2238–2246 *Nucleic Acids Research, 2006, Vol. 34, No. 8*
doi:10.1093/nar/gkl248

A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome

Chaolin Zhang[1,2], Zhenyu Xuan[1], Stefanie Otto[3], John R. Hover[3], Sean R. McCorkle[4], Gail Mandel[3] and Michael Q. Zhang[1,*]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA, [2]Department of Biomedical Engineering, State University of New York at Stony Brook, NY 11794, USA, [3]Howard Hughes Medical Institute, Department of Neurobiology and Behavior, State University of New York at Stony Brook, NY 11794, USA and [4]Biology Department, Brookhaven National Laboratory, Upton, NY 11973, USA

Received January 7, 2006; Revised February 15, 2006; Accepted March 29, 2006

ABSTRACT

Transcription factor binding sites (TFBSs) are short DNA sequences interacting with transcription factors (TFs), which regulate gene expression. Due to the relatively short length of such binding sites, it is largely unclear how the specificity of protein–DNA interaction is achieved. Here, we have performed a genome-wide analysis of TFBS-like sequences for the transcriptional repressor, RE1 Silencing Transcription Factor (REST), as well as for several other representative mammalian TFs (c-myc, p53, HNF-1 and CREB). We find a nonrandom distribution of inexact sites for these TFs, referred to as highly-degenerate TFBSs, that are enriched around the cognate binding sites. Comparisons among human, mouse and rat orthologous promoters reveal that these highly-degenerate sites are conserved significantly more than expected by random chance, suggesting their positive selection during evolution. We propose that this arrangement provides a favorable genomic landscape for functional target site selection.

INTRODUCTION

Transcription factor binding sites (TFBSs) are short stretches of DNA (usually 6–20 bp) often found in promoters (1). TFBSs are recognized specifically by the DNA binding domain of transcription factors (TFs), which in turn interact with other TFs or the basal transcription apparatus to activate or repress target gene expression. The genomic distribution of TFBSs and the interaction between TFs and their cognate binding sites are central in deciphering gene regulatory networks. Previously, both experimental (2–7) and computational (8,9) methods have been developed to identify TFBSs in different scales and at different resolutions.

Vertebrate TFBSs are usually degenerate and a number of sequence variations of a TFBS can be functionally related to binding affinity. This implies that nature has evolved a system for maintaining a robust response independent of binding specificity, such that the impact of most mutations within the site is relatively small and not lethal. In some cases difference in binding affinity plays a role in subtle regulation of gene expression. On the other hand, a wide dispersion of highly degenerate TFBS-like sequences exists in genomes that would seemingly result in a large number of nonspecific interactions or even promiscuous bindings to non-functional pseudosites (10). Therefore, mechanisms explaining robust transcriptional responses must take into account redundancy within and outside the cognate binding site. In support of this idea, the specificity of protein–DNA interactions is known to depend, at least in part, on the genomic context around target sites, which serves as a system for guiding target site selection. For example, the veracity of a potential TFBS correlates with the positioning of nucleosomes (11), GC content and CpG islands (12), co-operativity of *cis*-regulatory modules (13) and preferential spacing between TFBSs and transcription start sites (TSSs) (14).

Here, we present a systematic study of the distribution of inexact TFBS-like sequences, referred to as highly-degenerate TFBSs. Due to the high degeneracy and ubiquitous occurrences throughout the genome, these sequences are unlikely to bind TFs with enough affinity to achieve specificity. However, by definition, they do bear higher similarity with cognate binding sites than random sequences. Previous perspectives have ascribed to these sequences, which are

*To whom correspondence should be addressed. Tel: +1 516 367 8393; Fax: +1 516 367 8461; Email: mzhang@cshl.edu

assumed to be distributed randomly, the role of compromising efficient target site selection (10,15). In systematically examined highly-degenerate TFBSs for the transcriptional repressor REST, as well as for several other representative TFs (c-myc, p53, HNF-1 and CREB), we find surprisingly that highly-degenerate TFBSs are enriched significantly around the cognate TFBS. Comparative studies of human, mouse and rat orthologous promoters show that these highly-degenerate sites are conserved more than expected by chance, suggesting their positive selection during evolution. This implies that the nonrandom distribution of highly-degenerate TFBSs very likely contributes to a favorable genomic landscape facilitating specific target site recognition.

MATERIALS AND METHODS

Definition of proximal promoter region

The proximal promoter region was defined as the sequence from −2 to +2 kb relative to the TSS.

Collection of experimentally validated REST targets

Thirty-four unique REST target genes with experimental validation from human, mouse and rat were selected for the analyses. Orthologs of known REST target genes were also included with stringent criteria. A putative REST target gene was included in the dataset only if (i) it was present in human, mouse or rat genomes; (ii) at least one orthologous gene was experimentally validated as a REST target; (iii) the RE1 site was conserved in sequence, location and orientation with the experimentally validated RE1 in the orthologous gene (see below). As a result, 85 REST target genes representing the 34 unique genes were obtained (Supplementary Table S1). Among them, there are 22 triplets, 7 pairs and 5 singletons. The genomic sequences from −2 to +2 kb around TSS were extracted from UCSC genome browser (genome assembly versions: hg17, mm5 and rn3). To date, this represents the largest dataset of REST target genes validated by experimentation and their orthologs in human, mouse and rat genomes. They are called the validated set in the article for simplicity.

Orthologous promoters from human, mouse and rat

Previously, we generated a mammalian promoter database including human, mouse and rat (CSHLmpd, http://rulai.cshl.edu/CSHLmpd2/) as described in ref. (16). This study included all the 11 370 triplets of the orthologous promoters from human, mouse and rat.

Genomic search for RE1-like sequences

RE1-like sequences were identified using the RE1 matrix (accession no. M00256) and the MATCH program (17), both of which were from TRANSFAC [v9.1, http://www.biobase.de, ref. (18)]. In the MATCH program, each matrix was built from a set of validated TFBSs using Equation 1

$$I(i,B) = f_{i,B} \sum_{B} f_{i,B} \ln(4 f_{i,B}), \quad \mathbf{1}$$

where $i = 1,2,\ldots,L$ are L positions of the motif and f_{iB} is the frequency of observing a base $B \in \{A,T,G,C\}$ at the position i. Then the score of a sequence of length L can be given by Equation 2

$$s = \sum_{i=1}^{L} I(i,B_i), \quad \mathbf{2}$$

where B_i is the base at the position i. The score was then linearly scaled to 0 and 1 using the maximal/minimal possible score for the matrix.

