each position. As a threshold, we used the maximum score θ_{th} that left out <5% of the sites used to build the PWM (attempting to achieve a low false negative rate). The search was conducted on both strands.

Preliminary search for modules

For faster performance, we first compared the set of potentially co-regulated genes against a small background set of random genes before comparing against all genes in the genome (this reduces the number of modules to analyze in all genes in the genome by several orders of magnitude). For a set S with n_S potentially co-regulated genes, a background set of genes was generated by extracting a random set of n_b genes where $n_b = 20n_s$. We exhaustively explored all combinations

of motifs up to n_{motifs} (we used n_{motifs} = 2, 3 or 4) from the set of non-redundant motifs L. A preliminary module M was defined as a set of motifs $\{m_1, \ldots, m_n\}$, with $n < n_{motifs}$, $m_i \in L$, that fulfilled the following requirements. (i) The distance between adjacent motif occurrences was less than max_d. We explored max_d = 25, 50, 100 and 200 bp; these values are within the range of distances between binding sites observed in several experimentally validated studies in multiple species (3,24,31,38). (ii) The maximum overlap between adjacent motifs was half the motif length. (iii) The module had to be present in at least n_{tr} genes in S (n_{tr} = 4). (iv) The module had to be enriched in S with respect to the background set at $p < 0.01$ after Bonferroni correction by the total number of combinations (see definition of enrichment below). The motifs within a module were not required to be different and therefore this allowed the modules to represent homotypic interactions as well.

Comparison to all genes

For each preliminary module M, let x be the number of genes within S where the module was present ($n_{tr} \leqslant x \leqslant n_s$). We determined the frequency of occurrence of M in all genes as $P_g = e_g/n_g$, where e_g is the number of genes containing M in the set of n_g genes analyzed. For the enrichment with respect to 'all genes', n_g = 16 969 for mice, 17 689 for humans, 13 639 for flies and 6327 for yeast (the exact number of genes in each species is still not settled, particularly for humans and mouse, but we refer to this number as 'all genes in the genome' throughout the text). We assumed as a null hypothesis that M was randomly distributed across all genes. We therefore defined the enrichment as the probability that the number of genes in S where M is present, e_s, is larger than or equal to the observed value x assuming the frequency P_g in all genes. This follows a hypergeometric distribution (sampling without replacement from a finite population) which converges to the binomial distribution when n_s/n_g is small (39). The probability of enrichment can be expressed as:

$$P(e_s \geq x) = \sum_{i=x}^{i=n_s} \binom{n_s}{i} P_g^i (1 - P_g)^{n_s - i}$$

This enrichment probability was computed for all the modules M from the preliminary module search step. We report all modules with enrichment probability <0.01. Given that multiple hypotheses are tested, we applied the Bonferroni correction using the total number of hypotheses (40). For the comparison to all genes, the total number of hypotheses was given by the number of preliminary modules. For the definition of the preliminary modules, the total number of hypotheses corresponded to the total number of motif combinations.

Performance evaluation

We examined two main values to evaluate the performance of our algorithm. First we considered whether we could detect the known regulatory regions. We defined p_{known} as the proportion of known regions that were detected. Secondly, we examined the rate of false positives. It is not easy to accurately determine the false positive rate without making strong assumptions about the biology of the system under study. Here we assume that none of the hitherto uncharacterized regions play a biological role. The assumption that we know all the regulatory regions leads to an upper bound on the false positive rate. We define p_{FA}, the probability of false alarm, as the proportion of module predictions where the location of M overlaps with known regulatory regions in <50% of the genes. In other words, if an investigator were to conduct a follow-up experiment to study the putative regulatory regions predicted by each of the modules, then the probability of false alarms indicates the proportion of modules where more than half of the genes would not show any biological regulatory function. The 50% cut-off is arbitrary and Figure 4 and the supplementary information show the probability of false alarm for different values of this threshold. We separately report the false positive rate and p_{known} for all the predictions as well as for the top 10 predictions. We do not discuss here the computational performance of the algorithm; all the code was run on a Pentium IV, 2.8 GHz computer running Linux.

RESULTS

We incorporated the principles of combinatorial regulation and sparseness into an algorithm to search for *cis*-regulatory elements in a set of potentially co-regulated genes. We also added the power of evolutionary conservation to detect non-coding sequences from multiple species that show little variation through time. A schematic layout of the algorithm is shown in Figure 1. We tested the algorithm to search for putative regulatory sequences in both positive and negative control sets. The first positive control consisted of random sequences where clusters of motifs were artificially implanted. This served the purpose of calibrating the parameters of the algorithm and studying its performance for different degrees of degeneracy of the motifs, noise levels and distance constraints (data not shown). As a negative control, we selected random sets of genes and carried out the same analysis as with the other sets. This showed that the number of predictions expected by chance was small. We then studied the CLB2 gene set in yeast. This is a well-known case where the transcription of several genes in the set is regulated by the TFs SFF and MCM (30,41,42). Many characteristics of the yeast genome make searching for transcriptional regulatory signals easier than in higher eukaryotic organisms. Therefore, we next tested the algorithm in a set of genes involved in pattern formation in flies (24), as well as in a set of genes co-expressed in skeletal muscle in humans (31).

