Published online 13 April 2008

Nucleic Acids Research, 2008, Vol. 36, No. 10 ***3171–3184***
doi:10.1093/nar/gkn057

CSI-Tree: a regression tree approach for modeling binding properties of DNA-binding molecules based on cognate site identification (CSI) data

Sündüz Keleş[1,*], Christopher L. Warren[2], Clayton D. Carlson[2] and Aseem Z. Ansari[2,3]

[1]Department of Statistics, Department of Biostatistics and Medical Informatics, [2]Department of Biochemistry and [3]The Genome Center, University of Wisconsin, Madison WI, USA

Received November 6, 2007; Revised January 28, 2008; Accepted January 29, 2008

ABSTRACT

The identification and characterization of binding sites of DNA-binding molecules, including transcription factors (TFs), is a critical problem at the interface of chemistry, biology and molecular medicine. The Cognate Site Identification (CSI) array is a high-throughput microarray platform for measuring comprehensive recognition profiles of DNA-binding molecules. This technique produces datasets that are useful not only for identifying binding sites of previously uncharacterized TFs but also for elucidating dependencies, both local and nonlocal, between the nucleotides at different positions of the recognition sites. We have developed a regression tree technique, CSI-Tree, for exploring the spectrum of binding sites of DNA-binding molecules. Our approach constructs regression trees utilizing the CSI data of unaligned sequences. The resulting model partitions the binding spectrum into homogeneous regions of position specific nucleotide effects. Each homogeneous partition is then summarized by a position weight matrix (PWM). Hence, the final outcome is a binding intensity rank-ordered collection of PWMs each of which spans a different region in the binding spectrum. Nodes of the regression tree depict the critical position/nucleotide combinations. We analyze the CSI data of the eukaryotic TF Nkx-2.5 and two engineered small molecule DNA ligands and obtain unique insights into their binding properties. The CSI tree for Nkx-2.5 reveals an interaction between two positions of the binding profile and elucidates how different nucleotide combinations at these two positions lead to different binding affinities. The CSI trees for the engineered DNA ligands exhibit a common preference for the dinucleotide AA in the first two positions, which is consistent with preference for a narrow and relatively flat minor groove. We carry out a reanalysis of these data with a mixture of PWMs approach. This approach is an advancement over the simple PWM model and accommodates position dependencies based on only sequence data. Our analysis indicates that the dependencies revealed by the CSI-Tree are challenging to discover without the actual binding intensities. Moreover, such a mixture model is highly sensitive to the number and length of the sequences analyzed. In contrast, CSI-Tree provides interpretable and concise summaries of the complete recognition profiles of DNA-binding molecules by utilizing binding affinities.

INTRODUCTION

Elucidating the recognition properties of DNA-binding molecules such as transcription factors (TFs) is among the most challenging problems in computational biology. The importance of this problem is 2-fold. First, better characterization of TF binding sites (TFBSs) leads to more accurate predictions of their genomic binding. This is critical for both identifying TF target genes and constructing genome scale regulatory networks (1). The second aspect is related to the ability to design synthetic molecules that target specific sites in the genome and regulate the expression of desired genes (2–4). A crucial requirement in the creation of synthetic transcriptional regulators is the ability to program, with great precision, their DNA targeting properties.

Until recently, most effort for characterizing binding sites of DNA-binding molecules, which are on the order of

*To whom correspondence should be addressed. Tel: +1 608 263 4533; Fax: +1 608 262 0032; Email: keles@stat.wisc.edu

5–20 base pairs (bp), focused on learning position weight matrix (PWM) models from unaligned DNA sequences. These unaligned sequences are typically grouped together via analysis of data from gene expression, chromatin immunoprecipitation on microrray (ChIP-chip) experiments or comparative genomic analysis (5–7). The PWM model (8) is the backbone of numerous commonly used motif finding algorithms (9,10). This model assumes independence among positions of the binding site and views each position as being sampled independently from a distinct multinomial distribution. Another formulation of this model is presented by Foat *et al.* (11) within the context of learning recognition profiles by regressing sequence data onto *in vivo* binding intensity data from ChIP-chip experiments. Recent experiments have shown that position specific nucleotides exert unanticipated local as well as nonlocal interdependent effects on the binding affinity of the TFs (12,13). Motivated by these studies, several new probabilistic models have been proposed (14–17). These models, often suitable for aligned sequences, e.g. known instances of TFBSs, use Bayesian networks (14), variants of Markov models (permuted variable order) (15), or variable order Bayesian networks (16) to reveal better descriptions of recognition profiles. Although the inadequacy of the independent PWM model has become clear, the unavailability of good training data hindered the applicability of this richer class of models.

In this article we consider the detailed characterization of binding sites using a new type of microarray platform called the Cognate Site Identification array (CSI array) (2). This platform provides the comprehensive sequence recognition profiles of DNA-binding molecules individually or in cooperatively interacting pairs. These data are comprehensive and genome-independent. Recently, Berger *et al.* (18) generated such comprehensive binding data for five different TFs from yeast, mouse and humans an efficient microarray design. In addition to identifying the recognition properties of natural TFs, the CSI approach is particularly invaluable for rationally engineering synthetic molecules to target specific sequences in any genome. Moreover, the CSI platform also makes the consideration of more complex binding site models feasible. Warren *et al.* (1) generated CSI data by displaying every permutation of a duplex DNA sequence up to eight positional variants on a microfabricated array and determined the affinity of a DNA-binding molecule for every sequence on the array in a rapid and unbiased manner. This technology yields all the short *unaligned* sequences bound by the molecule of interest, and hence creates an unprecedented opportunity for studying dependencies throughout the positions of the binding sites. Similar small-scale data has been produced for zinc finger proteins by investigating their binding affinities using all permutations of 3mers (19). These data were rigorously analyzed by Lee *et al.* (20) utilizing a linear analysis of variance (ANOVA) model and the statistically significant dependencies between various positions were revealed. However, a linear ANOVA model is not directly applicable in most CSI applications as it requires the input sequences to be aligned.

Probabilistic models extending the simple PWM model are applied in cases where known examples of the TFBSs (15,16) or unaligned sequences presumably containing TFBSs are available (14). Importantly, these models are designed for sequence data alone and do not utilize the binding affinities of the unaligned sequences. Therefore, extensions of these models or new classes of models incorporating quantitative binding information obtained from the CSI data are needed. To address this important challenge, we developed a regression tree method named CSI-Tree that utilizes unaligned sequences and their binding affinities to characterize the recognition profiles of DNA-binding molecules. The tree fitting process is embedded in an Expectation-Maximization algorithm (21) which aims to align the sequences based on both the sequence and CSI data. The final regression tree partitions the sequence recognition space in a supervised manner by incorporating the binding intensities. Each leaf node in the tree is summarized with a leaf-specific PWM, thereby creating a rank-ordered collection of PWMs for representing the full recognition spectrum of a DNA-binding molecule. Position-specific nucleotide combinations appearing at the nodes of the tree highlight the important differences within the collection of PWMs. We also explored the use of a previously proposed mixture of PWMs model (14,22) for the analysis of CSI data. Both the data analysis and detailed simulations indicate that this mixture model fails to capture the dependencies discovered by the use of actual binding intensities. In contrast, CSI-Tree produces a representation of the recognition profiles in terms of a collection of binding intensity rank-ordered PWMs.

RESULTS

We will denote the binding intensities observed in 4^L Lmers by $Y_1, \ldots, Y_N$ and their corresponding sequence data by the vector $\mathbf{X}_i = \{X_{i1}, \ldots, X_{iL}\}$, $i = 1, \ldots, N$, where $X_{il} \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. In two recent papers, Warren *et al.* (2) and Puckett *et al.* (23) showed that fluorescent intensities from CSI experiments are linearly proportional to the binding affinities measured by solution assays (See Supplementary Table 1 and Figure 1 for a comparison of the fluorescent intensities from CSI arrays and the binding affinities $\mathrm{K_a}$ for the PA2 ligand used in this article). The current CSI technology allows an L value of up to 12 base pairs and in our applications $L = 8, 9$ base pairs. One of the challenges in the analysis of the CSI data is that the width W of the binding site is smaller than L. If W equals L, we are back to a similar case of the zinc finger example (19), namely a standard factorial design where each of the L positions can be considered as a factor in this experiment. Our interest lies in relating the actual binding measurement Y_i to a subsequence of $\mathbf{X}_i$ and characterizing the contribution of each position. Let Z_i denote the unobserved start site of the binding site in sequence i. Note that if W equals L, then we have $Z_i = 1$, $\forall i$. We are interested in the expectation (E) of the binding intensities conditional on the sequence data

$$E(Y_i \mid Z_i, X_{i1}, \ldots X_{iL}) = f(X_{iZ_i}^{Z_i+W-1}), \qquad \mathbf{1}$$

where $X_{iZ_i}^{Z_i+W-1}$ denotes the subsequence of the i-th Lmer starting at position Z_i and ending at position $Z_i + W - 1$. In the the case of the zinc finger study (20), $Z_i = 1, \forall i$ and the mean binding intensity function f is a linear function of the positions. Our goal is to estimate f as nonparametrically as possible and capture high level interactions between the positions of the binding site. A popular nonparametric regression technique is tree-based regression (24), which iteratively partitions the feature space, i.e. sequence space, into homogeneous regions and uses a constant regression line within each homogeneous region. When the mean binding intensity function f is estimated by a regression tree, its functional form is piecewise linear or piecewise constant. A regression tree is built through a *binary recursive partitioning* process. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches. The algorithm first tries to partition the data using every possible binary split on the positions of the binding site. Then, it chooses the split that partitions the data into two parts such that the sum of the squared deviations from the means in the resulting branches is minimized. This partitioning is then applied to each of the new branches. The process continues until each node reaches a user-specified minimum node size and becomes a terminal node, i.e. leaf node. The final outcome of the tree fitting procedure is a binary split tree where each split is based on an *if then* logical condition. For example, in Figure 1A, the first split in the tree is $X1 = \mathrm{A, C, G}$. This split criterion corresponds to the question 'Is the nucleotide at the first position of the sequence an A, C or G?'. If a given sequence has an A, C or G in the first position, it is moved down the left branch of the tree (branch with a '+' sign) and the sequence with a T is moved down the right branch. This process is repeated at the new branches until a leaf node is reached. Finally, each leaf node reports the mean binding intensity of all the sequences in that leaf node. The resulting tree is typically displayed as in Figure 1A, where both the fitted values and the number of observations in each leaf node are reported. We note that, in regression trees, binary splitting, rather than multiway splitting, is preferred for computational convenience. Allowing more than two partitions at each node rapidly grows the number of operations required as the number of variables and the number of categories in each variable grow. However, as we show later with cross-validation experiments, binary splitting is not a serious limiting factor in this type of analysis.

