

Bipartite pattern discovery by entropy minimization-based multiple local alignment

Chengpeng Bi^{1,2} and Peter K. Rogan^{1,2,3,*}

¹Laboratory of Human Molecular Genetics, Children's Mercy Hospital & Clinics, 2401 Gillham Road, Kansas City, MO 64108, USA, ²School of Computer Science and Engineering and ³School of Medicine, University of Missouri—Kansas City, Kansas City, MO 64110, USA

Received June 25, 2004; Revised August 11, 2004; Accepted August 26, 2004

ABSTRACT

Many multimeric transcription factors recognize DNA sequence patterns by cooperatively binding to bipartite elements composed of half sites separated by a flexible spacer. We developed a novel bipartite algorithm, *bipartite pattern discovery* (Bipad), which produces a mathematical model based on information maximization or Shannon's entropy minimization principle, for discovery of bipartite sequence patterns. Bipad is a C++ program that applies greedy methods to search the bipartite alignment space and examines the upstream or downstream regions of co-regulated genes, looking for *cis*-regulatory bipartite patterns. An input sequence file with zero or one site per locus is required, and the left and right motif widths and a range of possible gap lengths must be specified. Bipad can run in either single-block or bipartite pattern search modes, and it is capable of comprehensively searching all four orientations of half-site patterns. Simulation studies showed that the accuracy of this motif discovery algorithm depends on sample size and motif conservation level, but results were independent of background composition. Bipad performed equivalent with or better than other pattern search algorithms in correctly identifying *Escherichia coli* cyclic AMP receptor protein and *Bacillus subtilis* sigma factor binding site sequences based on experimentally defined benchmarks. Finally, a new bipartite information weight matrix for vitamin D₃ receptor/retinoid X receptor α (VDR/RXR α) binding sites was derived that comprehensively models the natural variability inherent in these sequence elements.

INTRODUCTION

Transcription factors (TFs) bind to specific DNA sequences; however, analyses of sequences recognized by the same factor often reveal considerable variability between the binding sites (1). This variability between different binding sites recognized by the same TF suggests that probabilistic models, e.g.

position weight matrices (PWMs), can be used to represent these motifs. These matrices can be used to predict previously unknown sites. Information weight matrices can be used to rank the binding affinities of different sites (1,2). The matrix elements represent contributions of the individual bases to the protein–DNA interaction. In many cases, TF binding proteins do not work alone, and regulation results from the *cis*-effects of multiple *trans*-acting factors (3). The affinity of the protein for any site depends on the sum of all the interactions between the protein and the cognate DNA segment. The interactions at individual nucleotide positions may or may not be independent of one another, however, independence between positions is often assumed, because these effects tend to be small for most TFs (1).

A variety of *in silico* approaches have been used to predict TF binding sites (TFBSs) (also called motif discovery), since experimental methods have been laborious, time consuming and prone to bias (1). It is still not practical to comprehensively define these motifs by laboratory tests on complete genomic sequences. Computational methods utilize a set of known binding sites to extract important residue patterns from DNA input sequences and provide a representation of the binding sites that can be used to locate new sites with reasonable reliability (1,4). Multiple sequence local alignments are a prerequisite for extracting these patterns as either single- or multi-block motifs. A one-block motif is deduced from alignment of a single, uninterrupted block of nucleotide sequences, and the corresponding PWM is referred to as a one-block model. A bipartite pattern consists of two adjacent blocks separated by a variable length nucleotide spacer (of unspecified sequence), and two PWMs are needed for the corresponding set of bipartite models. Representation of a binding site with a bipartite model is appropriate to the extent that overall conservation across the entire site is improved relative to a one-block model. In some instances, sequence analysis is required to elucidate the bipartite nature of an apparent one-block motif (see details below), since variable spacing between half sites may not be obvious from the experimental findings.

PWM-based one-block motif discovery algorithms that have been developed include CONSENSUS (5), MEME (6), Gibbs Motif Sampler (7–9) and AlignACE (10). Analogous objective functions are used in each of these methods, which, in general, maximize likelihoods or likelihood ratios; however, the methods primarily differ in their approaches to

*To whom correspondence should be addressed. Tel: +1 816 983 6511; Fax: +1 816 983 6515; Email: progan@cmh.edu

searching the multiple local alignment space. CONSENSUS is based on a greedy strategy that progressively adds subsequences to a set of alignments where each iteration extends a bounded number of partial alignments. MEME is an Expectation Maximization (EM)-based method that considers all sites of the training data simultaneously and over iterative training converges to a local maxima. Gibbs Motif Sampler and its variant AlignACE are based on the Gibbs sampling strategy. These algorithms are popular and reasonably accurate, but each has some limitations (11). A recent algorithm, GLAM (11), developed by enhancing Gibbs sampling strategies and implementing simulated annealing optimization methods, appears to perform better than other previous one-block motif discovery algorithms.

Bipartite models are essentially extensions of one-block models that incorporate an intervening gap of unspecified sequence between a pair of adjacent one-block motifs. The first pairwise motifs (i.e. direct repeats and inverted repeats) were defined by Staden (12). Subsequently, a fixed gap size motif algorithm was developed based on the EM method (13), but the underlying model was still the single-block motif, rather than a bipartite model. Previous bipartite pattern search algorithms include BioProspector, which has been extended from the Gibbs Sampler (14), and Co-Bind, which uses a Gibbs sampling strategy to model two-block cooperativity (3). BioProspector is the only available program for discovery of two-block motifs, however, neither of these approaches comprehensively recognizes all combinations of orientations of a bipartite pattern (Figure 1b).

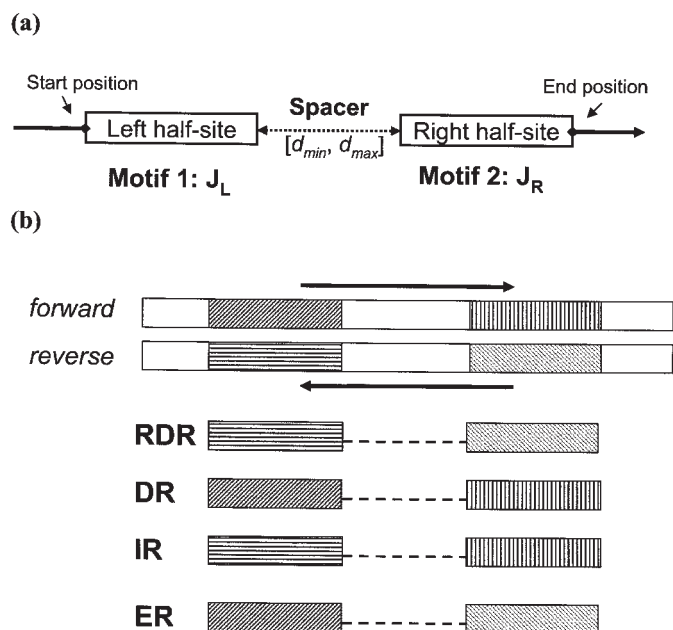


Figure 1. Bipartite patterns on double helical DNA. (a) A bipartite module is an independent functional unit on the upstream/downstream of a regulated gene and is recognized by a homodimer or heterodimer. We assume that the two subunits cooperatively bind to the module with constrained spacers. A bipartite pattern can be expressed as $J_L < D > J_R$. J_m is the width of motif m and D is the gap range as defined in the text. (b) Four possible types of a bipartite pattern [see (19)]. The arrows point from 5' to 3' direction. Filled areas are motifs. Four possible types of a bipartite pattern: RDR—reverse direct repeats, DR—direct repeats, ER—everted repeats and IR—inverted repeats.

Bayesian models for multiple elements and motifs were proposed by Liu *et al.* (8), which were subsequently implemented as a two-block motif search algorithm based on Gibbs sampling strategies in BioProspector (14). BioProspector uses a modified Gibbs sampling strategy with zero- to third-order Markov background models to identify two binding site motifs within a maximum allowable separation of 40 nt. BioProspector only detects direct repeats (one type of two-block patterns) on either the forward or the reverse strand, or a perfect palindrome.

The information theory-based bipartite model was first proposed and used for the analysis of ribosome binding sites in *Escherichia coli*, i.e. Shine–Dalgarno and translation initiation sites (15). The entropy of individual half-site motifs is minimized, rather than the combined entropy of both half-site motifs in a single integrated model. An initial one-block alignment is used to situate each of the half sites around the same coordinate; then one-block alignment of each half-side is carried out separately. Currently, this approach detects direct and imperfect repeats on a single strand, including bipartite patterns separated by long gaps of unspecified sequences.

We present an extension of the bipartite models introduced by Shultzaberger *et al.* (15), under the framework of Shannon information theory (16,17). There are notable differences between our respective approaches. First, we minimize the total entropy present in both half sites as an integral model (rather than as independent half sites). Second, we incorporate the length of the gap as a penalty (i.e. related to the surprisal of observing a particular gap length) in the objective function, and then present a greedy algorithm (9,18) to search for bipartite patterns in the multiple alignment space, based on principle of information maximization. The new search algorithm, known as *bipartite pattern discovery* (Bipad), searches a mixture of the four types of possible half-site orientations in the whole alignment space and generates bipartite models. We then validated this algorithm by constructing simulated, previously characterized and novel binding sites including (i) simulated bipartite datasets containing a mixture of four types of bipartite patterns, (ii) one-block and short-gapped bipartite patterns for cyclic AMP receptor protein (CRP) homodimeric binding sites from *E.coli*, (iii) long-gapped bipartite patterns for the sigma factor binding sites from *Bacillus subtilis* and (iv) human vitamin D₃ receptor/retinoid X receptor α (VDR/RXR α) heterodimeric binding sites. Finally, we compare the performance of Bipad with the other previously mentioned algorithms.

