Pacific Symposium on Biocomputing 6:127-138 (2001)

BIOPROSPECTOR: DISCOVERING CONSERVED DNA MOTIFS IN UPSTREAM REGULATORY REGIONS OF CO-EXPRESSED GENES

X. LIU
Stanford Medical Informatics, 251 Campus Dr. X215, Stanford University
Stanford CA 94305-5479 (xliu@smi.stanford.edu)

D.L. BRUTLAG
Department of Biochemistry, Beckman Center B400, Stanford University
Stanford CA 94305-5307 (brutlag@stanford.edu)

J.S. LIU
Department of Statistics, Science Center 610, Harvard University
Cambridge MA 02138 (jliu@stat.stanford.edu)

The development of genome sequencing and DNA microarray analysis of gene expression gives rise to the demand for data-mining tools. BioProspector, a C program using a Gibbs sampling strategy, examines the upstream region of genes in the same gene expression pattern group and looks for regulatory sequence motifs. BioProspector uses zero to third-order Markov background models whose parameters are either given by the user or estimated from a specified sequence file. The significance of each motif found is judged based on a motif score distribution estimated by a Monte Carlo method. In addition, BioProspector modifies the motif model used in the earlier Gibbs samplers to allow for the modeling of gapped motifs and motifs with palindromic patterns. All these modifications greatly improve the performance of the program. Although testing and development are still in progress, the program has shown preliminary success in finding the binding motifs for *Saccharomyces cerevisiae* RAP1, *Bacillus subtilis* RNA polymerase, and *Escherichia coli* CRP. We are currently working on combining BioProspector with a clustering program to explore gene expression networks and regulatory mechanisms. For a copy of the program and documentation for UNIX systems, please contact *xliu@smi.stanford.edu.*

1 Introduction

Over the last ten years, genomic sequencing has started in over 600 organisms, and more than 50 complete genomes are publicly available. The DNA microarray technology permits the measurement of gene expression in cultured cells[1]. An increasing number of laboratories are using the combination of these two methods to study gene expression on a genomic scale. After all the genes from an organism are clustered based on their expression patterns[2], an important next step is to examine the upstream region of genes in the same expression pattern group and look for sequence motifs. These motifs might be the regulatory signal (most likely a transcriptional regulatory site) that causes these genes to respond similarly to

developmental or environmental changes. Information on expressions, regulatory motifs and functions provides substantial insight to the understanding of gene networks[3]. The motivation of this research is to provide a gene expression data analysis tool — to look for regulatory sequence motifs in the upstream region of genes in the same expression group.

There are generally two strategies for DNA sequence motif finding that are explored in recent years – using enumeration to check the over-representation of all possible *w*-mers and using iterative processes to update a motif probability matrix. The second strategy calculates the expected frequency of each possible motif of width *w* based on background or input sequence distribution[4], then searches for *w*-mers that are much more abundant in the input than expected[5, 6]. This method is guaranteed to find motifs with the greatest *z*-scores, but it does not allow flexible substitutions in the matching segments. Also, the motifs that could be enumerated are limited in size (≤ 7 bases long). The third strategy employs a probability matrix for the motif, specifying the probability of each base at each motif position. An iterative procedure, implementing either an expectation maximization (EM)[7] or a Gibbs sampling[8] algorithm, is then applied to improve the matrix until convergence. In recent years, many modifications have been made on this methodology[9, 10]. Workman and Stormo[11] even tried motif finding using artificial neural network with alignment methods closely related to EM and Gibbs sampling. Our method also adopts the Gibbs sampling approach, with added improvements in flexibility and sensitivity.

