Published online January 2, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 1 ***189–200***
DOI: 10.1093/nar/gkh169

Finding functional sequence elements by multiple local alignment

Martin C. Frith[1], Ulla Hansen[1,2], John L. Spouge[3] and Zhiping Weng[1,4,*]

[1]Bioinformatics Program and [4]Department of Biomedical Engineering, Boston University, 44 Cummington Street, Boston, MA 02215, USA, [2]Department of Biology, Boston University, 5 Cummington Street, Boston, MA 02215, USA and [3]National Center for Biotechnology Information, National Library of Medicine, Building 38A, Bethesda, MD 20894, USA

Received August 28, 2003; Revised November 1, 2003; Accepted November 14, 2003

ABSTRACT

Algorithms that detect and align locally similar regions of biological sequences have the potential to discover a wide variety of functional motifs. Two theoretical contributions to this classic but unsolved problem are presented here: a method to determine the width of the aligned motif automatically; and a technique for calculating the statistical significance of alignments, i.e. an assessment of whether the alignments are stronger than those that would be expected to occur by chance among random, unrelated sequences. Upon exploring variants of the standard Gibbs sampling technique to optimize the alignment, we discovered that simulated annealing approaches perform more efficiently. Finally, we conduct failure tests by applying the algorithm to increasingly difficult test cases, and analyze the manner of and reasons for eventual failure. Detection of transcription factor-binding motifs is limited by the motifs' intrinsic subtlety rather than by inadequacy of the alignment optimization procedure.

INTRODUCTION

Sequence alignment is a fundamental means of detecting biologically significant patterns in biopolymers. Given a group of sequences that share a common biological property, multiple local alignment methods attempt to locate and align similar subsequences, which may confer this property. This approach has often been used to look for transcription factor-binding sites in similarly regulated promoters (1), but there are numerous other applications. For example, one could search for functional motifs at origins of DNA replication, in scaffold/matrix attachment regions, at RNA 3′-cleavage sites, at exon–intron boundaries or in alternatively spliced exons or introns, in 3′-untranslated regions of localized RNAs, in 5′-untranslated regions of translationally regulated RNAs or in stability regulated RNAs. Of course this technique can also be used to find domains and motifs shared by a set of proteins.

Although a plethora of multiple local alignment algorithms have been developed (1–12), all of these methods possess limitations. Most of them require the user to specify in advance the width of the aligned motif. Most methods provide no estimate of the statistical significance of their alignments, i.e. whether they are any better than alignments of random sequences. Most do not allow gaps in the alignments, or allow them only in a limited way. Many of these methods are heuristic searches for a mathematically optimum alignment, and they often return one result with little indication of whether it has achieved the mathematical optimum, or whether a repeat of the search from a different starting point would produce a completely different alignment. From a practical standpoint, the multiple alignment/motif finding problem is widely recognized to be difficult, and these algorithms often fail to find meaningful motifs when such patterns are known to be present. Our purpose in this study is first to overcome some of these limitations, and secondly to analyze the reasons why these approaches sometimes fail to find biological motifs, so that we can suggest which future research directions are most likely to be fruitful.

Some previous studies have taken steps towards automatic determination of the alignment width. A fragmentation approach permits the motif width to vary, by allowing variable spacing of a fixed number of 'informative columns' in the alignment (13). However, the number of informative columns must be specified in advance. The WCONSENSUS program automatically determines the alignment width, but since it is based on a scoring scheme that is non-decreasing with increasing width, the user must specify an *ad hoc* bias parameter which is subtracted from each column's score (5). Like our method described here, the MEME algorithm optimizes alignment width without requiring any nuisance parameters (14). However, since it also uses a non-decreasing scoring scheme, it employs complex statistical techniques to compare alignments of different widths. We believe our method is theoretically more transparent, and we present evidence that it works better in practice.

In another approach, Hertz and Stormo (5) estimate the statistical significance of a gapless multiple alignment of fixed width as follows. First, they calculate the *P*-value of the score produced when several short random sequences of the fixed width are aligned over their entire length. Then, they perform

*To whom correspondence should be addressed. Tel: +1 617 353 3509; Fax: +1 617 353 6766; Email: zhiping@bu.edu
Correspondence may also be addressed to John L. Spouge. Tel: +1 301 402 9310; Fax: +1 301 480 2288; Email: spouge@ncbi.nlm.nih.gov

a Bonferroni correction on the *P*-value, multiplying it by the number of ways of selecting a subsequence of the fixed width from each of the sequences under scrutiny. As they indicate, the resulting quantity is really an estimated *E*-value, which overestimates the corresponding *P*-value since it ignores correlations between alignments that overlap.

A non-parametric test for statistical significance has also been suggested (13). It matches each sequence of interest with a control sequence, concatenates the two sequences and aligns all the concatenated sequences. A Wilcoxon signed rank test is then used to decide whether the resulting multiple alignment involves significantly more test sequences than controls. On one hand, the method is computationally intensive and cannot declare significance if the number of sequences is small. On the other hand, it permits a flexible specification of the null hypothesis through the selection of the control sequences, whereas analytic methods are usually restricted to a null hypothesis of random independent nucleotides.

We present a modification to the Gibbs sampling alignment algorithm that allows the width of the aligned motif to be discovered automatically, and demonstrate that it chooses suitable widths for alignments of transcription factor-binding sites and SINE elements. Secondly, we implement and test the accuracy of the BLAST statistical method to estimate the significance of alignments. While the MACAW program (15) uses the same method, to our knowledge, the accuracy of its *E*-value estimates has not been examined. Unlike Hertz and Stormo's calculation, the BLAST method accounts for correlations between alignments on the same diagonal of the alignment matrix (see Methods). It ignores correlations between diagonals, although these correlations are known to become negligible in the limit of long sequences.

We also conduct failure tests of our algorithm by using it to search for motifs embedded in DNA sequences of increasing length, and analyze the manner of and reasons for eventual failure. In addition, we explore simulated annealing and rapid restart versions of the algorithm, and discover that the standard sampling method is not the most efficient way to optimize the alignment.

METHODS

Scoring scheme

In common with many other alignment techniques, our method has two components: we first define a scoring scheme to measure the quality of a gapless alignment, and then we specify a search algorithm to find the alignment with the maximum score. One scoring scheme that has been used previously is information content, but it possesses the drawback that it can never decrease as the width of the alignment increases (5). We use an alternative Bayesian scoring scheme that does not have this limitation (13). We suppose that each column in an alignment possesses some underlying propensities to contain each of the four nucleotides, A, C, G and T. These propensities reflect the function of that position in the motif. We denote these unknown propensities q_i, where i ranges over A, C, G, T and the q_i sum to 1. Furthermore, we supply a prior probability distribution over the q_i that indicates how common any set of propensities is. Let

$$S = \sum_{k=1}^{W} \ln \left[\frac{\int_{\{q_i\}} prior\{q_i\} \times \prod_i q_i^{c_{ki}}}{\prod_i p_i^{c_{ki}}} \right]. \quad \mathbf{1}$$

In equation **1**, W denotes the width of the alignment, c_{ki} is the count of nucleotide i in the kth column of the alignment, and p_i is the background abundance of nucleotide i.

Equation **1** for the alignment score is a log likelihood ratio, marginal for a particular window of length W. Its numerator gives the predictive probability of observing the nucleotides in column k under the hypothesis of relatedness, and its denominator gives the predictive probability of observing them under the hypothesis of unrelatedness (equation **1**).