METHODS

Defining a bipartite pattern

A bipartite pattern (Figure 1a) is an independent functional unit upstream or downstream of a co-regulated gene, which is often recognized by a homodimer and heterodimer protein complex. These patterns are typically characterized [(19,20) and references therein] by the four possible orientations of imperfect and perfect repeat units (Figure 1b): (i) Direct Repeat (DR), (ii) Everted Repeat (ER), (iii) Inverted Repeat (IR) and (iv) Reversed Direct Repeat (RDR, direct repeat on the reverse strand). Let q be a type of repeats. $q \in Q$, where $Q = \{DR, ER, IR, RDR\}$. Let type distribution be denoted by

$\psi(Q)$. All repeats are most probably imperfect and their widths could be different from each other. We define the term ‘repeat’ to mean a subunit, a block or a half-site motif of a bipartite pattern, but the half sites are not necessarily composed of identical or imperfect duplicates of the same sequence. In fact, two half-site motifs may have the same or different lengths (but contain related subsequences). The convention is useful to specify all possible orientations of the relationships between the half sites.

The gap distance between repeats is flexible. Our algorithm allows for a gap range of any size. Based on gap length, bipartite patterns can be divided into two classes: (i) short-gapped bipartite pattern, two blocks separated by a short gap distance (gap distance is ≤ 10 bp or a DNA helical turn, i.e. CRP, a homodimer, binding sites and VDR/RXR, a heterodimer, binding sites) and (ii) long gapped bipartite pattern, two blocks separated by long gap distance (gap size may be >10 bp, for instance, binding sites for sigma factors of *B.subtilis*). Although such distinctions are arbitrary, they may have biological significance, since they could correspond to structural features present in either the protein or the nucleic acid. A short gapped bipartite site is recognized and bound to a TF binding dimer, such as a homodimer or heterodimer. A long-gapped bipartite site is consistent with the possibility that two sub-sites (motifs) are recognized and bound initially by separate TFs and subsequently integrated into a single functional complex, e.g. by DNA looping or formation of secondary structures.

Bipartite model assumptions

A bipartite model built to simulate a bipartite pattern in genomic sequences has three components: left and right motif models, and an associated gap function, which is described by a probability distribution, $\omega(D)$, determined from the optimal bipartite alignment. The default sequence model is assumed to be OOPS (one bipartite site occurrence per sequence in the dataset), but ZOOPS (zero or one bipartite occurrence per sequence) is also available as an option. A bipartite pattern consisting of left (L) and right (R) half sites separated by a distance d is denoted as $L<d>R$. We assume that the two half-site motifs are independent but do not overlap. The gap size (d) is constrained based on experimental observation within the range $D = \{d: d_{\min} \leq d \leq d_{\max}\}$. We assume that *a priori*, $d \sim \text{Uniform}(D)$. This assumption implies that we deem every possible gap size equally likely while searching for an optimal pattern.

The left and right motif models are represented by two nucleotide frequency or weight matrices: M_L and M_R with sizes of $|A_X| \times J_L$ and $|A_X| \times J_R$ respectively. J_L and J_R are widths of left and right motif, respectively, $A_X = \{A, C, G, T\}$ and $|A_X| = 4$, the size of nucleotide set. Let $p_m(x_l)$ be the probability (frequency) of the nucleotide x at position l given motif $m \in \{L, R\}$, where $x_l \in A_X$. $p_m(x_l)$ is the element of frequency matrix (i.e. M_L or M_R). A bipartite model can be expressed as $M_L<D>M_R$ or $M_L<[d_{\min}, d_{\max}]>M_R$. A bipartite searching pattern is $J_L<D>J_R$ or $J_L<[d_{\min}, d_{\max}]>J_R$ that allows to search two fixed motif widths separated by a gap range (e.g. Table 3). The pattern format $M_L<d_c>M_R$ or $J_L<d_c>J_R$ is graphically displayed as a bipartite sequence logo (21) or bipartite logo (see Figures 6a–c and 7b and

Table 3) of two half-site motif logos separated by a gap d_c , where d_c is the predominant spacing observed. A bipartite motif found in a single sequence can be denoted as $a_L<d>a_R$ where a_L and a_R are start positions for left and right half-site motifs, respectively (e.g. see Table 2). Therefore, the relationship $a_R = a_L + J_L + d$ follows from this definition.

Probabilistic model for a bipartite pattern

It is assumed that two half-site motif models (M_L and M_R) in a bipartite pattern have independent probability distribution with a constrained gap size d following some distance probability distribution, $\omega(D)$. Each position in a motif is also assumed to be independent (22). Given a sequence x and bipartite model, the probabilistic model for a bipartite pattern can be expressed as

$$P(x | M_L, M_R, d) = \omega(d) \prod_{m \in \{L, R\}} \left\{ \prod_{l=1}^{J_m} \prod_{b \in A_X} p_m(x_l)^{I_{b, x_l}(x, m, d)} \right\} \quad 1$$

where the indicator function $I_{b, x_l}(x, m, d)$, which depends on x , m and d equals 0 if $x_l \neq b$ or $m \notin \{L, R\}$ or $d \notin D$ and equals 1 otherwise. We see that two motif sub-models are independent, but subject to a distance constraint. The indicator function simply shows whether or not a current bipartite pattern is a valid one. The pattern probability is computed once the bipartite model parameters, including the distance distribution $\omega(D)$, are estimated.

Objective function

We used a set of input sequences to estimate the parameters in Equation 1 based on maximizing information principle (23,24). Shannon’s entropy or uncertainty (in bits) was used to define the objective function, IC_{total} , which is a function of two half-site motifs separated by a constrained gap (Equations 2–4) as described previously (15). However, we maximize the IC_{total} as an integral function over the entire binding site rather than constructing separate models for each individual half site, as has been reported in (15). The algorithm assumes that two half-site motifs are independent of each other. The gap penalty (or surprisal) function $g(d)$ is derived from the gap probability distribution $\omega(d)$ which is determined from an optimal bipartite alignment. We assign a zero penalty to the most likely distance.

$$IC_{\text{total}} = IC(L, R, d) - g(d) = IC(L, d) + IC(R, d) - g(d) \quad 2$$

$$IC(m, d) = \sum_{l=1}^{J_m} (\log_2 |A_X| - H_{ml}(X | d) - e(n)) \quad 3$$

$$H_{ml}(X) = \sum_{x \in A_X} p_m(x_l) \log_2 \frac{1}{p_m(x_l)}, \quad A_X = \{A, C, G, T\} \quad 4$$

where $e(n)$ is a sample correction (15), n is the number of DNA sequences, $x_l \in A_X$, J_m is the width of motif $m \in \{L, R\}$, $p_m(x_l)$ is the probability of x at position l given motif m . $p_m(x_l)$ can be computed as a frequency or the estimated probability,

$\hat{p}_m(x_l)$, in a multiple local alignment, $\hat{p}_m(x_l) = (c(x_l) + \beta_x) / (N + \sum_{x \in A_X} \beta_x)$. $c(x_l)$ is the count of symbol x at position l , β_x is the pseudo-count (8,25) of x and N is the number of DNA input sequences in the alignment. Each pseudo-count β_x is set to 0.25×1.5 (11). It is obvious that the left and right motif sub-models are subject to a gap constraint. These two motifs are not permitted to overlap and the gap length range (D) is based on biological observation. The notation $H_{ml}(X|d)$ indicates that entropy is subject to a gap constraint d . The penalty function is simply $-\log_2(\omega(d))$.

Computing individual information content for a bipartite pattern

Given a bipartite model and a testing sequence $x = x_1, x_2, \dots, x_n$, where x_l is the nucleotide at position l . Let two sliding windows with size of J_L and J_R move along the sequence at offset of 1 bp, each time at a position each window covers both forward and reverse strands, and the corresponding information contents are computed separately. The distance between windows is constrained by $[d_{\min}, d_{\max}]$. Therefore, the individual information content (Ri) for a bipartite pattern is computed as

$$Ri(x) = \sum_{m \in \{L,R\}} \left\{ \sum_{l=1}^{J_m} \left(\log_2 |A_X| - \sum_{b \in A_X} \log_2 \hat{p}_m(x_l)^{I_{b,x_l}(x,m,d)} \right) \right\} - g(d) \quad 5$$

where $\hat{p}_m(x_l)$ is the estimated probabilities (frequency) of the nucleotide x_l for motif model m at position l . Because each window covers both strands, we end up with four types of patterns (Figure 1b) at a time. We use Equation 5 to scan a sequence for a new site such that its information content (Ri) is greater than a specified cutoff.