The improvements start with a better understanding of the dataset to be investigated. Since there may be more than one transcriptional mechanism involved within each group of sequences, some sequences in the input may have no copies of a motif while others may have multiple copies. Although this was addressed in the motif sampler by a mixture model[12], we used a different approach, called the *threshold sampler*, which is less susceptive to non-independence among input sequences. In some cases, simultaneous and proximal binding of two transcriptional factors or the binding of a homodimer may be required to initiate transcription. Therefore, considering the two binding blocks together may increase the signal strength. In addition, to better capture the characteristics of local DNA environment and structure, a Markov model for the background could be adopted which looks at successive duplet or triplet base pairs at a time[13]. Since the direction of transcription regulation is unknown, we need to check both forward and complementary strands of the input sequences. Finally, Gibbs sampling may report different motifs at different runs, so it is important to know the statistical significance of a reported motif.

2. Algorithm

2.1 Basic model

BioProspector is an algorithm for finding sequence motifs from a set of DNA sequences (Fig. 1). It takes the following input parameters:

- A file (F_{in}) with N DNA sequences in which the motifs are to be found.
- A file (F_{bg}) containing sequences or probabilities characterizing the background nucleotide distribution.
- The widths of the two motif blocks w_1 and w_2, and their gap range, [g_L, g_M]. In the case when a one-block motif is of interest, one can set w_2, g_L and g_M to 0.
- Whether each sequence has at least one copy of the motif.
- Whether the motif could occur in both DNA strands.
- Whether the motif has a palindromic pattern, in which case w_1 must be equal to w_2, and BioProspector checks both DNA strands automatically.

At the end, BioProspector outputs the following results:

- The motif score, significance value, and the number of aligned segments.
- A regular expression of the motif consensus and degenerate, as well as a probability matrix expression of the motif.
- The number of segments each input sequence contributes to the motif, the starting position and sequence of each segment.

Within each run of BioProspector, a process called threshold sampler is performed a number of times. Threshold sampler adopts the Gibbs sampling strategy, which initializes a motif probability matrix Θ by a random alignment of the input sequences and improves the matrix iteratively and stochastically by a predictive update method[8]. The predictive update formula used here, however, is based on the following important modifications of the underlying statistical model.

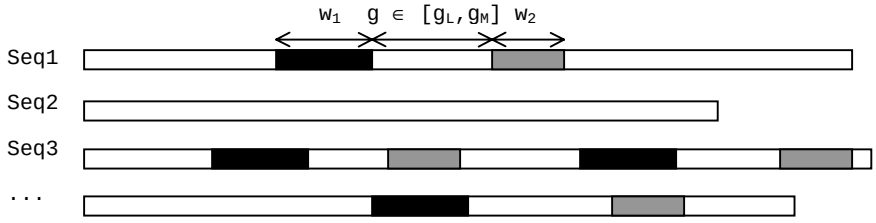

Figure 1. A graphical illustration of the model used for BioProspector. Suppose the input file consists of N DNA sequences, each containing 0 – n copies of a sequence motif. The motif has two binding blocks of width w_1 and w_2, respectively, which are separated by a gap of variable length ranging from g_L to g_M.

2.2 Scoring segments with background Markov dependency

In the original Gibbs sampler[8], every possible segment of width w within a randomly chosen sequence s (in F_{in}) is considered. A score $A_x = Q_x / P_x$ is computed and a new alignment position a_s is sampled with probability proportional to A_x. Here Q_x and P_x are the probability of generating segment x from the current motif matrix Θ and from the independent background model β, respectively. In DNA, however, the presence of a particular nucleotide usually has influence on its neighboring positions, so a better way to evaluate P_x is based on Markov background. For example, the probability of generating segment ATGTA from a third-order Markov background model β is calculated as:

$$P^3_{ATGTA} = p(\text{A}) \times p(\text{T} \mid \text{previous base is A}) \times p(\text{G} \mid \text{previous 2 bases are AT}) \times p(\text{T} \mid \text{previous 3 base are ATG}) \times p(\text{A} \mid \text{previous 3 bases are TGT})$$

BioProspector allows the user to specify one of the following two background file formats:

- A sequence file containing background sequences from which β is to be computed (it could be the same as the input sequence file F_{in}).
- A file with pre-computed background probabilities characterizing the complete genome (intergenic, ORF's, or the sum) of an organism (defaults to yeast).

