J. Mol. Biol. (1992) **223**, 159–170

Expectation Maximization Algorithm for Identifying Protein-binding Sites with Variable Lengths from Unaligned DNA Fragments

Lon R. Cardon†

Institute for Behavior Genetics, University of Colorado Boulder, CO 80309-0447, U.S.A.

and Gary D. Stormo

Department of Molecular, Cellular, and Developmental Biology University of Colorado, Boulder, CO 80309-0347, U.S.A.

(Received 20 May 1991; accepted 3 September 1991)

An Expectation Maximization algorithm for identification of DNA binding sites is presented. The approach predicts the location of binding regions while allowing variable length spacers within the sites. In addition to predicting the most likely spacer length for a set of DNA fragments, the method identifies individual sites that differ in spacer size. No alignment of DNA sequences is necessary. The method is illustrated by application to 231 *Escherichia coli* DNA fragments known to contain promoters with variable spacings between their consensus regions. Maximum-likelihood tests of the differences between the spacing classes indicate that the consensus regions of the spacing classes are not distinct. Further tests suggest that several positions within the spacing region may contribute to promoter specificity.

Keywords: promoters; DNA–protein; Expectation Maximum; multiple alignment; consensus sequences

1. Introduction

Transcription is often regulated by proteins binding to specific regions of DNA and affecting the expression of nearby genes. Because of the importance of the binding sites in regulation and of the time-consuming experiments necessary to identify them genetically or biochemically, several different statistical approaches have been developed to attempt to identify the binding sites using only the sequence data. There are many methods that utilize a set of known binding sites to extract important residue patterns from the DNA fragments and provide a representation of the binding sites that can be used to locate new examples with reasonable reliability (Berg & von Hippel, 1987; Mulligan & McClure, 1986; Staden, 1984; Stormo, 1988, 1990*a*).

A more difficult problem, but more valuable if solved, is to be able to identify the binding sites and determine their essential features from unaligned DNA fragments. The data for this type of problem are a set of DNA fragments each known to contain at least one binding site for some protein, perhaps from genetic mapping experiments or from binding studies on restriction fragments. The important point is that the binding site need only be known to be located somewhere within the fragment, and the actual alignment of the sites is unknown. The problem is difficult, in part, because of the degeneracy in the binding sites of most regulatory proteins. If the fragments were about 200 bp‡ long and the common binding site were a completely conserved hexamer, it would not be difficult to identify (assuming the fragments were otherwise unrelated). However, a typical regulatory protein may have binding sites consisting of some highly conserved domains, with some positions more highly conserved than others, and perhaps none of the positions in the sites is absolutely conserved. For example, in *Escherichia coli* promoter sites, the consensus patterns TTGACA and TATAAT

† Author to whom all correspondence should be addressed.

‡ Abbreviations used: bp, base-pair(s); EM, Expectation Maximization; CRP, cyclic AMP receptor protein.

0022-2836/92/010159-12 $03.00/0

typically appear at approximately 35 and 10 bases before the site of transcription initiation (Pribnow, 1975; Rosenberg & Court, 1979; Hawley & McClure, 1983), but none of these bases is absolutely conserved; in fact, only about 65% of all known promoters perfectly match even the most highly conserved region within these promoters (−10: TAxxxT: Harley & Reynolds, 1987). Furthermore, the spacing between the −35 and −10 is not completely conserved; 17 bases is the most common spacing but several others are observed. The diversity in consensus matching imposes some limitations upon methods that focus on specific consensus descriptions or on finding "words" within segments of DNA (Pearson & Lipman, 1988).

Some attempts have been made to solve the problem of finding optimal local alignments of multiple sequences, each with particular limitations. Two methods require that the binding sites be approximately aligned initially, and then use "word" matching criteria to find the most conserved regions (Galas *et al.*, 1985; Mengeritsky & Smith, 1987). These methods have the advantage that they can identify several conserved domains even if the spacing between them is variable, as long as all of the domains are initially aligned within a specified window of allowable shifting between the sequences. Another approach uses a "greedy" algorithm, which searches for an alignment of the fragments that maximizes the "Information Content" of the sites (Stormo & Hartzell, 1989; Hertz *et al.*, 1990). This method has the advantages that an initial alignment of the fragments is not required and it returns several alignments ranked by their significance. But it also has the disadvantage that only fixed-length domains are identified, so regulatory sites that consist of multiple regions separated by variable spacing will be missed. Further manipulations by hand, utilizing both optimal and suboptimal alignments, can be used to identify such binding sites (Stormo, 1990*b*), but this procedure is far from automated and may not be very reliable.

An Expectation Maximization (EM) method has also been applied to this problem (Lawrence & Reilly, 1990*a*). The initial EM method only allowed for fixed-length blocks and would, therefore, miss sites composed of multiple blocks with variable spacing between them. Here, we present an extention to the EM method that simultaneously predicts binding sites and aligns DNA fragments while allowing variability in the spacing regions between protein contact sites. We demonstrate the use of the method on *E. coli* promoters, a large, well-studied data set with the properties of conserved binding domains separated by variable length spacers (Harley & Reynolds, 1987). The present EM procedure also provides a method by which differential base conservation among different spacing classes may be tested using the robust statistical methods of maximum likelihood. We apply the method to *E. coli* fragments to address the issue of differential base conservation as a function of spacer length (O'Neill, 1989*a*,*c*).

2. The Expectation Maximization Algorithm

Expectation Maximization is a two-step iterative procedure for obtaining the global maximum likelihood parameter estimates for a model of observed data (Little & Rubin, 1987). The algorithm involves calculating expected parameter values that describe the data (step 1), then maximizing the likelihood of observing those values (step 2). These steps are repeated until the parameters that best explain the data are obtained. The details of our EM algorithm are given below, but a simple example can serve to illustrate the basic idea of EM. Given a collection of aligned binding sites for some protein, it is easy to determine a matrix that represents the specificity of that protein (for reviews, see Stormo, 1988, 1990*a*). On the other hand, if one is given a specificity matrix for that protein, it is easy to predict where the binding sites would be on a collection of unaligned fragments. In the problem under consideration, we are given a collection of fragments, each known to contain a binding site for some protein, but we know neither the alignment of the sites (their positions on each fragment) nor the specificity matrix of the protein and we want to determine both simultaneously. The EM approach is to alternate between the two methods just described, first using an initial guess as to the locations of the sites (we start by assuming all possible binding sites are equally likely) to derive a specificity matrix, and then using that matrix to re-estimate the locations of the sites. This procedure is iterated until convergence. While it is not difficult to make up data sets that lead EM to converge inappropriately, we think that most real problems will contain suitable data. We expect that the criteria for data being suitable are that the binding sites can be represented accurately by a specificity matrix, and that the proper alignment of the binding sites on the fragments give a more significant matrix than any other possible alignment. The statistical assumptions of the EM method with respect to DNA sequence data follow those described by Little & Rubin (1987) for finite mixture models.