Random sets of genes

Given the large number of combinations of motifs that had to be tested, it was important to put a bound on the probability of chance occurrence of putative modules. Therefore, we sought to determine whether it is possible to obtain putative modules that appear to be statistically significant in negative controls consisting of a random set of genes. The underlying assumption is that a random set of genes is unlikely to share a specific set of common regulatory elements. We randomly selected groups of N genes (N = 20, 30 or 40) from the mouse RefSeq collection and analyzed those genes as if they constituted a real set of potentially co-regulated genes by applying our algorithm and searching for regulatory modules. For each value of N, the procedure was repeated five times. The average

Table 1. Summary of performance for random gene sets and skeletal muscle set

n_{motifs}	30 Random mouse genes			Skeletal muscle		
	Modules	p_{mm}	p_{hs}	Modules	p_{mm}	p_{hs}
2	0.00			98.3	3.7E – 04	5.4E – 04
3	0.00			291.3	2.3E – 04	3.4E – 04
4	0.35	3.6E – 03	0.37	561.5	1.7E – 04	2.5E – 04

The number of modules and module enrichment for the negative controls consisting of 30 random genes from the mouse RefSeq collection (average of five iterations) and for the skeletal muscle gene set. n_{motifs} indicates the maximum number of motifs per module. The number of modules indicates the average over all iterations and the four values of max_d explored (25, 50, 100 and 200 bp). The p value shows the enrichment probability with respect to the whole mouse (mm) or human genome (hs) (median across different values of max_d and different iterations, see Materials and Methods). The results presented in this table were computed using a maximum upstream sequence length of 5000 bp and including the first exon and first intron (total sequence length: random genes, 7961 ± 419 bp; skeletal muscle, 6823 ± 1700 bp). The results for each iteration and each value of max_d as well as a description of the dependence on the sequence length and the number of genes are available at http://www.mit.edu/~kreiman/resources/cisregul/.

number of PWMs was 372 ± 29 (including 144 TRANSFAC mouse and human motifs plus the output of the motif finding programs; see Materials and Methods). The average sequence length was 7961 ± 419 bp. Table 1 summarizes the results of this analysis for the set of parameters that most closely matches the study of the human skeletal muscle set (results for other parameters are shown in supplementary information). Random sets of genes yield only a small number of module predictions compared to the positive controls. For a maximum of 4 motifs per module, the average number of modules from the random sets of genes was 0.4 ± 0.5 whereas the number of modules for the human skeletal muscle set was 561. This suggests that only a small fraction of the modules found in a set of genes can be explained by random co-occurrences of PWM hits. In addition to the low number, the quality of the modules obtained from random sets of genes, as assessed by the enrichment criterion, was poorer than the quality of those modules obtained from real sets of co-regulated genes (median p values: random sets of genes = 0.004 ± 0.005; skeletal muscle set = $7 \times 10^{-4} \pm 1.6 \times 10^{-4}$). Furthermore, the putative modules showed no enrichment in the human genome and therefore no evidence of evolutionary conservation of the *cis* elements. The median p value in the human genome for enrichment of the modules found in the random sets of mouse genes was 0.4 ± 0.3, whereas the median p value for enrichment in the mouse genome of the modules found in the human skeletal muscle set was $2.5 \times 10^{-4} \pm 9.5 \times 10^{-3}$. These observations show that, in spite of the large number of combinations, our analysis could still reveal interesting regulatory elements beyond chance expectations. The number of modules found for each iteration as well as results for other parameters are shown in supplementary information.

Initial exploration: a set of genes co-expressed in the yeast cell cycle

We studied a set of 32 genes (Table 2) comprising the CLB2 cluster in yeast. These genes show a pattern of expression that peaks in the M phase of the cell cycle (30). The transcription of several of these genes is known to be controlled by two TFs, SFF and MCM (38,41,42). We searched for combinations of two motifs from a list of 147 motifs that included those weight matrices in TRANSFAC (37) as well as the output from the alignACE motif search program (13) (the full list of motifs is available at http://www.mit.edu/~kreiman/resources/cisregul).

The top scoring module corresponded to the interaction between two motifs resembling the SFF and MCM binding sites. The SFF-like and the MCM-like motifs were found by running alignACE on the 1000 bp promoter region of the 32 genes (the output of 10 runs of alignACE was a total of 286 motifs; 116 motifs remained after removing redundant motifs). The correlation coefficient between the linearized PWM of our MCM-like motif and the MCM motif reported in the literature (30,43) was 0.54 and the correlation coefficient for the SFF-like motif was 0.94. These two motifs co-occurred 12 times in 12 genes from the CLB2 cluster (see supplementary information). Among these genes were CLB1, BUD4, SWI5, Swi5p and Ace2p, which are known to be transcriptionally controlled by MCM and SFF (30). This pair of motifs was also computationally identified by Pilpel and colleagues by considering the change in expression coherence during the M phase of the cell cycle for genes containing binding sites for both factors compared to either factor alone (43). The first motif alone occurred 57 times in 25 genes while the second motif appeared 39 times in 23 genes from the CLB2 cluster set. This emphasizes the power of combinatorial regulation by searching for co-occurrences of the two motifs. When scanning through all genes in the *S.cerevisiae* genome, the two motifs co-occurred 62 times (<1% of the total of 6327 genes searched) while either motif alone was present in >1000 genes. This illustrates the power of the sparseness principle, i.e. the enrichment in the set of co-regulated genes with respect to all genes in the genome. This almost 40-fold occurrence ratio in the CLB2 cluster compared to all genes corresponds to an enrichment p value $<10^{-17}$ (see Materials and Methods). This two-motif module was the top scoring module regardless of the value used for the maximum distance parameter (25, 50, 100 or 200 bp). Furthermore, this was also the top scoring module when allowing combinations of up to three motifs, indicating that another motif did not add to the specificity of this regulatory module.

