JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Number 1, 2004

Pp. 1–14

Determination of Local Statistical Significance of Patterns in Markov Sequences with Application to Promoter Element Identification

HAIYAN HUANG,[1,2,3] MING-CHIH J. KAO,[1,3] XIANGHONG ZHOU,[1]
JUN S. LIU,[1,2] and WING H. WONG[1,2]

ABSTRACT

High-level eukaryotic genomes present a particular challenge to the computational identification of transcription factor binding sites (TFBSs) because of their long noncoding regions and large numbers of repeat elements. This is evidenced by the noisy results generated by most current methods. In this paper, we present a *p*-value–based scoring scheme using probability generating functions to evaluate the statistical significance of potential TFBSs. Furthermore, we introduce the local genomic context into the model so that candidate sites are evaluated based both on their similarities to known binding sites and on their contrasts against their respective local genomic contexts. We demonstrate that our approach is advantageous in the prediction of myogenin and MEF2 binding sites in the human genome. We also apply LMM to large-scale human binding site sequences *in situ* and found that, compared to current popular methods, LMM analysis can reduce false positive errors by more than 50% without compromising sensitivity. This improvement will be of importance to any subsequent algorithm that aims to detect regulatory modules based on known PSSMs.

Key words: probability generating function, statistical significance, local genomic context, Position Specific Score Matrix (PSSM), transcription factor binding site.

INTRODUCTION

THE ELUCIDATION OF GENE FUNCTION, genetic network, and cellular processes requires the accurate identification of transcription factor binding sites (TFBSs). Experimental approaches, such as DNase footprinting (Galas and Schmitz, 1978) and gel mobility shift assay (Fried and Crothers, 1981; Garner and Revzin, 1981), are in general expensive and time consuming. Given the large number of transcription factors and the vast spans of noncoding genomic regions onto which they may bind, molecular characterization of transcription mechanisms will be facilitated by the prediction of transcription factor binding sites *in silico*.

[1]Department of Biostatistics, Harvard University, 655 Huntington Avenue, Boston, MA 02115.
[2]Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138.
[3]These authors contributed equally to this work.

Efforts on the computational prediction of TFBSs fall into two general approaches. The first seeks novel recurrent patterns in a set of DNA sequences, often the promoters of genes found to be coregulated in gene expression microarray experiments. A number of statistical models have been developed in the past decade for this purpose based on Bayesian models and Monte Carlo methods (Bailey and Elkan, 1994; Hughes *et al.*, 2000; Lawrence *et al.*, 1993; Lawrence and Reilly 1990; Liu *et al.*, 2001; Liu *et al.*, 2002; Roth *et al.*, 1998). They have been widely applied and found to be most successful in lower organisms such as bacteria and yeast. However, in higher organisms such as the human, these methods may yield noisy results because of the long noncoding regions and the large numbers of nonfunctional repeat elements (Lander *et al.*, 2001). A recent trend to improve upon these *de novo* methods is to incorporate the information from cross-species comparisons.

The other major approach to predict transcription factor binding sites makes use of prior knowledge on the binding sites. These methods evaluate individual candidate site sequences by their similarities to clusters of experimentally determined binding sites (Chen *et al.*, 1995; Hertz *et al.*, 1990; Quandt *et al.*, 1995; Stormo and Hartzell, 1989; Wingender *et al.*, 2000). These binding site sequences are most often summarized using *position-specific scoring matrices* (PSSMs), which are used to summarize the sequence patterns and to compare against candidate DNA segments. This is the approach of interest in this paper.

Various methods exist to score candidate segments for their similarities to known binding sites using PSSMs. We provide an example in Fig. 1 using the transcription factor myogenin. PSSM construction begins by using the alignment of known binding site sequences and tabulating the *nucleotide distribution matrix* (Fig. 1a). The counts are then transformed using either of two related schemes, *log-odds* (Fig. 1b) or *entropy* (Fig. 1c), to generate the PSSM. Candidate sites are scored against the PSSMs by summing over the corresponding scores of the nucleotides across the site sequence; i.e., the score of candidate site $S = S_1 \dots S_p$ against PSSM is $(w_{ij})_{p\times 4}$ is $S = \sum_{positioni} w_{iS_i}$. In practice, these scores are then compared to some predetermined cutoff values to generate computational TFBS predictions. Note that the most widely used database of transcription factor binding, TRANSFAC (Wingender *et al.*, 2000), is based on entropy-weighted PSSMs.

While probabilities are used in the construction of the PSSMs, the scores themselves cannot be interpreted statistically. This has led to the general difficulty with choosing the score cutoff values for each matrix, a problem that may have contributed to the large numbers of false positive predictions seen in practice. We propose a p-value based scoring scheme, which evaluates the statistical significance of the candidate site segment. This should apply to both the entropy-based and the log-odds–based scoring methods. However, in order to obtain a valid p-value, one needs to model the background sequence properly, which may serve either as the "null model" or a component in computing the log-odds scoring function.

In this paper, we model the background sequences, or the "null distribution," as a Markov chain. As in previous methods, candidate binding site sequences are scored by PSSMs. Each score is evaluated statistically by computing its p-value, that is, the probability that the background model can achieve a score at least as high as that observed. In order to calculate this p-value, we develop an efficient and exact algorithm based on probability-generating functions that can achieve up to 1000-fold speed up compared to Monte Carlo simulations. We note that in contrast to score-based evaluation, the p-values we generate can serve as a universal measure of statistical significance of all candidate binding sites regardless of their corresponding binding factor or of their genomic locations.

It has been known that the effectiveness of a binding site in recruiting its corresponding transcription factors can be dramatically affected by the genomic context that it is in. This can be attributed to a number of factors such as the local DNA bending, the accessibility of the binding site, or the positive or negative effects of neighboring TFBSs. We incorporate the local genomic context into the p-value–based scoring method and develop the Local Markov Method (LMM). The p-value for a candidate site provides a measure of its similarity to known binding sites and its contrast against the local genomic context. We first show that the incorporation of the local genomic context can be advantageous in the prediction of myogenin and MEF2 binding sites in the human genome, an advantage observed independently of the method of PSSM construction. We further compare the abilities of LMM and TRANSFAC to pick up 101 experimentally determined TFBSs from large tracts of human genomic sequences and find that LMM can identify TFBSs with more specificity (50% fewer false positive predictions) without compromising sensitivity. The LMM software is available upon request (*www.biostat.harvard.edu/complab/LMM*).