Below we use the example of *E. coli* promoters to illustrate the use of EM. In this example, we take advantage of prior knowledge about the size of promoters and their general characteristic of relatively conserved domains separated by a variable length spacer. In general, one might not have any prior knowledge of the binding sites being sought. In that case, the EM analysis could be repeated using different models for the binding sites. One advantage of the EM technique over other probability-based sequence analysis methods is that different models of the protein–DNA interaction can be compared directly and quantitatively to determine which is the best match to the data. For example, the EM model can account for structural properties of proteins involved in DNA binding by allowing correlated effects of residues across multiple binding site positions. This allows one to

Fragment	Sequence
frdabcd(16)	gatctcgtcaaATTTCAgacttatcgatcagacTATAATgttgtacctataaagga
tyrt/212(16)	gATCATAcctacacagctgaagaTATGATgcgcgcaggtcgtgacg
arabad(16)	ttagcggatcctacCTGACGctttttatcgcaactcTCTACTgtttctccatacccgtt
ampc/c16(17)	gctatcTTGACAgttgtcacgctgattggTATCGTtacaatctaacgtatcg
lambdapo(17)	tacctctgccgaagTTGAGTatttttgctgtatttgtCATAATgactcctgttgatagat
tn2661bla-pb(17)	cctcGTGATAcgcttatttttataggtTAATGTcatgataataatggttt
rpod-pb(18)	agccaggtCTGACCaccgggcaacttttagagCACTATcgtggtacaaat
cit.util-431(18)	gacaggcacagcaTTGTACgatcaactgatttgtgccAATAATtaaatgaaatcac
tn10pin(18)	tcattaagTTAAGGtggatacacatcttgtcaTATGATcaaatggtttcgcgaaa

Figure 1. Examples of *E. coli* DNA fragments containing protein-binding sites. Spacing classes are indicated in parentheses. Capitalized bases show consensus regions.

test whether a binding site is best represented by a direct or inverted repeat, or whether the total collection of binding sites is best separated into distinct classes, or other possible models of the protein–DNA interaction. The EM algorithm also generates estimates of residue frequencies for bases not included in the binding sites. Consequently, if the regions around the binding sites are characterized by overall base frequencies different from those within the sites, the EM technique will exploit the frequency differences.

The frequency estimates in the EM model are derived using the "missing information principle" in the context of the maximum likelihood estimation procedure (Edwards, 1972). Because DNA segments are unaligned when positional binding site information is unknown, the EM model treats the location of promoter sites as "missing". For a set of N DNA fragments, each with k possible promoter site locations, there are k^N possible combinations of promoters. The primary task is to identify the correct N binding sites and characterize the frequency of residues within those sites. This is performed by iteratively solving a series of probability/expectation equations and maximizing the likelihood of the equations with respect to the observed DNA segments.

Figure 1 shows some examples of DNA fragments from *E. coli* that may be examined by the EM algorithm. All of these fragments contain variations of the -35 and -10 consensus regions TTGACA and TATAAT (indicated in capital letters), but none is absolutely conserved. The EM technique is designed to identify the binding sites on each sequence and estimate the overall base frequencies for each position in the sites.

Lawrence & Reilly (1990*a*) applied the EM model to 18 DNA fragments from *E. coli* containing known binding sites for the cAMP receptor protein (CRP). Several constrained parameterizations of the EM algorithm also were applied to the CRP data (e.g. testing palindrome patterns by forcing equality of certain parameters). The simplest monoresidue model correctly identified 16 known primary binding sites in the 18 DNA fragments. Constrained representations of the binding sites permitted characterization of the major and minor groove openings in the fragments with respect to the CRP proteins. The findings reported by Lawrence & Reilly (1990*a*) illustrate the strength of the EM method for identifying binding sites in DNA fragments and characterizing structural features of the binding domains in relation to the binding proteins.

The original formulation of the EM algorithm contained no provision for variable spacing lengths between primary binding regions. If modified to allow variable spacing, the EM algorithm should be able to distinguish between promoters having different numbers of bases between consensus regions, such as the 16, 17 and 18 bp examples in

Figure 1, while retaining the ability to identify promoter start sites and characterize fragments by positional base composition. Given the utility of the EM method for analysis of binding sites in sequences having similar spacing lengths, and because of the importance of spacing in promoter function, it is reasonable to expect that the strengths of the EM procedure may be enhanced substantially if modified to account for different spacer lengths. Here, we describe an extension of the EM algorithm to account for such effects. Lawrence & Reilly (1990*b*) have independently developed a slightly different extension of EM to handle the case of variable length spacer regions. They show how it can be used to identify CRP binding sites with rare spacer lengths between the highly conserved domains of the binding sites.

3. Methods

(a) *EM algorithm with variable spacer lengths*

To allow variable spacer lengths in the EM algorithm, DNA fragments containing protein-binding regions to be identified are conceptualized as being composed of 2 structural components: (1) the conserved regions of protein binding containing an intervening segment of variable length; and (2) the segment of DNA outside the conserved regions. Recent analysis of the CRP data illustrate this classification scheme. Stormo & Hartzell (1989) calculated the "information content", a probability-based measure of protein specificity similar to the EM method, for each residue in a CRP-binding region 22 bases in length. Their findings revealed high information content values for positions 1 to 8 and 15 to 22 of the binding site, coupled with modest values for site positions 9 to 14. In terms of the present EM technique, positions 1 to 22 would be classified as conserved binding sites that contain a spacing region from positions 9 to 14 (category 1), and all unspecified regions of the DNA fragments would be classified as occurring outside the overall binding site (category 2). In order to allow variable spacer lengths in the EM algorithm, the probability of observing bases in both categories must be maximized while allowing the first category to vary in length.

