molecular pharmaceutics

reviews

A Comparative Study on Computational Two-Block Motif Detection: Algorithms and Applications

Chengpeng Bi,* J. Steven Leeder, and Carrie A. Vyhlidal

Bioinformatics and Intelligent Computing, Division of Clinical Pharmacology and Toxicology, Children's Mercy Hospitals and Clinics, 2401 Gillham Road, Kansas City, Missouri 64108, and School of Medicine, University of Missouri−Kansas City, Kansas City, Missouri 64110

Received August 17, 2007; Revised Manuscript Received October 22, 2007; Accepted October 25, 2007

Abstract: Since the completion of human genome sequencing, cataloging of all genomic functional elements has been one of the challenging problems in bioinformatics. Deciphering *cis*-regulatory elements in the human genome still remains elusive although much effort has been expended. This paper reviews a suite of methods for two-block motif discovery including mathematical modeling, *de novo* motif-finding based on multiple local alignment, and genomic sequence scanning method for putative sites. We formulate a general method to address this challenge and compare two major existing algorithms (i.e., greedy local search and Gibbs sampling) implemented to solve the popular two-block structured motif discovery issue. We demonstrate how to use this suite of methods and apply them to human nuclear receptor response elements (i.e., protein binding sites of several relevant nuclear receptors, HNF4α, CAR/RXR, and PXR/RXR).

Keywords: Gibbs sampling; nuclear receptors; two-block motif; structured motif; multiple local alignment; motif discovery; greedy search algorithm; gene regulation

1. Introduction

1.1. Deciphering Functional Elements in the Human Genome. The completion of human genome sequencing[1] demarcates a new genomic era: on one hand, it provides highly accurate DNA sequences for each of the 24 chromosomes, while on the other hand, it generates tremendous challenges to bioinformatics and pharmacogenomics research. Presently, we have an incomplete understanding of the protein-coding portions of the genome and much less understanding of both nonprotein-coding transcripts and genomic elements that temporally and spatially regulate gene expression.[2]

Chromosomal gene expression is regulated by recognition of *cis*-acting regulatory DNA sequences ("the promoter") typically upstream of the transcription start site (TSS) by proteins termed transcription factors (TFs). Each TF can bind to a range of DNA patterns in the promoter regions of genes to either induce or repress transcription of these genes, often by recruiting accessory proteins, which in turn, influence chromatin structure,[3] recruitment of RNA polymerase,[4] and promoter clearance.[5]

* To whom correspondence should be addressed. Mailing address: Bioinformatics and Intelligent Computing Pediatric Research Building, Third Floor, 2401 Gillham Road, Children's Mercy Hospitals and Clinics, Kansas City, MO 64108. Tel: (816) 983-6508. Fax: (816) 983-6515. E-mail: cbi@cmh.edu.

(1) Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409* (6822), 860–921.

(2) Birney, E.; Stamatoyannopoulos, J. A.; Dutta, A.; Guigo, R.; Gingeras, T. R.; Margulies, E. H.; Weng, Z.; Snyder, M.; Dermitzakis, E. T.; Thurman, R. E.; Kuehn, M. S.; Taylor, C. M.; Neph, S.; Koch, C. M.; Asthana, S.; Malhotra, A.; Adzhubei, I.; Greenbaum, J. A.; Andrews, R. M.; Flicek, P.; Boyle, P. J.; Cao, H.; Carter, N. P.; Clelland, G. K.; Davis, S.; et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **2007**, *447* (7146), 799–816.

10.1021/mp7001126 CCC: $40.75
Published on Web 12/13/2007

Deciphering *cis*-regulatory elements in the human genome still remains elusive despite the considerable effort that has been applied to the problem primarily because these elements are typically short (5–20 base pairs (bp) long), degenerate, and obey few rules.[6,7] Furthermore, in eukaryotes, combinatorial control of gene regulation is more complex than in prokaryotes requiring several TFs for the regulation of a single gene.[8] The eukaryotic binding motifs are often shorter and promoters are considerably longer (up to several kilobases in length) than prokaryotic. Eukaryotic upstream regions tend to contain regulatory modules composed of collections of adjacent binding sites for more than one TF. Transcriptional regulation not only relies on the combination of the individual TFs involved, but also on the context in which each recognition site in the regulatory region occurs (e.g., multiple copies of elements for the same TF or recognition sequences for cooperative TFs).

The magnitude of response for a particular target gene presumably depends on the degree to which a particular site is similar to a consensus sequence, the sequence with the highest binding affinity. Variation from the consensus binding site is likely to decrease binding affinity and thus occupancy by the TF, altering the degree of response. The observed sequence variability in a set of TF binding sites is biologically relevant for at least two reasons: (1) different levels are required for each protein expressed in the cell, and (2) it is necessary not only to positively regulate, but also to negatively regulate, gene expression in response to developmental and environmental cues.[9] In addition, nucleotide variation at each position of a motif for the same TF may be heterogeneous, and in such cases, heterogeneity may be a function of sequence constraint due to direct contact with the TF and accumulation of neutral mutations in base positions not directly involved in protein−DNA interactions. Such heterogeneous variation gives rise to a particular class of *cis*-regulatory elements referred to as structured motifs which are adapted to homodimeric or heterodimeric protein−DNA binding interaction.

A comprehensive catalog of encoded structural and functional components in the genome will be critical for understanding human biology and disease.[2,10] The challenges associated with identifying *cis*-regulatory elements in eukaryotes have stimulated the development of computational motif-finding algorithms for their identification. Used in combination with available gene expression data,[11,12] genome sequences for comparative analysis,[13,14] and binding sequences deduced from chromatin immunoprecipitation (ChIP) studies,[15,16] these computational algorithms have the potential to increase our understanding of the human genome.

1.2. Structured Motifs in Nuclear Receptor Binding Sites. A structured motif is a TF binding site that consists of two adjacent DNA recognition elements ("half-sites") separated by a spacer of variable length ("gap"). Among TF binding sites, structured motifs are recognized as important for gene regulation in both prokaryotes and eukaryotes. Examples of structured motifs include sequences recognized by the CRP homodimer[17] and the binding site for the xenobiotic receptor PXR (pregnane X receptor, NR1I2) and its heterodimer partner RXR (9-*cis*-retinoic acid receptor, NR2B1).[18–20] There are several synonymous terms commonly used to describe a structured motif including coopera-

(3) Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **2007**, *447* (7143), 407–12.
(4) Ptashne, M.; Gann, A. Transcriptional activation by recruitment. *Nature* **1997**, *386* (6625), 569–77.
(5) Sandaltzopoulos, R.; Becker, P. B. Heat shock factor increases the reinitiation rate from potentiated chromatin templates. *Mol. Cell Biol.* **1998**, *18* (1), 361–7.
(6) Michelson, A. M. Deciphering genetic regulatory codes: a challenge for functional genomics. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (2), 546–8.
(7) Tompa, M.; Li, N.; Bailey, T. L.; Church, G. M.; De Moor, B.; Eskin, E.; Favorov, A. V.; Frith, M. C.; Fu, Y.; Kent, W. J.; Makeev, V. J.; Mironov, A. A.; Noble, W. S.; Pavesi, G.; Pesole, G.; Regnier, M.; Simonis, N.; Sinha, S.; Thijs, G.; van Helden, J.; Vandenbogaert, M.; Weng, Z.; Workman, C.; Ye, C.; Zhu, Z. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **2005**, *23* (1), 137–44.
(8) Lemon, B.; Tjian, R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* **2000**, *14* (20), 2551–69.
(9) Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **2000**, *16* (1), 16–23.
(10) The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **2004**, *306* (5696), 636–40.
(11) Cho, R. J.; Campbell, M. J.; Winzeler, E. A.; Steinmetz, L.; Conway, A.; Wodicka, L.; Wolfsberg, T. G.; Gabrielian, A. E.; Landsman, D.; Lockhart, D. J.; Davis, R. W. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **1998**, *2* (1), 65–73.
(12) Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95* (25), 14863–8.
(13) Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **2000**, *16* (9), 369–72.
(14) Miller, W.; Makova, K. D.; Nekrutenko, A.; Hardison, R. C. Comparative genomics. *Annu. Rev. Genom. Hum. Genet.* **2004**, *5*, 15–56.
(15) Buck, M. J.; Lieb, J. D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **2004**, *83* (3), 349–60.
(16) Ren, B.; Robert, F.; Wyrick, J. J.; Aparicio, O.; Jennings, E. G.; Simon, I.; Zeitlinger, J.; Schreiber, J.; Hannett, N.; Kanin, E.; Volkert, T. L.; Wilson, C. J.; Bell, S. P.; Young, R. A. Genome-wide location and function of DNA binding proteins. *Science* **2000**, *290* (5500), 2306–9.
(17) de Crombrugghe, B.; Busby, S.; Buc, H. Cyclic AMP receptor protein: role in transcription activation. *Science* **1984**, *224* (4651), 831–8.
(18) Bi, C.-P.; Vyhlidal, C. A.; Leeder, J. S.; Rogan, P. K. A minimization entropy based bipartite algorithm with application to PXR/RXR binding sites. RECOMB 2004, San Diego, 2004, pp 453–454.
(19) Claessens, F.; Gewirth, D. T. DNA recognition by nuclear receptors. *Essays Biochem.* **2004**, *40*, 59–72.
(20) Handschin, C.; Meyer, U. A. Induction of drug metabolism: the role of nuclear receptors. *Pharmacol. Rev.* **2003**, *55* (4), 649–73.

tive binding sites,[21] two-block motif,[22] spaced dyad,[23] structured motif,[24] bipartite pattern,[25] double-box motif,[26] or gapped or structured motif.[27] The terms "two-block" or "bipartite structured motif" (or pattern) will be used interchangeably in this paper. In a two-block structured motif, each half-site is an independent submodel in which the half-sites may assume four possible relative orientations separated by a variable gap distance.[25] The half-sites may be oriented as a direct repeat (DR), reversed direct repeat (RDR), everted repeat (ER), or inverted repeat (IR) as shown in Figure 1B. Here the term "repeat" refers to a half-site motif or submotif, and the two elements do not have to be recurrent or strictly repetitive.