(a) Nucleotide count matrix

The number in each entry counts the frequency of each base A, C, G, or T, in the corresponding position of the aligned binding sites

Position	1	2	3	4	5	6	7	8
A	4	2	0	9	0	0	0	0
C	0	2	9	0	0	4	0	0
G	5	5	0	0	8	4	0	9
T	0	0	0	0	1	1	9	0

(b) log-odds PSSM

For each position i and for each base x = A, C, G, or T, a log odds is calculated as $\log\frac{m_i(x)}{q_x}$ to derive the Durbin PSSM (Durbin et al. 1998), where m_i is the probability for observing base x at position i from the nucleotide distribution matrix, and q_x is the probability of observing base x under a random model. The random model is often estimated using a large collection of intergenic regions and applied to all instances of candidate sites.

Position	1	2	3	4	5	6	7	8
A	0.49	-0.10	-1.70	1.24	-1.70	-1.70	-1.70	-1.70
C	-1.70	-0.10	1.24	-1.70	-1.70	0.49	-1.70	-1.70
G	0.69	0.69	-1.70	-1.70	1.13	0.49	-1.70	1.24
T	-1.70	-1.70	-1.70	-1.70	-0.61	-0.61	1.24	-1.70

(c) Entropy-weighted PSSM

The values in the PSSM are derived by weighting the counts in the nucleotide distribution matrix at each position using a entropy-related information measure, $\frac{100}{\ln 5}\cdot\sum_{\substack{x\in A,C,\\ GTgap}}\left(m_i(x)\ln m_i(x)+\ln 5\right)$, where m_i is the probability for observing base x at position i from the nucleotide distribution matrix (Quandt et al. 1995).

Position	1	2	3	4	5	6	7	8
A	202	56	0	900	0	0	0	0
C	0	56	900	0	0	122	0	0
G	252	141	0	0	599	122	0	900
T	0	0	0	0	75	30	900	0

FIG. 1. The construction of a position specific scoring matrix for myogenin binding sites.

MAIN RESULTS

p-value calculation

In order to calculate the probability that a given Markov background model can achieve a score at least as high as the observed score of the candidate site, we extend a previous method designed for a similar purpose but applicable only to the independent and identically distributed (*iid*) background sequence model (Staden, 1989). The key part for this method is the reformulation of the distribution of the score as a probability-generating function which leads to an efficient algorithm for its computation. We formulate the score probability-generating function under Markov models (detailed in the Detailed Methods section) and derive an algorithm with time complexity linear in the length of the PSSM, a dramatic improvement over the naive enumeration method which has time complexity exponential in the length of the PSSM.

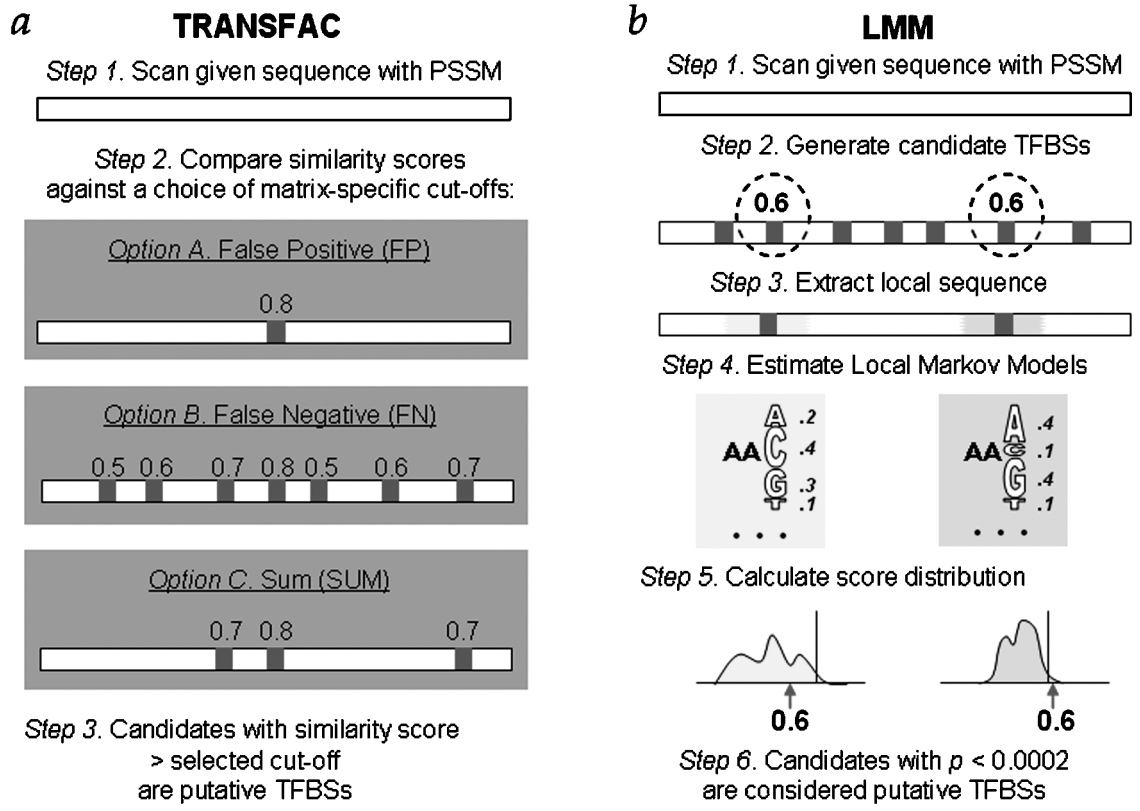


FIG. 2. TRANSFAC vs. local Markov model (LMM) in the identification of transcription factor binding sites (TFBSs) in a given genomic sequence. (**a**) TRANSFAC scans a genomic sequence, generates similarity scores of each subsequence against a given PSSM, and uses three matrix-specific cutoffs FN, FP, and SUM to make putative calls. The three sets of cutoffs attempt to minimize false-negative error, false-positive error, or the sum of these two errors, respectively. (**b**) LMM begins by selecting the top 0.1% candidate sites based on their PSSM similarity scores, since sites with low similarity scores are unlikely to be true binding sites. For each candidate TFBS, LMM models the DNA sequence segment of length L (e.g., 1,000) centered around the target site as a homogeneous Markov chain of orders $k = 0$, 1, 2, or 3. Under the estimated Markov model, LMM calculates the probability distribution of the similarity score using our algorithm. This distribution then allows us to assign statistical significance to the given candidate TFBS.

