

BEAM: A Beam Search Algorithm for the Identification of Cis-Regulatory Elements in Groups of Genes

JONATHAN M. CARLSON,* ARIJIT CHAKRAVARTY,* and ROBERT H. GROSS

ABSTRACT

The identification of potential protein binding sites (cis-regulatory elements) in the upstream regions of genes is key to understanding the mechanisms that regulate gene expression. To this end, we present a simple, efficient algorithm, BEAM (beam-search enumerative algorithm for motif finding), aimed at the discovery of cis-regulatory elements in the DNA sequences upstream of a related group of genes. This algorithm dramatically limits the search space of expanded sequences, converting the problem from one that is exponential in the length of motifs sought to one that is linear. Unlike sampling algorithms, our algorithm converges and is capable of finding statistically overrepresented motifs with a low failure rate. Further, our algorithm is not dependent on the objective function or the organism used. Limiting the space of candidate motifs enables the algorithm to focus only on those motifs that are most likely to be biologically relevant and enables the algorithm to use direct evaluations of background frequencies instead of resorting to probabilistic estimates. In addition, limiting the space of candidate motifs makes it possible to use computationally expensive objective functions that are able to correctly identify biologically relevant motifs.

Key words: motif finder, bounded search, transcription factor binding sites, coregulated genes, promoter motifs, gene regulation.

INTRODUCTION

MICROARRAYS ENABLE BIOLOGISTS TO OBTAIN a global view of the transcriptional activity of a cell by measuring the mRNA levels of thousands of different genes simultaneously. A major goal of microarray technology is the identification of mechanisms of transcriptional regulation (Lockhart and Winzler, 2000; Xu, 1999). In this context, the computational discovery of DNA binding sites for previously uncharacterized regulatory factors in groups of co-regulated genes assumes much practical significance. Such binding sites provide the biologist with a ready set of targets for mutational analyses.

A frequent first assumption in the *de novo* identification of binding sites for regulatory factors is to equate biological significance with statistical overrepresentation (the tendency of a short DNA sequence or motif to occur more often in a set of upstream sequences than expected given its background frequency

Department of Biology, Dartmouth College, Hanover, NH 03755.

*These authors contributed equally to this paper.

of occurrence in the genome; see appendix). At least fifteen different motif-finding algorithms have been developed over the years (Wolfertstetter *et al.*, 1996; Lawrence and Reilly, 1990; Bailey and Elkan, 1994, 1995; Hughes *et al.*, 2000; Liu *et al.*, 2001a, 2002; Thijs *et al.*, 2001; Jensen and Knudsen, 2000; Vanet *et al.*, 2000; van Helden *et al.*, 1998, 2000; Stormo and Hartzell, 1989; Hertz *et al.*, 1990; Hertz and Stormo, 1999; Brazma *et al.*, 1998; Sinha and Tompa, 2000; Narasimhan *et al.*, 2003). These algorithms fall into two general classes: sequence-driven methods and pattern-driven methods (Brazma *et al.*, 1997; Hudak and McClure, 1999). Sequence-driven methods try to identify conserved motifs by comparing the upstream sequences (for instance, by multiple sequence alignments) and looking for local similarities. These methods are typically based on position weight matrix models (Schneider *et al.*, 1986). Pattern-driven methods, on the other hand, seek to enumerate all possible words in the set of upstream sequences of co-regulated genes in order to identify conserved motifs. These methods employ a consensus model in which the cis-regulatory elements are represented by short words over the IUPAC nucleotide alphabet. It has been suggested that pattern-driven methods are better equipped to find cis-regulatory elements with low internal variation, while sequence-driven models are better equipped to find cis-regulatory elements with high internal variation.

Most pattern-driven methods are based on an exhaustive enumeration of the motifs that are present in a group of upstream sequences derived from the co-regulated genes. This results in the enumeration of many low scoring motifs that are unlikely to expand to a cis-regulatory element, as well as artificially high scoring artifacts that are contained within, or overlap, the actual cis-regulatory elements. In this paper, we present an enumerative algorithm that is based on a pruned breadth-first search, called a *beam search*. A search tree that focuses only on the high scoring submotifs at every length will be likely to identify the biologically relevant motifs while reducing the computational complexity from an exponential to a linear dependency on motif lengths. This effectively removes the motif length restrictions that limit pattern-driven methods.

Pruning the search space provides another potential advantage. While overrepresentation often works well in identifying cis-regulatory elements, many known cis-regulatory elements are not overrepresented in the groups of genes that they regulate. For this reason, there are a number of well-characterized regulons in *S. Cerevisiae* for which none of the current motif-finding programs is able to identify the known cis-regulatory element (Sinha and Tompa, 2003). In these cases, functions that measure positional biases in the occurrences of the motif, coverage of the gene group by a given motif, or even underrepresentation may be better suited to the identification of cis-regulatory elements. Aggressive pruning reduces the running time complexity enough to admit computationally more expensive objective functions that can be used to evaluate the biological relevance of motifs.

Beam search algorithms have been used with success in the field of natural language processing (Ratnaparkhi, 1997). While parsing (also known as part-of-speech tagging) can be implemented using an exhaustively enumerative algorithm, parsers that employ a beam search strategy demonstrate improved performance because the pruning method aggressively eliminates false positives from the list. Analogously, aggressive pruning of the search tree may lead to improved performance in motif-finding, since known *S. Cerevisiae* cis-regulatory elements have lower internal variation and are thus less likely to be eliminated via pruning of the search tree.

In what follows, we present the BEAM algorithm and show that the beam search heuristic is theoretically and practically robust to the various motifs that are encountered in DNA sequences. We then demonstrate the application of BEAM to other complex objective functions.

ALGORITHM

Motif tree

The set M of all possible substrings (motifs) over the DNA alphabet $A = \{a, c, g, t\}$ up to some maximum length h can be organized into a complete, rooted k -ary *motif tree* T of height h , where $k = 4$ (see Fig. 1). Each node m in T corresponds to a motif, and all nodes on the path from the root to are referred to as prefixes of m . A level of the tree at depth w corresponds to all motifs of length w (referred to as w -mers). While traditional pattern-driven algorithms pick a few levels of T and enumerate them in their

