BIOINFORMATICS

Vol. 17 no. 7 2001
Pages 608–621

Identifying target sites for cooperatively binding factors

Debraj GuhaThakurta and Gary D. Stormo

Department of Genetics, Washington University School of Medicine, 4566 Scott Avenue, Campus Box: 8232, St Louis, MO 63110, USA

Received on December 22, 2000; revised on February 27, 2001; accepted on March 6, 2001

ABSTRACT

Motivation: Transcriptional activation in eukaryotic organisms normally requires combinatorial interactions of multiple transcription factors. Though several methods exist for identification of individual protein binding site patterns in DNA sequences, there are few methods for discovery of binding site patterns for cooperatively acting factors. Here we present an algorithm, Co-Bind (for COperative BINDing), for discovering DNA target sites for cooperatively acting transcription factors. The method utilizes a Gibbs sampling strategy to model the cooperativity between two transcription factors and defines position weight matrices for the binding sites. Sequences from both the training set and the entire genome are taken into account, in order to discriminate against commonly occurring patterns in the genome, and produce patterns which are significant only in the training set.

Results: We have tested Co-Bind on semi-synthetic and real data sets to show it can efficiently identify DNA target site patterns for cooperatively binding transcription factors. In cases where binding site patterns are weak and cannot be identified by other available methods, Co-Bind, by virtue of modeling the cooperativity between factors, can identify those sites efficiently. Though developed to model protein–DNA interactions, the scope of Co-Bind may be extended to combinatorial, sequence specific, interactions in other macromolecules.

Availability: The program is available upon request from the authors or may be downloaded from http://ural.wustl.edu.

Contact: dg@genetics.wustl.edu;
stormo@genetics.wustl.edu

1 INTRODUCTION

Understanding the complex transcriptional regulatory network is an interesting and challenging problem. Towards that goal the elucidation of the basic regulatory apparatus, which is organized in the form of arrays of transcription factor (TF) binding sites on DNA, is of primary importance. Due to the laborious and time consuming procedure of elucidating TF binding sites through experimental methods, computational methods for identifying TF binding sites is an active area of research. Several methods for local multiple sequence alignment have been used to address the problem of identification of individual TF binding site patterns, e.g. Consensus (Hertz and Stormo, 1999), MEME (Bailey and Elkan, 1994), Gibbs Sampler (Lawrence *et al.*, 1993), ANN-Spec (Workman and Stormo, 2000). In many cases where binding sites for TFs are known from experiments, these programs have been shown to yield the known binding site patterns, indicating that the results of these methods can be useful in discovering unknown TF binding sites from a collection of sequences believed to contain a common binding site pattern. With the advent of technologies like DNA microarray (DeRisi *et al.*, 1997; Lockhart *et al.*, 1996), SAGE (Velculescu *et al.*, 2000) and various hybridization methods which can measure the mRNA expression levels of different genes, such a collection of sequences can now be readily obtained. Genes which have similar expression profiles, or are expressed in specific contexts, may be assumed to have similar transcription mechanisms governing their expression. Hence, upstream promoter regions of these genes might contain the binding sites for the same transcription factors. Application of local multiple sequence alignment methods on a set of unaligned promoter regions from such genes can help to identify novel TF binding sites.

Detailed experimental work on several individual genes (Fickett, 1996; Weintraub *et al.*, 1990; Yuh *et al.*, 1998; Zhong and Vershon, 1997) elucidate that the transcription regulatory mechanism in eukaryotes, and especially in higher organisms, is inherently much more complex than moderation of gene expression levels by binding of one single TF to the gene promoter region. In many (if not most) cases, transcription factors do not work alone, regulation results from the *cis*-regulatory action of several factors. TF binding sites are often organized in functional groups called modules (Yuh *et al.*, 1998; for a review see Werner, 1999) where TFs bind the promoter regions and regulate transcription as synergistic (cooperative) or antagonistic pairs (Arnone and Davidson, 1997; Fickett,

1996; Yuh *et al.*, 1998) (for more examples see the COMPEL database, Kel *et al.*, 1995). In synergistic or cooperative binding, simultaneous interactions of two factors with closely situated target sites can result in a non-additively high level of a transcriptional activation, whereas in antagonistic binding two factors interfere with each other so that competition for overlapping sites leads to a mutually exclusive binding of TFs (definitions taken from Kel *et al.*, 1995; http://compel.bionet.nsc.ru/compel/compel.html). Cooperative binding of factors to DNA and formation of a ternary complex of protein–protein–DNA has been shown in many cases (Weintraub *et al.*, 1990; Moreno *et al.*, 1995; Fickett, 1996; Muhlethaler-Mottet *et al.*, 1998).

Computational methods have recently begun to address these issues in gene regulation. Since cooperatively acting TFs need to be placed in close proximity to each other, the joint occurrence of two known binding site motifs within some distance constraints has been used for identification of transcription factor binding modules (Klingenhoff *et al.*, 1999). Methods have shown that utilization of information about coordinate or close positioning of known transcription factor binding sites present in the *cis*-regulatory elements of genes expressed in a specific context lead to more accurate prediction of novel genes which are likely to involve similar regulatory mechanisms (Wasserman and Fickett, 1998; Wagner, 1999). Thus, given two (or more) *known* binding site motifs and the information about their coordinate positioning, several methods can efficiently predict new regulatory regions involving those motifs. However, few methods exist to address the problem of *discovering* target site motifs for cooperatively binding TFs. Very recently, one method, Bio-Prospector, has been described for identification of two-block motifs (Liu *et al.*, 2001) which could potentially be used for identification of closely located binding site motifs for two cooperatively acting TFs.

Here we present a novel computational method, Co-Bind, for discovering binding site motifs for cooperatively acting factors. Our method models cooperative DNA binding by TFs by maximizing the joint likelihood of occurrence of two binding site motifs, in the process describing the Position Weight Matrices (PWMs) for the two binding motifs. Where the affinity of a given factor for a target site is low, cooperative interaction with another factor placed at an appropriate distance on the *cis*-regulatory region can increase both complex stability and specificity for the protein–DNA interaction. When addressing this problem computationally, this issue can be translated as follows: where the probability of observing one binding site is too small the joint probability of observing two binding sites may be high; or, from an information theoretic point of view, when the information content of a binding site motif is too small the information content of both binding site motifs taken together may become high enough for detection. We have tested Co-Bind on semi-artificial and real data sets from yeast. It is shown that the method can not only identify DNA binding site patterns within certain distance restrictions, but by virtue of modeling the cooperativity is also able to identify weak patterns for two TFs which would not have been identified if searched for individually using existing programs. Co-Bind may be applied to other cases of combinatorial, sequence specific, macromolecular interactions. We show that Co-Bind can effectively identify weak sequence signals for translation initiation in the *Escherichia coli* genome.

The relationship between information content of binding sites, spacing between the sites and expectation of binding site detection is discussed from an information theoretic point of view.

2 ALGORITHM AND IMPLEMENTATION

2.1 Overview

Two sequence data sets are given viz. the positive or training set, which represents a set of *cis*-regulatory elements, and a background set representing the genome. The positive set may be a collection of sequences believed to contain binding sites for two cooperatively binding TFs. The problem is to identify the binding site patterns for two TFs which bind cooperatively in the positive set. An objective function, which is derived from the thermodynamics of protein–DNA binding, is optimized to obtain PWMs for the two different DNA binding site patterns. In the objective function, sequences from both the training set and the genome are taken into account to discriminate against commonly occurring patterns in the genome and obtain patterns with higher specificity for the training set.

2.2 Description of the PWMs

A position weight matrix has previously been found to be a good model for describing protein binding sites in DNA (Stormo, 2000). An l long DNA binding site pattern may be described by a $4 \times l$ weight matrix, with four weights (for four DNA nucleotides) per pattern position (Figure 1A). Let us assume each weight in the matrix is the binding energy contribution of each nucleotide at a particular pattern position. With the additional assumption that protein–DNA contacts at individual residue positions in the binding site are independent of each other (Berg and von Hippel, 1987), the total binding energy for a TF molecule to a particular site is given by:

$$H_{\text{site}} = \sum_{k=0}^{l-1} \sum_{b=0}^{3} \omega_{k,b} \cdot x_{k,b}$$

where, ω denotes the PWM weights, x denotes the inputs from the site (DNA bases at different positions) k ranges

over the l positions of the site, b ranges over all four DNA bases. A simple weight matrix may also be looked upon as a simple, single layer, neural network (perceptron) (Stormo *et al.*, 1982), with one layer of weights between the input (site sequence) and the output (binding energy of TF to that site). Here, two weight matrices are used to represent the binding sites for two TFs and their combinatorial binding energy is given by the sum of individual binding energies (Figure 1B).