We implement the algorithm in C and incorporate the program into the local Markov method (LMM) program to study TFBSs *in situ* by evaluating each candidate binding site with respect to its local genomic context. A summary of the LMM method is in Fig. 2b. For comparison, we also describe in Fig. 2a the prediction program which accompanies the TRANSFAC database.

To assess the efficiency of our algorithm, we compare it against Monte Carlo simulations. Our experiments showed that the efficiency of our approach can be many times more efficient than Monte Carlo simulation. For example, at $p \leq 0.0001$, a sequence of length at least 10^9 basepairs needs to be simulated in order to obtain a sufficiently accurate cutoff value (relative error $\leq 1\%$), which needs more than 1000-fold more computing time than our exact algorithm (Table 1).

APPLICATIONS

MYL1 3′ enhancer myogenin binding site prediction

In human, mouse, and rat, there is a well-conserved 200bp-long skeletal muscle-specific enhancer about 24 kb 3′ of MYL1 (Rosenthal *et al.*, 1990; Wentworth *et al.*, 1991). Three myogenic determination factor binding sites A, B, and C are found in this region which are located 1267 bp, 1323 bp, and 1339 bp,

TABLE 1. RUNNING TIME COMPARISON OF LMM WITH MONTE CARLO SIMULATION[a]

		Significance at $p \leq 0.001$		*Significance at $p \leq 0.0001$*	
		Relative error		*Relative error*	
Simulation	*Time (sec)*	*Mean*	*S.D.*	*Mean*	*S.D.*
$N = 10^5$	0.06	6.5%	4.3%	24.0%	16.9%
$N = 10^6$	0.6	1.7%	1.2%	7.7%	4.5%
$N = 10^7$	5.7	1.0%	0.7%	2.6%	1.7%
$N = 10^8$	57	1.0%	0.5%	2.0%	0.8%
$N = 10^9$	570	0.7%	0.6%	0.6%	0.4%
$N = 2^{32}$	1228	0.6%	0.6%	0.6%	0.4%
LMM	0.43	0%		0%	

[a]For 10 randomly chosen intergenic regions, we estimate a 2nd-order Markov model for each sequence using maximum likelihood estimation. Under the 10 estimated Markov models, we use our algorithm to derive and Monte Carlo simulation to estimate the score cutoffs of the p53 PSSM at two significance levels, $p \leq 0.001$ and $p \leq 0.0001$, on a 1,500 MHz AMD Athlon machine running Linux. To assess the difference between the cutoffs C_p derived by our algorithm and the cutoffs C_p estimated by simulations, we consider the p-value $F(C_p)$ attained by C_p and the true p-value $F(C_p)$ of C_p derived using our algorithm. We assess the *relative errors* of the simulation estimate by calculating $(F(C_p) - F(C_p))/F(C_p)$.

respectively, downstream of the last exon of MYL1 in the human genome. Sites A and B are myogenin/myf4 binding sites (Rosenthal *et al.*, 1990), while site C is a MyoD binding site (Wentworth *et al.*, 1991), also considered to be a myogenin binding site (Fickett, 1996).

We applied the LMM to the 10,000 bp MYL1 downstream region (starting from the end of the last exon) to derive the local p-values for each candidate. The local p-value for each candidate is the statistical significance of observing its score (derived by both log-odds and entropy-related PSSMs) assuming that it is generated under a local random model, where Markov models of different orders (e.g., 0, 1, or 2) are used and with parameters estimated from the local 1,000 bp genomic sequence centered at the candidate. The top 10 score candidate sites derived using log-odds or entropy-weighted PSSMs are listed in Table 2a and 2b,

TABLE 2. INCORPORATING LOCAL SEQUENCE INFORMATION TO TRANSCRIPTION FACTOR BINDING SITE PREDICTION USING TWO TYPES OF PSSMS FOR MYOGENIN IN THE HUMAN MYL1 3′ ENHANCER (a,b), OR FOR MEF2 IN THE HUMAN PHOSPHOGLYCERATE MUTASE PROMOTER (c,d)[a]

(a) Using log-odds myogenin PSSMs

Position (bp from last exon of MYL1)	*Log-odds PSSM score*	*p-values of observed score under local background model*			
		iid	*1st Markov♣*	*2nd Markov*	*Binding site*
1267 (A)	556	0.000008	0.000017	0.000030	AGCAGGTG
1339 (C)	550	0.000015	0.000027	0.000055	GACAGGTG
1323 (B)	548	0.000033	0.000057	0.000112	ACCAGCTG
5434	556	0.000036	0.000074	0.000095	AGCAGCTG
2463	550	0.000059	0.000135	0.000179	GCCAGCTG
1235	531	0.000212	0.000354	0.000442	ACCATGTG
926	534	0.000181	0.000363	0.000468	TGCAGGTG
2574	536	0.000225	0.000416	0.000421	GGCAGATG
783	531	0.000274	0.000453	0.000537	AACATCTG
470	529	0.000404	0.000624	0.000731	GGAAGCTG

(*continued*)

TABLE 2. (*Continued*)