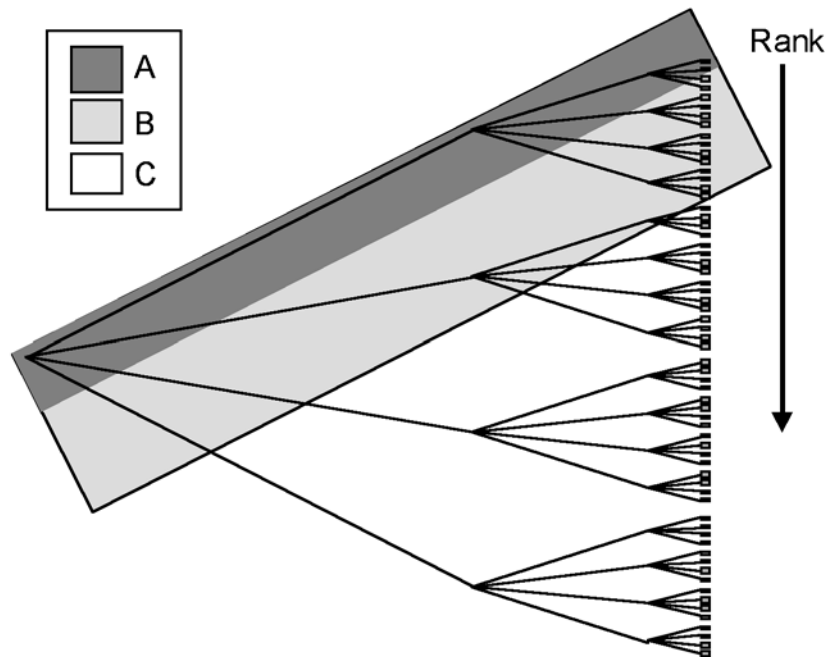


FIG. 1. A beam search tree. The motif tree T showing the region explored by the beam search algorithm. Biologically significant motifs within the zone of significance (set A) are a subset of the motifs contained within the beam width (set B). The majority of the motifs are pruned out of the tree (set C).

entirety, we search T in a pruned breadth-first manner, commonly referred to in the artificial intelligence literature as a *beam search*. This algorithm, called BEAM, is described in Algorithm 1.

Algorithm 1: BEAM. Enumerates only the highest-scoring nodes (motifs) in the motif tree. The score of each motif is computed in the generation process, allowing the nodes to be efficiently returned in sorted order by motif score. The function $\text{expand}(m)$ generates the $|A| = 4$ single-base extensions of m , thereby incrementing $\text{height}(T)$.

Require: objective function $f(m)$ measures a property of motif m (e.g., overrepresentation in some group)

Require: a beam-width β

Require: maximum height h equal to the longest motifs to enumerate

1. create motif tree T with a single node corresponding to the empty string
2. **while** $\text{height}(T) \leq h$ **do**
3. **for all** m in T at depth $h(T)$ **do**
4. compute $f(m)$
5. **for all** β -highest scoring m in T at depth $\text{height}(T)$ **do**
6. $\text{expand}(m)$

The algorithm makes it clear that the objective function $f(\cdot)$ will be computed at most $4\beta h$ times. Since h is the length of the longest motif of interest, we have reduced the identification of the highest scoring motifs from an exponential problem in h to a linear problem in h (see Fig. 1). This allows us to enumerate arbitrarily long motifs and to use significantly more complex objective functions. Furthermore, the heuristic for enumerating T does not depend on the size of a group of co-regulated genes. Note that the objective function f can be any function on m and that by drastically reducing the number of motifs that are evaluated we can relax the practical efficiency requirements of f .

The definition of the beam width β and the fact that we are concerned with only the highest scoring motifs leads to the notion of a *zone of significance*. Let M_w be the set of all motifs at depth w . For brevity, we write m_w to imply $m_w \in M_w$ and m_{w+i} to mean the concatenation of some prefix m_w with some suffix m_i .

We define the sets $B_w \subset M_w$ and $C_w \subset M_w$ for a given objective function f , such that:

$$M_w = B_w + C_w \tag{1}$$

and

$$B_w = \{m : \forall m \in B_w \wedge m' \in C_w, \text{rank}(m) > \text{rank}(m')\}. \tag{2}$$

The set B_w is thus the set of motifs that lie within the beam width $\beta = |B_w|$, and the set C_w is the set of motifs that will be pruned in the transition from level w to $w + 1$. Further, consider A_w , the zone of significance (the top ranking motifs in the set M_w),

$$A_w = \{m : \forall m \in A_w \wedge m' \in B_w - A_w, \text{rank}(m) > \text{rank}(m')\}. \tag{3}$$

These are the motifs that we are interested in, as these are the motifs that are most likely to be cis-regulatory elements. We can now define a *per classification error rate* (PCER) as the rate at which a motif will be erroneously pruned from the tree; that is, the probability that a motif of length $w + i$ will be in A even though the w -length prefix is in C :

$$\text{PCER} = \Pr(m_{w+i} \in A_{w+i} \wedge m_w \in C_w). \tag{4}$$

In general, the error rate of any motif-finding algorithm aimed at the identification of statistically overrepresented motifs is strongly dependent on the assumed background model. Our first objective was thus to gain an accurate empirical understanding of the validity of different background models in modeling yeast cis-regulatory elements in particular, and the upstream regions in general.

Background model for yeast cis-regulatory elements

Biological sequences are commonly modeled using Markov models, in which the probability of occurrence of a base is conditioned on the sequence of k bases that immediately precede it (k is referred to as the order of the Markov model). While other motif-finding programs have reported that low order (usually 1st–4th) Markov models yield a slight performance gain over a Bernoulli (or 0th order Markov) model (Thijs *et al.*, 2001; Liu *et al.*, 2002), it is not clear which orders, if any, more accurately model short motifs sampled from upstream sequences, or whether cis-regulatory motifs follow the same model as randomly sampled motifs at the genomic level.

We sought an empirical understanding of the accuracy of different orders of the Markov model in generating estimates for background frequencies of motifs in the genome. From an information-theoretic standpoint, when a Markov model (or any other stationary ergodic process) is used to model a probability distribution P that generated some data, the goodness-of-fit of the approximation model M is given by the cross-entropy of M on P (Kullback and Liebler, 1951). Intuitively, cross-entropy can be thought of as a measure of the amount of uncertainty that remains after the model has attempted to describe the data. The Shannon–McMillan–Breiman theorem (Breiman, 1957) can be used to provide an estimate for the cross-entropy by taking a large number of motifs sampled from P and computing the average base probability using M :

$$H(P, M) = \lim_{n \rightarrow \infty} \frac{1}{n \times w} \log_2 \prod_{i=1}^n M(m_i) \tag{5}$$

where m_i is a randomly sampled motif from P and w is the average motif length.

The results of a cross-entropy-based goodness-of-fit test are usually reported as perplexity, defined as $2^{H(P, M)}$, which can be roughly thought of as the number of viable options that remain after the model has been taken into account. Thus, a naive maximum entropy model, in which all four bases are equally likely, yields a perplexity of 4. This provides a simple way to compare models: given models M_1 and M_2 and a set of motifs M , the more accurate model for M is the one that generates a lower estimate for the perplexity.

We calculated the perplexity of Markov models of various orders against two motif distributions: known cis-regulatory elements (P_{CRE}) and motifs sampled randomly from upstream sequences (P_{rand} ; Fig. 2).

