Published online 22 April 2007

Nucleic Acids Research, 2007, Vol. 35, No. 10 ***3203–3213***
doi:10.1093/nar/gkm201

Position and distance specificity are important determinants of *cis*-regulatory motifs in addition to evolutionary conservation

Saran Vardhanabhuti, Junwen Wang and Sridhar Hannenhalli*

Penn Center for Bioinformatics, Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104-6021

Received December 5, 2006; Revised March 16, 2007; Accepted March 21, 2007

ABSTRACT

Computational discovery of *cis*-regulatory elements remains challenging. To cope with the high false positives, evolutionary conservation is routinely used. However, conservation is only one of the attributes of *cis*-regulatory elements and is neither necessary nor sufficient. Here, we assess two additional attributes—positional and inter-motif distance specificity—that are critical for interactions between transcription factors. We first show that for a greater than expected fraction of known motifs, the genes that contain the motifs in their promoters in a position-specific or distance-specific manner are related, both in function and/or in expression pattern. We then use the position and distance specificity to discover novel motifs. Our work highlights the importance of distance and position specificity, in addition to the evolutionary conservation, in discovering *cis*-regulatory motifs.

INTRODUCTION

Eukaryotic gene transcription is controlled by a network of transcription factor (TF) proteins (1,2). TFs bind to specific DNA *cis* elements near transcription start sites, and through cooperative interaction, guide Polymerase-II complex to the transcription start site. Identification of *cis*-regulatory elements for the TFs is an important first step towards deciphering regulatory networks. This, however, remains a practical challenge because TFs often bind to highly diverse sequences resulting in degenerate binding models or motifs, and searching for putative binding sites using these degenerate motifs results in too many false positives.

It is now well established that regions in the genome that have been conserved over long evolutionary periods are more likely to be functional (3). Fortunately, such highly conserved regions make up only a small fraction of the genome (4). Thus by restricting the search for putative binding sites in evolutionarily conserved sequences, one can drastically reduce false positives. This is exactly the premise underlying the, now well established, approach of *Phylogenetic Footprinting* (5–8). However, for a genomic region to be functional, evolutionary conservation is neither necessary (9,10) nor sufficient (11). Besides conservation, what are other important characteristics of functional *cis* elements?

The regulation of gene transcription depends on interactions among transcription factors and the polymerase. This imposes location constraints on the corresponding DNA elements. For example, several *cis* elements occur at a specific distance relative to the transcription start site (TSS) (12). Additionally, several *cis* elements occur in the same promoter with restricted spacing between them. For example, in the adenovirus 2 *E1B* promoter, increased spacing between the GC-box and the TATA-box diminishes *in vivo* transcription significantly (13). There are other examples of such positional and spacing restrictions (14–19). Previous works have exploited the co-occurrence of promoter motifs to predict interacting TFs (20,21), to model expression regulation (22,23), and to detect regulatory modules (24,25), and some of these works impose specific distance constraints between co-occurring motifs. Positional constraints provide distinguishing characteristics of *cis*-regulatory elements in addition to evolutionary conservation, but have not been systematically exploited for motif discovery.

Here, we show that in human promoters a large fraction of known motifs exhibit significant positional constraint and a large number of motif-pairs exhibit significant inter-motif distance constraint. The target genes that have position-specific motifs or the distance-specific motif-pairs tend to be co-expressed and have similar functions. A large majority of these positionally constrained motifs are not conserved between human and mouse; this underscores the importance of positional constraints in discovering *cis*-regulatory motifs. Finally, to discover novel motifs, we assess the position and distance specificity of all words (7 bases long) and

*To whom correspondence should be addressed. Tel: +215 746 8683; Fax: +215 573 3111; Email: sridharh@pcbi.upenn.edu

word-pairs in human promoters that do not overlap a known motif. After clustering of similar motifs, this resulted in 168 position-specific novel motifs and 3708 distance-specific pairs involving a novel motif. Several of these are highly correlated with specific expression and function of the target genes.

RESULTS

Data preparation

We extracted 600 bp human promoter sequences (+500, −100) corresponding to 30 927 transcription start sites from DBTSS version 5.2 (26). We also extracted the human–mouse conservation for these regions from UCSC's axtNet database (UCSC hg17 release). TRANSFAC v8.4 (27) describes 546 vertebrate TF positional weight matrices (PWM). Often PWMs corresponding to evolutionarily related TFs are highly similar. To minimize the bias caused by this redundancy we clustered the PWMs based on their similarity and then retained 175 representative PWMs (methods). For these 175 PWMs, we scanned the 600 bp promoter sequences using our PWM_SCAN tool (6) with a stringent *P*-value threshold of $e^{-9.21}$ (chance expectation of one hit every 10 kb of human genome). Ten of the 175 PWMs did not have any match in our promoter set; our analysis is based on the remaining 165 TRANSFAC motifs. We used the Novartis tissue survey data (28) for gene expression profiles and GO (29) for functional annotation of genes. We only use the GO 'biological process', and to avoid non-specific biological processes we only include processes that are associated with at most 500 genes. This includes 99% of all the GO terms and eliminates non-specific GO terms.

A generic approach to Z-score calculation

To quantify motif conservation, motif positional specificity and motif-pair distance specificity, we use a generic procedure. Let N be the total occurrences of a motif (or motif-pair). Among these let n be the number of 'successful' occurrences. An occurrence could be called 'successful' if for instance, it is conserved. Given the expected success rate p_0, we assume a binomial distribution for the number of occurrences and estimate the Z-score as $[n-(N\times p_0)]/[\sqrt{\ }(N\times p_0\times(1-p_0)]$. A similar procedure was used in (8). Precisely what we mean by 'successful' and how do we estimate p_0 depends on the context and will be described later.

