Bioinformatics Advance Access published October 26, 2006

Adding sequence context to a Markov background model improves the identification of regulatory elements

Nak-Kyeong Kim, Kannan Tharakaraman and John L. Spouge*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894 USA

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Many computational methods for identifying regulatory elements use a likelihood ratio between motif and background models. Often, the methods use a background model of independent bases. At least two different Markov background models have been proposed with the aim of increasing the accuracy of predicting regulatory elements. Both Markov background models suffer theoretical drawbacks, so this article develops a third, context-dependent Markov background model from fundamental statistical principles.

Results: Datasets containing known regulatory elements in eukaryotes provided a basis for comparing the predictive accuracies of the different background models. Nonparametric statistical tests indicated that Markov models of order 3 constituted a statistically significant improvement over the background model of independent bases. Our model performed slightly better than the previous Markov background models. We also found that for discriminating between the predictive accuracies of competing background models, the correlation coefficient is a more sensitive measure than the performance coefficient.

Availability: Our C++ program is available at: ftp://ftp.ncbi.nih.gov/pub/spouge/papers/archive/AGLAM/2006-07-19.

Contact: spouge@ncbi.nlm.nih.gov

Supplementary information: ftp://ftp.ncbi.nih.gov/pub/spouge/papers/archive/AGLAM/2006-07-19.

1 INTRODUCTION

In the post-genomic era, identifying regulatory elements is an important step to understanding the mechanisms of gene regulation. Regulatory motifs in DNA take considerable time and effort to identify experimentally, so the development of computational tools to discover them is of great interest in bioinformatics. Most computational tools for identifying regulatory elements use one of two methods, alignment or enumeration (Ohler and Niemann, 2001). On one hand, enumerative methods consider all words of a particular length, listing the over-represented words (Marino-Ramirez, *et al*., 2004; Pavesi, *et al*., 2004; Sinha and Tompa, 2002). Enumerative methods have difficulties, however, with word-lengths longer than twelve or with large datasets. On the other hand, alignment methods consider local alignments and maximize a score, usually with expectation maximization (Bailey and Elkan, 1995) or Gibbs sampling (Lawrence, *et al*., 1993; Liu, *et al*., 1995; Tharakaraman, *et al*., 2005). The score usually corresponds to a probability model and often is the sum of entries in a position-specific score matrix (PSSM). In this set-up, the PSSM represents the log-likelihood ratio between a statistical model for a motif and a background model. In the past, background models usually postulated random sequences whose letters were selected independently from a fixed distribution on the nucleotide alphabet {A, C, G, T}. Recently, however, researchers have proposed Markov background models, to extract more empirical information about nucleotide word frequencies. Presently, there are two different Markov background models extant (Liu, *et al*., 2001; Thijs, *et al*., 2001). Here, we examine the two Markov background models and compare them to a third model derived from fundamental statistical principles. We then compare the predictive accuracy of the different background models on a standard test dataset (Tompa, *et al*., 2005).

2 METHODS

The usual statistical approach posits two probability models for nucleotide sequences containing possible embedded motifs: a null model and an alternative model. In our null model, there is no regulatory element; in our alternative, each sequence contains a single element representing a transcription binding factor (TBF) motif. Tools implementing statistical models need not restrict themselves to a single element per sequence, and indeed, many do not (Hughes, *et al*., 2000; Liu, *et al*., 1995; Thijs, *et al*., 2001). Although the restriction to a single element is easily removed, we make it here for convenience and simplicity, principally to be able to compare the different Markov background models under controlled conditions.

Though the two probability models assume multiple sequences as an input, to start, let us consider a single random sequence $\mathbf{A} := (A_1, \ldots, A_n)$, i.e., a single string of length n. Let $\mathbf{a} := (a_1, \ldots, a_n)$ be a realization of the sequence $\mathbf{A}$. In the null model, the sequence $\mathbf{A}$ is generated by a Markov background model of order m. Let $w_k := (a_k, \ldots, a_{k+m-1})$ represent the k-th word of length m (m-letter word) in the realization $\mathbf{a}$. The probability of observing $\mathbf{a}$ is written as:

$$\mathbb{P}_0\{\mathbf{A} = \mathbf{a}\} = \pi(w_1)\prod_{k=1}^{n-m} p(w_k, w_{k+1}),$$

*To whom correspondence should be addressed.

where $\pi(w)$ is the equilibrium probability of the word w, and $p(u,v)$ is the transition probability given the word u to the word v. Naturally, $p(u,v) > 0$ only if the last $m-1$ letters of the word u and the first $m-1$ letters of the word v are the same.