We note that if we knew the motif start site within each Lmer, we could readily fit a regression tree to the binding data treating Y as the outcome and each position of the binding site as explanatory variables. However, neither the width of the binding site nor the start positions in each sequence are known *a priori*. Therefore, a regression tree fit is not readily applicable. To circumvent this problem, we design a pseudo Expectation-Maximization algorithm. This algorithm is motivated by the fact that if we knew the start position of the binding site within each Lmer,

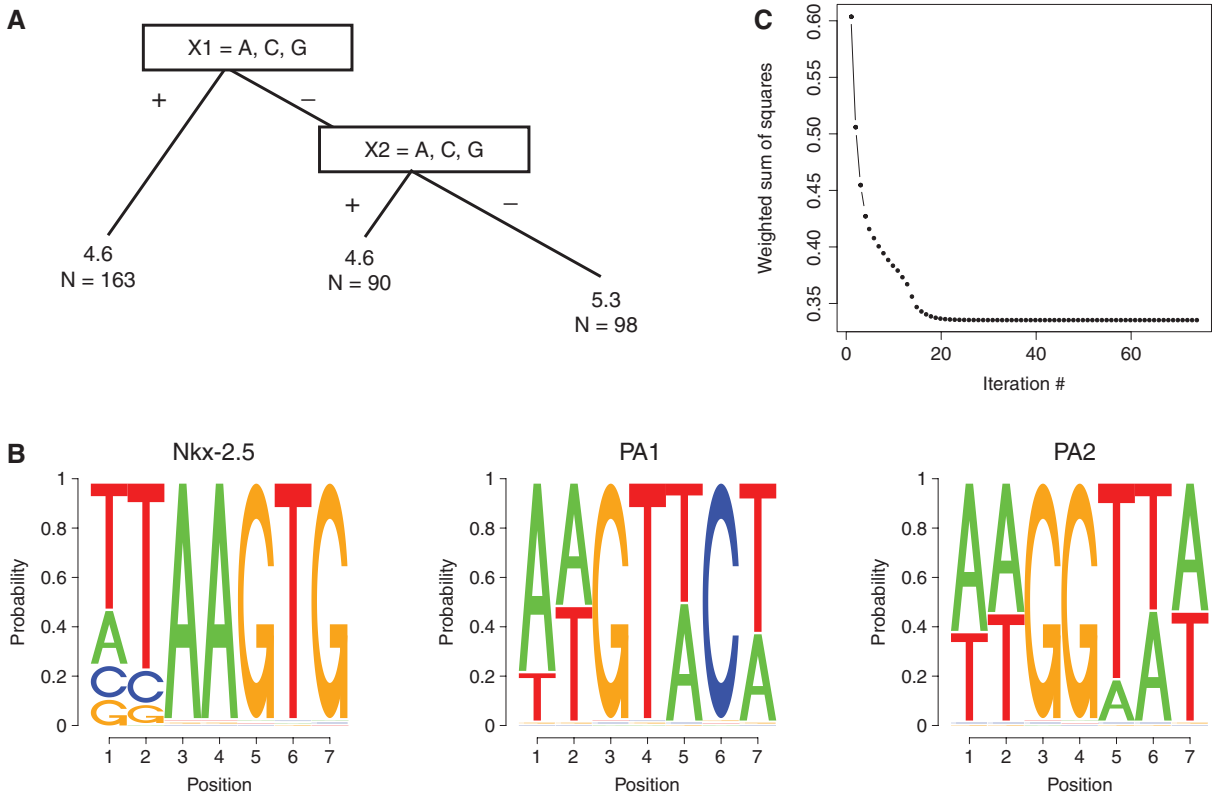

Figure 1. (**A**) An example of a regression tree with three leaf nodes. Number of observations n is displayed along with mean binding intensities of the leaf nodes. (**B**) Starting position weight matrices estimated by `cosmo` (26). (**C**) CSI-Tree-EM convergence for PA2. Mean weighted sum of squares is plotted as a function of the number of pseudo-EM iterations.

we could estimate f by a regression tree. Similarly, if we knew f, we could predict the most likely start site within each Lmer that would minimize the discrepancy between its observed and the tree-predicted binding intensities. This intuitive idea can be formalized within the framework of the following Gaussian model:

$$Y_i \mid Z_i, X_{i1}, \ldots, X_{iL} \sim \mathcal{N}(f(X_{iZ_i}^{Z_i+W-1}), \sigma^2), \quad \mathbf{2}$$

where f represents the, possibly logged, mean binding affinity as a function of the nucleotides within the binding site and Z_i, $i = 1, \ldots, n$ are unknown start positions. Within this model, the (t)-th iteration of the E-step is given by

$$\eta_{il}^{(t)} \equiv \Pr(Z_i = l \mid Y_i, \mathbf{X}_i, \hat{f}^{(t-1)}) \quad \mathbf{3}$$

$$= \frac{\exp\{-||Y_i - \hat{f}^{(t-1)}(X_{il}^{l+W-1})||^2\}}{\sum_{l'=1}^{L-W+1} \exp\{-||Y_i - \hat{f}^{(t-1)}(X_{il'}^{l'+W-1})||^2\}}, \quad \mathbf{4}$$

where $\hat{f}^{(t-1)}$ is the estimate of f from the $t-1$-th M-step. The (t)-th M-step for estimating f corresponds to the following maximization problem

$$\max_f \left(-\sum_{i=1}^{N} \sum_{l=1}^{L-W+1} \eta_{il}^{(t)} ||Y_i - f(X_{il}^{l+W-1})||^2 \right), \quad \mathbf{5}$$

where we consider f to be a tree function that partitions the binding spectrum into homogeneous regions utilizing the nucleotide composition at each position of the binding site. A solution to this M-step is achieved by estimating f with a regression tree. In our applications, we use R function (25) `rpart` to obtain such a tree estimate.

CSI-Tree algorithm

Since the CSI array contains double-stranded DNA, the binding site could be read from either of the strands. Allowing both strand information to be utilized by the CSI-Tree algorithm generates trees those are difficult to interpret due to the mixing of sites from forward and backward strands in the leaf nodes of the tree. Therefore, we fix the strand information with an initialization process as described below:

(1) *Preprocessing*. The CSI data is background corrected and normalized and the top N sequences are selected as described in the Methods Section and in (2). Since CSI-Tree algorithm utilizes the actual binding intensities, the choice of N can be flexible in the sense that the input sequences of the regression tree analysis can include oligonucleotides bound with a wide range of affinities.

(2) *Initialization*. Initialization step consists of running a *de novo* motif finding algorithm using the N Lmers from the preprocessing step. For the examples provided in this article, we utilized `cosmo` (26) allowing zero or one binding site occurrence in each Lmer. `cosmo` outputs an estimated PWM for the binding site enriched in the input sequences. Each subsequence within each Lmer is scored by summing the corresponding nucleotide specific contributions from the log transformed initial PWM. The strand with the highest scoring subsequence is selected from each Lmer for the CSI-Tree-EM step.

(3) *CSI-Tree-EM*. A regression tree is built using the N Lmer sequences as covariates and their CSI data as outcome in an iterative fashion. The pseudo-EM algorithm estimates the regression tree function f as follows. M-step implements the regression tree fit of minimum leaf size of 1. The start site posterior probabilities (η) of the E-step are used as weights in this step. This corresponds to using as many as $L - W + 1$ observations from each Lmer by adjusting their contribution to the regression tree fit by weights from the E-step.

(4) *CSI-Tree-Fit*. Once the pseudo-EM algorithm has converged to a single tree, subsequences from each Lmer are extracted using the positions with the highest posterior probability of being a start site to obtain a set of aligned sequences. Then, a regression tree is built using only these subsequences with their corresponding weights. The final regression tree is constrained to have at least 10 subsequences at each leaf and the optimal tree size is selected with 5-fold Monte Carlo cross-validation.

(5) *Leaf-specific PWM*. At each leaf node of the final tree, a PWM is constructed using the subsequences that are members of this leaf. Each subsequence is allowed to contribute to the PWM construction with a weight proportional to its binding intensity. The (a, j)-th position of the PWM at leaf node m equals

$$p_{aj} = \frac{\sum_{i=1}^{N} I(i \in \text{Leaf node } m) I(X_{i(\hat{Z}_i + j - 1)} = a) w_i}{\sum_{i=1}^{N} I(i \in \text{Leaf node } m) w_i},$$

where I represents the indicator function, $\hat{Z}_i$ is the predicted start site of the binding site in Lmer i and w_i represents the ratio of the binding intensity of the ith sequence to the maximum binding intensity observed in this leaf node. The indicator function $I(\ )$ takes on value 1 if the expression inside the parenthesis is true and 0 otherwise.