To formulate the EM algorithm using the classification scheme just outlined, we consider the overall binding region, which consists of consensus positions and the intervening spacer, to be a random variable of length J. We specify G as a similar random variable for the set of all possible spacer lengths within J. In *E. coli* promoters, the most highly conserved regions are expected to span 6 bases at each end of the site and the spacer is expected to vary between 15 and 21 positions, although most promoters have spacings in the range 16 to 18. Thus, $J = (27, 28, \ldots, 33)$, corresponding to $G = (15, 16, \ldots, 21)$. The model parameters to be estimated are $\rho_{b,j}$, the probabilities of observing each base, b, in each position, j, of the binding site J; and $\rho_{b,O}$, the probabilities of observing each base in the DNA region outside the overall binding sites. Additional parameters that define the probabilities of observing an overall binding site with spacer length g in the set $G[\rho_g = P(G = g)]$ are then added to the set of unknowns to fully characterize the model. Note that the overall residue frequencies, not at particular positions, are of interest for DNA regions outside binding sites ($\rho_{b,O}$), whereas position and frequency are potentially important considerations within binding sites ($\rho_{b,j}$).

The Expectation step of the EM algorithm is designed to calculate expected values for the parameters $\rho_{b,j}$, $\rho_{b,O}$, and ρ_g. Bayes' theorem (e.g. see Kendall & Stuart, 1977) provides the foundation for these probabilities. The probabilities are calculated by initially expressing the model in terms of the conditional probability of observing each sequence S_n, given that the binding site begins at position k of the nth sequence and the spacer is g bases long, then rearranging the conditional probability (using Bayes' theorem) to yield the probability that the binding site begins in position k, given the sequence S_n and spacer $g(1 \leq k \leq L_n - J + 1$; L_n is the number of observed positions in sequence S_n). It is useful to consider the former conditional probability as consisting of 2 components: (1) the probability of observing the binding region in sequence S_n, given that the site begins in position k of S_n and has spacer length g; and (2) the probability of observing the non-binding region in S_n conditional on the site beginning at position k and the spacer length g. These conditional probabilities may be formulated in the following manner.

Let $Y_{n,k}$, a position indicator, equal 1 if the binding site in S_n begins at k; $Y_{n,k} = 0$ otherwise. Then the first conditional probability described above may be expressed as the product of the probabilities of observing each base in the overall binding region:

$$P(S_{n,j} | Y_{n,k} = 1, G = g) = \prod_{j=1}^{J} \rho_{b,j}^{v_{bj'n}} \tag{1}$$

where $j' = j + k - 1$ and:

$$v_{bj'n} = \begin{cases} 1 & \text{if } S_{n,j'} \in J \text{ and } S_{n,j'} = b \\ 0 & \text{otherwise} \end{cases}$$

for each base b in each position j' of sequence S_n. These j' are the positions in S_n that constitute the binding site, from $j = 1$ to J. For each J, the second probability described above is the product of the probabilities of observing each base outside the binding region:

$$P(S_{n,O} | Y_{n,k} = 1, G = g) = \prod_{b=A}^{T} \rho_{b,O}^{\omega} \tag{2}$$

in which $\omega = \sum_{l \in \Delta} S_{n,O}^{(b)}$ for all bases, b, in positions, l, in the set of residues outside the binding site, Δ. Thus, the overall probability of observing sequence S_n conditional upon the binding start position k and the spacer length g is expressed simply as the product of expressions (1) and (2):

$$\begin{aligned} P(S_n | Y_{n,k} = 1, G = g) &= P(S_{n,j} | Y_{n,k} = 1, G = g) P(S_{n,O} | Y_{n,k} = 1, G = g) \\ &= \prod_{b=A}^{T} \rho_{b,O}^{\omega} \prod_{j=1}^{J} \rho_{b,j}^{v_{bj'n}}. \end{aligned} \tag{3}$$

It should be noted that when the spacer length is fixed (i.e. G is a unit set), equation (3) is equivalent to equation (A1) given by Lawrence & Reilly (1990*a*) for the fixed spacer length situation.

Bayes' theorem may then be applied to the conditional

probability shown in equation (3) to obtain the probability that the overall binding site begins in position k, given each observed DNA fragment S_n and a spacer length g:

$$\begin{aligned} P(Y_{n,k} = 1|S_n, G = g) &= \frac{P(S_n|Y_{n,k} = 1, G = g)P(Y_{n,k} = 1)\rho_g}{\sum_{k=1}^{L_n-J+1} P(S_n|Y_{n,k} = 1, G = g)P(Y_{n,k} = 1)\rho_g} \\ &= \frac{\prod_{b=A}^{T} \rho_{b,0}^{\omega} \prod_{j=1}^{J} \rho_{b,j}^{v_{bj'n}} P(Y_{n,k} = 1)\rho_g}{\sum_{k=1}^{L_n-J+1} \prod_{b=A}^{T} \rho_{b,0}^{\omega} \prod_{j=1}^{J} \rho_{b,j}^{v_{bj'n}} P(Y_{n,k} = 1)\rho_g} \\ &= \frac{\prod_{b=A}^{T} \rho_{b,0}^{\omega} \prod_{j=1}^{J} \rho_{b,j}^{v_{bj'n}} \rho_g}{\sum_{k=1}^{L_n-J+1} \prod_{b=A}^{T} \rho_{b,0}^{\omega} \prod_{j=1}^{J} \rho_{b,j}^{v_{bj'n}} \rho_g}. \end{aligned} \quad (4)$$

If a protein binding site is perfectly predicted by the EM model, this quantity will equal 1·0 for the position k that is the beginning of the site. For all other k, the probability will equal 0·0. The quantities $P(Y_{n,k} = 1)$ cancel in this derivation because we assume, *a priori*, that all possible binding start sites are equally likely within each DNA fragment $[P(Y_{n,k} = 1) = 1/(L_n - J + 1)]$ to avoid solutions near the boundaries.