In the first case, since Markov dependency order of f requires the estimate of 3×4^f parameters, the program picks f (≤ 3) so that the background sequences has about 1024×4^f bases. In the second case, a third-order Markov background model is automatically used because of the sufficient size of a genome.

2.3 Sampling new alignments with two score thresholds

In the original Gibbs sampler[8], a new motif alignment for a particular sequence s is chosen with probability proportional to A_x. However, this relies on the fact that each sequence contains a single copy of the motif. To deal with the problem that some input sequences contain no copies of the motif and some contain many copies, two thresholds, T_H and T_L, are introduced to the threshold sampler. During the sampling step, all the non-overlapping segments of sequence s with a score higher than T_H are automatically added to the motif, and their positions are added to the alignment a_s. For the rest of the segments in s with scores between $[T_L, T_H]$, one segment will be chosen with probability proportional to $A_x - T_L$. The high threshold T_H is chosen, based on a large deviation argument, to be proportional to the product of the average length of the input sequences and the motif width w. T_L is fixed at 0 for the first 10 iterations, and is linearly increased until it reaches $T_H / 8$ at the end of the procedure. If the user requests the program to search for motifs in both strands, the two strands

are considered as one sequence with only one segment sampled to the motif. The final alignment consists of all the segments with scores above T_H and the highest scoring segment between $[T_L, T_H]$ within each sequence.

Sampling only among segments with scores between $[T_L, T_H]$ helps the program converge more quickly. When the motif is near convergence, a sequence with multiple copies of the motif will have multiple segments above T_H, thus having all of the copies added to the motif. If a sequence does not have any segment with a score higher than T_L, it is considered as not containing the motif and no segment is sampled to the motif. In the case when the user specifies that each input sequence has at least one copy of the motif but a particular sequence has no segment with score above T_L, we sample a single segment in this sequence with probability proportional to A_x.

2.4 Finding two-block and palindromic motifs

For motifs with two blocks, BioProspector uses two probability matrices Θ_1 and Θ_2 to capture the two blocks. The matrices are initialized by randomly choosing the alignment positions (a_{s1}, a_{s2}) on the same strand from each sequence with a fixed gap $g_0 = (g_L + g_M)/2$. Two segments x_1 of width w_1 and x_2 of width w_2 within the gap range are scored as: $A_{x1,x2} = (Q_{x1}/P_{x1}) \times (Q_{x2}/P_{x2})$, in which Q_{x1} is the probability of generating x_1 by Θ_1 and Q_{x2} is the probability of generating x_2 by Θ_2. We sample x_1 from its marginal distribution, which is proportional to $A_{x1,*} = \Sigma_{x2} A_{x1,x2}$, where the sum is over all segments of width w_2 within $[g_L, g_M]$ downstream from x_1. Then segment x_2 is chosen with probability $A_{x1,x2}/A_{x1,*}$ conditioned on x_1.

When the two motif blocks are palindromic, we need only one motif probability matrix Θ. Each aligned sequence contributes two segments to the same matrix, one from each DNA strand.

2.5 Using motif score distribution to measure goodness of a motif

Kullback-Leibler information, also known as relative entropy, has been used to measure information content of a motif[14]. However, when *motif1* has 150 aligned segments whereas *motif2* has only 3, *motif2* can easily have better relative entropy although the former represents a more interesting and significant conserved motif. To resolve this dilemma, we introduce the following criterion to measure the goodness of a motif:

$$\textit{Motif Score} = \#seg \times \exp\{[\Sigma_{\textit{all positions } i}\ \Sigma_{\textit{all nucleotides } j}\ q_{i,j} \times \log(q_{i,j}/p_j)] / w\}$$

in which *#seg* is the number of aligned segments in the motif, $q_{i,j}$ is the probability of observing nucleotide *j* at position *i* of the motif matrix Θ, and p_j is the probability of observing nucleotide *j* from the background probabilities β.