(b) Using entropy-weighted myogenin PSSM

Position (bp from last exon of MYL1)	*TRANSFAC score*	*p-values of observed score under local background model*			
		iid	*1st Markov*♣	*2nd Markov*	*Binding site*
1267 (A)	4667	0.000008	0.000017	0.00003	AGCAGGTG
1339 (C)	4628	0.000018	0.000032	0.000059	GACAGGTG
5434	4667	0.000036	0.000074	0.000095	AGCAGCTG
1323 (B)	4581	0.000045	0.000077	0.000127	ACCAGCTG
2463	4628	0.000068	0.000152	0.000194	GCCAGCTG
2574	4596	0.000073	0.000177	0.000191	GGCAGATG
926	4463	0.000224	0.000414	0.000532	TGCAGGTG
7534	4377	0.000378	0.000534	0.000788	TACAGCTG
7156	4377	0.000346	0.00054	0.000686	CCCAGCTG
4895	4322	0.000829	0.001998	0.002045	CTCAGGTG

(c) Using log-odds MEF2 PSSMs

Position (bp from last exon of MYL1)	*Log-odds PSSM score*	*p-values of observed score under local background model*			
		iid	*1st Markov*♣	*2nd Markov*	*Binding site*
−2970	669	0.000199	0.000160	0.000228	ATTTTAAATA
−3115	671	0.000209	0.000183	0.000243	GTTATAAATA
−161	649	0.000355	0.000183	0.000322	ATTTTAAGCA
−2939	668	0.000233	0.000190	0.000266	TGTTTAAATC
−3151	663	0.000807	0.000655	0.000747	TGTTTAAGAA
−4767	656	0.000951	0.001009	0.001712	TTTTTATATA
−3433	649	0.003940	0.003099	0.003383	AAACTAAAAA
−3566	644	0.005710	0.004913	0.005231	TTTTTAAAGC
−3214	643	0.007155	0.005654	0.006459	AGTTTATATC
−3577	641	0.007363	0.006312	0.006625	GGTTTAACAT

(d) Using entropy-weighted MEF2 PSSM

Position (bp from last exon of MYL1)	*TRANSFAC score*	*p-values of observed score under local background model*			
		iid	*1st Markov*♣	*2nd Markov*	*Binding site*
−2970	591	0.000103	0.000082	0.000133	ATTTTAAATA
−161	550	0.000191	0.000101	0.000174	ATTTTAAGCA
−3115	590	0.000206	0.000181	0.000243	GTTATAAATA
−2939	554	0.001001	0.000807	0.001016	TGTTTAAATC
−3151	562	0.001091	0.000914	0.001083	TGTTTAAGAA
−4700	531	0.001451	0.001451	0.001451	TTGTTAAAGA
−3566	543	0.002961	0.002532	0.002676	TTTTTAAAGC
−3433	545	0.003271	0.002610	0.002788	AAACTAAAAA
−4444	532	0.003080	0.003200	0.004320	CATATAATTA
−3687	535	0.003761	0.003320	0.003671	GAAGTAAAGA

[a] Sorted in increasing order by column marked with ♣.

respectively, with the true sites A, B, and C labeled and shaded in gray, along with their local p-values. We find the PSSM scores to be less sensitive a measure than the local p-value: the true sites A, B, and C stood out under the local p-values, while they are not as distinct from the false predictions under the PSSM scores.

PGAM-M MEF2 binding site prediction

A major positive regulatory element is required for the muscle-specific expression of the muscle-specific subunit of the human phosphoglycerate mutase (PGAM-M) gene (Nakatsuji *et al.*, 1992). This element, located 161 bp upstream of the gene, is found to be bound by the transcription factor MEF-2.

We applied the LMM to the 5,000 bp PGAM-M upstream region using the MEF2_Q6 PSSM to derive the local p-values for each candidate. The top 10 score candidate sites derived using log-odds or entropy-weighted PSSMs are listed in Tables 2c and 2d, respectively, with the true site labeled and shaded in gray, along with their local p-values. We find that LMM behaves similarly as in the MYL1 enhancer.

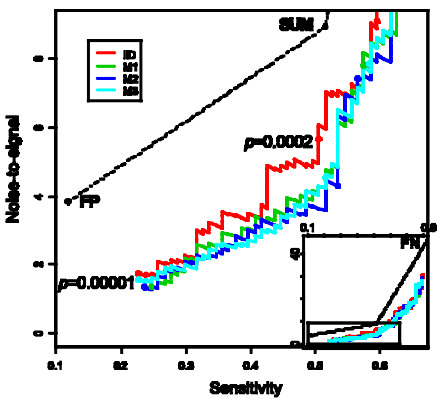
Overall, from Table 2, we see that by taking into account the local sequence composition we have reordered the candidate sequences in a way that is favorable to the true binding sites.

LARGE-SCALE VALIDATION

In order to evaluate the performance of LMM and to compare our local p-values to PSSM similarity scores, we apply both LMM and TRANSFAC to 101 known binding sites in the human genome obtained by mapping binding sites in the TRANSFAC database onto the human genome. We recorded and evaluated the extent to which LMM and TRANSFAC can capture this large collection of known binding sites in the human genome and the amount of noise generated in so doing.

In Figure 3a, the trade-off between sensitivity and noise is shown, in terms of the proportion of the known binding sites detected and the amount of concomitant noise generated. Noise is measured by the *noise-to-signal* ratio, which is defined as the number of binding site calls not known to be correct divided by the number of known binding sites found. For comparison, we show the tradeoffs achieved by TRANSFAC using its three matrix-specific similarity score cutoffs (FN, SUM, FP) along with that achieved by LMM under Markov models of orders 0, 1, 2, and 3 at various p-value cutoffs starting at the stringent $p = 0.00001$. From the inset graph, we see that at all levels of sensitivity, LMM outperformed TRANSFAC

(a)



(b)

	TRANSFAC			LMM							
				p=0.00001				p=0.0002			
	FN	FP	SUM	IID	M1	M2	M3	IID	M1	M2	M3
Accurate Calls	83	12	52	24	26	25	24	51	53	54	52
All Calls	3870	58	516	65	60	58	62	339	303	286	288
Noise to signal	45.6	3.8	8.9	1.7	1.30	1.3	1.6	5.6	4.7	4.3	4.5
Sensitivity	0.82	0.12	0.51	0.23	0.26	0.25	0.24	0.50	0.52	0.53	0.51