The difficulties inherent in two-block structured motif discovery can be illustrated using the nuclear receptor (NR) constitutive androstane receptor (CAR, NR1I3) as an example. CAR is a xenosensing transcription factor involved in regulating the biotransformation, transport, and elimination of several endogenous and xenobiotic compounds. It binds as a heterodimeric complex with RXR to *cis*-elements upstream of genes[28] to activate or repress gene expression. Figure 1 illustrates CAR/RXR heterodimer binding to a two-block structured motif sequence upstream of a regulated gene. CAR/RXR heterodimers have been shown experimentally to interact with diverse repeat elements including DR4 and DR5 motifs (DR with 4 or 5 bp gaps), ER6 (ER with a 6 bp gap), and IR2 motifs (IR with 2 bp gap).[29–33] Therefore, comprehensive modeling of CAR/RXR binding to DNA must be able to accommodate variability at many levels: DNA sequence variation, half-site orientation, and variability in the nucleotide gap between the two half-sites. Not all existing algorithms take these issues into consideration.

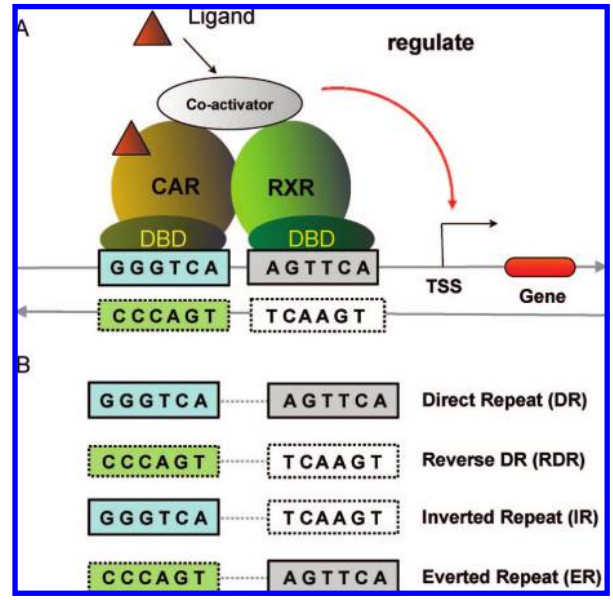


Figure 1. Gene regulation in drug metabolism. (A) The NRs CAR and RXR bind to a two-block structured motif (i.e., response cis element) upstream of the regulated gene. This modulation may either activate or repress the gene expression depending on the cofactor (either coactivator or corepressor). CAR and RXR bind to the submotifs 'GGGTCA' and 'AGTTCA' respectively. (B) Four orientations of a structured motif: direct repeat (DR), reversed direct repeat (RDR), everted repeat (ER), and inverted repeat (IR).

The purpose of this paper is to introduce the concepts and models required for representing structured motifs (section 2), discuss the scoring function of structured motifs for scanning binding sites given a genomic sequence (section

(21) GuhaThakurta, D.; Stormo, G. D. Identifying target sites for cooperatively binding factors. *Bioinformatics* **2001**, *17* (7), 608–21.
(22) Liu, X.; Brutlag, D. L.; Liu, J. S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* **2001**, *127*, 38.
(23) van Helden, J.; Rios, A. F.; Collado-Vides, J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.* **2000**, *28* (8), 1808–18.
(24) Marsan, L.; Sagot, M. F. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comput. Biol.* **2000**, *7* (3–4), 345–62.
(25) Bi, C.; Rogan, P. K. Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res.* **2004**, *32* (17), 4979–91.
(26) Favorov, A. V.; Gelfand, M. S.; Gerasimova, A. V.; Ravcheev, D. A.; Mironov, A. A.; Makeev, V. J. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics* **2005**, *21* (10), 2240–5.
(27) Chakravarty, A.; Carlson, J. M.; Khetani, R. S.; DeZiel, C. E.; Gross, R. H. SPACER: identification of cis-regulatory elements with non-contiguous critical residues. *Bioinformatics* **2007**, *23* (8), 1029–31.
(28) Baes, M.; Gulick, T.; Choi, H. S.; Martinoli, M. G.; Simha, D.; Moore, D. D. A new orphan member of the nuclear hormone receptor superfamily that interacts with a subset of retinoic acid response elements. *Mol. Cell. Biol.* **1994**, *14* (3), 1544–51.
(29) Echchgadda, I.; Song, C. S.; Oh, T.; Ahmed, M.; De La Cruz, I. J.; Chatterjee, B. The xenobiotic-sensing nuclear receptors pregnane X receptor, constitutive androstane receptor, and orphan nuclear receptor hepatocyte nuclear factor 4alpha in the regulation of human steroid-/bile acid-sulfotransferase. *Mol. Endocrinol.* **2007,** *21* (9), 2099–2111.
(30) Ferguson, S. S.; Chen, Y.; LeCluyse, E. L.; Negishi, M.; Goldstein, J. A. Human CYP2C8 is transcriptionally regulated by the nuclear receptors constitutive androstane receptor, pregnane X receptor, glucocorticoid receptor, and hepatic nuclear factor 4alpha. *Mol. Pharmacol.* **2005**, *68* (3), 747–57.
(31) Frank, C.; Gonzalez, M. M.; Oinonen, C.; Dunlop, T. W.; Carlberg, C. Characterization of DNA complexes formed by the nuclear receptor constitutive androstane receptor. *J. Biol. Chem.* **2003**, *278* (44), 43299–310.
(32) Goodwin, B.; Hodgson, E.; D'Costa, D. J.; Robertson, G. R.; Liddle, C. Transcriptional regulation of the human CYP3A4 gene by the constitutive androstane receptor. *Mol. Pharmacol.* **2002**, *62* (2), 359–65.
(33) Sueyoshi, T.; Kawamoto, T.; Zelko, I.; Honkakoski, P.; Negishi, M. The repressed nuclear receptor CAR responds to phenobarbital in activating the human CYP2B6 gene. *J. Biol. Chem.* **1999**, *274* (10), 6043–6.

3), describe two major algorithms for structured motif discovery (section 4), apply these two major algorithms to three NR binding sites and compare their performance (section 5), and finally, conclude with a discussion of issues to be considered when designing *in silico* motif discovery investigations.

2. Modeling Structured Motifs

2.1. Basics of the Motif Discovery Problem. The motif discovery problem can be formulated in several ways. The most common representation involves a set of DNA sequences that are believed, *a priori*, to be coregulated or conserved across species and thus likely to be bound by the same TF or TF complex. The problem is to uncover the motif parameters that could explain this binding activity. Mathematically, the motif-finding problem is often formulated as a maximization problem. A motif-finding algorithm takes as input potential binding sequences and locally aligns them to identify motif sites hidden within these sequences to maximize a specified function such as log-likelihood[34] or information content.[25] The most significant motif sites are output as the best candidate binding sites for the same TF.

2.2. Representing a One-Block Motif. Before choosing a particular algorithm for motif representation in a specific biological question, an investigator may consider the following characteristics: simplicity, interpretability, representational power, or computational convenience. Traditionally, consensus sequences have been used as the most convenient form of motif representation especially in a string search-based algorithm. A consensus sequence simply reflects the preferred (or most frequent) nucleotide at each position of a motif (adenine [A], cytosine [C], guanine [G], or thymine [T]), often employing the IUPAC ambiguity codes (W = A/T, S = C/G, R = A/G, Y = C/T, K = G/T, M = A/C, B = not A, D = not C, H = not G, V = not T, and N is any nucleotide[35]) to indicate degeneracy in the sequence. For example, the aligned motif of the left-half-sites in Figure 2A can be expressed as DBDNHH. A major limitation of this method is immediately evident: the nucleotide frequency information is not represented in the motif pattern. Furthermore, even though some nucleotide types do not appear in a short sequence alignment, it does not necessarily imply that they do not exist; they may be present at a low frequency that may only be detected if aligning a large data set. The major disadvantage of using consensus patterns to search genomic sequence for potential sites is under-representation of true sites.