2.3 Description of the objective function

Derivation of the objective function is based on thermodynamics of DNA–protein binding. We first describe the objective function for identification of one binding site pattern, which is then extended to the two pattern problem. The objective function the for one site problem has been described before (Workman and Stormo, 2000), but is described here again in brief to facilitate the derivation of the objective function for cooperatively acting TF binding sites.

Suppose we are given two sequence sets, viz. positive sequence set, $\mathcal{S}$, and background sequence set, $\mathcal{G}$. Let s_j, be a sub-sequence from any sequence S_i, in the training set with offset j. Instances of s_j may be present multiple times in both the positive and background sets; P_j is the likelihood of finding s_j in the genome. Let us assume a TF molecule is bound somewhere in the genome. The likelihood that *any* of the sub-sequences s_j is bound is related to the binding energy, H_j, of the factor to s_j and probability of observation of s_j. By following maximum entropy distribution and Boltzmann equation this likelihood is be given by:

$$F_j = \frac{P_j \cdot \mathrm{e}^{-H_j}}{\sum_j^n P_j \cdot \mathrm{e}^{-H_j}} = \frac{P_j \cdot \mathrm{e}^{-H_j}}{Y} \tag{1}$$

where, $Y = \sum_j^n P_j \mathrm{e}^{-H_j}$ is the partition function over the distribution of all sequences assuring $\sum_j^n F_j = 1$, n denotes the total number of all sub-sequences, like s_j, in the genome. The likelihood that *one particular instance* of s_j is bound is thus, $\left(\frac{F_j}{P_j}\right)$.

Suppose: (a) s_{j_i} is the binding site for the TF in sequence S_i, (b) there are total of p, sequences, like S_i in the training set, (c) for each of these sequences there is only one binding site starting at offset position j_i, (d) binding sites in all p sequences are occupied. The probability that a TF molecule is bound to its site, s_{j_i} in sequence, S_i, is given by:

$$b_i = \frac{F_{j_i}}{P_{j_i}} = \frac{\mathrm{e}^{-H_{j_i}}}{Y} \tag{2}$$

The probability that a TF molecule is bound to its site in *every* sequence of the positive set is given by the product of individual probabilities:

$$B = \prod_{i=1}^{p} b_i = \prod_{i=1}^{p} \left(\frac{F_{j_i}}{P_{j_i}}\right) = \prod_{i=1}^{p} \left(\frac{\mathrm{e}^{-H_{j_i}}}{Y}\right) \tag{3}$$

Our objective is to maximize this probability. We can instead maximize the logarithm:

$$\ln B = \ln \prod_{i=1}^{p} \frac{\mathrm{e}^{-H_{j_i}}}{Y} = \sum_{i=1}^{p} (-H_{j_i}) - p \cdot \ln Y \tag{4}$$

To have an expression independent of the number of sequences in the positive set, the objective function is defined as follows:

$$U = \frac{1}{p} \sum_{i=1}^{p} (-H_{j_i}) - \ln Y \tag{5}$$

When considering binding sites for cooperatively interacting factors, we assume the two factors are *simultaneously* bound to different sites or sub-sequences in each of the positive set sequences. The partition function in this case may be given by:

$$\mathcal{Y}^{\mathrm{c}} = \sum_i \sum_{j,j'} (P_j \cdot \mathrm{e}^{-H_j} * P_{j'} \cdot \mathrm{e}^{-H'_{j'}}) \tag{6}$$

where, superscript 'c' is for **c**ooperative binding, j and j' are offsets for two different, non-overlapping, sub-sequences in a sequence, P_j and $P_{j'}$ are the probabilities of observing these sub-sequences in the genome, H_j and $H'_{j'}$ are binding energies of the two TFs to these sub-sequences; i sums over all sequences in the genome, j and j' sums over all possible pairs of non-overlapping sub-sequences in each of these sequences. Analogous to equation (2), the probability that both factors simultaneously occupy their respective binding sites, s_{j_i} and $s_{j'_i}$ in a sequence, S_i, of the training set, may be given by:

$$b_i^{\mathrm{c}} = \frac{\mathrm{e}^{-H_{j_i}} \cdot \mathrm{e}^{-H'_{j'_i}}}{\mathcal{Y}^{\mathrm{c}}} = \frac{\mathrm{e}^{-(H_{j_i} + H'_{j'_i})}}{\mathcal{Y}^{\mathrm{c}}} = \frac{\mathrm{e}^{-\mathcal{H}_i}}{\mathcal{Y}^{\mathrm{c}}} \tag{7}$$

The probability that both sites are simultaneously occupied in *all* sequences of the training set, is thus $B^{\mathrm{c}} = \prod_i^p (b_i^{\mathrm{c}})$. Our objective is to maximize this probability, hence, analogous to the one site problem, the objective function to be maximized is given by:

$$\mathcal{U}^{\mathrm{c}} = \frac{1}{p} \sum_{i=1}^{p} (-\mathcal{H}_i) - \ln \mathcal{Y}^{\mathrm{c}} \tag{8}$$

2.4 Training the PWMs

The PWM or perceptron weights are fit to the training data by a gradient descent approach. Prior to the start of training we initiate two PWMs with an arbitrary set of weights. The following steps are then followed to train the perceptron weights:

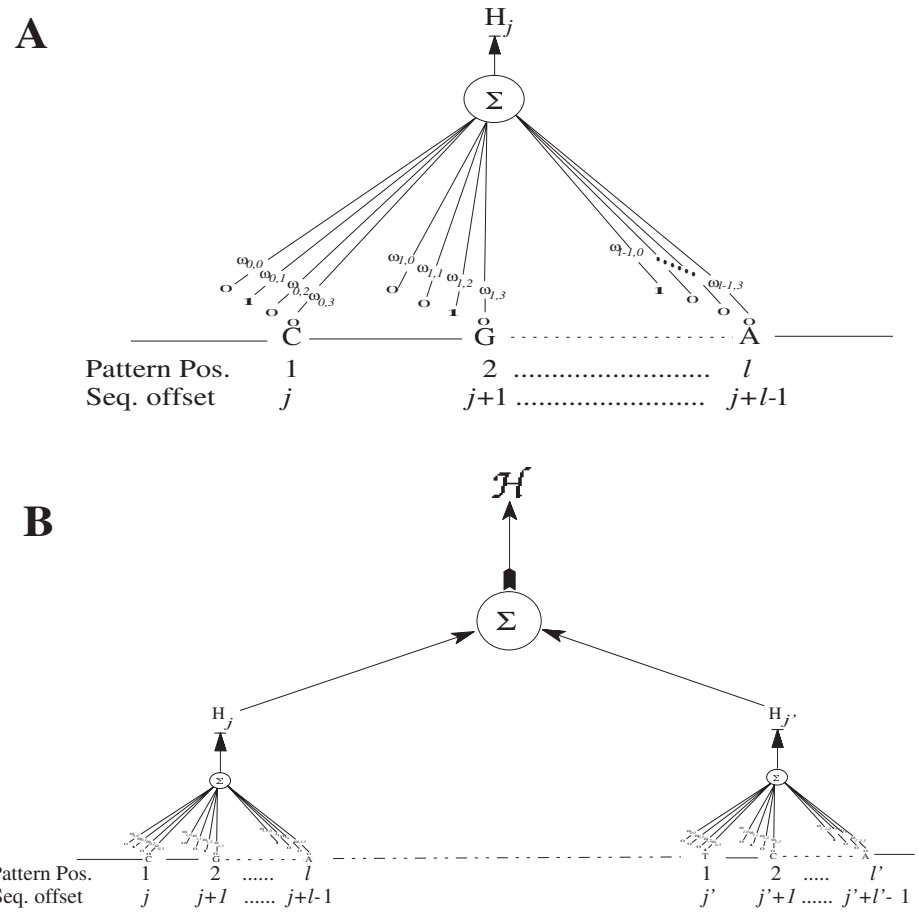

Fig. 1. (A) Schematic representation of a single DNA weight matrix or preceptron. $\omega_{k,b}$ denotes the weight for pattern position k, and base b. H_j is the output score from the PWM or perceptron for a given sub-sequence starting at position j in a particular sequence. (B) Schematic representation of two perceptrons for binding sites for cooperatively binding factors. Two perceptrons are allowed to sample non-overlapping sites from two different offset positions starting at j and j' on a particular sequence. $\mathcal{H}$, the cooperative output, is given by sum of individual outputs from the two perceptrons.