FIG. 3. Large-scale validation of TRANSFAC and LMM: Tradeoff between sensitivity and noise. (**a**) We compared the abilities of the two methods to detect the 101 known binding sites in the human genome by looking at their *sensitivity* and *noise-to-signal ratio.* The balance of the tradeoff between these two measures achieved at various significance levels by LMM are traced and compared to that attained by TRANSFAC. The inset graph shows the performance of LMM and TRANSFAC across all levels of sensitivity. (**b**) Detailed results for $p = 0.00001$ and 0.0002.

by producing significantly less noise. While the performance of LMM comes close to that of TRANSFAC as the p-value cutoff increases, in fact, by then, for both methods, the advantage of increased sensitivity has been nullified by the high level of accompanying noise, rendering them impractical. Overall, not only is the sensitivity of LMM comparable to TRANSFAC, its noise-to-signal ratio is also vastly superior. It should be noted that since only a limited number of true binding sites are known, not every unsupported binding site prediction is necessarily a false-positive prediction. Thus, the noise-to-signal ratio overestimates the true noise level, especially when stringent criteria are used to generate putative TFBSs with high sequence similarity to known binding sites. As the criteria relax, the large numbers of predictions over and above the known binding sites imply a high level of true background noise.

More detailed results for TRANSFAC using the three cutoffs and for LMM using different significance cutoffs, 0.00001 and 0.0002, and under different Markov models are summarized in Fig. 3b. While the FN cut off missed relatively few known binding sites, it generated more than 45 false-positive predictions for every accurate binding site call. On the other hand, FP made fewer false positives, but it detected only one in nine known binding sites. The SUM cutoff, designed as a balance of these inherent tradeoffs, did strike a reasonable compromise, having generated about nine false positives for every real binding site and detected more than half of the known sites.

At the stringent significance cutoff $p = 0.00001$, LMM detected about twice the binding sites than did the FP cutoff and on average produced about 60% fewer false-positive predictions for every correct prediction. At the more relaxed p-value cutoff $p = 0.0002$, the sensitivity of LMM is comparable to that of the SUM cutoff while only half of the noise is generated. The binding sites that were detected by LMM at $p \leq 0.0002$ but missed by TRANSFAC using the SUM cutoff include a MEF2 binding site over the desmin gene, an ATF1 (activating transcription factor 1) binding site over the TGFβ2 gene, a HIF (hypoxia-inducible factor) binding site over the VEGF gene, and an ICSBP (IFN consensus sequence binding protein) binding site over the OAS1 gene. We choose $p = 0.0002$ as the general significance cutoff for the application of LMM to mammalian genomic sequences, a cutoff with a sufficiently high sensitivity and an acceptable amount of noise. Overall, the LMM provides an advantageous tradeoff between noise-to-signal ratio and sensitivity.

In our validation experiment, we found that Markov models of orders 1, 2, and 3 have better combinations of high sensitivity and low noise than the *iid* model, confirming an earlier observation (Liu *et al.*, 2001) that Markov models can better capture the structure of biological sequences. In addition, we compared

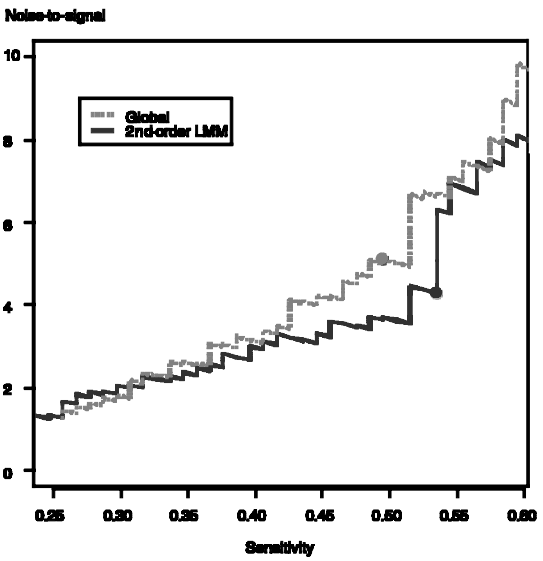


FIG. 4. The use of local sequence context is advantageous. The performance of the second-order LMM is compared against an analogous global Markov model with parameters estimated from a large collection of upstream regions. The performance is assessed in terms of the noise-to-signal ratio and sensitivity. At the recommended p-value cutoff 0.0002, LMM is more sensitive and less noisy.

the performance of the second-order LMM against an analogous global Markov model with parameters estimated from a large collection of upstream regions, in order to assess the ability of LMM to model the local sequence context information. We found that over the 101 known human TFBSs *in situ*, LMM generally outperforms the global Markov model, while they behave similarly at high and low sensitivity levels (Fig. 4). At high sensitivity levels, the lax p-value cutoffs produce large numbers of putative TFBS calls, overwhelming the advantage enjoyed by LMM. At low sensitivity levels, the stringent p-value cutoffs yield only putative TFBSs with undeniable sequence similarity to known binding sites. Thus, the noise-to-signal ratio may not reflect the true noise level in this region.

DISCUSSION

The work presented in this paper attemps to identify TFBSs by considering simultaneously both their similarity to the query PSSM and their differences from the local genomic context. Through the study of the human TFBSs in TRANSFAC, we show that LMM, which makes putative TFBS calls using local p-values, yields a much improved false-positive to true-positive ratio than that using the TRANSFAC or log-odds scores alone.

It has been known that neighboring nucleotide compositions can affect the interaction between a transcription factor and its binding site. To our best knowledge, however, there is no documented study on whether and how much an improvement can be made on the PSSM-based TFBS detection using a local background model. The result we present, which is based on more than 100 experimentally determined TFBS sequences in the human genome, shows a clear overall advantage for incorporating the local sequence context into PSSM-based TFBS search. There are various biological mechanisms that can explain this effect, which may lead to more complicated and more specific models. For instance, it may be that the local 1,000 bp genomic region does not contain DNA sequences similar to the true binding site because otherwise the target transcription factor may be competed away from its biologically meaningful binding site.

