Published online 29 May 2009 Nucleic Acids Research, 2009, Vol. 37, No. 13 e92
doi:10.1093/nar/gkp423

Measuring spatial preferences at fine-scale resolution identifies known and novel *cis*-regulatory element candidates and functional motif-pair relationships

Ken Daigoro Yokoyama[1,2,*], Uwe Ohler[2,3,*] and Gregory A. Wray[1,2]

[1]Biology Department, [2]Institute for Genome Sciences and Policy and [3]Departments of Biostatistics and Bioinformatics and Computer Science; Duke University, Durham, NC 27708, USA

Received February 12, 2009; Revised May 5, 2009; Accepted May 8, 2009

ABSTRACT

Transcriptional regulation is mediated by the collective binding of proteins called transcription factors to *cis*-regulatory elements. A handful of factors are known to function at particular distances from the transcription start site, although the extent to which this occurs is not well understood. Spatial dependencies can also exist between pairs of binding motifs, facilitating factor-pair interactions. We sought to determine to what extent spatial preferences measured at high-scale resolution could be utilized to predict *cis*-regulatory elements as well as motif-pairs binding interacting proteins. We introduce the 'motif positional function' model which predicts spatial biases using regression analysis, differentiating noise from true position-specific overrepresentation at single-nucleotide resolution. Our method predicts 48 consensus motifs exhibiting positional enrichment within human promoters, including fourteen motifs without known binding partners. We then extend the model to analyze distance preferences between pairs of motifs. We find that motif-pairs binding interacting factors often co-occur preferentially at multiple distances, with intervals between preferred distances often corresponding to the turn of the DNA double-helix. This offers a novel means by which to predict sequence elements with a collective role in gene regulation.

INTRODUCTION

Transcriptional initiation is a major point of control for gene expression (1–3), and considerable effort has been devoted to deciphering the code by which transcriptional regulation occurs. Although this aspect of the genotype–phenotype connection is central to many fundamental biological processes, our understanding of how this mechanism operates at the molecular level is far from complete.

Understanding transcriptional regulation requires knowledge about the individual *cis*-regulatory elements that affect gene expression. Several methods have been proposed to predict individual transcription factor-binding sites by detecting statistically overrepresented motifs within the promoter (4–15). In the absence of functional data, however, overrepresentation of DNA sequence elements is not a sufficient criterion for functionality for two basic reasons. First, binding sites are generally short (5–10 bp), which means that many instances are present by chance rather than for functional reasons. Second, many motifs occur at increased frequency as a result of mutational bias or dinucleotide fluctuations near the start of transcription. The widespread overrepresentation of these motifs frequently dominates the subtle indicators of regulatory function, thus limiting the efficacy of these approaches.

Several recent studies have used spatial preferences as a criterion to predict *cis*-regulatory elements (16–20). With one notable exception (21), most previous studies have taken the 'sliding window' approach, which measures position-specific overrepresentation within several independent windows of pre-determined width (e.g. 20–25 bp). Although low-resolution approaches have been useful, an inherent limitation is that larger windows fail to recover positional enrichment at very precise locations, while smaller windows overlook general trends and are susceptible to random noise. The sliding window approach cannot operate at high resolution while simultaneously detecting broadly distributed signals, which limits the sensitivity of these methods. A second difficulty not

*To whom correspondence should be addressed. Tel: +1 919 668 6249; Fax: +1 919 660 7293; Email: kdy2@duke.edu
Correspondence may also be addressed Uwe Ohler. Tel: +1 919 668 5388; Fax: +1 919 668 0795; Email: uwe.ohler@duke.edu

addressed by previous studies is that the dinucleotide composition fluctuates dramatically near the start of transcription (19). This can greatly affect the frequency of motif occurrence in a position-specific manner, and raises the concern that some of the motifs predicted to exhibit positional specificity are not true *cis*-regulatory elements.

As a result of the limitations described above, relatively little information exists about positional biases in regulatory sequences. Important questions include whether such biases are common, how they are distributed around genes, which transcription factors (TFs) are involved, and whether they are associated with functional classes of genes. In this study, we report that a large number of likely regulatory elements exhibit position-specific overrepresentation relative to the transcription start site (TSS). The model presented here predicts regulatory elements by measuring positional enrichment at single base pair resolution while considering the data collectively; this approach allows us to detect both broad and narrow ranges of positional enrichment. By using regression analysis and a likelihood ratio test, the model can differentiate noise in the data from true position-specific overrepresentation. The method also accounts for position-specific dinucleotide fluctuations that exist within the promoter by incorporating a non-uniformly distributed background (null) model based upon the dinucleotide composition across the regulatory region. We show that this method can be used to predict novel potential *cis*-regulatory elements exhibiting previously unrecognized instances of positional biases on a genome-wide scale.

Since the model provides a general measure for spatial preferences, it is not limited to individual motif prediction but can also be extended to predict pairs of motifs binding interacting TFs using biases in separation distances. Transcription is not driven by individual proteins working in isolation, but is instead produced by cooperative interactions between multiple protein factors (3,22–24). Previous studies have shown that mutual relationships exist between various motifs, such as paired co-occurrences and relative orientations to the TSS (25–28). Such relationships have been effectively utilized in a variety of applications, such as the study of condition-specific and time-dependent gene expression patterns, gene network analyses, and promoter region detection (25,29–31). However, such studies are frequently limited to analyzing sequence element relationships between either known binding site motifs or those predicted using standard motif overrepresentation methods. The study presented here effectively circumvents this limitation by extending our model to analyze spatial relationships comprehensively across all motif-pairs. This allows us to predict pairs of sequence elements that are putatively bound by interacting TFs *de novo*, without any prior knowledge about the sequences of the predicted motifs. We find that binding sites of putatively interacting TFs frequently co-occur preferentially at multiple distances, with the interval between preferred distances corresponding approximately to the number of nucleotides in one turn of the DNA double-helix. This suggests a tendency for certain factor-pair interactions to occur in a particular orientation relative to the turn of the DNA molecule. We use the periodic phasing of inter-motif distance preferences to predict motif-pairs bound by interacting proteins, predicting functional binding site relationships between both known and novel sequence elements.

METHODS

The MPF model

Our goal is to predict regulatory motifs using spatial enrichment as an indicator of functionality. The method is an application of non-linear regression: given a set of observed motif occurrences within a set of promoter sequences, we collectively estimate the underlying frequency of occurrence according to location. Spatial biases are modeled using a continuous function which we denote as a 'motif-positional function' (MPF). The current study defines two types of MPFs. The first, denoted as a ''motif locational function' (MLF), measures position-specific overrepresentation of a given motif in reference to a landmark such as the TSS. The second model, denoted as a 'motif-relational function' (MRF), measures spatial preferences between pairs of motifs.

For a given motif w, its MLF $g_w(x)$ represents the underlying probability of occurrence according to its position x. Suppose our data set s consists of N sequences each of length L: $s = \{s_1, s_2, \ldots, s_N\}$, where $s_i = s_i(1) \ldots s_i(L)$. We consider these sequences to be the observed outcome of an underlying biological process, and define a random variable S analogous to a single sequence in s, where $S = S(1) \ldots S(L)$. We then define a random variable $U_k(j)$ to be the k-mer starting at position j in S: $U_k(j) = S(j)S(j+1) \ldots S(j+k-1)$. Our model then defines the MLF $g_w(x)$ to be

$$g_w(x) = \Pr(U_{l_w}(x+t) = w) \quad \mathbf{1}$$

where x represents the position of the motif, t represents the position of the TSS, and l_w represents the length of w. Note that the position x is given relative to t, and thus the location of the TSS is given by $x = 0$. The value of $g_w(x)$ represents, for any individual position x, the underlying probability of occurrence of w at this precise location. The values of this function are not normalized across the values of x, and therefore the sum of the values do not, in general, equal 1 across the promoter.

In contrast to an MLF, an MRF provides a measure of inter-motif distance preferences between two motifs. For any pair of motifs w and v, we define an MRF $f_{w|v}(x)$ to be the frequency of w to occur exactly x bp from v. Thus we set:

$$f_{w|v}(x) = \Pr(U_{l_w}(x+i) = w | U_{l_v}(i) = v) \quad \mathbf{2}$$

We note that the position of v, given by i, defines the position $x = 0$. The function $f_{w|v}(x)$ is independent of i; i.e. MRFs are defined as a conditional, rather than joint, probability.

Both MLFs and MRFs are modeled as the sum of a 'background function', $C(x)$, and a 'signal function', $H(x)$.

Thus, for any MPF $y(x)$ (i.e. $y(x)$ can represent either $g(x)$ or $f(x)$ as defined above), we have

$$y(x) = C(x) + H(x) \quad (3)$$

The background function $C(x)$ represents the background frequency of the motif, namely, the frequency without explicit positional bias. This function is allowed to fluctuate according to the dinucleotide makeup of the promoter, and is modeled as a polynomial (see below). In contrast, the signal function $H(x)$ incorporates possible spatial bias(es) into the model. The signal function of an MLF is modeled as a single unnormalized Gaussian term times a coefficient a:

$$H(x) = a \cdot \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (4)$$

Thus, $H(x)$ is designed to incorporate a 'peak' into the function $y(x)$. The parameters a, μ, and σ are free parameters, where a and μ give the height and location of the peak, respectively, while σ reflects the width of the peak.

Spatial biases are modeled using non-zero values for $H(x)$, while the model where $H(x) = 0$ (i.e. $a = 0$) assumes no positional bias. Spatial preferences are predicted by fitting each of these models to the data and comparing log-likelihoods using a likelihood ratio test (see the 'Model selection' section below).

The signal function of an MRF is extended from that of the MLF model in order to incorporate multiple peaks into the model, as we have found that many motif-pairs co-occur preferentially at multiple distances. The signal function of each MRF is therefore modeled using a linear combination of unnormalized Gaussian terms:

$$H(x) = \sum_{j=1}^{M} a_j \cdot \exp\left[-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right] \quad (5)$$

where M represents the number of Gaussian terms. This model is similar to that of an MLF, as MRFs for which $H(x) = 0$ (i.e. $M = 0$) assume no spatial preferences. In contrast, motif-pairs exhibiting spatial preferences are modeled using one or more Gaussian terms ($M > 0$). For pairs of motifs exhibiting spatial preferences, the value for M reflects the number of inter-motif distances at which a pair of motifs tend to co-occur preferentially.

Background functions

The background function $C(x)$ represents the background frequency at position x (i.e. the frequency of occurrence without explicit spatial bias). $C(x)$ is estimated using a 'prototype background function', $c(x)$, which represents the expected frequency of occurrence. This expected frequency is determined according to the dinucleotide composition at each position within the promoters. We distinguish between the background frequency $C(x)$ and the 'expected' frequency of occurrence $c(x)$, since many motifs are either over- or under-represented with respect to their dinucleotide makeup.

For MLFs, we model the underlying expected frequency using a polynomial function:

$$c_w(x) = \sum_{k=0}^{K} h_k x^k \quad (6)$$

This function is obtained by conducting linear regression on the set of data points $\{\langle x, R_w(x)\rangle\}$, where $R_w(x)$ represents the expected frequency of occurrence at position x. The value $R_w(x)$ is determined independently at each nucleotide site according to the observed dinuculeotide frequencies at that particular location. The function $c_w(x)$ then gives the underlying probability of occurrence after fitting a polynomial to this data set. The degree of this polynomial K is unique to each motif, and reflects the expected amount of fluctuation in occurrence frequency according to changes in dinucleotide content across the promoters.

Formally, $R_w(x)$ for a motif w of length l_w, i.e. $w = w(1)...w(l_w)$, is given by a position-specific 1st order Markov-dependency model as described in Karlin *et al.* (32). The expected frequency $R_w(x)$ of w at position x is given by

$$R_w(x) = \frac{\prod_{i=1}^{l_w-1} R_{w(i)w(i+1)}(x+i-1)}{\prod_{i=2}^{l_w-1} R_{w(i)}(x+i-1)} \quad (7)$$

where $R_{w(i)w(i+1)}(x)$ gives the observed frequency of the dinucleotide $w(i)w(i+1)$ at position x; $R_{w(i)}(x)$ represents the analogous mono-nucleotide frequency.

Since the expected frequency of many motifs differs from the actual frequency of occurrence, the background frequency $C_w(x)$ of a k-mer w is allowed to deviate from $c_w(x)$. Namely, we allow for uniformly distributed over- and under-representation. We therefore model $C_w(x)$ as

$$C_w(x) = b + d \cdot c_w(x) \quad (8)$$

where b and d are free parameters. Thus, the background model is allowed to 'shift' and 'stretch' vertically using parameters b and d, respectively; this allows for uniformly distributed differences in the expected and observed occurrence frequencies.

The prototype background function $c_{w|v}(x)$ for an MRF is a simple extension of that of the MLF model. In this case, $c_{w|v}(x)$ represents the expected probability for motif w to occur x bp away from motif v. This expected frequency is estimated according to the background functions of each individual motif. We assume the two motifs will occur randomly with respect to each other, and estimate $c_{w|v}(x)$ using the conditional probability provided by Equation 2:

$$c_{w|v}(x) = \frac{\Pr(U_{l_w}(x+i) = w, U_{l_v}(i) = v)}{\Pr(U_{l_v}(i) = v)} = \frac{\int_i C_w(x+i)C_v(i)di}{\int_i C_v(i)di} \quad (9)$$

The $C(\cdot)$ functions on the right-most part of the equation represent those of the estimated background functions for each corresponding motif. The background function

$C_{w|v}(x)$ of the MRF is then derived similarly to that of an MLF; i.e, $C_{w|v}(x) = b + d \cdot c_{w|v}(x)$.