Recurrent motifs in a set of related DNA sequences are most commonly modeled by sequence patterns or by a position weight matrix (PWM). A PWM is a profile, sometimes referred to as position-specific scoring matrix (PSSM), in which the motif is represented as a matrix of nucleotide scores indexed by letter and position.[9] A PWM is essentially a table representing a motif matrix tabulating the frequencies of each of the four nucleotides at each position across motif positions. Although there is debate over whether the motif positions are dependent or independent, it is often assumed that nucleotide occurrence at each position is independent of all the other positions. A matrix of nucleotide probabilities is equivalent to a product multinomial distribution over the observed nucleotides across motif positions. Sequence logos are frequently used to graphically depict these types of motif matrices.[36]

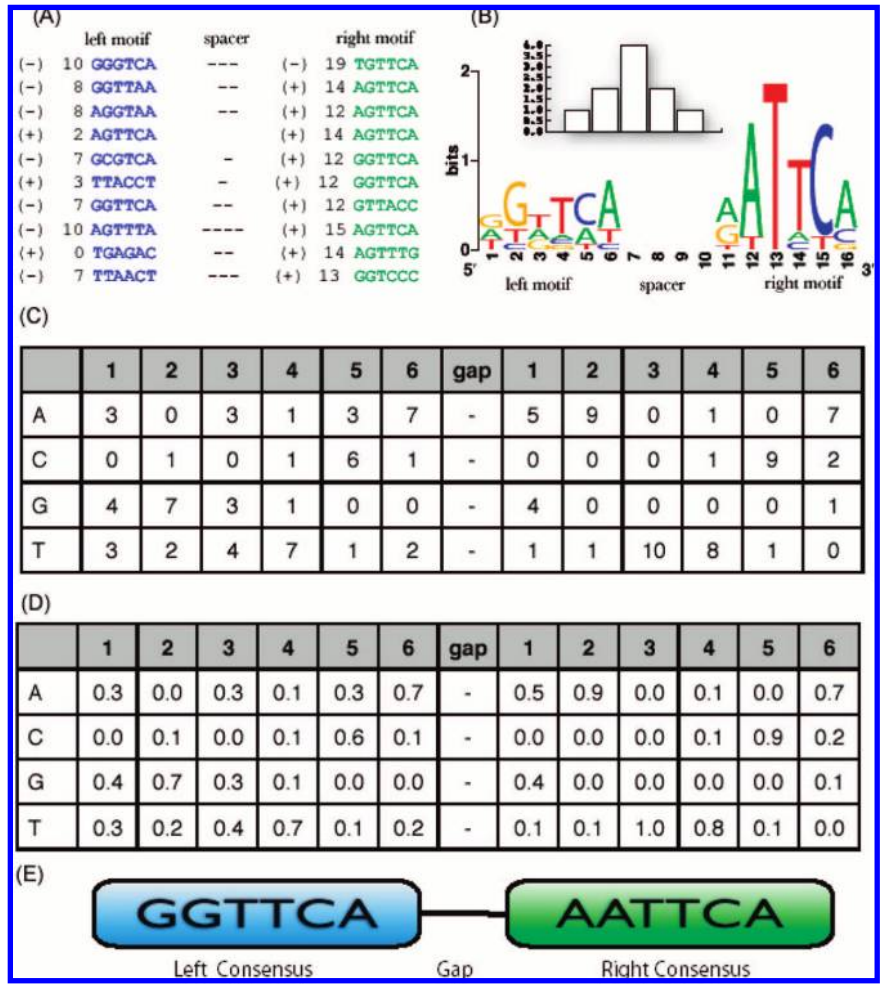
2.3. Representing a Structured Motif. A structured motif discovery algorithm is designed to find two-block binding sites among a set of unaligned genomic sequences from the regulatory regions of coregulated genes or that are enriched in sequences of ChIP-chip binding data or other types of orthologous sequences. Figure 2 illustrates a set of structured motif sites that have been aligned from a small set of verified CAR binding sequences (Figure 2A). The two-block structured motif sequence logo (Figure 2B) can be visualized using the BiLogo plotter program (refer to http://bipad.cmh.edu). To summarize the information presented in the structured alignment in Figure 2A, the nucleotides present in each column are counted; note that each column represents a discrete motif position in the alignment. Next, a profile that tabulates the count of each nucleotide (nucleotide counting matrix, Figure 2C) type in each position of the motif (i.e., column) is created. To determine the counting matrix, let c_{jb} denote the count of nucleotide b at position j, c_j represent the vector holding nucleotide counts on position j, and the matrices $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$ contain all counts on all positions in the left and right motifs, respectively. For example, the number of "G" nucleotides ("G-count") of the second position in the left motif is 7, that is to say, $c_{2G}{}^{(1)} = 7$. The table in Figure 2D is called the motif frequency matrix, which is derived from the occurrence or counting matrix $f_{jb}{}^{(1)} = c_{jb}{}^{(1)}/n$, where n is the total number of motif sequences aligned. Figure 2E manifests a typical structured motif using a consensus sequence (i.e., GGTTCAN$_x$AATTCA) derived from the frequency table. The line connecting the two blocks represents a flexible gap or nonconserved region (x) that varies from $d_{\min}$ to $d_{\max}$.

2.4. Motif Conservation and Information Content. To appreciate the relationship between information content and relative entropy as applied to protein−DNA interactions, several important concepts must be introduced. Given a DNA sequence and a TF that is capable of binding somewhere along that sequence (but is not yet bound to the DNA), there

(34) Lawrence, C. E.; Reilly, A. A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **1990**, *7* (1), 41–51.

(35) Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, (1), 4–8.

(36) Schneider, T. D.; Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **1990**, *18* (20), 6097–100.



	1	2	3	4	5	6	gap	1	2	3	4	5	6
A	3	0	3	1	3	7	-	5	9	0	1	0	7
C	0	1	0	1	6	1	-	0	0	0	1	9	2
G	4	7	3	1	0	0	-	4	0	0	0	0	1
T	3	2	4	7	1	2	-	1	1	10	8	1	0

	1	2	3	4	5	6	gap	1	2	3	4	5	6
A	0.3	0.0	0.3	0.1	0.3	0.7	-	0.5	0.9	0.0	0.1	0.0	0.7
C	0.0	0.1	0.0	0.1	0.6	0.1	-	0.0	0.0	0.0	0.1	0.9	0.2
G	0.4	0.7	0.3	0.1	0.0	0.0	-	0.4	0.0	0.0	0.0	0.0	0.1
T	0.3	0.2	0.4	0.7	0.1	0.2	-	0.1	0.1	1.0	0.8	0.1	0.0

Figure 2. Representation of structured motif sequences. (A) Alignment of structured motif sequences from a subset of verified CAR binding sequences. (B) Two-block structured motif (i.e., bipartite motif) sequence logo plotted using the BiLogo plotter program (refer to http://bipad.cmh.edu). Note that the inset is a graph showing the gap distance distribution. (C) Counting matrix of a structured motif derived from the two-block alignment. (D) Frequency matrix derived from the counting matrix. (E) Typical two-block structured motif.

is uncertainty as to where, exactly, the protein will bind. Once the binding event has occurred, there is less uncertainty as to the DNA sequence to which the protein is bound, and the information content is defined as this decrease in uncertainty.[37] Expressed in a slightly different way, the more tightly a protein binds to a specific DNA sequence, the more certain (less uncertain) we are that the DNA sequence is a true binding site. In analogous manner, unbound protein complexes, or nonspecifically bound complexes, possess energy that is dissipated to the surrounding area upon forming a stable protein−DNA complex at a cognate binding site. Thus, the stronger the binding, the greater the information content.[9,38] As organisms evolved to have TF binding sites, this process involved an increase in information that is now recognized as sequence conservation. The use of information content or relative entropy to reveal the conservation of a binding motif has been previously defined (see the Supporting Information, eqs 1–3).[9,39–42] For example, information content for the two-block ($t = 2$ and both half-sites are 6-mer) structured motif as shown in Figure 2D can be easily computed as 7.92 bits. In this case, the right-half-motif has

(37) Schneider, T. D. Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology* **1994**, *5*, 1–18.
(38) Stormo, G. D.; Fields, D. S. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* **1998**, *23* (3), 109–13.
(39) Frith, M. C.; Hansen, U.; Spouge, J. L.; Weng, Z. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.* **2004**, *32* (1), 189–200.
(40) Kullback, S.; Leibler, R. A. On information and sufficiency. *Ann. Math Statist* **1951**, *22* (1), 79–86.
(41) Lawrence, C. E.; Altschul, S. F.; Boguski, M. S.; Liu, J. S.; Neuwald, A. F.; Wootton, J. C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **1993**, *262* (5131), 208–14.
(42) Sjolander, K.; Karplus, K.; Brown, M.; Hughey, R.; Krogh, A.; Mian, I. S.; Haussler, D. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.* **1996**, *12* (4), 327–45.

more information content (5.32 bits) and thus is more conserved than the left-half-motif (2.60 bits).