The greedy algorithm

The goal is to maximize the IC_{total} in Equation 2 which can be reduced to minimize the total Shannon's entropy (Equations 3 and 4). We used a greedy algorithm (18) to search the multiple local bipartite alignment (MLBA) space (Θ) and find an optimal solution to the bipartite pattern search problem. Therefore, the new objective function is given by

$$(M_L^*, M_R^*, \omega(d)) = \arg \min_{(a,d) \in \Theta} \left\{ \sum_{m \in \{L,R\}} \left(\sum_{l=1}^{J_m} H_{ml}(X|d) \right) \right\} \quad 6$$

where $d \in D$, $\Theta = \{(a_L^{(i)} \in A_L^{(i)}, a_R^{(i)} \in A_R^{(i)}) : i = (1, 2, \dots, N)\}$, $A_L^{(i)} = \{a : 1 \leq a \leq \ell_i - J_L - J_R - d_{\min} + 1\}$, $A_R^{(i)} = \{a : J_L + d_{\min} + 1 \leq a \leq \ell_i - J_R + 1\}$ and ℓ_i is the length of sequence i . Obviously Θ is the whole MLBA space consisting of all possible combinations of two motif start positions for N sequences given a bipartite searching pattern $J_L < D > J_R$. We take one bipartite alignment (a, d) which can be computed as (M_L, M_R) at a time and hence a total entropy can be computed for each of these bipartite alignments (see Equation 6). A greedy algorithm (described in the next section) was applied to search the MLBA space (Θ) and to find the best alignment such that it

gets the minimum total entropy or maximum total information content. The best alignment corresponds to the best bipartite model $(M_L^*, M_R^*, \omega(d))$. The same idea can be applied to a one-block model without gap constraint ($d_i = 0$ for all i).

Algorithm implementation

The algorithm for building the bipartite model is shown in Figure 2. The Bipad program generates random seeds as the initial start positions of bipartite patterns. For each set of seeds, it sequentially picks up a sequence at a time and enumerates all possible start positions and all possible gap distances for two half-site motifs and computes their entropies. Entropies are stored with their coordinate indices in a vector \mathbf{E} . The total number of legal bipartite combinations for a sequence scanned is $|\mathbf{E}| = |D| \times (\ell_i - (J_L + J_R + |D|) + 1)$. We keep the minimum entropy for each sequence and update the frequency tables (Figure 2). Each pass of this operation is iterated for all input sequences until the total entropy difference between consecutive passes exceeds a very small negative threshold (δ). After each of these cycles, the bipartite alignment with the maximum information content is retained and the above procedure is restarted for a prescribed number of remaining cycles. The bipartite model is built based upon the overall best alignment found among all cycles.

The Bipad algorithm was implemented in C++ and compiled using GNU C++. Inputs include a DNA sequence file, motif widths and a gap range specified by the user. The sequence model can be either OOPS (the default) or ZOOPS. Other parameters, such as pattern output format and alignment modes, can be set through the command line. The program may be run to produce either one-block or bipartite models. Sites may be aligned on either the forward strand only or on both strands, and the motif and gap length range can also be specified. The initial gap length range is selected based on published experimental data and Bipad is executed (using Perl scripts) for multiple, different parameter settings. The results of these scripts are compared to select the models having the maximal information contents.

Performance evaluation for bipartite model

The *performance coefficient* (26) was the metric used to evaluate the performance of each algorithm for the same set of input data. Let $K_L^{(i)}$ and $K_R^{(i)}$ be the sets of known left and right motif positions, respectively, in a sequence i and $O_L^{(i)}$ and $O_R^{(i)}$ be the sets of left and right motif positions located by an algorithm. The performance coefficient for a bipartite pattern on sequence i , ρ_i , is computed as

$$\rho_i = \sum_{m=L}^R |K_m^{(i)} \cap O_m^{(i)}| / \sum_{m=L}^R |K_m^{(i)} \cup O_m^{(i)}|$$

$$\bar{\rho} = \sum_{i=1}^n \rho_i / n \quad 7$$

For a set of n DNA sequences, we compute the average performance ($\bar{\rho}$). If $\rho = 1$, then there is an exact match between predicted and validated site. If $\rho = 0$, then the valid site is missed. The coefficient for a one-block motif is computed by taking $m=R$ or $m=L$, which reduces to the definition given by Pevzner and Sze (26).

Input: N sequences, cycles, bipartite pattern $J_L < [d_{min}, d_{max}] > J_R$ and threshold δ

Output: bipartite model

```

while ( cycles-- > 0 ) {
  Randomly generate start positions for left motif of each sequence  $i$ :  $a_L(i)$ ;
  Set start positions for right motif of sequence  $i$ :  $a_R(i) = a_L(i) + d_{min}$ ;
  Generate frequency tables and calculate Entropies ( $H$ ) for half-site motifs;
  Initializing  $t = 0$  and compute  $H^{(t)} = H^{(t)}(L|d) + H^{(t)}(R|d)$ ,  $d \in [d_{min}, d_{max}]$ ;

  Find the initial minimum  $H_{min}^{(t)}$  in  $H^{(t)}$ ;

  do {
    for  $i = 1$  to  $N$  {
      Do bipartite alignment for each sequence  $i$ ;
      Update frequency tables for each half-site motif;
      Calculate entropies:  $H^{(t+1, i)} = H^{(t+1, i)}(L|d) + H^{(t+1, i)}(R|d)$ ,  $d \in [d_{min}, d_{max}]$ ;
      Store  $H^{(t+1, i)}$  and associated coordinates  $a_L(i)$  and  $a_R(i)$  in  $\mathbf{E}$ ;

      Update the alignment with minimum entropy:  $H_{min}^{(t+1, i)}$  in  $\mathbf{E}$ ;
    }
  } until  $H_{min}^{(t+1)} - H_{min}^{(t)} > \delta$ 
}

Update frequency tables and output the bipartite model

```

Figure 2. An algorithm for building bipartite model. A *bipartite alignment* for sequence i is to start from current positions and then move back and forth along the sequence until every possible combination of a bipartite pattern is enumerated. Each move is subject to a gap range. We record total entropy (H) and store it together with its associated motif start coordinates in a vector \mathbf{E} for each non-overlapping combination of patterns. The best alignment with the minimum entropy found in \mathbf{E} is kept after each of such scanning operations.

Other programs used for comparison

To compare the results of Bipad with other one-block algorithms, we ran the following available algorithms for one-block motif discovery: Gibbs motif sampler (<http://www.bioinfo.rpi.edu/applications/bayesian/gibbs/gibbs.html>) (7), CONSENSUS (<ftp://ftp.genetics.wustl.edu/pub/stormo/Consensus/>) (5) and GLAM (<http://zlab.bu.edu/glam/>) (11). BioProspector (<http://robotics.stanford.edu/~xsliu/BioProspector/>) (14) was compared with Bipad for searching bipartite or two-block motifs.

RESULTS

Analysis of results from simulated bipartite pattern datasets

In a DNA sequence set of bipartite patterns, a mixture of four types of patterns [*Imperfect* Direct Repeat (DR), Everted Repeat (ER), Inverted Repeat (IR) and Reverse strand Direct Repeat (RDR)] may coexist. Bipad defaults to searching for all of these patterns, however, the search may be limited by the user to specific patterns.

To test the ability of the algorithm to comprehensively search for bipartite patterns in all orientations on both strands (Figure 1b), we simulated binding sites in a variety of sequence contexts. Each sequence in the dataset is 100 bp long. The simulated datasets are made of the following combinations:

- (i) Sample size (s): 20 or 100 sequences.
- (ii) DNA sequence background distributions (K): (a) uniform (A, C, G and T are equally likely occurring), (b) AT richness (AT content = 60% and GC = 40%) and (c) GC richness (GC = 60% and AT = 40%). In (b) and (c), the nucleotides A and T, and G and C are equally likely.
- (iii) Gap distribution ($\omega(D)$): Let $P\{Y = d\} = \omega(d)$, $P\{Y = 3 \text{ bp}\} = 0.25$, $P\{Y = 4 \text{ bp}\} = 0.50$, and $P\{Y = 5 \text{ bp}\} = 0.25$.
- (iv) Motif conservation levels (z): (z) high, (b) mid and (c) low. A high conservation motif is formed such that at any position a dominant nucleotide has a probability of 0.91 and each of the rest is 0.03 (27) or the information content at position l (IC_l) = 1.42 bits. A mid conservation motif is formed such that at any position of a dominant nucleotide has a probability of 0.79 and each of the rest is 0.07 or IC_l = 0.93 bits. A low conservation motif is

Initializing \mathbf{s} , k , \mathbf{z} , $M_L < [d_{min}, d_{max}] > M_R$

for $i = 1$ to $|\mathbf{s}|$ do {

1. Generate a sequence s_i based on a background $k \in K$,
where $K = \{\text{Uniform, AT-Rich, GC-Rich}\}$;
2. Draw a pattern type q following type probability distribution $\psi(Q)$,
where $Q = \{\text{DR, ER, IR, RDR}\}$;
3. Generate two motifs (L, R) according to q , $\varphi(z_1, z_2)$ and frequency matrices
(M_L, M_R),
where $z_i \in \{\text{High, Mid, Low}\}$;
4. Draw a start position $a_i \sim \text{Uniform}(A_L)$,
where $A_L = \{a : 1 \leq a \leq l_i - J_L - J_R - d_{min} + 1\}$;
5. Plant the first motif (L) into the position a_i ;
6. Draw a gap distance on sequence i , $d_i \sim \omega(D)$, where $D = [d_{min}, d_{max}]$;
7. Insert the second motif (R) into position $(a_i + J_L + d_i)$;

}

Save simulation sequence dataset (\mathbf{s}) in a file.