Parameter estimation and statistical significance determination

As noted briefly above, parameter estimates (i.e. b, d, a, μ, σ) are obtained using likelihood-maximization. For the model assuming no positional specificity (i.e. $H(x) = 0$), only the parameters b and d must be estimated. In this case, since both b and d are linear parameters within the model, the likelihood can be maximized directly using linear regression. For models incorporating positional specificity, however, the parameters must be obtained through non-linear regression analysis. This is done by optimizing the log-likelihood $L(D;\ \theta_y)$ of the data D given the model y, where θ_y is the parameter vector of model y. Here, the data set D is given by: $D = \{\langle x_1,z_1\rangle,...,\langle x_n,z_n\rangle\}$, where z_i represents the number of motif occurrences at position x_i. As each MPF $y(x)$ represents the probability of motif occurrence at x, the log-likelihood of a single data point $L(\langle x_i,z_i\rangle;\theta_y)$ reflects the outcome of multiple Bernoulli trials with a 'success' being an occurrence of the motif at position x_i. Thus, the log-likelihood of this data point is given by the binomial distribution:

$$L(\langle x_i,z_i\rangle;\theta_y) \equiv z_i \cdot \log[y(x_i)] + (N_i - z_i) \cdot \log[1 - y(x_i)] \quad \mathbf{10}$$

where N_i is the number of 'trials'; i.e. the maximum possible value for z_i. The total log-likelihood $L(D;\theta_y)$ of the data is given by the sum of the log-likelihoods across all data points $\langle x_i,z_i\rangle$. This value is maximized using an iterative method called 'Broyden's method' (33) given an initial parameter estimate θ_0; interested readers are referred to (33–36). Several initial parameter vectors are used during each MPF estimation; the final parameter estimates are taken to be those producing the highest log-likelihood. The method by which the initial parameter vectors are determined as well as the high level of robustness for the parameter estimates is discussed in Supplementary Data S1.

MODEL SELECTION

Model selection for any given MPF involves determining both the degree K of the prototype background function as well as the number of Gaussian terms within the signal function (either 0 or 1 for MLFs, or any non-negative number M for MRFs). These are determined in the same manner; namely, we use a likelihood ratio test (F-test) to compare the log-likelihoods of the data given two possible models. To determine the presence or absence of positional enrichment using the MLF method, we compare the log-likelihood derived from the (null) model, where $H(x)$ is identically zero [i.e. $a = 0$ in Equation 4], to that of the (alternative) model where $H(x)$ takes on non-zero values ($a \neq 0$). Model selection involves comparing the log-likelihoods $L(D;\ \theta_{yA})$ and $L(D;\ \theta_{y0})$, where the MPF y_A allows for positional specificity, while its nested null model y_0 assumes no positional enrichment. The 'scaled deviance' $Z(\theta_{y_A},\theta_{y_0})$ given by

$$Z(\theta_{y_A},\theta_{y_0}) = 2 \cdot \left[L(D;\ \theta_{y_A}) - L(D;\ \theta_{y_0})\right] \quad \mathbf{11}$$

follows a x^2 distribution with $|\theta_{y_A}| - |\theta_{y_0}|$ degrees of freedom (35); see Supplementary Data S2 for discussion.

Our final statistic is

$$F = \frac{Z(\theta_{y_A},\theta_{y_0}) \cdot (n - |\theta_{y_A}|)}{Z(\theta_S,\theta_{y_A}) \cdot (|\theta_{y_A}| - |\theta_{y_0}|)} \quad \mathbf{12}$$

where n is the number of data points and model S is the 'saturated model', i.e. the model optimizing the log-likelihood at each data point without limits on the number of parameters. The value F follows the F-distribution with $|\theta_{y_A}| - |\theta_{y_0}|$ and $n - |\theta_{y_A}|$ degrees of freedom (35); P-values reflecting the significance of spatial enrichment are derived using this statistic.

The number of Gaussian terms M within an MRF are determined similarly. However, as the number of Gaussian terms can be larger than 1, the value of M is determined in an iterative fashion. Namely, we begin by comparing the model where $M = 0$ to the model where $M = 1$ and conduct the F-test in a similar manner as described above. Note that this is equivalent to the single model comparison for an MLF, as we are determining the presence or absence of a single Gaussian term. For motif-pairs producing significant P-values, we proceed to increment the value of M (i.e. comparing models for which $M = 1$ to that where $M = 2$, then $M = 2$ versus $M = 3$, etc.) until the P-value produced from the F-test is no longer significant. The final value of M is taken to be the last value of M that has produced a significant P-value.

Determining the order K of the polynomial $c(x)$ is also determined using an F-test in an incremental fashion. As the function $c(x)$ is derived using linear regression, the F statistic is obtained by comparing the sum of the squares for each of two (consecutive) models. Namely, if SS_y represents the sum of the squares produced from fitting model y to the data points, the F statistic is given by

$$F = \frac{(SS_{y_0} - SS_{y_A}) \cdot (n - K_A - 1)}{SS_{y_A}} \quad \mathbf{13}$$

where the degree of polynomial y_A is K_A, and the degree of the nested polynomial y_0 is $K_A - 1$. This value also follows the F-distribution with 1 and $n - K_A - 1$ degrees of freedom (35). In a manner similar to the estimation of M, we increase the value of K incrementally until the comparison no longer produces a significant P-value; the final value of K is then taken to be the last value of K that produces a significant P-value.

Motif clustering procedure (MLFs)

For MLF clustering analyses, 6-mer motifs are clustered for redundancy according to both sequence similarity as well as the position and width of enrichment. Clustering is conducted by considering motifs in rank order; at each step, an individual 6-mer motif is either placed in an existing cluster or else a new cluster is created. 6-mers matching at five of the six sites (i.e. containing only one mismatch, or no mismatches with a single bp offset) are clustered if their signal functions are similar according to their KL divergence (37). The KL divergence between two signal functions is calculated by converting each function into a discrete probability distribution $p(x)$ across each position within the promoter:

$$p(x) = \frac{H(x)}{\sum_i H(x_i)} \quad \textbf{14}$$

where the values for x are shifted according to any offset between the two motif sequences. Values of $H(x)$ are buffered by a minimum value of $1e$–45 to prevent extreme KL divergence values (i.e. 0 or infinity). The KL divergence V for two distributions, $p_{w1}(x)$ and $p_{w2}(x)$, is calculated to be

$$V = \sum\nolimits_x p_{w_1}(x) \cdot \log\left[\frac{p_{w_1}(x)}{p_{w_2}(x)}\right] \quad \textbf{15}$$

The V-value threshold was set to 0.2 during our analysis; motif-pairs with similar sequences were clustered if their KL divergence fell below this threshold.

Consensus sequence determination and known cis-regulatory motif comparisons

Motif clusters are condensed into a single consensus sequence according to the criteria derived from (38) and (39). Namely, each aligned site is assigned a single residue consensus if it comprises 50% of the aligned k-mers and occurs at least twice as frequently as every other nucleotide type. Double nucleotide degeneracy is applied to sites for which the two residues comprise 75% of the cases, with neither residue matching the criteria for a single site consensus. Sites not matching the criteria for either single or double nucleotide degeneracy are considered completely degenerate; triple degeneracy is not considered. During our analyses, comparisons to known regulatory elements in TRANSFAC v11.3 (39) were conducted using STAMP (40); only binding motifs found in humans were considered.

Data preparation

DNA sequences used during our analyses were taken from the UCSC Table Browser (http://genome.ucsc.edu) (41). Human analyses were conducted using the promoter sequences from the hg18, Build 36.1 assembly (42); mouse promoter data was taken from the mm9, NCBI Build 37 (43). Both data sets contained sequences comprising 500-bp upstream and 100-bp downstream of a known TSS in RefSeq (44,45). Sequence-pairs with at least 500 matching sites were filtered from the data sets. Genes without 5′ UTR annotations were excluded in order to eliminate TSS annotations caused by incomplete mRNA transcripts. The final data sets comprised a total of 20 609 non-redundant human promoters and 18 354 mouse promoters.

Program availability

The analyses presented here were conducted using the 'Functional Region Evaluation Engine' (FREE). Implementation of the program involves an initial formatting step; any number of subsequent analyses can be conducted after the formatting step is completed. The executable file is freely available at (http://www.biology.duke.edu/wraylab/); instructions for usage as well as an overview of user-defined parameters are included.

RESULTS

Motif Locational Functions (MLFs) provide a measure of positional enrichment

The MLF model presented here provides a measure of the amount of position-specific enrichment for any given motif. In contrast to previous models, our methodology is conducted at single-site resolution using regression analysis, allowing the data to be considered collectively across each position. MLFs are modeled using a continuous function $g(x)$, whose values represent the underlying probability of occurrence according to position x. This function is given as the sum of the background frequency $C(x)$ and a contribution of position-specific overrepresentation $H(x)$; thus $g(x) = C(x) + H(x)$. Positional bias, represented by $H(x)$, is modeled using a Gaussian term, incorporating a 'peak' into the MLF (Equation 4). The mean (μ) of the Gaussian term represents the central location of enrichment, while the standard deviation (σ) reflects the 'width' of this enrichment. This model is illustrated in Figure 1a which shows the MLFs of the TBP and SP1-binding motifs (TATA-box and GC box). These motifs are found overrepresented 30 and 65 bp prior to the TSS, respectively ($\mu = -29.6$ and $\mu = -65.3$). The TATA-box is found overrepresented only at a few sites within the promoter, thus producing a small σ-value ($\sigma = 1.9$), while the GC box is enriched across a much broader range upstream of the promoter ($\sigma = 52.2$).

For a given motif, we predict spatial preferences using a likelihood ratio test (F-test). Namely, we compare the model assuming no positional specificity to that allowing for spatial enrichment. The former is modeled by setting $H(x)$ to zero, and thus the underlying frequency of occurrence $g(x)$ simply equals the background frequency $C(x)$. We compare this model to the one where $H(x)$ takes on non-zero values, allowing for position-specific overrepresentation. Comparing the log-likelihoods given each of these two models produces a P-value reflecting the significance of positional enrichment. Positional specificity is then predicted for motifs for which this P-value falls below a given threshold.

Our model accounts for fluctuations in dinucleotide frequencies across the promoter by allowing the values of the background frequency $C(x)$ to vary according to position. For instance, as GC content rises near to the start of transcription, the background frequency $C(x)$ of GC-rich motifs likewise increases close to the TSS. In order to determine the background frequency, we first estimate each motif's expected frequency across the sequences.

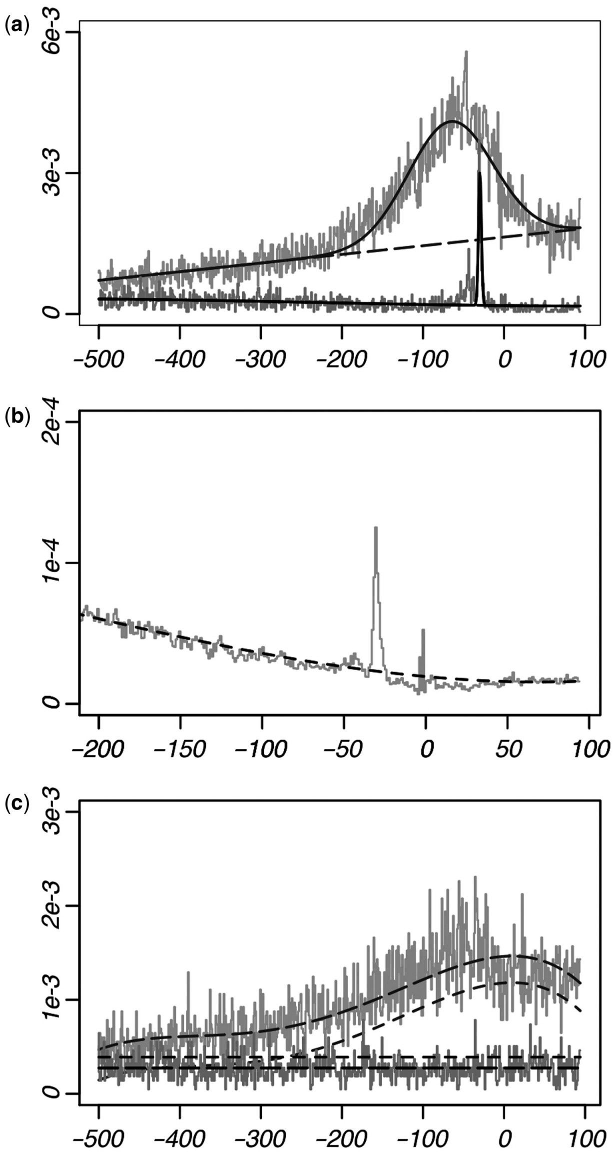

Figure 1. Raw data and MLFs of four example motifs. x-axis values denote the position within the promoter, where $x = 0$ represents the location of the TSS. The y-axis represents the frequency of occurrence. Solid plots represent the resulting MLFs ($g(x)$), long dashes show the background frequency of occurrence $C(x)$, and short dashes indicate the expected frequencies derived from dinucleotide composition ($c(x)$). (**a**) MLFs of the SP1 (gray) and TBP (black)-binding sites; significant amounts of positional enrichment are predicted for both. (**b**) Expected frequencies of the TBP-binding site. Each data point is derived according to the dinucleotide composition at each position. Note that $c(x)$ was designed to not incorporate sharp increases observed in the dinucleotide data, as such rises are often a byproduct of the overrepresentation of the motif itself. (**c**) Two motifs (GGGCGC, gray; TGCTTC, black) without positional enrichment. Note that without positional enrichment ($H(x) = 0$), the MLF $g(x)$ is the same as the background frequency $C(x)$. A comparison between $C(x)$ and $c(x)$ illustrates the ability of the background model to account for uniformly distributed over- and under-representation with respect to the expected frequency (according to dinucleotide composition). The gray plot shows a high amount of fluctuation in the promoter, although this is attributed to the dinucleotide makeup of the promoter [note $c(x)$]; this motif is therefore not predicted to exhibit positional enrichment.