(1) Calculate individual objective functions for the two perceptrons, U_1 and U_2, following equation (5), and the two site cooperative objective function, $\mathcal{U}^{\mathrm{c}}$, using equation (8). (While estimating the partition functions Y or $\mathcal{Y}^{\mathrm{c}}$, rather than calculating over all background sequences, one can sample a large number of sites at random to approximate the background.) Set current $\mathcal{U}^{\mathrm{c}}$ to $\mathcal{U}^{\mathrm{c}}_{\mathrm{max}}$ and current perceptron weights to best weights, i.e. $\omega_1 = \omega_{1\mathrm{best}}$ and $\omega_2 = \omega_{2\mathrm{best}}$. The best weights represent best possible descriptions for the two binding sites.

(2) Select two non-overlapping sites from each of the training set sequences by Gibbs sampling (Lawrence *et al.*, 1993) using the two representative weight matrices.

(3) Update the perceptron weights ω_1 and ω_2 by gradient descent following the equations:

$$\Delta\omega_1 = \eta_1 \cdot \left(\frac{\delta U_1}{\delta \omega_1}\right) - \lambda_1 \cdot \omega_1$$
$$\Delta\omega_2 = \eta_2 \cdot \left(\frac{\delta U_2}{\delta \omega_2}\right) - \lambda_2 \cdot \omega_2 \qquad (9)$$

where, η_1 and η_2 are the step sizes, λ_1 and λ_2 are decay parameters.

(4) Recalculate individual objective functions, U_1 and U_2 and the two site cooperative objective function, $\mathcal{U}^{\mathrm{c}}$. If current $\mathcal{U}^{\mathrm{c}}$ is greater than $\mathcal{U}^{\mathrm{c}}_{\mathrm{max}}$ set: $\mathcal{U}^{\mathrm{c}}_{\mathrm{max}} = \mathcal{U}^{\mathrm{c}}$; $\omega_{1\mathrm{best}} = \omega_1$; and $\omega_{2\mathrm{best}} = \omega_2$;

(5) Iterate through steps 2–4 for a fixed number of times. Report $\mathcal{U}^{\mathrm{c}}_{\mathrm{max}}$ and best weights, $\omega_{1\mathrm{best}}$, $\omega_{2\mathrm{best}}$.

2.5 Implementation

Co-Bind was implemented in C^{++} and was developed on Sun workstations running Solaris. The code is portable and can be compiled and run under the Unix environment. A typical program run on 30 training set sequences, each 500 nucleotides long, (background sequence set of ~6000, 500 nucleotide long sequences) takes ≃3 min on a Sun workstation with a 296 MHz processor; the memory requirement never exceeds 25 MB. The computation time is $O(N)$, where N is the number of sequences in the training set, because we do a fixed number of iterations rather than wait for convergence.

3 DATA

3.1 Semi-artificial sequences

Semi-artificial sequence data was generated for testing the program. Thirty ORFs were chosen from the yeast genome at random. Upstream regions of these genes (from −500 to −1 relative to the translation start site) were taken. Various sequence patterns were implanted in these sequences to yield different training sets. The background set consisted of identical regions (−500 to −1) from all ORFs in the yeast genome. All yeast ORF upstream sequences were obtained from the yeast promoter database, SCPD (Zhu and Zhang, 1999) (http://cgsigma.cshl.org/jian/). Several different training data sets were prepared:

(1) Two sequences of length ten were arbitrarily chosen, AATCGCGTTA and GGATATATCC. For each, an alignment of thirty sites was initially generated, each element of the alignment being an exact copy of the consensus. The alignment was then mutated a number of times depending on the specified mutation rate. The number of mutations introduced in the alignment is given by $(r \cdot p \cdot l)$, where r is the mutation rate, p is the number of sequences in the alignment and l is the length of the sites. Substitution probabilities are defined by the nucleotide priors for yeast (0.315 for As and Ts, 0.185 for Gs and Cs) and reflexive substitutions are disallowed (i.e. once a particular nucleotide of an element is mutated, its not mutated again). Mutation rates for AATCGCGTTA alignment varied from 0.18 to 0.33 and that for GGATATATCC ranged from 0.25 to 0.37 to give alignments with varying degrees of conservation. The information content of each mutated alignment is given in Table 1A; all information content values are given in nats ($\log_e$) instead of bits ($\log_2$). Each instance of the mutated alignment for both patterns (henceforth referred to as a binding site) are then implanted into the thirty positive set sequences in random orientations and random positions, but within a certain distance of each other. Since in cooperative binding of TFs binding sites for the factors are often located close to each other, the artificial binding sites were implanted within certain distances of each other. Several training sets were created with the elements from two different alignments implanted within 25, 50, 100, 150 or 200 nucleotides. Care was taken to see the implanted patterns did not overlap with each other.

(2) Training sets were created where elements from only one particular alignment, corresponding to either AATCGCGTTA or GGATATATCC, were implanted in the same 30 yeast ORFs.

Data set (1) was generated to test Co-Bind and Bio-Prospector (Liu *et al.*, 2001) programs. Data set (2) was generated to check how well the individual binding site patterns can be identified by other available programs.

3.2 Yeast genes regulated by two TFs

In order to test if Co-Bind could identify biologically relevant TF binding site patterns which would be missed using other pattern finding programs we obtained four sets of yeast genes which have been experimentally shown to be regulated by two factors:

(1) A set of eleven genes are regulated by the Cbfl–Met4p–Met28p complex and Met31p or Met32p in response to methionine (van Helden *et al.*, 1998). The individual binding site patterns for Cbfl–Met4p–Met28p complex and Met31p/Met32p can be given by the consensus TCACGTG and AAACTGTGG, respectively (van Helden *et al.*, 1998). Upstream regions (−700 to −1) were extracted from SCPD for these genes, as the relevant binding sites in all eleven genes were located in that sequence region. The background sequence set contained −700 to −1 regions from all yeast ORFs.

(2) Matα2 protein is involved in a regulatory system that specifies cell mating type in yeast *Saccharomyces cerevisiae* (Zhong and Vershon, 1997). In haploid α cells and diploid cells, Matα2 interacts with a general transcription regulatory factor, Mcm1, to repress expression of a-specific genes. A group of six genes were collected from the SCPD (Zhu and Zhang, 1999) that were regulated by binding of both Matα2 and Mcm1. The binding site patterns for Matα2 and Mcm1 may be represented by the consensuses sequences AATGA(A/C)(A/T)T and CCTAAT(A/T)GGG respectively. Upstream regions (−350 to −1) were taken from SCPD for these genes, as all relevant binding sites were contained in that region. The background sequence set consisted of −350 to −1 regions from all yeast ORFs.

(3) SCPD contains five genes which are regulated by transcription factors GCR1 and RAP1. GCR1 and RAP1 have been shown to act in concert to mediate high-level glycolytic gene expression in *S.cerevisiae* (Baker, 1991). Binding sites for GCR1 and RAP1 can be represented by consensus sequences ACCCAGACA(A/T) and GGGCTTCC respectively. Upstream sequences (−700 to −300) were taken from SCPD for these five genes, as all relevant binding sites were contained in that region. The background sequence set consisted of −700 to −300 regions from all yeast ORFs.

(4) In *S.cerevisiae* more than 25 characterized genes are expressed only during sporulation and are referred to as meiotic genes or sporulation-specific genes. These genes are in the early, middle, and late expression classes. Most early genes have a 5′ regulatory site, URS1, and one of two additional sequences, UASH or a T4C site (Mitchell, 1994). URS1 site is required both to repress meiotic genes during vegetative growth and to activate these genes during meiosis. UASH and the T4C site also contribute to meiotic expression. In some cases cooperation between URS1 and UASH sites has been shown to be required for full induction of expression (Prinz *et al.*, 1995). SCPD contains 11 genes regulated by both URS1 and UASH sites. The URS1 binding site contains a motif with a highly conserved GC-rich core GCCGCC; the UASH site is not as well conserved. In 10 out of the 11 genes URS1 and UASH binding sites are located within the upstream region −300 to −1; we obtained the −300 to −1 sequences for those 10 genes from SCPD. Background sequence set consisted of the −300 to −1 regions from all yeast ORFs.