In our alternative model, the sequence $\mathbf{A}$ is generated by a background equilibrium Markov model of order m, except for one random element, a t-word positioned somewhere within the sequence. Let J be the initial position of the random element, with a uniform prior over all its possible positions, before the sequence is known. The alternative model is

$$\mathbb{P}_1^{(2)}\{\mathbf{A}=\mathbf{a}; J=j\} \propto \left\{\pi(w_1)\prod_{k=1}^{j-m-1} p(w_k, w_{k+1})\right\} \times \left\{\prod_{k=1}^{t} q_k(a_{j+k-1})\right\}\left\{\pi(w_{j+t})\prod_{k=j+t}^{n-m} p(w_k, w_{k+1})\right\},$$

where $q_j(a)$ is the probability of finding the letter a at the j-th position within the element. We use the notation "$\mathbb{P}_1^{(2)}$" to differentiate this alternative model from two others, $\mathbb{P}_1^{(0)}$ and $\mathbb{P}_1^{(1)}$ below.

Starting with the probability of the sequence $\mathbf{A}$ containing an element at the j-th position conditional on the observed sequence,

$$\mathbb{P}_1^{(2)}\{J=j \mid \mathbf{A}=\mathbf{a}\} = \frac{\mathbb{P}_1^{(2)}\{\mathbf{A}=\mathbf{a}; J=j\}}{\mathbb{P}_1^{(2)}\{\mathbf{A}=\mathbf{a}\}} \propto \frac{\mathbb{P}_1^{(2)}\{\mathbf{A}=\mathbf{a}; J=j\}}{\mathbb{P}_0\{\mathbf{A}=\mathbf{a}\}} \propto \frac{\pi(w_{j+t})\prod_{k=1}^{t} q_k(a_{j+k-1})}{\prod_{k=j-m}^{j+t-1} p(w_k, w_{k+1})} \quad . \qquad (1)$$

In equation (1), $\mathbb{P}_0\{\mathbf{A}=\mathbf{a}\}$ can replace $\mathbb{P}_1^{(2)}\{\mathbf{A}=\mathbf{a}\}$, because both are constants given the data, the sequence $\{\mathbf{A}=\mathbf{a}\}$. Note: (A) although many factors cancel, equation (1) is derived from a model of the entire sequence $\mathbf{A}$; and (B) the probability in equation (1) depends on both of the m-words neighboring an element. We call the Markov background model in equation (1) the "context-2" model, where "2" represents two-sided dependency on the m-words neighboring an element. As a simple example of equation (1), consider an order-1 Markov model and a motif consisting of a single position 0 embedded within a sequence. Then, the equation (1) becomes

$$\pi(a_1) q_1(a_0) / p(a_{-1}, a_0) p(a_0, a_1) . \qquad (2)$$

In contrast, the equilibrium probability of an isolated m-word representing a putative TBF element is

$$\mathbb{P}_1^{(0)}\{J=j \mid \mathbf{A}=\mathbf{a}\} \propto \frac{\prod_{k=1}^{t} q_k(a_{j+k-1})}{\pi(w_j)\prod_{k=j}^{j+t-m-1} p(w_k, w_{k+1})} . \qquad (3)$$

Equation (3) is essentially the probability (Liu, *et al*., 2001) assign to a putative TBF element under their alternative hypothesis. Note: (A) equation (3) ignores information outside the element and considers only on the words within the element; and (B) the probability in equation (3) does not depend on either m-words neighboring an element. We call the Markov background model in equation (3) the "context-0" model, where "0" represents the absence of dependency on the m-words neighboring an element. As a simple example of equation (3), consider an order-1 Markov model and a motif consisting of a single position 0 embedded within a sequence, as above. Then, the equation (3) becomes

$$q_1(a_0) / \pi(a_0) . \qquad (4)$$

(Thijs, *et al*., 2001) proposed yet another Markov background, which in the present notation can be written

$$\mathbb{P}_1^{(1)}\{J=j \mid \mathbf{A}=\mathbf{a}\} \propto \frac{\prod_{k=1}^{t} q_k(a_{j+k-1})}{\prod_{k=j-m}^{j+t-m-1} p(w_k, w_{k+1})} . \qquad (5)$$