To see how significant an observed motif score is, we first use Monte Carlo simulations to estimate the null distribution of this score. More precisely, the program generates *M* independent and identically distributed sequence sets under the input sequence probability model, where each generated set is identical to the input file F_{in} in sequence number and length. For each generated sequence set, a number of threshold sampler runs are performed, and the highest motif score is recorded. A normal distribution is then fitted to the *M* recorded scores. With this score distribution, BioProspector runs the original sequence through the threshold sampler, and reports motifs that are *z* (defaults to 5) standard deviations above the motif score distribution mean.

3 Method

BioProspector is developed in C and runs on all UNIX systems. Currently, it takes about 20 seconds on a 400 MHz Sun station to finish a run of threshold sampler on a 60-sequence data set with an average sequence length of 800 bases.

We used Bioprospector to test three sets of data. The first set consists of 60 non-coding sequences that were shown to physically interact with the *Saccharomyces cerevisiae* telomere-binding protein Rap1p. DNA associated with Rap1p were identified by chromatin immunoprecipitation (IP) and purification of DNA fragments enriched by the IP, followed by labeling and hybridization of purified fragments to DNA microarrays containing all of the yeast intergenic regions. The binding site for RAP1 is well characterized, and although published determinations differ slightly in length and consensus, they all agree on the core site RMAYCCR[3, 16, 17, 18, 19]. The sequences analyzed ranged in length from 163 to 1339 base-pairs, and some do not contain the RAP1-binding motif, while others contained multiple copies of it. Three runs of BioProspector were performed, using the input sequence, a zero-order, and a third-order Markov model estimated from the yeast intergenic region to represent the background, respectively. For each sequence, both the forward and complementary strands were examined. We chose $M = 200$ for an accurate approximation of the motif score distribution, although $M = 40$ usually gives a reasonable estimate. To examine the performance of threshold sampler on the original data, we let it run 250 times and recorded the motif score and consensus of each.