3.3 *E.coli* translation initiation sites

The entire *E.coli* K-12 genome sequence and gene coordinates were obtained (http://www.genetics.wisc.edu/) (Blattner *et al.*, 1997). All genes annotated as 'hypothetical' or 'putative' were ignored. From the remaining, 30 genes were chosen at random. The −25 to +25 regions (numbering relative to annotated translation start sites) of these 30 genes were taken for sequence analysis. The background sequence set consisted of 4000 randomly generated sequences, each 50 nucleotides long. Base priors used for generating the background set was same as that of *E.coli*, 0.25 for all four bases.

4 METHODS

4.1 Identification of individual binding site patterns

To check whether the individual binding site patterns can be identified by well known methods, several programs were run viz. Consensus (Hertz and Stormo, 1999), MEME (Bailey and Elkan, 1994), Gibbs Sampler (Lawrence *et al.*, 1993) and ANN-Spec (Workman and Stormo, 2000). All programs were run with the specifications: (1) length of the pattern to be identified; (2) expected frequency of binding sites (one site per sequence, unless mentioned otherwise); and (3) appropriate strands of the DNA (just positive, or both positive and reverse complement) to be included in the search.

Consensus. The Consensus program identifies a pattern with the highest information content in a given set of sequences. Version 6.c of Consensus was used and the top scoring result was reported.

MEME. The MEME algorithm uses an expectation maximization algorithm for finding patterns in input sequences. MEME Version 2.2 was run over the MEME web-server (see reference). The top scoring result was reported.

Gibbs sampler. Charles Lawrence's Gibbs Motif Sampler (Version 1.01.009) was used, with the option 'site sampler'. 100 different 'seeds' or starting points were used, a maximum of 2000 iterations were performed for each run, and the highest scoring result was reported.

ANN-Spec. Version 1.0 was used. Due to the non-deterministic nature of the algorithm, multiple training runs are performed (100), with each run iterating 2000 times. The results are sorted by their best attained objective function values, $U_{\max}$, (see equation (5)). Weight matrices corresponding to the ten highest scoring runs are observed. The binding site pattern is said to be identified correctly in one of these runs if the consensus from the weight matrix matches the consensus of the known patterns. Results for ANN-Spec are reported in terms of the number of times (out of ten) binding site patterns were identified correctly.

4.2 Identification of binding site patterns for cooperative factors

Co-Bind. Co-Bind was used to identify target sites for cooperative factors. In order to optimize the parameters for Co-Bind (viz. step-sizes and decay parameters; equation (9)), it was run on many artificially generated data sets in which two different binding site patterns were implanted. The training data sets consisted of 30 artificial sequences, each 500 nucleotides long, with two different binding sites, of several different lengths, implanted at

random positions. The background set contained 3000 artificially generated sequences, each 500 nucleotides long, but with no implanted sites. Several values of step-size η (from 1 to 10) and decay factor λ (from 0.0 to 0.5) were tested. The step-size determines the extent to which the perceptron weights will move in direction of the $(\frac{\delta U}{\delta \omega})$ gradient (equation (9)) and the decay parameter allows the perceptrons to retain a fraction of its previous weights (e.g. a decay parameter of 0.1 will allow retention of 90% of previous weights). A large step-size and decay parameter allows the perceptron to wander more in sequence space, while smaller rates result in more local searches with the final result being more sensitive to the initial weight settings (Workman and Stormo, 2000). A step-size of 3 and decay factor of 0.06 for training of both perceptrons seemed to work well with a wide range of artificially generated training sets (data not shown). Hence, all Co-Bind results reported are with these fixed parameters. We perform 100 training runs for Co-Bind with each run iterating 2000 times. Runs were sorted by the best attained cooperative objective function, $\mathcal{U}^{c}_{\max}$ (see equation (8)), and the perceptron weights corresponding to the ten highest scoring runs were observed. Like in ANN-Spec, binding sites are said to be identified correctly in one of these runs if the consensus from the perceptron weights match the consensus of known patterns. Results for Co-Bind are reported in terms of the number of runs (out of top scoring ten runs) which identified *both* binding site patterns correctly.

BioProspector. The BioProspector program (Liu *et al.*, 2001) uses a modified Gibbs sampling strategy to identify two binding site motifs within certain distance constraints. The maximum allowable distance between the two motifs is 50 nucleotides. BioProspector uses zero to third-order Markov background models whose parameters are estimated from a given sequence file. The minimum distance between the two motifs was set to 0 nucleotides and maximum distance between two motif blocks was set to 25 or 50 depending on the training data set (see results). 50 runs were be performed in each case and all statistically significant motifs were reported.

5 RESULTS

5.1 Semi-artificial sequences

Consensus, MEME, Gibbs Sampler and ANN-Spec were run on training data sets, where only one kind of binding site pattern was implanted from a mutated alignment as explained in the data section. All programs were run with: (1) length of binding sites to be identified set to 10; (2) expected frequency of sites, one per sequence; (3) both positive and reverse complement of DNA sequence included in the search. Several representative results are shown in Table 1A. Most programs were able to identify the AATCGCGTTA and GGATATATCC binding site patterns correctly only when the mutation rate for the alignments were 0.18 and 0.25 respectively.

Low-complexity patterns (e.g. poly-A or poly-T) occur frequently in promoter regions of yeast. Often they tend to appear as the best results of pattern finding programs which do not consider the background sequence set. We observe a similar situation with some of our tests. In several instances Consensus and Gibbs Sampler fail to identify the implanted binding site patterns, identifying poly-A or poly-T as best patterns instead. ANN-Spec, by considering both the training and background sets, is able to discriminate against these commonly occurring patterns and find those which are present only in the positive set (Workman and Stormo, 2000; C.T.Workman and G.D.Stormo, unpublished observations).

Co-Bind was run on training sets where two different binding site patterns were implanted simultaneously. Length of the patterns to be identified were set to 10, and both the positive and reverse complement strands of DNA were included in the search. As the objective function for Co-Bind (equation (8)) is closely related to that of ANN-Spec (equation (5)), it is useful to compare results of the two programs in order to see whether modeling cooperativity between binding sites can help in identification of patterns which, by themselves, are too weak to be identified. ANN-Spec results show how well binding sites may be identified individually, whereas Co-Bind results show how well a binding site pattern can be identified in the presence of the other (Table 1B).

In cases where the first and second binding site patterns are too weak to be identified by Consensus, MEME, Gibbs Sampler or ANN-Spec, Co-Bind can identify the implanted sites correctly by looking for two binding sites within a certain distance of each other. For example, binding site patterns for AATCGCGTTA with mutation rates 0.29 and 0.33; and that for GGATATATCC with mutation rates 0.33 and 0.37 are not identified well by any of the methods. Presumably, the information in individual patterns is not strong, but the combinatorial information on both patterns is high enough for their simultaneous identification using Co-Bind. When binding sites are located close to each other (e.g. within 25 or 50 nucleotides) binding sites are identified with higher sensitivity.