This expected frequency is denoted as $c(x)$. The values of $c(x)$ are allowed to vary by position according to the dinucleotide makeup of the regulatory region, thus accounting for position-specific changes in dinucleotide composition. However, we distinguish between the 'background' and 'expected' frequency of occurrence [given by $C(x)$ and $c(x)$, respectively], as many motifs are either over- or under-represented with respect to their dinucleotide composition. Thus, the background frequency $C(x)$ allows for uniformly distributed over- or under-representation. Both $c(x)$ and $C(x)$ are important components of our model. Namely, we must allow for differences between the expected and observed frequency of occurrence while still incorporating dinucleotide fluctuations into the background frequency of occurrence. Thus, although the background frequency is allowed to deviate from the expected frequency, $C(x)$ is restrained to mimic the 'shape' of $c(x)$ in order to preserve the expected fluctuations according to dinucleotide composition. For instance, for motifs whose frequencies are expected to vary according to position, rises and drops in $C(x)$ are restricted to conform to those of $c(x)$. In contrast, motifs expected to occur at a constant frequency across the region [i.e. the values of $c(x)$ are uniform across all positions] likewise have a constant value for $C(x)$.

We model $c(x)$ in a continuous fashion (Equation 6). This function is determined by conducting linear regression on the set of data points representing the expected frequency of occurrence at each site. Fitting the function $c(x)$ to these data points then gives the underlying (expected) frequency of occurrence. An example is illustrated in Figure 1b, which shows the raw data and resulting function $c(x)$ for the TATA-box. We note that sharp rises in the observed dinucleotide frequencies at a particular location are not directly incorporated into $c(x)$, but instead remain outlier points after fitting this function to the data. This is an important aspect of our model, as overrepresentation of a motif can itself cause rises in dinucleotide frequency. Incorporation of such rises into $c(x)$ would therefore obscure the distinction between the signal and the background frequency at this location; this is discussed later in the Results section.

The ability of the background model to incorporate uniformly distributed over- and under-representation is illustrated in Figure 1c. Here, we show the MLFs for two motifs that do not exhibit positional enrichment. We note that although the gray plot fluctuates according to position, this would be expected according to the dinucleotide frequencies within the promoter [note the fluctuations in $c(x)$]. Thus, this motif is not predicted to exhibit biologically relevant positional specificity using our model.

The MLF method predicts position-specific over-representation for many motifs within human promoters

In order to determine which motifs exhibit positional enrichment within human promoters, we analyzed spatial enrichment for all 6-mers on a set of non-redundant RefSeq human promoters (44,45) collected from the UCSC Genome Browser (http://genome.ucsc.edu)

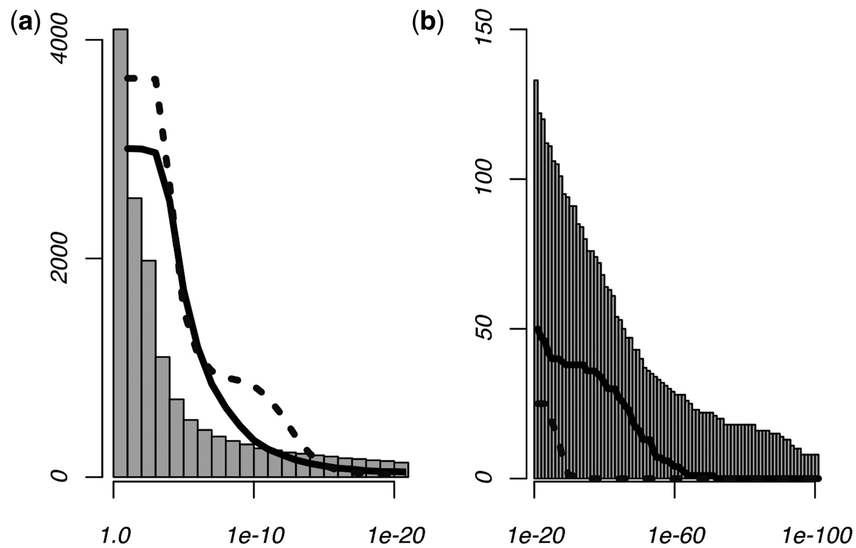

Figure 2. Results from the comprehensive MLF analysis. Histograms show the cumulative number of 6-mers with positional enrichment according to their *P*-values in human promoters. The plots give the number of 6-mers producing P_{sim}-values under the given thresholds during simulation analyses, where P_{sim} represents the most significant *P*-value for each individual 6-mer across 100 simulated data sets. Solid plots refer to simulations conducted according to the dinucleotide frequencies across each position within the promoter, while the dotted lines represent those generated using mono-nucleotide frequencies. *P*-value thresholds above 1*e*–20 are shown in (**a**), while the contrast between the results of the human and simulated analyses for which $P < 1e - 20$ is illustrated in (**b**). Note that the dinucleotide-generated simulated data sets produced a significantly larger number of predictions than the mono-nucleotide-generated simulated data, while the real human promoters produce more predictions than either of the control data sets.

(41,42). The data set consisted of 20 609 sequences, each comprising the region 500-bp upstream and 100-bp downstream of a known TSS. As expected, the vast majority of the motifs did not exhibit positional enrichment within the promoter data set. However, a few motifs showed highly significant position-specific overrepresentation, with 106 6-mers exhibiting spatial bias at a significance level of $P < 1e–25$.

To compare these results to those of a control data set, we repeated the analysis on a set of intergenic sequences, each comprising the 600-bp interval starting 2-kb upstream of a known TSS. Very few motifs were predicted to exhibit positional enrichment in this control data set, with less than 1% producing *P*-values under 1*e*–5. We then tested our model on two types of simulated data sets. The first was generated by considering the observed mono-nucleotide frequencies at each site, while the second was produced using dinucleotide frequencies at each position. One hundred data sets of both the mono- and di-nucleotide simulations were generated, with each individual data set comprising the same number of sequences as the human promoter data. For each type of simulation, we scanned for positional enrichment across all 100 data sets, recording the most significant *P*-value for each individual 6-mer. This *P*-value, denoted as P_{sim}, was thus unique to each 6-mer. Results of these analyses, as well as those for the human promoter analysis, are shown in Figure 2. Significantly more predictions were made during the human promoter analysis than either of the simulation analyses. We also note that the dinucleotide-generated data sets produced more predictions than those produced using only mono-nucleotide frequencies, suggesting that our use of a dinucleotide-based background model leads to a more conservative significance criteria.

We used the results of both the intergenic sequence analysis as well as those of the simulated data sets to set the prediction criteria for positional specificity within human promoters. The lowest *P*-value produced from the intergenic sequence analysis was found slightly under 1*e*–15; motifs above this threshold were excluded from the list of predictions. The remaining motifs producing *P*-values under their P_{sim}-value times a stringent multiple hypothesis correction factor of 1*e*–5 were then predicted to exhibit positional enrichment in the human RefSeq promoters. Thus, the prediction criteria ($p < P_{sim} \times 1e–5$) was unique to each 6-mer, subjecting motifs with lower *P*-values within the simulated data sets to a more stringent threshold.

The final list of predictions contained 166 6-mer motifs, representing 4% of the total number of possible 6-mers. Despite our stringent prediction criteria, the majority of the top-ranked motifs from the RefSeq promoter analysis were not filtered due to the simulation analysis. Out of the 50 top-ranked motifs exhibiting spatial bias within the real promoters, only one motif did not pass the *P*-value threshold determined as above.

Motif clustering predicts locational overrepresentation for both known and putatively novel *cis*-regulatory elements

We used the list of positionally enriched 6-mers to generate consensus motifs with degenerate sites and flexible lengths. Motifs were clustered computationally according to sequence similarity as well as the location and width of their positional enrichment. We then condensed each cluster into a single consensus sequence, generating a total of 48 consensus motifs exhibiting position-specific enrichment within human promoters. In order to test whether

Table 1. Positionally enriched motifs in human promoters

Location-specific motif clusters												
Human RefSeq data						Mouse RefSeq data			Previous studies			
Rank	p	TF	Consensus	μ	(σ)	Consensus	μ	(σ)	Fitz	Xi	Xie	Vard
1	7e-179	SP1(+)	AGGGGGCGGGG	-68.3	(52.2)	GRGGGGGGCGKG	-69.6	(42.8)	*	*	*	*
2	2e-106	NFY(-)	CTSATTGGCT	-78.8	(42.7)	ATTGGC	-100.0	(16.1)	*	*	*	*
3	1e-102	CREB	CGTGACGTC	-49.1	(39.0)	GTGACG	-44.5	(34.6)	*	*	*	*
4	3e-102	ZEB1(-)	CAGGTAAG	72.5	(31.6)	GGTAAG	71.6	(33.9)	*	--	--	*
5	5e-96	YY1	GATGGCGG	31.9	(22.1)	TGGCGG	23.8	(16.7)	*	*	--	*
6	6e-94	NFY(+)	AGCCAATCAG	-76.7	(40.7)	GCCAAT	-91.0	(21.9)	*	*	*	*
7	5e-91	*d*1	GTGAGTG	69.2	(36.4)	GTGAGTG	70.1	(32.6)	*	--	--	*
8	3e-90	NHLH1	CAGCGGCKGC	33.0	(40.9)	RGCGGCG	32.6	(44.4)	--	--	--	*
9	3e-87	SP1(-)	CGCCCC	*35.0	(32.2)	GCCCCC	-66.3	(33.7)	*	*	--	*
10	2e-83	ETS(+)	ACCGGAAGTG	-25.9	(32.3)	GCCGGAAGTG	-33.5	(37.0)	*	*	*	*
11	9e-83	TBP	ATATAAAWR	-30.6	(1.9)	ATATAAARGC	-30.9	(1.7)	*	*	*	--
12	4e-74	SP1(-)	GCCCCKCCCC	-76.2	(45.0)	SCYCCKCCCC	-78.7	(51.7)	*	*	*	*
13	7e-65	REST	CRCCATGGA	52.8	(38.0)	CGCCATGGCY	50.4	(34.9)	*	--	--	*
14	2e-58	ETS(-)	CACTTCCGGT	-24.3	(32.2)	CTTCCGG	-16.5	(16.0)	*	*	*	*
15	1e-54	HBP1	RCGTCAC	-47.0	(37.4)	CGTCAC	-53.2	(39.5)	--	--	*	*
16	3e-53	ZFP161	GCGCGC	-51.8	(95.0)	CGCGCGC	-32.6	(97.1)	--	--	--	--
17	1e-50	*d*2	TCTGCTGCT	51.0	(33.5)	CTGCTGCY	53.1	(37.0)	*	--	--	--
18	2e-48	YY1	CAAGATGG	22.9	(17.1)	CAAGATGG	14.5	(10.7)	*	*	--	--
19	3e-46	*d*3	TTTTTT	-12.7	(11.3)	--	--	--	--	--	--	--
20	3e-45	TBP	TWTATA	-29.9	(2.0)	ATATAW	-27.9	(1.8)	*	*	*	--
21	5e-44	NRF1	RTGCGCA	-53.7	(59.8)	TGCGCA	-57.8	(46.8)	--	*	--	*
22	5e-40	NRF1	GCGCATGC	-46.9	(38.0)	--	--	--	--	*	*	*
23	9e-39	Inr	GCTCAGTCC	-4.0	(0.2)	TCAGTC	-2.2	(0.5)	--	--	--	--
24	5e-37	MYC	CACGTG	-51.0	(50.7)	CACGTG	-53.3	(46.1)	*	--	*	*
25	7e-35	ZIC2	CCCACCC	-131.0	(70.2)	†CCCCCC	-117.6	(99.9)	--	--	--	--
26	1e-32	*d*4	TCCTCCT	-71.4	(82.9)	†CCCTCC	-61.9	(32.8)	--	--	--	--
27	8e-32	*d*5	GTGTGT	-325.6	(234.4)	TGTGTGT	-435.8	(212.5)	--	--	--	--
28	1e-25	TBP	AAAAGG	-27.3	(1.3)	--	--	--	*	--	--	--
29	2e-25	SRF	ATGGCC	53.6	(33.9)	GATGGC	26.9	(20.1)	--	--	--	--
30	5e-23	SOX9	CAATGG	-80.1	(23.9)	WCCAATGR	-85.7	(40.1)	--	--	--	--
31	2e-21	*d*6	GGCGTG	-62.5	(34.1)	--	--	--	--	--	*	--
32	2e-21	GTF2IRD1	CTCCCTC	-111.0	(100.6)	†CCCTCC	-61.9	(32.8)	--	--	--	--
33	3e-21	*d*7	AAAAAA	-165.0	(10.2)	--	--	--	--	--	--	--
34	2e-20	MEF2	AAAAAT	77.3	(23.1)	AAAATA	202.3	(78.0)	--	--	--	--
35	4e-20	*d*8	GCGCTC	-120.6	(174.9)	--	--	--	--	--	--	--
36	7e-20	*d*9	GCAGCA	47.5	(36.0)	GCAGCA	28.6	(15.4)	*	--	--	--
37	1e-18	Inr	CAGTTG	-1.2	(0.5)	†TCAGTC	-2.2	(0.5)	--	--	--	--
38	2e-18	Inr	GTCACT	-3.0	(0.1)	--	--	--	--	--	--	--
39	2e-18	*d*10	ACACACA	-12.6	(23.7)	--	--	--	--	--	--	--
40	3e-18	TBP	TAAAAA	-27.8	(0.9)	†TAAATAG	-28.8	(1.7)	--	--	--	--
41	6e-18	*d*11	AAGAAG	96.5	(55.5)	†GAAGGT	54.4	(38.3)	--	--	--	--
42	2e-17	TRIM63	TCACTT	-1.9	(0.5)	CACTTC	-1.0	(0.3)	--	--	--	--
43	3e-17	*d*12	AGTGCT	-529.4	(165.1)	--	--	--	--	--	--	--
44	8e-17	TBP	AAAAGC	-26.9	(0.9)	ATATAAARGC	-29.9	(1.7)	--	--	--	--
45	1e-16	Inr	CAGTGC	-1.0	(0.2)	--	--	--	--	--	--	--
46	2e-16	*d*13	GGACCC	78.7	(27.8)	GGACCC	102.1	(46.4)	--	--	--	--
47	3e-16	*d*14	GAGCCG	37.7	(36.2)	--	--	--	--	--	--	--
48	6e-16	PDX1	GTCATT	-3.0	(0.5)	--	--	--	--	--	--	--