Equation (5) conditions on the m-word to the left of a putative TBF element, but not on the m-word to the right. We call the Markov background model in equation (5) the "context-1" model, where "1" represents the dependency on the single m-word to the left of an element. As a simple example, for an order-1 Markov model and a motif consisting of a single position 0 embedded within a sequence, equation (5) becomes

$$q_1(a_0) / p(a_{-1}, a_0) . \qquad (6)$$

If the Markov order is 0, the three models are identical, but for higher Markov orders, important logical and statistical considerations distinguish the three models. Logically, the statistical analysis using a Markov model should not depend on the arbitrary direction chosen for the Markov transitions. Consider the reverse of the Markov chain considered above. The reverse chain has the same equilibrium probabilities as the forward chain, with its transition probabilities $\tilde{p}(\bullet,\bullet)$ satisfying $\pi(u)p(u,v) = \pi(v)\tilde{p}(v,u)$. Although the probabilities in equations (1) and (3) for the context-2 and -0 models are invariant under reversal, the probability in equation (5) for the context-1 model is not. The context-2 model, moreover, incorporates all available sequence information in accord with standard modeling procedures, including the sequence outside the putative TBF element, whereas the context-0 model does not.

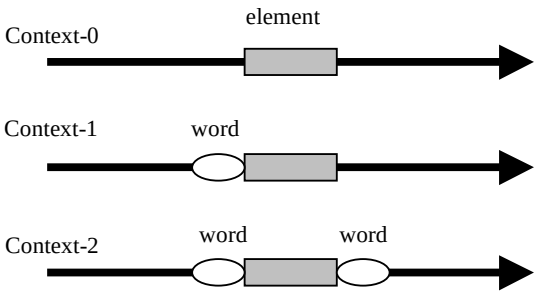

Fig. 1. A comparison of the context-0 model, the context-1 model, and the context-2 model. In the context-0 model, the probability of the element location depends only on the sequence within putative element. In the context-1 model, the probability of the element location is influenced by the word preceding the putative element, as well as the sequence within the element. In the context-2 model, the probability of the element location is influenced by both words neighboring the element, as well as the sequence within the element.

The simple examples in equations (2), (4), and (6) for Markov order 1 show that statistical inferences might differ among the three models described (see Figure 1). We implemented all three context-dependent Markov models inside the A-GLAM ("anchored gapless local alignment of multiple sequences") program (Frith, *et al*., 2004; Tharakaraman, *et al*., 2005). A-GLAM is a tool for finding regulatory motifs, based on a probability model. The probability model represents binding sites as a gapless alignment block, and each column of the alignment is assumed to follow a multinomial distribution over nucleotide alphabet {A, C, G, T}. An alignment that maximizes the likelihood function of the motif is reported as a candidate of binding sites. A-GLAM implements a Gibbs sampling algorithm to optimize the parameters of its probability model.

3 RESULTS

To test the context-dependent Markov models, we used a standard test dataset (Tompa, *et al*., 2005) consisting of human, yeast, mouse, and fly sequences with experimentally verified binding sites. (Tompa, *et al*., 2005) considered three types of data: "real", "generic", and "Markov". We used only real datasets, because the background sequences in the generic and Markov datasets were artificial and therefore not suited to the purpose of testing background models. To avoid special cases that unnecessarily complicate our study, we further removed three datasets having only one sequence or containing ambiguous characters, leaving 49 real datasets.

Transition probabilities in higher-order Markov model have to be estimated from some "background" dataset. Because the transition probabilities may differ from species to species, we constructed 4 background datasets, one for each species, by combining all datasets corresponding to that species. For example, there were 8 datasets from yeast, so the background transition probabilities for yeast were estimated from a combination of the 8 yeast datasets. This study examined order-3 Markov models, because background datasets for all 4 species were large enough to estimate order-3 transition probabilities, but not order 4. We therefore compared the three different Markov background models of order 3 to each other and to the Markov background model of order 0 (independent bases).