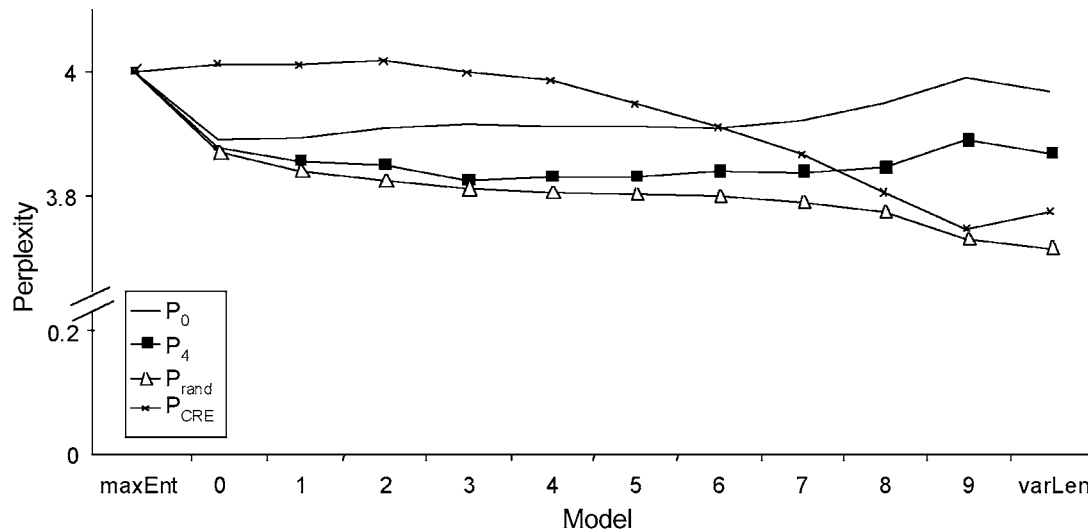


FIG. 2. Accuracy of motif representation models. Accuracy of models against motifs drawn from a database of known cis-regulatory elements, motifs sampled from upstream sequences, and motifs generated from either a 0th or 4th order Markov model that was drawn from the upstream sequences. The models tested were the maximum entropy model (maxEnt), Markov models of order 0 through 9, and the variable length Markov model (VarLen).

The perplexities provide us both with a measure of the accuracy of the various models and a measure of the similarity between P_{CRE} and P_{rand} . Twelve models were constructed: a maximum entropy model, in which each base is equally likely, 10 Markov models drawn from the set of all upstream regions (immediately 5' to the transcriptional start site of a gene) and ranging in order from 0 to 9, and a variable-length Markov model, which is equivalent to maximum likelihood estimation (MLE) generated by directly looking up the frequency of occurrence of each motif in all upstream regions. The P_{CRE} was estimated by randomly sampling 200 nondegenerate yeast cis-regulatory elements from the yeast transcription factor database TRANSFAC (Wingender *et al.*, 1996). These motifs ranged in length from 4 bp to 26 bp, with a median of 9 bp. For each selected cis-regulatory element, a motif of the same length was randomly sampled from the set of all upstream sequences. As a control, motifs of matching lengths were generated using 0th and 4th order Markov models drawn from the upstream sequences to represent P_0 and P_4 , respectively.

As expected, the motifs drawn from P_0 have a minimum perplexity when modeled by a 0th order Markov model, and the perplexity climbs steadily as the order increases (Fig. 2, unmarked line). Similarly, the perplexities on P_4 indicate that motifs generated from a 4th order Markov model are best modeled by a 4th order Markov model (Fig. 2, black squares). Remarkably, the perplexity for cis-regulatory elements modeled by low order Markov models generated from the upstream regions as a whole is greater than that of the naive MaxEnt model (Fig. 2, line with crosses). This suggests that P_{CRE} and P_{rand} are distinct distributions, even at the genomic level. Furthermore, an attempt to model an unknown motif with a low order Markov model includes an unknown amount of error that will depend on whether the motif is a cis-regulatory element—the very question that motif-finding programs seek to answer. The perplexities also indicate that both cis-regulatory elements and random motifs are most accurately modeled by high order Markov models, which are closely approximated by a variable order Markov model. Of particular interest is the fact that a 9th order Markov model drawn from upstream sequences models P_{CRE} and P_{rand} equally well, and that a variable length Markov model employing direct look-ups is a good approximation to the unwieldy 9th order Markov model. Thus, the probability that a motif m will be randomly drawn can be accurately approximated by $\Pr(M_{VL} = m)$, defined as

$$p_{VL}(m_w) = \frac{C(m_w)}{N(w)} \quad (6)$$

where $C(m_w)$ is the number of occurrences of m_w in all upstream regions and $N(w)$ is the number of instances of any motif of length w . Moreover, $p_{VL}(m_w)$ is accurate regardless of whether m_w came from P_{CRE} or P_{rand} .

Pruning the motif tree

Using $f(\cdot)$ of a prefix to prune the search tree corresponds to using a variable length Markov model. BEAM works because of the predictive value of high scoring prefixes in the context of identifying high scoring motifs. In terms of a variable-length Markov model, this predictive value is dependent on the correlation between the specific prefixes of Markov models of different length (that is, the correlation between $f(m_w)$ and $f(m_{w+i})$). This correlation can be defined as follows: for two Markov models of order k and $k + 1$, if the highest scoring motifs in M_{k+2} contain within them the highest scoring motifs in M_{k+1} , then the two Markov models of order k and $k + 1$ are said to be positively correlated (note that Markov models of length k directly model sequences of length $k + 1$).

In theory, Markov models of order k and $k + 1$ may be positively correlated, negatively correlated, or uncorrelated. In practice, because of specific properties of cis-regulatory elements, the variable length Markov models that encode them are very unlikely to be negatively correlated. Successive orders of Markov models that are negatively correlated would lead to spurious transcription factor binding sites that are overrepresented, while biological evidence suggests that spurious transcription factor binding sites are selected against (Hahn *et al.*, 2003). In the specific context of biological sequences, cis-regulatory elements with low internal variation are very likely to correspond to consecutive Markov models that are positively correlated with one another. On the other hand, cis-regulatory elements with high internal variation are likely to correspond to consecutive Markov models that are uncorrelated with each other. In the following sections, we show that BEAM works when the variable length Markov models are uncorrelated or positively correlated.

Motifs with high internal variation

Since the binding site for a regulatory factor is typically not a unique word, but a set of related words, it is possible for the binding site to allow considerable variation in terms of the bases allowed at every position. Such motifs with high internal variation represent a more challenging case for BEAM, and we will consider them first.

In the course of expanding the motif tree, BEAM generates motifs of length $w + 1$ that have the highest scoring w -length prefixes, as computed by the variable length Markov model of order $w - 1$. Since consecutive order Markov models are uncorrelated, the contribution of every single-base extension to the final objective score of the motif is linearly independent. In this case, the objective score of the motif as a whole will be dependent on the objective scores of its constituent single base extensions in the context of their respective Markov models of length $w - 1$ and can be described by some function F :

$$f(m_{w+i}) = F(g(b_1), g(b_1b_2), \dots, g(b_1b_2 \dots b_{w+i})) \tag{7}$$

where b_j is the j th base in m_{w+i} and $g(b_1b_2 \dots b_j)$ is some function proportional to the biological relevance (significance) of b_j in the context of $b_1 \dots b_{j-1}$.