Evolutionary conservation of TRANSFAC motifs

We say that a motif match is conserved if the mouse sequence aligned with the human site also matched the PWM with a *P*-value $\leq e^{-9.21}$. To estimate the expected conservation rate p_0, we permuted each column of PWM to shuffle the nucleotide preferences in each position and generate a set of control matrices (five for each of the 165 TRANSFAC PWMs with a total of 819). We use these control matrices to obtain an overall expected conservation rate p_0. Figure 1 shows the conservation

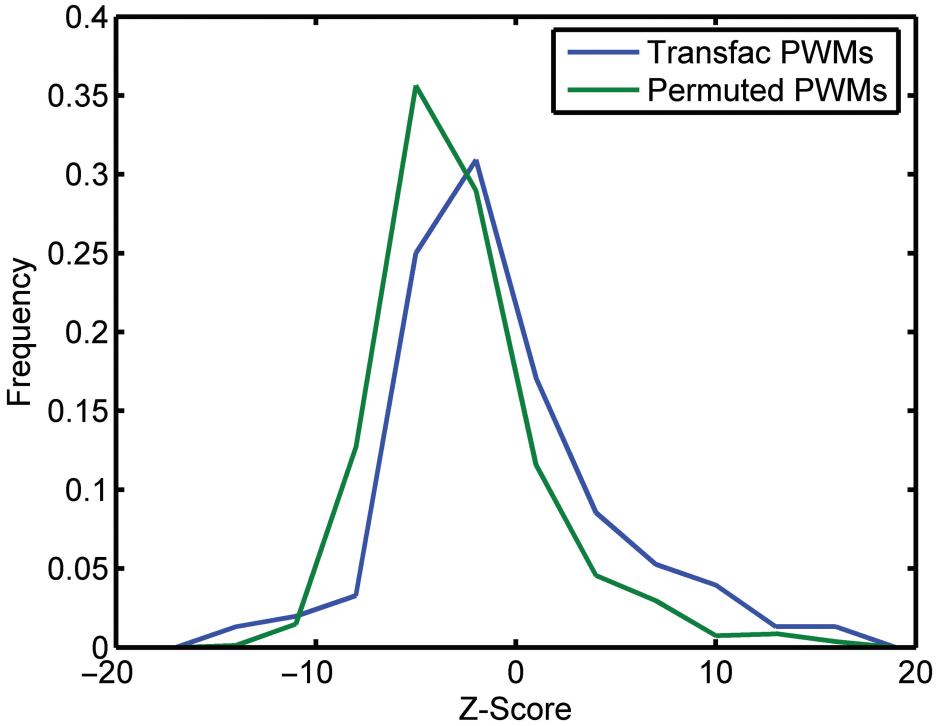

Figure 1. Conservation Z-scores for 165 TRANSFAC motifs. TRANSFAC motifs are generally more conserved compared to permuted motifs and majority of common core motifs are highly conserved. This two plots are significantly different; the Wilcoxon rank sum test based *p*-value $=2\times10^{-10}$. We have categorized motifs based on their conservation Z-scores. The high-conservation category ($Z>8$) has 27 motifs, the medium conservation category ($3\leq Z\leq 8$) has 27 motifs and the low conservation category ($Z<3$) has 111 motifs. Several core factors are conserved: CAAT box (100.3), Sp1 (90.4), Oct-1 (10.1), TATA (8.4).

Z-score distribution of 165 motifs. The core factors have high conservation Z-scores: CAAT-box (100.3), Sp1 (90.4), Oct-1 (10.1), TATA (8.4), etc. Figure 1 also shows the Z-score plot for the 819 randomized PWMs. Based on this plot, we arbitrarily categorize the motifs into three classes: (i) 27 highly conserved motifs (Z-score ≥ 8), (ii) 27 medium-conserved motifs ($3 \leq$ Z-score < 8) and (iii) remaining 111 non-conserved motifs.

Position specificity of TRANSFAC motifs

Here, we assess whether a motif preferentially occurs at a specific position relative to the transcription start site. Given the total occurrences of a motif and the subset of occurrences in a window (defined by the start position and the length), and the expected fraction of occurrences in the window, p_0, we compute the Z-score. We compute the Z-score for windows of length 20 bp starting at each position in the 600 bp promoter region and retain the maximum value for each motif among all windows; we call this the Z-max. An important concern in estimating p_0 is the GC-composition[1] of the motif and GC-composition of various parts of the promoter. We have experimented with three different controls (see Supplementary Data). Here, we report the results based on our most stringent control. To preserve the base composition of the motif, we randomly permute the columns of the PWM. To ensure a good representation we generate five permuted PWMs for a given TRANSFAC PWM. We estimate p_0 based on pooled occurrences of five permuted PWMs on the real promoter sequence. Note that p_0 is specific to each PWM and each window. It is easy to see that for instance, a G-rich motif is not going to differ from its permuted copies and thus will not have a high Z-score. At the risk of missing such cases, we decided to pursue this highly stringent control to minimize the risk of false discoveries. Figure 2a shows the distribution of Z-max for the 165 motifs. As a negative control for the Z-max distribution, we repeat the above process on randomized promoter sequences. The randomized promoter is generated so as to preserve the base composition at each position along the 600 bp region. As above, we estimate p_0 based on pooled occurrences of five permuted PWMs on the randomized promoter sequence. As shown in Figure 2a, a large fraction of TRANSFAC motifs occur in a position-specific fashion. As a reference, we show the positional Z-score distribution of three core motifs that exhibit high Z-max (Figure 2b).

Positional preferences of several core promoter motifs have been previously investigated. In our analysis, the GC-box binding TF, Sp-1 has a maximum Z-score at 66 bp upstream of the TSS, also observed in (8). We found CAAT-box binding TF NF-Y to have the maximum Z-score at position 86 bp upstream of TSS. Xie *et al.* have reported a preferred position of 89 bp upstream (8). TATA-box binding TF TBP is most frequent at ~35–31 bp upstream of TSS (30–32). However, the maximum Z-score in our analysis is achieved at 45 bp upstream of the TSS. Note that in contrast to frequency, we compute Z-score, which controls for the base composition. There is a peak of A + T frequency around 35–30 bp upstream of TSS (33), which would lower the Z-score exactly at those positions, and thus our Z-score peak is slightly shifted.

Based on these distributions, we define a set of 39 (23%) motifs to be position-specific (Z-max ≥ 5) (Supplementary Table T1) and another set of 38 motifs to be position-nonspecific (Z-max ≤ 3). Our numbers are consistent with the previous report where 25% of the known motifs were found to be position specific (8). Furthermore, we found that the position-specific motifs also tend to be conserved. Among the 39 position-specific motifs, 42% are highly conserved (conservation Z-score ≥ 8) whereas among the 38 position-nonspecific motifs, only 3% are highly conserved (chi-square P-value $= 3 \times 10^{-5}$). As expected, several known core factors like CAAT box (bound by NF-Y), Muscle TATA box, TBP, Sp1, etc. show a very high position specificity (Supplementary Table T1).