The conditional probability shown in equation (4) forms the basis for the Expectation step of the EM algorithm with variable spacer lengths. The expected number of bases in each position of the binding site(s) may be calculated as:

$$\begin{aligned} \varepsilon_{b,j} &= E(n_{b,j}|S) \\ &= \sum_{g \in G} \sum_{n=1}^{N} \sum_{k=1}^{L_n-J+1} v_{bj'n} P(Y_{n,k} = 1|S_n, G = g)\rho_g \end{aligned} \quad (5)$$

and:

$$\begin{aligned} \varepsilon_{b,0} &= E(n_{b,0}|S) \\ &= \sum_{g \in G} \frac{\sum_{n=1}^{N} \sum_{k=1}^{L_n-J+1} \omega^{(bn)} P(Y_{n,k} = 1|S_n, G = g)\rho_g}{L_n - J + 1}. \end{aligned} \quad (6)$$

Thus, for each position in the overall binding site, the expectations are calculated by adding the probability that the site begins in position k to the accumulating number of bases in each position of the site. For example, if the first 2 bases of the window that begins in position 15 of some sequence S_n are observed to be T and A, and the probability that the binding site begins in position 15 is 0·10 (from eqn (4)), then 0·10 is added to the accumulating number of expected T residues in the first position and A residues in the second position. This procedure is repeated for all possible windows in the set of observed sequences.

The expectations are used to calculate the probabilities for the qth iteration of the Maximization step in the algorithm:

$$\rho_{b,j}^{(q)} = \frac{\varepsilon_{b,j}}{N} \quad (7)$$

$$\rho_{b,0}^{(q)} = \frac{\varepsilon_{b,0}}{N}. \quad (8)$$

Finally, the probability of observing a binding region with spacer length g for the qth iteration is given by:

$$\rho_g^{(q)} = \frac{\sum_{n=1}^{N} \sum_{k=1}^{L-J+1} P(S_n|Y_{n,k} = 1, G = g)\rho_g^{(q-1)}}{\sum_{g \in G} \sum_{n=1}^{N} \sum_{k=1}^{L-J+1} P(S_n|Y_{n,k} = 1, G = g)\rho_g^{(q-1)}}. \quad (9)$$

This probability serves both as a frequency description of the different spacing groups and as a gap penalty in the alignment. Unlikely or uncommon spacer lengths yield small values of ρ_g, which decrease the probabilities of observing sequences with those uncommon spacers (see eqn (4)). Unusual spacer lengths will still be observed, provided that matches in the conserved domains are sufficient to compensate for the gap penalty.

The maximum likelihood estimates of the probabilities are those that maximize the log-likelihood equation:

$$\log L = \sum_{n=1}^{N} \left(\sum_{b=A}^{T} \left[\sum_{j=1}^{J_{\max}} f_{b,j} \ln(\rho_{b,j}) + (L_n - J_{\max}) f_{b,0} \ln(\rho_{b,0}) \right] + \sum_{g \in G} f_g \ln(\rho_g) \right). \quad (10)$$

In this equation, $f_{b,j}$ and $f_{b,0}$ are the observed base frequencies for all fragments examined. Frequencies of sequences with different spacing lengths are represented by f_g. If positional binding site data were known and the fragments aligned with respect to the binding information, the maximum likelihood estimates of $\rho_{b,j}$, $\rho_{b,0}$, and ρ_g would be the sample frequencies $f_{b,j}$, $f_{b,0}$, and f_g.

The free parameters in this model include $\rho_{b,j}$ and ρ_g; all other expectations and probabilities may be calculated using these unknown quantities, as shown in equations (4) to (8). Upon convergence of the EM algorithm, the free parameters and posterior probabilities for each DNA fragment contain valuable information about the location and length of the protein-binding site(s) in each segment of DNA. The $\rho_{b,j}$ parameters indicate the extent to which each residue located within the binding region is conserved, the posterior probabilities point to the most likely start position of the binding site in each DNA fragment, and the accompanying spacer length probabilities (ρ_g) indicate the most likely length of the conserved binding region. The obvious consequence of this treatment of variable spacing is that the most probable spacer length for all the DNA sequences examined will appear as the highest estimated spacer length parameter. A less obvious consequence is that each DNA segment has associated with it an expected protein-binding site for all possible spacer lengths in the *a priori* defined set G, which allows probable binding sites that contain spacers that differ in length from the "consensus spacer length" to be identified by the algorithm.

(b) *Application to* E. coli *promoters*

In order to assess the accuracy of the EM algorithm in locating, aligning and characterizing protein-binding regions, we applied the method to *E. coli* DNA fragments known to contain promoter regions. Promoters from *E. coli* provide an extremely useful set of sequences for examining different spacing classes because they comprise an extensive and well-defined group of DNA fragments. Harley & Reynolds (1987) have compiled a large set of promoters and have described the consensus properties of this reference group. Their alignment is useful, but cannot be considered proven experimentally because the polymerase–DNA contacts have been determined for only a few promoters. It is based on an initial alignment (Hawley & McClure, 1983) and an iterative procedure to find consistent "weight matrices" for the −10 and −35 regions. For the present application, 231 DNA fragments

from the Harley & Reynolds (1987) compilation were examined†. Of these 231 sequences, 50 (22%) were classified as having 16 bp spacing, 41 (18%) with 18 bp spacing, and 122 (53%) with 17 bases separating the conserved binding regions. A small number of the *E. coli* fragments also have spacing regions spanning 15, 19, 20 and 21 bp (4, 6, 1 and 7 sequences, respectively). All fragments roughly comprise positions −50 to +10 with respect to known transcriptional start points.

E. coli promoters are the largest well-characterized set of binding sites. For most of them, the actual positions of the initiation site of transcription are known. This information could be used to provide an approximate alignment of the promoters because it is known that the conserved regions occur within a narrow range of spacings (3 to 11 bp) upstream from those start points. However, we have decided to not use the initiation site information in our EM analyses (except in 1 analysis described later) because we want to test the method on a problem where no alignment information is provided. That is, we treat the data of the promoter-containing sequences (from Table 1 of Harley & Reynolds (1987) with the exceptions noted above) as though the promoter site could occur anywhere within the fragment. We do use the prior information that promoter sites vary in length from 27 to 33 bp (from the beginning of the −35 to the end of the −10) to avoid numerous exploratory models that would be needed if we knew nothing about the characteristics of the sites. We choose to ignore the fact that there are weak conservations observed for positions outside of the −35 to −10 region. We also have not tried to separate out promoters that require an activator protein to give substantial transcription, so the collection probably contains several promoters that have intrinsically low activity. Finally, in determining the alignment of promoters with different spacer lengths we have not attempted to scatter gaps around to provide any overall optimal alignment. Rather, we have always placed the gaps at defined positions in the middle of spacer regions. The promoters with 21 bp spacers have no gaps within them. These are aligned with the 20 bp spacer promoters by aligning a gap with the 9th position of the 21 bp spacer. For the promoters with shorter spacers, the gaps are clustered beginning at the same position. This means the 15 bp spacer class, the shortest, has the first 8 bases of the spacer aligned with those of the other classes, followed by 6 gap positions, and the last 7 bases aligned with the last 7 of all the other classes. This alignment is essentially identical with that reported by Harley & Reynolds (1987).