The location (μ) and width (σ) of enrichment are given to the right of each cluster. *P*-values given on the left pertain to the most significant 6-mer within the cluster. The third column shows factor names-binding to the known regulatory elements in TRANSFAC (39); putatively novel motifs are labeled d1–d14. Motifs found spatially enriched in mouse promoters are given to the right of the human analysis results. The right columns show comparisons to previous studies using the 'sliding window method' (18,19,47,48). Asterisks denote matches to non-redundant consensus motifs produced by these studies after *k*-mer clustering; only motifs predicted to be enriched at approximately the same location were considered matches. All sequence matches to TRANSFAC, mouse motif predictions, and those of previous studies were conducted using STAMP (40) (*E*-value threshold: 1*e*–6).
†Denotes a weak match.

the predicted motifs overlapped with known regulatory elements, we compared our results to known TF-binding sites in the TRANSFAC database (39) using STAMP (40). Thirty-four of the motif clusters matched known *cis*-regulatory elements, comprising a total of twenty known binding sites within TRANSFAC as well as the Inr sequence element (Table 1). Several of the motifs predicted were previously known to exhibit position-specific over--representation, including the TBP, SP1, NFY, CREB, ETS, NRF1 and MYC factor-binding sites (19,46–48). We also predicted several additional motifs whose positional enrichment had not been previously documented,

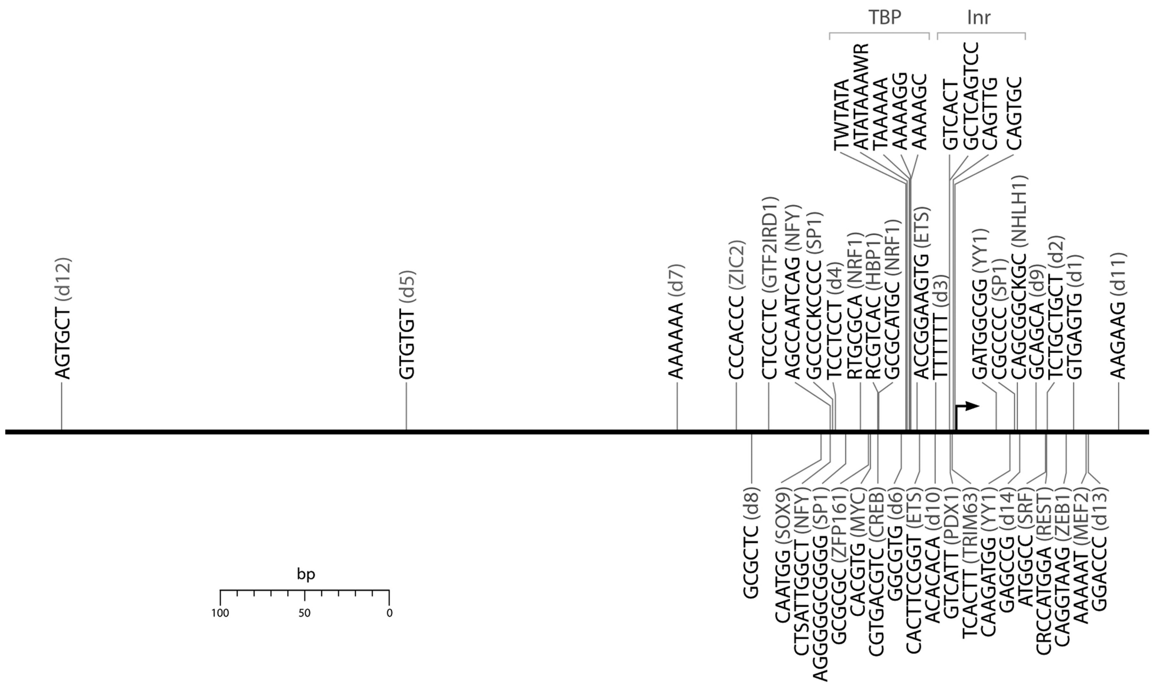

Figure 3. Location of enrichment for the 48 MLF motif predictions within human promoters.

including fourteen novel regulatory motif candidates, denoted as d1-d14. The location of enrichment for each of the predicted motif clusters is illustrated in Figure 3. Most of these motifs were found to be enriched close to the TSS, although a few were found farther upstream of the promoter. Motifs enriched far from the promoter were frequently found to be overrepresented over a large range of the regulatory region as shown in Table 1. This is to be expected, as it is unlikely that a regulatory element enriched far from the promoter would be constrained to a highly specific location. We note the precision of the method to predict related clusters at the same location, such as the TBP-binding site as well as the Inr sequence clusters.

Figure 4 shows the MLFs for six motifs with positional enrichment within the promoter. The MLFs of the GC-rich NHLH1 and ZFP161-binding sites are shown at the top of the figure. We note that the rise of GC content centers directly across the TSS, as indicated by the simulated data plots. However, the positional enrichment for each motif is found at other locations ($\mu = +33$ and -51, respectively), indicating that the positional bias of these motifs is not due to dinucleotide fluctuations within the promoter. The putatively novel d3 motif comprises a homopolymeric thymine tract. Such poly(T) sequences are known to alter DNA conformation, thereby affecting transcriptional regulation by displacing the nucleosome from the DNA molecule (49–51). Similarly, the novel d10 motif, comprising a CA-dinucleotide repeat, promotes left-handed Z-DNA conformations (52–54). The positional biases of these motifs may therefore reflect a functional role for each motif at these locations. The MLFs of the novel reverse complement motifs d2 and d9 are shown at the bottom of Figure 4; each orientation of this putatively novel regulatory element show positional enrichment at the same location downstream of the TSS.

Many positionally biased motifs are shared between human and mouse

We tested whether spatially biased motifs found in human promoters would also show positional enrichment within mouse promoters. We conducted a second comprehensive MLF analysis using a sequence data set of 18 354 non-redundant mouse promoters in RefSeq (43–45). We then compared the motif predictions between the two species according to sequence similarity as well as the location of positional overrepresentation.

Our analysis predicted a total of 49 consensus motifs to exhibit positional enrichment within mouse promoters (Supplementary Data S4). Comparisons of these results to those of the human promoter analysis showed a very significant amount of overlap between motif predictions across the two species. We found that 36 (75%) of the motif clusters identified in the human data set matched spatially biased motifs detected within the mouse promoters (Table 1). Such a significant overlap provides confidence in our new motif predictions, as these motifs were predicted during independent analyses using data from two highly diverged species. In addition, the location of enrichment for our motif predictions was often found to be highly conserved between the two lineages. Many motifs with well-documented positional enrichment were found overrepresented at very similar locations across the two species, particularly the TBP, SP1, NRF1 and CREB-binding sites. This was also found to be the case for many of our novel motif predictions. For instance, the novel d1

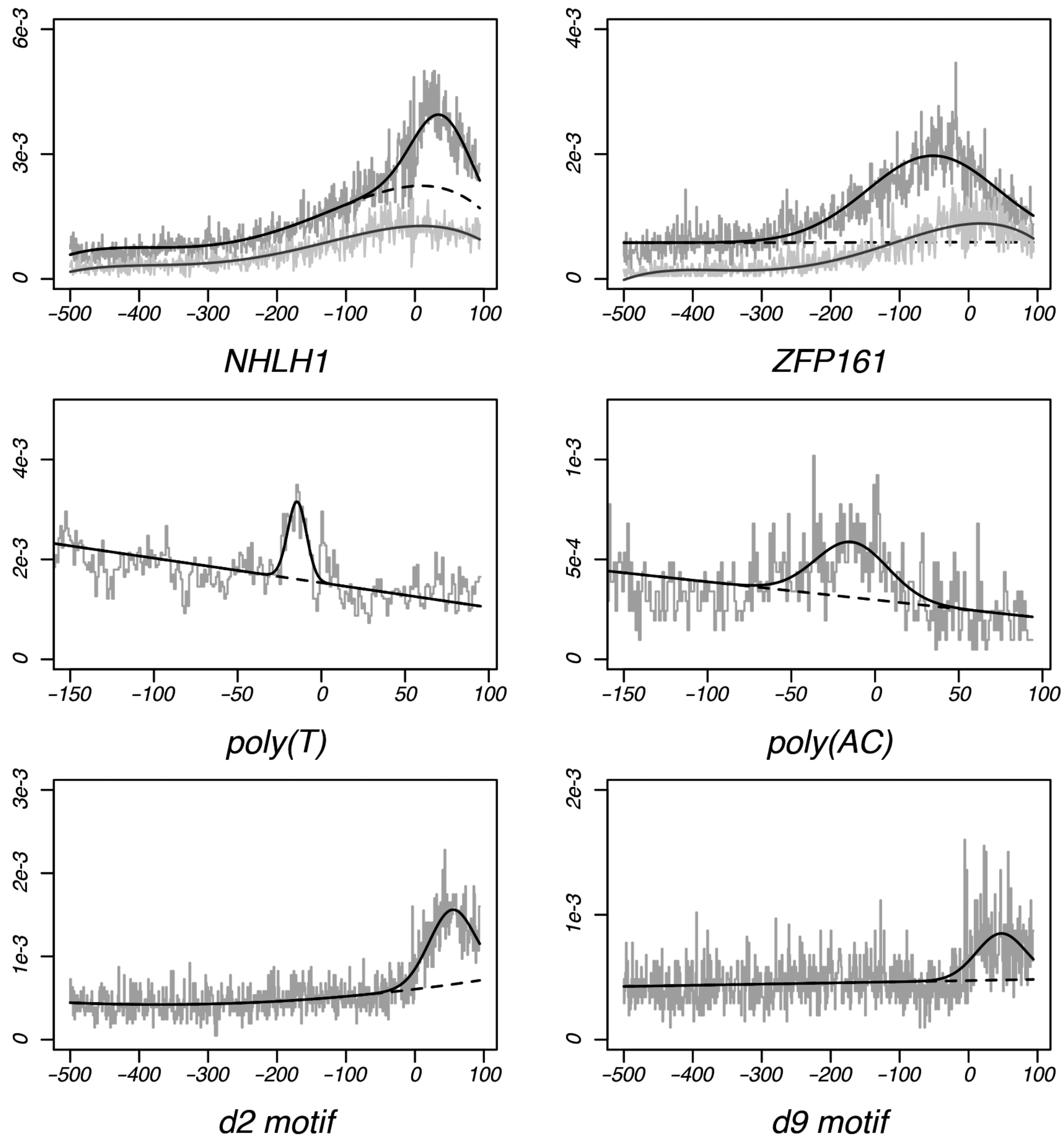

Figure 4. MLFs of six motifs exhibiting positional enrichment. Top: MLFs for the GC-rich NHLH1-binding site and the ZFP161-binding site. Each plot shows results for both human (black) and simulated (gray) data sets. Dashed lines denote the background functions $C(x)$. Middle: MLFs of the novel poly(T) 5-mer d3 and $(AC)_3$ motif d10. Bottom: MLFs for the GCT-repeat motif d2 and its reverse complement d9. Each shows significant amounts of positional enrichment ~50 bp after the TSS.

and d2 motifs exhibited overrepresentation peaks whose position differed by only 1 bp across the two lineages.

Study comparisons highlight differences in methodologies to previous studies

Several previous studies have analyzed spatial preferences of potential regulatory motifs within the promoter (17–21,47,48). Most previous analyses, with one exception (21), have used the 'sliding window' approach. In this approach, the promoter region is divided into several discrete bins of pre-determined width (e.g. 20–25 bp), and positional bias is then predicted by comparing the number of motif occurrences in each window to a background frequency of occurrence. A previous study conducted by FitzGerald *et al.* (19) used the sliding window approach, considering motif occurrences within separate windows of 20 bp. FitzGerald *et al.* predicted a total of 156 8-mers to exhibit positional bias prior to clustering. A direct comparison of our results to those of FitzGerald *et al.* showed that 97% of the 8-mers predicted by FitzGerald *et al.* matched one of our predicted 6-mers (Table 2). We also found that 85% of our individual 6-mers matched a 8-mer prediction made by FitzGerald *et al.* However, the vast majority of these matches were to redundant motifs that had been grouped according to sequence similarity during the clustering analysis. There were also cases in which distinct 6-mers found within different cluster groups matched a single 8-mer predicted by FitzGerald *et al*. For instance, one of the G-rich 8-mers predicted by FitzGerald *et al*. matched eight of our predicted 6-mers, although this group of 6-mers included representatives from three different motif clusters. These 6-mers clearly represented distinct regulatory elements, as their enrichment was found at significantly different locations within the regulatory region.

Thus, we looked to compare the non-redundant consensus sequences produced by both studies after clustering.