We analyzed the gene expression levels along the cell cycle of all the genes containing co-occurrences of these two motifs within 25 bp using the microarray cell cycle data of Spellman and colleagues (30). Of the 50 genes where the module was present (beyond the 12 genes already present in the CLB2 set), cell cycle expression data were unavailable for 4 genes, 29 genes had no apparent modulation of gene expression, 3 had strong cyclic modulation with a peak in G_1 or S phase (YFL037W, YLR194C and YNR009W), 5 genes showed

Table 2. List of genes in each set

Gene identifier	Symbol	Elements in searched region
Saccharomyces cerevisiae CLB2 set		
YLR131C	ACE2	Yes
YGL021W	ALK1	Yes
YNL172W	APC1	Yes
YCL014W	BUD3	Yes
YJR092W	BUD4	Yes
YLR353W	BUD8	Yes
YGL116W	CDC20	Yes
YMR001C	CDC5	Yes
YBR038W	CHS2	Yes
YGR108W	CLB1	Yes
YPR119W	CLB2	Yes
YOR025W	HST3	Yes
YPL242C	IQG1	Yes
YPL155C	KIP2	Yes
YIL106W	MOB1	Yes
YHR023W	MYO1	Yes
YDR150W	NUM1	Yes
YDR146C	SWI5	Yes
YML064C	TEM1	Yes
YCL063W	VAC17	Yes
YIL158W	YIL158W	Yes
YJL051W	YJL051W	Yes
YKL130C	SHE2	Yes
YLR057W	YLR057W	Yes
YLR084C	RAX2	Yes
YLR190W	MMR1	Yes
YML034W	SRC1	Yes
YML119W	YML119W	Yes
YMR032W	HOF1	Yes
YNL058C	YNL058C	Yes
YPL141C	YPL141C	Yes
YPR156C	TPO3	Yes
Drosophila melanogaster pattern formation set		
CG9786	hb	Yes
CG4717	kni	Yes
CG3340	kr	Yes
CG2328	eve	Yes
CG6494	h	Yes
CG1849	run	Yes
CG10325	abd-A	No
CG6464	salm	No
CG10388	ubx	No
CG7952	gt	Yes
CG3851	odd	Yes
CG6246	nub	Yes
CG12287	pdm2	Yes
Homo sapiens skeletal muscle set		
1140	CHRNB1[a]	Yes
1146	CHRNG[a]	Yes
1144	CHRND[a]	Yes
1145	CHRNE[a]	Yes
70	ACTC[a]	Yes
1158	CKM[a]	Yes
1674	DES[a]	Yes
6517	SLC2A4[a]	Yes
4656	MYOG[a]	Yes
4632	MYL1[a]	Yes
4635	MYL4[a]	Yes
7134	TNNC1[a]	Yes
7135	TNNI1[a]	Yes
7139	TNNT2[a]	No
4625	MYH7[a]	Yes
4624	MYH6[a]	Yes
58	ACTA1[a]	Yes
1410	CRYAB[a]	Yes
1339	COX6A2[a]	Yes
4634	MYL3[a]	Yes
4633	MYL2[a]	Yes
4151	MB[a]	Yes
5224	PGAM2[a]	Yes
5925	RB1[a]	No
6876	TAGLN[a]	Yes
226	ALDOA[a]	No
4878	NPPA[a]	Yes
1756	DMD[a]	No
2027	EN3[a]	Yes

For each of the three sets of biologically validated co-regulated genes that we analyzed, this list indicates the gene identifiers and symbols and whether the regulatory elements reported in the literature fall within the search areas that we included in the analysis.
[a]Locuslink.

weak modulation with a peak in M phase (YHL024W, YHL042W, YNL042W, YDR121W and YDR208W) and 9 genes showed expression modulation with a peak in M phase that resembled the expression patterns of other CLB2 genes (YAR018C, YGL008C, YHL028W, YJL157C, YKL043W, YML052W, YMR031C, YNL056W and YOR023C). The list of all genes where the module was found is available at http://www.mit.edu/~kreiman/resources/cisregul. These observations suggest that scanning for occurrences of regulatory modules may reveal genes that share similar expression patterns but were not detected in the gene expression analysis.

A more complex scenario: *Drosophila* pattern development

Searching for *cis*-regulatory elements in higher eukaryotes poses additional difficulties: intergenic regions are longer, gene structure includes longer introns and gene regulation seems to be more complex, requiring the interplay of more factors. Therefore, we next tested our algorithm in the search for *cis*-regulatory elements in 13 genes involved in development of the anterior–posterior axis in the *D.melanogaster* embryo (Table 2). These genes were taken from a previous study that combined computational and experimental work to study clustering of five known TFs, namely Bicoid, Caudal, Hunchback, Krüppel and Knirps (24). Here, we have assumed no knowledge about the involvement of these specific TFs. We ran our algorithm to study the potential regulatory mechanisms of these 13 genes by studying combinations from a list of 271 motifs (including 30 *Drosophila* PWMs from TRANSFAC and 241 PWMs obtained from the motif-finding algorithms after removing redundancies from an initial list of 1336 motifs; the full list of motifs is available at http://www.mit.edu/~kreiman/resources/cisregul).

Figure 2 shows an example where one of the top modules found by our algorithm located the known regulatory regions from six genes in the pattern formation set. This module constituted the best prediction for some but not all parameter combinations. For most of the parameters we explored (83%), this module occurred within the top 10 predictions. It was the top prediction, with an enrichment probability of 1.4×10^{-12}, for a maximum motif distance of 200 bp using a maximum of either three or four motifs. For this parameter combination, the