To measure the predictive accuracy of A-GLAM using the four different background models, we used the correlation coefficient (Tompa, *et al*., 2005), defined as:

$$CC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \qquad (7)$$

where TP (true positives) is the number of nucleotides in both known sites and predicted sites, FP (false positives) is the number of nucleotides not in known sites but in predicted sites, TN (true negatives) is the number of nucleotides in neither known sites nor predicted sites, and FN (false negatives) is the number of nucleotides in known sites but not in predicted sites. The correlation coefficient is the Pearson correlation coefficient of two binary variables, so it ranges from −1 to +1. The correlation coefficient of value +1 indicates perfectly correct prediction while value −1 indicates perfectly incorrect prediction. If sites were predicted randomly, the correlation coefficient would be 0. For completeness, we also examined another measure, the performance coefficient (Tompa, *et al*., 2005):

$$PC = \frac{TP}{TP + FN + FP}. \qquad (8)$$

We extended the A-GLAM program (Tharakaraman, *et al*., 2005) (written in C++), so it could calculate under each of the four background models. The program also calculates an E-value for each individual site by multiplying the p-value for candidate site by the number of site locations within a sequence (Huang, *et al*., 2004). Although many different tests of the four models are possible, to compare the models on an even and natural footing, we simply ran A-GLAM in its default setting for each model. In particular, the default setting for A-GLAM is a "double-stranded ZOOPS mode" (Zero or One Occurrence Per Sequence), where it searches the two strands corresponding to each input DNA sequence for zero or one instance of a sequence motif. Although arguments can be made for searching for more than one instance per sequence, searching for the "strongest" instance in each sequence provides a particularly straightforward evaluation of the different background models.

Table 1 shows an anecdotal result from a dataset, "hm06r". An alignment from an order-0 Markov model is in Table 1(a), and an alignment from an order-3 context-2 Markov model is in Table 1(b). The sequences in the table were sorted according to their E-values, in the same order as their scores. Because A-GLAM was run in the ZOOPS mode, each sequence was reported as containing zero or one site. For example, A-GLAM reported no site in seq_6 as shown in Table 1(a). The experimentally verified sites are underlined in the tables. An order-3 Markov model in Table 1(b) contains 3 known sites in the alignment while an order-0 Markov model in Table 1(a) contains only one known site. The correlation coefficient between the alignment in Table 1(a) and the known sites is 0.017 while the correlation coefficient between the alignment in Table 1(b) and the known sites is 0.202. Not only did

the order-3 Markov models identify three known sites, but Table 1(b) also shows those three sites received the smallest E-values. By contrast, the known site in Table 1(a) did not have the smallest E-value. For putative elements, the order-0 Markov model provided smaller E-values than the order-3 Markov model, which might mislead practitioners who judge the strength of a given alignment solely by looking at the E-values. Because an order-0 Markov model assigns a spuriously low background probability to repeats, Table 1(a) shows the order-0 Markov model wrongly predicted regulatory elements in C-rich regions. By contrast, the order-3 Markov model assigned larger background probabilities to repeats than the order-0 Markov model. Therefore, the order-3 Markov model successfully avoided predictions in C-rich regions and identified several known sites, as shown in Table 1(b).

Table 2 contains a systematic summary over the test datasets for the four different models. Among 49 datasets, the first 26 datasets are from human, next 8 are from yeast, following 11 are from mouse, and the last 4 are from fly. The table contains the correlation coefficients from the formula (7) along with the ranks of the correlation coefficients for each model within each dataset. Specifically, each row contains a dataset name and 4 correlation coefficients in the white columns and 4 ranks in the gray columns. For example, the dataset "hm05r" has four correlation coefficients: 0.051, -0.039, -0.038, and 0.082. Ranks are assigned according to relative magnitude. Since 0.082 is the largest of all, its rank is 1. The rank of the value 0.051 is 2, and so on. The smaller a rank is, the better the result. Ranks were used to compare the predictive accuracy of competing models, because an analysis based on ranks is robust against an occasional unusual magnitude among the correlation coefficients, thereby retaining fidelity to any overall trend within the data. Correlation coefficients and their ranks for each species are followed by averages. For completeness, Table 2 also lists the average correlation coefficients, but our analysis gives primacy to the comparison with average ranks.

