PROTEINS: Structure, Function, and Genetics 7:41–51 (1990)

An Expectation Maximization (EM) Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymer Sequences

Charles E. Lawrence and Andrew A. Reilly
Biometrics Laboratory, Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, New York 12201

ABSTRACT **Statistical methodology for the identification and characterization of protein binding sites in a set of unaligned DNA fragments is presented. Each sequence must contain at least one common site. No alignment of the sites is required. Instead, the uncertainty in the location of the sites is handled by employing the missing information principle to develop an "expectation maximization" (EM) algorithm. This approach allows for the simultaneous identification of the sites and characterization of the binding motifs. The reliability of the algorithm increases with the number of fragments, but the computations increase only linearly. The method is illustrated with an example, using known cyclic adenosine monophosphate receptor protein (CRP) binding sites. The final motif is utilized in a search for undiscovered CRP binding sites.**

Key words: DNA binding proteins, maximum likelihood, CRP, finite mixtures, transcription regulation

INTRODUCTION

The identification of common sites in multiple sequences is frequently encountered in the analysis of biopolymer sequence data. Examples are protein-binding sites in DNA sequences, T-cell binding sites, and antigenic epitopes of protein sequences. Although the methods we present are applicable to a broad class of such problems, we focus here on DNA protein-binding sites. Such sites have traditionally been identified by isolating cis-acting mutations that affect expression and determination of corresponding changes in the DNA of the mutant phenotypes.[1] More recent techniques include affinity purification of the DNA-binding regions and "footprinting" techniques.[2] These methods are time-consuming and provide partial information about the binding sites; the final determination of the sites usually requires the comparison of many examples. One aim of the methods proposed here is to reduce the required experimental work to the identification of restriction fragments that contain binding sites.

The diversity that arises in such fragments is illustrated by *Escherichia coli* promoter sites. Consensus sequences at −35 and −10 are, respectively, "TTGACA" and "TATAAT," but none of the positions are absolutely conserved. The most conserved bases at the −10 position are "TAxxxT," but only about 65% of all promoters match even this criterion. Deuschle et al.[3] have shown that promoters of identical strength exhibit different structures through optimization of different elements of the promoter sequence. This leads to a diversity of functioning sequences, all of which depart substantially from the consensus. Consensus descriptions are therefore likely to have important limitations. Methods that focus on matching identical substrings of "words" share in some of these limitations.[4,5] The identification of sites common to several sequences is consequently linked with the model employed to describe the diversity.

A better description of sites is a stochastic model of residue frequencies. This model assigns probabilities to each of the four bases at each position in the site. An information measure based on a matrix of these probabilities has been shown to provide a useful indication of how constrained the choice of bases is at each site.[6,7] The measure is highly correlated with binding affinities when binding sites are known.[8]

Recently, Stormo and Hartzell,[9] proposed an algorithm to identify protein binding sites in unaligned DNA fragments that uses this measure. They focus exclusively on a mononucleotide/monoresidue model: Each position in the site has residue probabilities independent of any other position. However, the structural features of the proteins involved in site selection transcend the features encompassed in a monoresidue model. The most important limitation of monoresidue models stems from their assumption of no correlative effects across multiple residues. For

Received July 10, 1989, revision accepted September 25, 1989.

Address reprint requests to Dr. Charles E. Lawrence, Biometrics Laboratory, Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, NY 12201.

example, in dimeric prokaryotic DNA binding proteins, residues from the helix-coil-helix motifs make contact with bases in two major groove openings. The symmetry of the dimer leads to a strong correlation in the base frequencies at equivalent yet nonadjacent positions in the major grooves, resulting in a palindromic pattern.

The method presented here is designed to capture and characterize such structural features by employing a set of models derived from the proposed motifs. These models and the associated statistical test procedures allow for the characterization of the binding motif and improved identification of the binding sites. Additionally, a mechanism for determining the length of the site is presented. To consider these various alternative models, we employ a generalization of the information measure presented by Storma and Hartzell, the log likelihood.

The log likelihood is at the heart of a very general procedure for data analysis, maximum likelihood estimation,[10] and the method presented here. Unaligned sequences contain no explicit information on the location of the sites of interest. In multiple sequence comparison problems, this leads to a high degree of positional uncertainty. We employ the "missing information principle" to develop an expectation maximization (EM) algorithm that overcomes this informational deficit. The EM algorithm is described by Dempster, Laird, and Rubin[11] and in a recent text by Little and Rubin.[12] For our purposes, it suffices that it has been shown to be applicable to a broad range of problems, including many problems not normally considered to arise from missing or incomplete data. The applications that are the closest to ours are latent class models[13] and estimation in finite mixture models.[14] We have found that the formulation of this problem as a special class of finite mixture models leads to a number of useful statistical and informational insights, to be presented elsewhere. The algorithm generates estimates of the probabilities that the sites are located in each possible position in each sequence and thereby predicts the most likely binding sites simultaneously with maximum likelihood estimates of the model parameters.

We begin with the mononucleotide model and then show how to incorporate specific motif characteristics through the use of models appropriate to each. The procedure also provides a means for selecting the most data-consistent model from the set of models specified. We apply the method using DNA fragments that are known to contain sites that bind to cyclic adenosine monophosphate receptor protein (CRP)). Using the final selected model, we scan a large database for undiscovered binding sites. This procedure locates four new sites and gives a frame of reference for judging the importance of secondary sites.

MATERIALS AND METHODS

The Problem

To facilitate the description of the algorithm, we begin by posing the problem using CRP binding sites as an example. This protein and its binding sites have been extensively studied, and much is known about its recognition sites.[15,16] CRP binds to several sites on the *E. coli* genome, where it can function either to enhance or to repress gene expression. Figure 1 shows several of these sites, each 22 bases long, from 18 sequences, and a tabular histogram of the occurrence of each base at each position.

Figure 2 shows the data to be analyzed. In these data, the sequences of 105 bases have been selected to position the sites randomly within the sequences.[9] The second row for each sequence marks the location of the start of each site with a 1. Solely to enhance the reader's ability to visualize the sites within the sequences, we have capitalized the bases within each site.