The default threshold for RE1s provided by MATCH was too strict. With the 'minimizing false positives' option, it reported only half of the sites which were used to derive the matrix. We therefore modified the threshold and defined a putative RE1, referred to as a high-score RE1, by an overall score of 0.86 or above (the core score remains the same). The highly-degenerate RE1 is defined by an overall score between $t_{\text{lower}} = 0.67$ and $t_{\text{upper}} = 0.86$, without requirement of the core score. The choice of the lower score is somewhat arbitrary but represents the most relaxed constraint.

Before performing the matrix search, repetitive regions were masked by RepeatMasker (A.F.A. Smit,R. Hubley and P. Green RepeatMasker at http://repeatmasker.org). A rigorous scheme was applied to handle overlapping counts of occurrences as follows: overlapping occurrences were clustered and only the one with the highest score was counted. This eliminated the possibility of over-counting caused by the degenerate palindromic or periodic pattern of RE1 sites.

Permutation of RE1

Permutations of RE1 were used to generate random motifs with the same information content and base composition as the real RE1. Permuted RE1 matrices were derived from the shuffling of columns of the RE1 matrix. The same thresholds were used to define high-score permuted RE1s and highly-degenerate permuted RE1s except that no requirement on the core score was set.

Conservation of RE1s and highly-degenerate RE1s

A number of promoters are divergent across human, mouse and rat, which makes it difficult to align orthologous promoters. In this study, an RE1 site or highly-degenerate RE1 site was defined as conserved if it satisfied the following three conditions. First, it occurred in all three sequences in same ortholog group; second, the three RE1 variants were conserved highly in sequences among all three species (allowed divergence in at most three bases, but could be divergent from the RE1 consensus for highly-degenerate RE1s); finally, the three variants also had similar coordinates relative to the TSS on the same strand (location difference $\leq$400 bp relative to TSS). This definition can avoid the difficulties in aligning multiple promoters.

To measure the significance of motif conservation, a similar approach as described previously (8) was used, with minor adaptations. The conservation rate p was defined as the ratio of conserved occurrences NC to the average of overall occurrences in human, mouse and rat, N. The conservation rate p was then compared to the expected rate p_0, which was estimated using random motifs generated by permutations. Permutations were run 10 times to obtain the average of the conserved occurrences NC_0 and the overall occurrences

N_0, permitting us to get a more robust estimation of p_0. Then the two-by-two contingency table (NC, N, NC_0, N_0) was tested by χ^2 test for association.

Other TFs

To test whether the nonrandom distribution of highly-degenerate TFBSs was a general mechanism, we also chose to look at the binding sites for c-myc, p53, HNF-1 and CREB.

As with the RE1 study, matrices were obtained from TRANSFAC [v9.1, http://www.biobase.de, ref. (18)]. Due to the small motif size or undefined thresholds for potential cognate sites, we did not attempt to distinguish high-score sites and highly-degenerate sites for these motifs. Instead, a low threshold was used to define the 'relaxed' binding site for each TF. For c-myc, p53 and HNF-1, the enrichment of relaxed binding sites was compared between promoters of validated or putative target genes for the corresponding TF and CSHLmpd promoters as a control. For CREB, the comparison was performed between putative CREB binding loci in the rat genome as selected by SACO (4) and uniformly selected random genomic loci. The details of data collections and matrices are listed in Supplementary Table S2. The same approach was used to compare the conservation rate of RE1s in two groups of sequences.

Statistical tests

To test the difference of enrichment (conservation rate) of two groups, χ^2 test was applied constructing the two-by-two contingency table (19).

RESULTS

To study the nonrandom distribution of highly-degenerate TFBSs, we first focused on the transcriptional repressor Repressor Element 1/Neuron-Restrictive Silencer Element (RE1/NRSE), the binding site of RE1 Transcription Silencing Factor/Neuron-Restrictive Silencer Factor (REST/NRSF) as a model (20–24). The RE1 site was chosen for a number of reasons, including (i) the important role of REST in nervous system development and as a tumor suppressor (20–22, 25–27); (ii) the RE1 motif is relatively long (21 bp) and conserved in validated REST target genes [see Figure 1A and B, ref. (22) and Supplementary Table S1]; (iii) the threshold for defining a potentially functional RE1 has been validated experimentally (22).

Genomic search for RE1s and highly-degenerate RE1s

In previous genomic searches for RE1 sites, other groups used consensus based methods, including blast (22) and regular expression search (24). In this study, the position specific score matrix (PSSM) was used since it is directly related to binding affinity and is more accurate (10,28). The RE1 matrix obtained from TRANSFAC (accession no. M00256, Figure 1A) (18) was derived from 28 RE1s (22). The MATCH program provided by TRANSFAC was then used to do the genomic search (17). We also tried to refine the matrix using the 85 validated RE1s collected from the literature and produced similar results (Figure 1B). All the results presented in this paper are based on the TRANSFAC matrix.

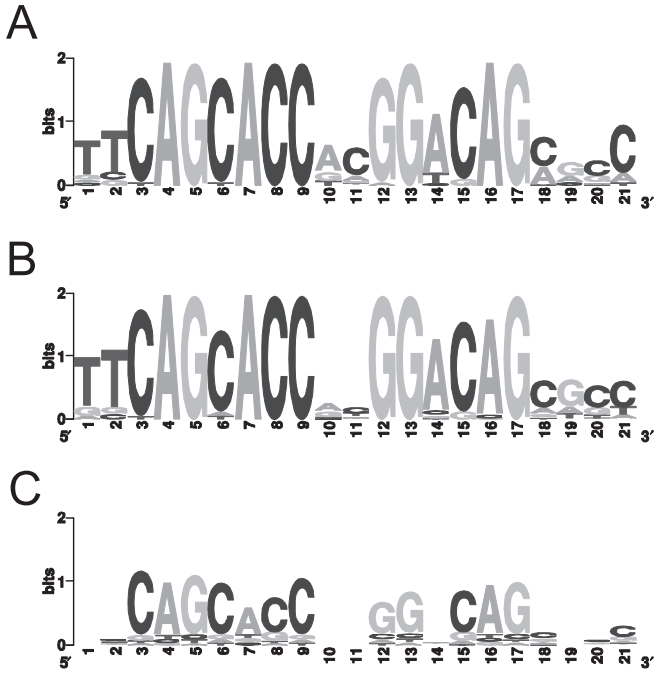

Figure 1. Pictogram of the RE1 motif. The letters on the top of each position represent the RE1 consensus; The height of each position reflects the information content at that position; The relative height of a letter in each position reflects the frequency of observing the nucelotide (48). (**A**) Derived from the TRANSFAC matrix. (**B**) Derived from 85 validated RE1s. (**C**) Derived from conserved highly-degenerate RE1s.