CSI-Tree-Fit step of the algorithm requires a minimum leaf node size as input. As the minimum number of sequences allowed in a leaf node decreases, it becomes possible to consider larger trees. We note that although this parameter will ultimately have an effect on the number of candidate trees to choose from, it does not automatically lead to over-fitting of the data as Monte Carlo cross-validation is used to select among the candidate tree sizes. As described above, the final summary from CSI-Tree consists of a PWM for each leaf node constructed from the sequences of the node. We empirically choose 10 as the minimum leaf node size since the PWMs built on too few sequences are not reliable and about 95% of the known PWMs in the recently curated JASPAR database (27) are based on at least 10 sequences. Moreover, in our applications (see next section), a large percentage of the leaf nodes have more than 20 sequences

indicating that splitting them further does not lead to an improvement in the model fit.

Applications

In this section, we investigate the operating characteristics and demonstrate the effectiveness of CSI-Tree algorithm with three applications. The datasets used in these applications correspond to DNA-binding by a mammalian transcription factor Nkx-2.5 and two synthetic DNA-binding molecules called *polyamides* (PA1 and PA2) (Figure 2). Nkx-2.5 is a NK-2 type homeodomain involved in heart development and is reported to bind to consensus 5′-TNNAGTG-3′ (N = A, C, G or T) (28). PA1 and PA2 hairpin polyamides are engineered to target specific DNA sequences using a recognition code based on minor groove hydrogen bonds (4, 29). The first polyamide (PA1) is designed to recognize the target sequence 5′-W WGWWCWW-3′ (W = A or T), whereas the second polyamide (PA-2) is engineered for 5′-WWGGWWW-3′. As a result of preprocessing (see Methods section), we used $N = 351$, 546 and 389 oligonucleotides from the CSI data of Nkx-2.5, PA1 and PA2 to build regression trees. CSI arrays for Nkx-2.5 and PA2 consisted of variable 9mers whereas PA1 CSI array was an 8mer array. The variable duplex regions are embedded in a 15-17mer duplex with constant 3bp flanking regions. These invariant flanking sequences buffer the variable core from thermal fraying of the duplex at one end and loop-induced structural distortions of the DNA hairpin at the other. The core 8 or 9mer duplex adopts a B-form DNA duplex that is indistinguishable from duplexes constructed from two linear complementary strands. In the analysis of Nkx-2.5 CSI data, we used 11mers extending the randomized 9mer oligonucleotides. As briefly mentioned above and explained in detail in the Methods section, each oligonucleotide appears on the array in the form of a DNA hairpin probe with constant flanking sequences 5′-CGC-3′ on either side. Since the last position of the Nkx-2.5 consensus is a G (as identified by cosmo (26) from 9mer sequences and is displayed in Figure 1B), the extended 11mers include the nucleotide C (reverse complement G) of the constant flanking sequences on either side of the randomized 9mer.

In what follows, we summarize general properties of the algorithm and elaborate on the biological and chemical implications of each of the case studies. PWMs to initialize the CSI-Tree algorithm are obtained by `cosmo` (26) and displayed in Figure 1B. These and the matrix pictorials at the leaf nodes of the trees are obtained by the logo plotting function `seqLogo` in the `R` package `cosmo` (26). The *y*-axes in these plots correspond to letter frequencies rather than the information content which is commonly used in sequence logos.

Convergence of the CSI-Tree algorithm

We investigate the convergence of the pseudo-EM algorithm to a stable f, i.e. regression tree, estimate. Figure 1C displays the mean weighted sum of squares as a

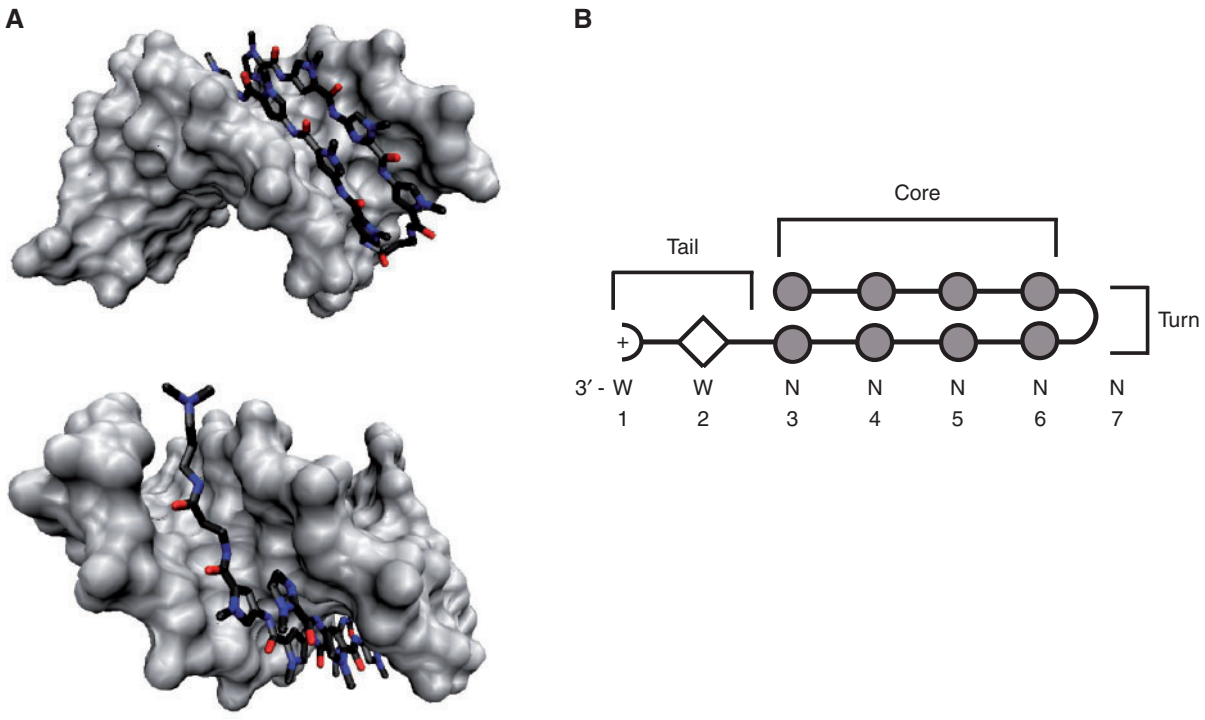

Figure 2. (**A**) Poylamide binding of DNA. Polyamides make hydrogen bond contacts along the base of the DNA minor groove. DNA is shown as a gray spacefill structure and polyamide is bound to DNA. For polyamide structure, red, blue and dark gray colors depict oxygen, nitrogen and carbon molecules, respectively [adapted from (45), PDB entry: 1m19]. (**B**) Schematic representation of polyamide interacting with DNA in minor groove. W represents A or T. N represents any nuclectide, gray circles depict heterocycle rings (typically pyrrole or imidizole), the diamond represents β-alanine, and + corresponds to a dimethylaminoproplyamide (See Supplementary Figure 2 for specific examples of chemical structures of polyamides used in this study).

function of the number of iterations in the pseudo-EM algorithm for PA2. As depicted in this figure, the algorithm converges to a stable tree in about 20 iterations. For PA1 and Nkx-2.5 datasets, the CSI-Tree-EM step did not alter the initial alignments of the sequences.

CSI-Tree analysis of the eukaryotic transcription factor Nkx-2.5

Our application with the CSI data of the eukaryotic TF Nkx-2.5 resulted in the tree given in Figure 3A. The tree size is based on Monte Carlo cross-validation with 5000 cross-validation iterations and the 1-standard-error rule. This 1-standard-error rule chooses the minimum size that has cross-validated error smaller than the minimum cross-validated error plus its SE. This rule is empirically shown to be more robust than using the size with minimum cross-validated error (30).

The regression tree depicted in Figure 3A can be summarized with the following algebraic expression:

$$\hat{Y} = 4.6I(X1, \in \{A, C, G\}) \quad [1]$$
$$+ 4.6I(X1, \in \{T\})I(X2 \in \{A, C, G\}) \quad [2]$$
$$+ 5.3I(X1, \in \{T\})I(X2 \in \{T\}), \quad [3]$$

where the numbers in square brackets after each expression represent the leaf node numbers in the tree going from left to right and the numbers before the indicator function are the mean binding intensities of the leaf nodes. In deriving the above expression, we utilized the displayed split rules in the tree. Nkx-2.5 CSI tree reveals that having