In all four species, and in terms of average ranks, the context-2 model of order 3 predicted most successfully while the order-0 Markov model predicted least successfully. With the single exception of yeast, all species show a trend, with the context-2 model of order 3 predicting best, followed by the context-1 of order 3, the context-0 of order 3, and the Markov model of order 0. Overall rank averages also reflect the same trend (the average ranks being in opposite order to the correlation coefficients). To evaluate statistical significance of our results, let μ represent the average rank of the correlation coefficient corresponding to a particular model. We examined various one-sided hypotheses on μ with the one-sample Wilcoxon test, e.g., $\mu_{order-0} < \mu_{context\_0}$, $\mu_{context\_0} < \mu_{context\_1}$, and $\mu_{context\_1} < \mu_{context\_2}$. The p-value for the test between the order-0 model and the context-0 model is 0.023; between the context-0 and the context-1, 0.338; and between the context-1 and the context-2, 0.002. Thus, the improvement of a Markov model of order 3 over the Markov model of order 0 is statistically significant at the level 0.05, as is the improvement of the context-2 model over the context-1 model. Our test datasets are all from eukaryotes, which might partially explain why these results contrast with a previous report based on datasets from prokaryotes (Hu, *et al.*, 2005). The previous study, however, took the performance coefficient as its measure of prediction accuracy, finding that differences between the predictive accuracy of an order-0 Markov model and higher-order Markov models (for example, order 3) were not obvious.

With this discrepancy in mind, we also evaluated our results using the performance coefficient (defined in equation (8)). Based on the average ranks of the performance coefficient, the context-2 model of order 3 predicted most successfully and the Markov model of order 0 predicted least successfully (see Table 1 in the supplementary material at our ftp site.), as before. The statistical test showed that the context-2 model significantly outperformed other three models with a p-value of 0.034. Differences in the other comparisons were not as dramatic, however, probably because the performance coefficient produced more ties than correlation coefficient.

To check that the comparison of the different background models was stable under incidental perturbations of the test conditions, we increased the number of iteration steps in the Gibbs sampling, to guard against A-GLAM converging prematurely to a local rather than global maximum. Although, e.g., the correlation coefficient of the context-1 model on hm23r (0.2774) degraded to a negative value (-0.051) over longer sampling, results remained generally stable and confirmed the results from A-GLAM's default settings. Most perturbations had similar, occasional, inconsequential effects, with one notable exception. Because the test datasets contain experimentally determined binding sites, all known sites are present in the strands corresponding to sequences in the datasets. Accordingly, we restricted the search space to the strand given in the dataset (as opposed to both the given strand and its complementary strand), to take advantage of strand-specific information that might occasionally be available in a practical situation. The difference between the order-0 Markov model and the order 3 Markov models remained statistically significant (p-value at most 0.024), but differences among the context-0, -1, and -2 models of order 3 were not significant. Note, however, that whereas the results from the order-0 Markov model, the context-0 model, and the context-1 model improved when the search was restricted to a single-stranded search, the result from the context-2 model did not improve. Although an entirely satisfactory explanation is lacking, possibly the context-2 model might have already reached a practical limit, leaving little room for improvement.

4 DISCUSSION

This article presents a context-dependent higher-order Markov model, the "context-2" model. We named the previous versions of the higher-order Markov model "context-0" and "context-1" model (see Figure 1). Intuitively, the context of a putative element should affect statistical inferences, under the following logic. In IUPAC notation (Suzuki, *et al.*, 2001), S denotes a base with strong pairing (G or C); and W, a base with weak pairing (A or T). Consider a hypothetical AT-rich (W-rich) motif embedded in a GC-rich (S-rich) regulatory region. In background GC-rich DNA, the transitions S → W or W → S at the motif boundaries are improbable. Our context-2 Markov model therefore sees the transitions from motif to background as evidence of a TBF element. In contrast, the context-0 Markov model discounts these transitions completely, and the context-1 Markov model ignores the transition on the right. Moreover, the asymmetry in the context-

1 Markov model is somewhat artificial, because logically, the directionality of the Markov model should not affect the analysis. Moreover, although the context-0 model correctly lacks any inherent directionality, it ignores potentially useful sequence information outside a putative regulatory element.

In our hands, computation on the order-3 Markov models required little additional computer time or space, when compared to an independent background model. When A-GLAM was run to find the binding sites in test datasets, order-3 Markov models, e.g., only doubled the computation time of the background model of order 0. We could discern no appreciable difference in computation time for the different context-dependent Markov models.

Other articles implicitly present higher-order Markov background models as improvements over an order-0 Markov model. To the best of our knowledge, however, these articles do not give a controlled comparison specifically testing their Markov "improvement". Here, nonparametric statistical tests evaluated potential improvements associated with higher-order Markov models, using as a "gold standard" 49 test datasets where regulatory elements had been experimentally identified.