While this improvement does not in itself render a solution to the much more difficult problem of detecting regulatory modules, by significantly reducing false-positive calls for single sites, the local p-value approach will contribute substantially to any subsequent algorithms aiming to detect combinatorial regulatory modules. The method we developed here is seen as a proof of principle and can be used as a component of a more complex approach. For example, considering that clusters of binding sites also often occur within small regions of about 200 bp to cooperatively recruit the transcription factors, a natural future development of LMM would be to take this distance effect into the background estimation and combine the LMM p-values of a few candidate PSSM sites. Many challenging problems in computational biology, e.g., translation initiation site identification, splice site recognition, and RNA secondary structure prediction, can be modeled in terms of the recognition of motifs. Our work may be adapted and extended to these problems as well. However, it should be noted that when applied to protein sequences, which are composed of a 20-letter alphabet, the performance of our algorithm may become an issue, especially when the order of the Markov chain k is large.

DETAILED METHODS

Data extraction for large-scale validation

To evaluate the performance of the LMM, we apply it to known TFBSs in the human genome. Known binding sites are extracted from the SITE table of the TRANSFAC database version 6.2. About half of the 12,262 binding sites in this table are experimentally derived from various species. The rest are generated from *in vitro* binding assays on artificial nucleotide sequences. Since LMM studies binding sites with respect to their genomic contexts, these artificial sequences, which do not correspond to any genomic region, cannot be used for our validation study. Of the 6,073 *in vivo* binding sites, 1,425 sites are based on the human genome. Of these, 149 (10.5%) are annotated with a corresponding PSSM. We use these binding sites for validation.

To locate the known TFBSs in the human genome, we focus on the 5,000 bp upstream sequences of all genes. We made use of the annotations provided by Ensembl (Hubbard *et al.*, 2002) and extracted 22,808 human gene promoters from the human genome assembly NCBI golden path 29 (*www.ensembl.org/Homo_sapiens*). Since heuristic sequence-mapping algorithms do not perform well on short sequences such as TFBSs, we use an exact-match algorithm based on suffix trees (Gusfield, 1997). We found that many binding site sequences are precisely mapped onto the promoters of the correct target genes. For those binding sites with mappings onto multiple promoters or with no mapping, we attempted to retrieve them by manual review. To find the correct one among multiple mappings, we made correspondences between the Ensembl gene name and the target gene name of the binding site as recorded by TRANSFAC. A review of some missed matches using inexact match algorithms revealed a small number of single-basepair differences between the recorded binding site sequences and the promoter sequences of the target genes, for example, the binding site HS$ALBU_06 over the human albumin promoter. After validating against the primary literature for the positions of these binding sites, we included these mappings as well. In total, we located 101 human TFBSs.

Local p-value calculation

Although the exact score distribution can be obtained by enumerating all possible binding site sequences under any "null" model for the observed nucleotide base pairs, the computational cost for a PSSM of length p is 4^p. Staden's method (Staden, 1989), which turns this into an order-p computation, is based on the PGF of the score under the simple null model that the base pairs are independent and identically distributed (*iid*). Recently, however, there are some evidences suggesting that Markov background models work better than the *iid* model for detecting TFBS (Liu *et al.*, 2002). By extending Staden's PGF method to dependent random variables, we present here the derivation of the PGFs under a first-order Markov model, the basis of the efficient algorithm for computing the exact score distribution.

Probability generating function derivation

In our study, we make use of the PSSMs constructed by TRANSFAC version 6.2. Given a PSSM $m = (w_{ij})_{p\times 4}$, where $i = 1, \ldots, p$ and $j = A,C,G,T$, the match score S and the similarity score S/S_{max} of a sequence $D_1 D_2 \ldots D_p$ is defined as (Quandt *et al.*, 1995)

$$S = \sum_{i=1}^{p} w_{iD_i} \text{ and } S/S_{max} = \sum_{i=1}^{p} w_{ij} \Big/ \sum_{i=1}^{p} \max_j\{w_{ij}\}.$$

Let S be a random variable taking integer values; then its probability generating function, $G(t)$, is the expected value of t^S, $G(t)$ is a polynomial, and the coefficient of the term t^n is the probability of the event $S = n$ (Gut, 1995).

Given a PSSM m of length p, under the assumption that the DNA sequence is *iid*, Staden provided the PGF of the match score in the form of a product of p polynomials (Staden, 1989): $G(t) = \prod_{i=1}^{p} \sum_{j=A,C,G,T} f_j t^{w_{ij}}$, where f_j is the frequency of letter j in the *iid* DNA sequence. For the first-order Markov case, $k = 1$, let the transition matrix be $\boldsymbol{P} = (f_{\alpha|\beta})_{4\times 4}$ and the stationary distribution of the Markov chain be π (viewed as a four-dimentional row vector). Then the PGF under the first-order Markov model is

$$G(t) = \pi \prod_{i=1}^{p} (\boldsymbol{P}M(i,t))I, \qquad (*)$$

where $M(i,t) = Diag(t^{w_{iA}}, t^{w_{iC}}, t^{w_{iG}}, t^{w_{iT}})$, and $I = (1,1,1,1)^T$ (proof provided at the end of this section).

Since a Markov chain of order k on set Γ is equivalently a first-order Markov chain on the set Γ^k, with a little modification on $M(i,t)$, we can generalize the above results to $k > 1$. An example of PGF for

$k = 3$ is in the online supplement (*www.biostat.harvard.edu/complab/LMM*). Using this representation for the PGF, we developed and implemented an algorithm using C++ to calculate the exact score distribution. Generally, for a kth-order Markov chain and a PSSM of length p, the time complexity of our algorithm is $O(4^k \cdot S_{max} \cdot p)$, linear in the matrix length but exponential in the order of the Markov chain. The source code is available upon request (*www.biostat.hsph.harvard.edu/LMM*).

Proof of equation (∗). For ease of notation and without loss of generality, we let p, the length of the PSSM, be 3.