Table 2. Between-studies comparisons of positionally enriched kmer predictions

		Fitz	Thara	Vard	MPF
Positionally enriched motifs: study comparisons					
Predictions	Number	156	1226	168	166
	kmer length	8	8	7	6
GC content	Fraction	63%	69%	60%	60%
GC rich	Number	28	387	19	48
	Expected	6	43	11	18
	(Numb/Exp)	4.7	9.0	1.7	2.7
AT rich	Number	3	39	0	16
	Expected	6	43	11	18
	(Numb/Exp)	0.5	0.9	0.0	0.9
FitzGerald *et al.*	Matches	—	521	60	141
	Fraction		42%	36%	85%
Tharakaraman *et al.*	Matches	149	—	101	156
	Fraction	95%		60%	94%
Vardhanabhuti *et al.*	Matches	125	507	—	103
	Fraction	80%	41%		62%
MPF	Matches	151	1004	84	—
	Fraction	97%	82%	50%	

Spatially enriched k-mers are compared between studies conducted by FitzGerald *et al.* (19), Tharakaraman *et al.* (21) and Vardhanabhuti *et al.* (18) as well as the MPF model. The total number of (unclustered) k-mer predictions are shown in the top row. The number of GC and AT rich motif predictions (those composed of G/C or A/T consensus sites at all but one site) are shown below, along with the expected number and the ratio of actual/expected. Here, the 'expected' number of predictions was determined by assuming a random collection of k-mers identical in size to the set of predictions. Bottom rows show the amount of overlap between predictions across the four studies. Overlapping predictions were determined by considering all consensus sites of the predicted motifs, allowing for any offset such that at most one consensus site of the smaller motif was not aligned to the larger k-mer. For instance, 149 (95%) of the 156 motif predictions made by FitzGerald *et al.* matched a prediction made by Tharakaraman *et al.* Note that the number of matches is not symmetrical, since a single *k*-mer may match more than one other motif prediction.

Motif clustering conducted by FitzGerald *et al*. resulted in nine non-redundant motif clusters. Eight of these clusters overlapped with one of our consensus motif cluster predictions, while our model attributed the putative spatial bias of the remaining cluster to dinucleotide fluctuations within the promoter. In contrast, less than half of our consensus sequences were detected by FitzGerald *et al*. Table 1 contains comparisons between our regulatory motif predictions to those of FitzGerald *et al*. as well as three other studies providing non-redundant motifs with spatial enrichment (18,47,48). We found that many of our motifs predicted with wider ranges of positional enrichment could not be detected using the sliding window approach. Our approach was also found to increase sensitivity to spatial biases occurring at very precise locations. For instance, FitzGerald *et al*., in addition to the three other studies included in Table 1, could not easily detect the well-known Inr sequence element. The Inr sequence has been previously characterized by the consensus motif YYAnWYY (55). This element is known to function specifically at a single nucleotide site at the start of transcription (55,56), and therefore it is difficult to detect using low resolution approaches. Out of 156 8-mer predictions made by FitzGerald *et al*. none included the YYAnW 5-mer with enrichment at the TSS. In contrast, our model identified seven 5-mers matching this consensus with significant enrichment at the start of transcription ($P < 1e{-}15$). The most common version of this motif was TCAGT, which was found overrepresented at the TSS more than seven and a half times over the background frequency ($P = 6e{-}48$). Despite the highly significant amount of positional overrepresentation exhibited by this motif, none of the studies using the sliding window approach detected any motifs containing this 5-mer (18,19,47,48). Figure 5 shows the occurrence data of this motif using 20 bp windows and using single-site resolution; we note the significant decrease of the signal when considering the data using windows of 20 bp.

Tharakaraman *et al*. (21,57) also scanned for positional biases within human promoters. However, their methodology allowed for varying window sizes, improving sensitivity of spatial enrichment considerably. Tharakaraman *et al*. predicted 1226 unclustered 8-mers to exhibit positional enrichment within the promoter. Despite such a large number of predictions made by Tharakaraman

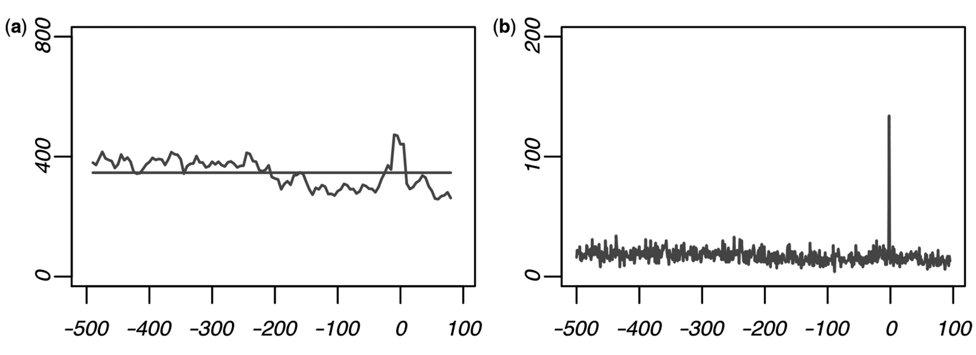

Figure 5. Occurrence frequency of the functional Inr sequence 5-mer TCAGT. The contrast is shown between occurrence data using (**a**) 20 bp windows and (**b**) single-site resolution.

et al., we found that 82% of their predicted motifs overlapped with our results (Table 2). However, their model assumed a uniform background frequency of occurrence across the promoter. Since GC mono- and di-nucleotide composition rises substantially near the start of transcription (19), about a third of the 8-mers predicted by Tharakaraman *et al.* were highly GC-rich, containing at least seven out of eight G/C consensus sites. This is nine times more than what would be expected from a random selection of 8-mers. In contrast, the number of GC rich motifs predicted during our analysis is only 2.7 times higher than would be expected by chance. As many GC-rich motifs do play functional roles in gene regulation, we looked to determine whether these GC-rich 8-mers do, in fact, comprise true regulatory elements. To assess the validity of these GC-rich predictions, we compared the predictions made by Tharakaraman *et al.* to known protein-binding sites found in humans. We found that, among the GC-rich predictions overlapping our results, over half matched human-binding elements found in the TRANSFAC database. This represents a significant enrichment of known regulatory elements, as only about a third of all GC-rich 8-mers match human binding sites in TRANSFAC. However, among the GC-rich predictions that did not overlap our results, only 19% matched known human binding sites. This is significantly less than would be expected by chance given a random selection of 8-mers. Although this evidence is not necessarily conclusive, we would still expect some amount of enrichment for known regulatory elements in this list of predictions. Thus, it is likely that a number of these predictions are simply the result of the rise of GC content near the TSS, rather than true regulatory elements.

In contrast to the analysis of Tharakaraman *et al.*, Vardhanabhuti *et al.* (18) controlled for changes in basepair composition across the promoter. In this analysis, the observed number of occurrences of a given motif was compared to an expected number of occurrences in each window of 20 bp. The expected frequency was estimated separately within each individual window by considering occurrence data of other motifs with identical basepair composition. That is, occurrence data was obtained for motifs whose columns were 'permuted' from the original motif, thus conserving base composition. The observed occurrences of these permuted motifs were then used to determine the expected frequency of occurrence in each individual window; both the 'observed' and 'expected' frequencies were thus unique to each window. Vardhanabhuti *et al.* first scanned for positional biases using known TF-binding sites in TRANSFAC, and subsequent analyses predicted spatial overrepresentation across all (novel) 7-mer motifs filtered for known binding sites in TRANSFAC. Although these latter predictions were presented as novel motifs, we found that a third of these 7-mers matched known regulatory elements. For example, the second and sixth highest-ranking motifs (ATTGGCT and AGCCAAT) match the NFY-binding site, each with a STAMP *E*-value under $E < 1e{-}7$ (Supplementary Data S3).

Between-studies comparisons showed consistently less overlap between the results of Vardhanabhuti *et al.* and those of other studies, including the one presented here (Table 2). It is likely that these differences can be explained by the methodology used to estimate the background frequency of occurrences. For instance, the occurrence frequency of a motif rich in a single nucleotide type will not be significantly different after permuting its columns, as the motif consensus itself will not be changed considerably. In particular, mono-nucleotide repeats are impossible to detect. As a result, sensitivity to many biologically relevant signals is decreased significantly. Vardhanabhuti *et al.* note within their study that their methodology predicts enrichment of the well-known TBP-binding element (TATA-box) at a location that differs from where it is known to function. This motif was predicted by Vardhanabhuti *et al.* to be enriched 45 bp prior to the TSS, although it is known to function at a very specific location 30-bp upstream of the TSS (55,56,58–60). The authors attribute this discrepancy to an increase of A/T nucleotide composition at this location, increasing the 'expected' number of occurrences within this window and therefore decreasing the observed/expected ratio. However, the increase of A/T nucleotide composition at this location is simply a result of the overrepresentation of the A/T rich TBP-binding site itself. This raises the concern that correcting for basepair composition in a position-specific manner can cause failure to detect real biological signals, as the signal itself can be incorporated into the background (expected) frequency. The method presented here effectively circumvents this problem, as the background frequency is modeled in a continuous fashion. Significant changes in the expected frequency caused by real biological signals remain outlier points after fitting the background model to the data (Figure 1b). We note that in the case of the TATA-box, the MLF model predicted enrichment at the correct location 30 bp prior to the TSS at a high level of confidence.

Motif relational functions (MRFs) detect spatial biases between motif-pairs

Transcription is driven by multiple regulatory elements acting in coordination, and knowledge of regulatory element interactions is essential to understand the mechanisms driving gene regulation. Since protein-protein interactions are inherently structure-specific, it is logical to expect that regulatory motifs binding interacting TF proteins preferentially co-occur non-randomly with respect to each other. Since the MPF model is designed to detect spatial biases of motif occurrence, we expanded the previous model in order to predict motif-pairs binding interacting TFs *de novo* according to their spatial relationships. The extended model measures inter-motif distance preferences between pairs of motifs; we denote this measure as a 'motif relational function' (MRF).

The MRF model represents a simple extension of the previous MLF model. In this new model, we are able to determine multiple instances of spatial biases for each individual motif-pair. This aspect of the model was designed to capture the phasing of inter-motif distance preferences, which would be expected according to the results of previous studies. Periodic distributions have been associated with DNA sequence features attributed to the structural conformations of the nucleosome (61), and TF-pair interactions are often known to occur at phased intervals around the histone complex or the winding of the DNA double-helix (3,62,63). This scheme is shown in Figure 6, which illustrates a potential preference for protein–protein interactions to occur in a specific orientation in relation to the turn of the double-helix. As explained in the methods section, the number of overrepresentation peaks (i.e. individual instances of spatial preferences) is not pre-defined but is instead estimated separately for each individual motif-pair. Specifically, the model is designed to detect each instance of spatial bias on an individual basis, continuing to add overrepresentation peaks until all statistically significant peaks have been incorporated into the model.

Figure 7 shows two MRFs which were both generated by motif-pairs that bind TFs with known interactions (64,65), namely, the NFY-NFY and NFY-SP1-binding motif pairs. Motif-pairs were often found to co-occur preferentially at multiple distances, with intervals separating preferred distances corresponding approximately to the turn of the DNA double-helix.

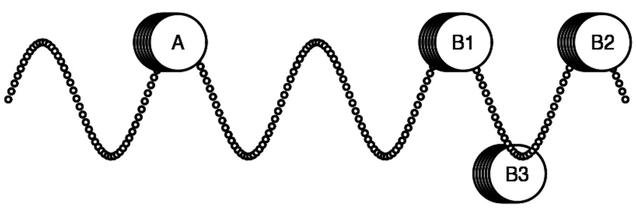

Figure 6. Functional motif-pair inter-relationships. Proteins must often be positioned in a particular orientation with respect to the DNA molecule to induce potential interactions (3,62,63). Interactions between protein A and protein B occur when the latter is positioned at B1. The same interaction frequently occurs one turn of the double-helix away from B1 (i.e. at B2), since the orientation of protein B is consistent relative to the turn of the DNA molecule. However, the interaction cannot occur when protein B is at B3 due to its inconsistent orientation. The distance between factors A and B is determined by the size of the proteins and is unique between different TF-pairs. In contrast, phasing intervals (i.e. the distance between B1 and B2) remains relatively consistent, as they correspond approximately to the number of nucleotides in a turn of the DNA double-helix.

In our particular application of the MRF method, we used the phasing intervals of inter-motif distance preferences as a criterion for predicting motifs with a cooperative role in transcriptional regulation. Namely, we predicted pairwise interactions between putative regulatory motifs by scanning for consistency in the intervals between overrepresentation peaks. In order to quantify the phasing of inter-motif distance preferences, we defined a ‘peak-separation value’, x_{sep}, to be the distance between any pair of overrepresentation peaks within the same motif-pair MRF. Thus, $x_{sep} = |\mu_i - \mu_i|$ for any two peaks i and j within the same MRF (Equation 5). We controlled for peaks potentially representing random outlier points by filtering peaks corresponding to a single-site location. We also filtered double motif occurrences separated by less than 20 bp in order to remove spurious peak phasing caused by repeat sequences and same-sequence dyads.

Our comprehensive MRF analysis was conducted across all possible pairwise combination of 5-mer motifs. Enumerating across all MRFs containing at least two overrepresentation peaks resulted in 619 MRFs exhibiting x_{sep}-values ranging between 7.5–9.5 bp (Figure 8). This is more than twice the expected number (i.e. the average number of MRFs producing x_{sep}-values within any other 2-bp range). While these values do not correspond precisely to the number of nucleotides in a turn of the double-helix (~10.5 bp), it is possible that this deviation can be explained by distortions of the DNA caused by protein binding, or by other similar mechanisms.