The correlation coefficient was our primary measure of the predictive accuracy of competing models. We also compared models with the performance coefficient, but it usually downplayed differences between models. To see why, assume that there are no true positives in a dataset. The corresponding performance coefficient is 0 (see equation (8)), whereas the correlation coefficient decreases from 0 as positives are predicted. For example, consider two hypothetical predictions with correlation coefficients of -0.1 and -0.2. On one hand, a correlation coefficient of -0.1 should be considered better than a correlation coefficient of -0.2, because the corresponding prediction has fewer false positives. On the other hand, both alignments have a performance coefficient of 0. In contrast to claims elsewhere (Hu, *et al*., 2005), we conclude that as a measure of predictive accuracy, the correlation coefficient is more sensitive than the performance coefficient, and therefore superior for distinguishing predictive accuracies.

Table 1. The alignments of "hm06r" produced by A-GLAM.

name	start	Alignment	end	strand	score	E-value
seq_4	288	CCCGCCCCCGGCTTCGCGCCCCGCCCTCCCCGCCCTGCGGCCGCCTCTCCCCGCCCTCCCCGCCC	224	–	43.8	2.36E-21
seq_8	386	GGACCCCGCCCCGTCCCGACCCCTCCCGGGTCCCCGGCCCAGCCCCCTCCGGGCCCTCCCAGCCC	450	+	39.7	1.07E-17
seq_5	267	ACAGCCCGCCCCGGCGCGCCTCGGGTTCCCGACTCCGCCGAGCCCTGGGCCGCTGCTGCCGGCGC	331	+	38.4	1.04E-16
seq_1	365	CCGCCCCGCCCCTGCGCCCTCCTTCCTCGCGTCTGCCCCTCTCCCCCACCCCGCCTTCTCCCTCC	429	+	37.3	6.34E-16
seq_0	392	CCCCTCCGCCCCCTTACACTCTTCGCCCTCCTCCCAGTCGAAGCACCTCCTGTCCGCCCCTCAGC	456	+	33.7	1.32E-13
seq_7	407	ACAGCCAGCCCCTGCGCGCCAGCCCTGGTGGCAGCCGGGAGGACGTCAGCAGGCCCTGCCAGAGC	471	+	33.2	2.62E-13
seq_3	383	GCAGCCCTCCTCCTCCACCTCCTCCTTCTCCTGTGATTGGGAGCAAGCGCGCTCCCAGCTCGCCC	319	–	31.6	2.13E-12
seq_2	232	AGTGTCCGCCGCGTTGAGAACCGCGCACCCTACCCATCGGCCACGTGACCAGTCCTTTTTAAAAA	296	+	29.1	4.47E-11
seq_6		Absent				

(a)

name	start	Alignment	end	strand	score	E-value
seq_0	469	GTCACGTGCCCAGAACGTCCGGCGTTCG	496	+	22.2	8.64E-10
seq_4	6	CCCACGTGGCCAGCACATCGGTCCTCCG	33	+	20.0	4.12E-08
seq_2	424	GTCACGTGGCCAGAAGCTGGCCAATCCG	451	+	19.0	1.96E-07
seq_1	462	GCTCCGAGCGCGGCGCATGCGCCGCTGG	435	-	18.3	5.48E-07
seq_5	109	CCTCCCGCGGCGGAAGCTCGGGCGTCCG	136	+	17.5	1.68E-06
seq_8	230	GGTCCGCCCGGAGCAGCTGCGCTGTCGG	257	+	16.4	7.13E-06
seq_6	284	ATCACCGGGCCAAACCCTTGGCTGTCTA	311	+	14.0	1.21E-04
seq_7	189	GCCGGCCCCCCAGAACAAGCGCCGCCTG	216	+	13.8	1.50E-04
seq_3	232	CCTAGCGCCGGGACGCCTGGGTTGCCTG	205	-	13.2	2.85E-04

(b)

(a) An alignment from an order-0 Markov model. (b) An alignment from order-3 Markov models (the context-1 and the context-2 models produce the same alignment). The underlined segments are known sites. The column "strand" contains either "+" or "–" depending on the site is found in the forward strand or in the reverse strand. The column "score" reports individual score for the site as defined in Tharakaraman *et al*. (2005), and the column "E-value" reports the corresponding E-values. The underlined segments are known sites. Note that the underlined segments for seq_2 are not same in (a) and (b) because this sequence contains multiple sites.