How should we select the prior? A conservative choice would be an uninformed prior, where all sets of q_i are equally probable. However, we do have some prior expectations concerning q_i: we suppose that in the motifs we are looking for, positions with roughly equal propensities for the four nucleotides will be less common on the whole, and positions with propensities biased to one or a few nucleotides will be more common. It is traditional to specify a Dirichlet function for the prior (equation **2**).

$$prior\{q_i\} = \frac{1}{Z} \prod_i q_i^{\alpha_i - 1} \quad \mathbf{2}$$

Z is a normalization constant, and different choices of α_i (referred to as pseudocounts) give different functions within the Dirichlet family. The only reason to favor this functional form over any other is that it leads to tractable mathematics. Setting all the α_i to 1 gives an uninformed prior, $\alpha_i > 1$ gives peaked distributions that favor choices for q_i close to one central, most favored choice, and $\alpha_i < 1$ gives anti-peaked distributions that favor choices at the extremes where some of the q_i are close to zero. With a Dirichlet prior, the integral in equation 1 can be solved analytically, leading to the final alignment scoring formula (equation **3**) (13).

$$S = \sum_{k=1}^{W} \ln \left[\frac{\Gamma(A)}{\prod_i \Gamma(\alpha_i)} \frac{\prod_i \Gamma(c_{ki} + \alpha_i)}{\Gamma(N + A)} \Big/ \prod_i p_i^{c_{ki}} \right] \quad \mathbf{3}$$

A is the sum of the α_i, N is the number of sequences in the alignment, and $\Gamma()$ is the gamma function.

The entropy inequality $\Sigma P_i \ln[Q_i/P_i] \leqslant 0$ ensures that the expected score per aligned position in equation **3** is negative. On average, the expectation of the sum in equation **3** therefore decreases with the alignment width. Thus, our scoring scheme is additive with a negative expectation, conditions that are required for the BLAST statistical method.

Estimation of pseudocounts

In order to select the pseudocounts in a principled way, we fit them to the transcription factor-binding site matrices contained in the TRANSFAC Professional database, release 6.3 (16). We excluded matrices whose nucleotide counts sum to 100, since these are probably percentages rather than counts, and matrices containing fractional numbers. The remaining matrices were then concatenated into one large $M \times 4$ matrix. Supposing that the pseudocounts are equal to one another, i.e.

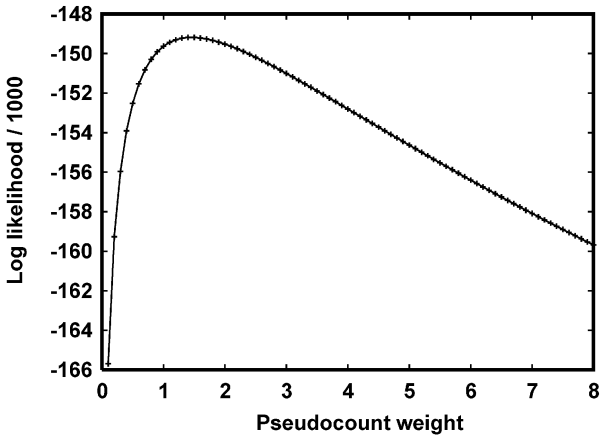

Figure 1. Log likelihood of observing TRANSFAC matrices as the pseudocount weight is varied.

equal to $\frac{1}{4}A$, the log likelihood of observing all of the columns in the TRANSFAC matrices is given by equation **4**. The best fit is the value of A that maximizes this formula.

$$\sum_{k=1}^{M} \ln\left[\frac{\Gamma(A)}{[\Gamma(\frac{1}{4}A)]^4}\frac{\prod_i \Gamma(c_{ki}+\frac{1}{4}A)}{\Gamma(N_k+A)}\right] \quad \mathbf{4}$$

Figure 1 graphs log likelihoods for a range of values of A. The log likelihood achieves a maximum when A is ~1.5, thus confirming our earlier prediction that appropriate values for α_i are less than 1. Plots obtained from only vertebrate or only insect matrices do not look significantly different from Figure 1, nor do plots using only matrices with at least 12 total counts per position (data not shown). We also performed a two-dimensional fit, using one pseudocount for nucleotides (C, G) and another for nucleotides (A, T), but this procedure did not result in a significant difference between the two pseudocounts (data not shown). We decided to use pseudocounts proportional to the background nucleotide abundances: $\alpha_i = 1.5 \times p_i$.

Search algorithm

Having chosen a scoring scheme, we require an algorithm to find the alignment with maximum score. We use the original Gibbs sampling technique (2), together with a novel procedure to adjust the width of the alignment. Beginning from a completely random alignment, the algorithm proceeds through many iterations. At each iteration, it randomly selects one sequence, and stochastically alters which segment of that sequence to include in the alignment. The probability of choosing each segment is proportional to exp(S), where S is the alignment score that would result from including that segment (equation **3**). Some previous Gibbs sampling papers describe choosing each segment with a probability proportional to its likelihood ratio given the other aligned segments (2): it can be shown that this approach is mathematically identical to ours (13). Thus choices that increase the alignment score are favored, but the stochastic aspect allows the algorithm to escape from local maxima. The algorithm may also choose to exclude the sequence from the alignment entirely, again with probability proportional to exp(S).

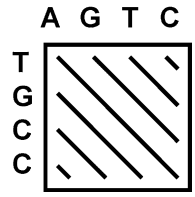

Figure 2. A pair-wise alignment matrix.

Our algorithm further incorporates two resizing moves. In the first kind of move, the left ends of the aligned segments in each sequence are held fixed, and the right ends are varied to change the width of the alignment. A new width is chosen with probability proportional to exp(S). For the second move, the right ends are held fixed and the left ends are varied. As an added benefit, these resizing moves overcome a problem of the fixed width algorithm, which can get stuck in alignments whose end points are shifted left or right relative to the optimal. We do not adjust both ends simultaneously, since the required number of moves would become proportional to length squared, which would make the search too slow.

The search continues until a certain number of iterations, n, have passed since finding the best alignment score so far. In addition, our program performs multiple runs of the search algorithm for a given alignment problem. If many of these runs converge to the same best alignment, we have increased confidence that it is indeed the global optimum.

Estimation of statistical significance

Our algorithm will always find an alignment, even for a set of unrelated sequences. In order to detect biologically meaningful signals, we would like to know whether an alignment is statistically significant, i.e. whether its score is greater than would be expected to occur by chance in a set of random sequences. For gapless alignments, there is a mathematical theory for the distribution of optimal alignment scores of random sequences, which is used by the BLAST program (15,17). This theory maps the alignment problem onto a one-dimensional random walk problem as follows.

A pair-wise alignment problem can be viewed as a matrix (Fig. 2), where each cell contains the score for aligning two letters, and the optimal alignment corresponds to the diagonal run of cells with maximum aggregate score. Similarly, a multiple alignment problem corresponds to a multidimensional matrix. We only consider alignments where all the sequences participate, and for the time being we do not consider reverse strands. The probability ρ_i of a cell containing any score σ_i can be calculated based on the background abundances of nucleotides and the scoring scheme (equation **3**). We can imagine joining all the diagonals of the matrix end-to-end to form a single line of cells with length ΠL_i, where L_i is the length of the ith sequence. The alignment problem corresponds approximately to finding the run of cells within this line that has maximum aggregate score S. If the cells are independent of one another and the expected score per cell is negative, this score follows a Gumbel distribution (equation **5**):

$$\mathrm{Prob}(S > X) = 1 - e^{-K\times\Pi(L_i)\times e^{-\lambda X}}, \quad \mathbf{5}$$

where K is given by a complex formula and λ is the unique positive solution to the equation $\Sigma\rho_i \exp(\lambda \sigma_i) = 1$ (17). Fortunately, the entropy inequality $\Sigma P_i \ln[Q_i/P_i] \leqslant 0$ mentioned above ensures that the expected score per cell is negative. Conveniently, moreover, for the scoring scheme in equation **3** [and, indeed, for any log likelihood ratio scheme (18)], $\lambda = 1$.