Suppose $g(\cdot)$ is a binary function, where $g(b_1b_2 \dots b_j) = 1$ if b_j is significant, and 0 otherwise. Furthermore, suppose that for a given motif m_{w+i} , $G(m_{w+i}) = \sum g(\cdot)$, and $f(m_{w+i}) = \Pr(G \geq G(m_{w+i}))$ is the probability of seeing at least $k = G(m_{w+i})$ significant extensions given the probability $p = \Pr(g(\cdot) = 1)$. This is given by the binomial cumulative distribution function:

$$\begin{aligned} f(m_{w+i}) &= 1 - \text{binomcdf}(k, i, p) \\ &= 1 - \sum_{j=k}^i \binom{i}{j} p^j (1-p)^{i-j} \end{aligned} \tag{8}$$

We are interested in calculating the probability of obtaining an objective score of a suffix that is sufficiently high that the full motif of length $w + i$ belongs to A_{w+i} even though the w -length prefix lies in C_w (and

hence will be pruned from the tree). Let $G(C_w^{\max}) = c$ be the maximum number of significant extensions for a motif to be in C_w . Similarly, let $G(A_{w+i}^{\min}) = a$ be the minimum number of significant extensions for a motif to be in A_{w+i} . Then, the probability of the prefix of length w lying in C is

$$\Pr(m_w \in C_w) = 1 - \text{binomcdf}(c, w, p) \quad (9)$$

and the probability of the motif lying in A_{w+i} is

$$\Pr(m_{w+i} \in A_{w+i}) = 1 - \text{binomcdf}(a, w + i, p). \quad (10)$$

We can then calculate the PCER as the probability of obtaining at least $a - c$ significant single-base extensions in the i -length suffix using

$$\text{PCER} = \Pr(m_{w+i} \in A_{w+i} \wedge m_w \in C_w) \quad (11)$$

$$= 1 - \text{binomcdf}(a - c, i, p) \quad (12)$$

which is the probability of having an i -length suffix with $a - c$ significant extensions. The binomial distribution has tails that taper rapidly, so the areas under the binomial curve can be used to set bounds for the error rate. As long as the beam width β is kept reasonably large with respect to the size of A , the error rates can be kept low. For example, for $p = 0.05$, $w = 5$, $i = 4$, $G(A_{w+i}^{\max}) = 4$, and $G(C_w^{\min}) = 1$, $\text{PCER} = 6.25 \times 10^{-6}$. That is, in looking for a motif of length 9, pruning motifs at length 5 results in a worst case error rate of 6.25×10^{-6} . This is equivalent to pruning all but the top 23 out of 1,024 motifs at length 5 and still being able to accurately identify the top 9 out of 262,000 motifs. This is a loose upper bound, as it estimates the probability of the highest scoring motif in C_w moving to the lowest scoring motif in A_{w+i} . Further, an objective function $f(\cdot)$ that correlates with biological significance is unlikely to be binary. Even so, the simple example outlined above makes clear the relevance of the kurtosis of the binomial distribution in setting stringent error bounds for pruning the motif tree.

Motifs with low internal variation

A substantial number of the yeast transcription factors characterized so far have been shown to bind short DNA sequences that are almost identical to each other. Such sequences are referred to here as having low internal variation (Zhu and Zhang, 1999). As shown in the earlier section, bona fide yeast cis-regulatory elements come from different distributions than randomly sampled upstream motifs. Taken together, these observations suggest that yeast cis-regulatory elements can be thought of as being planted intact by an outside process into the genomic background. In this case, consecutive orders of Markov models are highly likely to be positively correlated. That is, a highly overrepresented transcription factor binding site m will contain within it nested sub-motifs that are themselves highly overrepresented, since all the submotifs of m will have their underlying distributions altered by the planting process. In particular, each prefix m' of m will have its score raised, with the effect increasing as m' approaches m in length. We refer to this as the *nested motifs property*.

This nested motifs property can be demonstrated empirically by tracing out all the possible paths that arise during the expansion of all 256 4-mers in a dataset where five motifs have been planted at high levels of overrepresentation (Fig. 3). The objective function used here is the Sig value, which is a statistical measure of the degree of overrepresentation (see appendix). In this example, motifs corresponding to prefixes for the planted motifs rapidly rise above the background Sig values. At each motif length w , the highest scoring w -mers come from the highest scoring $(w - 1)$ -mers. As soon as the expansions pass the length of the planted motifs, the scores suddenly drop. This result is consistent with that of Van Helden *et al.* (1998), who traced out the significance level of the paths consisting of prefixes of various planted motifs. They found that every path had an optimal significance score which was associated with a motif of the same length as the planted motif.

In order to further investigate the degree of correlation between consecutive Markov model orders in the case of motifs with low internal variation, the following experiment was performed. Five randomly selected motifs from the TRANSFAC database were planted in each of five randomly selected genes.

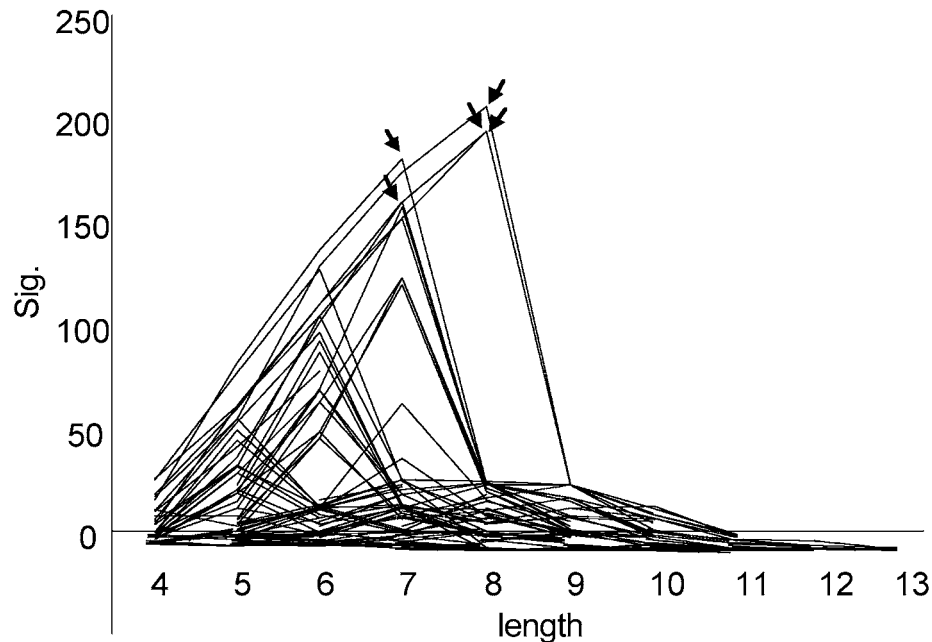


FIG. 3. High scoring motifs are derived from shorter high scoring motifs. The objective function scores for prefixes of planted motifs rise sharply above the background as these prefixes are expanded and continue to rise until these prefixes are expanded into their corresponding planted motifs, falling rapidly after such a point. (*Sig*: significance value, see appendix).