Table 2. A summary table for the Tompa's test datasets

dataset	order 0		order 3					
			context-0		context-1		context-2	
hm01r	-0.017	4	-0.015	2.5	-0.015	2.5	-0.015	1
hm02r	0.156	1	-0.019	3	-0.019	3	-0.019	3
hm03r	-0.028	4	-0.024	2	-0.025	3	-0.020	1
hm04r	-0.010	4	-0.009	3	-0.008	1	-0.008	2
hm05r	0.051	2	-0.039	4	-0.038	3	0.082	1
hm06r	0.017	4	0.182	3	0.202	1.5	0.202	1.5
hm07r	-0.031	4	-0.023	1.5	-0.023	3	-0.023	1.5
hm08r	-0.021	4	0.547	2	0.543	3	0.565	1
hm09r	0.048	3	0.049	2	-0.010	4	0.135	1
hm10r	-0.031	2	-0.046	3.5	-0.046	3.5	0.027	1
hm11r	-0.017	1	-0.024	3.5	-0.024	3.5	-0.023	2
hm12r	-0.103	4	0.149	2.5	0.149	2.5	0.165	1
hm13r	-0.031	4	-0.024	2.5	-0.024	2.5	0.285	1
hm14r	-0.055	4	0.197	1	0.145	3	0.192	2
hm15r	-0.016	1	-0.018	3	-0.018	4	-0.018	2
hm16r	-0.023	4	-0.021	2	-0.020	1	-0.022	3
hm17r	0.606	1	-0.042	4	0.023	3	0.287	2
hm18r	-0.011	4	-0.010	3	-0.010	2	-0.008	1
hm19r	0.128	1	-0.038	4	-0.021	2	-0.036	3
hm20r	-0.034	4	-0.030	3	-0.029	1	-0.030	2
hm21r	-0.031	4	0.164	1.5	0.164	1.5	0.075	3
hm22r	-0.069	4	0.091	1.5	-0.054	3	0.091	1.5
hm23r	-0.078	4	-0.051	2.5	0.277	1	-0.051	2.5
hm24r	0.116	1	-0.041	2	-0.043	3	-0.043	4
hm25r	-0.103	4	0.149	2.5	0.149	2.5	0.165	1
hm26r	0.131	1	0.117	2	-0.023	4	-0.022	3
average	0.021	3.00	0.045	2.58	0.046	2.58	0.074	1.85
vst01r	-0.020	4	-0.015	2	-0.016	3	0.316	1
vst02r	0.597	2	0.592	4	0.597	2	0.597	2
vst03r	-0.046	4	0.062	1	-0.036	3	0.051	2
vst04r	-0.025	4	0.020	1	-0.022	3	-0.020	2
vst05r	0.429	3	0.429	3	0.429	3	0.569	1
vst06r	0.331	3.5	0.355	1	0.331	3.5	0.343	2
vst08r	-0.037	4	0.570	1	0.568	2.5	0.568	2.5
vst09r	0.039	1	-0.015	4	-0.014	2.5	-0.014	2.5
average	0.159	3.19	0.250	2.13	0.230	2.81	0.301	1.88
mus01r	0.077	3	0.183	2	0.245	1	0.074	4
mus02r	-0.028	3.5	-0.028	3.5	-0.021	1.5	-0.021	1.5
mus03r	-0.001	2	0.042	1	-0.048	3.5	-0.048	3.5
mus04r	-0.041	4	0.021	3	0.026	1.5	0.026	1.5
mus05r	-0.050	4	-0.045	3	0.002	1.5	0.002	1.5
mus06r	-0.049	1.5	-0.050	3.5	-0.050	3.5	-0.049	1.5
mus07r	-0.021	4	-0.018	2.5	-0.018	1	-0.018	2.5
mus08r	-0.017	4	-0.016	2.5	-0.016	2.5	-0.016	1
mus09r	-0.083	4	-0.066	3	-0.061	2	-0.058	1
mus11r	0.176	1	0.045	4	0.118	3	0.152	2
mus12r	-0.083	4	-0.061	2.5	-0.061	2.5	-0.059	1
average	-0.011	3.18	0.001	2.77	0.011	2.14	-0.001	1.91
dm01r	-0.029	4	-0.023	1.5	-0.027	3	-0.023	1.5
dm03r	-0.020	4	-0.020	3	-0.020	1	-0.020	2
dm04r	-0.015	1	-0.022	4	-0.018	3	-0.017	2
dm05r	-0.017	4	0.056	3	0.059	2	0.064	1
average	-0.020	3.25	-0.002	2.88	-0.001	2.25	0.001	1.63
overall average	0.033	3.09	0.065	2.57	0.064	2.49	0.088	1.85