4. Results

The accuracy of the EM algorithm was assessed by applying the variable length formulation allowing 15 to 21 bp spacing to the 231 *E. coli* fragments. For this analysis, we consider a predicted promoter position to be correct if it occurs with a spacing of 3 to 11 from a known initiation position. This is the same criterion used by Harley & Reynolds (1987) in their alignment. For the 12 promoters without known initiation sites, we consider our prediction correct if it is consistent with that of Harley & Reynolds (1987). By these criteria, the full EM model on the pooled fragments resulted in correct identification of 87% (200/231) of the promoter positions. The maximum likelihood estimates of the frequencies of promoters with spacing lengths of 16, 17 and 18 bp were 0·25, 0·51 and 0·16. These estimates are quite similar to the expected values of 0·22, 0·53 and 0·18 based on the classifications by Harley & Reynolds (1987). Estimated spacing frequencies of the 15, 19, 20 and 21 bp promoters were 0·001, 0·001, 0·03 and 0·04, respectively, which also are similar to the expected frequencies of 0·02, 0·03, 0·004 and 0·03. The EM algorithm appears to account well for variability in spacing between protein-binding regions.

It should be noted that in the original compilation of these *E. coli* fragments, Harley & Reynolds (1987) pointed out that for several of the promoters, the most likely binding site differed depending on whether information about the transcriptional initiation site was included in the alignment. As presently specified, the EM algorithm does not take into account transcriptional start site information, although this is a natural extension of the method that we discuss below. Several of the missed promoter sites in the present application also were noted by Harley & Reynolds (1987) as being the best sites when initiation site information is ignored. Inclusion of these sites in our correct prediction set yielded 92% (212/231) correct promoter alignments.

The specific binding sites identified by the variable-spacing EM algorithm are of interest because the posterior probabilities generated by the method indicate how well the promoter start site is defined by the consensus bases: high values indicate a good match with consensus positions, and low values suggest that the sites are not very similar to the consensus regions. Our results show that many of the promoters yield probabilities approaching 1·0 for the start site, suggesting that the location of the promoter is quite certain. Other promoters yield probabilities much smaller for the most likely site, however, which suggests that the promoter site is not well defined by the consensus of residues. This may be due to the best site not being a good match to the consensus, or it could be due to there being several good matches to the consensus. Since the sum of the posterior probabilities is set to 1·0, several good matches to the consensus means none will have an especially high probability and this may indicate that the fragment actually contains more than one promoter. A sample of sequences and their predicted promoters and posterior probabilities are presented in Figure 2. Also shown are several of the sequences in which the predicted promoter differs from that noted by Harley & Reynolds (1987). It is interesting that in several of these disparate sequences the predicted promoter sites are completely incompatible with initiation site

† A slightly larger number of fragments was available, but mutant forms of promoters were omitted from the analysis. Also, 13 sequences were highly redundant with other fragments in the data set; in these cases, 1 copy of each redundant pair was arbitrarily eliminated. The fragments eliminated include argI, colE1l0.13, colicinE1-P3, NTP1rna100, p15primer, pBRRNAI, rnp(RNaseP), rrnG-P1, rrnG-P2, RSFrnaI, and Tn2660bla-P3.

Fragment	Sequence	Probability
clodfrnai	acacgcggttgctcTTGAAGtgtgcgccaaagtccggcTACACTggaaggacagatttgg	.989
ompf	ggtaggTAGCGAaacgttagtttgaatggAAAGATgcctgcagacacataaa	.948
psc101orip2	attatcaTTGACTagcccatctcaattggTATAGTgattaaaatcacctaga	.991
fuma	gtactagtctcagtTTTTGTtaaaaaagtgtgtaggaTATTGTtactcgctttaacagg	.442
tn10xxxp3	ccatgatagaTTTAAAataacataccgtcagtatgttTATGGTatcatgatgatgtggtc	.192
pyre-p2	gtaggcggtcataCTGCGGatcatagacgttcctgttTATAAAaggagaggtggaagg	.309
pcolviron-p2	tgtttcaacaccATGTATtaattgtgtttatttgTAAAATtaatttctgacaataa	.371
spc	ccgtttatttttctACCCAtatccttgaagcggtgtTATAATgccgcgccctcgata	.390
trps	cggcgaggctatcgATCTCAgccagcctgatgtaattTATCAGtctataaatgacc	.323

Figure 2. Selected promoter sites and site probabilities from the EM algorithm. Capitilized bases show consensus regions from the Harley & Reynolds (1987) alignment. Predicted sites from the EM algorithm are underlined. Transcriptional start points are shown in italics if the sites are known.

information (i.e. the alignment places the promoter and initiation sites in overlapping locations). These may be fragments that contain more than one promoter, but for only one of which has the initiation site been determined.

It has been reported that there is a weak consensus of CAT at the initiation site (the A is the usual first base of the transcript), occurring with variable spacing from the −10 region (Hawley & McClure, 1983). To see if the inclusion of a conserved pattern at the initiation position would improve the predictions, we modified the previous EM algorithm to include a pattern of three bases, separated by a 2 to 10 bp variable spacer following the last base of the −10 region. Note that in this analysis we are still not using the known initiation positions, but simply adding the *a priori* knowledge that an additional conserved pattern, three bases wide, is found at a variable distance downstream from the −10 region. Application of this modified EM model yields 217/231 (94%) promoters identified correctly, by the criterion of being consistent with known initiation sites. Of the 14 that are inconsistent, seven of those also were identified as alternative alignments by Harley & Reynolds (1987).

In a separate analysis, we forced the program to find promoter sites that are consistent with the known initiation sites. We did this by simply removing regions of the fragments that would permit an inconsistent site, and ran the EM algorithm (without the modification described above that includes the initiation site pattern). All of the promoters identified in this manner are, of course, consistent with the known initiation sites, and nearly all are the same as those identified by Harley & Reynolds (1987), but in a few cases the −10 and −35 regions we find are different from theirs. By log-likelihood analysis, our alignment and theirs are statistically equivalent; i.e. neither one can be said to be better than the other. This result emphasizes the point that, at least for some promoters, statistical evidence alone cannot unambiguously determine the proper position of the promoter. Indeed, it may be that for some promoters the RNA polymerase–DNA interaction is not always exactly the same.