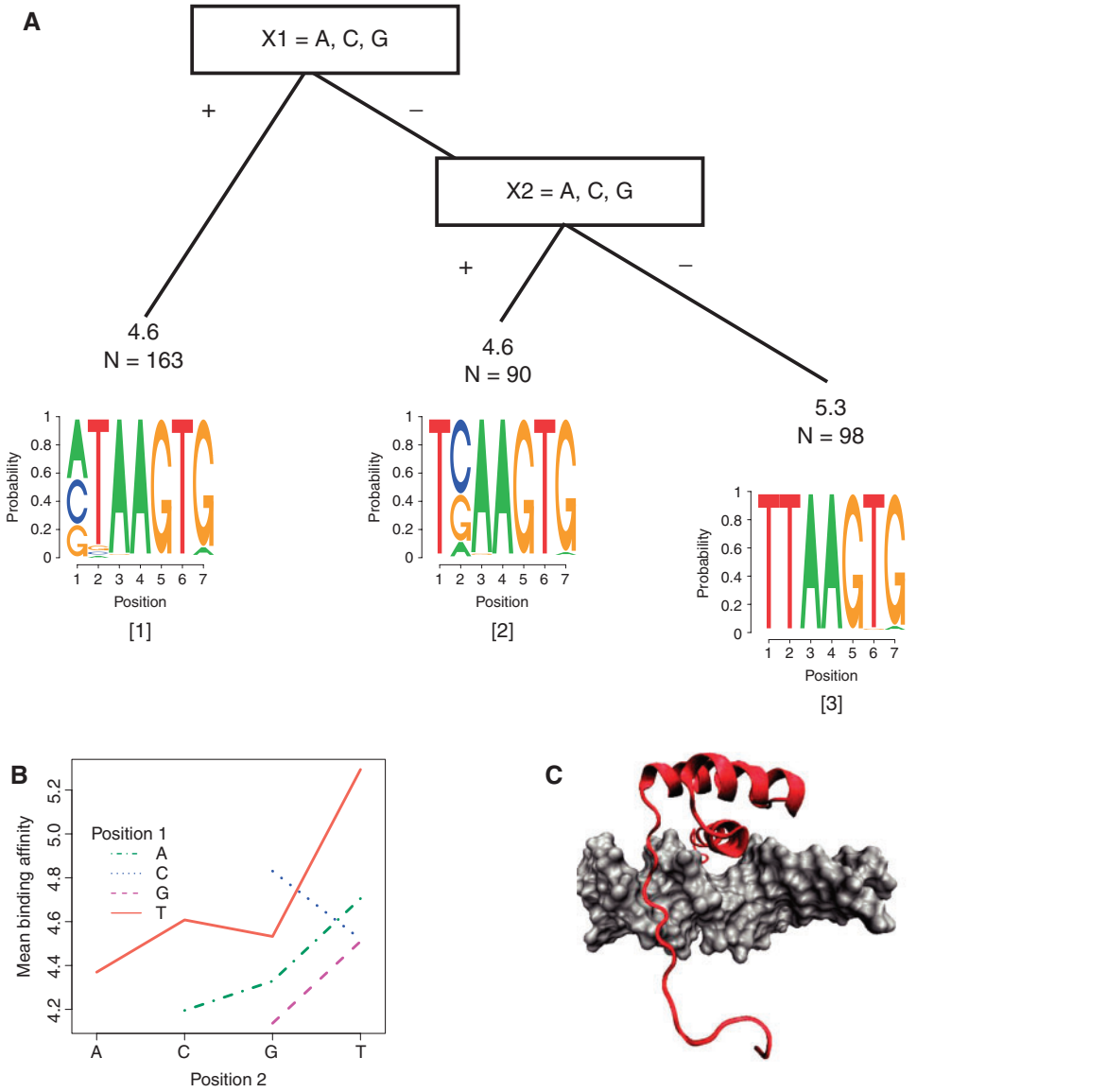

Figure 3. (**A**) CSI tree for the eukaryotic transcription factor Nkx-2.5. (**B**) Interaction plot for the first two positions of the Nkx-2.5 recognition profile. In generation of this plot, four subsequences, two with CA and two with CC in the first and second positions were excluded. Although, these had high intensities, we attributed these high intensities to noise due to the very small number of sequences (two in each group) in these groups. (**C**) Nucleic magnetic resonance (NMR) solution structure of a NK-2 type homeodomain (depicted in red color) bound to DNA (depicted in gray color) showing interactions in the minor groove with the first two nucleotides of the binding motif (PDB entry: 1NK-2).

a T in either the first or the second position leads to the similar binding affinities. The strongest binding affinities are observed when both of the positions bear T residues. This implies an interaction between these two positions. This interaction is also evident in the interaction plot of these positions as displayed in Figure 3B. The interaction plot displays the mean binding affinity at each level of the first position as a function of the nucleotide sequence at the second position. Highly nonparallel lines provide evidence for an interaction between the two positions and this is well supported with an analysis of variance (*P*-value of the corresponding *F*-test equals 9.20e-04).

The preference for T residues at the first and second positions revealed by CSI-Tree analysis is consistent with the known nuclear magnetic resonance (NMR) structures of the related NK-2 protein–DNA complex. Like most DNA-binding homeodomains, an alpha helix is inserted in the major groove in addition to conserved contacts and several buttressing interactions that drive affinity for DNA and a key tyrosine residue serves a major specificity determinant. In addition, the flexible and highly positively charged N-terminal arm of the homeodomain makes electrostatic interactions with the minor groove. In the structure it is apparent that the two residues (Lysine3-K3 and Arginine5-R5) insert into the minor groove and make contacts with the base edge (Figure 3C) (31–33). While R5 interaction with a T-residue is conserved among homeodomains the ability of the K3 to make base contacts is unusual. CSI-Tree analysis reveals that the best binding sequences display TT dinucleotide at the 5′ edge of the core motif. This dinucleotide would permit minor groove hydrogen bonding with two amino acid residues (K3/R5) in the flexible N-terminal arm of the protein. More importantly, in the absence of a TT dinucleotide, the CSI tree reveals equal preference for a T residue at the first or the second position. These two motifs TNAAGTG or NTAAGTG would have been collapsed in a single motif by standard motif searching methods. The averaged motif would have obscured the importance of the specific protein-DNA contacts in the minor groove. The distinct motifs identified by CSI-tree analysis likely arise due to the ability of the protein to make hydrogen bonds with position 1 or 2 by K3 or R5. The collection of three motifs more accurately defines the sequence preferences of the protein.

CSI-Tree analysis of the polyamides PA1 and PA2

PA1. Our application with the CSI data of polyamide engineered to target the sequence 5′-WWGWWCWW-3′ resulted in the tree displayed in Figure 4. This tree can be summarized with the following algebraic expression:

$$\hat{Y} = 4.4I(X2 \in \{\mathrm{T}\}) \quad [1]$$
$$+ 4.1I(X2 \in \{\mathrm{A}\})\mathrm{I}(X1 \in \{\mathrm{T}\}) \quad [2]$$
$$+ 5.7I(X2 \in \{\mathrm{A}\})I(X1 \in \{\mathrm{A}\}) \quad [3]$$

An interplay between the first two positions of the binding site is evident from this tree. The dinucleotide AA in the first two positions leads to strongest binding affinities. Decreased binding affinities are observed if either the first or the second position is replaced by a T residue. After the CSI-Tree alignment, positions 3, 4 and 6 consist of unique nucleotides (G3, T4 and C5), but positions 5 and 7, similar to position 1 and 2, allow for different nucleotides (A or T). We further complemented our tree analysis by fitting a linear ANOVA model using positions 1, 2, 5 and 7. The resulting linear ANOVA fit explains only 18% (multiple R^2) of the variability observed in the binding affinities by the positional information. However, the contribution of positions 1 and 2 and their interactions are highly significant as seen in Table 1. Positions 5 and 7 do not contribute significantly to explaining the variability observed in the binding affinities. This is also readily inferable from the resulting CSI tree as there are no splits that depend on either of these two positions.

PA2. The second polyamide was engineered to bind the 5′-WWGGWWW-3′ motif. The result of the CSI-Tree analysis is displayed in Figure 5A and the algebraic expression representing the tree is given as follows:

$$\hat{Y} = 4.4I(X5 \in \{\mathrm{A}\}) \quad [1]$$
$$+ 4.3I(X5 \in \{\mathrm{T}\})I(X1 \in\in \{\mathrm{C}, \mathrm{G}, \mathrm{T}\})I(X7 \in \{\mathrm{T}\}) \quad [2]$$
$$+ 3.9I(X5 \in \{\mathrm{T}\})I(X1 \in \{\mathrm{C}, \mathrm{G}\})I(X7 \in \{\mathrm{A}\}) \quad [3]$$
$$+ 5I(X5 \in \{\mathrm{T}\})I(X1 \in \{\mathrm{T}\})I(X7 \in \{\mathrm{A}\})I(X6 \in \{\mathrm{A}\}) \quad [4]$$
$$+ 6.4I(X5 \in \{\mathrm{T}\})I(X1 \in \{\mathrm{T}\})I(X7 \in \{\mathrm{A}\})I(X6 \in \{\mathrm{T}\}) \quad [5]$$
$$+ 4.4I(X5 \in \{\mathrm{T}\})I(X1 \in \{\mathrm{A}\})I(X3 \in \{C\}) \quad [6]$$
$$+ 4.9I(X5 \in \{\mathrm{T}\})I(X1 \in \{\mathrm{A}\})$$
$$I(X3 \in \{\mathrm{G}\})I(X7 \in \{\mathrm{T}\})I(X2 \in \{G, T\}) \quad [7]$$
$$+ 6.4I(X5 \in \{\mathrm{T}\})I(X1 \in \{\mathrm{A}\})I(X3 \in \{\mathrm{G}\})$$
$$I(X7 \in \{\mathrm{T}\})I(X2 \in \{\mathrm{A}\}) \quad [8]$$
$$+ 5.8I(X5 \in \{\mathrm{T}\})I(X1 \in \{\mathrm{A}\})I(X3 \in \{\mathrm{G}\})$$
$$I(X7 \in \{\mathrm{A}\})I(X2 \in \{G, T\})I(X6 \in \{\mathrm{A}\}) \quad [9]$$
$$+ 8I(X5 \in \{\mathrm{T}\})I(X1 \in \{\mathrm{A}\})I(X3 \in$$
$$\{\mathrm{G}\})I(X7 \in \{\mathrm{A}\})I(X2 \in \{\mathrm{G}, \mathrm{T}\})I(X6 \in \{\mathrm{T}\}) \quad [10]$$
$$+ 9.2I(X5 \in \{\mathrm{T}\})I(X1 \in \{\mathrm{A}\})$$
$$I(X3 \in \{\mathrm{G}\})I(X7 \in \{\mathrm{A}\})I(X2 \in \{\mathrm{A}\}) \quad [11].$$

Figure 5B displays boxplots of the binding intensities at each leaf node of the tree. As evident from this plot, leaf nodes labeled with different mean binding intensities exhibit statistically significant differences.