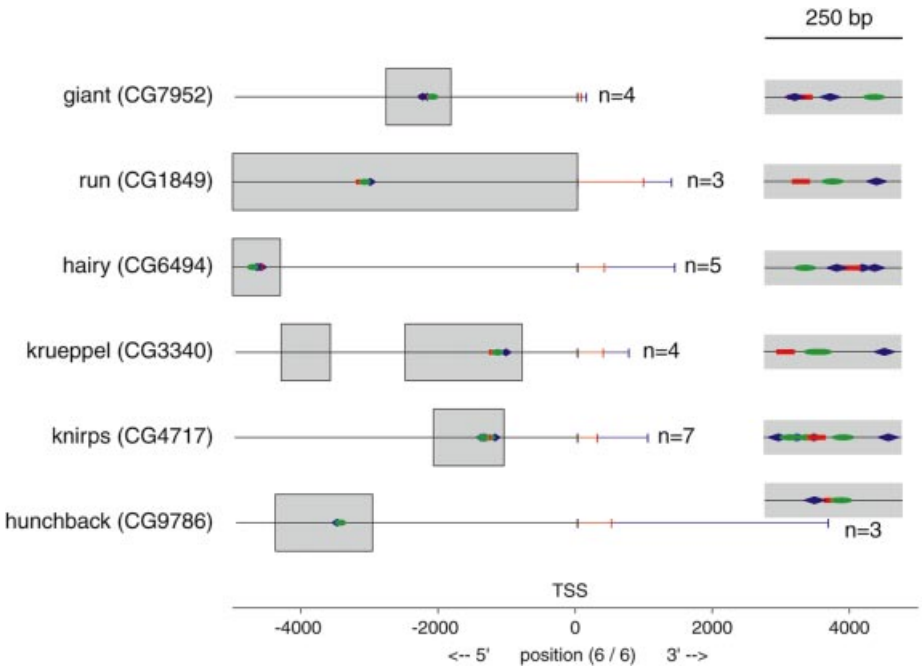

Figure 2. One of the top modules in the fly pattern formation set. Location of the three motifs from one of the top scoring modules found in six genes from the fly pattern formation set (Table 2) within the 5 kb region upstream of the TSS plus first exon and first intron. Shaded boxes correspond to the known regulatory regions as reported by Berman *et al*. (24). The number of total motif occurrences (n) is indicated next to each gene. Whenever $n > 3$, there were multiple occurrences of at least one of the motifs. To the right of each gene, we zoom in (8×) on the region including the module. The algorithm was run with the following parameters: maximum distance between motifs = 200 bp, maximum number of motifs = 3, no order constraint, minimum number of genes with module = 4. The enrichment p value (see Materials and Methods) was 10^{-12}.

top 10 predictions yielded a false positive rate of 60% (see Materials and Methods for definition of false positive rate) and 69% of the known regulatory modules were detected (out of a maximum of 77% given that the regulatory regions of three genes fell outside our search area; see Table 2). Using all module predictions instead of the top 10 predictions, the false positive rate increased to 80%, but we could detect all the known regulatory regions within the sequence search boundaries (see supplementary information available at http://www.mit.edu/~kreiman/resources/cisregul for other parameter combinations). Furthermore, considering the top 10 predictions for each parameter set, we detected at least one of the known regulatory regions in 92% of the cases. In many cases, there were multiple occurrences of some of the motifs within the modules (for example, the total number of motif occurrences was seven in the knirps gene for the module illustrated in Fig. 2). This is typical of many DNA signals and may improve the probability of DNA-binding proteins detecting their target sites and exerting their functions (2,24).

The three individual motifs in the module illustrated in Figure 2 occurred 30 times in 12 genes, 50 times in 13 genes and 97 times in 13 genes, respectively. This was reduced to the 26 occurrences in 6 genes illustrated in Figure 2 upon applying the constraint that the motifs had to cluster along the DNA. Examination of all 13 639 genes in the fly genome led to >2000 occurrences for each of the individual motifs. However, this module was present in only 48 genes (<0.5% of the total number of genes; see supplementary information available at http://www.mit.edu/~kreiman/resources/cisregul for the list of these 48 genes).

Human muscle regulatory regions

We further tested the algorithm to search for *cis* elements in complex regulatory systems by studying a set of genes expressed in skeletal muscle in humans (31). Several TFs have been shown to play a role in the regulation of gene expression in skeletal muscle, including Myf, Mef-2, SRF, Tef and Sp-1 (31). We applied our algorithm to a set of 29 genes with skeletal muscle expression (Table 2) without assuming any knowledge of the specific factors that regulate these genes. We performed our *de novo* search for *cis* elements by considering combinations from a list of 406 motifs (including 144 *Mus musculus* and *Homo sapiens* PWMs from TRANSFAC and 262 PWMs obtained from the motif-finding programs after removing redundancies from an initial set of 4578 motifs; the full list of all motifs is available at http://www.mit.edu/~kreiman/resources/cisregul).

An example of the results obtained is shown in Figure 3. This module, formed by SP1, SRF, TEF and another putative motif, locates the known regulatory regions in four muscle genes and showed an enrichment probability of 2×10^{-9} ($max_d = 100$ bp, $n_{motifs} = 4$). We searched for the occurrences of this module in the mouse genome (using exactly the same PWMs). We observed that it was also enriched within the upstream regions of the mouse orthologs of the human skeletal muscle genes of Table 2. The enrichment p value in mouse with respect to the whole mouse Refseq collection was 8×10^{-6}. For this combination of max_d and n_{motifs} parameters, 72% of the known signals were detected (four genes in the set did not have any annotated regulatory elements; see Table 2) and the false alarm rate was 80% (Fig. 4; see supplementary information).