It has been suggested that *E. coli* promoters in different spacing classes also have distinct base compositions at various positions in the binding sites (O'Neill, 1989*a*). If this conjecture is correct, then the EM algorithm will do better by separating the promoters into spacing classes first. This will allow the positional base frequencies for each class to be determined separately and will show an increased significance compared to the previous analysis in which the positional base frequencies are a combination from all spacing classes mixed together. To test this possibility, we applied the EM model to the fragments in each of the 16, 17 and 18 bp spacing classes, as defined by Harley & Reynolds (1987), and compared the results to those obtained from analyses of the same promoters pooled together†.

† The small number of fragments within the 15, 19, 20 and 21 bp spacing classes precluded similar comparisons for the fragments in these groups.

Table 1

Tests of sequence homogeneity for E. coli *promoters with spacer lengths of 16, 17 and/or 18 bases*

Model description	LL	N Param	*versus* model	χ^2	df	p
(1) 16 bp promoters with fixed spacer ($N = 50$)	−3577·596	84				
(2) 17 bp promoters with fixed spacer ($N = 122$)	−8980·618	87				
(3) 18 bp promoters with fixed spacer ($N = 41$)	−3029·513	90				
(4) All promoters with variable spacer lengths (16 to 18; $N = 213$)	−15,859·096	92	1+2+3+k†	126·724	169	>0·95

LL, log-likelihood; N Param, number of parameters estimated by the model.

† k is a constant added to the fixed spacer models that represents the observed frequencies of each spacing class: $k = N\sum_{g=16}^{18} f_g \ln(f_g)$. For this constant, N equals 50+122+41 = 213, and the frequencies (f_g) of the 16, 17 and 18 spacing classes are 0·24, 0·57 and 0·19, respectively. Thus $k = -208·007$.

Likelihood ratio tests formed the basis for these comparisons.

The likelihood ratio test is calculated as −2 times the difference between the log-likelihoods of two nested models and is asymptotically distributed as a chi-square statistic with degrees of freedom equal to the difference in the number of parameters estimated by the two models. Statistical significance of chi-square values may be assessed by comparison to critical values found in most statistics textbooks (e.g. see Beyer, 1988) to determine the consistency of the model with the observed DNA fragments. In the present application, the likelihood ratio test was used to assess the similarity, or homogeneity, of sequence compositions by obtaining log-likelihood values (using eqn (10)) for sequences in each spacing class and comparing the sum of those values with the log-likelihood generated from the pooled sequences with variable spacer lengths. In addition to the test of spacing class homogeneity, the positions in the spacer regions were examined to identify those that may be important components of promoter specificity.

The results of applications of the EM model to specific and combined promoter spacing classes are presented in Table 1. Models 1, 2 and 3 represent applications to the 16, 17 and 18 bp spacer length promoters, respectively. In these models, the spacer lengths were fixed on the basis of known spacing class and all positions within the spacer regions were allowed to exhibit different frequencies. In model 4, the 213 sequences in these groups were combined and the spacer length allowed to vary from 16 to 18 bases. The chi-square homogeneity test of the data indicated that the sequences are, indeed, similar as regards positional base composition of protein binding sites ($\chi^2_{169} = 126·724$, $p > 0·95$). These results suggest that the overall promoter regions in *E. coli* fragments do not have distinct consensus sequences and, therefore, need not be separated on the basis of spacing class.

The probabilities for each spacing class and for the pooled data are listed in Table 2. As expected from the apparent sequence similarities, the −35 and −10 regions of high conservation do not differ dramatically between the 16, 17 and 18 bp spacing groups. These similarities were noted in the original compilation of the fragments by Harley & Reynolds (1987). In general, the base frequencies within the spacer regions also are similar for all spacer lengths, although some positions differ between spacing groups. For example, in position 9, the 16 and 17 class promoters reveal somewhat high frequencies of T, but the 18 class fragments show thymine as the least frequent base in this position. Results from the likelihood ratio test suggest that these differences are within the range of expected fluctuations and that, overall, there is insufficient evidence to justify the separation of promoters into distinct classes. However, the EM algorithm does not identify every promoter correctly, both when performed on each class separately and on the combined set, and it may be that properly aligned promoters would show significant spacing-class dependent positional base frequencies.

To further explore the finding of compositional sequence similarity for promoters having different spacer lengths, we examined the data in the original published alignments (Harley & Reynolds, 1987). Contingency tables of "spacing class" by "base frequency" by "position" were tested using the method of log-linear analysis (Bishop *et al.*, 1975). All positions in the fragments, both within and outside the consensus regions (−50 to +10 relative to transcription start sites), were included in this analysis. The log-linear method provides a likelihood ratio chi-square value that indicates whether or not the distributions of residues across positions are contingent upon spacing class. Results of the log-linear analysis support those of the EM model comparisons in revealing little evidence for differential base frequencies as a function of promoter spacing class ($\chi^2_{304} = 324·58$, $p > 0·20$).

In addition to the global test of spacing-class related frequency differences, we calculated contingency chi-square values for the fragments, examining each position separately to assess specific position-composition differences. Although this type

Table 2

Expected base frequencies for E. coli *promoters by spacing class (16, 17 or 18)*