This argument has made two approximations: we have neglected that alignments cannot cross the join between two diagonals, and we have assumed that the cells are all independent when in fact neighboring diagonals are correlated (19). An edge correction has been proposed to compensate for the first approximation, where adjusted sequence lengths are calculated, and used instead of L_i in the significance calculation (equation **6**) (20):

$$L'_i = L_i - \ln(K \times \prod L_i)/H \qquad \mathbf{6}$$

H is the expected score per column in an optimal alignment of random sequences (20,21). The correlated diagonals error does not seem to be too great for typical pair-wise alignment problems, but as we shall see it becomes increasingly serious as the number of sequences increases.

This statistical method can be extended to double-stranded alignments. In this case, alignments can occur not only along leading diagonals of the alignment matrix (top-left to bottom-right in Fig. 2), but along any diagonal direction. An N-dimensional matrix has 2^{N-1} diagonal directions. So for a double-stranded alignment of N sequences, we multiply ΠL_i in equations **5** and **6** by an extra factor of 2^{N-1}.

RESULTS

We developed a C++ program to perform gapless local alignment of multiple sequences (GLAM), which can be downloaded from: http://zlab.bu.edu/glam/. The only required input is a file of DNA sequences in FASTA format, although there are many options for modifying the program's behavior. The program performs r (default: 10) alignment 'runs' on the sequences, and outputs the best alignment found, the score and E-value of this alignment, and the number of runs which converged to this alignment. It also prints the marginal score of each sequence's aligned segment, i.e. the score difference between the alignment including this segment and the alignment excluding this segment. The marginal scores are useful indicators of how well each segment matches the rest of the alignment. If some of the runs converged to lower scoring alignments, details of these alignments are also printed. The program can align sequences in either OOPS (one occurrence per sequence) or ZOOPS (zero or one occurrence per sequence) modes (14), and it can perform single- or double-stranded alignment. Upper and lower bounds on the alignment width can be set if desired. All results described here were obtained with default options ($r = 10$, ZOOPS, no bounds on width) unless otherwise specified. The n parameter, defining how many iterations to persist for without improving the alignment, was generally made large enough so that at least three runs converged to the best alignment found. For the sake of comparison, MEME alignments were performed with default options, except that the width bounds were set to the maximum allowed extent (2–300 bp). Data sets used in this paper are available at http://zlab.bu.edu/glam/sup/.

Tests of alignment width optimization

We tested the ability of our algorithm to determine suitable widths for alignments by analyzing sequences containing known transcription factor-binding sites. We obtained 50 bp genomic sequences surrounding 25 mammalian estrogen response elements (EREs), 19 vertebrate LSF-binding sites, 27 mammalian E2F-binding sites, and 35 bicoid- and 27 Krüppel-binding sites from *Drosophila*. Regulation by E2F, bicoid and Krüppel has been extensively analyzed using experimental approaches, and their binding sites have also been studied in previous computational analyses (22,23). Known binding sites for LSF and ERE were collected sites from experimental literature (24–26) (R.B.O'Lone, M.C.Frith and U.Hansen, submitted). The consensus sequences of all these motifs (Fig. 3, see below) are based not only on alignments of known binding sites, but also on mutational analyses to determine both the critical positions and nucleotide preferences for binding of the transcription factors, on chemical and enzymatic 'footprinting' experiments to determine base pairs in contact with the proteins and, in the case of the estrogen receptor, on the structure of the protein–DNA complex determined by X-ray crystallography (27). These various methods agree on determination of the critical DNA–protein contact regions, and in general on the extents of the DNA-binding sites.

In every case, the GLAM alignment is very similar to consensus sequences of the motif established by experimental data, differing in width by a few base pairs at most (Fig. 3) (22,24,25,28). For instance, in the case of the ERE, mutagenesis studies have confirmed that the extent of sequence important for binding affinity is indeed the 13–15 bp region indicated (25). The MEME program aligns three of the motifs (E2F, bicoid and Krüppel) just as successfully as GLAM, but it returns excessively wide alignments for ERE and LSF.

Alignment of Alu elements

It proved surprisingly instructive to apply GLAM to a random set of human DNA sequences, where it aligns the ubiquitous Alu element (Fig. 4). The initial alignment (gray shapes above the lines) covered half an Alu sequence, which is the greatest extent possible without exceeding the bounds of two elements that are truncated. We then masked the nucleotides within this alignment (replaced them with 'n'), and applied GLAM a second time, recovering the other half Alu sequence. Surprisingly, we found no other program that can align these elements as cleanly as GLAM. MEME also identifies the Alu elements, but its alignments do not cover the full width of the element (Fig. 4). Dialign aligns the full extent of the sequences rather than picking out the Alu elements (29). Although BLAST can align the Alu elements in a pair of sequences, it is not straightforward to combine pairwise alignments into a multiple alignment. This example demonstrates that GLAM works for a range of multiple alignment problems. In addition, these alignments were fast (tens of seconds on an 1.1 GHz Pentium III CPU), demonstrating that GLAM is fast not only when the data set is small, but also when the signal to be aligned is strong (as is the case for Alu elements).