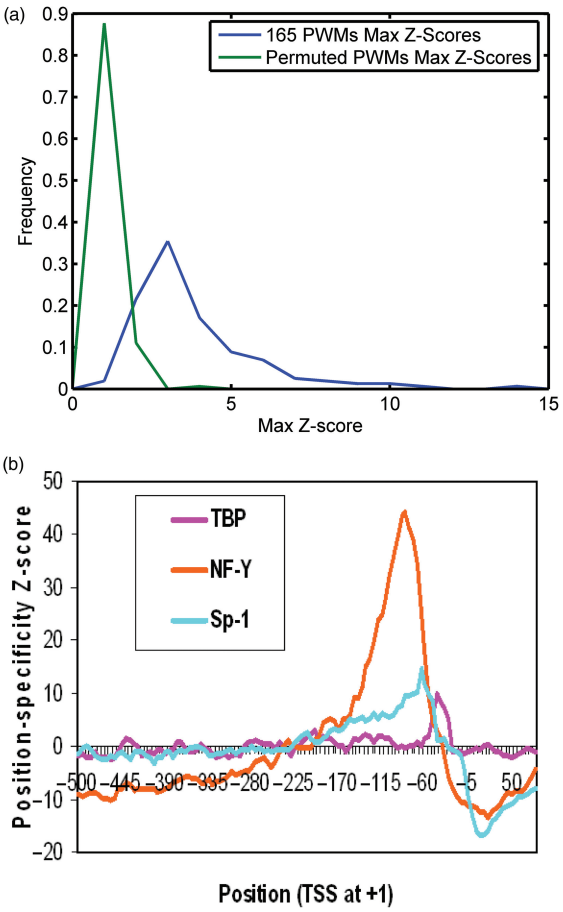

Figure 2. (**a**) Position specificity Z-max distribution for the 165 motifs. Also shown is the Z-max for random sequences with positionally matched base composition used as promoters. We define a set of 39 motifs to be position specific (Z-max ≥ 5) and a set of 38 motifs to be position nonspecific (Z-max ≤ 3). (**b**) Position specificity Z-score distribution for three core motifs. Transcription start site is at position +1.

[1] Throughout the manuscript, by 'CG' we mean '(C + G)' and by 'GC' we mean the GC dinucleotide.

Functional relevance of position-specific motifs

We investigated whether the presence of a position-specific motif in the gene promoter is correlated with the gene's function or expression pattern. We assessed the functional coherence of a set of target genes based on GO annotation using the Ontologizer tool (Robinson *et al.*, 2004). For each target gene-set, we also randomly select the same number of genes and subject them to same analysis. To assess how significantly the position- specificity of a motif correlates with the target gene function, we used three criteria to select the target gene-sets: (i) genes that contain position-specific motifs at specific position range where Z-max is achieved, (ii) genes that contain position-specific motifs at locations other than Z-max positions and (iii) genes that contain position-nonspecific motifs at any location. Each of the gene-sets also has a corresponding random gene-set of the same size. The GO Ontologizer tool compares a given gene-set with several functional categories for significant overlaps. To control for multiple testing, we pool the GO *P*-values for all gene-sets and their corresponding randomized gene-sets for the three criteria and based on the pooled set of *P*-values we estimate a *P*-value cutoff corresponding to a false discovery rate, FDR $\leq 5\%$ (34). Table 1 shows the fraction of motifs under each of the three criteria whose target genes show significant GO association and also the average number of associated GO processes. These numbers are also shown for the matched random gene-set. Additionally, among the motifs that do show a significant GO association, the table also shows the fraction that is conserved. As shown in Table 1, GO association is the greatest for position-specific occurrence of motifs and some of these could not be detected using conservation criterion alone. We have listed the significant GO associations under criterion A in Supplementary file 'Motif2GOAssociation'. These include mRNA processing and metabolism, protein localization, transcription, etc.

Next we investigated whether the gene targets of position-specific motifs are differentially expressed in specific tissues. For each of the 79 tissues from the Novartis dataset, we assessed using Wilcoxon rank sum test, whether a target gene-set had differential (up or down) expression relative to all other genes. Each pair of gene-set and tissue results in a *P*-value and based on pooled *P*-values as before we estimate a cutoff corresponding to FDR $\leq 1\%$[2]. We consider a motif to have differential expression if in at least one of the 79 tissues the target genes are differentially expressed with a significant *P*-value. Approximately 1% of the random gene-sets show significant differential expression at a FDR cutoff of 1%. Thus if 1% of the *P*-values are significant by chance, we expected $\sim 55\%$ $(1 - 0.99^{79})$ of the motifs will show differential expression in at least one tissue by chance alone. The relative-enrichment number (column 6 of Table 1) is the ratio of the 'actual% of motifs that show differential expression' to 55. The average number of tissues in which the gene-set is differentially expressed is also of interest. Table 1 shows the fraction of motifs under each of the three criteria whose target genes are differentially expressed, as well as the average number of tissues. These numbers are also shown for the matched random gene-set. Additionally, among the motifs that are differentially expressed, the table also shows the fraction that is conserved. Similar to our conclusions based on GO analysis, the tissue-specific differential expression is most prevalent among the position-specific occurrence of motifs and many of these motifs could not be detected using conservation criterion alone. Only about half of tissue-associated motifs are conserved. The top five most significant tissues under criterion A are SuperiorCervicalGanglion, Skin, PrefrontalCortex, PB-CD8 + TCells, Ovary and Atrioventricular node.

Thus our results highlight the importance of position specificity of cis elements and that it should be used in conjunction with conservation to identify *cis*-regulatory motifs. Supplementary Table T1 lists 39 representative position-specific TRASNFAC motifs. Most of these are known to be involved in condition-specific regulation (as opposed to basal transcription).

[2] At a higher FDR threshold, a random gene-set shows significant enrichment in at least one of the 79 tissues, simply by chance.

Distance specificity of TRANSFAC motif-pairs

Next we assessed how often a motif-pair preferentially occurs at a specific distance from each other. Given the total number of occurrences of a motif-pair and the subset that falls within a distance range, defined by the minimum distance and the range, and the expected fraction, p_0, we compute the Z-score. Here, the locations of individual motifs are irrelevant and only the distance between them is of concern. Much like position specificity analysis, we compute the Z-score for 50 bp ranges starting at each position and retain for each motif-pair the maximum value, the Z-max. Our control is analogous to that for the position specificity analysis, i.e. we randomly permute the columns of each PWM to generate the background matches. Figure 3 shows the distribution of Z-max for the 21 777 motif-pairs on real promoters and permuted PWMs on the random promoter sequences with positionally conserved GC composition. Two position-specific motifs will obviously manifest as distance specific. To unambiguously reveal distance specificity, we require that at least one of the motifs in the pair must be position nonspecific. Figure 3 also shows the distribution of Z-max for this reduced set of motif-pairs, and for comparison the motifs pairs where both motifs are position specific. From the reduced set (pairs with at least one non-position-specific motif), based on Z-max distributions, we define a set of 915 motif-pairs to be distance specific (Z-max ≥ 4) (listed in Supplementary Table T2) and another set of 865 motif-pairs to be distance nonspecific (Z-max ≤ 2). The distance-specific pairs involve 41 position-specific motifs and 38 position-nonspecific motifs. For comparison and subsequent analyses, we also included pairs where both motifs are position specific, using the same Z-score cutoff ($Z \geq 4$) we get 410 motif-pairs.