Figure 3. Procedure for generating bipartite dataset. This is an iterative process until a given number of sequence $|\mathbf{s}|$ is generated.

formed such that at any position of a dominant nucleotide has probability of 0.70 and each of the rest is 0.10 or $IC_l = 0.64$ bits. Let $P(z)$ be the probability of z level conservations on half-site motifs. For example, a bipartite combination (High, High) has the probability $P(z_1 = \text{high}, z_2 = \text{high}) = 1.0$, meaning that both left and right half-site motif conservation levels are high.

- (v) Bipartite pattern conservation level distribution $\varphi(z_1, z_2)$:
(a) (High, High); (b) (Mid, Mid); and (c) (Low, Low).
- (vi) Bipartite pattern implanted $(J_L[d_{min}, d_{max}]J_R): 9 < [3,5] > 9$.
- (vii) Type distribution of bipartite patterns $\psi(Q)$:
 $\psi(\text{DR}) = 0.70$, $\psi(\text{ER}) = 0.10$, $\psi(\text{IR}) = 0.10$ and $\psi(\text{RDR}) = 0.10$.

The simulation procedure for generating a bipartite dataset is presented in Figure 3. The simulated datasets were used to test Bipad's capability to detect a mixture of four types of bipartite patterns planted in different background sequences.

Bipad successfully located the correct positions of the embedded bipartite sites (Table 1). The bipartite search pattern is set to $9 < [3,5] > 9$, the pattern we embedded during the simulation. We did not compare simulations of Bipad with BioProspector, because the latter failed to detect almost all cases in the simulated datasets. BioProspector was designed to search for one type in a run, either direct repeats (DR/RDR) or palindromic sequence pairs, not a mixture of four possible types. Bipad scanned both forward and reverse strands, looking for four possible types of a bipartite pattern (Figure 1b). Based on the experimental ground truth, the average performance with >0.5 signifies successful detection of a bipartite pattern in the dataset. In simulated bipartite datasets, we generated a mixture of all four types of bipartite patterns

Table 1. Results for simulated bipartite datasets^a

Sample size ^b	Bipartite conservation, level distribution	Background distribution		
		Uniform	AT-rich	GC-rich
Small	(High, High)	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	(Mid, Mid)	0.93 ± 0.06	0.96 ± 0.03	0.96 ± 0.02
	(Low, Low)	0.73 ± 0.16	0.44 ± 0.20	0.31 ± 0.19
Large	(High, High)	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
	(Mid, Mid)	0.95 ± 0.05	0.93 ± 0.06	0.94 ± 0.03
	(Low, Low)	0.77 ± 0.16	0.76 ± 0.17	0.80 ± 0.14

^aBipartite alignments were applied on both strands. The number in the cell is the average performance calculated as in Equation 7 plus its standard error. The performance coefficient ≤ 0.50 is considered as failure detection (number in boldface).

^bSmall sample size is 20 sequences and large sample size is 100 sequences.

(Figure 1b) and Bipad successfully detected planted sites in each of these circumstances (Table 1).

Simulation studies also showed that the performance or accuracy of motif discovery algorithms depends not only on sample size but also on the motif conservation level. Bipad successfully detected binding sites embedded in a uniform background sequence, and its performance increased to 1.0 (exact match with all implanted sites). Bipad's performance increases from 0.73 to 1.0 as the conservation level increases from low (0.64 bits per base), through mid (0.93 bits per base) to high (1.42 bits per base). The bipartite pattern conservation level was increased in both small and large sample datasets (see Table 1 for details). Mid- and high levels of conservation had very similar performance criteria. In large datasets, Bipad successfully detected all patterns of nine total combinations and had nearly similar performance, independent of the background distributions.

The performance also relies to a limited degree on background composition, especially for small datasets. Although our bipartite models assumed a uniform background, algorithmic performance was also adequate for non-uniform backgrounds (Table 1). In AT-rich sequence sets, Bipad behaved the same way as it did in the uniform background, except for the failure to detect a sufficient percentage of binding sites in a small sample size and low bipartite pattern conservation level dataset (performance is 0.44). In simulated binding sites within a GC-rich background, the Bipad performance was similar to that of the AT-rich background. Simulation results showed that Bipad was able to detect a mixture of four types of bipartite patterns in all three backgrounds. The accuracy increases as the degree of motif conservation increases. Larger sample sizes are needed to correctly detect a set of binding sites with a lower overall conservation. The predicted information content (bits per base pair) was very close to the expected IC (Figure 4) for all three background compositions. In each case where a low conservation level was expected, the predicted IC value was a little higher than what was anticipated. Figure 5 shows the bipartite motif IC distributions for three different background sequences. In each case we can see that highly conserved motif IC curves almost separated from the background distribution. This explained why Bipad was able to detect all embedded sites of high conservation patterns ($\bar{p} = 1.0$) regardless of background. On the other hand, in each case, the IC distribution for motifs with low levels of conservation overlapped to some extent with the corresponding distribution of background sequences. There is a greater overlap between these distributions for AT- and GC-rich backgrounds, suggesting that there is less power to detect

embedded patterns with low levels of conservation in these backgrounds for small samples (boldface values in Table 1).

Algorithm comparison for CRP dataset

Cyclic AMP receptor protein. CRP protein is a positive control factor necessary for the expression of catabolite repressible genes. It is a prokaryotic dimeric DNA binding protein that binds to adjacent DNA major grooves in a bipartite pattern. It is known that there is at least one CRP-binding site in each of the 18 sequences and the location of these binding sites have been determined by DNA footprinting studies (28). Each sequence is 105 bp in length and each determined motif width is 22 bp long. The nucleotide composition of the CRP-binding sequences is A (30.26%), C (18.25%), G (20.90%) and T (30.58%), a background rich in A and T. This CRP-binding site data has been commonly used as a golden standard for testing motif discovery algorithms (9,14,28–30). We used the dataset to verify our algorithm and for comparison with other one- and two-block motif discovery algorithms (Table 2). The Bipad algorithm was able to recognize one-block or bipartite patterns, regardless of which type of motif was selected.

We set the motif length to 22 bp for one-block motif models based on experimentally defined single-block sites. For bipartite models, we initially set each of half-site motif width to 8 bp and allowed for a gap ranging from 2 to 6 bp (denoted as $8<[2,6]>8$ in bipartite pattern format). Subsequently, the distribution of average information across the binding site justified reducing the bipartite search pattern to $5<[2,9]>8$ (see below). This constraint permits the length of the entire pattern

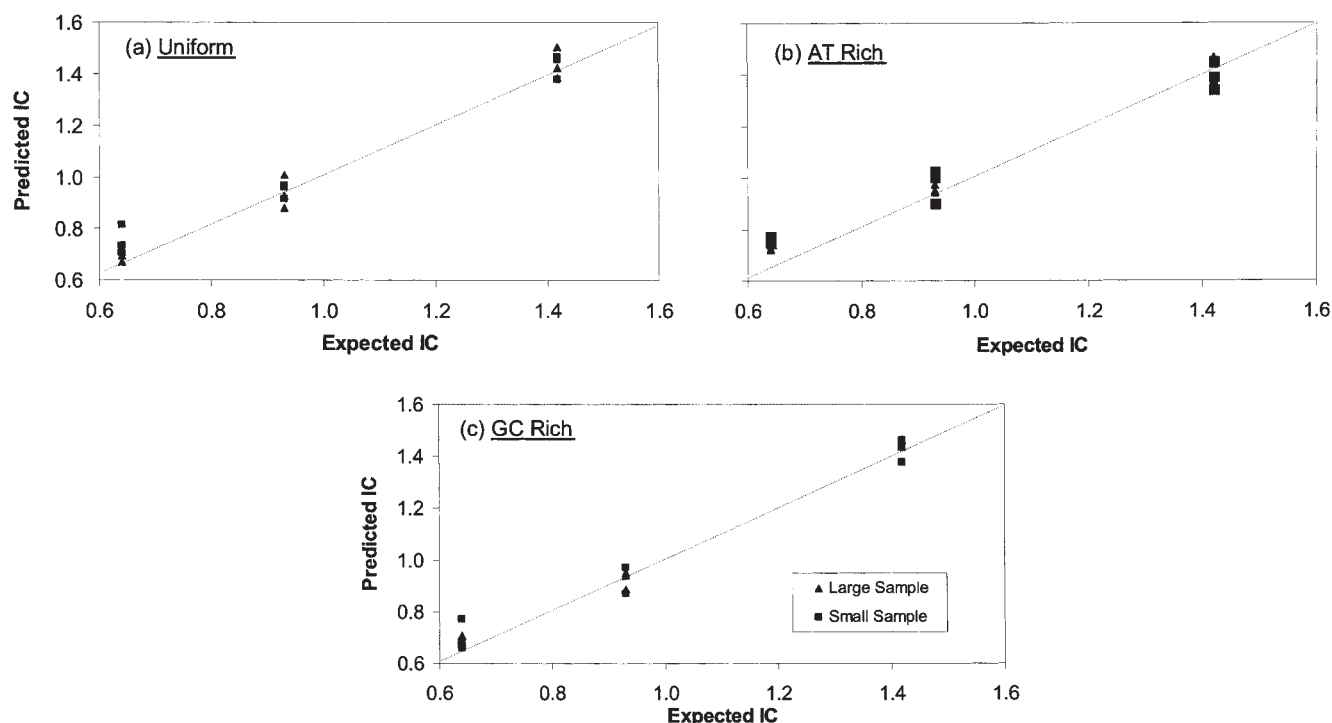


Figure 4. Plots of the predicted information content (IC, bits per base pair) versus the expected IC for three different background sequence compositions: (a) Uniform; (b) AT-Rich; and (c) GC-Rich. Each graph shows three repetitive simulations for large (100 sequences) and small sample (20) sizes.