BEAM was run on this test set using a beam width of $\beta = 1,000$. For various levels of the tree generated in this process, the fraction of w -mers that contain as prefixes any one of the top 1,000 $(w - 1)$ -mers was measured directly (Fig. 4). Close to 100% of the top 10 motifs at any length contain a prefix from the top 1,000 motifs of length $w - 1$. For the top 50 or 100 motifs of any given length, more than 90% of the motifs at any length contain a prefix from the top 1,000 motifs of length $w - 1$. In contrast, only 40% of the top 1,000 motifs of length 9 contain within them a prefix from the top 1,000 motifs of length 8. This shows that, for a beam width of 1,000, the turnover between C_w and B_{w+i} occurs largely at the boundary, and the top of the ranked list of motifs at every length w is remarkably stable. Thus, for a large enough size of β relative to α (the size of the zone of significance), the error rate can be made extremely small. When consecutive orders of Markov models are strongly positively correlated with each other, the binomial process outlined in the previous section becomes an upper bound on the error rate of the BEAM algorithm, since single base extensions that make a substantial contribution to the objective score are more likely to be associated with high scoring single-base extensions further along the same expansion path. Since yeast cis-regulatory elements tend to follow the nested motifs property, those motifs that exist in A but have prefixes in C tend to be false positives, because the significance of the prefix and suffix are sufficiently different to imply a lack of positive correlation. This is consistent with the findings in the natural language processing literature and supplies an added benefit for setting aggressively low beam widths.

RESULTS

Synthetic data (low internal variation)

This section deals with the validation of the underlying statistical framework of the algorithm using planted datasets. The planted datasets were built based on a motif model derived from experimentally verified yeast motifs listed in the *S. Cerevisiae* TRANSFAC database (Wingender *et al.*, 1996). Random groups of *S. Cerevisiae* genes were selected, and motifs from the TRANSFAC database were planted in these groups. A full description of the motif model and parameter ranges can be found in the appendix.

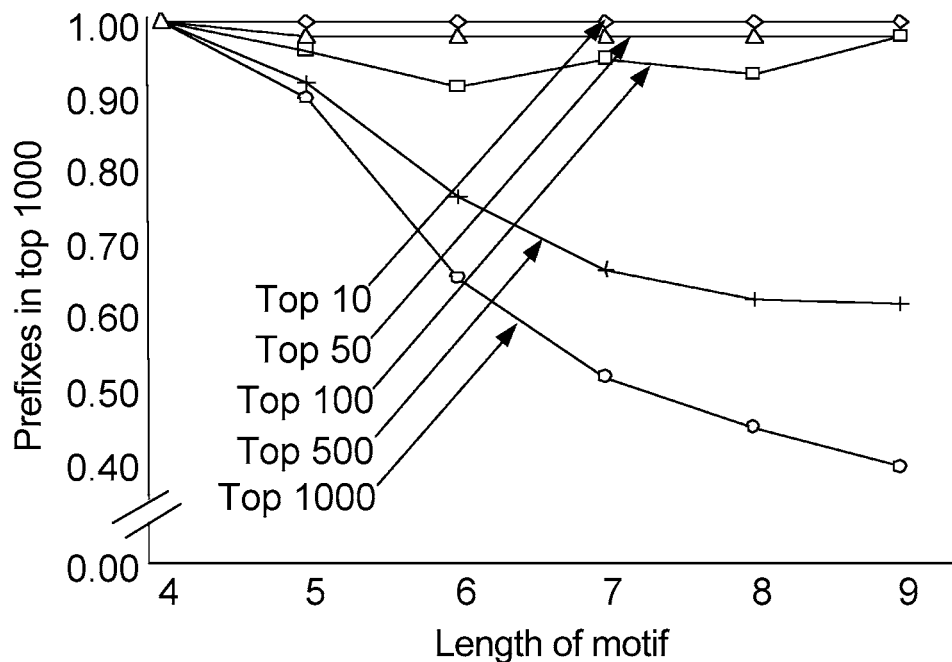


FIG. 4. Effect of beam width on motif selection. The proportion of the top $\alpha = \{10, 50, 100, 500, 1000\}$ motifs of length w in the list whose prefixes ranked in the top 1,000 motifs of length $w - 1$.

In the experiments that follow, the performance of our beam search algorithm was benchmarked by comparing its output to the motifs identified by exhaustive enumeration of all possible motifs. The planted datasets contained motifs that were highly statistically overrepresented. Ten groups of genes were generated, each with between 5 and 20 randomly selected genes. The entire space of motifs of length 4 to 10 was enumerated. The top scoring motifs from all A_w , $4 \leq w \leq 10$, were compiled into a set called A^* . For each motif m in A^* , all the prefixes were identified and the per classification error rate was calculated based on the number of prefixes that were in C . Table 1 shows the results for different values of $|A^*|$ (referred to in the table as the cutoff) and $|B_w|$, the beam width. The test was repeated for a set of random genes with no planted motifs, a set with a motif planted at a highly significant value, and a set of genes known to be co-regulated in yeast. While the error rates vary on the strength of the highest scoring motif, the 50 highest scoring motifs of all motifs ranging from 4 to 10 base pairs long can be identified without errors using a beam width of 1,000. This is a significant reduction in the search space, as 21,000 motifs are now sufficient to search a space of 1.4×10^6 motifs with no loss in accuracy (a 50-fold reduction in the size of the search space). In a biologically relevant set of genes, the space can be pruned even more aggressively. In the presence of a strongly overrepresented signal, the beam width can be reduced to as low as 10 or 50 motifs in some cases with no loss of accuracy relative to a full enumeration of the motif search tree. Further, a comparison of the accuracy of the algorithm at a beam width of 500 between random groups of genes and gene groups with planted motifs (Table 1) shows that BEAM performs better on motifs with positive correlation between prefixes (planted groups) than on motifs with no correlation between prefixes (random groups), though the performance on both groups is strong.

The performance of the algorithm was then tested with a wide range of beam widths, using 15 different datasets consisting of groups of random genes with motifs from the TRANSFAC database planted in them. The results are shown in Fig. 5. The performance was assessed in terms of false positive (FP) and false negative (FN) rates, where a false positive is defined as a motif that was not planted, but that appears in the top q motifs returned by BEAM, where q motifs were planted in the dataset. A false negative, on the other hand, is defined to occur when a planted motif does not appear in the top q motifs returned by BEAM. In terms of FP and FN rates, the algorithm is robust to broad changes in the beam width. Beam widths of 1,000 and above show no statistically significant differences in the FP and FN rates relative