Table 1. Functional relevance of known motifs. Gene targets were obtained for motifs based on three different criteria

Criterion	Number of Motifs	Number associated with GO process with FDR ≤ 5%		Average number of associated GO processes		Percent of GO-associated motifs that are conserved	Relative enrichment in tissue-associated motifs		Average number of tissues		Percent of tissue-associated motifs that are conserved
		Real	Random	Real	Random	Real	Real	Random	Real	Random	Real
A	39	6 (15%)	0 (0%)	3	0	4/6 (67%)	1.6	0.3	31	8	17/35 (49%)
B	39	1 (3%)	0 (0%)	1	0	1/1 (100%)	0.9	0.3	22	4	11/21 (52%)
C	38	1 (3%)	1 (3%)	1	2	0 (0%)	0.6	0.1	9	0	1/14 (7%)

(A) Gene promoters containing position-specific motif in the preferred window, (B) Gene promoters containing position-specific motif at any position, (C) Gene promoters containing position-nonspecific motif at any position. For each 'Real' target gene-set a 'Random' gene-set of the same size was selected. Table shows for each criterion: col2: number of gene-sets, col3: fraction of gene-sets that associated with a GO process (FDR ≤ 5%), col4: the average number of GO processes, col5: among the motifs that associated with a GO process, the fraction that was conserved, col6: ratio of the 'actual% of motifs that show differential expression (FDR ≤ 1%)' to the expected fraction of 55 (see text for how 55 was calculated), col7: the average number of such tissues, col8: among the motifs that associated with tissue, the fraction that was conserved. For each of these columns we show the figures for both the 'Real' and the 'Random' gene-set.

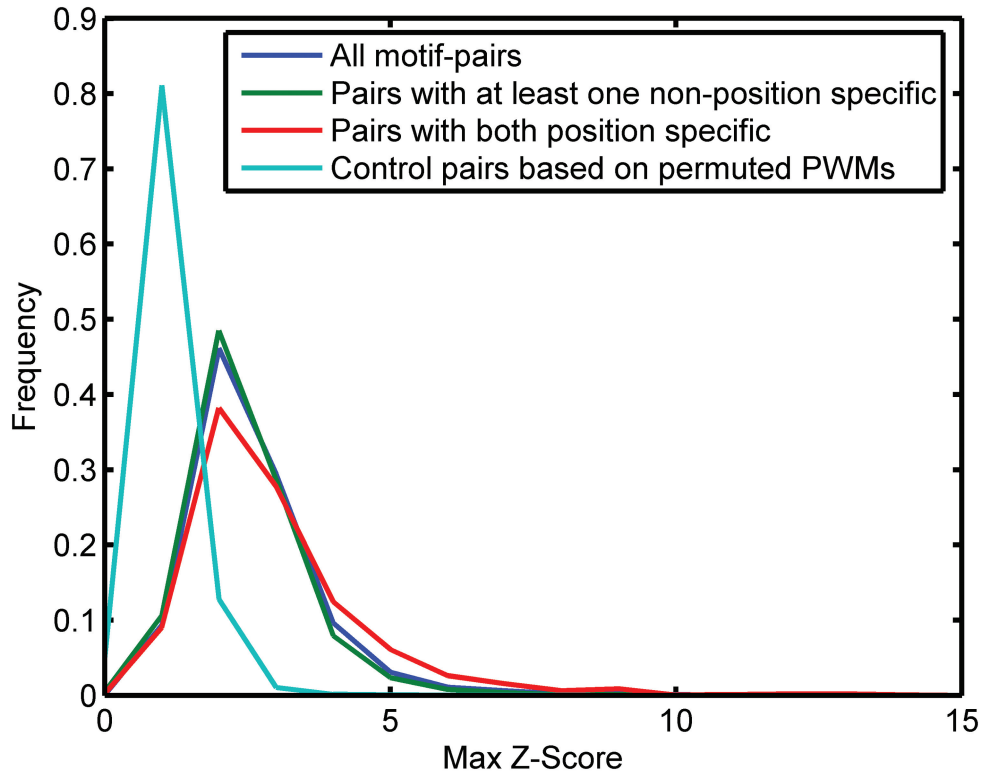

Figure 3. Distribution of the distance-specificity Z-max distribution for the 21 777 motif-pairs on real promoters are shown in blue. The Z-max distribution for permuted PWM pairs on random promoter sequences is shown in cyan. Distribution for pairs with at least one non-position-specific motif (7804 pairs) is shown in green and pairs with both position-specific motifs comprise (1618 pairs) is shown in red.

Functional relevance of distance-specific motif-pairs

We investigated whether the presence of a motif-pair at a specific distance range is correlated with the gene's function or expression pattern. To assess the correlation between distance specificity and the target gene function, we used four criteria to select the target gene-sets: (i) genes that contain distance-specific motifs at Z-max distance where both the motifs are position specific, (ii) genes that contain distance-specific motifs at the Z-max distance and at least one of the motif is position nonspecific, (iii) genes that contain distance-specific motifs at any distance other than the Z-max distance where at least one of the motif is position nonspecific and (iv) genes that contain distance-nonspecific motifs at arbitrary distance. We also generate a random gene-set for each of the gene-sets obtained using the above four criteria.