Given these trends, we hypothesized that motifs producing consistent x_{sep}-values corresponding to the turn of the DNA double-helix would act as protein-binding sites. We chose to use a stringent criterion for motif-pair predictions by specifically focusing on motifs exhibiting significant concentrations of x_{sep}-values, whose consistency was unlikely to be due to chance. Thus, we calculated the distribution of peak phasing intervals for each individual 5-mer motif across all possible motif partners, testing for significant concentrations within their x_{sep}-value concentrations. Namely, at each iteration, we chose an individual 5-mer to be the fixed motif, and x_{sep}-values were then determined across all MRFs

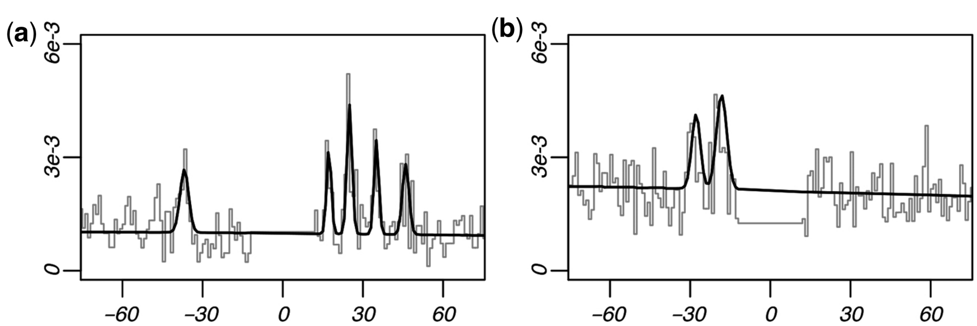

Figure 7. MRFs of two motif-pairs-binding interacting TFs (64,65). A known occurrence of the reverse-strand NFY-binding site defines the position $x = 0$; x-axis values denote the position of the (**a**) plus-strand NFY-binding site and the (**b**) minus-strand SP1-binding site. y-axis values show the frequency of occurrence of these partner motifs. Each motif-pair exhibits more than one preferred distance between motifs, with intervals between peaks being ~8–10 bp. This is consistent with the scheme illustrated in Figure 6, where the location of the peaks represents the positions of B1 and B2, and the position $x = 0$ corresponds to the position of factor A.

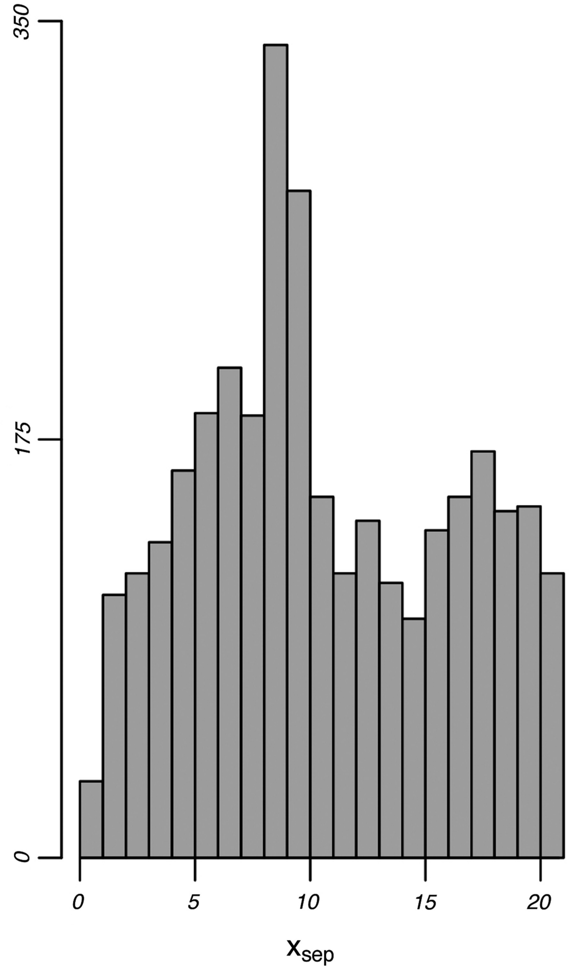

Figure 8. MRF peak separation distributions. The distribution of peak separation values (x_{sep}-values) shown were produced by enumerating all peak-pairs across all possible 5-mer motif-pairs. Peak separation values are defined as $x_{sep} = |\mu_i - \mu_j|$ between overrepresentation peaks i and j within the same MRF; this value is analogous to the distance between B1 and B2 in Figure 6. Note the strong concentration of x_{sep}-values close to ~8–9 bp.

produced from this fixed motif and one of the possible variable motif partners. x_{sep}-values were then accumulated across all variable motif partners; fixed motifs producing a significant x_{sep}-value concentration within one of the 2-bp windows centered at $x_{sep} \approx 8,9,10$ were predicted to be functional.

After correcting for multiple hypothesis testing, thirteen 5-mers were found to have significant x_{sep} concentrations within one of these regions ($P < 1e$–5). Clustering these 5-mers according to sequence similarity and their corresponding x_{sep} distributions produced nine consensus motifs (Table 3). Six of the nine consensus sequences matched known TF-binding sites in TRANSFAC, namely the NRF1, NFY, EVI1, and MADS-box protein family-binding sites. The NFY and MADS-box protein family-binding motifs were predicted on both strands, while the NRF1-binding sequence was palindromic. Three additional motifs, denoted as y1-y3, did not match any known binding sequences in TRANSFAC,

Table 3. Motifs exhibiting consistent phasing intervals between preferred inter-motif distances

Rank	Consensus	TF	x_{sep}	P	Partners
Motifs with consistent phasing intervals					
1	TTTGTA	*y*1	9	7*e*–26	19
2	ATTTTT	MADS (−)	8	3*e*–21	24
3	AAAAAT	MADS (+)	8	4*e*–19	16
4	GCATGC	NRF1	9	7*e*–19	23
5	ATTGC	*y*2	8	3*e*–12	8
6	TCTTG	EV11	9	le–11	7
7	GAGCT	*y*3	10	2*e*–10	7
8	ATTGG	NFY(−)	10	4*e*–8	5
9	CCAAT	NFY(+)	10	le–7	3

Phasing intervals (x_{sep}) were considered across all MRFs produced by the (fixed) motifs shown above and one of the possible 5-mer partners. x_{sep}-values denote the interval between distance preferences (peaks) within the same MRF for a pair of motifs ($x_{sep} = |\mu_i - \mu_j|$). x_{sep}-value concentrations were determined across all 2-bp intervals centered around 8–10 bp. *P*-values (fifth column) correspond to the significance of this concentration for the top-ranking 5-mer in each cluster. Tfs binding to known motifs in TRANSFAC (39) are shown in the third column [STAMP (40) *E*-value threshold: 1*e*–5]; novel regulatory element predictions are labeled y1–y3. The number of predicted partner clusters is given in the right column.

and therefore represent novel *cis*-regulatory element candidates. Figure 9 shows the x_{sep}-value distribution for the highest-ranking 5-mer in four of the predicted motif clusters. These include the reverse-strand MADS-box protein family-binding site, the NRF1-binding site, the novel y1 motif, and the reverse-strand NFY-binding site. Note the highly significant concentration of x_{sep}-values around ~8–10 bp for each motif.

Periodic phasing of inter-motif distance preferences detects known and novel regulatory element relationships

We extended the analysis in order to predict binding site partners for each of the motifs exhibiting significant x_{sep}-value concentrations. Each fixed motif predicted during the previous MRF analysis was paired with multiple partner motifs by considering each individual fixed/partner motif-pair MRF. A 5-mer was predicted to pair with the fixed motif if the motif-pair produced phased distance preferences corresponding to the fixed motif's x_{sep}-value concentration. The predicted partner motifs were then clustered according to sequence similarity as well as the location of their overrepresentation peaks. This procedure produced a total of 112 motif partner clusters pairing with one of the nine fixed motifs predicted in the previous analysis.

Partner motif predictions for the NFY and MADS-box protein family-binding sites are given in Tables 4 and 5, respectively. Only a few motif clusters were predicted to pair with the NFY-binding motifs; each of the partner clusters corresponded to either the NFY or the SP1-binding sequences. Both factors are known to have direct interactions with NFY (64,65).

The MADS-box protein family-binding sites were predicted to pair with more partner motifs than the NFY-binding element. Sixteen and 24 partner clusters are

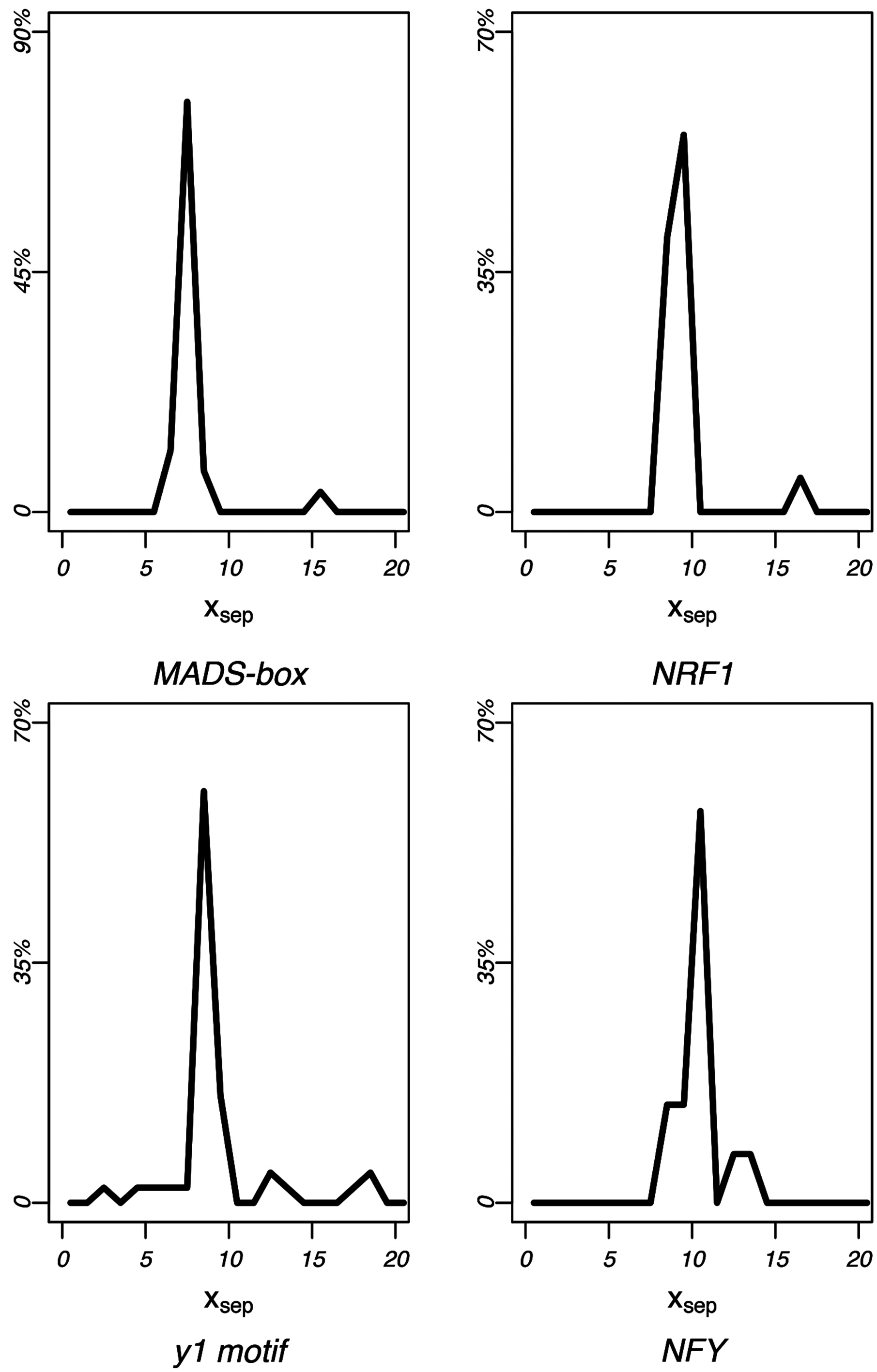

Figure 9. MRF peak separation concentrations. Peak separation distributions are shown for four motifs with significant x_{sep}-value concentrations. Each panel shows the x_{sep} distributions for the most significant 5-mer of the reverse-strand MADS-box family-binding motif, the NRF1-binding motif, the reverse-strand NFY-binding motif and the novel regulatory element prediction y1.

predicted to pair with the forward- and reverse-strand MADS-box motif, respectively. A total of 24 (60%) of the partner clusters were found to match known TF-binding sequences in TRANSFAC, comprising a total of 19 known regulatory elements. Several of these regulatory elements bind proteins with direct interactions to SRF, a member of the MADS-box protein family. These include binding motifs of TCF3 (66), CEBP (67), NFY (68) and ATF6 (69). We found that partner motifs pairing with the MADS-box-binding site were frequently predicted in both orientations. There were eight reverse complement matches between partner motifs pairing across opposing strands of the MADS-box-binding motif, comprising half of the motif predictions pairing with the forward-strand

MADS-box parent motif (Table 5). We found that the mutual directionality of the fixed/partner motif-pairs was highly conserved, with each separate strand of the MADS-box motif pairing with only a single strand of the partner motif. There were no same-strand matches between partner motifs across opposing strands of the parent MADS-box-binding element; i.e. individual strands of the partner motifs were not found to pair to both orientations of the MADS-box-binding motif.

DISCUSSION

Measuring spatial overrepresentation at high-scale resolution provides a powerful, but not yet frequently applied, approach to predicting functional motifs on a genome-wide scale. The MPF model provides a general measure of position-specific enrichment for a motif at single-site precision in relation to a given reference point. The methodology is distinct from previous methods that use general overrepresentation as a criterion for regulatory function. Although such motif finders have been useful for certain applications, assessments of these methods have shown that the efficacy of such approaches is somewhat limited (70). The method presented here is designed to predict either individual regulatory elements or functional motif-pair relationships using spatial biases, rather than the overall frequency of occurrence, as a criteria for functionality. The MLF model shows that many known regulatory elements exhibit strong locational preferences relative to the TSS. The MRF model predicts motif-pair relationships *de novo* by analyzing inter-motif distance preferences without prior knowledge about the corresponding motif sequences. Our implementation of this model predicts *cis*-regulatory elements and their pairwise interactions using periodically phased distance preferences between pairs of motifs.