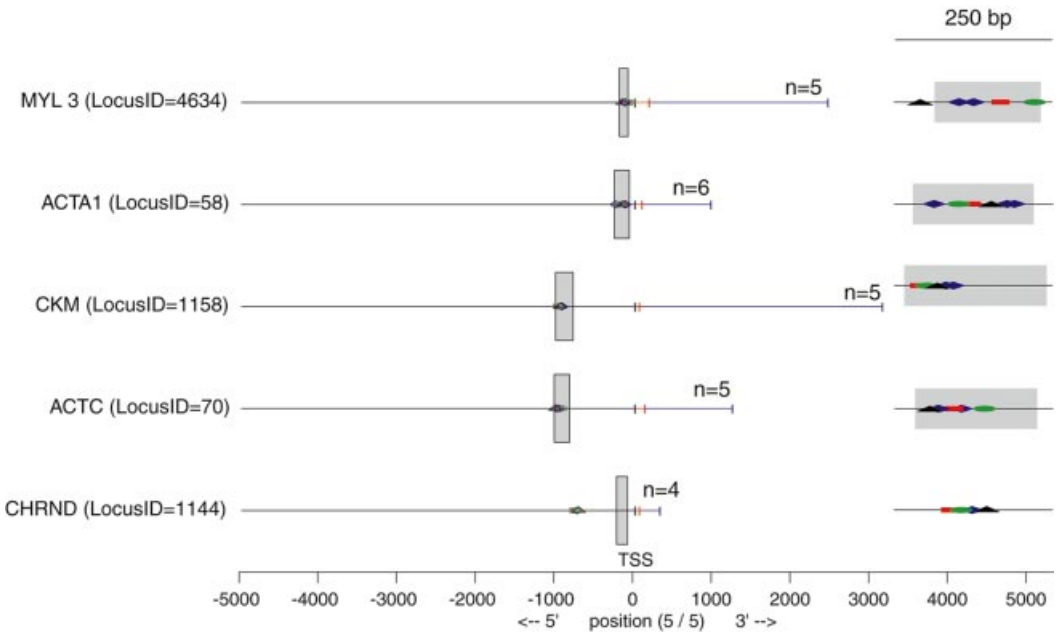

Figure 3. One of the top results in the human skeletal muscle set. Location of four motifs (red rectangle, SP1-like; blue diamond, SRF-like; green oval, TEF-like; black triangle, putative motif) from one of the top scoring modules found in five genes from the human muscle set. Shaded boxes correspond to the known regulatory regions as reported by Wasserman and Fickett (31). The format follows that in the previous figure. The algorithm was run with the following parameters: max_d = 100 bp, maximum number of motifs = 4, no order constraint, minimum number of genes with module = 4. The enrichment p value was 2×10^{-9}.

We detected this module in 175 genes (<1% of the total of 17 689 genes in the human RefSeq collection). Whether this module can control the expression of any of these other genes or not requires further experimentation. As a coarse preliminary exploration, we analyzed the expression patterns of these genes in two independent DNA microarray studies of gene expression across tissues (44–46). The list of tissues included the human and mice skeletal muscle. Expression values for some of the genes were not available (19 and 38%, respectively). Also, DNA microarray data can yield false negatives; for example, the CHRND gene, which was present in the skeletal muscle set, did not show enrichment in either microarray study. This emphasizes the importance of other complementary experimental techniques such as *in situ* hybridization to study gene expression (47). However, on average, the expression level in skeletal muscle of the genes with this module was larger than that of all genes. The ratio of the Affymetrix mean expression levels of genes containing this module to that of all genes was 4.0 for humans and 4.6 for mice ($p < 10^{-3}$ for humans and $p < 10^{-5}$ for mice, t-test). Furthermore, several genes that were not included in the human muscle skeletal set showed higher expression levels in muscle compared to other tissues in both humans and mice (see supplementary information).

Parameter landscape

Throughout the algorithm, there are several parameters requiring choices by the user. We discuss those parameters here and we show that the search results were robust to most (but not all) of these arbitrary choices.

One of the most sensitive decisions concerns where to search for regulatory elements. Shorter sequences reduce the amount of noise but the chances of missing true sites are significantly increased. In the CLB2 gene set in yeast, we always used a fixed window of 1000 bp. In the fly, mouse and human studies, we chose a compromise of using 5000 bp upstream of the TSS. Shorter regions (of 1000 or 2000 bp) missed many important regulatory elements in the fly gene set. In the analysis in flies, known regulatory regions were present between 5 and 10 kb upstream of the TSS in two genes and beyond 10 kb upstream in four genes (24). Furthermore, two genes had regulatory modules downstream of the first intron. Therefore, our search area encompassed 52% of the regulatory modules and 77% of the genes in this set. Restricting the search to only the first 2 kb upstream region would miss 74% of the known regulatory modules in the fly gene set. Our algorithm was also able to detect the known regulatory regions upon extending the upstream segment to 10 kb, albeit with a considerable increase in computation time (the number of putative motif binding sites is proportional to the total sequence length and the computation time is polynomial on the number of binding sites). The performance of the algorithm was quite poor with upstream regions of 50 kb (not shown). However, once a particular module (or a small number of modules) is found and accurately defined, it is possible to search for occurrences of the module throughout the whole genome (20,24,48). Restricting the search to the upstream sequences (i.e. ignoring the first intron and first exon) would not have missed any other regulatory elements in the example in flies. However, in the human muscle set, three genes (10%) had regulatory regions in the first exon or intron.