To state the problem succinctly, we seek to reproduce Figure 1 and second rows of Figure 2, using only the sequence information, i.e., the top rows, of Figure 2. Suppose it is known that a protein binds to at least one site of length J, 22 positions in this example, in each of these fragments, but that the position of the binding site(s) in each fragment is unknown. If the fragments are L, here 105, bases long, then there are (L − J), here 83, positions outside the site. Since the location of the site in any of the sequence is unknown, the site could be in any of (L − J + 1), 84, positions in each sequence. There are consequently 84^{18} combinations of segments of 22 bases, from which the correct 18 must be chosen. If we knew where the sites where, i.e., if complete data were available, then the base probabilities, $\rho_{b,j}$, $j = 1,2 \ldots J$; $b = A,C,G,T$ for the positions within the site, and the base composition for all positions outside the site, $\rho_{b,O}$, can be estimated from the collection of marked subsequences (see Fig. 1). Our problem is that the site location information is missing. We are challenged to find the site locations and the base probabilities, $\rho_{b,j}$, using only the sequence data $S_1 \ldots S_N$.

The Algorithm

The EM algorithm simplifies the analysis of problems with missing information by iteratively solving a sequence of problems in which expected information is substituted for missing information. This expected information is used at each step to solve the more straightforward problem associated with having complete information, by maximum likelihood. Thus the first step is to find the form of the solution as if one had complete information.

In our case, we are missing positional information. Thus we begin by formulating the problem as if we had the missing positional information. Given the

Footprint Sites

Col E1 site 2	T T T T T T G A T C G T T T T C A C A A A A
Col E1 site 1	T T T T G T G G C A T C G G G C G A G A A T
ara site 2	T T A T T T G C A C G G C G T C A C A C T T
ara site 1	A A A A G T G T C T A T A A T C A C G G C A
Bgl R mut1	A A C T G T G A G C A T G G T C A T A T T T
crp	G T A T G C A A A G G A C G T C A C A T T A
cya	A G G T G T T A A A T T G A T C A C G T T T
deo P2 site 2	T T A T T T G A A C C A G A T C G C A T T A
deo P2 site 1	A A T T G T G A T G T G T A T C G A A G T G
gal	T A A T T T A T T C C A T G T C A C A C T T
ilv B	A A A C G T G A T C A A C C C C T C A A T T
lac site 2	T A A T G T G A G T T A G C T C A C T C A T
lac site 1	G A A T T G T G A G C G G A T A A C A A T T
mal E	T T C T G T A A C A G A G A T C A C A C A A
mal K	T T C T G T G A A C T A A A C C G A G G T C
mal T	A A T T G T G A C A C A G T G C A A A T T C
omp A	A T G C C T G A C G G A G T T C A C A C T T
tna A	G A T T G T G A T T C G A T T C A C A T T T
uxu AB	T G T T G T G A T G T G G T T A A C C C A A
Pbr P4	C G G T G T G A A A T A C C G C A C A G A T
cat	A A A A T G A G A C G T T G A T C G G C A C

A)

Base frequencies in footprint sites

A	8	10	9	2	0	0	4	15	8	5	3	10	3	7	1	2	15	4	14	4	7	6
C	1	0	3	2	1	1	0	1	5	8	5	1	4	3	2	18	1	15	1	7	1	3
G	3	3	3	0	14	2	15	3	2	5	6	5	10	6	3	0	4	1	5	4	0	1
T	9	8	6	17	6	18	2	2	6	3	7	5	4	5	15	1	1	1	1	6	13	11

B)

Mononucleotide model results

A	4	4	5	0	0	0	4	15	7	5	1	8	1	4	0	1	14	2	13	5	9	7
C	1	0	2	2	3	2	0	2	4	5	3	0	4	3	1	14	0	14	1	5	0	1
G	2	4	5	1	11	0	12	0	2	4	6	3	9	4	4	0	3	0	1	1	0	0
T	10	9	5	14	3	15	1	0	4	3	7	6	3	6	12	2	0	1	2	6	8	9

Fig. 1. CRP experimentally determined sites from the 18 loci listed in Figure 2. Although not all these sites were identified by footprinting experiments, to facilitate the presentation we refer to them in the text as footprint sites. A is compiled from these experimentally determined sites. B is compiled from the results of the mononucleotide model.

locations of the sites in each sequence, and a model, say, the monoresidue model, the probabilities at each position in the site can be estimated. The following generalization of Stormo's and Hartzell's information measure, the log likelihood, forms the basis for the required maximum likelihood estimates:

$$\log L = N\sum_{j=1}^{J}\sum_{b=A}^{T} f_{b,j}\log_e(\rho_{b,j}) + N(L-J)\sum_{b=A}^{T} f_{b,0}\log_e(\rho_{b,0}), \quad (1)$$

where $\rho_{b,O}$ are the unknown population base probabilities, parameters, for all positions outside of the site; $f_{b,O}$ are observed base frequencies; $n_{b,O}$ are the base counts outside the site; and $\rho_{b,j}$ are the parameters and $f_{b,j}$, are the base frequencies and $n_{b,j}$ the base counts for each position in the site. Note that Equation 1 differs from previous information measures used for this problem in two ways: 1) it is a generalization, since arbitary models for base frequencies may be employed, rather than only position-specific base frequency models; 2) it encompases all the data in each sequence, not just the data in the site. Thus the second term is added to the first to describe bases not in the site. As a result, if one of the features that charterize the site is a tendency for a different overall base composition than the rest of the sequence, this effect will be exploited by improvements in the second term of the log likelihood. In a sample of N segments of length L, each position within the site will yield N observed bases, but data from all (L − J) nonsite positions yeild N(L − J) nonsite observations. To obtain the maximum likelihood estimates, we find the values for the parameter estimates $\hat{\rho}_{b,j}$ that maximize the log likelihood. These are easily obtained in this case, since the values that maximize log L are the sample frequencies, i.e.

$$\hat{\rho}_{b,0} = f_{b,0} = n_{b,0}/(N[L-J]) \qquad \hat{\rho}_{b,j} = f_{b,j} = n_{b,j}/N. \quad (2)$$