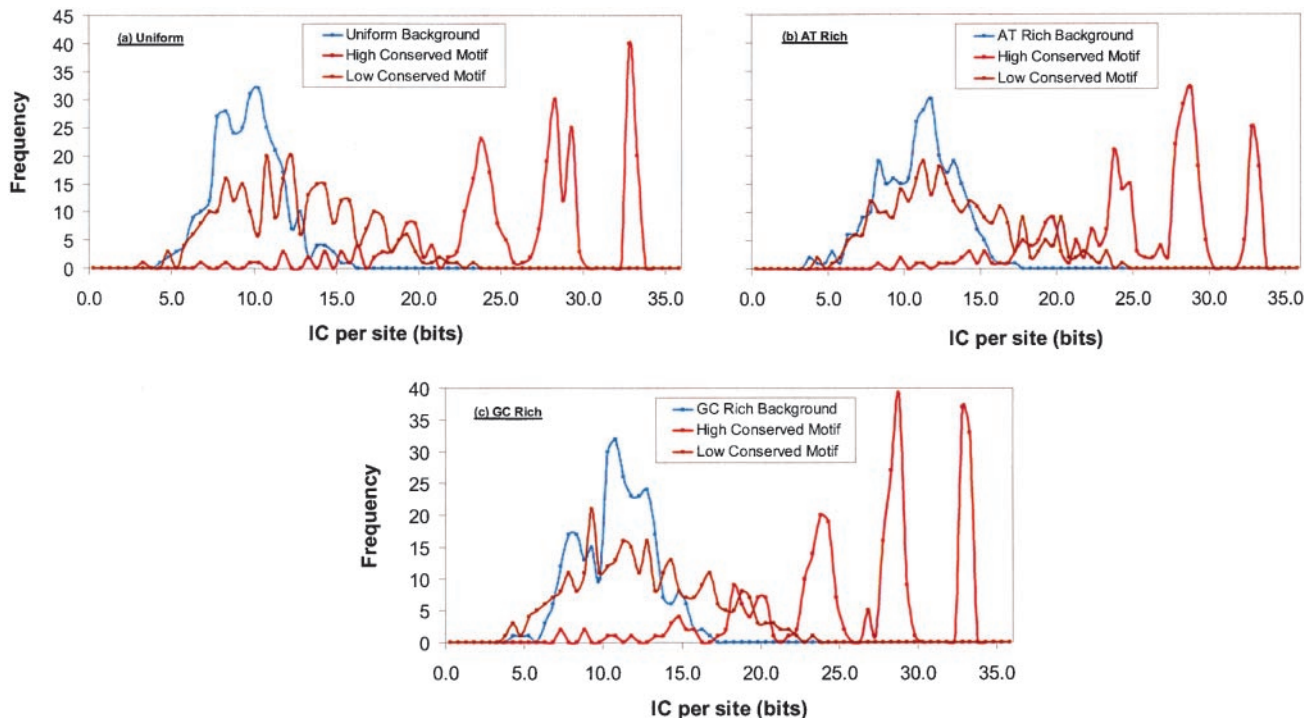


Figure 5. Plots of the motif information content (bits) distributions for three different simulated backgrounds where bipartite motifs were implanted: (a) uniform; (b) AT-rich and (c) GC-rich. Each graph shows the IC distributions of bipartite motifs found by Bipad in pure background sequences (blue line), patterns with either high (red line) or low (orange line) degree of conservation embedded in the same background.

Table 2. Comparison results for CRP-binding data^a

Sequences	Footprint coordinates	Single-block motif (width = 22 bp)				Bipartite motif ^b	
		GLM	Gibbs	Consen	Bipad	BioProspector	Bipad
cole1	17, 61	61	64	64	62	64<3>72	64<6>75
ecoarabop	17, 55	55	58	58	56	58<3>66	58<6>69
ecobglr1	76	76	79	79	77	79<3>87	79<6>90
ecocrp	63	63	66	66	64	64<5>74	64<8>77
ecocya	50	50	44	47	49	14<2>21	53<6>64
ecodeop	7, 60	7	54	4	59	10<3>18	10<6>21
ecogale	42	42	45	27	43	45<3>53	47<4>56
ecoilvbpr	39	39	42	42	40	16<7>28	42<8>55
ecolac	9, 80	9	12	12	8	12<3>20	84<5>94
ecomale	14	14	8	11	13	17<3>25	17<6>28
ecomalk	29, 61	61	55	58	60	32<9>46	64<8>77
ecomalt	41	41	35	38	40	44<9>58	44<6>55
ecoompa	48	48	42	45	47	47<7>59	51<6>62
ecotnaa	71	71	74	74	72	74<3>82	74<6>85
ecouxul	17	17	20	20	16	80<3>88	20<8>33
pbr-p4	53	53	56	56	52	56<3>64	56<6>67
trn9cat	1, 84	84	78	81	83	2<5>11	2<5>12
tdc	78	78	72	81	77	81<3>89	81<6>92
$\bar{\rho}$ ^c	—	1.0	0.69	0.74	0.92	0.67	0.78
Standard error	—	0.0	0.09	0.13	0.0	0.31	0.06

^aThe optimal bipartite searching pattern used here was 5<[2,9]>8 and we set 500 cycles were run for all training procedures.

^bA bipartite motif on a sequence is expressed as $a_L<d>a_R$. The number between the brackets (d) is the gap size, the number on the left-side of the bracket (a_L) denotes the first motif start position and on the right-side is the second motif start position (a_R).

^c $\bar{\rho}$ is the average performance, $\bar{\rho} = \sum_{i=1}^n \rho_i/n$, where ρ_i is the performance coefficient for sequence i . The performance coefficient for a bipartite pattern was calculated as if it were a one-block motif, because the second motif positions are unknown in this case.

to range from 15 to 22 bp. Since one motif per sequence is assumed, if a program output contained more than a single site in a sequence, the one most closely resembling previously identified sites was selected for comparison.

One-block patterns. The results of one-block pattern comparisons (listed in Table 2) show that both Bipad and GLAM (11) correctly located all motif sites in each of the 18 sequences. The average performance for both Bipad and GLAM (0.92 and