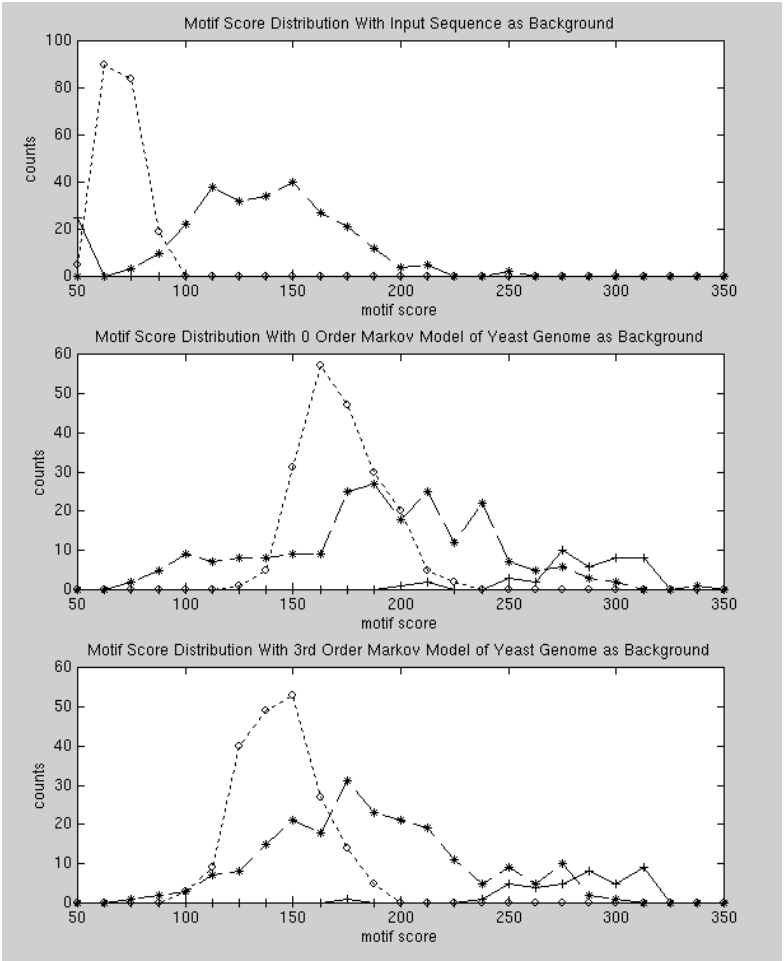

Figure 2. Motif scores of the RAP1-binding data. The top, middle and bottom plots show the motif score distribution using input sequence F_{in}, a zero-order, and a third-order Markov model estimated from yeast intergenic region as the background, respectively. Within each plot, the dotted line (with 'o' markers) represents the motif score distribution of the 200 generated sequence sets and each dotted line approximates a normal distribution; the dashed line (with '*' markers) and solid line (with '+' markers) represent score distribution of the 250 motifs found from original input sequences, in which the dashed line represents the false positive motif scores and the solid line represents the true positive motif scores. With a third-order Markov background model estimated from yeast intergenic region, all the motifs with a score above 305 are correct.

The second data set contains 136 σ^A-dependent promoter sequences (mostly at positions [-100, 15]) from *Bacillus subtilis*[20]. Each sequence has one RNA polymerase-binding motif on the forward strand, otherwise known as the TATA box. This is a two-block motif: the first block with consensus TTGACA mostly occurs at position –35, and the second block with consensus TATAAT mostly occurs at position –12. BioProspector performed motif finding on this data with a specified gap range of [15, 20].

The third data set consists of 18 *Escherichia coli* sequences of length 105 which are known to contain CRP-binding sites. CRP is a prokaryotic dimeric DNA-binding protein that binds to adjacent DNA major grooves in a palindromic pattern. One-block motif models using both EM[7] or Gibbs sampling[21] have been applied to this data and yielded satisfactory results, although both mispredicted one or two sites. We ran BioProspector on this data to search for a palindromic two-block motif. Since the first and last (22) positions of the binding site are not complementary (i.e., not a palindromic match), the parameters were specified to be $w_1 = w_2 = 8$ with a gap range of [1, 4].

4 Result

4.1 RAP1 site: background Markov dependency and motif score distribution

The 200 motif scores obtained from the 200 generated sequence sets were approximated by a normal distribution, no matter how the background model β was estimated (Fig. 2). When using the background model estimated from F_{in}, none of the reported motifs agree with the published RAP1 consensus. When using an independent background model estimated from yeast intergenic region, most of the high-scoring motifs are correct, although there are some high-scoring false positive motifs. When using a third-order Markov background model estimated from yeast intergenic region, the distributions of true positive and false positive motifs separate very well. At scores above 305, all the 9 motifs reported contain a consensus of ACACCCA which agrees with the published result.

4.2 TATA-box: two-block motifs

Among the 136 *B. subtilis* sequences containing the two-block TATA-box motif, BioProspector correctly found 70% of the sites and accurately identified the motif consensus as TTGACA, TATAAT. This motif is not very well conserved; many of the missed sites are significantly different from the consensus. For example, the

following four missed sites are so variable that the sites predicted by BioProspector match with the consensus better:

	Correct site		Site found	
ald	AAGAAT	TACACT	TTTCCA	TAAAAA
cspB	TTGTTT	TGGAGT	ATTACT	TATTTT
menE	AATACA	GATGAT	TTGAGA	TCTTTT
odhA	TTGTGA	CAAATT	TTTACT	TAGAAT

For the following 3 sequences, besides finding the correct sites, BioProspector also found a second site closely matching the consensus. In fact, the second site of sequence *veg* matches exactly with the TATA-box consensus, which is even better than the correct site.