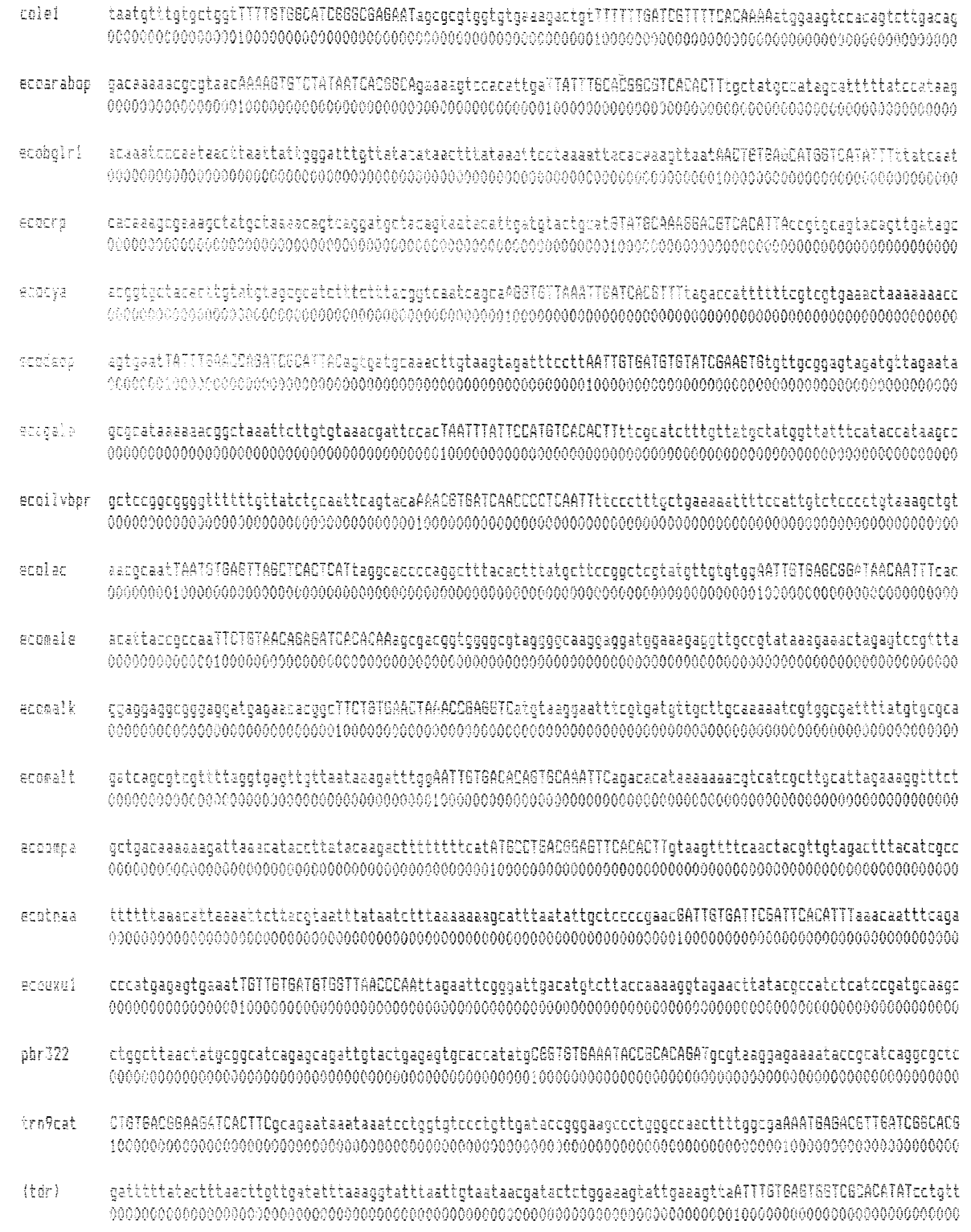

Fig. 2. Sequences from the 18 loci that contain CRP binding sites and site location vectors. The first 17 sequence are taken directly from Genbank rev 55. In new revisions of Genbank eco lmale and eco lmalk are combined in eco malba. (tdc) is not in Genbank; it was taken directly from Stormo and Hartzell.[9] Sequences have been chosen to place the sites at random positions.

These estimate are not available when information on the positions of the sites is missing; the $n_{b,j}$ are unknown.

EM algorithms are named for their two iterative steps, the expectation (E) step and the maximization (M) step, which are alternately repeated until a convergence criterion is satisfied. In the following description, we assume that we have completed some number of these iterative cycles, say (q − 1).

E Step

The site, which is J = 22 bases long, can start at any of L − J + 1 = 84 positions. We are missing the information that specifies the location of the start of each site. However, at the beginning of the E step of the q^{th} iteration, the current values of the population frequency estimates from the previous iteration of the M step are available. These values taken together specify the current estimate of the model parameters. With these values, we calculate the probability of observing the data in each sequence assuming that the site starts in each of the possible L − J + 1 (84) positions. These probabilities can now be used to calculate the probability that the site starts in each of the possible positions by using Bayes formula. Appendix A gives the formulas for this calculation.

Now, using the probabilities that the site starts in each of the possible positions as weights, add across the positions to find the expected number of the bases at each position in the site. For example, assume that there is an A in the first position of the window that starts at position 50 of, say, the third sequence. If the probability that the site starts at position 50 in the third sequence is 0.01: add 0.01 A's to the accumulating expected number of A's in the first position of the site. These expected values may be formally represented as follows:

$$\epsilon_{bJ}^{(q)} = E(n_{b,j} | \rho^{(q-1)}, S); \quad j = 0, \ldots, J; b = A, C, G, T, \tag{3}$$

where S indicates the N available sequences.