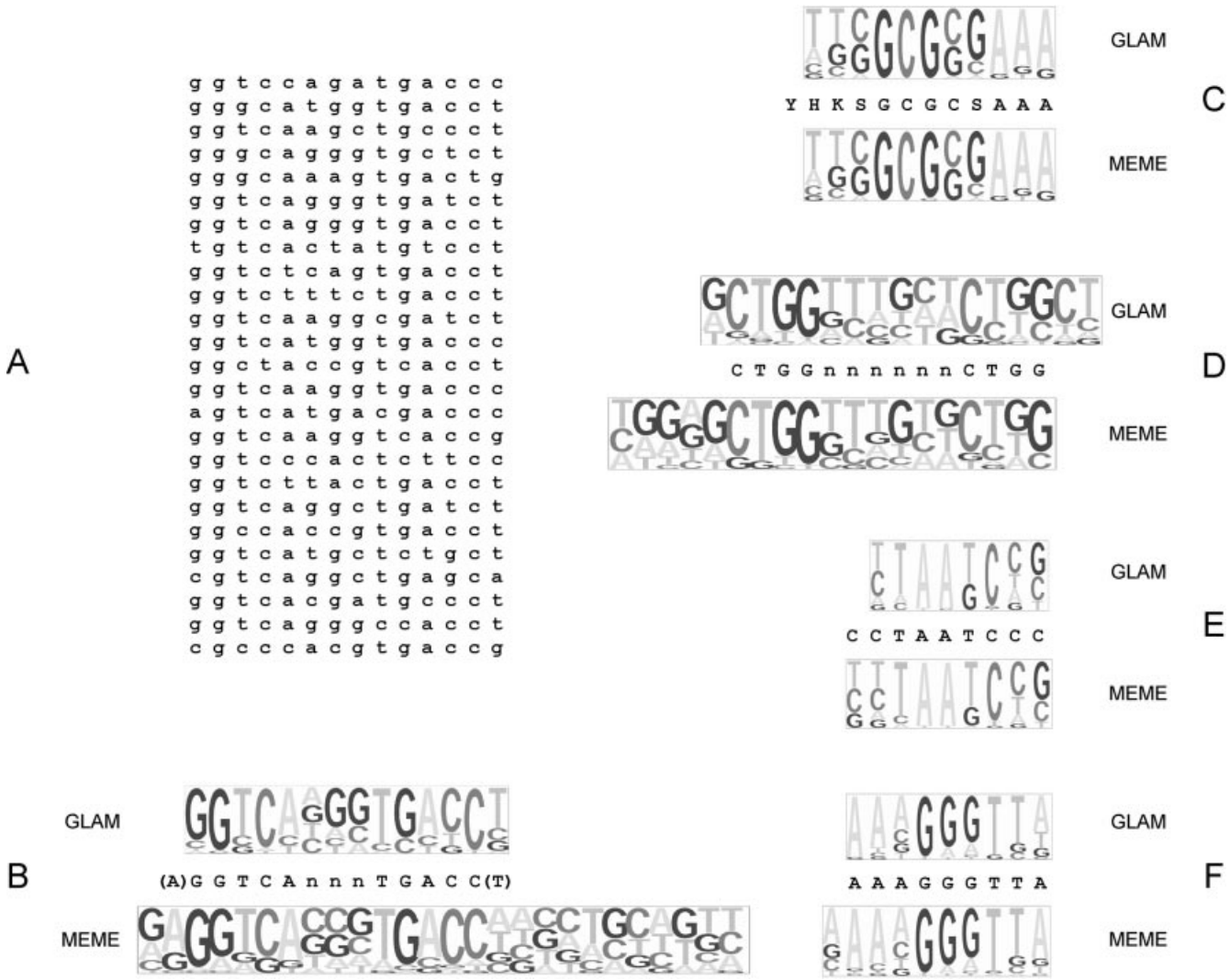

Figure 3. Alignments of transcription factor-binding sites. (**A**) Alignment of 25 EREs by GLAM. The remaining panels show pictogram representations (C. Burge and F. White, http://genes.mit.edu/pictogram.html) of GLAM and MEME alignments, compared with independently established consensus sequences of these motifs. (**B**) EREs (same as in A). (**C**) E2F-binding sites. (**D**) LSF-binding sites. (**E**) Bicoid-binding sites. (**F**) Krüppel-binding sites.

Tests of the statistical significance calculation

Since GLAM will always return an alignment even for unrelated sequences, it is extremely useful to know whether an alignment score is statistically significant, i.e. greater than would be expected by chance for random sequences. To this end, we implemented a multiple sequence generalization of the BLAST statistical calculation, and tested its accuracy by aligning sets of random DNA sequences (in OOPS mode) and observing the empirical score distribution (Fig. 5). For alignments of five sequences (Fig. 5A), the calculated and empirical score distributions generally agree. When aligning greater numbers of sequences, however, the calculated score distribution becomes increasingly conservative. This means that if an alignment score is calculated to be statistically significant, then it certainly is. However, if it is not calculated to be significant, we do not know whether or not it really is. We performed a similar test for double-stranded alignments of five sequences, obtaining an even better agreement between theory and observation than for the single-stranded case (30). In conclusion, this method is quite useful because it is accurate for small numbers of sequences, and it provides upper bounds of *E*-values for large numbers of sequences. Presumably, a correction for correlated diagonals would make the calculation more accurate. Plots of one minus the cumulative distribution function zoomed into the upper tail are provided as Supplementary Material available at NAR Online.

The above analysis has a potential weakness: the BLAST statistical theory applies to optimal alignments, but we have tested it using a heuristic algorithm which is not guaranteed to find them. We present two pieces of evidence arguing that GLAM has, in fact, found the optimal alignment most of the time. If GLAM is robustly finding optimal alignments, we would predict that if we repeat an alignment several times, we will obtain the same result on each (or most) of these attempts (in spite of the algorithm being stochastic). We tested this prediction by performing 10 runs of each of the 1000 alignments and recording how many of the runs converged

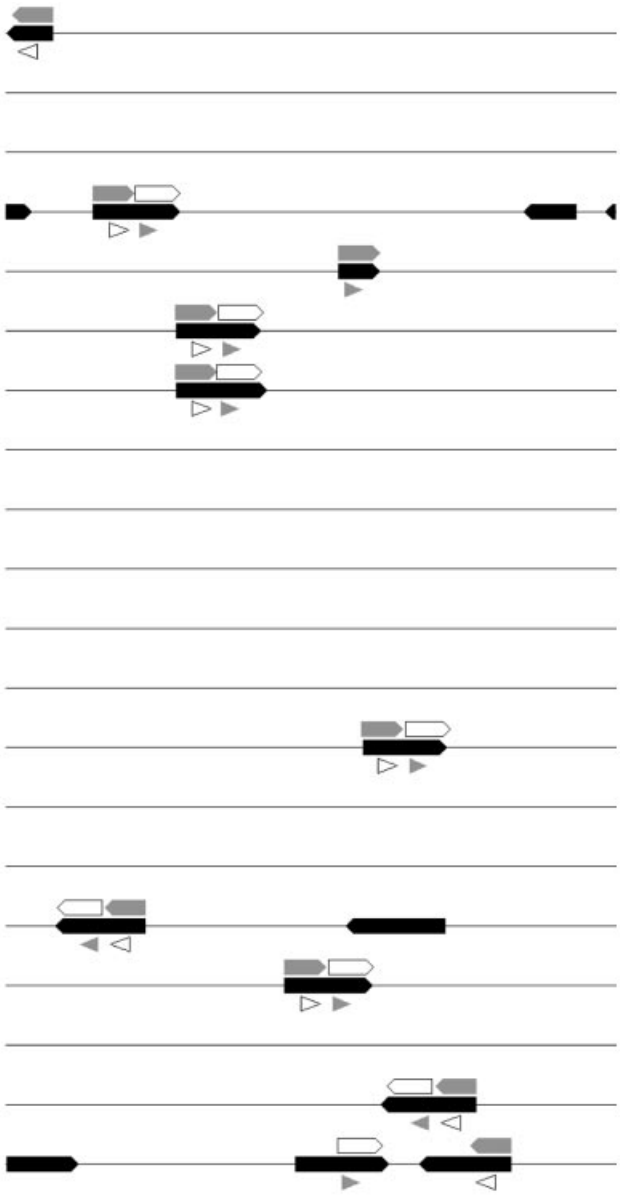

Figure 4. Alignment of Alu elements within 20 human DNA sequences. The horizontal lines represent the sequences (2000 bp each) and the black shapes indicate the positions of the Alu elements within them. The gray and hollow shapes above the lines indicate the first and second alignments returned by GLAM, respectively. The gray and hollow triangles below the lines indicate the first and second alignments returned by MEME, respectively.

to the same best alignment in each case (Fig. 5). In most cases, all 10 runs gave the same result. This agreement of runs is slightly less good for the alignments of 20 sequences, which may explain the score distribution's appearance of having eroded from the Gumbel shape.

We also considered the possibility that the global optimum alignment has a very narrow 'basin of attraction', somewhat like the hole on a golf course. The following alignment problem was constructed to mimic this scenario: the sequence TATTAATTTAAA was inserted once into each of 20 random sequences of 500 bp consisting only of G and C. In this problem, there are no sites outside the embedded motifs, which make it favorable for the algorithm to select the motifs. Nevertheless, GLAM was able to identify the embedded sequences rapidly (tens of seconds on an 1.1 GHz Pentium III CPU). In general, GLAM seems to have most difficulty when the global optimum lies within a very broad and shallow basin where many alignments have almost the same score. This is typically the case when there is no significant motif to be found.

Failure testing GLAM

We applied GLAM to a range of increasingly difficult problem scenarios, to learn the limits of its applicability and to study the reasons for eventual failure. In particular, we were interested to learn whether failure is caused by the search scheme failing to find the highest scoring alignment, or by the biologically meaningful alignment failing to have the highest score. Various numbers of ERE sites were selected, and embedded in randomly generated DNA sequences of various lengths. The sites were embedded in synthetic rather than real sequences because real sequences are likely to contain many unknown biological signals, making it hard to tell whether an alignment is successful or not. Some tests were made harder, and probably more realistic, by including 'decoy' sequences (randomly generated sequences lacking EREs). To measure GLAM's accuracy in identifying the embedded sites, we use the correlation coefficient (equation **7**).