TABLE 1. EFFECT OF BEAM PARAMETERS ON PER CLASSIFICATION ERROR RATES^a

		10	20	50
A: Random group	10	7.0×10^{-4}	1.0×10^{-3}	3.4×10^{-3}
	50	4.0×10^{-4}	5.0×10^{-4}	2.0×10^{-3}
	100	1.0×10^{-4}	5.0×10^{-4}	1.0×10^{-3}
	500	1.0×10^{-4}	0.0	1.0×10^{-4}
	1000	0.0	0.0	0.0
B: Planted motifs	10	4.6×10^{-5}	2.8×10^{-4}	1.7×10^{-3}
	50	0.0	0.0	1.4×10^{-4}
	100	0.0	0.0	4.7×10^{-5}
	500	0.0	0.0	0.0
	1000	0.0	0.0	0.0
C: SCB regulon	10	4.1×10^{-4}	8.3×10^{-4}	1.7×10^{-3}
	50	1.4×10^{-4}	2.8×10^{-4}	7.9×10^{-4}
	100	4.7×10^{-5}	9.4×10^{-5}	3.3×10^{-4}
	500	0.0	0.0	0.0
	1000	0.0	0.0	0.0
D: MCB regulon	10	9.2×10^{-5}	6.4×10^{-4}	1.7×10^{-3}
	50	0.0	3.7×10^{-4}	1.0×10^{-3}
	100	0.0	1.9×10^{-4}	7.5×10^{-4}
	500	0.0	0.0	1.0×10^{-4}
	1000	0.0	0.0	0.0

^aColumns represent different sizes of the zone of significance (zone A), while rows represent different sizes of beam width B_w . A: PCER for a random group of genes. B: PCER for a group of genes containing a single motif planted at a *Sig* value of 120. C: PCER for the *S. Cerevisiae* SCB regulon listed in the SCPD database. D: PCER for the MCB regulon.

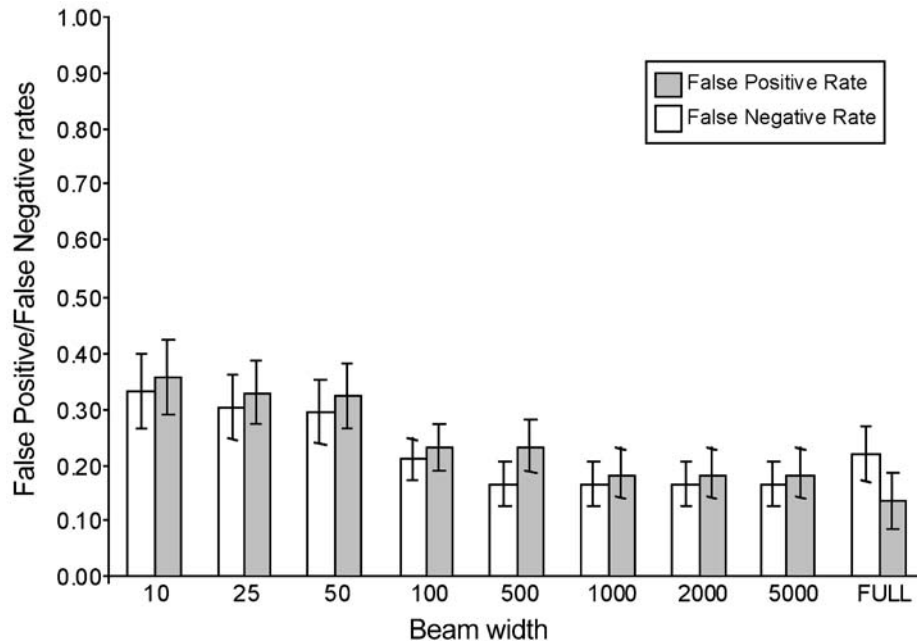


FIG. 5. Accuracy of BEAM. The relationship between the false positive and false negative rates of BEAM and the selected beam width.

to a full enumeration. These results are consistent with the results in the per classification error rates experiment.

Known transcription factor binding sites (low and high internal variation)

A per classification error rate analysis was performed using sets of yeast genes (regulons) that have been determined to be coregulated at the level of transcription and that contain an experimentally determined common transcription factor binding site (Zhu and Zhang, 1999). The results were comparable to those obtained with planted datasets. Table 1c shows the classification error rate for a cis-regulatory element with low internal variation, while Table 1d shows the classification error rate for a known cis-regulatory element having high internal variation. In both cases, the results suggest that a beam width of 100 is sufficient to find these motifs with accuracy comparable to a full enumerative search. In both cases, the motifs are significantly statistically overrepresented (ranked 1st and 4th in the total list of motifs).

We then tested the overall accuracy of the motifs returned by BEAM, running the algorithm on *S. Cerevisiae* regulons. We assessed the performance of BEAM using the Φ score (Pevzner and Sze [2001]; see appendix, section Phi Scores, for details). This metric has been employed in a number of published comparisons of motif-finding algorithms (Pevzner and Sze, 2000; Sinha and Tompa, 2001, Shinozake *et al.*, 2003). Each of the top three motifs was evaluated in terms of its Φ score, and the best score was reported. When the biologically relevant motifs are over-represented, such as in the RAP1 and REB1 regulons, BEAM run with an overrepresentation objective function returns Φ scores that are considerably higher than other motif finding programs (see Table 2).

Alternative objective functions

A further consideration in motif-finding is that some biologically relevant motifs are not statistically overrepresented. Working from the hypothesis that the biologically relevant motifs must differ from the surrounding upstream sequences in some way in order for them to be biologically active, we tried two other objective functions on a number of different yeast regulons from the SCPD database for which other motif finding programs had failed to find the biologically relevant motifs. The objective functions were the Kolmogorov–Smirnov (K–S) statistic, which measures differences in the positional bias in the occurrence of a motif between the selected group of genes and the genome as a whole (see appendix), and statistical underrepresentation (see appendix). In the case of the K–S statistic, for two regulons (BAS1, and TBP, whose reported motif is the TATA box), BEAM was able to find motifs with a 0.24 and 0.37 Φ score, respectively. This is in comparison to the performance of three other motif finding programs, AlignACE, YMF, and MEME (reported in Sinha and Tompa [2003]), all of which returned Φ scores of between 0 and 0.03 for these two motifs. This result is encouraging and is consistent with the finding that the TATA box is strongly positionally restricted. By narrowing its focus to the most promising candidate motifs, BEAM is able to identify biologically characterized motifs using a computationally expensive objective function.

TABLE 2. PERFORMANCE COMPARISON OF BEAM TO OTHER MOTIF FINDERS^a

<i>Regulon</i>	<i>Sig</i>	<i>UR</i>	<i>KS</i>	<i>SO</i>	<i>YMF</i>	<i>MEME</i>	<i>AlignACE</i>
HAP2	−12.5	0.38	0.00	0.04	0.00	0.00	0.02
PHO2	−17.3	0.18	0.00	0.00	0.00	0.00	0.00
BAS1	−12.0	0.00	0.24	0.00	0.02	0.03	0.02
TBP	−11.3	0.00	0.37	0.00	0.00	0.00	0.00
RAP1	8.9	0.00	0.00	0.52	0.09	0.31	0.23
REB1	8.2	0.00	0.00	0.75	0.39	0.34	0.01

^aPerformance of BEAM and other motif finding programs on six real regulons from the SCPD database. Column headings: *Sig*, significance value of known biologically relevant motif; *UR*, BEAM run with underrepresentation objective function; *KS*, BEAM run with KS objective function; *SO*, BEAM run with statistical overrepresentation fitness function; *YMF*, *MEME*, *AlignACE*, scores reported in Sinha and Tompa (2000) for these regulons.