The CSI tree for PA2 is far more textured, revealing additional dependencies that were not apparent from the data for PA1. At the first level of discrimination between sequences, CSI-Tree readily identified the preference of PA2 for a T, instead of a W (A or T) at position 5. This is consistent with the preference of the imidazole-pyrrole rings for a GT dinucleotide rather than a GW (2,4,34). The preference for a GT dinucleotide was also manifest in the PA1 recognition motifs (GT in positions 3 and 4 in Figure 4). In the PA1 CSI tree, GT preference did not emerge as one of the split rules because the best alignment for the oligonucleotides already included a GT dinucleotide at the third and fourth positions. The more obvious commonality between binding of PA2 and PA1 to their cognate sites is that the *tail* of the polyamide prefers the

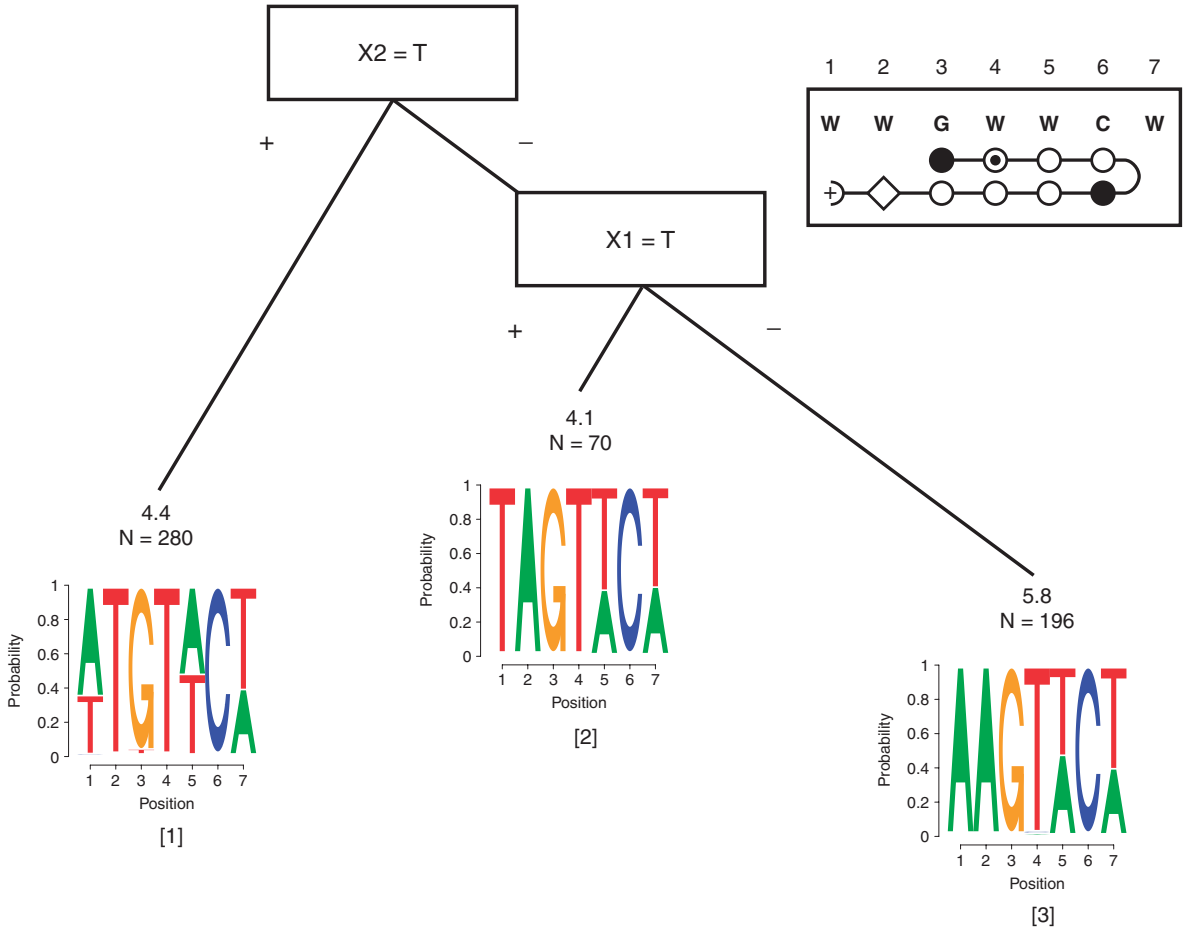

Figure 4. CSI tree for polyamide PA1. Upper right box shows predicted polyamide PA1-binding motif. Motif positions are indicated above the box. Here, W = {A, T}, closed circle ≡ imidazole (im) compound, open circle ≡ pyrrole (py) compound, circle with a center dot ≡ pyrrole compound attached to fluorescent dye, open diamond ≡ β-alanine and half circle with a plus ≡ dimethylaminopripylamide. The hairpin is connected with a γ butyric acid turn. (See Supplementary Figure 2A for chemical structure.)

AA dinucleotide in both cases (leaf node 11 in Figure 5A and leaf node 3 in Figure 4, see Figure 2B for nomenclature of polyamide elements). This preference may be explained by the propensity of the 5′-AA-3′ dinucleotide to further narrow the minor grove and thereby increase the van der Waal's interactions with the tail of the polyamide [see Figure 2B, (35)]. Closer inspection of the structures also suggests the possibility of hydrogen bonding between the polyamide tail and the T nucleotide at the second position (on the complementary strand, Figure 2). This could explain the enriched AA in the first two positions for these two polyamides. On the other end of PA2, the *turn* element (see Figure 2B and Supplementary Figure 2B) also appears to display sequence preference for an A over T residue at position 7 (Figure 5A, compares leaf node 11 with leaf node 8 and leaf node 5 with leaf node 2). This is not apparent in the PA1 where the turn element accepts a T or an A, consistent with previous studies (35). In this context, the preference of PA2 turn for an A at position 7 is more in keeping with the nonlocal structural effects of neighboring residues at positions 5 and 6. It is known that the structural properties of T-tracts differ from mixed sequence DNA. T-tracts tend to be more rigid with narrowing minor groove dimensions compared with mixed sequence DNA (36–40). Moreover, structural and computational modeling studies suggest that the sequences flanking a T-tract may be bent. Since PA2 requires a T residue at position 5, it discriminates against the consecutive occurrence of T residues at positions 6 and 7. Thus, nonlocal dependence between position 5 and position 7 is identified by CSI Tree and is explained by well-studied structural properties of the underlying sequence.

The role of sequence-dependent DNA micro-structure revealed by CSI-Tree also clarifies the basis of molecular

Table 1. Analysis of variance decomposition of the linear model for PA1 CSI data

Position	*F*-value	*P*-value
1	2.22e+01	5.62e−10
2	4.26e+01	1.58e−10
5	4.03e−01	5.26e−01
7	1.75e+00	1.86e−01
1.2	3.39e+01	1.02e−08

1:2 refers to the interaction between positions 1 and 2.