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

TP (true positives) is the number of nucleotides contained in the sites and also included in the alignment returned by GLAM; FP (false positives) is the number of nucleotides not in a site but in the alignment; FN (false negatives) is the number of nucleotides in a site but not in the alignment; TN (true negatives) is the number of nucleotides neither in a site nor in the alignment. The correlation coefficient varies between +1 and –1, with +1 indicating perfect performance, and 0 indicating a performance no better than random.

GLAM's ability to find the sites decreases as the sequences become longer, and improves when there are more EREs (Fig. 6). Interestingly, the accuracy deteriorates in a gradual rather than a catastrophic manner as the sequence length increases, especially for larger numbers of EREs. We studied the alignments in detail to understand this behavior. Many of the EREs are so weak that as the sequence length increases, stronger motifs appear by chance in some of the sequences and get selected by GLAM. The resulting alignments still resemble an ERE, cover the full extent of the motif, and often have the correct width, though they sometimes are wider by several base pairs. It is worth emphasizing that alignments with correlation coefficients as low as ~0.3 still very much resemble EREs. For example, the alignment of 20 EREs in 2000 bp sequences with five decoys has a correlation coefficient of only 0.341, but eight of the EREs are perfectly recovered by this alignment. Moreover, the two highest marginal scores in this alignment belong to embedded EREs.

Decoy sequences almost always get included in the alignments, the exceptions occurring when the sequences are

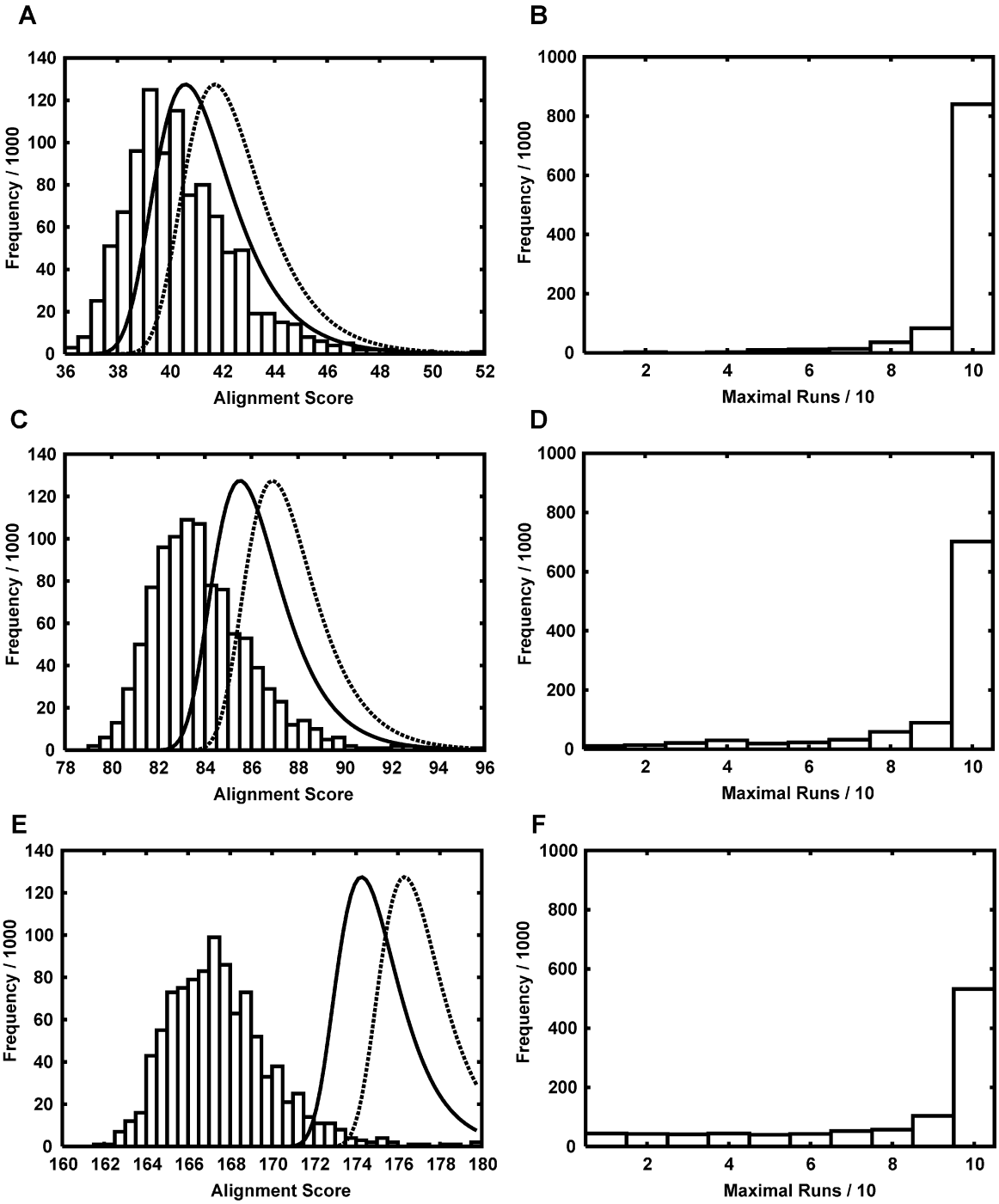

Figure 5. Tests of the statistical significance calculation. (**A**) One thousand sets of five random DNA sequences, each 500 bp long, were generated. Each set was aligned using GLAM, and the alignment scores are plotted as a histogram. The solid curve indicates the distribution of scores expected according to the statistical theory. The dashed curve indicates the theoretical distribution without the edge correction. (**B**) The alignment scores for (A) were obtained by applying 10 GLAM 'runs' to each sequence set and keeping the highest score. The number of runs that converged to the same alignment with this score ('maximal runs') are plotted. (**C** and **D**) Likewise, using sets of 10 DNA sequences. (**E** and **F**) Likewise, using sets of 20 DNA sequences.

very short. We do not think this indicates a fundamental flaw in the method, since our alignment of Alu repeats excluded all sequences that lack this element (Fig. 4). The ERE motif, unlike the Alu, is simply so weak that sequences that resemble it more closely than they resemble random DNA are likely to occur by chance. The segments of decoy sequences that appear in the alignments tend to have low marginal scores, which might provide a criterion for filtering them out.

While the correlation coefficient decreases, the alignment score relentlessly increases as the sequences become longer (Fig. 6). This observation strongly suggests that the accuracy is not suffering because the search algorithm fails to find the