In addition, we obtained the surprising and somewhat counterintuitive result that when BEAM is run with an underrepresentation objective function, it is able to find the biologically relevant motif in regulons where none of the popular motif finding programs were successful. In the case of the HAP2 and PHO2 regulons from the SCPD database, BEAM run with an underrepresentation objective function returns motifs with Φ scores of 0.38 and 0.18, respectively. Again, these Φ scores represent a significant improvement over the Φ scores reported by other programs in an earlier study, which range from 0 to 0.04.

CONCLUSIONS

BEAM is an efficient, generalized motif finding algorithm that, with low probability of error, will find the most significant motifs of any length given any biologically relevant, pattern-based objective function. The success and flexibility of our algorithm is related to the manner in which it exploits the properties of variable length Markov models. Since BEAM is capable of working under both positively correlated and uncorrelated Markov models, it is able to accurately identify motifs with both low and high internal variation. BEAM is able to identify only unambiguous instantiations of such motifs. However, in a separate study, we demonstrate that the BEAM framework can successfully be extended to return ambiguous motifs (Carlson *et al.*, in press).

Aggressively limiting the search space not only increases the efficacy of the algorithm, it also reduces the asymptotic running time, transforming a problem that has traditionally been exponential in the length of motifs searched to a problem that is linear. This effectively removes the length restrictions that are common in pattern-driven methods and allows us to use objective functions that are much more complex than those that are traditionally used (though the specific objective function may increase the run-time complexity).

For example, we have used the reduced search space that BEAM exploits to develop a more accurate overrepresentation objective function that uses direct lookups and variable length Markov models to estimate expected motif frequencies rather than the low order Markov models that are traditionally used. Our results suggest that low order Markov Models are actually poor estimators for genomic motif frequencies—especially when the motif in question is a cis-regulatory element.

Most motif finding programs (with AlignAce a notable exception) are focused on the identification of motifs that are common to, or statistically overrepresented within, a group of genes. Here we have shown that BEAM is capable of correctly identifying cis-regulatory elements using objective functions other than statistical overrepresentation, such as positional bias and statistical underrepresentation. To the best of our knowledge, these cis-regulatory elements found by BEAM's alternative objective functions cannot be found by any other motif finding program. It is possible that a complex objective function that combines several such measures may prove to be a good model of biological significance. The small search space of BEAM, coupled with a suffix array that allows rapid lookup of every position of every motif, is ideally suited for such complex objective functions.

We have demonstrated that BEAM can find the highest scoring motifs given any pattern-based objective function. This algorithm can be easily applied to any organismal background to find cis-regulatory elements. Our work is currently focused on identifying objective functions that better differentiate between cis-regulatory elements and random motifs. We hope this process will provide us with novel biological insights into the mechanism of transcriptional regulation in eukaryotes.

APPENDIX

Overrepresentation objective function

The most common objective functions used in motif prediction are based on overrepresentation. Generally, a motif m is overrepresented in a group of upstream sequences if it occurs more often in the group than is expected given some background distribution of motifs. This notion is made concrete in a Bayesian context: Given a set C of the upstream regions of a set of (presumably co-regulated) genes and some

background distribution of motifs described by the random variable M , the likelihood of seeing at least $N_C(m)$ occurrences of m in C given M is

$$L(m) = \Pr(n \geq N_C(m) \mid M = m) \quad (13a)$$

$$= \sum_{k=N_C}^N \binom{N}{k} p^k (1-p)^{N-k} \quad (13b)$$

$$\approx \sum_{k=N_C}^N \frac{\lambda^k e^{-\lambda}}{k!} \quad (13c)$$

for $\lambda = Np$ and $p = \Pr(M = m)$. Equations (13b) and (13c) are simply the binomial and Poisson distributions, respectively.

The function $L(m)$ is thus a probability that provides a natural way to compare two motifs by their degrees of overrepresentation and suggests that a motif can be considered *significantly overrepresented* if $L(m) \leq \alpha$, where $\alpha = 0.05$ is a customary cutoff. However, in the context of an enumerative approach to motif identification, $L(m)$ will be computed for all motifs of a given length w . Thus, α must be adjusted to account for the number of motifs that will be scored. This is typically done using a Bonferroni correction factor (for example, see Van Helden *et al.* [2000]), defined as

$$B_w = \frac{\alpha}{N_w} \quad (14)$$

where N_w is the number of w -mers that are enumerated, given by

$$N_w = 4^w - r^{\lfloor \frac{w}{2} \rfloor}. \quad (15)$$

The relative significance of motifs of varying length can thus be computed by converting $L(m)$ into a negative log expectation using N_w :

$$f_{\text{over}}(m_w) = -\log(L(m_w) \times N_w). \quad (16)$$

The Bonferroni correction was first applied to motif finding by Van Helden *et al.* (2000) and is equivalent to using an unnormalized prior in the Bayesian approach. Following Van Helden, we refer to $f_{\text{over}}(m_w)$ as the *Sig* value of m_w . Note that *Sig* is 0 when exactly one motif is expected to have a given $L(m)$ value and is positive for overrepresented motifs. Most importantly, the *Sig* values of motifs of very different lengths are immediately comparable.

Underrepresentation objective function

The overrepresentation function can be trivially changed to underrepresentation by defining $f_{\text{under}}(m) = -f_{\text{over}}(m)$. Of course, this will converge on motifs that do not exist at all in the set of genes. To avoid this problem, we add the additional requirement that a motif exist in at least c_{min} of the total genes in the group. The fraction c_{min} was set to 0.5 on an empirical basis.

Kolmogorov–Smirnov objective function

The Kolmogorov–Smirnov (K–S) statistic is a nonparametric test that measures the probability that two sample distributions are drawn from the same distribution. Let X be the sample distribution that we wish to compare to some reference distribution Y . The two-tailed K–S statistic is defined to be the maximum absolute difference between the unbiased cumulative distribution functions of X and Y . The K–S statistic has a well defined distribution from which a p -value can be easily computed (see Press *et al.* [2002]). An objective function was created from this statistic by taking the distribution X of motif m in the set of upstream sequences from a group of co-regulated genes and comparing it against the distribution Y of m in the upstream sequences of all genes in the genome, such that each value $x \in X$ and $y \in Y$ was a negative integer in the range -800 to -1 specifying the position of m relative to the start of the associated

gene. Thus, $f_{KS}(m)$ is a measure of how m is localized *differently* in the set of co-regulated genes than in the genome as a whole.