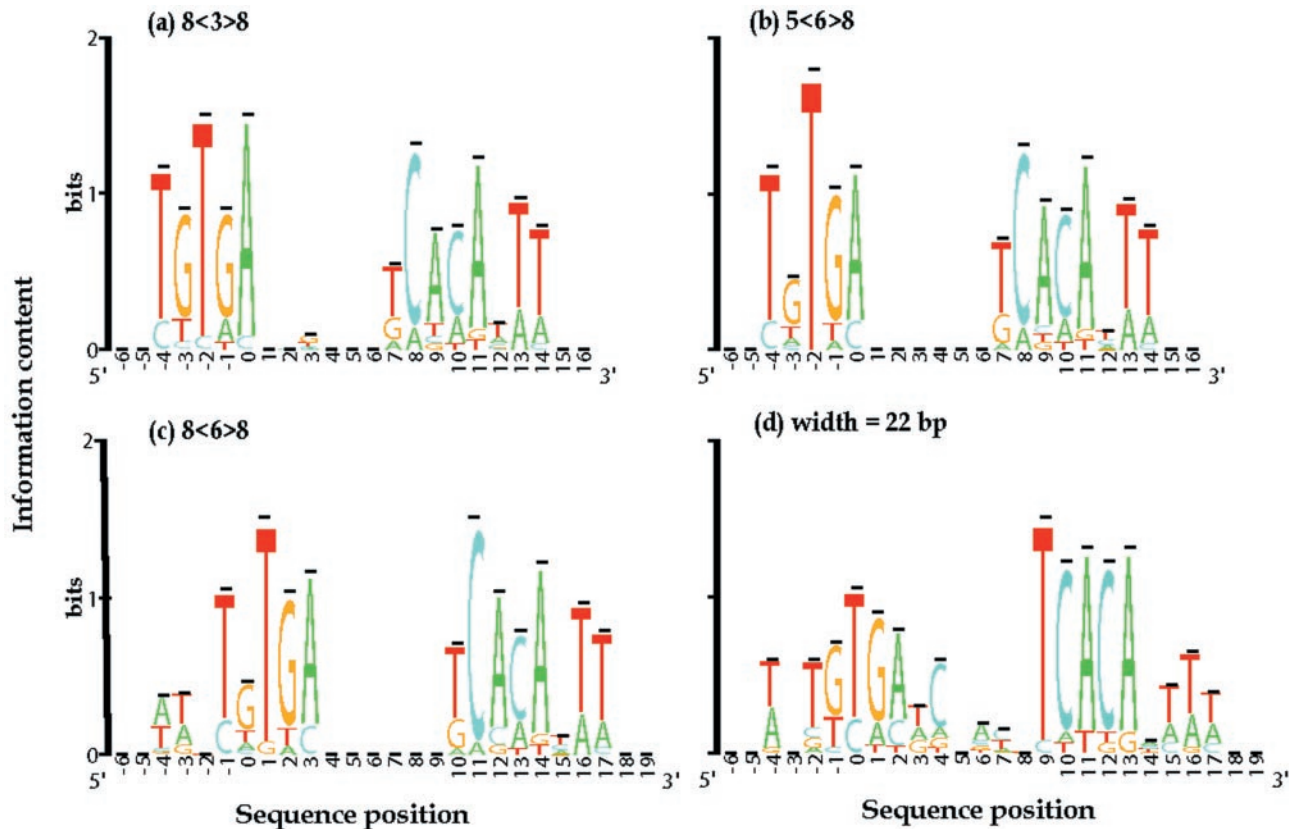


Figure 6. Sequence logos for one-block and bipartite models of CRP-binding sites. (a) Bipartite pattern $8<3>8$, dominant spacer length is 3 bp; information in left half site is dominated by five positions (–4 to 0) with positions 1, 2, and 3 considered noise; total information for left site is $R_{\text{seq}}(\text{left}) = 6.03$ bits per 8 bp and for the right-site $R_{\text{seq}}(\text{right}) = 6.62$ bits per 8 bp; after removing these noise signals the total information for both half sites increased and the model was reduced to that seen in panel (b). (b) Bipartite pattern $5<6>8$, dominant spacer lengths are 6 bp; total information for left site is $R_{\text{seq}}(\text{left}) = 5.65$ bits per 5 bp and for right-site $R_{\text{seq}}(\text{right}) = 7.02$ bits per 8 bp. (c) $5<6>8$ extended to $8<6>8$ with $R_{\text{seq}}(\text{left}) = 6.04$ bits and $R_{\text{seq}}(\text{right}) = 7.19$ bits. (d) Motif width of one-block model is 22 bp as demonstrated in DNA footprint studies; the total information for this model is $R_{\text{seq}}(\text{one-block}) = 10.42$ bits per 22 bp. Thus, the bipartite model contains >2 bits per site more than the one-block model and it is obvious from the distribution of information in the logos that the pattern can be subdivided into two half sites.

1.0, respectively) were excellent and nearly equivalent. Bipad predicted an alignment that was a single nucleotide offset from the footprints of experimentally identified CRP sites (Table 2). This occurred because the experimentally determined motifs exhibit lower conservation at their terminal positions (Figure 6c), but exhibit approximately equal information contents at these positions. Because Bipad maximizes information across the site, small differences in information content at these positions can result in realignment of experimentally defined binding sites. In contrast, the average coefficient for CONSENSUS was 0.74 and has one predicted miss, as indicated by a larger variance (0.13). The lower performance of CONSENSUS resulted from the fact that most positions it located had a 3 bp offset from identified sites (Table 2); and the whole pattern was shifted 3 bp downstream relative to the original data. The Gibbs motif algorithm failed to locate a significant percentage of sites (average performance is 0.69). The most highly conserved positions (shown in boldface) of the consensus sequence derived from the one-block search option of Bipad, **WNTGTGADCNAYNTCACADWWW**, form a palindrome separated by a 6 nt gap.

Short gapped bipartite pattern. BioProspector correctly locates sites specified as perfect palindromes with a

constrained gap distance ranging from 1 to 4 bp (14). However, as we can see in Figure 6a–c, the bipartite binding site lacks perfect symmetry and is a partially imperfect palindrome, possibly because of the inherent asymmetry of the operon promoter region. Unlike BioProspector, Bipad does not require an assumption that the sequence pattern has a palindromic structure, since the algorithm recognizes all four types of patterns including perfect and imperfect palindromes.

Bipad and BioProspector were compared by searching the CRP data with the $5<[2,9]>8$ search pattern on the same strand (results shown in Table 2). Bipad correctly located all sites (average performance = 0.78), whereas BioProspector had a performance of 0.67. A larger variance (0.31) indicated three predicted misses for BioProspector. For the initial search of the $8<[2,6]>8$ pattern, BioProspector missed seven predicted sites, while Bipad missed a single site (data not shown). After removing noise (defined as IC value of ~ 0 bits) at positions within the left site (position 1, 2 and 3; see Figure 6a), the performance of both Bipad and BioProspector were improved for the resulting bipartite search pattern, $5<[2,9]>8$ (Table 2). BioProspector's previously reported optimal search pattern was a $8<[1,4]>8$ palindrome (14). When Bipad was used to search this data, the performance coefficient increased to 0.79.

Considering that CRP homodimerizes, we extended our bipartite pattern to 8<6>8 which further increased the performance coefficient to 0.91 (Figure 6c). Although the 5<6>8 pattern exhibits the maximum information content and is therefore considered optimal, extension of the 5 nt motif to 8 nt better fits the experimentally identified binding sites. The 8<6>8 model does not contain any more information per base pair than the 5<6>8 model (see below). The consensus bipartite binding sequence deduced by Bipad is **WWNTGTGA<6>TCACANWW**, with the subsequences in boldface forming a palindrome. The 8mer half sites form an imperfect palindrome. The logo shows that the most highly conserved positions correspond to critical contacts required for formation of the complex with CRP (31).

Analysis of the CRP-binding data showed that the bipartite model generated with Bipad exhibited a higher level of sequence conservation than the one-block model. Figure 6 shows the sequence logos for CRP one-block and bipartite models. L and R half-site motifs are evident in the 22 bp one-block model. The total information in the one-block model is $R_{\text{seq}}(\text{one-block}) = 10.42$ bits in the 22 bp site. Sequence logos showed that the bipartite model 8<3>8 can be reduced to an optimal pattern 5<6>8, as very little information, is contained in the terminal 3 bp of the L motif of the 8<3>8 logo. The nucleotide at position -2 in the 8<3>8 model is random and positions -3 and -4 exhibit very little (~ 0.3 bits) information (Figure 6c). In the 8<3>8 model (Figure 6a), the information content for left site is $R_{\text{seq}}(\text{left}) = 6.04$ bits per 8 bp site, whereas $R_{\text{seq}}(\text{right}) = 6.62$ bits for the 8 bp R site. The new model, after removing the noise from the left site, is shown in Figure 6b. Despite removing three positions from the PWM, both half sites gained a small amount of information: $R_{\text{seq}}(\text{left}) = 5.65$ bits for 5 bp L site and $R_{\text{seq}}(\text{right}) = 7.02$ bits for the 8 bp site. The bipartite model gained more than 2 bits of information per site in comparison with the corresponding one-block model.

The optimal pattern found by Bipad is very close to 5<[7,9]>7, similar to the results reported by BioOptimizer (27). Bipad built these two bipartite models and ended up with similar results (Figure 6a–c), except that the site information content increased. Results also showed that the most probable width of CRP-binding sites is 19 bp (5 + 6 + 8), with the pattern shifted 3 bp downstream from the original site (Table 2). We used the experimentally determined single-block motifs to evaluate bipartite pattern recognition. Although a lower performance coefficient was obtained (because the experimentally determined bipartite coordinates were not available), Bipad detected all the identified single block sites. Interestingly, the optimal motif length inferred from Bipad is in agreement with that automatically detected by GLAM in a one-block model (11).

Bipartite modeling on binding sites of sigma factors

Long gapped two-block pattern: sigma factors. We examined datasets for six different two-block sigma transcription factors σ^B , σ^D , σ^E , σ^F , σ^G and σ^H from *B. subtilis*. All the binding site sequences for sigma factors were extracted from the DBTBS transcriptional database (<http://dbtbs.hgc.jp/>) (32). The site length varies among sequences recognized by each of these factors. The bipartite patterns reported for the σ^E consensus

sequences are **Ata<[16,18]>cATAcanT**, the σ^F bipartite binding patterns are **GywTA<15>GgnrAnAnTw** and the bipartite pattern of σ^H binding sequences are **RnAGGAawWW<[11,12]>RnnGAAT** (27,33).








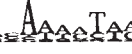




The relative performances of Bipad and BioProspector were compared for 53 binding sites for σ^B (RNA Polymerase general stress factor sigma), 33 sites for σ^D (RNA polymerase flagella, motility, chemotaxis and autolysis sigma factor), 62 sites for σ^E [RNA polymerase sporulation mother cell-specific (early) sigma factor], 14 sites for σ^F [RNA polymerase sporulation forespore-specific (early) sigma factor], 33 sites for σ^G [RNA polymerase sporulation forespore-specific (late) sigma factor] and 23 sites for σ^H (RNA polymerase vegetative and early stationary-phase sigma factor). In our analyses, sequences containing uncharacterized, unknown binding sites were excluded and all bipartite search patterns (column 3 in Table 3) were determined based on sites identified previously.

As the sites identified are known to reside on the same strand, we searched for the DR type bipartite pattern on a single strand. Table 3 shows that both Bipad and BioProspector detected almost all the long-gapped bipartite patterns among the sigma factor sites with various performance coefficients. In those instances where the low performance detection was evident, the predicted sites were shifted a few base pairs upstream of the validated sites, as we selected the longest identified site for the search pattern length. The bipartite logos in Table 3 were generated from the Bipad output. For purposes of comparison with documented motifs, the consensus sequence derived from each logo is also indicated in Table 3; however, the weight matrix more accurately represents these motifs, especially at positions containing little or no information. The gap length between motifs represented the central gap size as expressed in $M_L < d_c > M_R$. Bipad generated the sigma factors' bipartite patterns resembling their prior consensus bipartite patterns (i.e. σ^E , σ^F and σ^H).