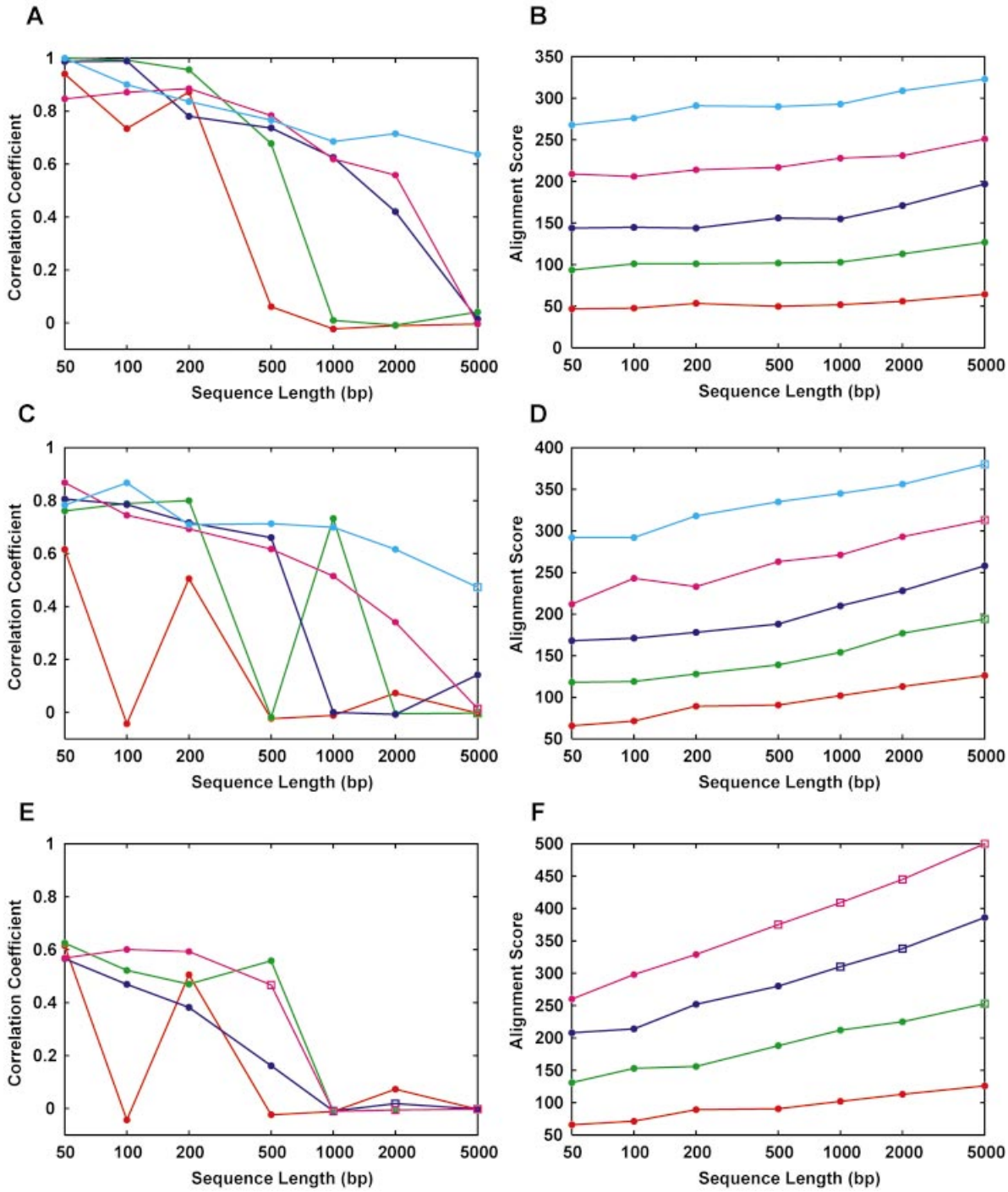

Figure 6. Tests of GLAM's ability to find EREs embedded in random DNA sequences of varying lengths. In all panels, red lines indicate the results when five EREs were present; green, 10 EREs; dark blue, 15 EREs; purple, 20 EREs; light blue, 25 EREs. Open boxes indicate tests where fewer than three out of 10 GLAM runs converged to the same maximal alignment. (**A**) Each ERE was embedded in a random DNA sequence, and these were presented to GLAM. The *y*-coordinate (correlation coefficient) indicates how closely the alignment found by GLAM corresponds to the EREs. (**B**) The scores of these alignments. (**C** and **D**) Five decoys: five random sequences lacking EREs were added to each sequence set. (**E** and **F**) Double decoys: each sequence set contains a number of decoys equal to the number of ERE-containing sequences.

highest scoring alignment, but because the embedded sites increasingly do not constitute the optimal alignment. Finally, we observe that alignments which fall short of the global optimum can still be useful. We performed 10 runs of GLAM on the set of 25 EREs in 5000 bp sequences with five decoys, obtaining eight different alignments, six of which have correlation coefficients >0.4.

We investigated whether GLAM's calculated *E*-value can distinguish biologically meaningful from random alignments, by plotting the *E*-values of all the alignments shown in Figure 6

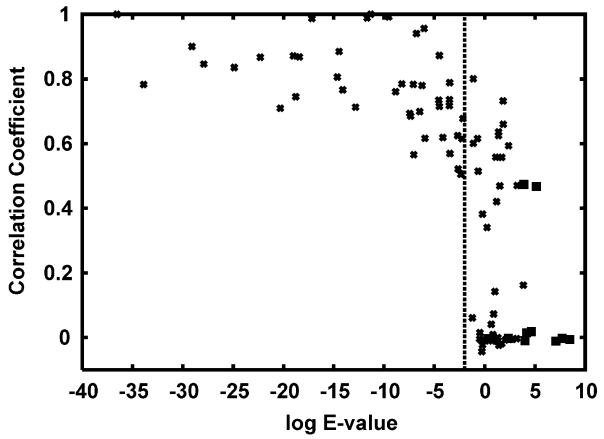

Figure 7. Relationship between statistical significance and accuracy of GLAM alignments. The *E*-value versus correlation coefficient is plotted for each alignment in Figure 6. The open boxes indicate cases where fewer than three out of 10 GLAM runs converged to the same alignment.

against their correlation coefficients (Fig. 7). All of the alignments with *E*-values <0.01 have correlation coefficients >0.5, indicating that they recover the embedded EREs to a considerable extent. On the other hand, there is a 'twilight zone' of alignments that have good correlation coefficients but insignificant *E*-values. These cases can be partly explained by the conservative nature of the *E*-value calculation, and presumably some of these alignments really are on the verge of becoming statistically insignificant. The failure tests were repeated using E2F instead of ERE motifs, which confirmed the observations made above (data not shown).

Assessment of various alignment strategies

We explored whether modifications to the search algorithm might optimize alignments more efficiently. As explained in Methods, the basic algorithm is a stochastic search of the alignment space, where new alignments are selected with probability proportional to exp(*S*), where *S* is the score of the new alignment. A closely related approach is simulated annealing, where a 'temperature' parameter *t* is introduced, and new alignments are selected with probability proportional to exp(*S*/*t*). Low temperatures exaggerate score differences, so that the probability of selecting higher scoring alignments increases at the expense of lower scoring ones. High temperatures have the opposite effect. We experimented with three annealing schedules: constant temperature; geometric cooling, where the temperature is multiplied by a parameter $c < 1$ at each iteration; and a more complex technique known as the 'modified Lam schedule' (31). The modified Lam schedule aims for a target accept rate, i.e. rate of changing the alignment versus leaving it unchanged per iteration. During the first 15% of an alignment run, the target accept rate decays geometrically from 100 to 44%, then it remains fixed at 44%, and finally it decays geometrically to 0% over the last 35% of the run. At each iteration, the temperature is either multiplied or divided by *c* to force the real accept rate towards the target. For this schedule, *n* specifies the total number of iterations, rather than how long to persist without improving the alignment.

We applied these strategies to a difficult case: aligning 40 DNA sequences of length 1000 bp each (one of the double decoy sets described above). Interestingly, the default strategy ($t = 1$) does not achieve the highest alignment scores (Fig. 8A). Among the constant temperature strategies, lower temperatures perform better when the run time is short, and higher temperatures perform better when given more time, up to $t = 0.9$ at 10 000 s on an 1.1 GHz Pentium III CPU. It is conceivable that $t = 1$ would become the best strategy given even more time. Constant temperatures greater than 1 perform extremely poorly (data not shown). The geometric schedules generally perform worse than the constant temperature strategies (Fig. 8B). The modified Lam schedule does not obviously perform better than constant *t*, which is disappointing given its extra complexity (Fig. 8B). Strong values of the forcing parameter *c* work better for short run times, and weaker values work better for long time scales.

Another strategy is to perform very many very fast searches. We applied this strategy using low temperatures ($t = 0.00001$ and $t = 0.5$) so that each search quickly reaches a local optimum, with $r = 100$, 1000, 10 000 or 100 000 restarts (Fig. 8C). When $t = 0.00001$, it is better to use high *r* and low *n* (i.e. more and quicker searches), whereas when $t = 0.5$, it is better to use low *r* and high *n*. Overall, the rapid restart approach is not the best, suggesting that the problem has an extremely large number of local maxima.