For our analysis, the proximal promoter region was defined as the sequence from −2 to +2 kb relative to the TSS. Searches for RE1-like sequences were performed in the repeat-masked promoters of 85 validated REST target genes collected from published literature (Supplementary Table S1), as well as of 11 370 promoters with orthologs in human, mouse and rat from the mammalian promoter database CSHLmpd (16). High-score RE1s (potential cognate RE1s) are defined by a motif score of $t_{upper} = 0.86$ or above. This corresponds approximately to five mismatches within the consensus sequence, which is an empirical threshold for a functional RE1 observed from experimental data (22). This threshold included all the experimentally validated RE1s. However, a high specificity can still be expected because the RE1 motif is relatively long. The estimation of false positive rate is <5% by motif permutation experiment (see Materials and Methods, data not shown).

The highly-degenerate RE1 was defined by an overall motif score below $t_{upper} = 0.86$ and above a very low threshold of $t_{lower} = 0.67$. The choice of the lower bound is somewhat arbitrary but sets a very relaxed constraint. The frequency of highly-degenerate RE1 under this threshold is approximately one site/kb throughout the genome. Most of these are extremely degenerate and are unlikely to have high binding affinity and specificity to REST.

To eliminate possible over-counting caused by the weak palindromic or periodic pattern of RE1 sites, a rigorous scheme for handling overlapping sites was applied. Overlapping sites were clustered and only the one with the highest score was counted.

Enrichment of highly-degenerate RE1s

Many validated REST target genes (67 of 85) have high-scoring RE1s in the promoter regions (others have RE1s outside the promoters). Among 11 370 orthologous triplets of CSHLmpd promoters, 562 promoters have high-score RE1s (191 in human, 181 in mouse and 190 in rat) and were identified as putative REST targets. The other promoters without high-scoring RE1s were used as controls. We compared the enrichment of highly-degenerate RE1s in the