3. In Search of Structured Motif Sites on a Genomic Sequence

3.1. Defining Information Weight. Because a motif site is embedded in background sequences, a weight matrix should include the background information in addition to motif conservation. The information weight function was originally defined by Kullback and Leibler in 1951.[40] This function has been applied to express PWM information weight,[43,44] a logarithm ratio of motif and background probability (Supporting Information, eq 4). Note that the expectation of the weight function defines the information content or relative entropy. This function simply indicates how similar a motif nucleotide probability is relative to the background and implies that the contributions of the individual positions of a site to the total binding free energy of the binding protein are consistent with the investigation into the DNA–protein interaction biophysical model.[38,45] If their similarity is 1.0 (i.e., background and motif signals are the same), then the information weight is zero. The larger their difference, the greater is the information weight. A special case of information weight function that has been widely adopted in sequence logo plotting[36] was defined in an early bioinformatics research paper[46] by assuming that the distribution of background information is uniform.

3.2. Scoring Function. To scan a genomic sequence for potential structured motif sites, an information weight matrix must be computed. The aligned frequency matrices in Figure 2C can be readily converted into information weight matrices (Figure 3A). Given a DNA sequence of length L, the scoring function for scanning two-block structured motif is defined by the information weight matrices of two submotifs and the gap penalty function between two adjacent half-site motifs (Supporting Information, eq 5). However, an optimum solution to the issue of gap function remains unresolved. In general, we can assume that the gap distance approximates a normal distribution, with the gap mean and variance being derived from a structured motif alignment (Figure 2A). The probability of a gap of specified size can then be calculated, and higher probability events can be assigned a smaller penalty while lower probability gap sizes are assigned a larger penalty. The penalty should be normalized to avoid overestimation. Most commonly, the gap distance that occurs at the highest frequency is assigned a zero penalty while increasing penalties are assigned as the gap size (d) increases or decreases relative to the highest frequency event. More complex strategies for simulating the gap function, such as hidden Markov model,[47] have been proposed. One approach to defining the gap frequency (f) uses a cosine function as: $f(d,m) = 1.0 + cos(2\pi(d - m)/B)$, where B is a DNA helical repeat (10.4 bp/turn) and m is the most frequent gap. This gap function can be applied to two half-sites separated by short gaps (i.e., $d \leq 10$ bps). The inset in Figure 2B shows the gap distance distribution, which can be converted into a penalty table as shown in Figure 3B. Note that the central gap in this example is 2 bp long (i.e., $m = 2$).

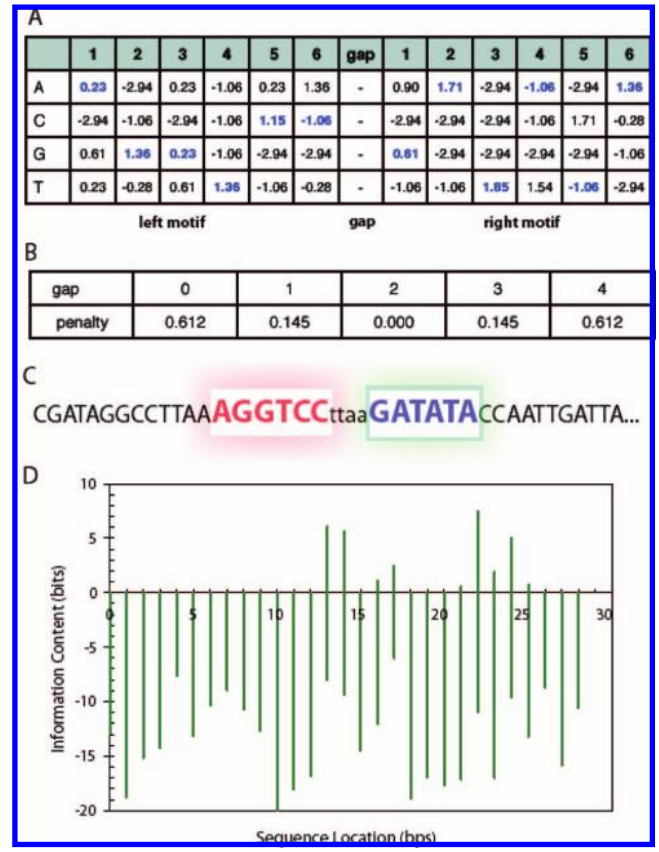


Figure 3. Scanning a genomic sequence for putative structured motif sites: (A) two-block position weight matrices (PWM) converted from Figure 2C using eq 5 (Supporting Information); (B) gap penalty table computed using eq 6 (Supporting Information), the gap range and gap center are determined from the inset in Figure 2B; (C) sample target DNA sequence (40 bp) (note that the highlighted area is a DR-orientated two-block structured motif with 4-bp gap); (D) plot of putative DR motif sites with gap ranging from 0 to 4 bp.

3.3. A Scoring Example. Given two position weight matrices (e.g., the information weight matrices in Figure 3A),

(43) GuhaThakurta, D. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res.* **2006**, *34* (12), 3585–98.
(44) Stormo, G. D.; Hartzell, G. W. 3rd, Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86* (4), 1183–7.
(45) Berg, O. G.; von Hippel, P. H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **1987**, *193* (4), 723–50.
(46) Schneider, T. D.; Stormo, G. D.; Gold, L.; Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **1986**, *188* (3), 415–31.
(47) Sandelin, A.; Wasserman, W. W. Prediction of nuclear hormone receptor response elements. *Mol. Endocrinol.* **2005**, *19* (3), 595–606.

a penalty table (e.g., Figure 3B), and a target DNA sequence (e.g., Figure 3C), the target DNA sequence can then be scanned and scores computed for every potential combination of all two-block structured motif sites. For example, highlighted in Figure 3C is a DR structured motif with a 4-bp gap starting at position 13 (by convention, the first position of the motif is designated as position 0). If the left motif model (i.e., the left motif table in Figure 3A) is applied to the left half-site (AGGTCC), the information score is calculated as $0.23 + 1.36 + 0.23 + 1.36 + 1.15 - 1.06 = 3.27$ (bits). The same calculation can be applied to the other half-site (GATATA) using the right motif table: $0.61 + 1.71 + 1.85 - 1.06 - 1.06 + 1.36 = 3.41$ (bits). The penalty for a 4-bp gap is 0.612 bits as shown in Figure 3B. Thus, the total score for this two-block structured motif site is $3.27 + 3.41 - 0.612 = 6.07$ (bits). Figure 3D is a plot of all potential DR motif sites across the target sequence given a cutoff of -20 bits. Note that for each DR orientation, the gap varies from 0 to 4 bp. The sequence location in Figure 3D is the starting position of the left half-site, and gap information is not included in the plot.

4. Algorithms for *de Novo* Structured Motif Discovery

4.1. Overview of Structured Motif-Finding Algorithms. Although over 50 different algorithms have been developed to perform *de novo* motif discovery,[43,48–50] most motif-finding methods assume a contiguous motif and thus do not explore the properties of a discontinuous structured motif. As a consequence, these strategies are unable to correctly align structured motifs with variable numbers of intervening nucleotides. Nevertheless, several structured motif finders recently have been developed to specifically deal with the characteristics of these structured motifs. The algorithms designed to tackle this problem can be divided into two categories: enumerative methods and alignment- or PWM-based methods.

The first class of structured motif discovery algorithms is based on word enumeration. Enumerative methods typically involve exhaustive enumeration of words up to some maximum size in a data set and are thus best suited to consensus sequence motif models. Once the words are cataloged, they can be scored using an appropriate measure of statistical significance, and the most statistically significant motifs are then reported. Many enumerative methods use tradeoffs on the alphabet size and the number of allowable errors to make these searches computationally feasible.[23,24,51]

Examples of enumerative methods include several programs such as MITRA,[52] RSAT,[23] and YMF.[53] MITRA is based on a k-mismatch flexible motif model whereas both RSAT (dyad-analysis) and YMF exhaustively enumerate motifs using a consensus motif model. They are also enhanced to specifically search for gapped motifs.[23,53] However, neither program is able to identify weak base preferences in the spacer region, motifs of arbitrary length, or full degeneracy in the binding regions at the ends. More recently, another word-counting program named SPACER demonstrated its effectiveness in identifying "gapped" and highly degenerate motifs[27] by treating the "gap" as a region of low conservation. In general, the word-enumeration methods capture all the significant gapped words, but they do not consider gap function. Furthermore, no enumerative method treats the gapped motif in all possible orientations, and they suffer from low specificity for predicting new sites.

Based on these considerations, we will focus on the second class of structured motif discovery algorithms based on a PWM motif model including BiPad,[25,54] BioProspector,[22] CoBind,[21] and most recently SeSiMCMC.[26] BiPad is a deterministic algorithm via greedy local search and it can be viewed as a special version of EM (expectation maximization) motif-finding algorithm. Both SeSiMCMC and BioProspector are Gibbs sampling algorithms.