We performed the GO and tissue expression analysis using same procedure as that for the analysis of position-specific motifs above and the result summarized in Table 2 is organized similarly to Table 1. The gene promoters containing distance-specific motif-pairs that are both position specific at a preferred distance have highest association with GO process and highest

Table 2. Functional relevance of known motif pairs. Gene targets were obtained for motif-pairs based on four different criteria

Criterion	Number of motifs	Number associated with GO process with FDR ≤ 5%		Average number of associated GO processes		Percent of GO-associated motifs that are conserved	Relative enrichment in tissue-associated motifs		Average number of tissues		Percent of tissue-associated motifs that are conserved
		Real	Random	Real	Random	Real	Real	Random	Real	Random	Real
A	245	61 (25%)	12 (5%)	4	1	37/61 (61%)	1.6	0.6	36	10	78/217 (36%)
B	321	35 (11%)	16 (5%)	3	1	4/35 (11%)	1.5	0.4	25	6	8/266 (3%)
C	321	31 (10%)	12 (4%)	2	1	2/31 (7%)	1.1	0.4	18	7	16/196 (8%)
D	417	24 (6%)	15 (4%)	2	1	1/24 (4%)	1.1	0.3	16	7	3/254 (1%)

For each criterion, for all the gene-sets with significant GO processes (FDR ≤ 5%) were computed. Four criteria were used to select target gene-sets: (A) genes that contain distance-specific motifs at Z-max distance and both the motifs are position specific, (B) genes that contain distance-specific motifs at the Z-max distance and at least one of the motif is position nonspecific, (C) genes that contain distance-specific motifs at any distance other than the Z-max distance where at least one of the motif is position nonspecific and (D) genes that contain distance-nonspecific motifs at arbitrary distance. For each 'Real' target gene-set, a 'Random' gene-set of the same size was selected. See Table 1 legend for the description of the columns 2–8.

fold-enrichment for differential expression in tissues. By far, most of these motifs are more conserved compared with the gene-sets using other criteria. Particularly of interest, criterion B, that includes gene promoters containing distance-specific motif-pairs where at least one motif is non-position specific, show significant association with GO processes and differential expression in tissues. Moreover, the functional and expression associations for criterion B is stronger than that for criterion C, which includes all occurrences of the motif-pair, as opposed to only distance-specific occurrence in criterion B. Supplementary Table T2 lists the top 100 of the 7804 distance-specific motif-pairs where at least one of the pair is position nonspecific. In fact in 826 pairs, both motifs are position nonspecific. A cursory inspection of literature shows support for many of these. Muscle-specific motif (derived from actin promoter among other genes) has a Z-score of 11.63 with NF-Y which is known to form a complex on the alpha-actin-4 promoter (35). Ets and Pax5 (BSAP) with Z-score = 6.91 are known to interact physically to regulate a B-cell specific promoter (36). SREBP-1 and NF-Y show a Z-score of 5.6; expression of mouse gene ACBP is induced in hepatocytes by SREBP1 and this induction also requires a NF-Y-binding site (37). We have listed the significant GO associations under criteria A and B in Supplementary file 'Motif2GOAssociation'. These include protein localization/transport, regulation of lipid metabolism, biopolymer metabolism, RNA processing/metabolism, regulation of transcription, negative regulation of biological process, etc. The top five most significant tissues under criteria A and B include several CD cells, dendritic cell, T cell, ColorectalAdenocarcinoma, PrefrontalCortex, OccipitalLobe, Subthalamicnucleus and Hypothalamus. This is consistent with previous report by Xie *et al.*, where the two main groups of motif-associated tissues were found to be brain and immunity related (8).

Novel motifs in human promoters

For a majority of human transcription factors, their DNA-binding specificities are not known. *De novo* motif discovery thus remains important in analyzing transcriptional networks. Our analysis of known motifs shows that position and distance constraints, in addition to conservation, are important attributes of *cis*-regulatory motifs and thus can be used to detect novel motifs. To ensure that the motifs we detect do not correspond to any of the 175 representative TRANSFAC motifs, we start by masking all positions in all human promoters that matched any of the TRANSFAC motifs based on *P*-value threshold (see above). We then extract all 7-mers from the unmasked portion of the promoters, while allowing for at most two bases overlap with the masked portion on either side. We then cluster these 7-mers; two 7-mers were clustered if either they had at most 1 mismatch or they had 6 identical bases (after 1 base shift) including the reverse complement. All 7-mers within a cluster were aligned and a PWM was derived from each cluster. The set of 661 PWMs were then subject to same analysis as the TRANSFAC motifs to compute their conservation, position specificity and distance specificity. Seventy four of these 661 were conserved ($Z \geq 8$) and 168 were position specific ($Z \geq 6$) and 3708 pairs were distance specific (see below). We report a subset of 74 novel conserved motifs in Supplementary Table T3 along with their corresponding position and distance specificity properties. The Logos of the top 10 conserved motifs are shown in Supplementary Table T3a. We found that a large fraction of position-and distance-specific motifs are conserved and this fraction increases with increasing position or distance specificity. Figures 4 and 5 show the relationship between conservation and respectively the position and the distance specificity.

Functional analysis of novel motifs

We analyzed the target gene-sets of the novel motifs for their functional coherence and for tissue-specific differential expression. For each novel motif, to define the gene targets, we applied each criterion that it was detected by. For instance, if a motif was detected by position specificity, we consider a gene promoter as a target if the motif is at specific position. Each gene target

set thus obtained is subject to GO and expression analysis as for the TRANSFAC motifs described above.

We categorized the 661 7-mer motifs into 168 (25%) position-specific motifs ($Z \geq 6$) and 123 position-nonspecific motifs ($Z \leq 3$). In a previous analysis, Xie *et al.* have reported 35% of novel motifs to be position specific (Supplementary Table T4 shows these 168 motifs and associated information). The Logos of the top 10 position-specific motifs are shown in Supplementary Table T4a. We applied the previously described three criteria to select the target gene-sets and Table 3 shows the results of the GO and tissue expression analysis. As was the case for the TRANSFAC motifs, very few motifs show strong GO association, but a large fraction show a strong association to tissue expression. The trend across the three criteria is, however, similar to that for the TRANSFAC motifs (compare Table 1 and Table 3). We have listed the significant GO associations under criterion A in Supplementary file 'Motif2GOAssociation'. These include RNA/mRNA metabolism/processing, reproduction and pregnancy. The top five most significant tissues under criterion A are PrefrontalCortex, OccipitalLobe, Hypothalamus, Amygdala and Thyroid. This is consistent with what was reported in (8).

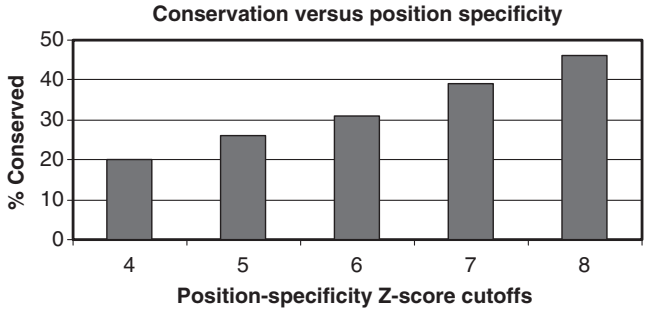

Figure 4. For the 661 novel motifs, as we increase the threshold for the position specificity, the fraction of qualifying motifs that are conserved increases. At a stringent position-specificity Z-score ≥ 8, 46% of motifs are conserved compared to only 20% among the motifs that have position-specificity Z-score ≥ 4.