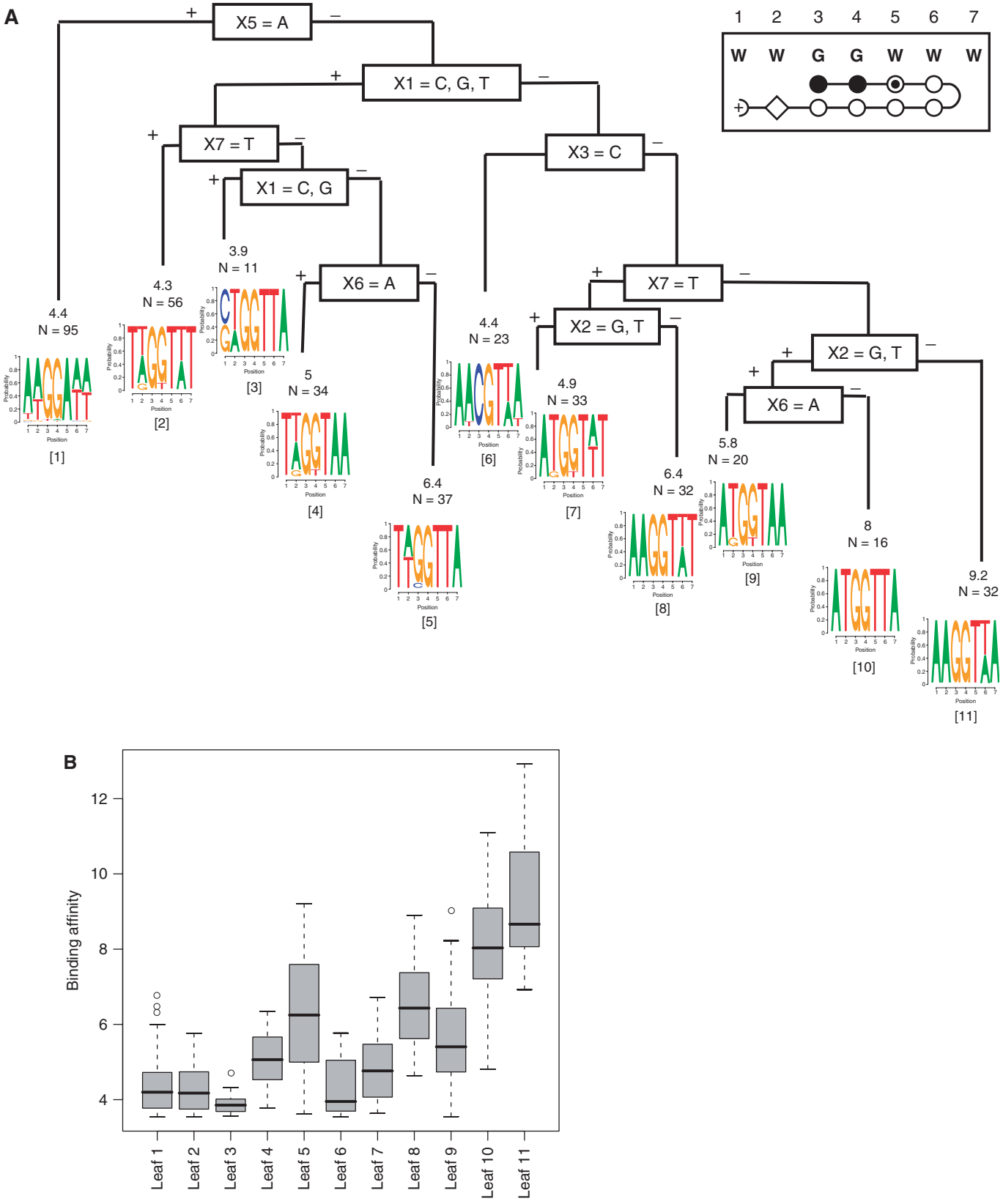

Figure 5. (**A**) CSI tree for polyamide PA2. Upper right box displays predicted polyamide PA2 binding motif. For chemical structure see supplementary Figure 1B. (**B**) Box-plots of the binding intensities at the leaf nodes of the PA2 CSI tree.

recognition by the engineered molecules. Deciphering the energetic penalties for altering flanking sequences has a significant impact on precisely tailoring the specificity of the engineered molecules for their desired target DNA sites.

Cross-validation experiments for comparisons with an ANOVA approach and MatrixREDUCE of Foat *et al.* (11)

To assess the predictive performance of CSI-Tree, we compared it with a simple ANOVA approach on the aligned sequences and the MatrixREDUCE method of Foat *et al.* (11). For the ANOVA approach, we included both a main effects only model, ANOVA(0), corresponding to simple additive model where each position of the binding site contributes independently and a higher order interaction model, ANOVA(+), where the order is determined by sequential *F*-tests. For MatrixREDUCE, we allowed one motif of length range between 5 bp and a maximum equal to that of the length of the oligonucleotides (L). Since we *a priori* know that the factors under the study do not bind to homodimers, we set `max_gap` and `flank` parameters of MatrixREDUCE to 0. MatrixREDUCE model fits a univariate regression of binding intensities on scores that are calculated from a position specific affinity matrix (PSAM). This matrix is estimated simultaneously with the slope and the intercept of the linear model during fitting. We used estimates of the PSAM, the intercept and the slope from the model fit to predict the binding intensities on the validation datasets in the cross-validation experiments.

The results of the 5-fold cross-validation experiments are reported in Table 2. As cross-validation criteria, we report both the averaged mean squared prediction error (MSE) and the Pearson correlation ($\hat{\rho}$) among the predicted and observed binding intensities over the validation sets. CSI-Tree has the best cross-validation performance with the minimum cross-validated MSE and the maximum cross-validated Pearson correlation. For PA2 and Nkx datasets, the ANOVA approaches outperform MatrixREDUCE. There are two potential reasons for this. First, ANOVA approach starts with a well-aligned set of sequences and does not have the added complexity of the alignment. Second, MatrixREDUCE tends to underestimate the motif width with one base in the Nkx-2.5 dataset and with two bases in the PA2 dataset. As an additional analysis, we repeated these cross-validation experiments using the natural log transformed intensities as suggested by Lee *et al.* (20). Supplementary Table 2 reports the results of this analysis and supports the general conclusions drawn here.

Table 2. Cross-validation experiments

Factor	Criteria	CSI-Tree	ANOVA(0)	ANOVA(+)	MatrixREDUCE
Nkx-2.5	MSE	0.311	0.403	0.383	0.583
	$\hat{\rho}$	0.709	0.509	0.549	0.368
PA1	MSE	1.662	–	–	3.405
	$\hat{\rho}$	0.735	–	–	0.451
PA2	MSE	1.240	1.544	1.322	2.129
	$\hat{\rho}$	0.740	0.740	0.764	0.685

ANOVA(0): ANOVA model with only main effect terms (each position of the binding site is contributing independently); ANOVA(+): ANOVA model where the model complexity, i.e. inclusion of higher order interactions, is based on a sequential *F*-test. MSE refers to averaged mean squared prediction error over the validation sets; $\hat{\rho}$ refers to averaged Pearson correlation between the observed binding intensities and the predicted intensities over the validation sets. For PA1, cross-validation criteria for the ANOVA methods are not reported as ANOVA method does not provide prediction if a position in a sequence has a level (nucleotide) that has not been encountered in the training dataset.

Analysis with the mixture of position weight matrices

Next, we set out to assess whether a mixture of PWMs model, one of the simple extensions of the PWM model, can be utilized to analyze data from CSI arrays. The simple trees of Nkx-2.5 and PA1 suggests that, the subsequences that are bound by these factors can be viewed as being generated by two sets of mixtures of three PWMs. The mixing proportions of the PWMs corresponding to leaves 1, 2 and 3 are estimated as (0.46, 0.26 and 0.28) and (0.51, 0.13 and 0.36), respectively for PA1 and Nkx-2.5, by the proportion of sequences at these leaf nodes. Note that since CSI experiments aim to generate a comprehensive set of sequences bound by the DNA-binding molecule, the number of bound sequences we have for both factors (351 sequences for Nkx-2.5 and 546 sequences for PA1), roughly represent all the sequences bound tightly *in vitro* by these factors. These sample sizes are ~4–6 times larger than that of the largest TF specific sequence dataset available in the TRANSFAC database (41). Therefore, they provide good test cases for assessing the performance of mixture of PWMs model on unaligned sequences. There are two difficulties in fitting mixture of PWMs. First is the tuning of K, the number PWMs to allow. Within the context of regular mixture models, this is typically addressed by Bayesian Information Criterion (BIC), Akaike Information Criteria (AIC) or Cross-Validation (CV). Practical performances of these selection criteria are highly variable. We experimented with all three and finally settled on cross-validation as the sample sizes for both datasets are quite large and the likelihood-based cross-validation has appealing optimality properties (42). The second challenge is the initialization of the Expectation-Maximization algorithm for fitting the mixture model. This is crucial as the algorithm can get stuck in local optima with poor starting choices. To bypass this, we initialized the PWMs using the ranked ordered PWMs obtained by running MEME (9) on unaligned sequences. In cases where MEME could not estimate enough number of matrices to initialize all of the components, we generated starting values based on frequently occurring Wmers.

For Nkx-2.5 CSI data, selecting the number of PWMs K with 5-fold cross-validation leads to six PWMs. However, two of the matrix classes have <10 sequence members according to the maximum posterior probability rule. Therefore, we truncate the number of selected PWMs to four. The sequence logos of these PWMs are displayed

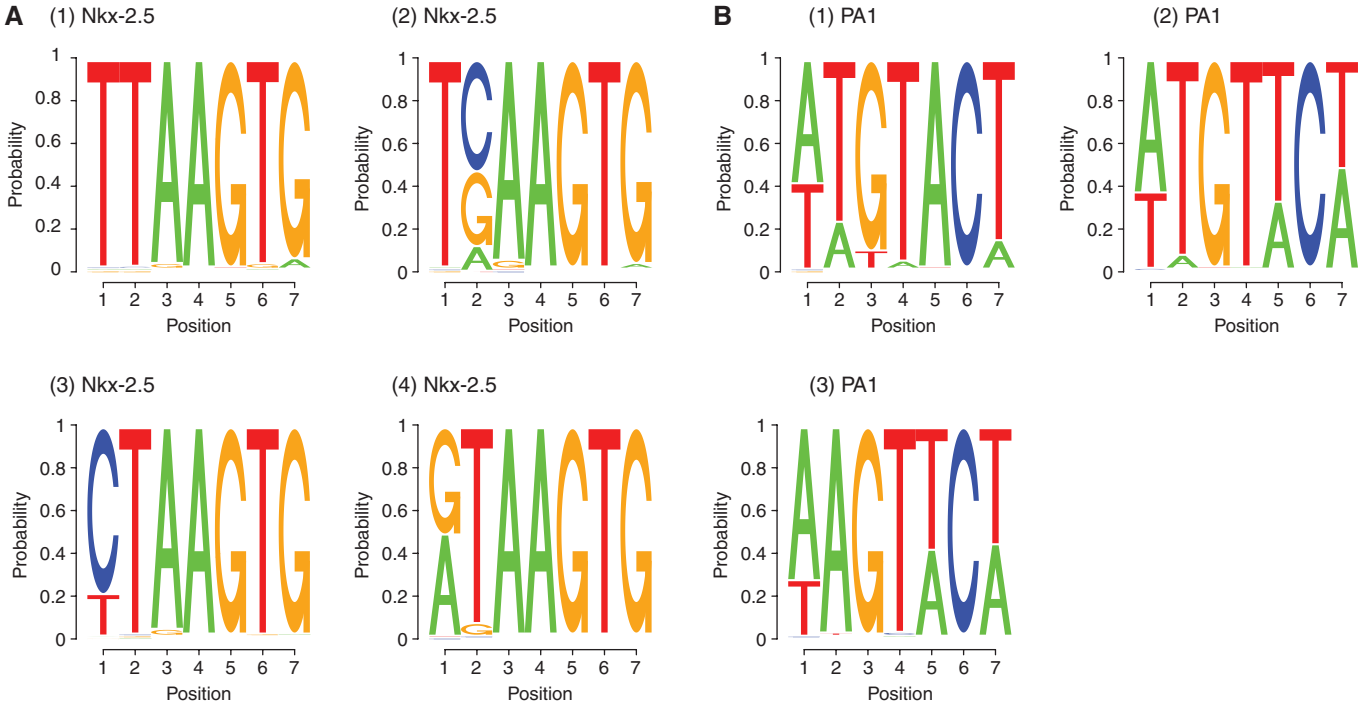

Figure 6. (**A**) Nkx-2.5 position weight matrix estimates for the four component mixture of PWMs model. (**B**) PA1 position weight matrix estimates for the three component mixture of PWMs model.