M Step

Recall from Equation 2 that the maximum likelihood estimates for the population frequencies are just the sample frequencies when complete data are available. In the M step, substitute the expected number of bases for each position in the site from the E step for the unavailable directly observed number of bases into Equation 2:

$$\hat{\rho}_{bJ}^{(q)} = {}^{(q)}\epsilon_{bJ}/N \qquad j = 1, \ldots, J; \; b = A, C, G, T \tag{4}$$

$$\hat{\rho}_{b,0}^{(q)} = \epsilon_{b,0}^{(q)}/(N[L - J]) \qquad b = A, C, G, T.$$

The algorithm converges when the parameter estimates stop changing, i.e., when

$$\hat{\rho}^{(q)} = \hat{\rho}^{(q-1)} = \hat{\rho}^{(*)}. \tag{5}$$

At convergence the algorithm yields estimates of the population base probabilities, $\hat{\rho}_{b,j}$, and a set of posterior probabilities that indicate the probability that the site is at each of the possible positions in each sequence.

Motif Characterization

Although we have employed the mononucleotide model to illustrate the formulation of the algorithm, the algorithm encompasses a much broader set of motifs. To incorporate alternate motifs in our analysis, we consider alternate models for the population base frequencies for the positions within the site. For example, when a palindromic sequence is appropriate, expected numbers in the palindrome including bases, and their reverse complements are employed in Equations 3 and 4. Variable-length gaps within the site can be added to the algorithm through the use of a second missing variable, the gap length. This will result in the modification of the E step by including as candidate sites all locations with all allowed gaps lengths.

When the binding motif is known a priori, then we need employ only the single model that corresponds to this motif. However, when the binding motif is not fully specified, the approach proposed here provides a means for choosing between the candidate motifs: We progressively impose more restrictions to yield a set of progressively more specific models. This sequential process is terminated when added restrictions reduce our ability to predict the data more than would be expected from chance alone. The likelihood ratio statistic is used to assess the limits of chance variation.

Application to CRP: The Data

We tested the algorithm by using CRP binding sites as an example. Beginning the analysis with only the assumption that CRP is a prokaryotic dimeric DNA binding protein yields a set of specific hypotheses about the binding motif. A palindromic binding motif at the positions in adjacent major groove openings implies a probability model specifying reverse complementarity of the bases separated by about six positions for the intervening minor groove. This model is implemented by requiring that the bases at positions 4–8 be the same as the probability of corresponding complementary bases in positions 19–15 (denoted in Table II as palindromic motifs). For example, the probability of a *T* at position 6 should match the probability of an A at position 17.

The bases in the minor groove that intervenes between the two major groove openings in the site are less accessible to the CRP protein. It has been suggested that the interaction is such that, at most, double-bonded base pairs (AT/TA) can be distin-

TABLE I. Starting Positions of the Sites*

Sequence	Footprint sites	Two most likely sites: Mononucleotide First	Mononucleotide Second	Final model (D) First	Final model (D) Second
cole 1	17, 61	61	45	61	17
eco arabop	17, 55	55	76	55	17
eco bglrl	76	76	40	76	42
eco crp	63	63	73	63	45
eco cya	50	50	15	50	15
eco deop	7, 60	7	39	7	60
eco gale	42	24	76	42 (.92)	24 (.07)
eco ilvbpr	39	39	20	39 (.91)	20 (.09)
eco lac	9, 80	9	73	9	73
eco male	14	14	12	14	12
eco malk	29, 61	61	29	61	29
eco malt	41	41	11	41	51
eco ompa	48	48	12	48	82
eco tnaa	71	71	34	71	7
eco uxul	17	17	26	17	25
pbr-p4	53	53	84	53	51
trn9cat	1, 84	5	66	84 (.94)	5 (.05)
(tdc)	78	78	76	78 (.93)	76 (.07)

*Footprint sites are those indicated in Stormo and Hartzell,[9] except for site 61 in eco malk, which has been recently identified as a stronger footprint site than eco malk at position 29.[21] The two most likely sites are the two sites in each sequence with the highest posterior probability of being a binding site. Values within the parentheses are the posterior probabilities of the positions being at these sites. For all other sites the probability for the most likely site > 0.995 and for the second < 0.005.

guished from triple-bonded base pairs (CG/GC).[17] A model that considers two, (AT/TA) vs. (CG/GC), rather than four categories at each position may be appropriate in this case (denoted in Table II as AT position-specific motifs). Alternatively, the lower melting point of AT/TA, pairs, which impart flexibility, may allow better accessibility for the protein. If this is the relevant sequence feature in some minor groove positions, then only the overall concentration of (AT/TA) pairs would affect binding in these positions (denoted in Table II AT concentration motifs).

Additionally, since dimeric DNA binding proteins make base-specific contacts only in two adjacent major grooves, the DNA recognition region may span no more than 16 positions. Classic phosphate ethylation experiments indicate a binding site of 22 bases for CRP. However, using new gel electrophoretic analysis methods, Liu-Johnson et al.[18] present evidence of a thermodynamically defined binding domain of 26–30 bases for the CRP site in the eco lac promoter. To address site length issues, motifs that include/exclude bases from the ends of the site are called for. Since these additional residues are from minor grooves, various combinations of mononucleotide, AT position-specific, and AT concentration models are appropriate candidates.

RESULTS

The mononucleotide model is a natural starting point. Figure 1B gives the final estimates of the population percentage frequencies for each of the four bases in all 22 positions in the site for the mononucleotide model. Note that there is good agreement between the frequency of the bases from the footprint results and those estimated by the algorithm. Table I identifies the two top choices for the sites selected by the algorithm, i.e., the two sites for each sequence with the highest posterior probabilities. The algorithm's first choice correctly identifies a site reported by footprinting for 16 of the 18 sequences. In the top two choices, we identify 17 of the 24 footprinting sites. The results of the examination of alternative motifs are given in Table II. In comparing motif B with motif A, there is no evidence in the data to reject the palindromic motif ($P = 0.37$). Furthermore, in comparing motif C with motif A, only the double-bonded bases pairs are justified by the data in the minor grooves ($P = 0.73$). As is indicated by the comparisons of motif D and motif C, we can further relax the requirements for the outer minor grooves. Motif D employs only overall AT concentration in these outer minor groove positions. However, as indicated by the comparison of motif E and motif D, in the central minor groove, we reject the model that considers only overall (AT/TA) vs. (CG/GC) concentration. We find rather that double bond (AT/TA) vs. triple bond (CG/GC) specificity at each of the six central minor groove positions is justified by the data.