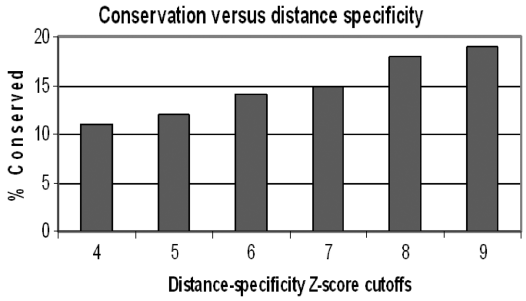

Figure 5. For all 79 576 motif-pairs (at least 1 motif is non-position specific), as we increase the threshold for the distance specificity, the fraction of qualifying motifs that are conserved increases. At a stringent distance-specificity Z-score ≥ 9, 19% of motifs are conserved compared to only 11% among the motifs that have distance-specificity Z-score ≥ 4.

For distance specificity between these 661 7-mer motifs, we have a total of 285 762 motif-pairs that have at least 100 non-overlapping occurrences in all the promoter regions. A further requirement that at least one of the motifs be position-nonspecific results in 79 576 pairs to analyze. We categorized these motif-pairs into 3708 distance-specific motif-pairs ($Z \geq 5$) and 9204 distance-nonspecific motif-pairs ($Z \leq 2$). The distance-specific motif-pairs are provided in a Supplementary file. Because the numbers are very large, we randomly selected 200 pairs from each group and did the GO and tissue expression analysis. We applied the previously described four criteria to select the target gene-sets and Table 4 shows the results of the GO and tissue expression analysis. Overall the results are consistent with that for the TRANSFAC motif-pairs (compare Table 2 and Table 4). We have listed the significant GO associations under criteria A and B in Supplementary file 'Motif2GOAssociation'. These include regulation of transcription, regulation of cellular/biological process, protein/biopolymer modification, phosphorylation/phosphate metabolism, chromatic modification, reproduction, pregnancy, etc. The top five most significant tissues under criteria A and B include several CD cells, dendritic cell, B cell and prefrontal cortex.

DISCUSSION

Comprehensive identification of genomic *cis*-regulatory elements is an important long-term goal. *Cis*-regulatory

Table 3. Functional assessment of novel motifs. Gene targets were obtained for motifs based on four different criteria

Criterion	Number of Motifs	Number associated with GO process with FDR $\leq 5\%$		Average number of associated GO processes		Percent of GO-associated motifs that are conserved	Relative enrichment in tissue-associated motifs		Average number of tissues		Percent of tissue-associated motifs that are conserved
		Real	Random	Real	Random	Real	Real	Random	Real	Random	Real
A	168	2 (1%)	0 (0%)	3	0	1/2 (50%)	1.6	0.3	32	9	48/146 (33%)
B	168	0 (0%)	0 (0%)	0	0	0/0 (0%)	1.1	0.3	20	10	35/98 (36%)
C	123	0 (0%)	0 (0%)	0	0	0/0 (0%)	0.5	0.1	7	7	2/29 (7%)

For each criterion, for all the gene-sets with significant GO processes (FDR $\leq 5\%$) were computed. Three criteria were used to select target gene-sets: (A) gene promoters containing position-specific motif in the preferred window, (B) gene promoters containing position-specific motif at any position and (C) gene promoters containing position-nonspecific motif at any position. For each 'Real' target gene-set and 'Random' gene-set of the same size was selected. See Table 1 legend for the description of the columns 2–8.

Table 4. Functional assessment of novel motif-pairs. Gene targets were obtained for motif-pairs based on four different criteria

Criterion	Number of motifs	Number associated with GO process with FDR ≤ 5%		Average number of associated GO processes		Percent of GO-associated motifs that are conserved	Relative enrichment in tissue-associated motifs		Average number of tissues		Percent of tissue-associated motifs that are conserved
		Real	Random	Real	Random	Real	Real	Random	Real	Random	Real
A	200	23 (12%)	8 (4%)	3	1	6/23 (26%)	1.7	0.9	44	15	44/189 (23%)
B	200	17 (9%)	6 (3%)	3	1	4/17 (24%)	1.6	0.6	32	10	14/172 (8%)
C	200	16 (8%)	5 (3%)	3	1	3/16 (19%)	1.3	0.6	29	13	15/148 (10%)
D	200	3 (2%)	3 (2%)	1	1	2/3 (67%)	1.0	0.4	16	10	0/107 (0%)

For each criterion, for all the gene-sets with significant GO processes (FDR ≤ 5%) were computed. Four criteria were used to select target gene-sets: (A) gene promoters containing distance-specific motif-pairs that are also position specific at the preferred distance, (B) gene promoters containing distance-specific motif-pairs (at least one motif is non-position specific) at the preferred distance, (C) gene promoters containing distance-specific motif-pairs (at least one motif is non-position specific) at any distance and (D) gene promoters containing distance-nonspecific motif-pairs at any distance. For each 'Real' target gene-set, a 'Random' gene-set of the same size was selected. See Table 1 legend for the description of the columns 2–8.

elements are often characterized by evolutionary conservation. In order to reduce false positives, the search for *cis*-regulatory elements is traditionally restricted to evolutionarily conserved regions of the genome. However, both, lack of conservation among *cis*-regulatory elements, as well as lack of functionality among conserved elements has been previously reported (9–11). Here, we have assessed the importance of two additional attributes of *cis*-regulatory elements—their position specificity from the TSS, as well as the spacing between them (14–19).

Several previous works have shown *cis*-regulatory motifs to be positionally constrained. Xie *et al*. have reported a large number of motifs in the human genome based on multiple genome comparison (8). Although they have shown that some of the novel motifs are position specific relative to the TSS, this position specificity of the motifs was not used in the discovery process itself. In another work, motifs in *Escherichia coli* were ranked based on the enrichment of specific spacing between them and were experimentally validated for their functionality in binding sites (38). We have previously reported a promoter model that captures the position and distance specificities of motifs (39). Other previous works have exploited the co-occurrence of promoter motifs to predict TF interactions and TF modules (20–25). Here, we have explicitly and extensively assessed the importance of two attributes of *cis*-regulatory elements—position and distance specificity—independent of evolutionary conservation. Our results indicate that even though evolutionary conservation is the most important attribute of *cis*-regulatory elements, these additional attributes are important, especially to detect the species-specific elements. We emphasize that our work does not represent a novel motif discovery tool, in the traditional sense of the term. Traditionally, the term 'motif discovery' refers to identification of motifs potentially mediating the regulation of a set of co-expressed genes (40). Our work, by exploiting the positional and distance constraints, however does identify a global set of motifs in the human gene promoters that potentially mediate transcriptional regulation.