Only a small fraction of the putative binding sites found by the motif-finding algorithms (alignACE, MEME and MotifSampler) were incorporated into modules that occurred

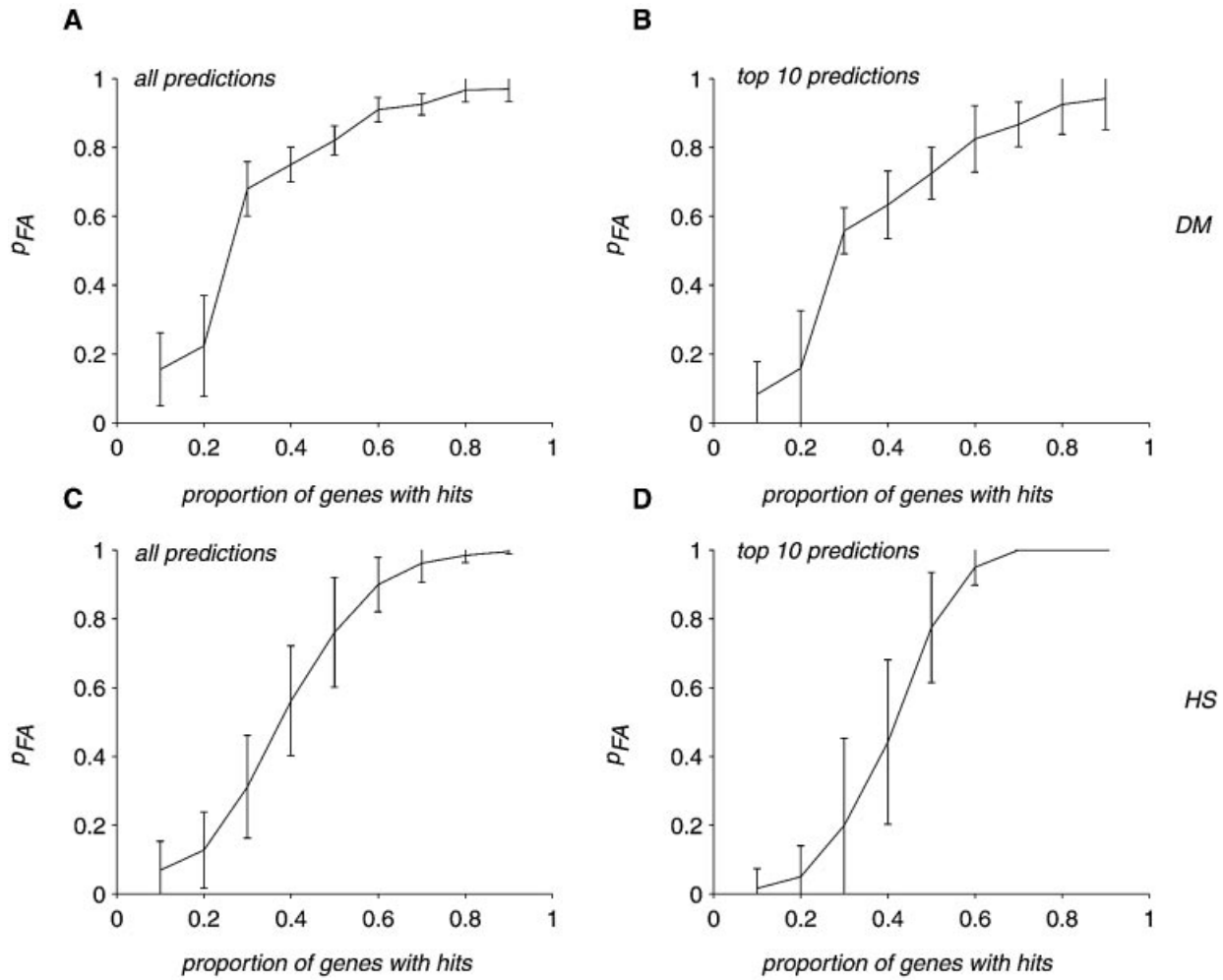

Figure 4. Average false positive probabilities. False positive probability, p_{FA}, as a function of the threshold for the proportion of genes where the module correctly predicts the known regulatory regions (see Materials and Methods for definitions). (**A**) *Drosophila* pattern formation set, all predictions. (**B**) *Drosophila* pattern formation set, top 10 predictions. (**C**) Human skeletal muscle set, all predictions. (**D**) Human skeletal muscle set, top 10 predictions. Error bars correspond to standard deviations. The values here were averaged over all combinations of max_d and n_{motifs} (the individual values for each parameter combination are available at http://www.mit.edu/~kreiman/resources/cisregul/).

within the known regulatory regions (2.5% for the human skeletal muscle gene set, 4% for the fly pattern formation set and 8% for the yeast CLB2 cluster). It is unclear whether any of the remaining motifs play any biological role or not. This small fraction suggests that further research is necessary to improve the detection of novel putative individual binding sites in complex regulatory systems with long sequences. The threshold for the comparison of motifs to eliminate redundancies (see Materials and Methods) and the parameters used to filter out some of the poorly defined PWMs did not affect the results of the analysis.

Detecting the binding sites of a motif given its PWM requires a threshold parameter. The trade-off between sensitivity and specificity for the detection of binding sites has been discussed previously (see for example 49). We observed that we could use a low threshold in the motif scanning step (increasing the false positive rate but reducing the number of missed binding sites) because of the subsequent filtering steps imposed by motif clustering to form modules. We used the maximum score value that left out <5% of the sequences used to define the PWM.

The definition of the modules also required several parameters. These included the maximum number of interacting motifs, the maximum and minimum distance between motifs and the statistical threshold. Increasing the number of interacting motifs beyond two did not improve the results for the yeast case. In contrast, for the human and fly gene sets, there was a significant increase in the proportion of known regions detected with three or four motifs. In the yeast case, the maximum distance parameter did not influence the results (from 25 to 200 bp). For the fly and human cases, better performances were obtained for max_d = 100 or 200 bp. The constraint that the binding sites cluster in a short region of the DNA is very important (9,20,24). Requiring only that the motif combination is present in the set of genes, regardless of the distances between motifs, led to poor results. For example, running the algorithm on the human skeletal muscle set without imposing any distance constraint between motifs, for

the case $n_{motifs} = 2$, yielded 9 putative modules (versus 49 with $max_d = 100$), the top scoring module had an enrichment p value of 4×10^{-5} (versus 6×10^{-8} with $max_d = 100$) and showed no conservation between humans and mouse (enrichment p value in mouse = 0.1366 in contrast to the $max_d = 100$ case where the top module had a p value in mouse = 6×10^{-4}).