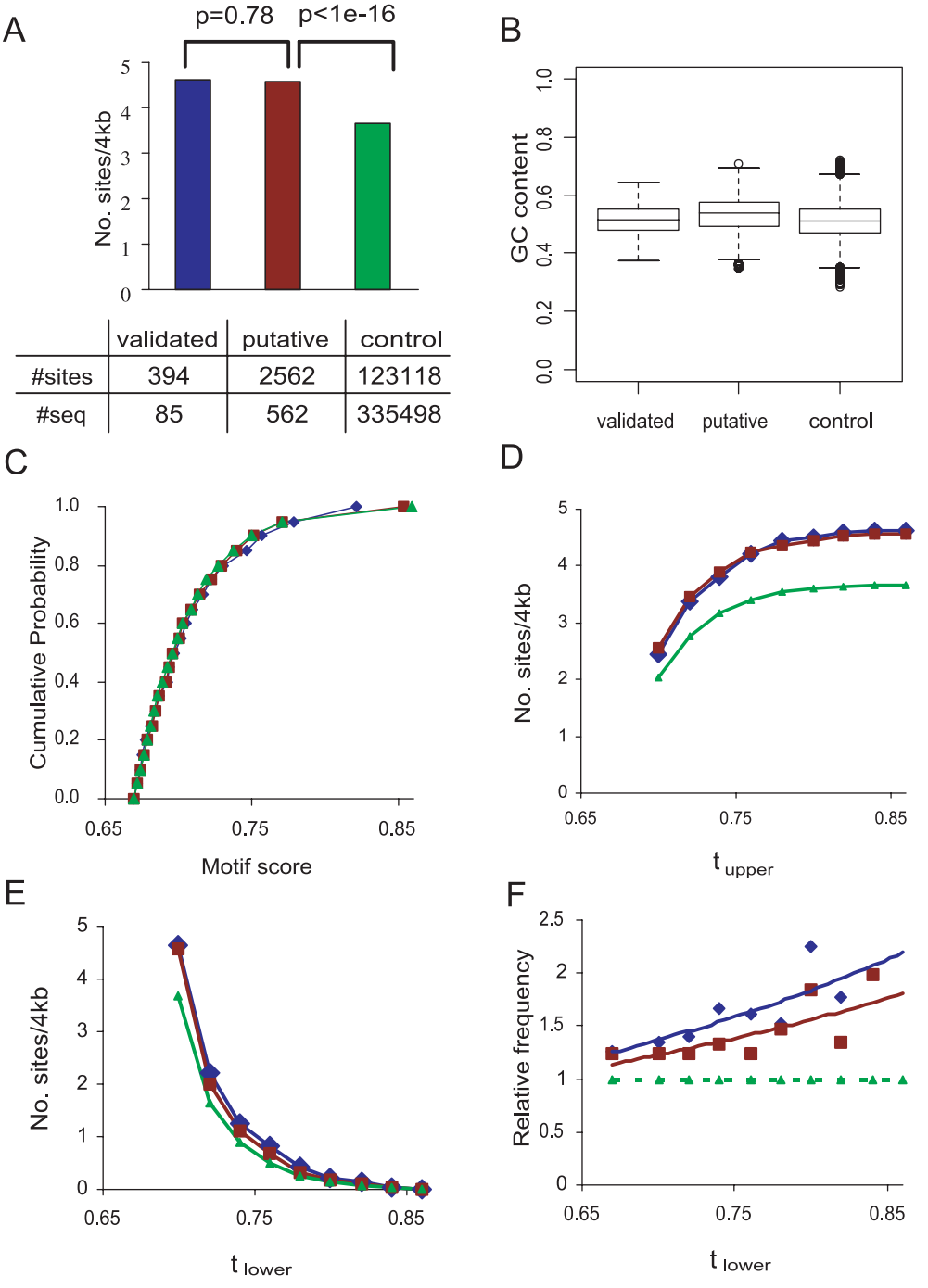

Figure 2. Highly-degenerate RE1s are enriched in REST target genes. (**A**) Enrichment of highly-degenerate RE1s in promoters of validated REST target genes (black), putative REST target genes (red) and control genes (green). The number of highly-degenerate RE1 sites and the number of sequences for each group are given in the chart (lower). The enrichment (average number of sites per sequence) is shown in the bar-plot. The enrichment is similar in validated and putative target genes ($P = 0.78$). In contrast, the enrichment is significantly higher in putative (and validated) REST target genes than in control genes ($P < 10^{-16}$). (**B**). The distribution of GC content for the three groups of promoters. (**C**). Distribution of motif scores of highly-degenerate RE1s derived from the three groups of promoters are shown in cumulative probability functions. (**D**) Enrichment of highly-degenerate RE1s in the three groups of genes changes as a function of the parameter t_{upper}. (**E**). Enrichment of highly-degenerate RE1s in the three groups of genes changes as a function of the parameter t_{lower}. (**F**). The relative enrichment of highly-degenerate RE1s in validated and putative REST targets normalized by the controls. The color coding in (**C**–**F**) is the same as in (**A**).

validated targets, putative targets and controls. The numbers of highly-degenerate RE1s in the three groups are shown in Figure 2A. The frequency of highly-degenerate RE1s is very similar among validated targets (4.6 sites per promoter) and putative targets (4.6 sites per promoter) ($P = 0.78$). In contrast, the frequency is significantly higher in putative REST targets (and also in validated REST targets) than in random controls (3.7 sites per promoter) ($P < 10^{-16}$) (Figure 2A). The three groups of promoters have similar GC content (validated: 52 ± 6%, putative: 53 ± 7%, controls: 51 ± 6%, reported as mean ± std) (Figure 2B). This eliminates the possibility that the difference in site frequencies between REST targets and controls is simply due to the difference in the GC content.

It was important to examine whether this enrichment of highly-degenerate sites in validated and putative REST targets was an artifact of parameter choices. In particular, if the upper threshold of highly-degenerate RE1s, t_{upper}, was too high, a functional cognate site might also be classified as being highly-degenerate. In this case, if the REST targets were significantly enriched in RE1s of moderately high motif scores (but below t_{upper}), the higher frequency of highly-degenerate RE1s might simply reflect the clustering of functional cognate RE1s in REST targets. These homotypic *cis*-regulatory modules could also promote cooperative binding, thereby improving the specificity of target site selection. Although not tested in this study, to exclude this possibility, the distributions of motif scores of highly-degenerate RE1s were compared. If the validated targets and the putative targets are disproportionally enriched with moderately high-scoring RE1s, this would be reflected in the difference in the distributions of the motif scores. The three groups, namely validated targets, putative targets and controls, have very similar distributions ($P = 0.26/0.30/0.58$ between each pair, Kolmogorov–Smirnov test) (Figure 2C). Indeed, the enrichment of highly-degenerate RE1s in REST targets persists with a wide range of different thresholds (t_{upper} from 0.7 to 0.86, Figure 2D). In particular, when thresholds $t_{upper} = 0.70$ and the same $t_{lower} = 0.67$ were used, the frequency of highly-degenerate RE1s is still significantly higher in validated targets (2.43 sites per promoter) and putative targets (2.56 sites per promoter) than in controls (2.04 sites per promoter) (validated versus control, $P = 0.01$; putative versus control, $P < 10^{-16}$). Therefore, even for very degenerate RE1s, the frequency is significantly higher in REST targets than in controls.

However, when the lower score $t_{lower} = 0.67$ is used (t_{upper} is fixed at 0.86), the relative frequency of highly-degenerate RE1s in REST targets and controls ($4.6/3.7 = 1.2$) is moderate, although significant statistically. This is because the threshold is extremely low. We would expect many of them are random matches and do not play functional roles. As the threshold increases, a larger fraction of matches would not be due to chance and more likely to be functional. This is indeed the case. As t_{lower} increases from 0.67 to 0.85, the relative frequency increases from 1.2 to 2 (Figure 2F). For example, when $t_{lower} = 0.75$, the relative frequency is about 1.5. However, this threshold still defines very degenerate sites, which are ubiquitous throughout the genome (~0.5 sites/4 kb, Figure 2E).

Conservation of highly-degenerate RE1s

We further reasoned that sequences conserved across species were more likely to be functional because random mutations were eliminated during evolution. Previous studies have reported that functional regulatory elements tend to be conserved when compared to a random motif or to neutral background, as is true for REST [see ref. (8) and Supplementary Table S1]. Therefore, we focused on those highly-degenerate RE1s conserved across species.

Due to the difficulty of aligning divergent promoters, our definition of a conserved motif does not require sequence alignment, but is derived from direct motif sequence comparisons. This is possible because the RE1 is relatively long. Briefly, three sites, one from each promoter in the same orthologous triplet, are considered to be orthologous if they have very similar motif sequences, locations relative to TSS and the same orientation (see Materials and Methods). To evaluate the effectiveness of our method, we first applied this approach to the CSHLmpd promoters to identify conserved high-score RE1s. Among the 206/190/199 high-score RE1s in human/mouse/rat, 76 RE1s (38%) from 74 triplets were conserved across the three species (two triplets have two conserved RE1s each). A complete list of the 47 genes with existing annotations is shown in Supplementary Table S3. Most of them are expressed specifically in neuronal cells, which is consistent with the known role for REST. Among them, GLRA2, GABRG2 and SYP are well established REST targets (22,23). Others carry the core RE1 motif proven to have specific binding affinity to REST *in vitro* by gel shift (24). Of the 23 high-score RE1s verified to have binding affinity in gel shift, 19 sites were conserved across human, mouse and rat; in contrast, among the seven high-score RE1s that showed no binding affinity in gel shift, only one was conserved. This comparison suggests that functional RE1s are more likely to be conserved than the non-functional RE1s ($P = 5.1 \times 10^{-7}$). Therefore, the constraint of species conservation greatly improves the specificity of cognate site prediction. Further examination shows that more RE1s (48 of 76, 63%) are located downstream of TSS, which is consistent with the previous observation that RE1s occurs more often in the 5′-untranslated region (5′-UTR) (22).