Although we have controlled for base composition, we have not controlled for dinucleotide composition, especially for CG dinucleotide frequency. Clearly, using more stringent di- and tri-nucleotide controls will result in increased specificity in motif detection, however at the risk of decreased sensitivity. The compositional bias in certain genomic regions may have been preserved over evolutionary period precisely for the maintenance of *cis*-regulatory elements and thus using the extremely stringent local composition as a control will result in failed detection of these *cis*-regulatory motifs on statistical grounds. We have mentioned the example of TATA box earlier, where, because of a local (A + T) frequency peak at ~35 bp upstream, where TATA box is most abundant, we detect the position of most specificity at a slightly shifted position of 45 bp upstream. The CG dinucleotides are of special concern because of their association with DNA methylation and transcriptional regulation (41). Several of the detected position-specific motifs have one or more CG dinucleotides; indeed there is an enrichment of CG-containing motifs among the position-specific motifs, both in known and novel motifs. Similar to previous observation, these position-specific motifs mostly occur in 100 bp upstream of the TSS (8). We have performed a number of cautionary analyses to ensure that CG-containing motifs are not detected simply because of lack of control for CG dinucleotides. We have summarized these analyses in the Supplementary material.

For both, known and novel motifs, our analysis shows the importance of position specificity in determining the functional and tissue association of the target genes. These findings are consistent with the shifting view of core promoter as an active participant in the regulation of eukaryotic gene expression (42). Besides the top five tissues mentioned earlier that are enriched for targets of position-specific motifs, fetal brain and fetal liver are also among the significant tissues; this is also true for distance-specific motif-pairs. Although our result shows a greater tissue enrichment for position- and distance-specific motifs, the specific tissues revealed by our analysis may be over-interpreted; permutation-based test for tissue enrichment may be more stringent. The relative enrichment of tissue-associated motifs is much higher when the position-specific motifs occur at their preferred positions

(compare column 6 for rows A and B in Table 1 and Table 3) thus underscoring the importance of position specificity for *cis*-regulatory motifs. Furthermore, although there is a strong correlation between position specificity and the conservation (Figure 4), more than half of the position-specific motifs that drive tissue-specific expression are not conserved (last columns in Table 1 and Table 3), thus underscoring the importance of position specificity for motif discovery independent of conservation.

Similar conclusions can be made regarding the importance of distance specificity in the discovery of *cis*-regulatory motifs. Distance-specific motif-pairs that are comprised of position-specific motifs (row A in Table 2 and Table 4) by far show the most functional association, tissue association and conservation. In fact there is a correlation between distance specificity and conservation (Figure 5). Nevertheless the distance-specific motif-pairs that include non-position-specific motifs (row B) show a comparable tissue association and slightly lower functional association, and yet a large majority of these are not conserved and would be missed by a conservation-only based approach. Thus, based on an unbiased comprehensive analysis we have shown the importance of position and distance specificity in discovering *cis*-regulatory motifs beyond the use of evolutionary conservation alone, which is likely to miss species-specific *cis*-regulatory motifs.

METHODS

Clustering of TRANSFAC PWMs to obtain representative PWMs

We have previously reported an information-theoretic approach to compute the similarity between a pair of PWMs (43). Here, we introduce the approach briefly and provide the details in the Supplementary Data. To compute the similarity between a pair of PWMs, we use a symmetric derivative of the standard relative entropy measure (44). This measure is transformed into a Z-score, and eventually into a *P*-value, based on empirically derived distributions. We allow for shifts between the PWMs and our measure accounts for the PWM widths. Using an appropriate *P*-value threshold for the pairwise similarity, we then compute clusters of similar PWMs; we use bi-connected components (in contrast to single-linkage clusters) as our clusters. Finally, for each cluster, we select the median PWM as the cluster representative. See Supplementary Data for further details.

Z-score computation

We use Z-score to quantify motif conservation, motif positional specificity and motif-pair distance specificity. Assuming a binomial distribution, the Z-score is defined as:

$$Z = \frac{[n - (N \times p_0)]}{\left[\sqrt{(N \times p_0 \times (1 - p_0)}\right]}$$

For conservation. N = total number of occurrences of a motif.

n = number of conserved occurrences of a motif.

p_0 = background conservation rate, i.e. the expected fraction of occurrences that are conserved based on permuted PWMs.

For position specificity. N = total number of occurrences of a motif.

n = number of occurrences of a motif at a particular bin position.

p_0 = expected fraction of occurrences at a particular bin based on permuted PWMs.

For distance specificity. N = total number of occurrences of a motif-pair.

n = number of motif-pair occurrences where the motifs occur at a specific distance range, or distance bin.

p_0 = expected fraction of occurrences where the motifs occur at a specific distance bin based on permuted PWMs.

GO analysis

For both, position specificity and distance specificity, gene-sets are generated for particular motifs or motif-pairs according to different criteria (see Results Section). For each of these gene-sets, a matched random control set of genes is generated with equal number of genes. These sets of genes are then fed into GO Ontologizer tool (45) and *P*-values are generated for significance of motif to a GO-term association. A 5% FDR cutoff are applied to these *P*-values. We only used the biological process GO terms that have at most 500 genes to avoid ubiquitous classes.

Tissue differential expression analysis

Similar to GO analysis, real and random set of genes are generated for particular motifs or motif-pairs. For each gene-set, and for each of the 79 tissues in the GNF tissue survey data (28), we test whether the genes in the set are differentially expressed in the tissue using Wilcoxon rank sum test. Each pair of gene-set and tissue results in a *P*-value and based on pooled *P*-values, we estimate a cutoff for FDR $\leq 1\%$. We consider a gene-set differentially expressed, if it is differentially expressed (FDR $\leq 1\%$) in at least one of the 79 tissues.

Clustering novel 7-mers to form PWMs

All 7-mers motifs that are not covered by a TRANSFAC match are ranked according to their conservation Z-score with highly conserved motifs appearing nearer to the top of the list. We then walked down the list and clustered current motif with motifs we have already encountered according to the following criteria. Two 7-mers were clustered if either they had at most 1 mismatch (including the reverse complement) or they had 6 identical bases (after 1 base shift). All 7-mers within a cluster were aligned and a PWM was derived from each cluster.

ADDITIONAL FILE

There are three additional files. They are as follows:

File1: PositionalMotifs_AdditionalFile, this includes Supplementary results.

File2: Motif2GOAssociation, this includes the significant GO associations for known and novel position/distance-specific motifs.