DISCUSSION

Transcriptional control constitutes one of the most ubiquitous regulatory mechanisms; it is present in every species examined so far and is part of many important biological processes. Here we have shown that two simple principles about the organization of regulatory regions in the DNA could lead to the identification of *cis* elements in a group of co-regulated genes. These principles are: (i) the combinatorial nature of transcriptional regulation; (ii) the sparseness of the regulatory modules. The combinatorial arrangement of multiple TFs allows cells to finely control gene transcription and integrate multiple signal transduction pathways (1,3,8,16,50). Thus, a small number of TFs can control a large array of biological processes. Sparseness permits the cell to manipulate the expression of different genes in a distinct way.

The exploration of the parameter landscape suggests that the algorithm's performance is robust to many of the parameters and thresholds. Performance is very sensitive to the choice of the search space. Without additional knowledge about the location of the regulatory modules, short sequence segments (e.g. <2 kb for higher eukaryotes) are likely to miss many of the important regulatory elements. However, very long sequences (e.g. >10 kb) significantly increase the levels of noise. Our compromise was to use 5000 bp upstream of the TSS and to include the first exon and first intron. Once a module is found in this sequence segment, the algorithm searches for occurrences of the module in all genes in the genome. The enrichment criterion assumes as a null hypothesis that the scanned sequences of genes in the set under study are not distinct from those in all genes. Differences that may or may not play a direct role in specific regulation of transcription, such as different content of CpG islands in tissue-specific genes (51), would also be detected by our algorithm.

In addition to improving our understanding of transcriptional regulation, an accurate model of the *cis*-regulatory elements in a set of genes can lead to the identification of other genes that share the same regulatory mechanisms. A typical case involves DNA microarray studies where some genes may be missed due to thresholds in the analysis or to low expression values that are below the detection limits. For example, the module shown in Figure 2 was also present in several TFs in the fly genome, including tll, can and fkh (see supplementary information for the whole list). The tll TF is known to play a role in controlling the expression of several of the genes in the fly pattern formation gene set (24). In the human skeletal muscle gene set, the module shown in Figure 3 was present in MAPK12 (mitogen-activated protein kinase 12), which is enriched in human skeletal muscle compared to the median expression across all tissues in two independent microarray data sets (44,45). Furthermore, MAPK12 is also enriched in mouse skeletal muscle (44). This module was also present in other genes highly expressed in skeletal muscle, including MAPKAPK2, MAPKAPK3, CA3, PTGDS, CSDA, SMG1 MRPL28 DMPK, TUBA8 and LARGE. However, it is hard to assess without further biological experiments whether expression of the other genes that contained the module and were not in the original set is actually regulated by these putative *cis* elements.

Other investigators have included requirements for clustering of motifs within their algorithms to search for *cis* elements (20,24,43,52,53). Several of these algorithms consider a combination of motifs to be significant when its presence cannot be accounted for by a simple model (like a Poisson model) assuming independence for occurrences of separate motifs. Important evolutionary forces such as duplication and transposition are not well accounted for by independence assumptions made by many parametric models of motif clustering (as a trivial example, the sequence AAAAAAAAAAAA occurs much more frequently than would be predicted by independent occurrences of AAAAAA in the yeast 1 kb sequences before the TSS). Our implementation of the sparseness principle requires fewer assumptions and may be more relevant to the situation in the cell. In particular, it does not require determining a specific distribution for the motif binding sites.

Gene expression also depends on chromatin structure (2,4,54), on other regulatory mechanisms that are not seen at the DNA sequence level (e.g. the phosphorylation of a TF) and on elements far from the TSS (7). At the moment, our algorithm does not account for any of these. The high false positive rates observed with this as well as other algorithms stresses the fundamental role of biological experimentation. However, the use of computational algorithms to search for *cis*-regulatory elements significantly reduces the search space to one that is manageable with current laboratory techniques.

Understanding transcriptional regulatory networks in higher eukaryotes is a complex problem. As in other difficult problems in other research areas, it is important to identify some of the basic organizational structure that can account for the majority of the examples (albeit not necessarily all of them). Combinatorial regulation, sparseness and evolutionary conservation can guide the search for *cis* elements, which constitutes one of the first key steps towards a more detailed model of regulatory networks.

ACKNOWLEDGEMENTS

We would like to thank John Hogenesch, Sayan Mukherjee, Uwe Ohler, Tommy Poggio and Mariela Zirlinger for discussions and comments on the manuscript. G.K. is supported by a Whiteman Fellowship at MIT.

REFERENCES

1. Johnson,P. and McKnight,S. (1989) Eukaryotic transcriptional regulatory proteins. *Annu. Rev. Biochem.*, **58**, 799–839.
2. Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J.D. (1994) *Molecular Biology of The Cell*, 3rd Edn. Garland Publishing, New York, NY.
3. Davidson,E., Rast,J., Oliveri,P., Ransick,A., Calestani,C., Hood,L. and Bolouri,H. (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.