4.2. Technicalities of *de Novo* Motif Discovery. The PWM-based motif discovery algorithms gained their popularity due to their expeditious search capability and powerful motif representation. Basically, they assume that the motif positions are unobserved (i.e., *de novo*) but hidden somewhere in the sequence data set. Two well-known discovery methods use EM[55] and Gibbs sampling[56] algorithms in combination with the multiple local alignment (MLA) methods.[57] A basic introduction to the three technical terms (i.e., MLA, EM, and Gibbs) is described in the following sections.

4.2.1. Multiple Local Alignment. The multiple local alignment methods take on a wide variety of forms, but often

(48) Bulyk, M. L. Computational prediction of transcription-factor binding site locations. *Genome Biol.* **2003**, *5* (1), 201.
(49) MacIsaac, K. D.; Fraenkel, E. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput. Biol.* **2006**, *2* (4), e36.
(50) Wasserman, W. W.; Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **2004**, *5* (4), 276–87.
(51) Li, H.; Rhodius, V.; Gross, C.; Siggia, E. D. Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (18), 11772–7.
(52) Eskin, E.; Pevzner, P. A. Finding composite regulatory patterns in DNA sequences. *Bioinformatics* **2002**, *18* (Suppl 1), S354–63.
(53) Sinha, S.; Tompa, M. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* **2002**, *30* (24), 5549–60.
(54) Bi, C.; Rogan, P. K. BIPAD: a web server for modeling bipartite sequence elements. *BMC Bioinformatics* **2006**, *7*, 76.
(55) Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B* **1977**, *39*, 1–38.
(56) Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 721–741.

involve development of a probabilistic model of the observed sequence data and maximization to find motifs common to all input sequences. A formal statistical model for the PWM-based methods was originally described by Lawrence and Reilly.[34] In this model, the sequence data are represented as an array **S**, where S_{ij} is the residue symbol (i.e., a nucleotide or amino acid) at position j of sequence i. The N input sequences can be represented by N random variables $\{S_1, \ldots, S_i, \ldots, S_N\}$. Each residue can take a symbol from its alphabet K, for example, K = {A, C, G, T} for nucleic acids. Within **S**, one can postulate that there are substrings or words of length w (or w-motif) that are sites describing an unknown motif model. The locations of these sites are unobserved or missing, so a missing array of indicators **A** is introduced, where A_{ij} is either one or zero, indicating whether or not position j in sequence i is the starting point of a motif site. **A** is a particular realization of multiple local sequence alignment, giving us a subset of **S**, denoted as **S**(**A**), which consists only of the residues in the motif segments, and the complementary subset **S**($\mathbf{A}^c$), contains the remaining background sequences. For a moderate number of sequences, the background sequences are often treated as if they were fixed to reduce the computing time. Given a local alignment (**A**), one can compute a motif model and its background distribution by simply counting the residues in each column (as presented in Figure 2).

A common scenario for PWM-based motif-finding algorithms is to initialize an alignment (**A**) and then progressively update the PWM matrix until convergence is achieved. The generative model describes a family of parametrized probability distributions, and the motif is simply a parameter of this distribution. Any number of optimization techniques may be used to search for the parameter setting that maximizes the log-likelihood of the observed sequence data.[49,58] Two frequently used techniques to perform this search are the EM algorithm and Gibbs sampling.[34,41,55]

4.2.2. Expectation Maximization. The EM algorithm is a general approach for maximizing a log-likelihood function given an incomplete data set[55] composed of observed and unobserved data. In the case of alignment-based motif discovery applications, the observed data are the coregulated DNA sequences and the unobserved data are the motif locations hidden in the set of input sequences. In principle, EM algorithms consist of two steps:[55] the expectation and maximization steps. In the expectation or E step, the expected log-likelihood being observed in each input sequence is calculated based on the current setting of the parameters (randomly generated for the first iteration). In the maximization or M step, the current parameters are recalculated and updated to maximize the expected-likelihood function. EM is a local optimization procedure that has been proven to progressively (monotonically) improve the expected likelihood, however, such maximization is also dependent on its initialization seed and thus the EM algorithm is subject to getting trapped in a local maximum. For this reason, it is recommended that motif discovery programs using EM algorithms run with multiple initialization seeds, each of which is randomly generated to improve the probability of converging to the near-optimum. Multiple runs also improve the possibility of finding biologically relevant motifs that may not necessarily correspond to the maximum log-likelihood. Motif-finding algorithms often output a certain number of solutions or alignments for biological justification.

The first EM motif-finding algorithm, developed by Lawrence and Reilly,[34] has given rise to other, diverse EM variations.[58] Enhanced strategies to improve the quality of the alignments and to minimize the risk of local maxima have been proposed as alternatives to random number generation for selecting reasonable initialization points. For example, the MEME algorithm[59,60] and random projection method[61] are based on prior knowledge of word statistic on the given sequence set. The BiPad motif-finding algorithm[18,25] is a greedy algorithm via local search. It can be viewed in the light of a special version of regular EM algorithms in that BiPad maximizes the information content which is equivalent to maximum log-likelihood used in EM algorithms.[62] Note that both MEME and BiPad are deterministic in search of the alignment space. Therefore, each algorithm will come up with the same motif output when it is initialized with the same seed. However, MEME and BiPad may produce different motif outputs if given the same starting seed.

4.2.3. Gibbs Sampling. Gibbs sampling is one of the most popular techniques for Markov Chain Monte Carlo simulation to perform parameter estimation.[56] It is a special case of Metropolis−Hastings algorithms[58,63–65] and was one of the earliest motif-finding algorithms.[41] However, it is different from the EM motif-finding algorithms in that it is a stochastic and global search over the alignment space and is therefore less likely to be trapped in a local maximum.[64] In the context of motif discovery, Gibbs sampling involves drawing random samples of the location of the binding motif

(57) Durbin, R.; Eddy, S. R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*; Cambridge University Press: Cambridge, 1999.
(58) Bi, C. SEAM: a Stochastic EM-type Algorithm for Motif-finding in biopolymer sequences. *J. Bioinform. Comput. Biol.* **2007**, *5* (1), 47–77.
(59) Bailey, T. L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1994**, *2*, 28–36.
(60) Bailey, T. L.; Elkan, C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* **1995**, *21*, 51–80.
(61) Buhler, J.; Tompa, M. Finding motifs using random projections. *J. Comput. Biol.* **2002**, *9* (2), 225–42.
(62) Bailey, T. L. *Likelihood vs information in aligning biopolymer sequences* **1993**, CS93–318.
(63) Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57* (1), 97–109.
(64) Liu, J. S., *Monte Carlo Stragegies for Scientific Computing*; Springer-Verlag: New York, 2001.
(65) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21* (6), 1087–1092.

from a conditional probabilistic distribution. The parameters are re-estimated based on the randomly drawn samples, and then the procedure is performed iteratively. Gibbs sampling is computationally intensive, and it is recommended that the algorithm be run with multiple chains to adequately represent the multimodal log-likelihood surfaces typically encountered in motif discovery.[49,58]

The original implementation of Gibbs sampling was done in the site sampling mode, which assumes that there is exactly one motif element (notably a TF binding site) located in each input sequence. Since its initial introduction, a few improvements have been reported in the literature including: (i) motif sampling allowing zero or multiple motif elements to be located in each sequence;[66] (ii) incorporation of a higher-order Markov background model;[22,67] and (iii) incorporation of phylogeny information,[68] among others.

Although the above motif discovery algorithms focus on one-block motif, their methodology can be easily extended to perform structured motif search. In the following sections we apply these search strategies to solve the structured motif discovery problem by introducing two popular methods: a greedy search-based algorithm, BiPad, and a Gibbs sampling algorithm, BioProspector. These two algorithms perform not only two-block structured motif discovery but can also accommodate one-block motif searching as a special application.

4.3. A Greedy Search Algorithm for Bipartite Motif Discovery. The conventional motif-finding algorithms encounter difficulty in aligning structured motif sequences because they do not allow gapped or structured motifs. BiPad is the first greedy algorithm to solve the two-block or bipartite structured motif problem given a set of unaligned DNA sequences and gap range according to multiple sequence local alignment.[18] In essence, a greedy algorithm (also known as a deterministic method) attempts to make the best decision based on as much information as is immediately available. It is initiated from a random starting point and then progresses to search for a local optimum in each step until convergence is achieved.[58] BiPad performs a structured local alignment, or so-called bipartite alignment, considering not only a two-block motif structure and flexible gap range but also simultaneously aligning the sequences in four possible orientations. The BiPad algorithm has at least three new features that simulate these biological findings: (1) aligning bipartite motif with four orientations (Figure 1B), (2) maximizing the total information content of two submotif models as an integral function, and (3) considering a gap penalty function. The goal is to find a two-block structured motif alignment with the highest information content, which is equivalent to maximizing the log-likelihood function.[62] Implicit in this process is calculation of the entropy function that indicates the degree of chaos or randomness at each position of each individual motif; the higher the entropy, the lower the conservation (or information content). As a result, the objective function can be reduced to minimize the total Shannon's entropy (Supporting Information, eq 6). However, because the total entropy depends on the aligned motif locations, bipartite-structured multiple local alignment is performed to get the best alignment that minimizes the total entropy. Like most motif alignment algorithms, BiPad requires input of motif widths. Details of the algorithm have been described previously.[18,25] However, an automatic width selection procedure has also been proposed and implemented[54] (for details, see http://bipad.cmh.edu).