For a DNA sequence $D_1D_2D_3$, its match score against PSSM m is $w_{1D_1} + w_{2D_2} + w_{3D_3}$, and the probability of the occurrence of $D_1D_2D_3$ is $f_{D_1} f_{D_2|D_1} f_{D_3|D_2}$. By definition, the PGF of match score against m is

$$\sum_{D_1,D_2,D_3} f_{D_1} f_{D_2|D_1} f_{D_3|D_2} t^{w_{1D_1}+w_{2D_2}+w_{3D_3}} = \sum_{D_1,D_2,D_3} f_{D_1} t^{w_{1D_1}} \cdot f_{D_2|D_1} t^{w_{2D_2}} \cdot f_{D_3|D_2} t^{w_{3D_3}}.$$

In the following, we derive the PGF in the alternative form of a product of p matrices. First,

$$\begin{aligned}
&\sum_{D_1,D_2,D_3} f_{D_1} t^{w_{1D_1}} \cdot f_{D_2|D_1} t^{w_{2D_2}} \cdot f_{D_3|D_2} t^{w_{3D_3}} \\
&\quad = \sum_{D_1,D_2,D_3} \sum_a f_a \cdot f_{D_1|a} t^{w_{1D_1}} \cdot f_{D_2|D_1} t^{w_{2D_2}} \cdot f_{D_3|D_2} t^{w_{3D_3}} \\
&\quad = \sum_a f_a \cdot \sum_{D_1,D_2,D_3} f_{D_1|a} t^{w_{1D_1}} \cdot f_{D_2|D_1} t^{w_{2D_2}} \cdot f_{D_3|D_2} t^{w_{3D_3}} \\
&\quad = (f_a, f_C, f_G, f_T) \cdot \left(\sum_{D_1,D_2,D_3} f_{D_1|A} t^{w_{1D_1}} \cdot f_{D_2|D_1} t^{w_{2D_2}} \cdot f_{D_3|D_2} t^{w_{3D_3}}, \right. \\
&\qquad \left. \ldots, \sum_{D_1,D_2,D_3} f_{D_1|T} t^{w_{1D_1}} \cdot f_{D_2|D_1} t^{w_{2D_2}} \cdot f_{D_3|D_2} t^{w_{3D_3}} \right)^T.
\end{aligned}$$

For the component of the second vector corresponding to base A,

$$\begin{aligned}
&\sum_{D_1,D_2,D_3} f_{D_1|A} t^{w_{1D_1}} \cdot f_{D_2|D_1} t^{w_{2D_2}} \cdot f_{D_3|D_2} t^{w_{3D_3}} \\
&\quad = \sum_{D_1} f_{D_1|A} t^{w_{1D_1}} \cdot \sum_{D_2,D_3} f_{D_2|D_1} t^{w_{2D_2}} \cdot f_{D_3|D_2} t^{w_{3D_3}} \\
&\quad = (f_{A|A} t^{w_{1A}}, f_{C|A} t^{w_{1C}}, f_{G|A} t^{w_{1G}}, f_{T|A} t^{w_{1T}}) \\
&\qquad \cdot \left(\sum_{D_2,D_3} f_{D_2|A} t^{w_{2D_2}} \cdot f_{D_3|D_2} t^{w_{3D_3}}, \ldots, \sum_{D_2,D_3} f_{D_2|T} t^{w_{2D_2}} \cdot f_{D_3|D_2} t^{w_{3D_3}} \right)^T.
\end{aligned}$$

We apply similar arguments to the components corresponding to bases C, G, and T and obtain

$$\sum_{D_1,D_2,D_3} f_{D_1|C}t^{w_{1D_1}} \cdot f_{D_2|D_1}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}}$$

$$= (f_{A|C}t^{w_{1A}}, f_{C|C}t^{w_{1C}}, f_{G|C}t^{w_{1G}}, f_{T|C}t^{w_{1T}})$$

$$\cdot \left(\sum_{D_2,D_3} f_{D_2|A}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}}, \ldots, \sum_{D_2,D_3} f_{D_2|T}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}} \right)^T,$$

$$\sum_{D_1,D_2,D_3} f_{D_1|G}t^{w_{1D_1}} \cdot f_{D_2|D_1}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}}$$

$$= (f_{A|G}t^{w_{1A}}, f_{C|G}t^{w_{1C}}, f_{G|G}t^{w_{1G}}, f_{T|G}t^{w_{1T}})$$

$$\cdot \left(\sum_{D_2,D_3} f_{D_2|A}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}}, \ldots, \sum_{D_2,D_3} f_{D_2|T}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}} \right)^T,$$

$$\sum_{D_1,D_2,D_3} f_{D_1|T}t^{w_{1D_1}} \cdot f_{D_2|D_1}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}}$$

$$= (f_{A|T}t^{w_{1A}}, f_{C|T}t^{w_{1C}}, f_{G|T}t^{w_{1G}}, f_{T|T}t^{w_{1T}})$$

$$\cdot \left(\sum_{D_2,D_3} f_{D_2|A}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}}, \ldots, \sum_{D_2,D_3} f_{D_2|T}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}} \right)^T.$$

Therefore, for the first position, we have

$$\left(\sum_{D_1,D_2,D_3} f_{D_1|A}t^{w_{1D_1}} \cdot f_{D_2|D_1}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}}, \ldots, \sum_{D_1,D_2,D_3} f_{D_1|T}t^{w_{1D_1}} \cdot f_{D_2|D_1}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}} \right)^T,$$

$$= \begin{pmatrix} f_{A|A}t^{w_{1A}} & f_{C|A}t^{w_{1C}} & f_{G|A}t^{w_{1G}} & f_{T|A}t^{w_{1T}} \\ f_{A|C}t^{w_{1A}} & f_{C|C}t^{w_{1C}} & f_{G|C}t^{w_{1G}} & f_{T|C}t^{w_{1T}} \\ f_{A|G}t^{w_{1A}} & f_{C|G}t^{w_{1C}} & f_{G|G}t^{w_{1G}} & f_{T|G}t^{w_{1T}} \\ f_{A|T}t^{w_{1A}} & f_{C|T}t^{w_{1C}} & f_{G|T}t^{w_{1G}} & f_{T|T}t^{w_{1T}} \end{pmatrix} \cdot \begin{pmatrix} \sum_{D_2,D_3} f_{D_2|A}t^{w_{2D_2}} f_{D_3|D_2}t^{w_{3D_3}} \\ \sum_{D_2,D_3} f_{D_2|C}t^{w_{2D_2}} f_{D_3|D_2}t^{w_{3D_3}} \\ \sum_{D_2,D_3} f_{D_2|G}t^{w_{2D_2}} f_{D_3|D_2}t^{w_{3D_3}} \\ \sum_{D_2,D_3} f_{D_2|T}t^{w_{2D_2}} f_{D_3|D_2}t^{w_{3D_3}} \end{pmatrix}$$