5.2 Yeast genes regulated by two factors

Out of four sets of genes which are regulated by binding of two factors (see data), in three cases the patterns for both sites can be obtained from the upstream regions of the genes regulated by them, using one or more of the programs—Consensus, MEME, ANN-Spec or Gibbs Sampler, in a sequential manner. When run on promoter regions of the genes, the programs first identified the

Table 1. (A) Identification of individual patterns. The consensus for implanted patterns are shown along with the mutation rate (r) and Information Content (IC) of the alignment in nats. Consensus from the weight matrices produced by the Consensus program, Gibbs Sampler, MEME are shown. A pattern is said to be identified correctly when consensus from identified pattern matches the consensus of implanted pattern with no more than two mismatches. All correctly identified patterns are marked with *. For ANN-Spec, the result is given by a number indicating times (out of top scoring 10 runs) the correct pattern is identified. The weight matrix consensus from all top scoring ANN-Spec runs are also shown for reference. (B) Results of Co-Bind runs and comparison with ANN-Spec results. ANN-Spec results are from (A). Co-Bind program was run on sequences with two different implanted patterns. Co-Bind results are given by the number of times *both* implanted patterns are identified correctly. Criterion for deciding correct results is same as in (A), i.e. not more than two mismatches between identified pattern consensus and implanted pattern consensus

Table 1A

Binding Site	AATCGCGTTA			GGATATATCC		
Mutation Rate, r	0.18	0.29	0.33	0.25	0.33	0.37
I.C. (nats)	7.61	6.23	5.76	6.57	5.55	5.22
Consensus	AATCGCGTTA*	TTTTTCTTTT	TTTTCTTTTT	GGATATATCC*	TTTCTCTTTT	TTTCTCTTTTC
MEME	CCACGCGTGG	GCGCATGCGC	GCACATGTGC	GGATATATCC*	GGGCATGCCC	CTGCCGGCAG
Gibbs Sampler	AATCGCGTTA*	TTTTTTCTTT	AAAGAGAAAA	AAAGAAAAAA	AAAGAAAAAA	AAAGAAAAAA
ANN-Spec	10	2	0	10	4	0
	AATCGCGTTA*	AATCGCGTTA*	AGCTAGCTTT	GGATATATCC*	GGATATATCC*	TTAAGCGGAG
	AATCGCGTTA*	ATCGCGTTAT*	CGGGATTGCC	GGATATATCC*	GGATATATCC*	AGCTAGCTTT
	AATCGCGTTA*	AGCTAGCTTT	CTCGCGGGGT	GGATATATCC*	GGATATATCC*	CTCGCGGGGT
	AATCGCGTTA*	CGGGATTGCC	TGGGGGTACT	GGATATATCC*	GGATATATCC*	CGGGATTGCC
	AATCGCGTTA*	CTCGCGGGGT	GAGTGTTTTT	GGATATATCC*	AGCTAGCTTT	GGGCTAGGAA
	AATCGCGTTA*	GGGCTAGGAA	CATCATCATC	GGATATATCC*	CTCGCGGGGT	GAGTGTTTTT
	AATCGCGTTA*	AGATCGTGAG	AGATCGTGAG	GGATATATCC*	GGGCTAGGAA	GAGGAACTTA
	AATCGCGTTA*	CCTGCAAATA	GATCATGCTC	GGGATATATC*	GAGGAACTTA	CCTGCAAATA
	AATCGCGTTA*	AATCCAGAGA	GGGCTAGGAA	GGATATATCC*	CGCCCCTTGC	AATCCAGAGA
	AATCGCGTTA*	AAATAACTTT	AAGGATTACC	GGATATATCC*	AATCCAGAGA	CATGATCACC

Table 1B

Implanted Patterns						Co-Bind results				
First Pattern			Second Pattern			Maximum possible separation between implanted sites				
Pattern Consensus	Mut. Rate	ANN-Spec	Pattern Consensus	Mut. Rate	ANN-Spec	25	50	100	150	200
AATCGCGTTA	0.18	10	GGATATATCC	0.25	10	10	10	10	10	10
AATCGCGTTA	0.29	2	GGATATATCC	0.25	10	10	10	10	10	10
AATCGCGTTA	0.33	0	GGATATATCC	0.25	10	10	10	10	10	10
AATCGCGTTA	0.29	2	GGATATATCC	0.33	4	10	10	10	10	10
AATCGCGTTA	0.33	0	GGATATATCC	0.33	4	10	10	9	8	10
AATCGCGTTA	0.33	0	GGATATATCC	0.37	0	10	8	4	4	1

stronger of the two patterns. The highest scoring sites corresponding to the weight matrix of the first pattern were then deleted from each sequence in the set. The programs were re-run a second time on the promoter sequences to identify the second binding site pattern. Binding patterns which can be identified in this manner include Cbfl–Met4p–Met28p complex and Met31p/Met32p (in data-set 1); Matα2 and Mcm1 (in data-set 2); GCR1 and RAP1 (in data-set 3). In all these data-sets Co-Bind is able to identify both patterns with high sensitivity (data not shown). In data-set 4 one of the patterns is not identified using any other method except Co-Bind. Detailed results for the fourth data-set are described below.

Table 2 summarizes the information about the genes regulated by both URS1 and UASH sites; the experimentally reported sites, the position of sites relative to the translation start, and the distance between the sites. All of the above information were obtained from SCPD. We have not considered the HOP1 gene in our analysis for the following reasons: (1) the URS1 site is placed much further upstream compared to all other genes; (2) the mutual distance between URS1 and UASH sites is 336, which

Table 2. DNA binding information for URS1 and UASH. Binding sites and positions for eleven gene upstream regions from yeast are shown. All positions are relative to the annotated translation start sites of respective genes. Distances between binding sites are given. In boxes on the right are given the mean (μ) and standard deviation (σ) of distances between URS1 and UASH binding sites. Statistics are given in separate boxes for group 1 (genes 1–5), and group 2 (genes 6–10), and group 1 and 2 taken together

Gene #	Gene ID	Name	URS1		UASH		Dist.
			Position	Mapped Site	Position	Mapped Site	
1	YDR285W	ZIP1	-22	TCGGCGGCTAAAT	-42	GATTCGGAAGTAAA	20
2	YER044C-A	MEI4	-98	TGGGCGGCTAAAT	-121	TCTTTCGGAGTCATA	23
3	YER179W	DMC1	-143	AAATAGCCGCCCA	-175	TTGTGTGGAGAGATA	32
4	YHR014W	SPO13	-100	AAATAGCCGCCGA	-119	TTTTCTGAATAAAC	19
5	YNL210W	MER1	-115	TTTTAGCCGCCGA	-152	GGTTTTGTAGTTCTA	37
6	YHR153C	SPO16	-90	TGGGCGGCTAAAA	-201	CATTGTGATGTATTT	111
7	YHR157W	REC104	-93	TTGGCGGCTATTT	-182	CAATTTGGAGTAGGC	89
8	YLR263W	RED1	-165	TCAGCGGCTAAAT	-355	ATTTCTGGAGATATC	188
9	YMR133W	REC114	-94	TGGGCGGCTAACT	-288	GATTTTGTAGGAATA	194
10	YOR351C	MEK1	-150	ATGGCGGCTAAAT	-233	TCATTTGTAGTTTAT	83
11	YIL072W	HOP1	-534	AATTAGCCGCCGA	-198	TGTGAAGT	336

Gr. 1 μ_{26} σ_{8}

Gr. 2 μ_{133} σ_{53}

μ_{80} σ_{67}

is substantially larger than that in any other genes in the set; (3) in all cases other than HOP1, URS1 site is downstream compared to UASH. These reasons make HOP1 an exception as far as positioning of the two sites are concerned. The average distance between URS1 and UASH sites in the remaining 10 genes is 80 nucleotides with a standard deviation of 67 nucleotides. Since Co-Bind performance decreases when the distance between two sites is large, we wished to see whether the program would be successful in detecting both URS1 and UASH sites from the remaining 10 genes where the distance between sites is smaller compared to the HOP-1 gene. If Co-Bind did identify those sites effectively, HOP1 would then be included in the training set. Based on average distance between binding sites, the 10 genes can be divided equally into two groups of five genes each. The average distance between URS1 and UASH sites in group 1 (genes 1–5, Table 2) is much smaller (26 nucleotides) than the average distance between sites in group 2 (genes 5–10) (133 nucleotides).

Alignment of mapped binding sites using Consensus indicate a length of 10 and 7 would be appropriate for URS1 and UASH sites respectively while searching for those patterns in upstream regions. We aligned the experimentally reported sites by the Consensus program in order to determine an appropriate length of the binding sites, because the factor binding sites are usually mapped using a multitude of techniques including DNase and hydroxyl radical footprinting methods which frequently overestimate binding site lengths. All URS1 sites were 12–13 long (Table 2), but appeared to have a highly conserved GCCGCC core. Most UASH sites were longer than 14 nucleotides, however, a site of length 8 is reported for HOP1 gene in SCPD. The alignment gave a consensus of T(A/G/T)GCCGCCTA (Information content in nats, IC = 8.1) for URS1 and TTTGGAG (IC = 4.2) for UASH when considering sites from all 10 genes; TAGCCGCC(G/T)A (IC = 6.8) and TT(C/T)GGAG (IC = 3.9) for the same sites when considering only the five group 1 genes. Thus, the UASH site pattern is significantly weaker compared to URS1.