Like EM algorithms, BiPad is a greedy search algorithm, and it is initiated from random starting points; therefore, multiple iterations with different initialization points are strongly recommended to find the best possible solution. BiPad may generate different models depending on the motif widths and gap range specified (based on biological observations). BiPad is particularly well-suited to building several alternative models that can then be tested experimentally. The program assumes a uniform background distribution; although this treatment simplifies the models and reduces computational complexity, difficulty may be encountered where the multiple sequence alignment is ambiguous. In addition, BiPad is easily implemented in parallel computation especially in aligning large data sets in that each round of large-scale alignment can be sent to a different node and the collected results merged.

4.4. Basic Version of Gibbs Sampling Motif Algorithms. The basic version of Gibbs sampling algorithms searches for a PWM characterizing the best conserved pattern of fixed length w without a gap. It is assumed that the pattern occurs in each sequence included in the analysis. The algorithm is carried out in iterations. The result of each iteration is a set of subsequences, one subsequence from each sequence in the input data set. This set of subsequences represents the occurrences of the pattern of interest, and a PWM characterizing the pattern can be computed. The residual sequences after extraction of the motif constitute the background sequences. A multinomial distribution is often used to characterize the background model; however, higher order Markov models have also been employed and generally provide some improvement.[22,67]

Gibbs sampler is iteratively applied in two steps, the sampling (S) and update (U) steps, until a convergence (equilibrium). The S step draws a sample according to the Gibbs target function, and it is sequentially or randomly applied to each sequence followed by an update step (Supporting Information, eq 7). Note that the Gibbs target function is defined as the ratio of two distributions (motif and background). A higher ratio means that the motif model is different from the background. The Gibbs sampler

(66) Neuwald, A. F.; Liu, J. S.; Lawrence, C. E. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* **1995**, *4* (8), 1618–32.

(67) Thijs, G.; Lescot, M.; Marchal, K.; Rombauts, S.; De Moor, B.; Rouze, P.; Moreau, Y. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **2001**, *17* (12), 1113–22.

(68) Siddharthan, R.; Siggia, E. D.; van Nimwegen, E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.* **2005**, *1* (7), e67.

essentially aims to maximize the objective function in a stochastic way that can avoid the local optimum problem in EM motif algorithms. In fact, the Gibbs target function is very similar to the BiPad objective function except for the background assumption. A recently published Gibbs sampling algorithm also solves a two-box structured motif problem,[26] but its output model is essentially a one-block motif and is therefore not considered further.

4.5. BioProspector. BioProspector[22] is an extension of the one-block Gibbs motif sampler[41] that models gapped motifs in the form of either DR or palindromic patterns. It uses zero- to third-order Markov models to simulate the background distribution, and its target function is a two-block extension to the one-block Gibbs sampler. Both BiPad and BioProspector are subject to the gap constraint, and in this paper we use zero-order background for both of them. If a palindromic pattern is assumed, BioProspector merges the two-block matrices into a single new PWM model, essentially treating the two-block motif as a one-block model. BioProspector performs a Gibbs sampling procedure as described above except that the target function is described by two-block extension to the general equation for the Gibbs motif sampler rather than the Gibbs motif sampler itself (Supporting Information, eq 8). Furthermore, BioProspector adopts a so-called threshold sampling method to expedite convergence.[22] High (T_H) and low (T_L) thresholds are determined internally by the BioProspector program to scan a sequence using the two-block extension equation.[22] All segments with scores higher than T_H are automatically added to the alignment; for the segments with scores between T_L and T_H, they are drawn according to their probability distribution. The BioProspector web server is available at http://robotics.stanford.edu/~xsliu/BioProspector/.

5. Applications for Detection of Hormone Response Elements

5.1. Motif Discovery Using BiPad. In this section, we use the BiPad program to discover the two-block binding sites of three nuclear receptors.

5.1.1. HNF4α. Initially identified as a TF required for liver-specific gene expression, HNF4α is also highly expressed in kidney, intestine, and pancreas, but at only low levels in the testis.[69] HNF4α binds as a homodimer to DRs with 1 or 2 bp gaps. The structured motif model in Figure 4A is derived from 63 validated binding sequences from multiple genes and species.[70,71] BiPad set the structured motif search pattern as 6-[0,2]-6, and one cycle was needed to maximize the two-block structured alignment. Assuming that

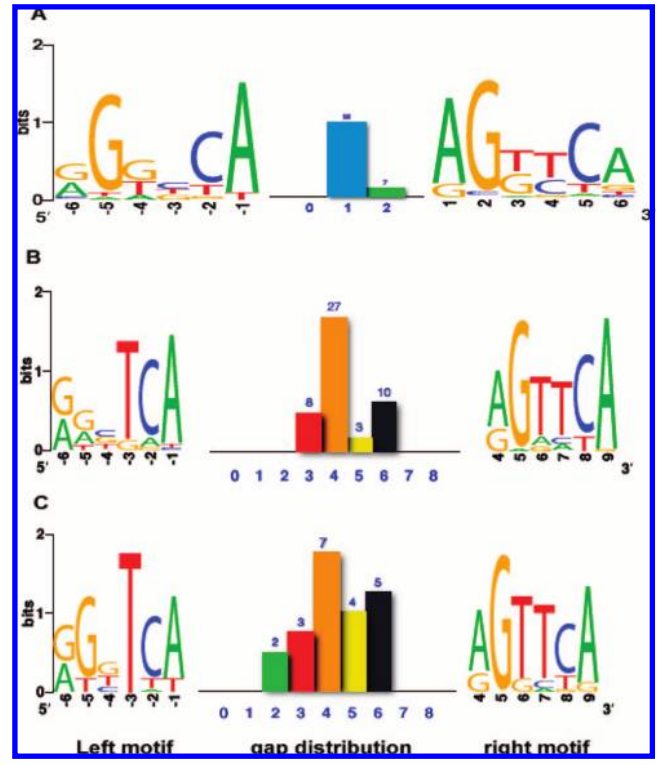


Figure 4. Structured motif detection in protein binding sites of three TFs (homo- or heterodimers) involved in drug metabolism: (A) structured motif logo of HNF4α homodimer binding sites; (B) structured motif logo of PXR/RXR heterodimer binding sites; (C) structured motif logo of CAR/RXR heterodimer binding sites.

all orientations could be bound, nearly all of the sites identified by BiPad were DR. The total information content of the HNF4α binding site is 11.2 bits on average including 5.3 bits for the left and 5.9 bits for the right half-sites.

5.1.2. PXR. PXR; NR1I2, a key regulator of genes encoding several major cytochrome P450 (CYP) enzymes and transporter proteins,[20,72] is activated by a structurally diverse collection of xenobiotics, including both prescription drugs (e.g., macrocyclic antibiotics, antimycotics, glucocrticoids) and compounds derived from plants. PXR was originally shown to regulate the expression of CYP3A isozymes by binding as a heterodimer with RXR to xenobiotic response elements located in the regulatory regions of these genes.[73–78]

Figure 4B shows the PXR/RXR-structured motif model that was built from 48 published and experimentally verified binding sites from the human genome as previously de-

(69) Giguere, V. Orphan nuclear receptors: from gene to function. *Endocr. Rev.* **1999**, *20* (5), 689–725.
(70) Ellrott, K.; Yang, C.; Sladek, F. M.; Jiang, T. Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics* **2002**, *18* (2), S100–9.
(71) Sladek, F. M.; Seidel, S. D., Hepatocyte nuclear factor 4 alpha. *Nuclear receptors and genetic disease*; Academic Press: New York, 2001.
(72) Goodwin, B.; Redinbo, M. R.; Kliewer, S. A. Regulation of cyp3a gene transcription by the pregnane x receptor. *Annu. Rev. Pharmacol. Toxicol.* **2002**, *42*, 1–23.
(73) Bertilsson, G.; Berkenstam, A.; Blomquist, P. Functionally conserved xenobiotic responsive enhancer in cytochrome P450 3A7. *Biochem. Biophys. Res. Commun.* **2001**, *280* (1), 139–44.

scribed.[79] BiPad set the structured motif search pattern as 6-[3,6]-6, and it required 129 runs to find the best bipartite alignment. The PXR/RXR two-block motif can be expressed in DR-orientation consensus as AGTTCA$(N)_{3-6}$AATTCA. The total information content is 13.9 bits including 7.9 bits for the right and 6.0 bits for the left half-site. The dominant gap spacer is 4 bp long (gap distance distribution is shown in between two sublogos in Figure 4B).