The first group consists of 26 datasets from human, the second group consists of 8 datasets from yeast, the third group consists of 11 datasets from mouse, and the last group consists of 4 datasets from fly. Each dataset contains at least two nucleotide sequences and the sites were verified experimentally. The first column contains the dataset names. The white columns in 2nd, 4th, 6th, and 8th contain correlation coefficients. The gray columns in 3rd, 5th, 7th, and 9th contain the ranks of the preceding correlation coefficient in each line. The averages for each species are reported as well as the overall averages at the bottom. The overall average ranks show that the context-2 model produced best alignments.

In our tests, the higher-order Markov models were superior to a background model of independent letters, the superiority of all three order-3 Markov models being statistically significant. The context-2 Markov was consistently superior to the other higher-order Markov models, although the statistical significance of the improvement occasionally disappeared as conditions of the testing varied. Because repeats are often composed of repeated one- or three-letter words, Markov models of order 3 should capture some of their essence. One might therefore expect that part of the superiority of the order-3 Markov models could be due to modeling repeats as part of the background. Our results did not bear that expectation out, however. The superiority of Markov background models was not as pronounced in human (where DNA has abundant repeats) as in some other species (see Table 2).

Statistical tests showed the clear superiority of some models, but all correlation coefficients remained disquietingly low. A-GLAM's performance is comparable to any other motif-finding tool, however (see Table 2 in the supplementary material). The low coefficients therefore suggest at least two possibilities. First, (although perhaps unlikely) any tool presently available might be only slightly more accurate than a random guess. Second, the best standard test dataset available for evaluating motif prediction (Tompa, *et al.*, 2005) might not provide a good "gold standard": its real biological sequences could contain large numbers of unidentified motifs, and experiments might not have accurately identified the boundaries of its known motifs. The fact that it is difficult to quantify the second possibility accurately indicates that the science of finding regulatory motifs is still very incomplete.

ACKNOWLEDGEMENTS

The authors thank Sergey Sheetlin for helpful discussion. This research was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health.

REFERENCES

Bailey, T.L. and Elkan, C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization, *Machine Learning*, 21, 51-83.

Frith, M.C., Hansen, U., Spouge, J.L. and Weng, Z. (2004) Finding functional sequence elements by multiple local alignment, *Nucleic Acids Res*, 32, 189-200.

Hu, J., Li, B. and Kihara, D. (2005) Limitations and potentials of current motif discovery algorithms, *Nucleic Acids Res*, 33, 4899-4913.

Huang, H., Kao, M.C., Zhou, X., Liu, J.S. and Wong, W.H. (2004) Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification, *J Comput Biol*, 11, 1-14.

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.

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

Liu, J.S., Neuwald, A.F. and Lawrence, C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies, *J. Amer. Statistical Assoc.*, 90, 1156-1169.

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.

Marino-Ramirez, L., Spouge, J.L., Kanga, G.C. and Landsman, D. (2004) Statistical analysis of over-represented words in human promoter sequences, *Nucleic Acids Research*, 32, 949-958.

Ohler, U. and Niemann, H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches, *Trends in Genetics*, 17, 56-60.

Pavesi, G., Mereghetti, P., Mauri, G. and Pesole, G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes

An algorithm for finding signals of unknown length in DNA sequences, *Nucleic Acids Res*, 32, W199-203.

Sinha, S. and Tompa, M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation, *Nucleic Acids Res*, 30, 5549-5560.

Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., Suyama, A., Sakaki, Y., Morishita, S., Okubo, K. and Sugano, S. (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes, *Genome Research*, 11, 677-684.

Tharakaraman, K., Marino-Ramirez, L., Sheetlin, S., Landsman, D. and Spouge, J.L. (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements, *Bioinformatics*, 21, I440-I448.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P. and Moreau, Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling, *Bioinformatics*, 17, 1113-1122.

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. and Zhu, Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites, *Nat Biotechnol*, 23, 137-144.