Table 4. Motif partners for the NFY-binding element

TF partner	Forward strand	Reverse strand
NFY-binding site partners		
NFY (+)	CCAAT	GCCAATC
(−)	GATTGGC	ATTGG
	CGATT	
SP1 (+)	GGCGG	GGGCGG
(−)	CCGCC	

Partner motifs for the forward and reverse strand NFY-binding motif are shown in the second and third columns, respectively. Each partner motif binds either the NFY or SP1 factors (left column). Both NFY-NFY and NFY-SP1 factor-pairs exhibit known interactions (64,65).

The MLF method predicts positional enrichment for 48 consensus motifs

The comprehensive MLF analysis predicted 48 consensus motifs to exhibit positional enrichment within the promoter (Table 1). Thirty-four (71%) of these consensus sequences matched previously documented *cis*-regulatory

Table 5. MADS-box family-binding site partners

Forward strand: AAAAAT				Reverse strand: ATTTTT			
	Consensus	TF	RC		Consensus	TF	RC
MADS-box motif partners							
1	AGACC			1	GAACTCCT	NR1I2	
2	CAGCTAC	TOPORS		2	AGCCT		Rl
3	AGGCTG		rl	3	AGTGC	HMX3	R8
4	CAGCC			4	ATCCG		
5	CCTGTA	AR		5	ATGTT		
6	CGCCA	E2F1		6	CACCA	NFY (*)	R7
7	CTACTC			7	CCACG	ATF6 (*)	
8	GCTGAG	NFE2	r2	8	CCCAA	IKZF1	
9	GAACC			9	CCTCC	MAZ	R3
10	GGCAGG	TCF3 (*)		10	CCTGA		
11	GGAGG	MAZ	r3	11	TCGAAC	XBP1	
12	GGTTG		r4	12	GCTGGGACA	PITX2	
13	TGTAATCCCA	CEBP (*)	r5	13	GATCC		
14	GTGGC		r6	14	GGATTACA	CEBP (*)	R5
15	TGGTG	NFY (*)	r7	15	GCCAC		R6
16	GCACT	HMX3	r8	I6	GGGTTT	TERF2IP	
				17	TCAAG	NKX2	
				18	TGACC	ESR1	
				19	TGATC		
				20	AGCCA	PCBP2	
				21	CAACC		R4
				22	CTCGG	ZNF569	
				23	TCAGC	NFE2	R2
				24	TGCCT		

TFs binding to the known *cis*-regulatory elements in TRANSFAC (39) are shown in the third columns. Binding factors with known direct interactions to SRF, a MADS-box family member, are labeled with asterisks. Reverse complements across opposing strands of the MADS-box fixed motif, as determined by STAMP (*E*-value threshold: 1*e*–5), are labeled r1–r8 and R1–R8 (e.g. r1 is the reverse complement of R1, etc.).

elements, comprising a total of 22 known regulatory elements. The remaining 14 comprise putatively novel *cis*-regulatory element predictions without known binding partners.

While most previous studies predict spatial preferences by counting motif occurrences within multiple independent windows of ~20–25 bp (18–20,47,48), the MLF model considers the data collectively at single-site resolution, estimating the underlying frequency of occurrence according to position. Although prior studies have shown positional enrichment of motifs, the increased spatial resolution of our approach provides evidence for finer-scale functional constraints on the position of motifs. Inspection of the results of FitzGerald *et al.* (19) and Vardhanabhuti *et al.* (18) showed that the most common Inr 5-mer consensus (TCAGT) did not appear in the predictions of either study. Occurrences of this motif are constrained to a single nucleotide site, and thus the positional specificity of this motif could not be detected using larger 20-bp windows. Previous methodologies cannot operate at single-site resolution without a loss of sensitivity to spatial preferences occurring across larger ranges of the regulatory region. We also note that several of our motif predictions not predicted by these previous methods were found overrepresented across wide ranges of the upstream flanking region. For instance, the ZFP161-binding motif shows clear enrichment across a range of ~200 bp (Figure 4); this consensus motif had not previously been detected using the sliding window method. We show that considering the data collectively leads to a significant increase of sensitivity, and allows us to detect positional biases occurring either across broad portions of the promoter or constrained to a single nucleotide site.

The model presented here accounts for dinucleotide fluctuations found in and around the promoter. We show from the results of two simulation analyses that changes of basepair composition within the promoter can produce spurious motif predictions due to the rise of GC content near the TSS. This effect was found to be more prominent in simulated data sets conserving dinucleotide frequencies at each site than those generated using only mono-nucleotide basepair composition. Comparisons to previous studies suggest that models not accounting for nucleotide composition are sensitive to the significant rise of GC content near the TSS. Methods assuming a uniform background frequency are prone to predict GC-rich motifs which are likely to reflect only changes in basepair composition, rather than motif functionality. However, the cross-studies comparisons also show that accounting for fluctuations in mono- and dinucleotide frequencies must be conducted collectively across all positions, rather than independently at each location. For instance, Vardhanabhuti *et al.* (18) estimated the expected frequency of occurrence separately at each location within the promoter. However, overrepresentation of a positionally enriched motif can itself increase the basepair composition at the same location. Estimating the expected frequency within each separate window therefore incorporates such increases in basepair composition into the background frequency, although this rise in the expected frequency is simply a direct byproduct of the biologically relevant overrepresentation. As discussed by the authors of this study, their method was unable to predict positional specificity of the TATA-box at its correct functional location (48,58–60); it is highly likely that many other existing signals went undetected using this model. In contrast, the model presented here estimates the background frequency of occurrence collectively in a continuous fashion. In this way, increases of the expected frequency resulting from true biological signals remain outlier points after fitting the expected frequency to the data, and they are therefore not directly incorporated into the background frequency of occurrence (Figure 1b). We note that in the case of the TATA-box, our model predicts enrichment at its correct location 30 bp prior to the TSS at a high level of significance. Thus, we find that considering the data collectively is essential in order to control for dinucleotide fluctuations within the promoter.

The results from our analyses indicate that many regulatory motifs exhibit positional enrichment in or near the promoter. It is likely that these motifs may perform specific functions at their respective locations of enrichment. We note that some of the known motifs are also known to function at a diverse number of locations, and thus functional occurrences of these motifs are not restricted to occur within their location of enrichment. For instance, functional occurrences of the SRF-binding motif are not limited to its predicted location of enrichment (71). However, SRF binding is known to be facilitated by the YY1 factor, with each protein occupying the same binding location on the DNA, albeit on opposite strands of the molecule (72,73). As YY1 has been shown to function at specific locations within the promoter (48), it is not surprising that the SRF-binding motif also exhibits positional enrichment near that of YY1. Thus, although functional occurrences of the SRF-binding element occur at other locations, it is likely that SRF may play a unique role within its location of enrichment in concert with the YY1 factor. It is possible that other factors also known to bind at a diverse number of locations may also perform specific functional roles at which their binding motifs preferentially occur, and future studies may elucidate the biological relevance regarding the roles of these motifs within their area of enrichment.

Our predictions also include 14 motifs (d1–d14) that do not match any known human TF-binding motifs in TRANSFAC. Although these motifs do not match any known TF-binding sequences, some have been reported to play alternative roles in gene regulation. For instance, the novel d5 and d10 motifs consist of $(GT)_n$ and $(CA)_n$ dinucleotide repeats, respectively. These sequence elements are known to promote left-handed Z-DNA structures that affect DNA supercoiling, and subsequently transcription (52–54). Similarly, the d3 motif consists of a homopolymeric thymine [poly(T)] tract. Such poly(T) tracts are known to alter the conformation of the DNA molecule, thereby disrupting nucleosome positioning (51). Both d3 and d10 show a significant overrepresentation peak centering near the start of the transcription bubble (74). A recent study conducted by Kaplan *et al.* (75) has shown that nucleosome occupancy is significantly decreased at this location. It is therefore likely that occurrences of these motifs in this region of the promoter may make the area more accessible to the RNA-polymerase machinery. Thus, while many of our novel regulatory element predictions are likely to function as protein-binding sites, others may affect transcriptional regulation through alternative mechanisms.

Applying our method to a set of mouse promoters showed that the majority of the motifs exhibiting spatial preferences in humans also show positional enrichment in mouse. Thirty-six (75%) of the motif clusters were shared across the two species, and the location of positional enrichment was frequently found to be highly conserved between the two lineages. Supplementary Data S4 provides results from a second mouse promoter analysis conducted on a smaller, high-quality data set provided by the RIKEN Institute. This analysis was conducted on 1354 mouse promoters generated using CAGE-tag data (56,76), a significant decrease in the number of sequences used during our previous analyses. Thirty-one motifs predicted during the human promoter analysis matched predictions from the RIKEN data set, representing a significant overlap between results produced from the two data sets. However, the positional bias of some infrequently occurring motifs could not be detected using the smaller RIKEN data set, including several regulatory elements whose positional bias is well-documented, such as the CREB, MYC, NRF1 and ETS-binding sites (19,47,48). Thus, we note that it is usually preferable to use larger data sets than smaller ones in order to minimize the amount of random noise relative to the background frequency. This increases sensitivity to the spatial bias of infrequently occurring motifs, and it also eliminates spurious peaks caused by random outlier points. We have found that our approach is not greatly affected by noise when applied to large data sets, and that the MLF method is efficient at detecting instances of spatial bias when using an adequate number of sequences.

MRF peak separation analyses predict individual and putatively interacting *cis*-regulatory elements

The MRF model presented here analyzes inter-motif distance preferences between pairs of sequence elements. Our application of this method predicted *cis*-regulatory element-pairs that putatively bind interacting TFs by considering periodically phased spatial preferences. We found that pairs of TF-binding sequences with collective roles in transcription often show elevated frequencies of co-occurrence at multiple separation distances. The model presented here explicitly uses multi-modal characteristics of inter-motif distance frequencies, and is therefore inherently different from previous uni-modal models, which have generally relied on the sliding window method or maximum-distance approaches (16–18,31,77–79). We have found that individual instances of spatial preferences are generally constrained to widths of only ~2–3 bp, and that single overrepresentation peaks often exhibit only minimal amounts of significance. However, despite the subtlety of each individual instance of spatial preference between motif-pairs, overall trends in the phasing intervals between preferred inter-motif distances are highly significant. Our model was intentionally designed to account for spurious overrepresentation peaks by considering peak separation distances collectively across a comprehensive list of all pairs of 5-mers. We explicitly focused on motifs with distinguishing phasing intervals between preferred distances, and whose consistency was unlikely to be due to chance.

We have found that for many motif-pairs exhibiting multiple preferred separation distances, the intervals between preferred inter-motif distances are often found near the range of ~8–9 bp. This corresponds to approximately 2 bp less than the number of nucleotides in one turn of the DNA double-helix (~10.5 bp). This finding may reflect a structural requirement for interacting TF proteins to be positioned in a particular orientation relative to the turn of the DNA molecule, albeit at a distortion of the helical structure upon protein binding. The deviation from the expected number of 10.5 bp is worth noting, particularly due to the robustness of the signal occurring across such a large number of motif-pairs. There are some possible explanations available from the literature. Structural analyses have shown that protein binding, and the binding of multi-protein complexes in particular, distort the conformation of the DNA (80–83), thus affecting the helical characteristics of the DNA. An alternative, although not necessarily exclusive, interpretation is that the occurring protein-protein interactions may either be stabilized by alterations of the DNA molecule or require them for collective binding. Selective binding of proteins to DNA involves not only sequence-specific elements within the DNA, but also topological characteristics of the DNA molecule (84–86); this is known to be particularly true during the recruitment of multiple interacting proteins to the DNA (85,86). Thus, although the biological explanation of the observed pattern remains unclear, these results are not inconsistent with our current knowledge of protein–protein and protein–DNA interactions.

Nine motif clusters were found to exhibit highly consistent phasing intervals around 8–10 bp. Six of these consensus motifs matched known binding sites in TRANSFAC, including the binding motifs for the NFY, NRF1, EVI1 and the MADS-box protein family. We subsequently used our phasing criteria to predict multiple partner motifs for each of these nine motifs, resulting in 112 motif-pairs predicted to bind interacting TFs (Supplementary Data S5). We have illustrated several examples of our motif-pair predictions in Tables 4 and 5, which show the partner motifs pairing with the NFY-binding element and the MADS-box protein family-binding sequence. All partners predicted to pair with the NFY-binding motif correspond to either the NFY or SP1-binding elements; both NFY–NFY and NFY–SP1 factor-pair interactions are documented in the literature (64,65). The MADS-box family consensus sequences predicted during the analysis bind both the myocyte enhancer factor 2 (MEF2) and the serum response factor (SRF) (87). These two factors are known to be involved in complex extra-cellular signaling pathways, playing multiple roles involving cell differentiation and development (88–90). Both MEF2 and SRF regulate gene expression through the recruitment of multiple accessory co-factors whose presence or absence within the complex cause differential expression of their target genes (66,67,86), and therefore we would expect a large number of partner motifs to pair with their binding elements. Many predicted partner motifs were predicted on both orientations, with eight reverse complement-pairs occurring across opposing strands of the MADS-box-binding motif. The mutual directionality of the MADS-box/partner motif-pairs was found to be highly conserved, with each individual strand of the partner motif pairing with one, but not both, orientations of the MADS-box-binding motif. The majority of partner motifs pairing with the MADS-box-binding element match known TF-binding sites. In addition, 12 motif partners bind proteins known to be involved in either signal transduction pathways or developmental processes. Three such factors belong to the homeobox family, whose members play a crucial role in early development (91–93). Many of the remaining partner motifs may play unknown functional roles in concert with one of the MADS-box protein factors.