in Figure 6A. A closer examination of these logos reveal that components 1 and 2 correspond to leaf nodes 3 and 2 of the Nkx-2.5 CSI tree. Components 3 and 4 are, however, a mixture of leaf nodes 1 and 3, which is an indication that the dependence of the first two positions is not completely deconvolved. As these components indicate, the mixture model is able to decipher the partitioning of the sequence recognition space into regions captured by the CSI tree of Figure 3A to some limited extend.

For the PA1 CSI data, selecting the number of components K with 5-fold cross-validation leads to 3 components. The pictorials of these matrices are displayed in Figure 6B. Component 2 corresponds to leaf node 1 of the PA1 CSI tree, component 3 is a mixture of leaf nodes 2 and 3 and component 1 is actually a mixture of all the three leaf nodes. The interpretation that the first position is more likely to be an A if the second position is an A can still be inferred from these collection of three PWMs. This is evident from considering the last PWM (PA1-3) in which the second position is an A and the probability of having an A in the first position is higher compared with the other two PWMs. However, these three matrices are highly overlapping in terms of the types of sequences that they cover. In contrast, the CSI-Tree generates PWMs with nonoverlapping sequence space by utilizing the binding affinities, thereby leading to unambiguous and robust interpretations.

Next, we wondered how the number of components selected would change if we had much smaller sample sizes and/or longer sequences. Underestimating the number of components leads to observing a PWM which is a fusion of one more PWMs therefore does not lead to full characterization of the sequence recognition profiles. Barash *et al.* (14) analyzed experimentally verified (and aligned) binding site data for 95 TFs from the TRANSFAC database. The mean sample size of these datasets is 35 and the maximum sample size is 88. We generated 50 datasets of sizes 35 and 90 by randomly sampling from the original set of sequences identified from the CSI experiments while keeping the proportion of mixture components the same as that of observed in the CSI-Tree analysis. For one set of the runs, we used the sampled sequences as they are. In another set of runs, we extended them to 25mers by sampling from a multinomial distribution fitted to all positions combined over the dataset to investigate how the performance is affected when the binding sites are embedded within longer sequences. We then fitted a mixture of PWMs model to each data set with $K = \{1, \ldots, 6\}$ and reported the size selected by cross-validation. If a selected component had <5 members, then the number of such classes are deducted from K. This experiment was repeated for Nkx-2.5 and PA1 separately. Figure 7 displays the proportion of number of components selected under each scenario. We note that when the sample size is 35, the dominating number of components selected for the PA1 experiment is 2 and 3. For Nkx-2.5, $K = 2$ is selected for most of the time and this is followed by $K = 3$. When the width is increased at this sample size, only 1 component is typically selected for PA1 whereas 1 or 2 components are selected for Nkx-2.5. Although increasing the sample size leads to an increase in the frequency of selecting 3 components, there is a clear decline in the number of components

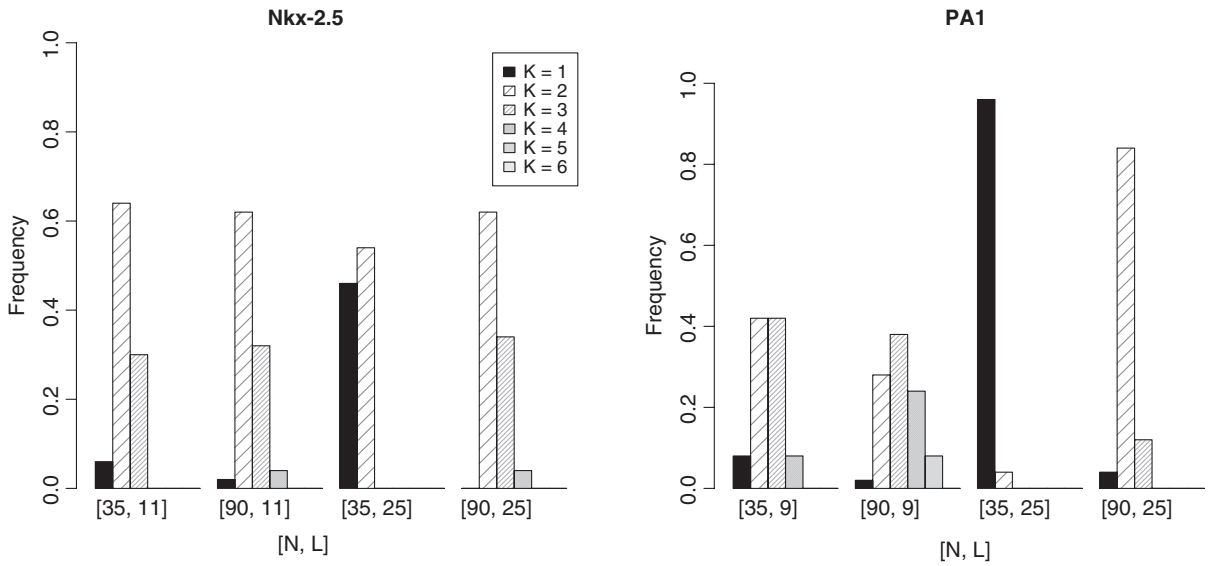

Figure 7. Simulation results. Each bar represents the proportion of times, out of 50 simulations, different number of components K (pattern coded) are selected. [*N*, *L*] refers to *N* sequences of width *L*.

selected as the length of the sequences increases at this larger sample size. These simulations illustrate that with a small number of sequences, which is the case for most of the datasets in TRANSFAC, it is difficult, if not impossible, to identify the true number of mixture components even if the proportion of the sequences from different mixture components comply with the true proportions on the observed data.

DISCUSSION

We have presented a regression tree analysis method named CSI-Tree for inferring recognition properties of DNA-binding molecules from CSI data. The advantage of CSI-Tree over currently available 'sequence only' methods is that it can capitalize on the quantitative binding information available from the CSI array experiments in a nonparametric fashion. As a result, the interactions between different positions can be interpreted as leading to increasing or decreasing binding affinities. Another key advantage of CSI-Tree is that it generates a series of binding intensity rank-ordered PWMs. Furthermore, the ability to deal with unaligned bound sequences from the CSI data makes CSI-Tree highly practical when the actual binding site is smaller than the length of the analyzed oligonucleotides.

Once we have aligned the sequences at the `CSI-Tree-EM` step, it is possible to consider a linear ANOVA model to explain the variability in the binding intensities as a function of sequence composition as an alternative to building a regression tree. One difficulty with this approach is that the alignment process by design produces unbalanced designs where each nucleotide combination exists in different proportions across the positions in the dataset. Unbalanced designs in general lack the computational simplifications, which exist in balanced designs. In balanced designs, the successive addition of higher order interaction terms leaves the preceding estimates unchanged and provides means for exploring higher order interactions without estimating the full model. However, this is not the case for unbalanced designs. This makes the application of linear ANOVA somewhat unattractive. Whenever possible (where we had enough degrees of freedom), we fitted linear ANOVA models after the CSI-Tree alignment and verified the apparent interactions of the resulting regression tree. Overall, our cross-validation experiments indicated that the regression tree approach has better predictive performance than those of simple linear modeling approaches.

We have explored the possibility of utilizing a mixture of PWMs for characterizing dependencies between the positions of the recognition profiles based on the bound sequences from the CSI data. Both using all the tightly bound sequences and various sized subsets of these indicated that fitting such a mixture model is quite difficult and often leads to underestimation of the number of mixture components. The performances we observe depend highly on the actual structure (length and conservation) of the PWMs. Both the Nkx-2.5 and the PA1 PWMs are quite conserved with information contents >90th percentile of the information contents of the TRANSFAC PWMs. We expect that with more degenerate matrices, it will be even more challenging for the mixture model to identify relevant components. However, as we illustrate here, direct use of binding intensities with a regression tree approach as in CSI-Tree helps to decipher positional dependencies and can lead to better designs of synthetic molecules.

METHODS

CSI array design

CSI microarray consists of duplex DNA sequences. The synthesized sequence is designed to form a DNA hairpin. Each hairpin probe is composed of a permuted hairpin

stem, which is one of the permutations of the Wmer with a 3 bp flanking sequence (CGC) on either side. For a generic Wmer, the hairpin has the sequence 5′-CGC-Wmer-CGC-TCCT-GCG-RWmer-GCG-3′, where RWmer represents the reverse complement of the Wmer. More details on the design are available in (2).