File3: DistanceSpecificMotifs, this includes all novel distance-specific motif-pairs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

This work and the Open Access publication charges were funded by NIH grant R21GM078203.

Conflict of interest statement. None declared.

REFERENCES

1. Ptashne,M. (2004) *A Genetic Switch3rd edn*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
2. Kadonaga,J.T. (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, **116**, 247–257.
3. Thomas,J.W., Touchman,J.W., Blakesley,R.W., Bouffard,G.G., Beckstrom-Sternberg,S.M, Margulies,E.H, Blanchette,M., Siepel,A.C, Thomas,P.J *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.
4. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
5. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
6. Levy,S. and Hannenhalli,S. (2002) Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome*, **13**, 510–514.
7. 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.
8. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
9. Plessy,C., Dickmeis,T., Chalmel,F. and Strahle,U. (2005) Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends Genet.*, **21**, 207–210.
10. Emberly,E., Rajewsky,N. and Siggia,E.D. (2003) Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics*, **4**, 57.
11. Nobrega,M.A., Zhu,Y., Plajzer-Frick,I., Afzal,V. and Rubin,E.M. (2004) Megabase deletions of gene deserts result in viable mice. *Nature*, **431**, 988–993.
12. Lim,C.Y., Santoso,B., Boulay,T., Dong,E., Ohler,U. and Kadonaga,J.T. (2004) The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.*, **18**, 1606–1617.
13. Grace,M.L., Chandrasekharan,M.B., Hall,T.C. and Crowe,A.J. (2004) Sequence and spacing of TATA box elements are critical for accurate initiation from the beta-phaseolin promoter. *J. Biol. Chem.*, **279**, 8102–8110.
14. Kadonaga,J.T. (2002) The DPE, a core promoter element for transcription by RNA polymerase II. *Exp. Mol. Med.*, **34**, 259–264.
15. Butler,J.E. and Kadonaga,J.T. (2001) Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev.*, **15**, 2515–2519.
16. Spek,C.A., Bertina,R.M. and Reitsma,P.H. (1999) Unique distance- and DNA-turn-dependent interactions in the human protein C gene promoter confer submaximal transcriptional activity. *Biochem. J.*, **340**(Pt 2), 513–518.
17. Wu,L. and Berk,A. (1988) Constraints on spacing between transcription factor binding sites in a simple adenovirus promoter. *Genes Dev.*, **2**, 403–411.
18. Sugiyama,T., Scott,D.K., Wang,J.C. and Granner,D.K. (1998) Structural requirements of the glucocorticoid and retinoic acid response units in the phosphoenolpyruvate carboxykinase gene promoter. *Mol. Endocrinol.*, **12**, 1487–1498.
19. Senger,K., Armstrong,G.W., Rowell,W.J., Kwan,J.M., Markstein,M. and Levine,M. (2004) Immunity regulatory DNAs share common organizational features in Drosophila. *Mol. Cell*, **13**, 19–32.
20. Hannenhalli,S. and Levy,S. (2002) Predicting transcription factor synergism. *Nucleic Acids Res.*, **30**, 4278–4284.
21. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.
22. Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.
23. Thompson,W., Palumbo,M.J., Wasserman,W.W., Liu,J.S. and Lawrence,C.E. (2004) Decoding human regulatory circuits. *Genome Res.*, **14**, 1967–1974.
24. Segal,E. and Sharan,R. (2004) In *Proceedings of the Eigth Annual International Conference on Computational Molecular Biology*. ACM Press, New York, NY, USA, San Diego, CA, pp. 141–149.
25. Hannenhalli,S. and Levy,S. (2003) Transcriptional regulation of protein complexes and biological pathways. *Mamm. Genome*, **14**, 611–619.
26. Suzuki,Y., Yamashita,R., Sugano,S. and Nakai,K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res.*, **32**, D78–D81.
27. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
28. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, **101**, 6062–6067.
29. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
30. Gershenzon,N.I. and Ioshikhes,I.P. (2005) Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics*, **21**, 1295–1300.
31. Ohler,U., Liao,G.C., Niemann,H. and Rubin,G.M. (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, **3**, RESEARCH0087.
32. Juven-Gershon,T., Hsu,J.Y. and Kadonaga,J.T. (2006) Perspectives on the RNA polymerase II core promoter. *Biochem. Soc. Trans.*, **34**, 1047–1050.
33. Schug,J., Schuller,W.P., Kappen,C., Salbaum,J.M., Bucan,M. and Stoeckert,C.J.Jr (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, **6**, R33.
34. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, **100**, 9440–9445.
35. Poch,M.T., Al-Kassim,L., Smolinski,S.M. and Hines,R.N. (2004) Two distinct classes of CCAAT box elements that bind nuclear factor-Y/alpha-actinin-4: potential role in human CYP1A1 regulation. *Toxicol. Appl. Pharmacol.*, **199**, 239–250.
36. Fitzsimmons,D., Hodsdon,W., Wheat,W., Maira,S.M., Wasylyk,B. and Hagman,J. (1996) Pax-5 (BSAP) recruits Ets proto-oncogene family proteins to form functional ternary complexes on a B-cell-specific promoter. *Genes Dev.*, **10**, 2198–2211.

37. Neess,D., Kiilerich,P., Sandberg,M.B., Helledie,T., Nielsen,R. and Mandrup,S. (2006) ACBP–a PPAR and SREBP modulated housekeeping gene. *Mol. Cell. Biochem.*, **284**, 149–157.
38. Bulyk,M.L., McGuire,A.M., Masuda,N. and Church,G.M. (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in Escherichia coli. *Genome Res.*, **14**, 201–208.
39. Wang,J. and Hannenhalli,S. (2006) A mammalian promoter model links *cis* elements to genetic networks. *Biochem. Biophys. Res. Commun.*, **347**, 166–177.
40. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
41. Cross,S.H. and Bird,A.P. (1995) CpG islands and genes. *Curr. Opin. Genet. Dev.*, **5**, 309–314.
42. Burke, T.W. and Kadonaga, J.T. (1997). The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. Genes Dev., **11**(22), 3020–31.
43. Everett,L., Wang,L.S. and Hannenhalli,S. (2006) Dense subgraph computation via stochastic search: application to detect transciptional modules. *Bioinformatics*, **22**, e117–e123.
44. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.
45. Robinson,P.N., Bohme,U., Lopez,R., Mundlos,S. and Nurnberg,P. (2004) Gene-Ontology analysis reveals association of tissue-specific 5′ CpG-island genes with development and embryogenesis. *Hum. Mol. Genet.*, **13**, 1969–1978.