	Correct site		Second site	
abrB	TTGACG	TAGTCT	CTGACT	TACAAT
veg	TTGACA	TACAAT	TTGACA	TATAAT
ϕ105	TTTACA	TACAAT	TTGACG	TACAAT

4.3 CRP site: palindrome motifs

The gold standard of this test is based on footprint experiments which identified 24 CRP-binding sites in the 18 sequences. However, the aligned segments are not very conserved, especially at the ending positions. Expectation maximization and Gibbs sampling with one-block motif model succeeded in finding most of the sites, although both mispredicted one or two sites (Table 1). With a two-block palindromic motif model, all the sites found by BioProspector are correct. The base shifts of the starting position of the first block were caused by our specification of a shorter block width and a flexible gap between the two blocks. The resulting probability matrix shows a much more conserved motif with a consensus of WTGTGAWM.

5 Discussion

As we have shown with the RAP1 binding motif data, using different background models can greatly influence the performance of BioProspector. We tried this dataset on a couple of web servers that implement Gibbs sampling, neither of which allow the user to choose different background models, and we failed to find the

correct RAP1-binding motif. Furthermore, it is sometimes helpful to estimate the background model from a set containing "contrasting" sequences. For example, the background file may contain sequences with a very popular motif *A*; whereas the input file F_{in} contains sequences with not only motif *A*, but also a less popular motif *B*. In this case, BioProspector would be able to find motif *B*.

In the RAP1 experiment, many of the high-scoring false positive motifs agree with a motif of consensus CTTACCCTAC. In fact, this motif is the highest scoring motif using zero-order Markov model estimated from yeast genome as background. This motif occurs quite frequently in the input sequences, and its score is highly significant. It is possible that besides RAP1, another protein binds to this group of sequences.

Table 1. Result comparison for the CRP data. a. Motif probability matrix obtained from 36 aligned palindromic CRP binding segments by BioProspector. b. Starting position of CRP sites by footprint experiment, EM and Gibbs with one-block motif model, and BioProspector with palindromic two-block motif model.

a. Motif probability matrix using BioProspector

%	1	2	3	4	5	6	7	8
A	36.1	5.5	5.5	0.0	5.5	80.5	30.5	25.0
G	25.0	2.7	72.2	0.0	86.1	0.0	16.6	22.2
C	11.1	11.1	2.7	2.7	0.0	16.6	22.2	44.4
T	27.7	80.5	19.4	97.2	8.3	2.7	30.5	8.3

b. Starting position of CRP sites

	Footprint	1-block Motif		2-block Palindrome		
Sequence	Sites	EM	Gibbs	Start1	Start2	Gap
cole 1	17, 61	61	61	63	73	2
eco arabop	17, 55	55	55	57	67	2
eco bglrl	76	76	76	78	88	2
eco crp	63	63	63	63	75	4
eco cya	50	50	50	52	62	2
eco deop	7, 60	7	7	9	19	2
eco gale	42	24	42	44	54	2
eco ilvbpr	39	39	39	41	53	4
eco lac	9, 80	9	9	83	92	1
eco male	14	14	14	16	26	2
eco malk	29, 61	61	61	63	75	4
eco malt	41	41	41	43	53	2
eco ompa	48	48	48	50	60	2
eco tnaa	71	71	71	73	83	3
eco uxul	17	17	17	19	29	2
pbr-p4	53	53	53	55	65	2
trn9cat	1, 84	5	5	1	12	3
Tdc	78	78	78	80	90	2

The two-block motif model works well when transcription regulation depends on the binding of two proteins (or a dimer). As long as the two proteins are spatially

close to each other, insertion and deletion between the two DNA binding sequences can be well tolerated. By allowing for a variable gap between the blocks instead of forcing the segments with insertions and deletions to align, BioProspector has a better ability to find the correct CRP-binding sites. As for the TATA-box experiment, the reason why BioProspector failed to detect 30% of the binding sites is probably because the RNA-polymerase-binding mechanism also depends on the specific physical-chemical and structural characteristics surrounding the two consensus sequences[22]. It would be of interest to see whether RNA polymerase would bind to the sites BioProspector predicted if the correct sites were deleted, especially for the 3 sequences predicted to have 2 copies of the motif.