Preprocessing of the CSI data

The datasets are preprocessed following the procedure in (2). Within and across array normalizations were carried out by loess (43) and quantile normalizations (44) respectively. Average intensities of each spot across arrays were transformed into z-scores by using the mean and SE estimates of the unbound spots. SE estimates are obtained by taking mirror images of the histogram of all the normalized array intensities from the lower end of the mode of the histogram. Sequences with z-scores >95-th percentile of all the z-scores were then used in the CSI-Tree analysis.

Mixture of position weight matrices

A simple extension of the PWM model that aims to accommodate positional dependencies of the binding sites is a mixture of PWMs model (14,22). This model assumes that the sequence data is generated from a mixture of K PWMs where the columns of the PWMs are represented by independent but not identical multinomial distributions with four cell probabilities, one for each nucleotide. Let p_{aw}^{k} denote the probability of observing nucleotide $a \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ in position $w \in \{1, \ldots, W\}$ of the binding site under the PWM $k \in \{1, \ldots, K\}$. Then, the likelihood of observing a sequence of length W, $X_i = (X_{i1}, \ldots, X_{iW})$, is given by

$$Pr(\mathbf{X}_i) = \sum_{k=1}^{K} \pi_k \prod_{w=1}^{W} \prod_{a \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}} p_{aw}^{I(X_{iw}=a)},$$

where π_k is the mixing proportion of the k-th PWM and denotes the prior probability of a binding site being generated from the k-th component. Hannenhalli *et al.* (22) use this model specifically with $K = 2$ to analyze aligned sequence data and Barash *et al.* (14) allow extension of this model with $K = 2$ to unaligned sequences by allowing background nucleotides to be generated from a background component described by a multinomial distribution.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online. CSI-Tree R package and the accompanying data are available at http://www.stat.wisc.edu/~keles/CSI-NAR/CSI-NAR.html.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the NIH (1-R01-HG03747-01a) and PhRMA Foundation Research Starter Awards to S.K.; NIH (GM069420), March of Dimes, W. M. Keck Foundations, USDA/Hatch grant and Shaw Scholar and Vilas Associate awards to A.Z.A. C.L.W. was supported by an NIH/NLM pre-doctoral fellowship and C.D.C. was supported by the American Heart Association predoctoral fellowship. S.K. is supported in part by an NIH grant (1-R01-HG03747-01a) and a PhRMA Foundation Research Starter Grant. Funding to pay the Open Access publication charges for this article was provided by NIH Grant (1-R01-HG03747-01a) awarded to S.K.

Conflict of interest statement. None declared.

REFERENCES

1. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N., Macisaac,K.D., Danford,T.D., 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.
2. Warren,C., Kratochvil,N.C., Hauschild,K.E., Foister,S., Brezinski,M.L., Dervan,P.B., Phillips,G.N. and Ansari,A.Z. (2006) Defining the sequence-recognition profile of dna-binding molecules. *Proc. Natl Acad. Sci. USA*, **103**, 867–872.
3. Ansari,A.Z. and Mapp,A.K. (2002) Modular design of artificial transcription factors. *Curr. Opin. Chem. Biol.*, **6**, 765–772.
4. Dervan,P.B. and Edelson,B.S. (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. *Curr. Opin. Struct. Biol.*, **13**, 284–299.
5. Spellman,P., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identincation of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell.*, **9**, 3273–3297.
6. Wang,T. and Stormo,G. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
7. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
8. Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.
9. Bailey,T. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.*, **21**, 51–80.
10. 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.
11. Foat,B.C., Morozov,A.V. and Bussemaker,H.J. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, 141–149.
12. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
13. Bulyk,M.L., Johnson,P.L.F. and Church,G. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
14. Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in protein-DNA binding sites. *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*. Annual Conference on Research in Computational Molecular Biology, ACM Press, New York, NY, USA, pp. 28–37.
15. Zhao,X., Huang,H. and Speed,T.P. (2005) Finding short DNA motifs using permuted markov models. *J. Comput. Biol.*, **12**, 894–906.
16. Ben-Gal,I., Shani,A., Gohr,A., Grau,J., Arviv,S., Shmilovici,A., Posch,S. and Grosse,I. (2005) Identification of transcription factor binding sites with variable-order bayesian networks. *Bioinformatics*, **2**, 2657–2666.

17. Zhou,Q. and Liu,J. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.
18. Berger,M.F., Philippakis,A.A., Qureshi,A., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **11**, 1429–1435.
19. Bulyk,M.L., Huang,X.H., Choo,Y. and Church,G. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.
20. Lee,M.-L.T., Bulyk,M.L., Whitmore,G.A. and Church,G.M. (2002) A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays. *Biometrics*, **58**, 981–988.
21. Dempster,A.P., Laird,N. and Rubin,D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B Met.*, **39**, 1–38.
22. Hannenhalli,S. and Wang,L.-S. (2005) Enhanced position weight matrices using mixture models. *Bioinformatics*, **21**, i204–i212.
23. Puckett,J.W., Muzikar,K.A., Tietjen,J., Warren,C.L., Ansari,A.Z. and Dervan,P.B. (2007) Quantitative microarray profiling of DNA-binding molecules. *J. Am. Chem. Soc.*, **129**, 12310–12319.
24. Breiman,L., Friedman,J., Stone,C.J. and Olshen,R. (1993) *Classification and Regression Trees*. Chapman & Hall, Boca Raton, Florida, USA.
25. Ihaka,R. and Gentleman,R. (1996) R A language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–341.
26. Bembom,O., Keleş,S. and van der Laan,M.J. (2007) Supervised detection of conserved motifs in DNA sequences with cosmo. *Stat. Appl. Genet. Mol. Biol.*, **6**, Article 8.
27. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–4.
28. Chen,C. and Schwartz,R. (1995) Identification of novel DNA binding targets and regulatory domains of a murine tinman homeodomain factor, nkx-2.5. *J. Biol. Chem.*, **270**, 15628–15633.
29. Trauger,J.W., Baird,E.E. and Dervan,P.B. (1996) Recognition of DNA by designed ligands at subnanomolar concentrations. *Nature*, **382**, 559–561.
30. LeBlanc,M. and Tibshirani,R. (1998) Monotone shrinkage of trees. *J. Comput. Graph. Stat.*, **7**, 417–433.
31. Gruschus,J.M., Tsao,D.H.H., Wang,L.-H., Nirenberg,M. and Ferretti,J.A. (1997) Interactions of the vnd/NK-2 homeodomain with DNA by nuclear magnetic resonance spectroscopy: basis of binding specificity. *Biochemistry*, **36**, 5372–5380.
32. Weiler,S., Gruschus,J.M., Tsao,D.H.H., Yu,L., Wang,L.-H., Nirenberg,M. and Ferretti,J.A. (1998) Site-directed mutations in the vnd/nk-2 homeodomain: basis of variations in structure and sequence-specific DNA binding. *J. Biol. Chem.*, **273**, 10994–11000.
33. Gruschus,J.M., Tsao,D.H.H., Wang,L.-H., Nirenberg,M. and Ferretti,J.A. (1999) The three-dimensional structure of the vnd/NK-2 homeodomain-DNA complex by NMR spectroscopy. *J. Mol. Bio.*, **289**, 529–545.
34. White,S., Baird,E.E. and Dervan,P.B. (1996) Effects of the A-T/T-A degeneracy of pyrrole-imidazole polyamide recognition in the minor groove of DNA. *Biochemistry*, **35**, 12532–12537.
35. Swalley,S.E., Baird,E.E. and Dervan,P.B. (1999) Effects of γ-turn and β-tail amino acids on sequence-specific recognition of DNA by hairpin polyamides. *J. Am. Chem. Soc.*, **121**, 1113–1120.
36. Mack,D.R., Chiu,T.K. and Dickerson,R.E. (2001) Intrinsic bending and deforma-bility at the T-A step of CCTTTAAAGG: a comparative analysis of T-A and A-T steps within A-tracts. *J. Mol. Biol.*, **312**, 1037–1049.
37. Beveridge,D.L., Dixit,S.B., Barreiro,G. and Thayer,K. (2004) Molecular dynamics simulations of DNA curvature and flexibility: Helix phasing and premelting. *Biopolymers*, **73**, 380–403.
38. Rohs,R., Bloch,I., Sklenar,H. and Shakked,Z. (2005) Molecular flexibility in ab initio drug docking to DNA: binding-site and binding-mode transitions in all-atom Monte Carlo simulations. *Nucleic Acids Res.*, **33**, 7048–7057.
39. Rohs,R., Sklenar,H. and Shakked,Z. (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure*, **13**, 1499.
40. Siggers,T., Silkov,T. and Honig,B. (2005) Bending in the right direction. *Structure*, **13**, 1400.
41. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhäuser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.
42. van der Laan,M.J., Dudoit,S. and Keleş,S. (2004) Asymptotic optimality of likelihood-based cross-validation. *Stat. Appl. Genet. Mol. Biol.*, **3**, article 3..
43. Dudoit,S., Yang,Y.H., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.
44. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185–193.
45. Suto,R.K., Edayathumangalam,R.S., White,C.L., Melander,C., Gottesfeld,J.M., Dervan,P.B. and Luger,K. (2003) Crystal structures of nucleosome core particles in complex with minor groove DNA-binding ligands. *J. Mol. Biol.*, **326**, 371–380.