In conclusion, the use of a constant temperature slightly less than 1 and the modified Lam schedule are the most effective search methods. We also tested the strategies on two further sequence sets, 20 sequences of 5000 bp each and 100 sequences of 1000 bp each, which confirmed all of the observations made above (data not shown).

DISCUSSION

Our studies of gapless local alignment of multiple sequences demonstrate a straightforward way to optimize the alignment width, indicate that the BLAST statistical method is useful but not perfect for multiple alignments, and suggest enhancements to the alignment optimization procedure.

Fundamental limits to motif discovery?

The failure tests we have performed indicate that the limiting factor in our ability to discover transcription factor-binding sites is the motifs' weakness relative to the scoring scheme. Therefore, improved search algorithms will not bring about fundamental improvements. Perhaps scoring schemes better attuned to these motifs could be devised. One suggestion is to search for palindromic motifs, since many transcription factors bind as homodimers to palindromes (32). However, this approach would miss many non-palindromic motifs, e.g. LSF, bicoid and Krüppel (Fig. 3). Another idea is to use 'shape priors' to favor alignments where highly constrained positions lie next to one another, since real motifs are alleged to exhibit this pattern (M.B.Eisen, personal communication). We expect this approach to produce incremental rather than order-of-magnitude improvements.

However, there is no reason why it must be possible to detect binding sites *ab initio* in arbitrarily long sequences. This problem is not one the cell itself has to solve, since the transcription factors are structurally programmed to recognize

their sequence binding preferences. On the other hand, we are encouraged that GLAM's performance improves as the number of motif-containing sequences increases. For studies of basal promoter elements, intron splice signals or mRNA 3′-end processing signals for instance, it should be possible to obtain hundreds or thousands of sequences that potentially share functional motifs. We also note the many potential applications of multiple local alignment to RNA and protein sequences, which are generally limited to a few thousand residues in length: within the range where GLAM identifies transcription factor-binding site-like motifs.

Phylogenetic footprinting?

One potentially powerful approach for motif discovery is to focus on sequence regions that are conserved between species, such as human and mouse (33). However, studies by ourselves and others suggest that many transcription factor-binding sites are not conserved between these species (R.B.O'Lone, M.C.Frith and U.Hansen, submitted) (34). Moreover, we found that 2 kb regions centered on human EREs often show little sign of conservation relative to mouse, although estrogen regulation could be atypical in this regard (R.B.O'Lone, M.C.Frith and U.Hansen, submitted).

Dealing with repetitive elements

In order to recognize the presence or absence of an interesting alignable pattern, one must avoid being misled by alignments of ubiquitous repetitive elements. This problem is much more severe for multiple than for pair-wise alignments. For example, GLAM's third alignment of the 20 human Alu-containing sequences (Fig. 4) is a highly A-rich motif with an *E*-value of 6×10^{-26}. A variety of approaches could be used to filter out repetitive elements. They can be masked prior to alignment using programs such as RepeatMasker, nseg and dust (A.F.A.Smit and P.Green, personal communication; R.Tatusov and D.Lipman, personal communication) (35). Alternatively, the alignment program can be used iteratively to find several different alignments, and the uninteresting ones can be flagged and ignored (36). This approach requires a robust criterion for defining suboptimal alignments, which is discussed below. Another class of methods builds awareness of uninteresting motifs into the alignment scoring scheme, for example using higher order Markov models of background DNA (8,9), a Bayesian segmentation method (33), or by use of a background sequence set (37). All of these methods face a non-trivial trade-off between filtering repetitive elements aggressively enough and not filtering too many interesting motifs.

Suboptimal alignments

Since sequences may share more than one alignable pattern, we are generally interested in suboptimal as well as optimal alignments. However, the second highest scoring alignment is usually a trivial variant of the highest scoring one, making it

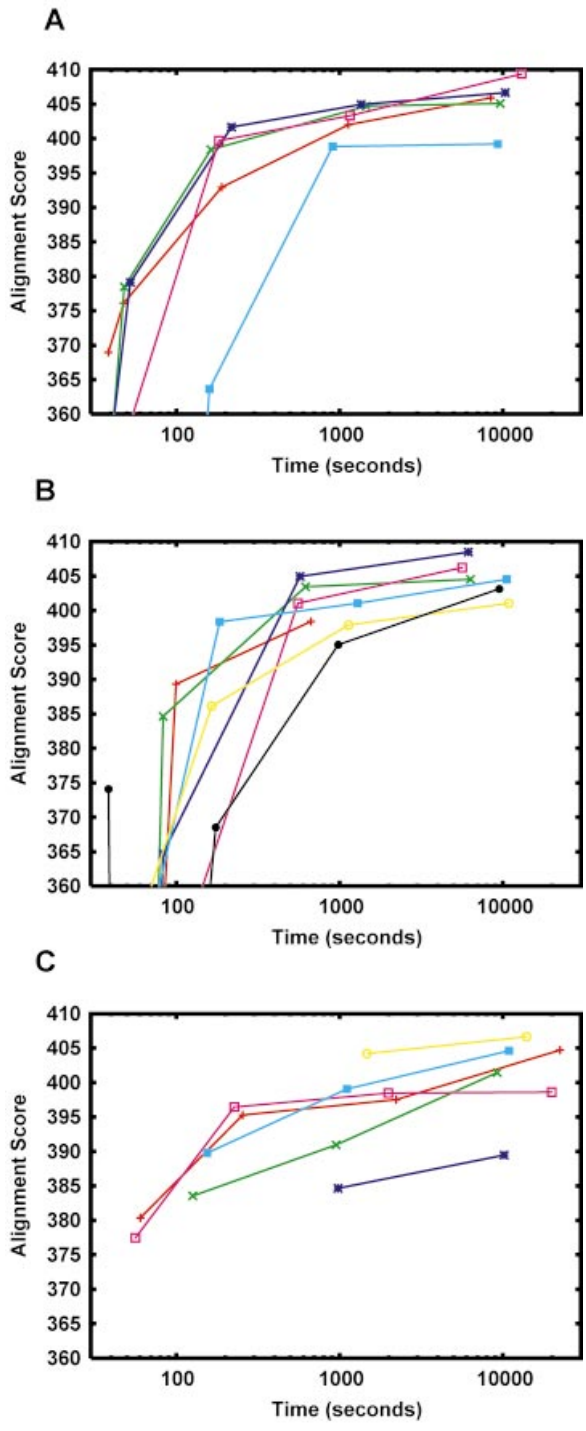

Figure 8. Alignment score achieved versus run time (on 1.1 GHz Pentium III CPUs) for various alignment strategies. (**A**) Fixed temperature strategy. GLAM's *t* parameter was fixed at certain values, and the running time was varied by adjusting the *n* parameter. Red, $t = 0.6$; green, $t = 0.7$; dark blue, $t = 0.8$; purple, $t = 0.9$; light blue, $t = 1.0$. (**B**) Simulated annealing strategies. The 'modified Lam' and geometric cooling schedules were used to lower the temperature during alignment runs (starting with $t = 1$ for Lam and $t = 2$ for geometric). The running time was varied by adjusting the *n* parameter. For the geometric schedule, the cooling factor *c* was set so that the temperature would have a specific half-life relative to *n*. Red, Lam, $c = 0.999$; green, Lam, $c = 0.9999$; dark blue, Lam, $c = 0.99999$; purple, Lam, $c = 0.999999$; light blue, geometric, half-life = *n*/2; yellow, geometric, half-life = *n*/3; black, geometric, half-life = *n*/5. (**C**) Rapid restart strategy. The *n* and *t* parameters were fixed at certain values, and the running time was varied by adjusting the *r* parameter (number of restarts). Red, $t = 0.00001$, $n = 100$; green, $t = 0.00001$, $n = 1000$; blue, $t = 0.00001$, $n = 10000$; purple, $t = 0.5$, $n = 100$; light blue, $t = 0.5$, $n = 1000$; yellow, $t = 0.5$, $n = 10\,000$.

necessary to develop a criterion for how different a suboptimal alignment must be before we are interested in it. We currently use the most aggressive criterion: forbidding every base pair that participates in an alignment from participating in any subsequent alignment. This approach risks missing interesting patterns that slightly overlap a higher scoring alignment. The AlignACE program masks only the most information-rich column of its alignments (38). This method is sometimes too lenient, suffering from 'mask variants', and in some cases may be too aggressive. Therefore, we propose using a generalization of Waterman and Eggert's pair-wise criterion (39): any pair of residues that appear in the same column of an alignment must not appear in the same column of any subsequent alignment.