To address the question of site size, we examined the effect of describing residues at the end of the site as bases whose frequency was described by the frequency of bases in nonsite positions. In comparing motif F with motif D, we can reject the hypothesis that positions (1,22,2,21,3,20) are no different from nonsite positions. The base composition at positions 3 and 20 is not much different from that of the nonsite positions. Thus the primary reason that these outside minor groove sites cannot be excluded is that the characteristic overabundance of double-bonded bases in outer minor groove positions (1,22,2,21) is conserved over the sites. Thus these positions apparently play a role in the DNA–protein complex. On the other hand, in comparing motif G and motif H to model D, we see that there is no significant evidence in these data for a site size as large as 26 or 30 bases. In summary, the these sequence data are consistent with a binding domain of 22 bases composed of three distinct regions: the outer minor grooves, the major grooves, and the central minor groove. These regions follow three different binding patterns, as shown in Table III.

As is indicated in Table II, the agreement between the known footprint sites and the sites selected by the algorithm improves as we converge on the model

TABLE II. Alternative Models/Motifs*

Model/motif	Log l	LRT (comparison model)	*P* value (degrees of freedom)	Number of footprints	
				First	First and Second
A. Mononucleotide	−2,391.7			16	17
B. Palindromic all positions	−2,410.1	36.9 (vs. A)	0.37 (33)	17	20
C. Palindromic major AT-specific minor	−2,410.9	38.2 (vs. A)	0.73 (41)	17	20
D. Palindromic major AT-specific central AT concentration ends	−2,413.5	5.54 (vs. C)	0.13 (3)	18	22
E.				†	†
Palindromic major AT concentration minor	−2,464.0	100.9 (vs. D)	<0.001 (6)	11	15
F. Sixteen positions Palindromic major AT-specific central	−2,426.4	25.8 (vs. D)	<0.001 (7)	17	20
G. Twenty-six positions Palindromic major AT-specific central AT concentration ends	−2,412.6	1.8 (vs. D)	0.18 (1)	16	20
H. Thirty positions Palindromic major AT-specific central AT concentration end minor Mononucleotide end major	−2,405.1	16.9 (vs. D)	0.20 (13)	16	21

*The values in the column labeled Log l are the log likelihood values from Equation 1. The values in the column labeled LRT are the values of the likelihood ratio test, which is equal to −2 times the differences between the log liklihoods of the two models indicated in parentheses. Likelihood ratio statistics are calculated using the method presented by Aitkin and Rubin[14] to overcome potential problems with the assumption of asymptotic normality of the estimators. For all models, the optimal solutions were obtained by adding 1 to each cell in the M step to attain an initial optimal. Then, the 1 was removed to allow convergence to the final optimal solutions presented here. The results considered after the initial optimal solutions are not remarkably different from those presented here. It is well known that mixture problems have multiple optimal solutions. Thus it is strongly recommended that multiple starting values be employed.[23] Since this is a mixture problem, multiple optimals can and sometimes do occur for the method presented here as well. In this application, when we started from an initial solution with a probability of 0.25 for all bases in all positions, i.e., $\rho_{bj} = 0.25$, b = A, C, G, T; j = 0, 1 . . . J, we always converged to solutions that identified the indicated experimentally determined sites shown in Table I or to solutions that were at most one base off. Spacing of minor and major grooves is as proposed by Liu-Johnson et al.,[18] six positions for the central minor groove and five each for the surrounding major and minor grooves.
†For model E, these identified sites disagree with the footprint sites by plus or minus one position.

most consistent with the data. For the final model, motif D in Table II, the algorithm's first choice agrees with the footprint data for all 18 sequences; 21 of the 23 footprint sites are included in the top two choices. The two sites that we did not identify are trn9cat at position 1 and eco lac at position 80. The trn9cat site at position 1 shows much better agreement with the other footprint sites, with seven bases between the major grooves positions,[9] and eco lac at position 80 shows better agreement when five bases are used.[15] Thus the algorithm identified all footprint sites spaced with six bases between the major groove positions. There are 1,512 (84 × 18) segments, and 84^{18} combinations of candidates segments of length 22 in the sequences. Since all but 23 of these segments are not footprint sites, these results illustrate the ability of the algorithm to discriminate sites from nonsites.

In an effort to identify new candidate binding sites in these loci, we gathered the entire sequences for each of the 17 loci that were available in Genbank. This yielded some 31,614 segments of 22 bases, for which the probability that each was drawn from the population of sites, as described by model D, was calculated. This procedure is not an additional validation step; the model was tuned to the best sites in the segments of 105 bases in Figure 2. All but a few segments had very small probabilities. Figure 3 displays the cumulative probability distribution of the 50 sites with the highest probabilities. Not surprisingly, the best sites include the primary sites identified by the algorithm. The top 19 sites, those above the indicated cut-off, include 16 of the 17 primary known footprint sites. The rapid increase in the predicted probabilities in the vicinity of the cut-off indicates that this small fraction of observations, approximately 0.06% (100 × 20/31,614), are outliers and are not drawn from the same population as the bulk of the observations. The only sequence with a footprint site not included in this group of outliers is trn9cat. As has been pointed out, neither of the footprint sites in trn9cat shows good homology with the remaining sites.[15]

The other three outlying sites include two seg-

TABLE III. Description of the Final Model (D)*

Groove	Position	Base		Base			
		A or T	C or G	A	T	C	G
Outer minor groove	1	0.876	0.124				
	2	0.876	0.124				
	21	0.876	0.124				
	22	0.876	0.124				
Major groove	4 & (19)			0.030	0.806	0.109	0.056
	5 & (18)			0.028	0.139	0.057	0.777
	6 & (17)			0.053	0.833	0.113	0.000
	7 & (16)			0.115	0.056	0.000	0.829
	8 & (15)			0.830	0.028	0.114	0.029
Central minor groove	9	0.614	0.386				
	10	0.397	0.603				
	11	0.441	0.559				
	12	0.781	0.219				
	13	0.219	0.781				
	14	0.509	0.491				

*Estimated population base frequencies from model D. In the major groove for the positions given in parenthesis, the probabilities are for the complementary base. The base probabilities in the nonsite positions provide a rudimentary description of the context in which the sites occur. The estimates for these are A (.303), T (.300), C (.187), and G (.211). Positions 3 and 20 are on the margin of the major groove and the end minor grooves. They were equally well described by several models. We selected the simplest mononucleotide model for these positions.