Multiple local alignments of VDR/RXR heterodimeric binding sequences

The VDR is a member of the class of nuclear hormone transcription factors that binds as a heterodimer complex with RXR to bipartite vitamin D response elements (VDREs) to activate transcription of downstream target genes. A new model of VDREs was developed using Bipad. Previous models of VDR/RXR binding sites derived from SELEX experiments select for strong binding sites resembling a consensus sequence (34). To elucidate the VDR/RXR binding patterns from a set of natural VDREs, we collected 26 sites from published studies demonstrating experimentally validated VDREs (35–41), and extracted extended versions of these sequences in their natural context. Figure 7a shows that Bipad generated one-block sequence logo for the natural binding sites. While it is comparable to the logo displayed in the JASPAR database (34), the overall level of conservation is lower in the model composed of natural sites. The average information content for the 16mer motif is $R_{\text{seq}} = 10.59$ bits and the one-block consensus sequence is **RRGKTCANNR-RGKTCA**. The VDR/RXR bipartite logo and alignment derived by Bipad are indicated in Figure 7b and c,

Table 3. Comparison of bipartite pattern recognition algorithms for Sigma factor binding sites in *B.subtilis*

<i>sigma</i> factors	No. site	Search pattern	Performance*		Consensus sequence	Bipartite logos†	
			BioPros	Bipad			
B	53	8<[11,16]>8	0.69 ± 0.19	0.81 ± 0.22	KGTTTAAA<13> GGGWAWAN		<13> 
D	33	6<12,16]>9	0.72 ± 0.14	0.78 ± 0.18	NTWAAW<15>C CGATATAA		<15> 
E	62	9<11,15]>10	0.84 ± 0.12	0.85 ± 0.09	GTCATATTT<13> >CATAYAWTD W		<13> 
F	14	7<[15,19]>10	0.73 ± 0.11	0.73 ± 0.11	ANGTWTA<15> GAVAWMTW R		<15> 
G	33	7<[15,19]>8	0.75 ± 0.20	0.85 ± 0.19	GYATAAW<15> MAWAATAA		<15> 
H	23	9<[9,16]>8	0.57 ± 0.14	0.58 ± 0.15	RVAGGAWWW <11>WRMGAA T		<11> 

*The data in each cell are the average performance and its standard error. The average performance is calculated as $\bar{\rho} = \sum_{i=1}^n \rho_i/n$, where ρ_i is the performance coefficient for sequence i .

†A bipartite logo is expressed as $M_L < d_c > M_R$, M_L and M_R is the frequency matrices for left and right motifs with width J_L and J_R respectively, d_c is the central gap distance. Half-site sequence logos were drawn based upon frequency matrices M_L and M_R respectively and had gap d_c labeled in between. Bipartite logos were generated based upon the bipartite models from Bipad. A bipartite logo is drawn with our program Bipad_logo.pl which is based on two programs, makelogo (21) or glam_logo.pl (<http://zlab.bu.edu/glam/>), both of which generate one-block sequence logos (21).

respectively. The bipartite search pattern was set to $7 < [0,6] > 7$, resulting in information contents for left and right half sites of 6.76 and 6.57 bits, respectively. The total information in the bipartite model exceeds the one-block motif by 2.74 bits relative, indicating that it more accurately depicts sequence conservation across the binding site, and therefore, the contacts recognized by the heterodimer. We noticed that the one-block model in the JASPAR database consists predominantly of strong VDRE binding sites (because it is based entirely on 10 binding sites derived from SELEX experiments) and has information content, R_{seq} (15mer motif) = 15.70 bits. Both single-block and bipartite models based on natural sites more comprehensively depict the potential of VDR/RXR to recognize binding sites in the genome *in vivo*, since both strong and weak binding sites were included. For example, in the JASPAR one-block model, an invariant 'T' was detected at the equivalent position +8 in the sequence logos shown in Figure 7. However, oligonucleotides with either a 'T' or 'G' at that coordinate have approximately the same affinity for VDR/RXR heterodimer binding (35), as manifested in our models (see Figure 7a and b on position 8). The VDR/RXR bipartite consensus sequence is derived as **DRGKTCA<2>DRGKTCA**. The predominant nucleotide spacer between the half sites is 2 bp in length and ranges from 0 to 6 nt in length. While the core **RGKTCA** consensus repeat element is very similar to that obtained from a set of experimentally derived (SELEX) high affinity sites, the naturally derived binding sites showed significantly greater sequence variation.

DISCUSSION

We developed and tested an algorithm for modeling bipartite DNA binding sites based on minimizing Shannon entropy across the entire site. Although these models are characterized by their simplicity and generalizable assumptions, the performance of these models for actual or simulated binding site data is comparable or improved over other, more complex approaches. Successful detection of binding sites with Bipad illustrates the preference for using the simplest pattern recognition methods to explain the available data (42).

We applied a greedy algorithm to search the bipartite alignment space for the patterns using information maximization methods. Since the models are based on Shannon's information theory, the background composition is uncorrected for non-uniformity, however, biased distributions did not significantly impact either the models or binding site detection. Using simulated binding site distributions, this assumption could be generalized, since it represents the average situation. In comparison with other popular algorithms using real binding site datasets (CRP and sigma factors), Bipad has good performance in both one- and two-block motif discovery. Bipad detected a more flexible VDRE core repeat than traditional one-block alignment methods.

The greedy search with cyclic randomized initiation can avoid the occurrence of potential local minima, which has been observed using MEME (6). A single cycle starting from a set of random seeds is not sufficient to approximate an optimal solution, because of the possibility that it may

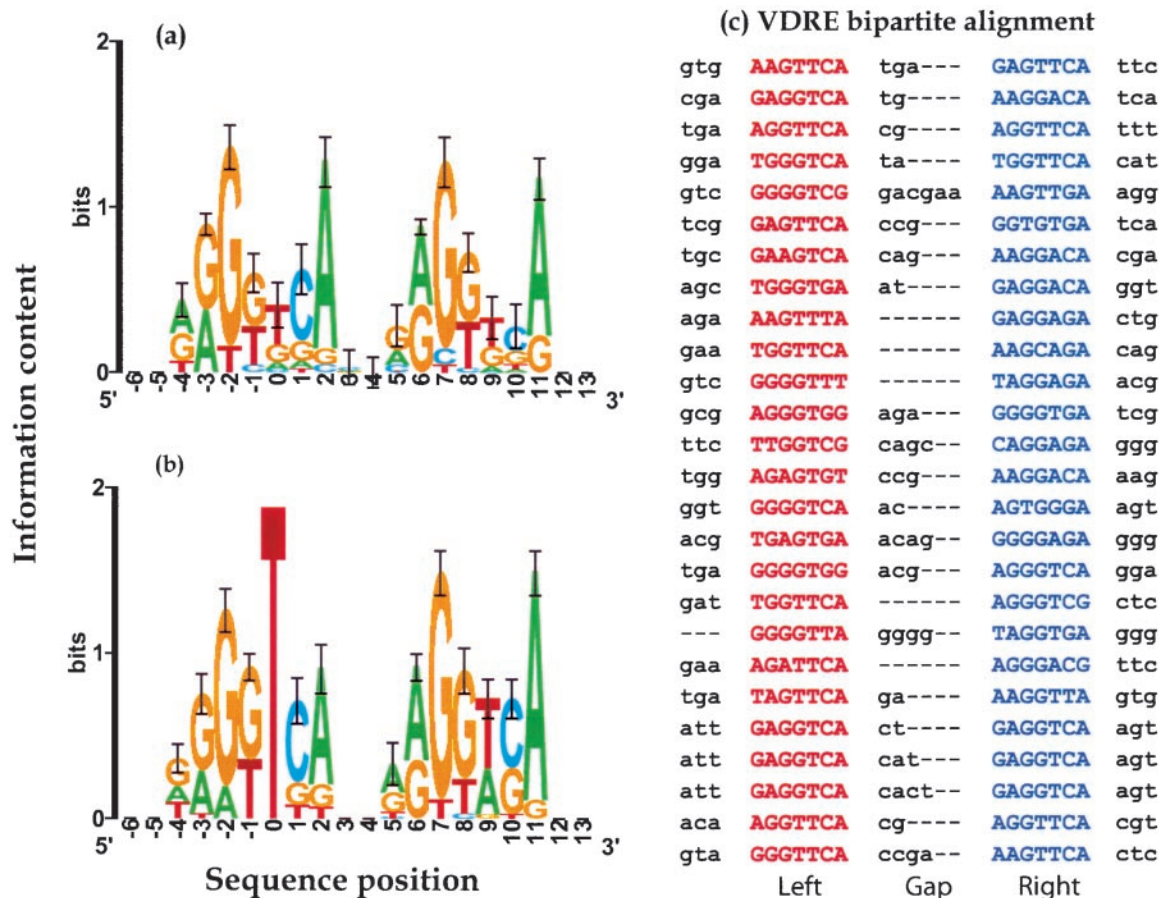


Figure 7. Sequence logos of VDR/RXR heterodimeric binding sites. (a) Sequence logo based on 26 VDR/RXR binding sites. One-block alignment was performed by Bipad. R_{seq} (16-mer motif) = 10.59 bits. (b) Twenty-six natural binding sites were aligned by Bipad as a bipartite pattern. The bipartite pattern 7<[0,6]>7 has the maximum information content of all possible patterns. R_{seq} (7mer left motif) = 6.76 bits and R_{seq} (7mer right motif) = 6.57 bits. (c) Bipartite multiple alignment result with flexible spacing between directly repeated motifs. The left and right motifs are shown in red and blue upper-case, respectively. Nucleotides within the gap interval between two motifs are indicated in lower-case. Dashes are added to facilitate alignment of the half sites and to indicate the lengths of the spacers in each binding site.

converge to local minima. Additional cycles decrease the likelihood of such false convergence. Our simulation experiments showed that, in general, 500 cycles were sufficient to approximate an optimal solution. Approximately 15 s were required to complete bipartite training on the CRP dataset using a 1.2 GHz computer. In each cycle, the greedy search guarantees a convergence (e.g. Figure 2) and the computing cost is $O(N \times |E|)$ per cycle.