Consensus, MEME, Gibbs Sampler and ANN-Spec were run on the promoter regions of (a) all 10 genes, or (b) group 1 genes, with appropriate pattern lengths (10 for URS1 and 7 for UASH); only the positive strand of DNA was included in the search. In either case, binding site pattern for URS1 was efficiently obtained. URS1 sites were deleted from the promoters and the programs run again to identify a second pattern. However, the UASH site pattern was not obtained. Data for group 1 genes is shown in Table 3.

We realized physical deletion of URS1 sites can introduce an artificial pattern in the sequences if regions around the deleted sites are conserved. To check whether this leads to the failure in programs identifying UASH sites, we observed different program runs from Consensus and MEME. Consensus allows the user to ignore specific portions of input sequences (here, URS1 sites) while searching for a pattern, and the MEME program allows the user to search for two (or more) non-overlapping patterns from a given data set. These results (not shown) indicate deletion of URS1 sites did not lead to artifactual failure in identification of the UASH pattern from promoter sequences. The UASH site pattern is too weak to be identified from the promoter regions.

Co-Bind was able to efficiently identify both URS1 and UASH patterns from group 1 genes. The Co-Bind program was initially run on the promoter sequences of all 10

Table 3. Pattern identified from the upstream region of 5 group 1 genes. For ANN-Spec, as in Table 1, the consensus from the weight matrices for top ten scoring runs are shown. Results which match the consensus pattern for known sites are marked with *. For Co-Bind, consensus patterns from the two different weight matrices are shown for top ten scoring runs. Eight times out of ten *both* URS1 and UASH patterns are identified correctly by Co-Bind

Programs \ Identified Patterns	Pattern 1	Pattern 2
Consensus	TAGCCGCC(G/T)A*	GCGCCAT
MEME	GCCGCCCAAG*	GCGCCAT
Gibbs Sampler	TAGCCGCC(G/T)A*	AGAAAAC
Ann-Spec	10	0
	TAGCCGCCGA*	AGCGCCA
	TAGCCGCCGA*	AGCGCCA
	TAGCCGCCTA*	GCGCCAG
	TAGCCGCCGA*	AGCGCCA
	GCCGCCGAAA*	AGCGCCA
	GCCGCCGAAA*	GCGCCAG
	GCCGCCGTAA*	AGTTGAG
	GCCGCCCTAA*	TAAACGG
	GCCGCCCAAA*	AGTTGAG
	GCCGCCGACA*	GCGCAAG
Co-Bind	GCCGCCGACA*	TTCGGAG*
	TGGCCGCCGA*	TTTGGAG*
8	TGGCCGCCGA*	TTCGGAG*
	GGCCGCCTAA*	GTTCGGA*
	GCCGCCCAAA*	GTTCGGA*
	AGCCGCCGAA*	TTGGAGT*
	AGTCGAGTAC	GCGCCAT
	AGCCGCCGAC*	TTCGGAA*
	AGCCGCCGAC*	TTGGAGT*
	ATAGCCGCCG*	CTCGGAA

genes. Again, only the positive DNA strand was included in the search. Length of binding patterns to be identified were set to 10 and 7 respectively. In each sequence, the second perceptron was allowed to sample for sites within a distance of 200 nucleotides upstream of the site sampled by the first perceptron, since all UASH sites were within that distance upstream of URS1 sites. While using all 10 genes, the first perceptron efficiently identified the URS1 pattern but the second perceptron did not identify the UASH pattern (data not shown). Co-Bind was consequently run only on group 1 gene upstream regions. In each sequence, the second perceptron was allowed to sample for sites within a distance of 50 nucleotides upstream of site sampled by the first perceptron, since in this set, all UASH sites were within 50 nucleotides upstream of URS1 sites. Both URS1 and UASH sites were identified in this case with high sensitivity. The results of program runs on group 1 gene upstream regions are summarized in Table 3.

5.3 Identification of translation initiation sites

The thermodynamic principles on which Co-Bind is based can be applied to other cases of cooperative, sequence-specific macromolecular interactions. Below we show, from a group of *E.coli* gene upstream regions, Co-Bind can identify the translation initiation sites by combining the sequence signals in the start codon and ribosome binding region.

From a purely sequence recognition point of view translation initiation is analogous to recognition of DNA binding sites by two cooperatively acting TFs. Here, the two binding components, the initiator tRNA and ribosome (or 16S ribosomal RNA), utilizes the sequence information in the mRNA start codon and the Shine–Dalgarno (SD) sequence (Shine and Dalgarno, 1974) to recognize the correct translation initiation sites (Gualerzi and Pon, 1990). The cooperativity between the two mRNA binding components can be imagined to be mediated by the ribosome itself and the Initiation Factors (IFs) (Gualerzi and Pon, 1990). Individually, the sequence information in either the start codon or the SD may be poor for the binding events to occur and formation of the pre-initiation complex. However, the combined information in the start codon and SD sequence along with an optimal distance range between the two sites are sufficient for recognition of the initiation sites, and effective binding of the initiator tRNA and ribosome for translation initiation (Barrick *et al.*, 1994).

Though the binding event involves the mRNA, the sequence information required for the mRNA binding is encoded in the DNA. Consensus, MEME, Gibbs Sampler and ANN-Spec were run on 50 long DNA sequences from 30 randomly chosen *E.coli* genes (−25 to +25, relative to annotated start sites). Positive strand of DNA was used as input. We used searching lengths of 3 (corresponding to the start codon) and 6, corresponding to the most conserved region of SD sequence, AGGAGG (Ringquist *et al.*, 1992). With searching length 3, most programs efficiently identify the ATG start codon. When pattern searching length was 6 all program results included the ATG start codon (Table 4A). The start codons were then deleted from all 30 sequences and programs re-run. None of the programs were able to find the SD region from the sequences. We verified that deletion of the more conserved ATG sites were not responsible for artifactual failure in identification of the SD pattern using several Consensus and MEME runs. So, the SD region appears to have insufficient information to be detected by the above programs.

Co-Bind was run on the same sequences. As before, only the positive strand of DNA was used, and lengths of patterns to be identified were set to 3 and 6 respectively. In each sequence, the second perceptron was allowed to sample sites within 15 nucleotides upstream of sites sampled by the first perceptron. This is because, in *E.coli* most SD regions are located within 15 nucleotides upstream of the start codon (Ringquist *et al.*, 1992). Out

Table 4. Recognition of translation initiation sites. (A) 'First pattern' indicates 3 or 6 long patterns identified by different programs from the 50 long *E.coli* sequences. 'Second pattern' indicates patterns identified from the data upon deletion of ATG start sites from the sequences. Patterns are consensus representations of weight matrices produced by the programs. For ANN-Spec, all top ten scoring runs were consistent and are represented by only one consensus sequence. (B) Co-Bind results from *E.coli* sequences. Top four of Co-Bind runs identified both the start site and the SD region correctly. Consensus from the two weight matrices from these four runs are shown

Table 4A

Identified Patterns / Programs	First Pattern		Second Pattern
	3-long	6-long	
Consensus	ATG	CATGAA	TTT(G/T)T(C/G)
MEME	ATG	ATGAAA	TTGTTG
Gibbs Sampler	ATG	ATGAAA	TT(G/T)TT(C/G)
Ann-Spec	AAA	ATGAAA	GAAAAA

Table 4B

First Pat.	Second Pat.
ATG	AGAGGA
ATG	AGGAGT
ATG	AGGAGT
ATG	AGGAGT

of the top scoring 10 runs, the four highest scoring runs yielded the correct patterns for both the start site and SD region (Table 4B), and for those runs, both sites were correctly identified in all individual sequences.