After validating the method, the conservation of highly-degenerate RE1s was examined. Three datasets representing the experimentally validated REST targets, the putative REST targets and the controls are constructed as follows. For the validated targets, 22 triplets with orthologous promoters in human, mouse and rat are used. For the putative targets, 74 triplets from the orthologous CSHLmpd promoters with conserved high-score RE1s are used. For the controls, the triplets of orthologous CSHLmpd promoters without high-score RE1s in any of the three species are used. It should be noted that high-score RE1s were removed before searching for conserved highly-degenerate RE1s. The significance of motif conservation is measured essentially as described previously (8). The conservation rate is calculated from the ratio of conserved motif occurrences to the average of overall motif occurrences in human, mouse and rat. We first compare the conservation rate of the highly-degenerate RE1s in the three datasets. As shown in Figure 3A, the conservation rate is similar for the validated targets and the putative targets (0.117 versus 0.123,

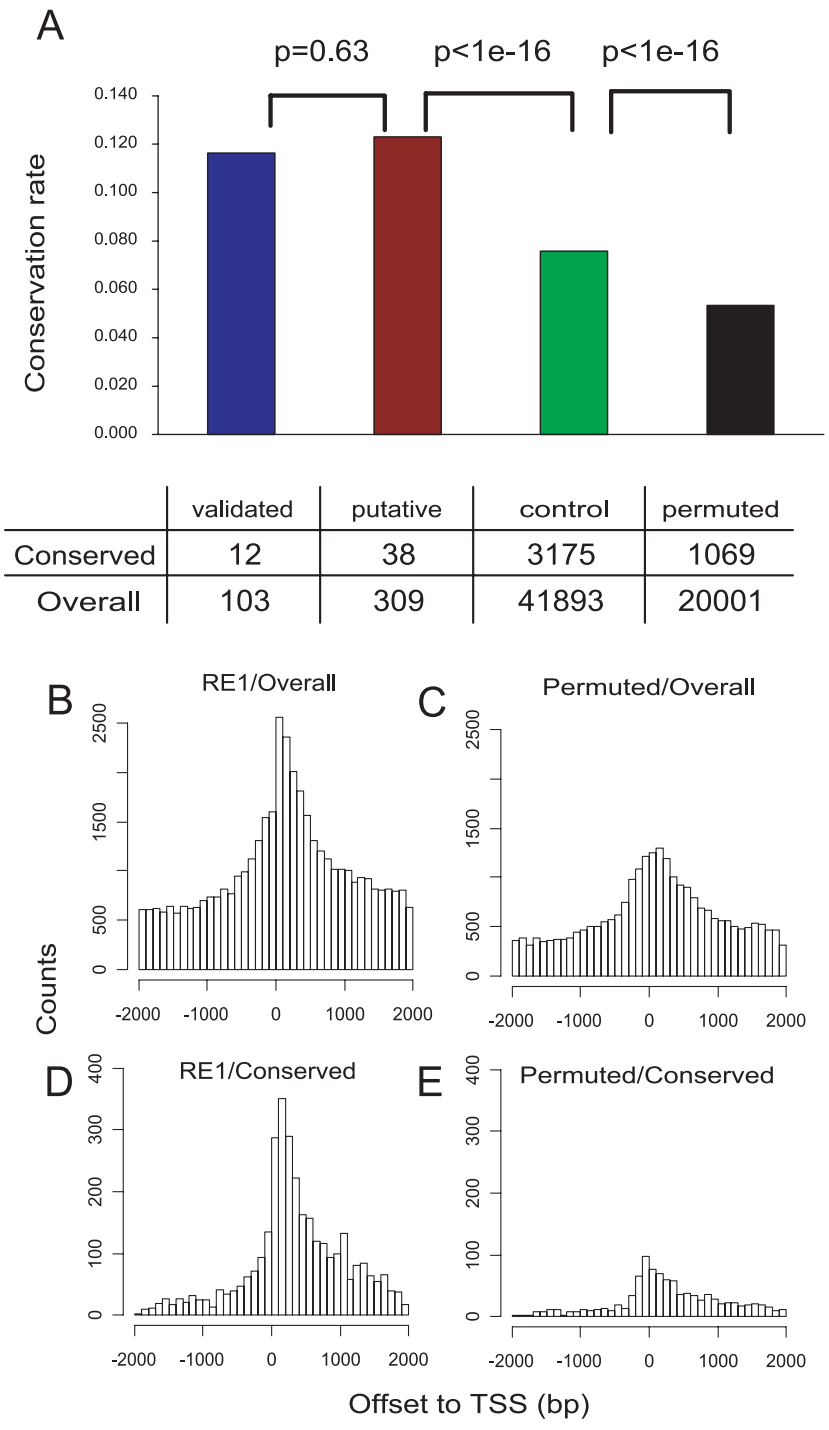

Figure 3. Conservation of highly-degenerate RE1s and random motifs in human, mouse and rat. (**A**) Conservation rate of highly-degenerate RE1s for validated REST targets (blue bar), putative REST targets (red bar), controls (green bar) and the conservation rate of highly-degenerate random motifs (black bar). The chart (bottom) shows the number of conserved highly-degenerate RE1s and overall highly-degenerate RE1s before and after motif permutation. Conservation rate is measured by the ratio of the number of conserved highly-degenerate RE1s to overall number of highly-degenerate RE1s. The conservation of highly-degenerate RE1s is similar in validated and putative REST targets ($P = 0.63$). In contrast, highly-degenerate RE1s in REST targets are significantly more conserved than those in control promoters ($P < 10^{-16}$), which are in turn significantly more conserved than random motifs ($P < 10^{-16}$). (**B–E**) Distribution of the location of highly-degenerate motifs relative to TSS. (**B**) Overall highly-degenerate RE1s. (**C**) Overall highly-degenerate permutedRE1s. (**D**) Conserved highly-degenerate RE1s. (**E**) Conserved highly-degenerate permuted RE1s.

$P = 0.63$). In contrast, the conservation rate in the REST targets is significantly higher than that in the controls, which is 0.076 ($P < 10^{-16}$). The conservation rate is then compared with that of random motifs which are derived by permuting the real motif. The conservation rate of random motifs (in all CSHLmpd triplets) is 0.053. This conservation rate is even lower compared with that of highly-degenerate sites without permutation in controls ($P < 10^{-16}$). Therefore,

the conservation of the highly-degenerate RE1s in the REST targets has a ~50% increase compared with the control promoters and a ~120% increase compared with random motifs.