5.1.3. CAR. The bipartite algorithm is well suited for modeling the CAR/RXR recognition element, as the heterodimer is known to recognize DR, RDR, IR, and ER patterns.[29–33] The model presented in Figure 4C was built using sequences of 21 human CAR/RXR binding sites extracted from published sources.[20] Alignment of half-sites on both strands was permitted, consistent with published binding studies indicating that all possible orientations should be considered. BiPad set the structured motif search pattern as 6-[2,6]-6 and it required 62 runs to find the best bipartite alignment. The half-site and gap range lengths were set to 6-[0,8]-6; a single cycle was needed to find the best alignment (Figure 4C). The total information content is 14.2 bits including 6.6 bits for the left and 7.6 bits for the right half-site. The discovered patterns are consistent with the experimentally verified sites (RKKTCA$(N)_{2-6}$RKKTCA).[20,31]

5.2. Comparison of BiPad and BioProspector Algorithms. A performance comparison of BiPad and BioProspector was carried out for all three binding-site motifs. Similar results were obtained for HNF4α and the PXR/RXR binding-site motifs using either BiPad or BioProspector (data not shown). In contrast, motif discovery of the CAR/RXR binding site was different for BiPad and BioProspector. Analysis of the same set of CAR/RXR binding sites with BioProspector produced a similar alignment of the right-side motif compared with BiPad but displayed a lower overall conservation (12.1 bits) (Figure 5B.). The very low information content in the left half-site in the CAR/RXR motif detected with BioProspector reflects the fact that BioProspector does not detect all four possible orientations of the two half-sites, a significant limitation for this TF. Analysis of the individual binding sites for CAR/RXR heterodimers further demonstrate the superior performance of BiPad over BioProspector. With BiPad, only two sites with phase shifts were observed (Figure 5A, sequences 10 and 21, cyan underlined elements) versus 11 sites with phase shifts observed using BioProspector (Figure 5A, red underlined elements). BioProspector consistently detected only one half-site and was unable to detect either half-site of the IR2 (Figure 5A, sequence 20). It seems that BioProspector is biased to align one half-site with very high information content without balancing the other half-site as part of a structured motif (refer to the two-block sequence logo predicted by BioProspector in Figure 5B). Collectively, these data suggest that for complex heterodimers, such as CAR/RXR, the BiPad motif discovery algorithm outperforms BioProspector.

Previous studies show that the performance of all motif-finder algorithms decreases as the sequence length increases.[54] To make a fair comparison of BiPad and BioProspector, we computationally embedded 35 known DR binding sites within 35 simulated promoter sequences (one site per sequence). The sequence length ranged from 100 to 1000 bp with a step size of 100 bp. Note that the motif subtlety increases as the sequence becomes longer. If at least one base overlap between the known site and a predicted site occurs, a hit is counted. The simulation results show that in the length range 100–200 bp, both algorithms can detect all sites. In the range of 300–600 bp, BiPad performance varied from 83.0 to 94.0% whereas BioProspector performance dramatically dropped from 60.0% at 300 bp to 6.0% at 600 bp. BioProspector failed to detect the binding motif when the input sequences were longer than 600 bp. However, even in the 1000-bp sequence data set, BiPad still exhibited high performance (77.0%). This simulation implies that the greedy algorithm BiPad is more powerful in subtle motif discovery than the stochastic counterpart Gibbs sampler.

5.3. Genome-Wide Scanning. The CAR/RXR-structured motif model (Figure 4C) can be used to scan defined regions upstream of genes of interest to identify potential two-block motifs. This process has been applied to 10 kb of regulatory sequences upstream of the *CYP3A4* gene. The structured motif scanner was implemented using eqs (5) and (6). Figure 6A shows the *CYP3A4* gene location on chromosome 7 and its 10 kb upstream region (Figure 6B) that has been scanned using a two-block structured motif scanner. Figure 6C shows the plot of information content vs genomic location with negative values for information content representing motifs present on the antisense strand. Around −7.8 kb there is a

(74) Bertilsson, G.; Heidrich, J.; Svensson, K.; Asman, M.; Jendeberg, L.; Sydow-Backman, M.; Ohlsson, R.; Postlind, H.; Blomquist, P.; Berkenstam, A. Identification of a human nuclear receptor defines a new signaling pathway for CYP3A induction. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95* (21), 12208–13.

(75) Blumberg, B.; Sabbagh, W., Jr.; Juguilon, H.; Bolado, J., Jr.; van Meter, C. M.; Ong, E. S.; Evans, R. M. SXR, a novel steroid and xenobiotic-sensing nuclear receptor. *Genes Dev.* **1998**, *12* (20), 3195–205.

(76) Goodwin, B.; Hodgson, E.; Liddle, C. The orphan human pregnane X receptor mediates the transcriptional activation of CYP3A4 by rifampicin through a distal enhancer module. *Mol. Pharmacol.* **1999**, *56* (6), 1329–39.

(77) Lehmann, J. M.; McKee, D. D.; Watson, M. A.; Willson, T. M.; Moore, J. T.; Kliewer, S. A. The human orphan nuclear receptor PXR is activated by compounds that regulate CYP3A4 gene expression and cause drug interactions. *J. Clin. Invest.* **1998**, *102* (5), 1016–23.

(78) Pascussi, J. M.; Jounaidi, Y.; Drocourt, L.; Domergue, J.; Balabaud, C.; Maurel, P.; Vilarem, M. J. Evidence for the presence of a functional pregnane X receptor response element in the CYP3A7 promoter gene. *Biochem. Biophys. Res. Commun.* **1999**, *260* (2), 377–81.

(79) Vyhlidal, C. A.; Rogan, P. K.; Leeder, J. S. Development and refinement of pregnane X receptor (PXR) DNA binding site model using information theory: insights into PXR-mediated gene regulation. *J. Biol. Chem.* **2004**, *279* (45), 46779–86.

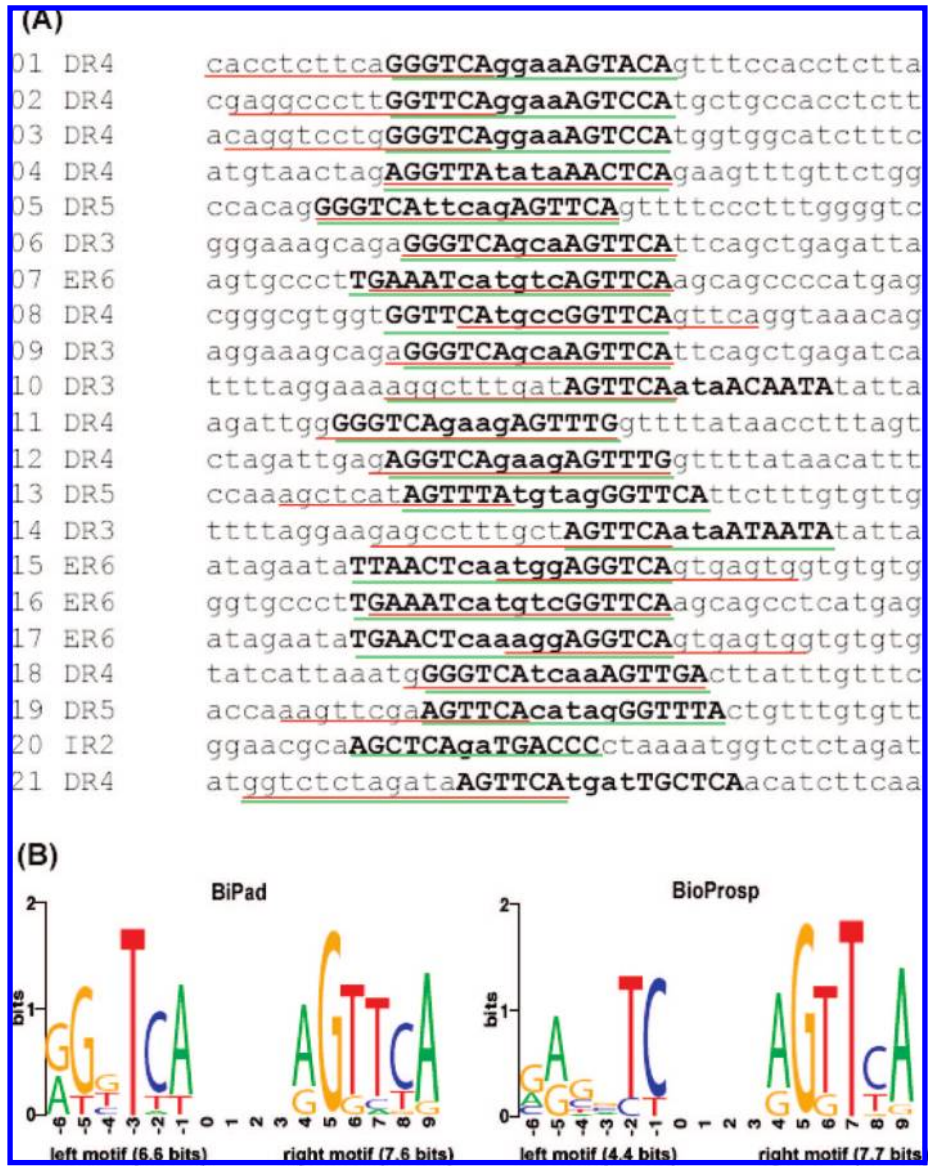


Figure 5. Comparison of BiPad and BioProspector in detecting the CAR/RXR binding site motif. Subsequences in bold are experimentally verified CAR/RXR binding sites. (A) Motif sites detected by BiPad (cyan) and BioProspector (red) in known CAR/RXR binding sites. Note that BioProspector was unable to detect either half-site in sequence 20, an IR2. (B) Two-block structured motif logos derived from the alignments by BiPad and BioProspector. The motif detected by BiPad has higher information content whereas information content detected by BioProspector is biased toward one half-site.