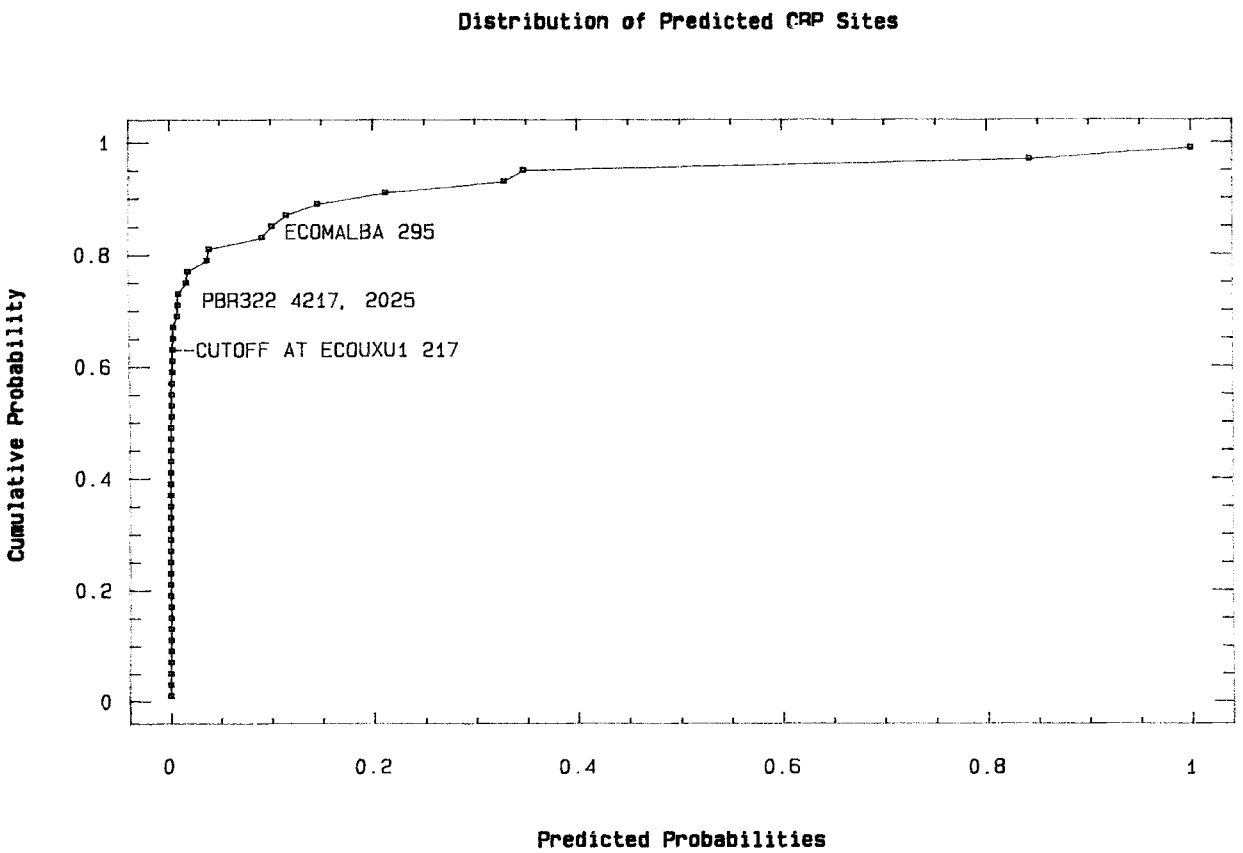

Fig. 3. The cumulative distribution function for the 50 segments with the highest probabilities of being drawn from the population of binding sites as described in model D. The cut-off point is indicated for the last consecutive primary experimentally determined site. The prb322 sites and the eco malba site at the alternative sites described in the text.

ments from pbr322. The segment at position 4217 is in the P3 promoter region of the β-lactamase gene, 42 positions upstream from the transcriptional start site.[19] The segment at position 2025 is 68 bases upstream from the replication factor Y effector site. Factor Y is one of the proteins in the multienzyme unit known as the primosome. The primosome gains entry into the DNA at a specific site, and the specificity of this interaction resides in factor Y.[20] These two segments are in regulatory regions of the plasmid and thus are in plausible positions for a CPR binding site.

The loci eco malk and eco male as listed in Figure 2 constitute segments of a joint regulatory region for both genes. This joint region has since been included in Genbank as the locus eco malba. Raibaud et al.[21] have recently presented experimental evidence of four CRP footprint sites in the joint regulator region of locus eco lmalba. The segment at positions 295 in eco malba, which overlaps the end of the eco malk locus listed in Figure 2, was identified as a footprint site to which CRP binds more tightly than to the other two CRP sites listed in Table I for eco malk. Raibaud et al. have also given the order of filling of these four sites. These sites correspond with the three sites in eco malk mentioned above and the site identified in eco male in Table I. Furthermore, the order in which these sites are filled in the footprint experiment agrees fully with the order of the predicted probabilities from the final model, motif D.

The secondary site in locus eco deopl is the 25th most likely binding site. All other secondary footprint sites have predicted binding probabilities well below the outlier cut-off value. For example, the site at position 29 in eco malk in Figure 2 has a lower predicted binding probability than 180 other sites. On the other hand, the other three footprint sites in eco malba are included in the group of outliers.

It has recently been suggested[22] that Laplace's law of succession be used to smooth sampling zeros in a motif derived, as here, from a small database in order to search a large one. We therefore repeated the search of the 31,614 potential segments with the smoothed version of motif D. The results were essentially identical to those from the unsmoothed search except that a new site, trn9cat position 493, now appears above the cut-off. This sequence has a G at position 6, previously unobserved in motif D. This site is in the coding region of the catl gene.