We note that the MRF method comprises a general tool that can be used to analyze spatial preferences, and that our present analysis represents only one of the many possible applications of the model. Expanding our knowledge about the nature of collective protein binding is crucial to further the understanding of gene regulation, and future studies are necessary to demonstrate the full efficacy of the method presented here.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Sayan Mukherjee, Jeff Thorne, the Wray lab and STU labs for their comments.

FUNDING

National Institute of General Medical Sciences; National Institutes of Health [grant numbers HG004065 and 5P50-GM-081883]. Funding for open access charge: National Institutes of Health grant number HG004065 and 5P50-GM-081883.

Conflict of interest statement. None declared.

REFERENCES

1. Pedersen,A.G., Baldi,P., Chauvin,Y. and Brunak,S. (1999) The biology of eukaryotic promoter prediction – a review. *Comput. Chem.*, **23**, 191–207.
2. Latchman,D.S. (1990) Eukaryotic transcription factors. *Biochem. J.*, **270**, 281–289.
3. Wray,G.A., Hahn,M.W., Abouheif,E., Balhoff,J.P., Pizer,M., Rockman,M.V. and Romano,L.A. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, **20**, 1377–1419.
4. Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.
5. Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs biopolymers using expectation maximization. *Machine Learning*, **21**, 51–80.
6. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
7. 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.
8. van Helden,J., Rios,A.F. and Collado-Vides,J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **28**, 1808–1818.
9. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouze,P. and Moreau,Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.
10. Sinha,S. and Tompa,M. (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical over-representation. *Nucleic Acids Res.*, **31**, 3586–3588.
11. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
12. Bussemaker,H.J., Li,H. and Siggia,E.D. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. USA*, **97**, 10096–10100.
13. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.
14. Down,T.A., Bergman,C.M., Su,J. and Hubbard,T.J. (2007) Large-scale discovery of promoter motifs in Drosophila melanogaster. *PLoS Comput. Biol.*, **3**, e7.
15. Keles,S., van der Laan,M. and Eisen,M.B. (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167–1175.
16. Hannenhalli,S. and Levy,S. (2002) Predicting transcription factor synergism. *Nucleic Acids Res.*, **30**, 4278–4284.
17. Wang,J. and Hannenhalli,S. (2006) A mammalian promoter model links cis elements to genetic networks. *Biochem. Biophys. Res. Commun.*, **347**, 166–177.
18. Vardhanabhuti,S., Wang,J. and Hannenhalli,S. (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.*, **35**, 3203–3213.
19. FitzGerald,P.C., Shlyakhtenko,A., Mir,A.A. and Vinson,C. (2004) Clustering of DNA sequences in human promoters. *Genome Res.*, **14**, 1562–1574.
20. FitzGerald,P.C., Sturgill,D., Shyakhtenko,A., Oliver,B. and Vinson,C. (2006) Comparative genomics of Drosophila and human core promoters. *Genome Biol.*, **7**, R53.
21. Tharakaraman,K., Bodenreider,O., Landsman,D., Spouge,J.L. and Marino-Ramirez,L. (2008) The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. *Nucleic Acids Res.*, **36**, 2777–2786.
22. Beachy,P.A., Varkey,J., Young,K.E., von Kessler,D.P., Sun,B.I. and Ekker,S.C. (1993) Cooperative binding of an Ultrabithorax homeodomain protein to nearby and distant DNA sites. *Mol. Cell Biol.*, **13**, 6941–6956.
23. Biggin,M.D. and McGinnis,W. (1997) Regulation of segmentation and segmental identity by Drosophila homeoproteins: the role of DNA binding in functional activity and specificity. *Development*, **124**, 4425–4433.
24. Elkon,R., Linhart,C., Sharan,R., Shamir,R. and Shiloh,Y. (2003) Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.*, **13**, 773–780.
25. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

26. Keles,S., van der Laan,M.J. and Vulpe,C. (2004) Regulatory motif finding by logic regression. *Bioinformatics*, **20**, 2799–2811.
27. 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.
28. Li,L., Cheng,A.S., Jin,V.X., Paik,H.H., Fan,M., Li,X., Zhang,W., Robarge,J., Balch,C., Davuluri,R.V. *et al.* (2006) A mixture model-based discriminate analysis for identifying ordered transcription factor binding site pairs in gene promoters directly regulated by estrogen receptor-alpha. *Bioinformatics*, **22**, 2210–2216.
29. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.
30. Ma,X.-T. and Tang,H.-X. (2004) Predicting polymerase II core promoters by cooperating transcription factor binding sites in eukaryotic genes. *Acta Biochimica et Biophysica Sinica*, **36**, 250–258.
31. Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.
32. Karlin,S., Burge,C. and Campbell,A.M. (1992) Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.*, **20**, 1363–1370.
33. Broyden,C. (1965) A class of methods for solving nonlinear simulationeous equations. *Mathematic Comput.*, **19**, 577–593.
34. Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (2002) *Numerical Recipes in C: The Art of Scientific Computing*. 2nd edn. Cambridge University Press, New York.
35. Davison,A.C. (2003) *Statistical Models*., Cambridge University Press, New York.
36. Burden,R.L. and Douglas,J.F. (2001) *Numerical Analysis*. 7th edn. Wadsworth Group, Pacific Grove, CA.
37. Kullback,S. and Leiber,R.A. (1951) On information theory and sufficiency. *Ann. Mathematical Stat.*, **22**, 79–86.
38. Cavener,D.R. (1987) Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. *Nucleic Acids Res.*, **15**, 1353–1361.
39. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
40. Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.
41. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
42. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
43. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
44. Maglott,D.R., Katz,K.S., Sicotte,H. and Pruitt,K.D. (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28**, 126–128.
45. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
46. Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
47. 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.
48. Xi,H., Yu,Y., Fu,Y., Foley,J., Halees,A. and Weng,Z. (2007) Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res.*, **17**, 798–806.
49. Crothers,D.M., Haran,T.E. and Nadeau,J.G. (1990) Intrinsically bent DNA. *J. Biol. Chem.*, **265**, 7093–7096.
50. Hizver,J., Rozenberg,H., Frolow,F., Rabinovich,D. and Shakked,Z. (2001) DNA bending by an adenine—thymine tract and its role in gene regulation. *Proc. Natl Acad. Sci. USA*, **98**, 8490–8495.
51. Segal,E. and Widom,J. (2009) Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.*, **19**, 65–71.
52. Rich,A., Nordheim,A. and Wang,A.H. (1984) The chemistry and biology of left-handed Z-DNA. *Annu. Rev. Biochem.*, **53**, 791–846.
53. Naylor,L.H. and Clark,E.M. (1990) d(TG)n.d(CA)n sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. *Nucleic Acids Res.*, **18**, 1595–1601.
54. Rothenburg,S., Koch-Nolte,F., Rich,A. and Haag,F. (2001) A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc. Natl Acad. Sci. USA*, **98**, 8985–8990.
55. Yang,C., Bolotin,E., Jiang,T., Sladek,F.M. and Martinez,E. (2007) Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, **389**, 52–65.
56. Ponjavic,J., Lenhard,B., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y. and Sandelin,A. (2006) Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol.*, **7**, R78.
57. Tharakaraman,K., Marino-Ramirez,L., Sheetlin,S., Landsman,D. and Spouge,J.L. (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics*, **21(Suppl. 1)**, i440–i448.
58. Aso,T., Conaway,J.W. and Conaway,R.C. (1994) Role of core promoter structure in assembly of the RNA polymerase II preinitiation complex. A common pathway for formation of preinitiation intermediates at many TATA and TATA-less promoters. *J. Biol. Chem.*, **269**, 26575–26583.
59. Kaufmann,J. and Smale,S.T. (1994) Direct recognition of initiator elements by a component of the transcription factor IID complex. *Genes Dev.*, **8**, 821–829.
60. Martinez,E., Chiang,C.M., Ge,H. and Roeder,R.G. (1994) TATA-binding protein-associated factor(s) in TFIID function through the initiator to direct basal transcription from a TATA-less class II promoter. *EMBO J.*, **13**, 3115–3126.
61. Ioshikhes,I., Trifonov,E.N. and Zhang,M.Q. (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc. Natl Acad. Sci. USA*, **96**, 2891–2895.
62. Lewin,B. (2000) *Genes VII*. Oxford University Press, Oxford.
63. White,R.J. (2001) *Gene Transcription: Mechanisms and Control*. Blackwell Science, Malden, Mass.
64. Liberati,C., di Silvio,A., Ottolenghi,S. and Mantovani,R. (1999) NF-Y binding to twin CCAAT boxes: role of Q-rich domains and histone fold helices. *J. Mol. Biol.*, **285**, 1441–1455.
65. Benfante,R., Antonini,R.A., Vaccari,M., Flora,A., Chen,F., Clementi,F. and Fornasari,D. (2005) The expression of the human neuronal alpha3 $Na+,K+$-ATPase subunit gene is regulated by the activity of the Sp1 and NF-Y transcription factors. *Biochem. J.*, **386**, 63–72.
66. Groisman,R., Masutani,H., Leibovitch,M.P., Robin,P., Soudant,I., Trouche,D. and Harel-Bellan,A. (1996) Physical interaction between the mitogen-responsive serum response factor and myogenic basic-helix-loop-helix proteins. *J. Biol. Chem.*, **271**, 5258–5264.
67. Hanlon,M., Sturgill,T.W. and Sealy,L. (2001) ERK2- and p90(Rsk2)-dependent pathways regulate the CCAAT/enhancer-binding protein-beta interaction with serum response factor. *J. Biol. Chem.*, **276**, 38449–38456.
68. Yamada,K., Osawa,H. and Granner,D.K. (1999) Identification of proteins that interact with NF-YA. *FEBS Lett.*, **460**, 41–45.
69. Zhu,C., Johansen,F.E. and Prywes,R. (1997) Interaction of ATF6 and serum response factor. *Mol. Cell Biol.*, **17**, 4957–4966.
70. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
71. Cooper,S.J., Trinklein,N.D., Nguyen,L. and Myers,R.M. (2007) Serum response factor binding sites differ in three human cell types. *Genome Res.*, **17**, 136–144.

72. Natesan,S. and Gilman,M. (1995) YY1 facilitates the association of serum response factor with the c-fos serum response element. *Mol. Cell Biol.*, **15**, 5975–5982.
73. Chen,C.Y. and Schwartz,R.J. (1996) Recruitment of the tinman homolog Nkx-2.5 by serum response factor activates cardiac alpha-actin gene transcription. *Mol. Cell Biol.*, **16**, 6372–6384.
74. Pal,M., Ponticelli,A.S. and Luse,D.S. (2005) The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II. *Mol. Cell*, **19**, 101–110.
75. Kaplan,N., Moore,I.K., Fondufe-Mittendorf,Y., Gossett,A.J., Tillo,D., Field,Y., LeProust,E.M., Hughes,T.R., Lieb,J.D., Widom,J. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.
76. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G., Frith,M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
77. 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.
78. 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.
79. Segal,E. and Sharan,R. (2005) A discriminative model for identifying spatial cis-regulatory modules. *J. Comput. Biol.*, **12**, 822–834.
80. Tomovic,A. and Oakeley,E.J. (2008) Computational structural analysis: multiple proteins bound to DNA. *PLoS ONE*, **3**, e3243.
81. Jones,S., van Heyningen,P., Berman,H.M. and Thornton,J.M. (1999) Protein-DNA interactions: a structural analysis. *J. Mol. Biol.*, **287**, 877–896.
82. Prabakaran,P., Siebers,J.G., Ahmad,S., Gromiha,M.M., Singarayan,M.G. and Sarai,A. (2006) Classification of protein-DNA complexes based on structural descriptors. *Structure*, **14**, 1355–1367.
83. Pan,Y. and Nussinov,R. (2008) p53-Induced DNA bending: the interplay between p53-DNA and p53-p53 interactions. *J. Phys. Chem. B*, **112**, 6716–6724.
84. Harrington,R.E. (1992) DNA curving and bending in protein-DNA recognition. *Mol. Microbiol.*, **6**, 2549–2555.
85. Kerppola,T.K. (1998) Transcriptional cooperativity: bending over backwards and doing the flip. *Structure*, **6**, 549–554.
86. West,A.G. and Sharrocks,A.D. (1999) MADS-box transcription factors adopt alternative mechanisms for bending DNA. *J. Mol. Biol.*, **286**, 1311–1323.
87. Huang,K., Louis,J.M., Donaldson,L., Lim,F.L., Sharrocks,A.D. and Clore,G.M. (2000) Solution structure of the MEF2A-DNA complex: structural basis for the modulation of DNA bending and specificity by MADS-box transcription factors. *EMBO J.*, **19**, 2615–2628.
88. Miano,J.M., Long,X. and Fujiwara,K. (2007) Serum response factor: master regulator of the actin cytoskeleton and contractile apparatus. *Am. J. Physiol. Cell Physiol.*, **292**, C70–C81.
89. Chai,J. and Tarnawski,A.S. (2002) Serum response factor: discovery, biochemistry, biological roles and implications for tissue injury healing. *J. Physiol. Pharmacol.*, **53**, 147–157.
90. Potthoff,M.J. and Olson,E.N. (2007) MEF2: a central regulator of diverse developmental programs. *Development*, **134**, 4131–4140.
91. Whitelaw,E. (1989) The role of DNA-binding proteins in differentiation and transformation. *J. Cell Sci.*, **94** (Pt 2), 169–173.
92. Lonai,P. and Orr-Urtreger,A. (1990) Homeogenes in mammalian development and the evolution of the cranium and central nervous system. *FASEB J.*, **4**, 1436–1443.
93. Nunes,F.D., de Almeida,F.C., Tucci,R. and de Sousa,S.C. (2003) Homeobox genes: a molecular link between development and cancer. *Pesqui Odontol Bras*, **17**, 94–98.