The distribution of highly-degenerate RE1s and conserved highly-degenerate RE1s with respect to TSS is shown in Figure 3B and D. The distribution of permuted RE1s is also shown in Figure 3C and E. There was a peak of occurrence frequency near TSS, partially due to the fact that RE1 is a GC rich TFBS and the statistical elevation of GC content around TSSs (29). However, the peak of highly-degenerate RE1s was significantly higher than that of the permuted version. Moreover, the peak of the highly-degenerate RE1s was shifted to the downstream of TSS, which is consistent with the distribution of validated RE1s and barely seen in the permutations. Under the constraint of conservation, most of the occurrences far from TSS are automatically eliminated. These observations taken together suggest that the highly-degenerate RE1s are positively selected and fixed during evolution. Therefore, they are likely to play important roles in protein–DNA interactions and transcriptional regulation. The motif pictogram derived from the conserved highly-degenerate RE1s is shown in Figure 1C.

Experiments with other TFs

The results detailed above are based on REST and its binding element RE1. However, they point to the importance of the unique landscape provided by highly-degenerate binding sites as another general mechanism for protein–DNA interactions. To assess this probability, we examined the distribution of highly-degenerate binding sites for several other representative TFs including c-myc, p53, HNF-1 and CREB. These TFs have distinct functional roles and are expressed in a variety of different tissues. The binding sites for these TFs also differ in terms of their motif characteristics including palindromic pattern, motif size and base composition. We did not attempt to distinguish high-score sites from highly-degenerate sites for these motifs for reasons explained in Materials and Methods. Instead, a very low threshold was employed to define 'relaxed' binding sites for each TF (Supplementary Table S2). To correct for the fact that a larger fraction of promoters which are the targets for a TF are expected to bear corresponding TFBSs than control promoters, the enrichment of relaxed TFBSs was measured using only promoters carrying corresponding relaxed sites. Therefore, the enrichment here measures only the clustering effect of relaxed sites. For all of these TFs tested, highly-degenerate binding sites were significantly more enriched in the promoter region of validated or putative target genes (for p53, c-myc, HNF-1) or around potential *in vivo* binding sites (for CREB) than expected by chance (Figure 4). Therefore, the enrichment of highly-degenerate sites around cognate sites is not specific for REST binding sites and is likely to be a general phenomenon (at least) in mammalian genomes.

DISCUSSION

Due to the relative short length of TFBSs, it is unclear how the robustness and efficiency of TF-DNA (or more generally, protein–DNA) interactions is achieved. The local genomic context around the cognate binding site allows for combinatorial regulation and epigenetic modifications, which have been shown to be important for regulating functional activity. In this study, we find that the highly-degenerate TFBSs are not distributed randomly but rather are enriched significantly around the cognate sites. They are also conserved across species suggesting the evolutionary selection pressure. These findings suggest that the highly-degenerate TFBSs play important roles in the accurate and efficient recognition of the cognate site. Several possibilities, which are not necessarily mutually exclusive, can be envisioned to explain the functional importance of the TFBS sites. For example, in one scenario, they could serve as 'backup' sites, which could be activated if the cognate site was mutated (30). The extreme degeneracy of these TFBSs, however, likely precludes their ability to interact specifically with TFs. In another scenario, multiple copies of TFBS close to each other could form a homotypic *cis*-regulatory module, which could interact with multimerized TFs. In this case, all copies would have a relatively high motif score and would be separated by relatively small distances. This scenario does not explain the observation that highly-degenerate TFBSs can extend hundreds of base pairs beyond cognate sites (data not shown). Additionally, not all TFs function as multimers. For example, there is no evidence that REST can multimerize with itself to facilitate cooperative binding.

We postulate that the nonrandom distribution of highly-degenerate TFBSs throughout the genome provides a unique and favorable genomic landscape for facilitating a protein's ability to identify its cognate binding site. In some cases, the importance of highly-degenerate sites clustering around cognate sites has been validated experimentally. For example, the promoters of several REST target genes contain multiple RE1 sites with varying homologies to the consensus and the clustering of these sites appears to enhance REST occupancy (24,31,32). Qiang *et al*. (31) demonstrated further that deletion of the secondary, highly-degenerate RE1 sites partially derepressed the target gene. This hypothesis takes on more significance with regard to the advances in theoretical modeling of protein–DNA interactions over the past two decades. The canonical view of molecular interactions is 3D diffusion. Protein is frequently released from the DNA, diffuses in solution and then rebinds to another genomic locus. However, the rate of target site recognition measured experimentally by Riggs *et al*., in the case of LacI repressor and its operator on DNA, is 1000 times faster than the theoretic limit imposed by the 3D diffusion model (33,34). To resolve this discrepancy, the 'sliding' model was proposed (35–37) and later received experimental support (38). In this model protein walks forward or backward along the DNA without disassociating from the DNA. Recently, several groups suggested that optimal search efficiency is achieved by a combination of sliding and 3D diffusion (39–41). According to these models, protein contacts DNA with low affinity, quickly scans a DNA segment and then hops to other random loci until the target site is successfully located. This was experimentally validated to be true at least in a particular case (42). Our findings of the nonrandom distribution of highly-degenerate TFBSs are consistent with these recent results and provide a possible mechanism of TF-DNA or protein–DNA interactions. Compared with the 3D diffusion model, the local context of highly-degenerate binding sites is more crucial for sliding models because of synergistic effects. Highly-degenerate binding