Phi scores

The Φ score, first proposed by Pevzner and Sze (2000), measures the degree of overlap between the actual instances of two motifs m_1 and m_2 in the set of co-regulated upstream sequences. The Φ score can be defined as follows: let U be a unique numbering of all the bases in the upstream sequences of a given set, and $I_U(m) \subseteq U$ be the set of bases that are covered by actual instances of m in U . Then Φ is defined as

$$\Phi_U(m_1, m_2) = \frac{I_U(m_1) \cap I_U(m_2)}{I_U(m_1) \cup I_U(m_2)}. \quad (17)$$

On the one hand, the Φ score is a very stringent metric. If, for example, $m_1 = TGCCGA$ and $m_2 = TGCCGAT$, we would expect that only 25% of the instances of m_1 would be followed by a T giving $\Phi(m_1, m_2) = 0.25$ despite the high degree of textual overlap (m_1 contains 6 out of the 7 bases in m_2). On the other hand, the Φ score provides an objective, biologically relevant method for defining the similarity between two highly degenerate motifs that cannot easily be visually compared.

Implementation of look-ups

Calculating variable length Markov models and motif distributions for the K-S statistic requires a data structure that permits the efficient look-up of every position of any substring. A suffix array is ideally suited for this task. First described by Manber and Myers (1993), the suffix array is a space and time efficient data structure for the identification of any pattern P of length w over a fixed string S of length L . As the name implies, it is an array of integers in the range 1 to L specifying the lexicographic order of the L suffixes of S (for precise definitions and a review, see Gusfield [1997]). Thus the number of occurrences of any P in S can be retrieved in $O(w \log L)$ time using a binary search over the suffix array. For the purpose of motif look-ups in all upstream sequences, S can be created as the concatenation of all upstream sequences, where each $s \in S$ is prefixed by some number $a(s)$ of some filler character $s_1 \notin \{a, c, g, t\}$ and affixed with some spacer character $s_2 \notin \{a, c, g, t\}$ such that $\text{length}(s_1^{a(s)} s s_2)$ is constant over all s . Then the exact gene and position within each gene of every occurrence of P in S can be computed in linear time with respect to the number of occurrences of P in S .

Scaling

As noted above, BEAM scales linearly as a function of the beam-width β , the length of the longest motifs examined h , and the average running time $O(f(\cdot))$ of the objective function f . When f_{Sig} is used to find overrepresented motifs, the running time is $O(\beta h^2 \log_2(L))$, where L is the combined length of all upstream sequences in the genome. In practice, a beam width of $\beta = 100$ and maximum motif length $h = 20$ run on a test set of 30 yeast regulons yields average running times of 2–4 seconds per regulon on a 600 Mhz SGI machine.

ACKNOWLEDGMENTS

First of all, we wish to thank Prof. Chris Langmead for his perceptive suggestion of using a pruned search space to tackle this problem. Acknowledgements are also due to Prof. Melanie Mitchell for her guidance with the genetic algorithm, the intellectual ancestor of this project, and to Prof. Doug McIlroy for suggesting the use of suffix arrays to facilitate rapid look-ups. The authors would also like to thank Prof. Mark McPeck for his assistance with some of the statistical issues related to motif finding and Prof. Langmead and Prof. McIlroy for their critical readings of and suggestions on the manuscript. This research was supported by a grant to RHG from the National Science Foundation, DBI-0223719.

REFERENCES

- Bailey, T.L., and Elkan, C.P. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Intelligent Syst. Mol. Biol.* 2, 28–36.
- Bailey, T.L., and Elkan, C. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* 21, 51–80.
- Blanchette, M., and Sinha, S. 2001. Separating real motifs from their artifacts. *Bioinformatics* 17, S30–S38.
- Brazma, A., Jonassen, I., Eidhammer, I., and Gilbert, D. 1998. Approaches to the automatic discovery of patterns in biosequences. *J. Comp. Biol.* 5(2), 279–305.
- Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. 1998. Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.* 8(11), 1202–1215.
- Breiman, L. 1957. The individual ergodic theorem of information theory. *Ann. Math. Stat.* 28(3), 809–811.
- Gusfield, D. 1997. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, New York.
- Hahn, M.W., Stajich, J.E., and Wray, G.A. 2003. The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol.* 20(6), 901–906.
- Hertz, G.Z., Hartzell, G.W. III, and Stormo, G.D. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* 6, 81–92.
- Hertz, G.Z., and Stormo, G.D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15, 563–577.
- Hudak, J., and McClure, M.A. 1999. A comparative analysis of computational motif-detection methods. *Pac. Symp. Biocomput.* 4, 138–139.
- 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.
- Jensen, L.J., and Knudsen, S. 2000. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics* 16, 326–333.
- Kullback, S., and Leibler, R.A. 1951. On information and sufficiency. *Ann. Math. Stat.* 22, 79–86.
- 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.* 6, 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. *Nature Biotechnol.* 20, 835–839.
- Manber, U., and Myers, G. 1993. Suffix arrays: A new method for on-line search. *SIAM J. Comput.* 22, 935–948.
- Narasimhan, C., LoCascio, P., and Uberbacher, E. 2003. Background rareness-based iterative multiple sequence alignment algorithm for regulatory element detection. *Bioinformatics* 19, 952–963.
- Pevzner, P.A., and Sze, S.H. 2000. Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8, 269–278.
- Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. 2002. *Numerical Recipes in C*, Cambridge University Press, New York.
- Ratnaparkhi, A. 1997. A linear observed time statistical parser based on maximum entropy models. *Proc. 2nd Conf. on Empirical Methods in Natural Language Processing (EMNLP '97)*, 1–10.
- Schneider, T.D., Stormo, G.D., Gold, L., and Ehrenfeucht, A. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188, 415–431.
- Shinozaki, D., Akutsu, T., and Maruyama, O. 2003. Finding optimal degenerate patterns in DNA sequences. *Bioinformatics* 19(Suppl. 2), II206–II214.
- Sinha, S., and Tompa, M. 2000. A statistical method for finding transcription factor binding sites. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8, 344–354.
- Sinha, S., and Tompa, M. 2003. Performance comparison of algorithms for finding transcription factor binding sites. *Third IEEE Symp. on Bioinformatics and Bioengineering*, 214–220.
- Stormo, G.D., and Hartzell, G.W. III. 1989. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA* 86, 1183–1187.
- 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.
- van Helden, J., Andre, B., and Collado-Vides, L. 1998. Extracting regulatory sites from upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281, 827–842.

- van Helden, J., Rios, A.F., and Collado-Vides, J. 2000. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucl. Acids Res.* 28, 1808–1818.
- Wingender, E., Dietze, P., Karas, H., and Knuppel, R. 1996. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucl. Acids Res.* 24, 238–241.
- Wolfertstetter, F., Frech, K., Herrmann, G., and Werner, T. 1996. Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput. Appl. Biosci.* 12(1), 71–80.
- Zhu, J., and Zhang, M.Q. 1999. A promoter database of yeast *Saccharomyces cerevisiae*. *Bioinformatics* 15, 607–611.

Address correspondence to:

*Robert H. Gross
Dept. of Biology
Gilman Labs 104
Dartmouth College
Hanover, NH 03755*

E-mail: robert.h.gross@Dartmouth.edu