Location	Spacing class	A	C	G	T	Location	A	C	G	T
1	16	0·06	0·08	0·08	0·78	14	0·20	0·26	0·28	0·26
	17	0·05	0·07	0·06	0·81		0·24	0·28	0·22	0·26
	18	0·00	0·20	0·03	0·78		0·30	0·15	0·20	0·36
	Combined	0·04	0·08	0·08	0·80		0·26	0·24	0·26	0·24
2	16	0·04	0·12	0·00	0·84	15	—	—	—	—
	17	0·11	0·05	0·07	0·77		0·28	0·21	0·22	0·29
	18	0·12	0·00	0·07	0·80		0·19	0·29	0·12	0·39
	Combined	0·09	0·04	0·07	0·81		0·25	0·24	0·18	0·33
3	16	0·02	0·08	0·70	0·20	16	—	—	—	—
	17	0·01	0·11	0·72	0·15		—	—	—	—
	18	0·07	0·19	0·51	0·22		0·34	0·22	0·05	0·39
	Combined	0·03	0·13	0·67	0·17		0·53	0·14	0·15	0·18
4	16	0·50	0·12	0·06	0·31	17	0·26	0·12	0·32	0·30
	17	0·56	0·18	0·06	0·20		0·25	0·26	0·20	0·28
	18	0·68	0·10	0·12	0·10		0·20	0·12	0·39	0·29
	Combined	0·56	0·18	0·05	0·21		0·28	0·21	0·21	0·29
5	16	0·22	0·58	0·10	0·10	18	0·31	0·22	0·20	0·28
	17	0·30	0·48	0·10	0·12		0·35	0·15	0·17	0·34
	18	0·22	0·63	0·00	0·15		0·22	0·05	0·41	0·32
	Combined	0·26	0·51	0·10	0·13		0·30	0·16	0·18	0·36
6	16	0·46	0·10	0·18	0·26	19	0·38	0·18	0·18	0·26
	17	0·53	0·04	0·15	0·28		0·27	0·15	0·19	0·39
	18	0·44	0·12	0·32	0·12		0·39	0·12	0·12	0·37
	Combined	0·48	0·10	0·16	0·27		0·31	0·12	0·19	0·37
7	16	0·34	0·24	0·16	0·26	20	0·28	0·23	0·18	0·32
	17	0·28	0·17	0·16	0·39		0·20	0·26	0·16	0·38
	18	0·39	0·22	0·15	0·24		0·27	0·19	0·03	0·51
	Combined	0·27	0·22	0·13	0·39		0·27	0·25	0·12	0·36
8	16	0·26	0·35	0·18	0·22	21	0·48	0·06	0·20	0·26
	17	0·30	0·23	0·17	0·30		0·23	0·24	0·22	0·31
	18	0·27	0·37	0·12	0·25		0·27	0·29	0·22	0·22
	Combined	0·25	0·29	0·17	0·29		0·28	0·21	0·19	0·33
9	16	0·18	0·19	0·16	0·47	22	0·06	0·31	0·30	0·32
	17	0·27	0·19	0·18	0·36		0·16	0·17	0·32	0·36
	18	0·21	0·29	0·32	0·17		0·32	0·12	0·22	0·34
	Combined	0·20	0·21	0·21	0·39		0·13	0·19	0·34	0·35
10	16	0·16	0·22	0·25	0·37	23	0·16	0·27	0·31	0·26
	17	0·22	0·21	0·17	0·39		0·15	0·12	0·45	0·27
	18	0·22	0·22	0·22	0·34		0·32	0·12	0·31	0·25
	Combined	0·17	0·22	0·22	0·39		0·16	0·19	0·41	0·24
11	16	0·26	0·16	0·26	0·32	24	0·16	0·26	0·39	0·20
	17	0·29	0·19	0·27	0·25		0·24	0·18	0·31	0·27
	18	0·19	0·27	0·34	0·19		0·29	0·27	0·22	0·22
	Combined	0·19	0·27	0·34	0·19		0·20	0·19	0·34	0·27
12	16	0·36	0·19	0·24	0·22	25	0·06	0·05	0·06	0·83
	17	0·29	0·22	0·19	0·30		0·00	0·14	0·12	0·74
	18	0·22	0·29	0·32	0·17		0·00	0·02	0·15	0·83
	Combined	0·30	0·23	0·23	0·24		0·06	0·13	0·10	0·71
13	16	0·24	0·29	0·21	0·26	26	0·92	0·06	0·00	0·02
	17	0·27	0·24	0·26	0·23		0·89	0·01	0·00	0·09
	18	0·20	0·31	0·25	0·24		0·63	0·00	0·22	0·15
	Combined	0·29	0·26	0·27	0·18		0·83	0·03	0·04	0·10

Table 2 *(continued)*

Location	Spacing class	A	C	G	T	Location	A	C	G	T
27	16	0·20	0·06	0·11	0·64	29	0·60	0·13	0·15	0·12
	17	0·26	0·09	0·16	0·50		0·43	0·26	0·10	0·21
	18	0·34	0·27	0·05	0·34		0·49	0·10	0·22	0·19
	Combined	0·22	0·13	0·14	0·51		0·51	0·27	0·10	0·13
28	16	0·72	0·10	0·10	0·08	30	0·00	0·08	0·00	0·92
	17	0·55	0·14	0·18	0·13		0·04	0·07	0·00	0·89
	18	0·54	0·12	0·08	0·27		0·10	0·03	0·00	0·87
	Combined	0·55	0·10	0·21	0·13		0·03	0·00	0·00	0·97

The −35 and −10 conserved regions are underlined. Position 1 refers to the start of the −35 consensus; position 25 represents the start of the −10 binding site. Base probabilities for non-site positions are A(0·29), C (0·22), G (0·22) and T (0·27).

of examination suffers from repeated assessments of the data, thus increasing the likelihood of observing a significant result by chance, it presents a simple illustrative procedure for testing distribution differences of specific positions. Using the Harley & Reynolds (1987) sequence alignments, these chi-square tests are essentially a replication of O'Neill's (1989*c*) analyses, but using a sample that is approximately five times larger. The contingency analyses identified only one position in the entire −50 to +10 region as having substantially different base compositions for the three spacing classes. The one position that appeared to differ among spacing classes is not one known to be important for protein binding (it is located −11 relative to the −35 promoter region, $\chi^2_6 = 22{\cdot}797$, $p = 0{\cdot}001$). Thus, these results strongly support those of the EM algorithm in yielding little justification for separation of *E. coli* promoter fragments on the basis of spacing class.

To examine the spacing region of the *E. coli* sequences in more detail, and to illustrate the utility of likelihood-based model comparisons in the EM algorithm, we conducted a series of model comparisons to determine if certain spacer positions contribute to the specificity of binding sites, as some analyses have indicated (Deuschle *et al.*, 1986; O'Neill, 1989*a*). For these tests, our intention was to locate which spacer positions, if any, have base frequencies that differ significantly from those of non-site positions. Results from the series of EM model comparisons are presented in Table 3.