Right now the most time consuming step in BioProspector is finding the motif score distribution using Monte Carlo. We are exploring methods for calculating the statistical significance of an alignment directly from input sequence base distribution and background base distribution[23], especially with third-order Markov background model and two-block motif models.

Once BioProspector finds a correct motif and calculates its probability matrix Θ, we can use Θ to screen the whole genome of an organism. For each sequence in the genome, the score ΣA_i (A_i is the ratio between the probability of a segment being generated by Θ and the probability of it being generated by the background β) could be a good measure of the overall binding likelihood of the sequence by a particular transcription-regulator protein. We can then use this score to recluster genes and provide insight in understanding the gene network of an organism.

Acknowledgments

The authors thank the Brown Lab at Stanford, especially Dr Jason D. Lieb, for the valuable RAP1-binding sequence data. X. Liu is supported by the Lucille P. Markey Biomedical Research Fellowship. This work is also supported by NSF grant DMS-9803649, and NIH grant LM05716-05.

References

1. DeRisi JL, Iyer VR, Brown PO, *Science*. **278**(5338):680-6 (1997).
2. Eisen MB, Spellman PT, Brown PO, Botstein D, *Proc Natl Acad Sci U S A*. **95**(25):14863-8 (1998).
3. Zhu J, Zhang MQ, *Pac Symp Biocomput*. **5**:476-487 (2000).
4. Atteson K, *Ismb*. **6**:17-24 (1998).

5. Tompa M, *Ismb*. **7**:262-71 (1999).
6. Brazma A, Jonassen I, Vilo J, Ukkonen E, *Genome Res*. **8**(11):1202-15 (1998)
7. Lawrence CE, Reilly AA, *Proteins*. **7**(1):41-51 (1990).
8. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC, *Science*. **262**(5131):208-14 (1993).
9. Bailey T, Elkan C, *Machine Learning*. **21**(1-2):51-80 (1995).
10. Hughes JD, Estep PW, Tavazoie S, Church GM, *J Mol Biol*. **296**(5):1205-14 (2000).
11. Workman CT, Stormo GD. *Pac Symp Biocomput*. **5**:467-78 (2000).
12. Liu, JS, Neuwald, AF, and Lawrence, CE. *J Amer Statist Assoc.* **90**:1156-1170 (1995).
13. Dickerson RE, *Methods Enzymol*. **211**:67-111 (1992).
14. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A, *J Mol Biol*. **188**(3):415-31 (1986).
15. Lieb JD, Brown PO. Unpublished data.
16. Buchman AR, Kimmerly WJ, Rine J, Kornberg RD, *Mol Cell Biol*. **8**(1):210-25 (1988).
17. Graham IR, Chambers A, *Nucleic Acids Res*. **22**(2):124-30 (1994).
18. Idrissi FZ, Pina B, *Biochem J*. **341**(Pt 3):477-82 (1999).
19. Lascaris RF, Mager WH, Planta RJ, *Bioinformatics*. **15**(4):267-77 (1999).
20. Helmann JD, *Nucleic Acids Res*. **23**(13):2351-60 (1995).
21. Liu JS, *J Amer Statist Assoc*. **89**(427): 958-966 (1994).
22. Lisser S, Margalit H, *Eur J Biochem*. **223**(3):823-30 (1994).
23. Hertz GZ, Stormo GD. *Bioinformatics*. **15**(7-8):563-77 (1999).