5.4 Comparison of Co-Bind and BioProspector results

Since BioProspector (Liu *et al.*, 2001) is the only other available method for identifying two closely placed patterns in sequences, we compared results of Co-Bind and BioProspector program runs on yeast semi-artificial data sets where two different artificial motifs were implanted in each sequence of the set. BioProspector allowed a maximum distance of 50 nucleotides between two binding site motifs, so it was run only on those data sets where the distance between two implanted sites were either 25 or 50 nucleotides. Both strands of DNA were included in the search and the expected frequency of either motif was set to one per sequence. Whereas Co-Bind identifies both implanted patterns correctly from these training sets (Table 1B), in most cases no statistically significant patterns were obtained by BioProspector. In no cases did BioProspector identify both patterns correctly. In one instance one binding site pattern was identified correctly (viz. AATCGCGTTA where its mutation rate was 0.18) (details not shown). In BioProspector, significance of each motif found is judged based on a motif score distribution estimated by a Monte Carlo method. Only those motifs with scores greater than five standard deviations above the motif score distribution mean are reported by the program. We decreased the threshold score for motif reporting from five to three standard deviations above the mean. This did not increase the number of correctly identified patterns, and again, in no case were both patterns identified correctly. Thus, when tested on identical training sets derived from yeast sequences, Co-Bind shows significantly improved performance over BioProspector in identification of closely positioned sequence motifs.

6 DISCUSSION

6.1 Information content and binding energy

Information content of binding sites may be directly related to the binding energy of TFs to those sites (for a review see Stormo and Fields, 1998), and thus to the objective functions we define in equations (5) and (8). It has been shown that in random genomes with no compositional inhomogeneities, where the probability of observing a site can be approximated by the genome base priors, the average binding energy of a TF to a collection of its binding sites is related to the information content, I_{sites}, of an alignment of those sites by the equation:

$$\langle \Delta G \rangle = -RT(I_{\text{sites}}) \tag{10}$$

$I_{\text{sites}} = \Sigma_{b,k} f(b,k) \ln \frac{f(b,k)}{p(b)}$, where, $p(b)$ is the genome composition for each base, and $f(b,k)$ is the fraction of each base present in each position, k, of the site (Stormo, 1998). In the case where binding of two TFs are considered, suppose the average binding energy of one TF is given by: $\langle \Delta G_1 \rangle$ and that of the other is given by $\langle \Delta G_2 \rangle$. The average combined binding energy may then be given by the sum of the binding energies, $\langle \Delta G_{1,2} \rangle = \langle \Delta G_1 \rangle + \langle \Delta G_2 \rangle = -RT(I_1 + I_2)$. This would be the case if there were no uncertainty in the positioning of binding sites with respect to each other. However, in a case where there is such uncertainty, the above combined binding energy is an overestimate, since the positional uncertainty has not been taken into account.

6.2 Positional uncertainty and spacing between sites

If the second site could be located anywhere within a window of J nucleotides of the first site, the loss in information due to this positional uncertainty can be given by (Schneider *et al.*, 1986):

$$I_{\text{pos}} = -\sum_{J} \frac{1}{J} \ln\left(\frac{1}{J}\right) = -\ln\left(\frac{1}{J}\right) = \ln(J) \tag{11}$$

and the average combined binding energy can be given by:

$$\begin{aligned} \langle \Delta G_{1,2} \rangle &= \langle \Delta G_1 \rangle + \langle \Delta G_2 \rangle - \langle \Delta G_{\text{pos}} \rangle \\ &= -RT(I_1 + I_2 - I_{\text{pos}}). \end{aligned} \tag{12}$$

The objective function defined for Co-Bind is directly related to the combined binding energy (equation (8)). Hence, the smaller the value of J, the greater the combined binding energy and the greater is the maximum value of objective function expected from the training data. This consequently means the perceptron weights will be better able to define the two binding site patterns in the training set if J is small. Equations (10)–(12) are true only for random genomes, however, they may explain why the sensitivity of correct answers from Co-Bind decreases when the binding sites are further apart (in semi-artificial data Table 1b, or in case of URS1 and UASH sites).

Coordinate positioning is an important aspect of combinatorial DNA binding by TFs (Fickett, 1996; Wasserman and Fickett, 1998; Wagner, 1999). Results of Co-Bind show that weak binding sites are more efficiently identified where binding sites are located close together. In some instances, positional constraints for DNA binding sites of cooperatively acting TFs have been observed to be much more stringent (Fickett, 1996). Thus, whereever possible, imposing such positional constraints on the two perceptrons may improve the sensitivity of binding site identification by Co-Bind.

6.3 Expectation of pattern identification

The amount of information needed to identify γ sites out of a possible Γ may be given by: $I_\gamma = -\ln\left(\frac{\gamma}{\Gamma}\right)$ (Schneider *et al.*, 1986). In semi-artificially generated test sets, binding sites are 10 long, occur once in every 500 long sequence and may be placed in either of the two DNA strands. In this case, only one out of possible $(491 * 2 =)$ 982 positions could be the starting position of a real site (thus, $\gamma = 1$ and $\Gamma = 982$). The average information required to find sites is then: $-\ln\left(\frac{1}{982}\right)$ or, 6.8 nats. For semi-artificial sequences, Table 1a shows, only those patterns which have more than 6.5 nats in information are identified efficiently by multiple programs.

Now we describe a relationship between information content of binding sites, spacing between them and expectation of pattern identification. For the same dataset as above, we ask what is the amount of information needed to identify the second site, having found the first. Let the maximum permissible distance between two sites be D nucleotides. The total window length, J, within which the second site can be located, is equal to $2D$, since ordering of the two sites with respect to each other is random (i.e. the second site can be placed in either direction, upstream or downstream, of the first site). The loss of information due to the uncertainty in the location of the second site with respect to the first is: $-\ln\left(\frac{1}{J}\right)$ or $-\ln\left(\frac{1}{2D}\right)$ (equation (11)). Also, the second site can be placed on either of the two DNA strands with respect to the first. The loss of information due to the uncertainty in the DNA strand, is $-\ln\left(\frac{1}{2}\right)$. Thus, the minimum amount of information required to detect two sites within a distance of D nucleotides of each other is given by:

$$I_{\text{req}} = -\ln\left(\frac{1}{982}\right) - \ln\left(\frac{1}{2D}\right) - \ln\left(\frac{1}{2}\right) \qquad (13)$$

The above arguments are exactly valid for random genomes, and is only an approximation for non-random genomes. However, at least qualitatively, they give an idea about how much information might be required to identify patterns in a given set of sequences, and how much spacing variability can be allowable before we start failing to identify the signals for binding sites.

Table 5 shows, for a pair of binding site motifs, the amount of information content required for binding site identification (I_{req}), the amount of information actually present in the binding site patterns (I_{actual}, which is the sum of information contents of individual patterns), and the efficiency of identification of those binding sites using Co-Bind. In theory, one would expect the efficiency of binding site identification to decrease with decreasing I_{diff} (defined as $I_{\text{actual}} - I_{\text{req}}$), and the efficiency to be high when I_{diff} is positive and poor when I_{diff} is negative. We see in Table 5, as I_{diff} decreases, the efficiency of binding site identification also decreases, and when I_{diff} is a large negative value (e.g. -2.5) efficiency of binding site identification by Co-Bind is poor. However, in a few instances Co-Bind is able to identify sites efficiently even though I_{diff} is a negative value (e.g. -0.4 and -1.1). The likely reasons for this are: (1) the genome is non-random; and (2) Co-Bind is not designed to obtain patterns with maximum information content. Co-Bind partitions against the whole genome in order to identify sites with high specificity for the training set. In the yeast genome there might be some frequently occurring sites (e.g. poly-A or poly-T) with high information content in all gene upstream regions. Though information rich, these sites are thus not specific for the training set and are not identified by Co-Bind. The implanted sites may have lower information content but could be more specific for the training set and hence are identified by Co-Bind.