DISCUSSION

The ability of the algorithm to examine motifs that encompass the correlative effects of multiple residues, which may be nonadjacent in sequence, is a feature not available in previous algorithms. This feature greatly expands the range of problems that can be investigated. The ability of the methods presented here to characterize the binding motif as it identifies the sites through the application of successive models is another new feature. A rigorous statistical test, the likelihood ratio test, allows for rejection vs. acceptance of models as (in)consistent with the data. As is illustrated in the example, the stepwise application of this test allows for the characterization of the binding motif from unaligned sequence data alone. The ability to determine the length of the site most consistent with unaligned sequence data is another feature not available in previous algorithms.

As is illustrated in the CRP example, the incorporation of such features into more specific models improved the identification of footprint sites. This improvement stemmed from the fact that the more specific models have fewer free parameters to be estimated from the data. Specifically, the mononucleotide model required 66 parameters, i.e., 66 degrees of freedom, whereas the final model required only 28 parameters. Consequently, the more specific final model brought substantially more á priori information, in the form of structural considerations, to the data analysis problem; substantially less information need come from the data. In this example, the mononucleotide model identified a substantial proportion of the footprint sites reflecting the substantial information in the sequence data. In cases in which there are fewer sequences available or the sites have lower similarity, we expect more specific models to show an even larger advantage in the correct identification of the binding sites. When binding sites are unknown, there is a danger of employing a model whose specific characteristics are inappropriate for the unidentified sites. However, in that event, the likelihood ratio tests provide evidence to reject the inappropriate model, as was illustrated by model E in Table II.

Nearly all other methods used for the identification of features in sequences focus on the identification of the single best site. This approach brings with it the requirement that order statistics and extreme theory be employed for statistical inferences. The approach we have taken avoids this problem by modeling the problem as a mixture of models at all possible sites. Thus, the contributions from all possible sites are included in all the parameter estimates, and the need for the selection of the best site is avoided. Statistical inferences for mixtures are also frequently problematic due to the discontinuity of the log likelihood at the boundaries of the parameter space. However, as was first described by Aitkin and Rubin,[14] these discontinuities are circumvented through the use of a Bayesian approach. We have employed their approach here to maintain the validity of the assumptions underlying the likelihood ratio test.

The method proposed here is directly applicable to the identification of common sites in protein sequence, given a reasonable description of the motif hypothesized to be shared by the sites. With pro-

teins, the need for nonmonoresidue models is apparent from the large number of residues, 20, at each position. The use of alternate alphabets, which group residues by a common characteristic, such as charge, provides an initial step. However, several characteristics at once may be required to describe a motif adequately, in which case a multivariate motif model may be required. The interaction of nonadjacent residues in proteins is frequently important to there structure. Thus the ability of the methods given here to incorporate the correlative effects of nonadjacent residues may be of considerable value in this respect. The use of variable-length gaps may also be important.

The memory requirements for this algorithm are relatively small and are linearly dependent on both the number of sequences and their length. Only three small arrays need to be stored: the data ($O[N \times L]$); the matrix of expected sufficient statistics (three elements for each element in the site); the vector of posterior probabilities (O[L] elements). In contrast to the algorithm of Stormo and Hartzell,[9] the power and convergence properties of this algorithm do not depend on the amount of memory employed to store these arrays. This problem is a member of the class of problems known as finte mixtures. The convergence properties of the EM algorithm for finite mixtures have been reviewed at length.[23] Stated briefly, for these problems, the EM algorithm more consistently converges than do Newtom-type methods, but the convergence is slower. Also, the log likelihood for these problems can have multiple optimals. There is currently no algorithm that can ensure convergence to the global optimal. Thus multiple starting points are recommended. We have found that, for the problem presented here, the algorithm consistently converges to a solution that identified the majority of the binding sites to within plus or minus one base.

CONCLUSIONS

The finding of a palindromic motif for the major grooves is consistent with the dimeric form of CRP. However, as indicated by the differences in the central minor groove, the overall binding motif described by our final model is not symmetric. This is consistent with evidence that the two subunits are not fully symmetric in cocrystals with cAMP.[24] The finding that only pairs rather than bases can be distinguished in the minor grooves is consistent with the available hydrogen bonding contacts in the minor grooves of B-DNA.[17] We note, however, that pair specificity does not necessarily require the formation of such bonds. Rather, this pair specificity could also arise from the bending requirements for the DNA in complex with CRP.[24,25] Our findings on the size of the site agree with the ethylation experiments. We were unable to confirm electrophoretic measurements indicating a site larger than 22 bases.[18]

The observation that the primary footprint sites in all but one of the target sequences are outliers (i.e., they are not drawn from the population that represents nearly all other segments of 22 bases in the *E. coli* genome) suggests that CRP selectively binds to these sites. The finding that there are many segments with higher site probabilities than the secondary footprint sites indicates that there is a set of "pseudosites" that CRP may bind to as well as it does to secondary footprint sites.[15]

As the quantity of DNA sequence data grows, methods for the identification and description of important features will become increasingly important. Projects such as the Human Genome Initiative[26] will greatly expand the requirement for such tools. We present here a statistical method that goes beyond existing methods in its ability simultaneously to identify and to characterize sites in unaligned sequences. The identification of sequence patterns with variable-length gaps is a natural extension of the analysis presented here by the introduction of a second missing variable, the gap length. Thus extensions to multiple alignment problems appear promising.

ACKNOWLEDGMENTS

We thank Gary Stormo for providing us with the data and for helpful suggestions. We appreciate the valuable suggestions made by Sam Karlin and David Lipman.