$$= \boldsymbol{P} \cdot Diag(t^{w_{1A}}, t^{w_{1C}}, t^{w_{1G}}, t^{w_{1T}})$$

$$\cdot \left(\sum_{D_2,D_3} f_{D_2|A}t^{w_{2D_2}} \cdot f_{D_2|D_2}t^{w_{3D_3}}, \ldots, \sum_{D_2,D_3} f_{D_2|T}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}} \right)^T$$

$$= \boldsymbol{P} \cdot M(1,t) \cdot \left(\sum_{D_2,D_3} f_{D_2|A}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}}, \ldots, \sum_{D_2,D_3} f_{D_2|T}t^{w_{2D_2}} \cdot f_{D_3|D_2}t^{w_{3D_3}} \right)^T.$$

Further, applying the above arguments to positions 2 and 3, we have

$$\left(\sum_{D_2,D_3} f_{D_2|A}t^{w_{2D_2}}\cdot f_{D_3|D_2}t^{w_{3D_3}},\ldots,\sum_{D_2,D_3} f_{D_2|T}t^{w_{2D_2}}\cdot f_{D_3|D_2}t^{w_{3D_3}}\right)^T$$

$$=\boldsymbol{P}\cdot Diag(t^{w_{2A}},t^{w_{2C}},t^{w_{2G}},t^{w_{2T}})\cdot\left(\sum_{D_3} f_{D_3|A}t^{w_{3D_3}},\ldots,\sum_{D_3} f_{D_3|T}t^{w_{3D_3}}\right)^T$$

$$=\boldsymbol{P}\cdot Diag(t^{w_{2A}},t^{w_{2C}},t^{w_{2G}},t^{w_{2T}})\cdot\boldsymbol{P}\cdot Diag(t^{w_{2A}},t^{w_{2C}},t^{w_{2G}},t^{w_{2T}})\cdot(1,1,1,1)^T$$

$$-\boldsymbol{P}\cdot M(2,t)\cdot\boldsymbol{P}\cdot M(3,t)\cdot I.$$

Above all, $G(t)=\pi\prod_{i=1}^{p}(\boldsymbol{P}M(i,t))I$.

ACKNOWLEDGMENTS

The work of H.H., X.Z, and W.H.W is supported by NSF grants DBI0196176 and DMS-0090166. The work of H.H. and J.S.L. is supported by NSF grant DMS-0204674 and NIH grant R01 HG02518-01. The work of M.-C.J.K. is supported by the Howard Hughes Medical Institute predoctoral fellowship.

REFERENCES

Bailey, T.L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.

Chen, Q.K., Hertz, G.Z., and Stormo, G.D. 1995. MATRIX SEARCH 1.0: A computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.* 11, 563–566.

Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. 1998. *Biological Sequence Analysis: Probalistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK.

Fickett, J.W. 1996. Coordinate positioning of MEF2 and myogenin binding sites. *Gene* 172, GC19–32.

Fried, M., and Crothers, D.M. 1981. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucl. Acids Res.* 9, 6505–6525.

Galas, D.J., and Schmitz, A. 1978. DNAse footprinting: A simple method for the detection of protein–DNA binding specificity. *Nucl. Acids Res.* 5, 3157–3170.

Garner, M.M., and Revzin, A. 1981. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: Application to components of the *Escherichia coli* lactose operon regulatory system. *Nucl. Acids Res.* 9, 3047–3060.

Gusfield, D. 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press, Cambridge, England.

Gut, A. 1995. *An Intermediate Course in Probability*, Springer-Verlag, New York.

Hertz, G.Z., Hartzell, 3rd, G.W., and Stormo, G.D. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* 6, 81–92.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. 2002. The Ensembl genome database project. *Nucl. Acids Res.* 30, 38–41.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296, 1205–1214.

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., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray,

S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., Szustakowki, J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., and Chen, Y.J. 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

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

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

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

Liu, X.S., Brutlag, D.L., and Liu, J.S. 2002. An algorithm for finding protein DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*

Nakatsuji, Y., Hidaka, K., Tsujino, S., Yamamoto, Y., Mukai, T., Yanagihara, T., Kishimoto, T., and Sakoda, S. 1992. A single MEF-2 site is a major positive regulatory element required for transcription of the muscle-specific subunit of the human phosphoglycerate mutase gene in skeletal and cardiac muscle cells. *Mol. Cell Biol.* 12, 4384–4390.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. 1995. MatInd and MatInspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.* 23, 4878–4884.

Rosenthal, N., Berglund, E.B., Wentworth, B.M., Donoghue, M., Winter, B., Bober, E., Braun, T., and Arnold, H.H. 1990. A highly conserved enhancer downstream of the human MLC1/3 locus is a target for multiple myogenic determination factors. *Nucl. Acids Res.* 18, 6239–6246.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16, 939–945.

Staden, R. 1989. Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.* 5, 89–96.

Stormo, G.D., and Hartzell, 3rd, G.W. 1989. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA* 86, 1183–1187.

Wentworth, B.M., Donoghue, M., Engert, J.C., Berglund, E.B., and Rosenthal, N. 1991. Paired MyoD-binding sites regulate myosin light chain gene expression. *Proc. Natl. Acad. Sci. USA* 88, 1242–1246.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. 2000. TRANSFAC: An integrated system for gene expression regulation. *Nucl. Acids Res.* 28, 316–319.

Address correspondence to:
Jun S. Liu, Wing H. Wong
Department of Statistics
Science Center 6th floor
1 Oxford Street
Cambridge, MA 02138

E-mail: {jliu, wwong}@stat.harvard.edu