Where individual patterns have less information, the combined information content of two weak patterns becomes high enough for identification. In semi-artificial data sets, the information content required to identify one 10 long binding site motif by itself (without the second site) is 6.8 nats. But in the example where two sites are present together and the maximum distance between the two sites is $D = 25$ nucleotides, individual binding sites can have 5.7 nats of information and still be identified successfully when searched for their joint occurrence ($D = 25$ and $J = 50$, giving $I_{\text{req}} = 11.4$ nats (equation (13))). For URS1-UASH regulated genes, the information content required to identify individual

Table 5. Information content of binding sites and expectation of pattern identification using Co-Bind. The training set used is where two patterns, AATCGCGTTA and GGATATATCC, with mutation rates 0.33 and 0.37 respectively, are implanted in 30, 500 nt long, yeast sequences (refer to Table 1B, last row). Information contents for the two patterns are 5.8 and 5.2 respectively. Maximum possible distance (D in discussion) between the two sites is given along with information content (IC) and efficiency (Eff) of pattern identification with Co-Bind. I_{req} is the information required to find both sites, as calculated from equation (13) in the discussion Section 6.3. I_{actual} is the sum total of information content contained in the two binding site motifs. $I_{diff} = I_{actual} - I_{req}$. Efficiency of Co-Bind in determining both sites correctly is taken directly from Table 1B (last row)

Max Dist / I.C./Eff	25	50	100	150	200
I_{req}	11.4	12.1	12.7	13.2	13.5
I_{actual}	11.0	11.0	11.0	11.0	11.0
I_{diff}	-0.4	-1.1	-1.7	-2.2	-2.5
Co-Bind Efficiency	10	8	4	4	1

site patterns from the 300 long sequences is roughly 5.7 nats. That may explain why UASH sites cannot be identified unless in conjunction with URS1 sites using Co-Bind (note, alignment of UASH sites gives information content around 4.0 nats). In the case of *E.coli* sequences, we observe from Co-Bind results which identify both patterns correctly, the information content of both ATG sites and the SD sequences are around 3.5 nats. The minimum information content required to identify sites from that training set is about 3.9 nats. ATG start codons are frequently identified by other methods but not the SD region (Table 4). Here again, by taking advantage of combinatorial information content, Co-Bind is able to identify both sites correctly. Thus Co-Bind models the situation where binding of individual factors to respective binding sites may be weak, but cooperativity between two factors lead to efficient binding by increasing complex stability.

6.4 Future improvements

Currently, specific pattern lengths need to be input in the program. Automatically determining the appropriate binding site lengths for the patterns should be very useful. As is common with gradient descent approaches, the objective function can get stuck in a local minimum. Hence, the program is run multiple times and top scoring runs are considered. Several methods for avoiding such a problem are known (Baldi and Brunak, 1998, and references therein), some of which could be implemented and examined.

7 CONCLUSION

Results of Co-Bind presented indicate that it is able to model the synergy between the binding factors. It can identify weak patterns, which cannot be identified by other available methods, by combining the sequence information in those patterns. Currently Co-Bind models cooperative binding by two factors, but extension to more than two factors is possible. Given the universal nature of transcriptional regulation by combinatorial binding of TFs, Co-Bind could prove to be useful in discovering new regulatory sites for synergistically acting TFs and understanding transcriptional regulatory mechanisms. The principles of macromolecular binding on which Co-Bind is based are general, hence, it may be extended beyond the scope of protein–DNA binding to other cases of macromolecular interactions, as evidenced from identification of translation initiation sites in *E.coli*.

ACKNOWLEDGEMENTS

Compaq Computers and Ray Hookway are gratefully acknowledged for making available CPU cycles for our computational purposes. Chris Workman is thanked for his useful input during development of the method. We thank Chip Lawrence for providing the 1.01.009 version of Gibbs Motif Sampler and instructions on its use. We thank Xiaole Liu and Jun Liu for providing the BioProspector program. One unknown reviewer is thanked for bringing to our attention several relevant publications. The work was supported by grant (HG00249) from National Institutes of Health to GDS.

REFERENCES

Arnone,M.I. and Davidson,E.H. (1997) The hardwiring of development: organization and function of genomic regulatory sequences. *Development*, **124**, 1857–1864.

Bailey,T.L. and Elkan,C.P. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Intelligent Syst. Mol. Biol.*, **2**, 28–36. (http://meme.sdsc.edu/meme/website/meme-adv.html/)

Baker,H.V. (1991) GCR1 of *Saccharomyces cerevisiae* encodes a DNA binding protein whose binding is abolished by mutations in the CTTCC Sequence motif. *Proc. Natl Acad. Sci. USA*, **88**, 9443–9447.

Baldi,P.F. and Brunak,S. (1998) *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, USA.

Barrick,D., Villanueba,K., Childs,J., Kalil,R., Schneider,T.D., Lawrence,C.E. and Stormo,G.D. (1994) Quantitative analysis of ribosome binding sites in *E.coli*. *Nucleic Acids Res.*, **22**, 1287–1295.

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**, 723–750.

Blattner,F.R. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Butscher,W.G., Powers,C., Olive,M., Vinson,C. and Gardner,K. (1998) Coordinate transactivation of the interleukin-2 CD28 response element by c-Rel and ATF-1/CREB2. *J. Biol. Chem.*, **273**, 552–560.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Fickett,J.W. (1996) Coordinate positioning of MEF2 and myogenin binding sites. *Gene*, **172**, GC19–GC32.

Gualerzi,C.O. and Pon,C.L. (1990) Initiation of mRNA translation in prokaryotes. *Biochem.*, **29**, 5881–5888.

van Helden,J., André,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Kel,O.V., Romaschenko,A.G., Kel,A.E., Wingender,E. and Kolchanov,N.A. (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res.*, **23**, 4097–4103.

Klingenhoff,A., Frech,K., Quandt,K. and Werner,T. (1999) Functional pomoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics*, **15**, 180–186.

Lawrence,C.E., Altschul,S.F., Bogusky,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Liu,X., Brutlag,D. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.

Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotech.*, **14**, 1675–1680.

Mitchell,A.P. (1994) Control of meiotic gene expression in *Saccharomyces cerevisiae*. *Microbiol. Rev.*, **58**, 56–70.

Moreno,C.S., Emery,P., West,J.E., Durand,B., Reith,W., Mach,B. and Boss,M. (1995) Purified X2 binding protein (X2BP) cooperatively binds the class II MHC X box region in the presence of purified RFX, the X box factor deficient in the Bare Lymphocyte Syndrom. *J. Immunol.*, **155**, 4313–4321.

Muhlethaler-Mottet,A., Berardino,W.D., Otten,L.A. and Mach,B. (1998) Activation of the MHC Class II Transactivator CIITA by interferon-gamma requires cooperative interaction between Stat1 and USF-1. *Immunity*, **8**, 157–166.

Prinz,S., Klein,F., Auer,H., Schweizer,D. and Primig,M. (1995) A DNA binding factor (UBF) interacts with a positive regulatory element in the promoters of genes expressed during meiosis and vegetative growth in yeast. *Nucleic Acids Res.*, **23**, 3449–3456.

Ringquist,S., Shinedling,S., Barrick,D., Green,L., Binkley,J., Stormo,G.D. and Gold,L. (1992) Translation initiation in *Escherichia coli*: sequences within the ribosome-binding site. *Mol. Microbiol.*, **6**, 1219–1229.

Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.

Shine,J. and Dalgarno,L. (1974) The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementary to nonsense triplets and ribosome binding sites. *Proc. Natl Acad. Sci. USA*, **71**, 1342–1346.

Stormo,G.D. (1998) Information content and free energy in DNA–protein interactions. *J. Theo. Biol.*, **195**, 135–137.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Stormo,G.D. and Fields,D.S. (1998) Specificity, free energy and information content in protein–DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.

Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982) Use of the 'Perceptron' algorithm to distinguish translation initiation sites in *E.coli*. *Nucleic Acids Res.*, **10**, 2997–3011.

Velculescu,V.E., Vogelstein,B. and Kinzler,K.W. (2000) Analyzing uncharted transcriptomes with SAGE. *Trends Genet.*, **16**, 423–425.

Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.

Weintraub,H., Davis,R., Lockshon,D. and Lassar,A. (1990) MyoD binds cooperatively to two sites in a target enhancer sequence: occupancy of two sites is required for activation. *Proc. Natl Acad. Sci. USA*, **87**, 5623–5627.

Werner,T. (1999) Models for prediction and recognition of eukaryotic promoters. *Mam. Genome*, **10**, 168–175.

Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*, **5**, 464–475.

Yuh,C.-H., Bolouri,H. and Davidson,E.H. (1998) Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**, 1896–1902.

Zhong,H. and Vershon,A.K. (1997) The yeast homeodomain protein Matα2 shows extended DNA binding specificity in complex with Mcm1. *J. Biol. Chem.*, **272**, 8402–8409.

Zhu,J. and Zhang,M.Q. (1999) A promoter database of yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.