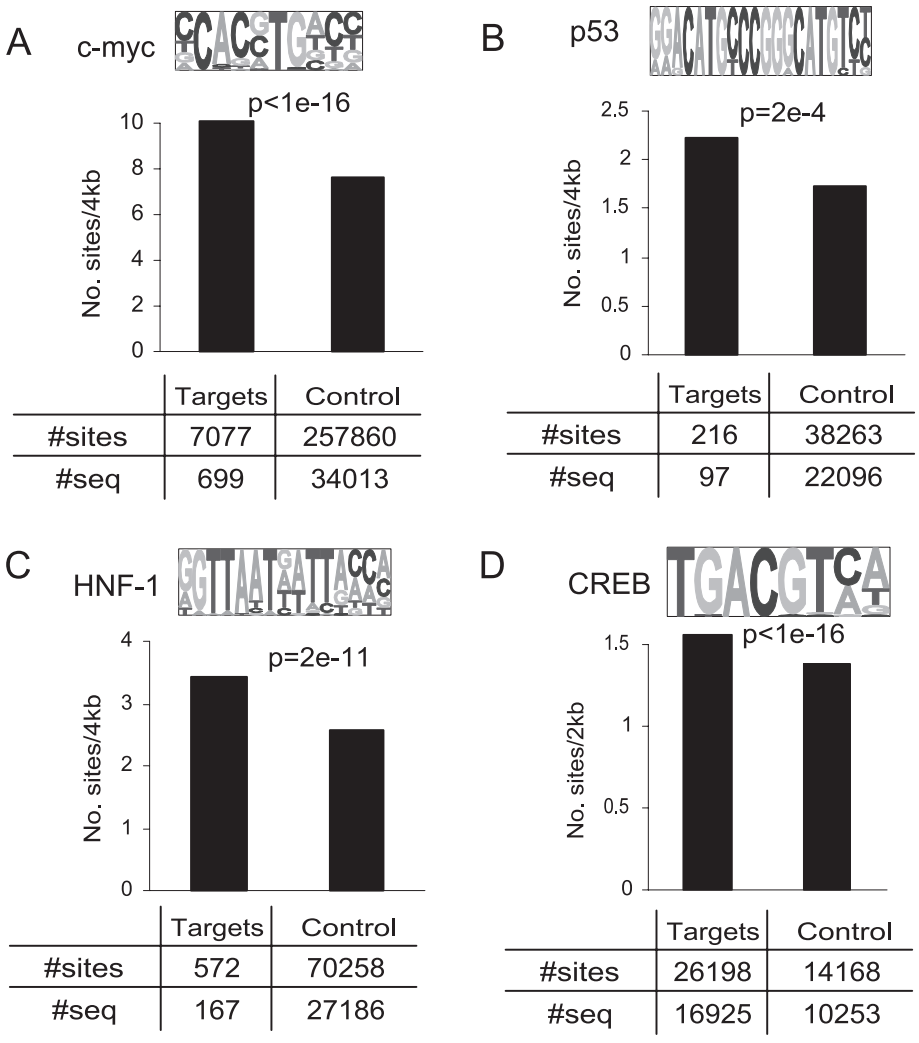

Figure 4. Enrichment of highly-degenerate binding sites for c-myc (**A**), p53 (**B**), HNF-1 (**C**) and CREB (**D**) in promoters of the target genes and control genes. For each Panel, the number of sites and sequences is shown in the chart. The enrichment (average number of sites per sequence) is shown in the bar-plot. The difference of site enrichment between two the groups is tested by χ^2 test. The resulting *P*-values are also shown in the bar-plot.

sites near cognate TFBSs may slow down the protein sliding process as the corresponding TFs approach the binding loci, which could help enrich the concentration of TFs within the neighborhood. At the same time, the interaction between TFs and their highly-degenerate binding sites may also alter the 3D conformation of protein and/or the DNA, thus changing the accessibility of the cognate sites. However, more experimental evidence is required to confirm this theory.

In addition to contributing to a mechanistic understanding of protein–DNA interactions, this study has another potential important application. Current TFBS prediction algorithms suffer greatly from a high rate of false positives. Incorporation of sequence context information generally improves prediction accuracy (43–47). The nonrandom distribution of highly-degenerate sites is another source of information for these algorithms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank A.D. Smith for helpful comments of the manuscripts. The authors also thank anonymous reviewers for their constructive suggestions to improve the manuscript. This work was supported by a grant from the National Institutes of Health to G.M. and by National Institutes of Health to M.Q.Z. under grant HG01696. G.M. is an Investigator of the Howard Hughes Medical Institute. C-Z. is also partly supported by the Dart Neurogenomic Alliance. Funding to pay the Open Access publication charges for this article was provided by the NIH grant HG01696.

Conflict of interest statement. None declared.

REFERENCES

1. Lewin,B. (2003) *Genes VIII*. Prentice Hall, pp. 597–614.
2. Galas,D. and Schmitz,A. (1978) DNAse footprinting: a simple method for the detection of protein–DNA binding specificity. *Nucleic Acids Res.*, **5**, 3157–3170.
3. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.-B., Reynolds,D.B., Yoo,J. *et al.*