4. Lee,T.I. and Young,R.A. (2000) Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, **34**, 77–137.
5. Stormo,G.D. and Fields,D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.
6. Schleif,R. (1992) DNA looping. *Annu. Rev. Biochem.*, **61**, 199–223.
7. Blackwood,E.M. and Kadonaga,J.T. (1998) Going the distance: a current view of enhancer action. *Science*, **281**, 60–63.
8. Adhya,S. (1989) Multipartite genetic control elements: communication by DNA loop. *Annu. Rev. Genet.*, **23**, 227–250.
9. Arnone,M. and Davidson,E. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development*, **124**, 1851–1864.
10. Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, **8**, 1202–1215.
11. Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouze,P. and Moreau,Y. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, **9**, 447–464.
12. Bailey,T. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learn.*, **21**, 51–80.
13. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
14. Bussemaker,H., Li,H. and Siggia,E. (2001) Regulatory element detection using correlation with expression. *Nature Genet.*, **27**, 167–171.
15. van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.
16. Kel-Margoulis,O., Romaschenko,A., Kolchanov,N., Wingender,E. and Kel,A. (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res.*, **28**, 311–315.
17. Martone,R., Euskirchen,G., Bertone,P., Hartman,S., Royce,T.E., Luscombe,N.M., Rinn,J.L., Nelson,F.K., Miller,P., Gerstein,M., Weissman,S. and Snyder,M. (2003) Distribution of NF-kappa B-binding sites across human chromosome 22. *Proc. Natl Acad. Sci. USA*, **100**, 12247–12252.
18. Sinha,S., van Nimwegen,E. and Siggia,E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19**, i292–i301.
19. Bailey,T. and Noble,W. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19**, ii16–ii25.
20. Frith,M.C., Spouge,J.L., Hansen,U. and Weng,Z.P. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3214–3224.
21. Johansson,O., Alkema,W., Wasserman,W. and Lagergren,J. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, **19**, i169–i176.
22. Aerts,S., Van Loo,P., Thijs,G., Moreau,Y. and De Moor,B. (2003) Computational detection of cis-regulatory modules. *Bioinformatics*, **19** (Suppl. 2), ii5–ii14.
23. Halfon,M.S., Grad,Y., Church,G.M. and Michelson,A.M. (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, **12**, 1019–1028.
24. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) 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**, 757–762.
25. Markstein,M., Markstein,P., Markstein,V. and Levine,M.S. (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA*, **99**, 763–768.
26. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
27. Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human–mouse genome comparisons to locate regulatory sites. *Nature Genet.*, **26**, 225–228.
28. Boffelli,D., McAuliffe,J., Ovcharenko,D., Lewis,K.D., Ovcharenko,I., Pachter,L. and Rubin,E.M. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.
29. Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin,E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12**, 832–839.
30. Spellman,P., Sherlock,G., Zhang,M., Iyer,V., Anders,K., Eisen,M., Brown,P., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
31. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
32. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I., Zeitlinger,J., Jennings,E.G., Murray,H.L., Gordon,D.B., Ren,B., Wyrick,J.J., Tagne,J.B., Volkert,T.L., Fraenkel,E., Gifford,D.K. and Young,R.A. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
33. Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.
34. Pruitt,K. and Maglott,D. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
35. Bucher,P. (1990) Weight matrix descriptions of 4 eukaryotic RNA polymerase-II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
36. Berg,O. and von Hippel,P. (1987) Selection of DNA binding sites by regulatory proteins: statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.
37. Wingender,E., Kel,A., Kel,O., Karas,H., Heinemeyer,T., Dietze,P., Knuppel,R., Romaschenko,A. and Kolchanov,N. (1997) TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation. *Nucleic Acids Res.*, **25**, 265–268.
38. Koranda,M., Schleiffer,A., Endler,L. and Ammerer,G. (2000) Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters. *Nature*, **406**, 94–98.
39. Keeping,E.S. (1995) *Introduction to Statistical Inference*. Dover, New York, NY.
40. Shaffer,J. (1995) Multiple hypothesis testing. *Annu. Rev. Psychol.*, **46**, 561–584.
41. Zhu,G.F., Spellman,P.T., Volpe,T., Brown,P.O., Botstein,D., Davis,T.N. and Futcher,B. (2000) Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, **406**, 90–94.
42. Pic,A., Lim,F., Ross,S., Veal,E., Johnson,A., Sultan,M., West,A., Johnston,L., Sharrocks,A. and Morgen,B.A. (2000) The forkhead protein Fkh2 is a component of the yeast cell cycle transcription factor SFF. *EMBO J.*, **19**, 3750–3761.
43. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.*, **29**, 153–159.
44. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A., Patapoutian,A., Hampton,G.M., Schultz,P.G. and Hogenesch,J.B. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
45. Shmueli,O., Horn-Saban,S., Chalifa-Caspi,V., Shmoish,M., Ophir,R., Benjamin-Rodrig,H., Safran,M., Domany,E. and Lancet,D. (2003) Genenote: whole genome expression profiles in normal human tissues. *C. R. Biol.*, **326**, 1067–1072.
46. Su,A., Wiltshire,T., Batalov,S., Lapp,H., Ching,K., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G., Cooke,M., Walker,J. and Hogenesch,J. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
47. Zirlinger,M., Kreiman,G. and Anderson,D. (2001) Amygdala-enriched genes identified by microarray technology are restricted to specific amygdaloid sub-nuclei. *Proc. Natl Acad. Sci. USA*, **98**, 5270–5275.
48. Markstein,M. and Levine,M. (2002) Decoding cis-regulatory DNAs in the *Drosophila* genome. *Curr. Opin. Genet. Dev.*, **12**, 601–606.

49. Fickett,J. (1996) Quantitative discrimination of MEF2 sites. *Mol. Cell Biol.*, **16**, 437–441.
50. Yuh,C., Bolouri,H. and Davidson,E. (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**, 1896–1902.
51. Larsen,F., Gundersen,G., Lopez,R. and Prydz,H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
52. Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.
53. Hannenhalli,S. and Levy,S. (2002) Predicting transcription factor synergism. *Nucleic Acids Res.*, **30**, 4278–4284.
54. Emerson,B.M. (2002) Specificity of gene regulation. *Cell*, **109**, 267–270.