Future directions

Faster search algorithms would expand the range of problems for which reproducible alignments can be obtained. Techniques that could be explored include ordered over-relaxation and parallel tempering (40,41), or use of oligomer counting algorithms to seed Markov Chain Monte Carlo methods. A more accurate statistical significance calculation would assist interpretation of results, as would statistical estimates for ZOOPS alignments and for marginal scores of aligned segments. Algorithms that can align more than one segment from the same sequence might help to detect motifs that occur in clustered repeats. However, this approach massively increases the search space of possible alignments, which both makes the search harder and increases the likelihood of high scoring random alignments. Ultimately we would like a fully general gapped alignment method, to find motifs that tolerate insertions and deletions. The alignment of Alu elements (Fig. 4) suggests an interesting generalization: to allow sequences to participate in only part of a more extensive alignment. This approach can lead to arbitrarily complex alignment networks (42). For alignment of proteins, it would be wise to incorporate prior knowledge of the chemical similarities among amino acids, for example using a Dirichlet mixture prior (43). Some RNA motifs may exhibit conservation of intramolecular base pairing rather than primary sequence. Existing programs to align such motifs do not scale to large problems (44), and it is worth exploring advanced combinatorial optimization techniques such as simulated annealing.

Conclusions

We have developed several useful enhancements to the Gibbs sampling alignment method: automatic detection of the alignment width, calculation of statistical significance (albeit conservatively), and more efficient optimization of the alignment via simulated annealing. We have shown that transcription factor-binding site discovery is limited by the motifs' weakness rather than inadequacy of the search algorithm. The great variety of untapped applications and the many fascinating avenues for future development suggest that multiple local alignment methods have yet to demonstrate their full potential.

SUPPLEMENTARY MATERIAL

Supplementary material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Joel Graber, Jun Liu, Alexander Favorov and John Straub for helpful discussions, Johannes Jaeger and John Reinitz for advice about simulated annealing, and Dmitriy Leyfer, Prashanth Vishwanath, Kavitha Venkatesan and Yutao Fu for beta-testing. This work was partly supported by an NSF Major Research Instrument grant (DBI-0116574). U.H. and M.C.F. were supported in part by funding from the National Institutes of Health (R01-CA81157). Z.W. and M.C.F. were supported in part by NSF grant DBI-0078194. U.H. and Z.W. were partially funded by NIH grant 1P20GM066401-01. M.C.F. is a Howard Hughes Medical Institute Predoctoral Fellow.

REFERENCES

1. 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.
2. 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.
3. 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.
4. Yada,T., Totoki,Y., Ishikawa,M., Asai,K. and Nakai,K. (1998) Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences. *Bioinformatics*, **14**, 317–325.
5. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
6. Lukashin,A.V. and Rosa,J.J. (1999) Local multiple sequence alignment using dead-end elimination. *Bioinformatics*, **15**, 947–953.
7. Kielbasa,S.M., Korbel,J.O., Beule,D., Schuchhardt,J. and Herzel,H. (2001) Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics*, **17**, 1019–1026.
8. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing*, 127–138.
9. 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.
10. Horton,P. (2001) Tsukuba BB: a branch and bound algorithm for local multiple alignment of DNA and protein sequences. *J. Comput. Biol.*, **8**, 283–303.
11. Keich,U. and Pevzner,P.A. (2002) Finding motifs in the twilight zone. *Bioinformatics*, **18**, 1374–1381.
12. Buhler,J. and Tompa,M. (2002) Finding motifs using random projections. *J. Comput. Biol.*, **9**, 225–242.
13. Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.
14. Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 21–29.
15. Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) A workbench for multiple alignment construction and analysis. *Proteins*, **9**, 180–190.
16. 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. *Nucleic Acids Res.*, **28**, 316–319.
17. Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
18. Yu,Y.K., Bundschuh,R. and Hwa,T. (2002) Hybrid alignment: high-performance with universal statistics. *Bioinformatics*, **18**, 864–872.
19. Park,Y. and Spouge,J.L. (2002) The correlation error and finite-size correction in an ungapped sequence alignment. *Bioinformatics*, **18**, 1236–1242.

20. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
21. Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.
22. Kel,A.E., Kel-Margoulis,O.V., Farnham,P.J., Bartley,S.M., Wingender,E. and Zhang,M.Q. (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. *J. Mol. Biol.*, **309**, 99–120.
23. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
24. Frith,M.C., Hansen,U. and Weng,Z. (2001) Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.
25. Klinge,C.M. (2001) Estrogen receptor interaction with estrogen response elements. *Nucleic Acids Res.*, **29**, 2905–2919.
26. Sanchez,R., Nguyen,D., Rocha,W., White,J.H. and Mader,S. (2002) Diversity in the mechanisms of gene regulation by estrogen receptors. *Bioessays*, **24**, 244–254.
27. Schwabe,J.W., Chapman,L., Finch,J.T. and Rhodes,D. (1993) The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. *Cell*, **75**, 567–578.
28. Lifanov,A.P., Makeev,V.J., Nazina,A.G. and Papatsenko,D.A. (2003) Homotypic regulatory clusters in *Drosophila*. *Genome Res.*, **13**, 579–588.
29. Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
30. Frith,M.C., Graber,J.H., Weng,Z. and Spouge,J.L. (2002) Gapless local alignment of multiple sequences. *Genome Informatics*, **13**, 436–437.
31. Boyan,J.A. (1998) *Learning Evaluation Functions for Global Optimization*. PhD thesis, Carnegie Mellon University.
32. McCue,L., Thompson,W., Carmack,C., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, **29**, 774–782.
33. Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human–mouse genome comparisons to locate regulatory sites. *Nature Genet.*, **26**, 225–228.
34. Dermitzakis,E.T. and Clark,A.G. (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, **19**, 1114–1121.
35. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
36. McGuire,A.M., Hughes,J.D. and Church,G.M. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.*, **10**, 744–757.
37. Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pacific Symposium on Biocomputing*, 467–478.
38. 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.
39. Waterman,M.S. and Eggert,M. (1987) A new algorithm for best subsequence alignments with application to tRNA–rRNA comparisons. *J. Mol. Biol.*, **197**, 723–728.
40. Holmes,I. and Bruno,W.J. (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17**, 803–820.
41. Geyer,J.C. (1991) Markov chain Monte Carlo maximum likelihood. In Keramidas,E.M. (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Interface Foundation, pp. 153–163.
42. Lee,C., Grasso,C. and Sharlow,M.F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.
43. Brown,M., Hughey,R., Krogh,A., Mian,I.S., Sjolander,K. and Haussler,D. (1993) Using Dirichlet mixture priors to derive hidden Markov models for protein families. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1**, 47–55.
44. Gorodkin,J., Heyer,L.J. and Stormo,G.D. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.*, **25**, 3724–3732.