(2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
4. Impey,S., McCorkle,S.R., Cha-Molstad,H., Dwyer,J.M., Yochum,G.S., Boss,J.M., McWeeney,S., Dunn,J.J., Mandel,G. and Goodman,R.H. (2004) Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell*, **119**, 1041–1054.
5. Laniel,M.A., Beliveau,A. and Guerin,S.L. (2000) Electrophoretic mobility shift assays for the analysis of DNA-protein interactions. *Methods Mol. Biol.*, **148**, 13–30.
6. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
7. Sabo,P.J., Hawrylycz,M., Wallace,J.C., Humbert,R., Yu,M., Shafer,A., Kawamoto,J., Hall,R., Mack,J., Dorschner,M.O. *et al.* (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl Acad. Sci USA*, **101**, 16837–16842.
8. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′-UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
9. Matlin,A.J., Clark,F. and Smith,C.W.J. (2005) Understanding alternative splicing: towards a cellular code. *Nature Rev. Mol. Cell. Biol.*, **6**, 386–398.
10. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins : statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–743.
11. Yuan,G.-C., Liu,Y.-J., Dion,M.F., Slack,M.D., Wu,L.F., Altschuler,S.J. and Rando,O.J. (2005) Genome-scale identification of nucleosome positions in *S.cerevisiae*. *Science*, **309**, 626–630.
12. Ioshikhes,I.P. and Zhang,M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nature Genet.*, **26**, 61–63.
13. Istrail,S. and Davidson,E.H. (2005) Gene regulatory networks special feature: logic functions of the genomic cis-regulatory code. *Proc. Natl Acad. Sci USA*, **102**, 4954–4959.
14. Marino-Ramirez,L., Spouge,J.L., Kanga,G.C. and Landsman,D. (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res.*, **32**, 949–958.
15. Von Hippel,P.H. and Berg,O.G. (1986) On the specificity of DNA–Protein interactions. *Proc. Natl Acad. Sci USA*, **83**, 1608–1612.
16. Xuan,Z., Zhao,F., Wang,J., Chen,G. and Zhang,M. (2005) Genome-wide promoter extraction and analysis in human, mouse, and rat. *Genome Biol.*, **6**, R72.
17. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
18. 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(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
19. Sokal,R.R. and Rohlf,F.J. (1995) *Biometry*. W. H. Freeman and Company, NY, pp. 695–697.
20. Mori,N., Stein,R., Sigmund,O. and Anderson,D.J. (1990) A cell type-preferred silencer element that controls the neural-specific expression of the Scg10 gene. *Neuron*, **4**, 583–594.
21. Kraner,S.D., Chong,J.A., Tsay,H.J. and Mandel,G. (1992) Silencing the type-Ii sodium-channel gene: a model for neural-specific gene-regulation. *Neuron*, **9**, 37–44.
22. Schoenherr,C.J., Paquette,A.J. and Anderson,D.J. (1996) Identification of potential target genes for the neuron-restrictive silencer factor. *Proc. Natl Acad. Sci USA*, **93**, 9881–9886.
23. Lunyak,V.V., Burgess,R., Prefontaine,G.G., Nelson,C., Sze,S.-H., Chenoweth,J., Schwartz,P., Pevzner,P.A., Glass,C., Mandel,G. *et al.* (2002) Corepressor-dependent silencing of chromosomal regions encoding neuronal genes. *Science*, **298**, 1747–1752.
24. Bruce,A.W., Donaldson,I.J., Wood,I.C., Yerbury,S.A., Sadowski,M.I., Chapman,M., Gottgens,B. and Buckley,N.J. (2004) Genome-wide analysis of repressor element 1 silencing transcription factor/ neuron-restrictive silencing factor (REST/NRSF) target genes. *Proc. Natl Acad. Sci USA*, **101**, 10458–10463.
25. Chong,J.H.A., Tapiaramirez,J., Kim,S., Toledoaral,J.J., Zheng,Y.C., Boutros,M.C., Altshuller,Y.M., Frohman,M.A., Kraner,S.D. and Mandel,G. (1995) Rest: a mammalian silencer protein that restricts sodium-channel gene-expression to neurons. *Cell*, **80**, 949–957.
26. Westbrook,T.F., Martin,E.S., Schlabach,M.R., Leng,Y., Liang,A.C., Feng,B., Zhao,J.J., Roberts,T.M., Mandel,G., Hannon,G.J. *et al.* (2005) A genetic screen for candidate tumor suppressors identifies REST. *Cell*, **121**, 837–848.
27. Zuccato,C., Tartari,M., Crotti,A., Goffredo,D., Valenza,M., Conti,L., Cataudella,T., Leavitt,B.R., Hayden,M.R., Timmusk,T. *et al.* (2003) Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nature Genet.*, **35**, 76–83.
28. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
29. Yamashita,R., Suzuki,Y., Sugano,S. and Nakai,K. (2005) Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene*, **350**, 129–136.
30. Simpson,P. (2002) Evolution of development in closely related species of flies and worms. *Nature Rev. Genet.*, **3**, 907.
31. Qiang,M., Rani,C.S.S. and Ticku,M.K. (2005) Neuron-restrictive silencer factor regulates the N-methyl-D-aspartate receptor 2B subunit gene in basal and ethanol-induced gene expression in fetal cortical neurons. *Mol. Pharmacol.*, **67**, 2115–2125.
32. Nakatani,T., Ueno,S., Mori,N. and Matsuoka,I. (2005) Role of NRSF/ REST in the molecular mechanisms regulating neural-specific expression of trkC/neurotrophin-3 receptor gene. *Brain Res. Mol. Brain Res.*, **135**, 249–259.
33. Riggs,A.D., Bourgeoi,S. and Cohn,M. (1970) Lac repressor-operator interaction 3. Kinetic Studies. *J. Mol. Biol.*, **53**, 401–417.
34. Riggs,A.D., Suzuki,H. and Bourgeoi,S. (1970) Lac repressor-operator interaction 1. Equilibrium studies. *J. Mol. Biol.*, **48**, 67–83.
35. Berg,O.G., Winter,R.B. and Von Hippel,P.H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids .1. *Models and Theory Biochem.*, **20**, 6929–6948.
36. Richter,P.H. and Eigen,M. (1974) Diffusion controlled reaction-rates in spheroidal geometry—application to repressor-operator association and membrane-bound enzymes. *Biophys. Chem.*, **2**, 255–263.
37. Von Hippel,P.H. and Berg,O.G. (1989) Facilitated target location in biological-systems. *J. Biol. Chem.*, **264**, 675–678.
38. Kim,J.G., Takeda,Y., Matthews,B.W. and Anderson,W.F. (1987) Kinetic-studies on cro repressor operator DNA interaction. *J. Mol. Biol.*, **196**, 149–158.
39. Slutsky,M. and Mirny,L.A. (2004) Kinetics of Protein–DNA interaction: facilitated target location in sequence-dependent potential. *Biophys. J.*, **87**, 4021–4035.
40. Halford,S.E. and Marko,J.F. (2004) How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.*, **32**, 3040–3052.
41. Gerland,U., Moroz,J.D. and Hwa,T. (2002) Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc. Natl Acad. Sci USA*, **99**, 12015–12020.
42. Gowers,D.M., Wilson,G.G. and Halford,S.E. (2005) Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA. *Proc. Natl Acad. Sci USA*, **102**, 15883–15888.
43. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci USA*, **99**, 757–762.
44. Gupta,M. and Liu,J.S. (2005) *De novo cis*-regulatory module elicitation for eukaryotic genomes. *Proc. Natl Acad. Sci USA*, **102**, 7079–7084.
45. Kel,A.E., Kel-Margoulis,O.V., Farnham,P.J., Bartley,S.M., Wingender,E. and Zhang,M.Q. (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. *J. Mol. Biol.*, **309**, 99–120.
46. 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**, i440–i448.
47. Wagner,A. (1998) Distribution of transcription factor binding sites in the yeast genome suggests abundance of coordinately regulated genes. *Genomics*, **50**, 293–295.
48. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.