REFERENCES

1. Gilbert, W., Gralla, J., Majors, J., Maxam, A. (eds). "Protein–Ligand Interactions." Berlin: Walter de Gruyter, 1975: 193.
2. Kadonaga, J.T., Jones, K.A., Tjian, R. Promoter-specific activation of RNA polymerase II transcription by Spl. Trends Biochem. Sci. 11:20–23, 1986.
3. Deuschle, U., Kammerer W., Reinerg G., Hermann B. Promoters of Escherichia coli—A hierarchy of in vivo strength indicates alternate structures. EMBO. J. 5:2987–2994, 1986.
4. Karlin, S., Morris, M., Ghandour, G., Leung, M. Efficient algorithms for molecular sequence analysis. Proc. Natl. Acad. Sci. USA 85:841–845, 1988.
5. Pearson, W.R., and Lipman, D.J. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85:2444–2448, 1988.
6. Stormo, G.D. Computer methods for analyzing sequence recognition of nucleic acids. Annu. Rev. Biophys. Biophys. Chem. 17:241–263, 1988.
7. Mulligan, M.E., Hawley, D.K., Entriken, R., McClure, W.R. Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity. Nucleic Acids Res. 12:789–800, 1984.
8. Berg, O.G., von Hippel, P.H. Selection of DNA binding sites by regulatory proteins. J. Mol. Biol. 193:723–750, 1987.
9. Stormo, G.D., III G.W. Hartzell Identifying protein-binding sites from unaligned DNA fragments. Proc. Natl. Acad. Sci. USA 86:1183–1187, 1989.
10. Kendall, M., Stuart, A. "The Advanced Theory of Statistics." Vol. 2. New York: Macmillan Publishing Co, Inc., 1979: 1–716.
11. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum like-

lihood from incomplete data via the EM algorithm. J. R. Statist. Soc. Series B 39:1–38, 1977.
12. Little, R.J.A., Rubin, D.B. "Statistical Analysis With Missing Data." New York: John Wiley & Son, 1987: 1–278.
13. Goodman, L.A. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61:215–231, 1974.
14. Aitkin, M., Rubin, D.B. Estimation and hypothesis testing in finite mixture models. J. R. Statist. Soc. Series B 47: 67–75, 1985.
15. Berg, O.G., von Hippel, P.H. Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. J. Mol. Biol. 200:209–223, 1988.
16. Crombrugghe, B.D., Busby, S., Buc, H. Cyclic AMP receptor protein—role in transcription activation. Science 224: 831–838, 1984.
17. Seeman, N.C., Rosenberg, J.M., Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. Proc. Natl. Acad. Sci. USA 73:804–808, 1976.
18. Liu-Johnson, H.N., Gartenberg, M.R., Crothers, D.M. The DNA binding domain and binding angle of E. coli CAP protein. Cell 47:995–1005, 1986.
19. Brosius, J., Cate, R.L., Perlmutter, A.P. Precise location of two promoters for the β-lactamase gene of PBR 322. J. Biol. Chem. 257:9205–9210, 1982.
20. Marians, J.M., Soeller W., Zipursky, S.L. Maximal limits of the Escherichia coli replication factor Y effector site sequences in pBR 322 DNA. J. Biol. Chem. 257: 5656–5662, 1982.
21. Raibaud, O., Vidal-Ingigliardi, D., Richet E. A complex nuculeoprotein structure involved in activation of transcription of two divergent Escerichia coli promoters. J. Mol. Biol. 205:471–485, 1989.
22. O'Neill, M.C. Consensus methods for finding and ranking DNA binding sites. JMB 207:301–310, 1989.
23. Redner, R.A., Walker, H.F. Mixture densities maximum likelihood and EM algorithm. SIAM Rev. 26:195–239, 1984.
24. Warwicker, J., Engelman, B.P., Steitz, T.A. Electrostatic calculations and model-building suggest that DNA bound to CAP is sharply bent. Proteins Struct. Funct. Genet. 2: 283–289, 1987.
25. Weber, I.T., and Steitz, T.A. Model of specific complex between catabolite gene activator protein and β-DNA suggested by electrostatic complementarity. Proc. Natl. Acad. Sci. USA 81:3973–3977, 1984.
26. Lewin, R. Genome projects ready to go. Science 240:602–604, 1988.

APPENDIX

The conditional probability that the site of interest begins at position k in the n^{th} sequence, given the estimates of the parameters at the end of the q^{th} iteration, $\rho^{(q)}$, and the sequence data is calculated as follows.

If the site starts in position k, then the l^{th} base in the sequence will be in the site if $k \leq l \leq k + J$. Otherwise, it will be outside of the site. Also, the j^{th} base in the site will be located at position $l = j + k - 1$ of the sequence. Let $\nu_{b,l,n}$ be an indicator variable for base b in position l of sequence n, and let Δ represent the set of positions not in the site. Also, let $Y_{n,k}$, a position indicator variable, equal 1 if a site starts a position k in sequence n and O otherwise. Then the probability of observing the data in the n^{th} sequence, S_n, if the site starts at positions k and the population base frequency estimates are $\rho^{(q)}$ is:

$$P(S_n|\ Y_{n,k} = 1, \rho^{(q)}) = \prod_{j=0}^{J} \prod_{b=A}^{T} \rho_{bj}^{\nu_{bj'n}} \qquad (A1)$$

where

$$j' = j + k - 1, \text{ and } \upsilon_{bj',n} = \upsilon_{bj+k-1,n} \text{ for } j = 1, 2, \ldots J$$

and

$$\upsilon_{bj',n} = \sum_{l \epsilon \Delta} \upsilon_{b,l,n} \text{ for } j = 0.$$

Applying Bayes formula, we calculate the probability that the site starts at position k as follows:

$$P(Y_{n,k} = 1|\hat{\rho}^{(q)}, S_n) = \frac{P(S_n|\ Y_{n,k} = 1, \hat{\rho}^{(q)}) * P^0(Y_{n,k})}{\sum_{k=0}^{L-1} P(S_n|Y_{n,k} = 1, \hat{\rho}^{(q)}) * P^0(Y_{n,k})}, \qquad (A2)$$

where $\hat{\rho}^{(q)}$ are the estimates of the population base frequencies after (q) iterations of the algorithm, and $P^O(Y_{n,k})$ are the prior probabilities that the site is located at position k in sequence n. To avoid solutions near the boundaries, we do not estimate these values. Thus the method presented here is a special case of the methods presented by Aitkin and Rubin.[14] In this application, we assume that at the outset we are completely ignorant about the positions of the sites in all sequences, i.e., that the site is equally likely to occur in any position, $P^O(Y_{n,k}) = 1/(L - J + 1)$, $k = 1 \ldots (L - J + 1)$.