cluster of sites with highly dense information. This is an indicator of potential binding sites because real sites are frequently clustered.[80] For example, highlighted in Figure 6D are a DR3 (13.9 bits) and ER6 (12.8 bits) within the previously verified distal xenobiotic responsive enhancer module (XREM)[76] and the proximal PXRE (ER6, 14.9 bits) at −272.[74,75,77]

6. Discussion

Since the completion of sequencing of genomes from multiple species, most efforts have focused on increased understanding of genomic regions that code for functional products (RNA, protein). However, a major challenge in predicting phenotype from genomic sequence data lies in the understanding of the functional contributions of noncoding regions of the genome, including gene promoter and regulatory elements. Sophisticated tools and rich new data sources allow for greater success than ever before in identifying gene promoter sequences and TF binding sites. Recently developed bioinformatic techniques assist in determining the context-specific effects of sequence motifs on gene expression and offer the possibility of accurately associating regulatory proteins with their cognate binding sites. These advances open up the potential for building accurate mechanistic models of gene regulation. However,

(80) Zhou, Q.; Wong, W. H. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (33), 12114–9.

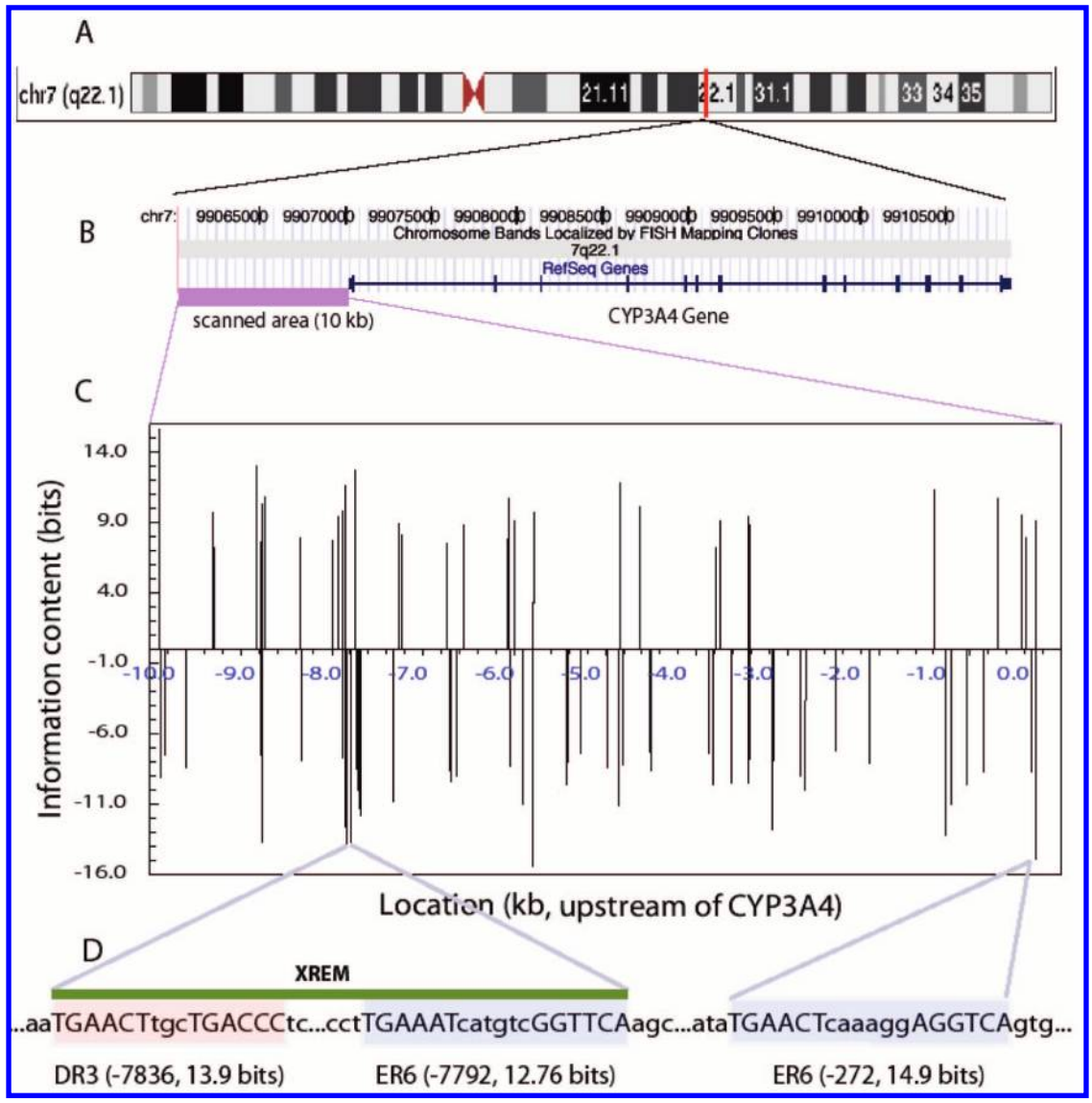


Figure 6. Scanning upstream of the *CYP3A4* gene for putative structured motif binding sites: (A) chromosome 7 showing the *CYP3A4* gene position; (B) sequence region showing the *CYP3A4* gene and its upstream 10 kb that are scanned using a two-block structured motif scanner; (C) information content is plotted versus genomic location in the *CYP3A4* gene promoter (note that the chromosome coordinates in A and B are from the Human March 2006 (hg18) assembly and that negative information in C indicates that the first-block motif is detected on the reverse strand); (D) part of the 10 kb upstream sequence highlighting two regions: the left region is the XREM composed of a DR3 (13.9 bits) and an ER6 (12.8 bits) structured motifs, and the right highlighted region is the proximal PXRE (ER6) that have been experimentally characterized.

effective implementation of motif-finding tools and scanning programs requires an understanding of the strengths, weaknesses, and potential limitations associated with each individual algorithm.

A priori knowledge of the TF of interest should help in selecting an appropriate motif discovery algorithm. For example, algorithms particularly suited to identifying single-block motifs may perform as well as more complex algorithms for TFs that function as monomers or homodimers. However, in reality, a single strategy may not identify all possible motifs in the situations most likely to be encountered by investigators. Therefore, taking advantage of multiple motif discovery tools leverages the strengths of different algorithms and can greatly improve results.[7,49] Furthermore, consideration of additional biological information may help overcome this obstacle and thus decrease the detection of false positives.[81,82] Several ways to take advantage of additional biological information include: (1)

(81) Bae, S. H.; Tang, H.; Wu, J.; Xie, J.; Kim, S. dPattern: transcription factor binding site (TFBS) discovery in human genome using a discriminative pattern analysis. *Bioinformatics* **2007,** *23* (19), 2619–21.

(82) Tsukahara, T.; Kim, S.; Taylor, M. W. REFINEMENT: a search framework for the identification of interferon-responsive elements in DNA sequences--a case study with ISRE and GAS. *Comput. Biol. Chem.* **2006**, *30* (2), 134–47.

analyzing coregulated genes from microarray data; (2) comparative genomics by aligning genomes across species;[83–85] (3) identification of clustered binding sites in gene regulatory regions;[80] and (4) searching for multiple motifs for TFs that are recruited together to cooperatively regulate gene expression.[80,86]

In conclusion, advances in computational motif detection algorithms will facilitate the genome-wide mapping of *cis*-acting regulatory elements for heteromeric complexes that bind to variably spaced two-block motifs. However, as illustrated by the CAR/RXR example, careful consideration of the strengths and weaknesses of the algorithms available for motif discovery and sequence scanning are crucial for proper data interpretation and to appropriately design studies to test the hypotheses generated.

Acknowledgment. This research is supported in part by the Katharine B. Richardson Foundation.

Supporting Information Available: Additional methodologic details. This material is available free of charge via the Internet at http://pubs.acs.org.

MP7001126

(83) Fang, F.; Blanchette, M. FootPrinter3: phylogenetic footprinting in partially alignable sequences. *Nucleic Acids Res.* **2006**, *34* (Web Server issue), W617–20.

(84) Li, X.; Wong, W. H. Sampling motifs on phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (27), 9481–6.

(85) Li, X.; Zhong, S.; Wong, W. H. Reliable prediction of transcription factor binding sites by phylogenetic verification. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (47), 16945–50.

(86) Gupta, M.; Liu, J. S. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (20), 7079–84.