Low-complexity patterns, such as poly(A) or poly(T) tracts, may frequently occur in some training sets and in eukaryotic genomes. The results of our simulation studies imply that backgrounds enriched for these sequences may mislead information content-oriented pattern discovery algorithms. These types of repetitive elements should be masked with programs such as RepeatMasker (<http://www.repeatmasker.org/>) (43), nseg (<ftp://ftp.ncbi.nih.gov/pub/seg/nseg/>) and dust (<ftp://ftp.ncbi.nih.gov/pub/tatusov/dust/>), since such sequences have been found to affect multiple local alignments of target binding sites (11).

A reasonable hypothetical bipartite search pattern is required to perform meaningful sequence alignments and thus to accurately locate real binding sites. Bipad can automate the refinement procedures by minimizing entropy and

removing potential sources of noise, increasing the likelihood that the bipartite model will be biologically significant. It has been suggested that simulated annealing is efficient and better for one-block motif discovery than Gibbs sampling (11). Our greedy searching algorithm for bipartite pattern discovery could be seen as a special case of simulated annealing method. While it converges much faster than the simulated annealing method in a single cycle, it requires a greater number of cycles to reach the global minima.

The individual information content of a binding site is related to the enthalpy of its interaction with the protein recognizer (44). The information contents of eukaryotic binding sites identified with one-block and bipartite models built with Bipad have also been related to experimentally measured binding affinities (45,46).

ACKNOWLEDGEMENTS

This work was supported by the grant PHS ES10855-02 from the National Institute of Environmental Health Sciences and by the Merck Genome Research Foundation.

REFERENCES

- Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
- Gadiraju, S., Vyhldal, C., Leeder, J.S. and Rogan, P.K. (2003) Genome-wide prediction, display and refinement of binding sites with information theory-based models. *BMC Bioinformatics*, **4**, 38.
- GuhaThurta, D. and Stormo, G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
- Staden, R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.
- Hertz, G.Z. and Stormo, G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
- Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, Stanford, CA, AAAI Press, Bethesda, MD, pp. 28–36.
- Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
- Liu, J.S., Neuwald, A.F. and Lawrence, C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.
- Liu, J.S. (2001) *Monte Carlo Strategies for Scientific Computing*. Springer Verlag, NY.
- Roth, F.R., Hughes, J.D., Estep, P.E. and Church, G.M. (1998) Finding DNA regulatory motifs with unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
- Frith, M.C., Hansen, U., Spouge, J.L. and Weng, Z. (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.*, **32**, 189–200.
- Staden, R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *CABIOS*, **5**, 89–96.
- Cardon, L.R. and Stormo, G.D. (1992) An expectation maximization algorithm for identifying protein binding sites with variable gaps from unaligned DNA fragments. *J. Mol. Biol.*, **223**, 159–170.
- Liu, X., Brutlag, D.L. and Liu, J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.
- Shultzaberger, R.K., Bucheimer, R.E., Rudd, K.E. and Schneider, T.D. (2001) Anatomy of *Escherichia coli* ribosome binding sites. *J. Mol. Biol.*, **313**, 215–228.
- Shannon, C.E. and Weaver, W. (1949) *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana, IL.
- Cover, T.M. and Thomas, J.A. (1991) *Elements of Information Theory*. John Wiley & Sons, Inc., London.
- Cormen, T.H., Leiserson, C.E. and Rivest, R.L. (2001) *Introduction to Algorithms*. The MIT Press, Cambridge, MA.
- Handschin, C. and Meyer, U.A. (2003) Induction of drug metabolism: the role of nuclear receptors. *Pharmacol. Rev.*, **55**, 649–673.
- Kliwer, S.A., Goodwin, B. and Willson, T.M. (2002) The nuclear pregnane X receptor: a key regulator of xenobiotic metabolism. *Endocr. Rev.*, **23**, 687–702.
- Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
- 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.
- Schneider, T.D. and Mastrorade, D.N. (1996) Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. *Discr. Appl. Math.*, **71**, 259–268.
- Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press UK, London.
- Brown, M., Hughey, R., Krogh, A., Mian, I.S., Sjolander, K. and Haussler, D. (1993) Using Dirichlet Mixture Priors to Derive Hidden Markov Models for Protein Families. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 47–55.
- Pevzner, P.A. and Sze, S.-H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, La Jolla, CA. AAAI Press, Heidelberg, Germany, pp. 269–278.
- Jensen, S.T. and Liu, J.S. (2004) BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics*, **20**, 1557–1564.
- Stormo, G.D. and Hartzell, G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.
- Liu, J.S. (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.*, **94**, 958–966.
- Lawrence, C.E. and Reilly, A.A. (1990) An expectation maximization algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Struct. Func. Genet.*, **7**, 41–51.
- Gunasekera, A., Ebright, Y.W. and Ebright, R.H. (1992) DNA sequence determinants for binding of the *Escherichia coli* catabolite gene activator protein. *J. Biol. Chem.*, **267**, 14713–14720.
- Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res.*, **32**, D75–D77.
- Helmann, J.D. and Moran, C.P. (2002) RNA polymerase and sigma factors. In Sonenshein, A.L., Hoch, J.A. and Losick, R. (eds), *Bacillus Subtilis and Its Closest Relatives*. ASM Press, Washington, DC.
- Sandelin, A., Alkema, Engström, P., Wasserman, W. and Lenhard, B. (2004) JASPAR: an open access database for eukaryotic transcription factor binding profiles *Nucleic Acids Res.*, **32** (Database issue).
- Colnot, S., Lambert, M., Blin, C., Thomasset, M. and Perret, C. (1995) Identification of DNA sequences that bind retinoid X receptor-1,25(OH)₂D₃-receptor heterodimers with high affinity. *Mol. Cell. Endocrinol.*, **113**, 89–98.
- Jimenez-Lara, A. and Aranda, A. (1999) The vitamin D receptor binds in a transcriptionally inactive form and without a defined polarity on a retinoic acid response element. *FASEB J.*, **13**: 1073–81.
- Papagerakis, P., Hotton, D., Lezot, F., Brookes, S., Bonass, W., Robinson, C., Forest, N. and Berdal, A. (1999) Evidence for regulation of amelogenin gene expression by 1,25-dihydroxyvitamin D(3) *in vivo*. *J. Cell. Biochem.*, **76**, 194–205.
- Toell, A., Polly, P., Carlberg, C. (2000) All natural DR3-type vitamin D response elements show a similar functionality *in vitro*. *Biochem. J.*, **352**, 301–309.
- Takeshita, A., Ozawa, Y. and Chin, W. (2000) Nuclear receptor coactivators facilitate vitamin D receptor homodimer action on direct repeat hormone response elements. *Endocrinology*, **141**, 1281–1284.
- Fujisawa, K., Umehara, K., Kikawa, Y., Shigematsu, Y., Taketo, A., Mayumi, M. and Inuzuka, M. (2000) Identification of a response element for vitamin D₃ and retinoic acid in the promoter region of the human fructose-1,6-bisphosphatase gene. *J. Biochem.*, **127**, 373–382.
- Thummel, K., Brimer, C., Yasuda, K., Thottassery, J., Senn, T., Lin, Y., Ishizuka, H., Kharasch, E., Schuetz, J. and Schuetz, E. (2001) Transcriptional control of intestinal cytochrome P-450 3A by 1 α , 25-dihydroxy vitamin D₃. *Mol Pharmacol.*, **60**, 1399–406.
- Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern Classification*. John Wiley & Sons, Inc., London.
- Bedell, J.A., Korf, I. and Gish, W. (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**, 1040–1041.
- Schneider, T.D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427–441.
- Bi, C.-P., Vyhldal, C.A., Leeder, J.S. and Rogan, P.K. (2004) A minimization entropy-based bipartite algorithm with application to PXR/RXR α binding sites. *Proceedings of Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004)*, San Diego, CA, pp. 453–454.
- Vyhldal, C.A., Rogan, P.K. and Leeder, J.S. (2004) Development and refinement of pregnane X receptor DNA binding site model using information theory: insights into PXR mediated gene regulation. *J. Biol. Chem.*, doi:10.1074/jbc.M408395200.