The first model listed in Table 3 is identical with the final model of Table 1, in which all consensus and spacer positions are allowed to have different frequencies. The second model provides an omnibus test of the difference between frequencies of spacer residues and those outside the binding regions by constraining all spacer parameters to equal the non-site frequencies and comparing the resulting log-likelihood with that of the full model (model 1). The chi-square value for this test is highly significant, suggesting that at least some of the positions in the spacer region are significantly different from non-site positions. To identify the spacer positions most important in this respect, the equality constraints were successively relaxed on specific positions until the likelihood did not differ significantly from that obtained in the full model. The order in which this procedure was undertaken was dictated by the probabilities shown in Table 2; the positions were examined in decreasing order of base specificity (specifically, the positional order was 16, 23, 10, 7, 9, 19, 20 and 22 in relation to Table 2). This ordering of model comparisons is not guaranteed to find a unique set of important positions, but is expected to yield a minimum number of positions that may be important for promoter specificity. The final model from the series is listed as model 3 in Table 3. The procedure identified eight of the 18 spacer positions

Table 3

Tests of spacer parameters for E. coli *promoters*

Model description	LL	*N* Param	*versus* model	χ^2	df	*p*
(1) All spacer bases unconstrained	−15,859·096	92				
(2) All spacer bases equal to overall base frequencies	−16,022·326	38	1	326·460	54	<0·0001
(3) Model 2, with spacer bases at 7, 9, 10, 16, 19, 20, 22 and 23 unconstrained	−15,880·636	62	1	43·080	30	≈0·06

Position 1 refers to the −35 site; position 25 represents the −10 site. LL, log-likelihood; *N* Param, number of parameters estimated by the model.

as having base compositions that differ from non-site regions. In conjunction with the −35 and −10 binding regions, these results generate the overall consensus:

TTGACA TxTTxxxxxAxxTTx$\frac{G}{T}$Gx TATAAT,

although it is important to emphasize that the spacer positions do not necessarily show a very high degree of conservation, but only a significant difference from the composition of bases outside the promoter regions. The low conservation also implies that mutations in this region may not have large effects. In order to obtain mutations occurring in the spacer it may be necessary to select from an already weak promoter, or do mutagenesis likely to create several simultaneous mutations.

5. Discussion

An extension of the EM algorithm (Lawrence & Reilly, 1990*a*) has been developed that allows for variable spacer lengths in promoter regions of DNA. The variable spacer length model has the ability simultaneously to locate, align and characterize protein-binding sites in biopolymer sequences. The primary feature of the algorithm that is not available in previous methods is the ability to identify binding sites from a group of DNA sequences that share conserved regions but differ in spacing between the binding positions. The method also allows for rigorous comparisons of different models for the binding interaction. An additional practical advantage of the EM method over other search protocols is that the computer memory requirements are slight (Lawrence & Reilly, 1990*a*). Only the observed data [$O(N \times L_{max})$], the vectors of estimated parameters, and the posterior probabilities [$O(NG \times L_{max})$] need to be stored in computer memory.

In applications of the EM model to a large set of *E. coli* promoter fragments, the algorithm appeared to perform very well in identifying promoter sites while allowing variation in spacing between the protein contact regions. Approximately 87% of the known promoter sites were located by the algorithm in the absence of information about the initiation sites, or even of information concerning the weakly conserved pattern that occurs around the initiation site. The predicted frequencies of fragments having different spacing lengths closely resembled those from previous alignments, which did utilize initiation site information (Harley & Reynolds, 1987). Including the information that a weakly conserved pattern exists, with variable spacing, downstream from the −10 region improved the prediction accuracy to 94%. Also, the variability in the probabilities assigned to promoter sites by the EM algorithm, and the result that slightly different alignments are equally good statistically, illustrate the fact that the precise location of some of the promoters are not known with absolute certainty.

Using the maximum likelihood basis of the EM method, we conducted statistical tests of possible differences in promoter base composition that may be associated with different spacing lengths between conserved protein-binding regions. Our findings of consensus similarities between promoters having 16, 17 or 18 bases in the spacer region support the conclusions of Harley & Reynolds (1987) in analysis of this data set but differ somewhat from those reported more recently. O'Neill (1989*a*,*b*,*c*) has examined small subsets of the *E. coli* compilation and found differences between spacing classes for the positional base frequencies. However, these findings were based on very small sample sizes, approximately one-fifth of that presently examined, and were concerned additionally with base positions extending far upstream from the −35 contact region and downstream from the −10 site. Although our findings strongly suggest similarity of protein-binding sites and intervening spacer regions ($p > 0.95$), they also identified at least one position upstream from the −35 region that differs as a function of spacing length. Thus, the extended positions may confer the spacing class specificity noted by O'Neill.

Our assessments of positional importance within the spacing region indicated that as many as eight positions may contribute to promoter specificity. Base frequencies for spacer regions typically have been ignored, because the level of conservation is much less than that of the protein contact sites. However, there is increasing evidence suggesting that the residues in spacer regions play a role in binding sites even if there are not specific contacts (Koudelka *et al.*, 1987). The results of the present application lend further support for such effects.

The present application of EM to *E. coli* promoters has been facilitated somewhat by available information concerning promoter length and its variation, but the EM method is not restricted to these types of well-defined problem. The present algorithm may be very useful for situations in which little is known about protein-binding regions, since the maximum likelihood procedure allows for a large number of different hypotheses to be tested for consistency with the data. Thus, the EM technique seems promising for locating new protein-binding sites and describing their essential features, as well as for adding information about known binding sites.

We recognize that, in addition to sequence composition and spacer length, other features of DNA fragments may be important for promoter specificity. Our findings suggest that inclusion of transcriptional start site information enhances promoter site identification. It is possible that multiple interactions of residues in adjacent and non-adjacent sequence positions may also contribute to promoter uniqueness. Lawrence & Reilly (1990*b*) have presented some preliminary work in this area using an alternative extension of the original EM procedure.

The EM algorithm presented here expands the range of problems and hypotheses in DNA binding specificity that may be investigated using non-

biochemical methodology. Questions concerning protein-binding motifs, sequence similarity, and promoter specificity as a function of base conservation and segment structure may be studied with little expense of time and effort. The method also is applicable to examining protein sequences to determine the common motifs associated with particular functions. As the amount of sequence data increases, methods such as the EM approach may become increasingly useful for description and characterization of important sequence regions.

We thank Calvin Harley for providing us with the promoter sequences in computer-readable form. We thank Charles Lawrence and Andrew Reilly for many helpful discussions and for providing us with the Technical Report describing their extension to the EM algorithm. This project was conducted while L.R.C. was supported by NICHD training grant HD-17053 awarded to David W. Fulker of the Institute for Behavior Genetics, University of Colorado, Boulder, to whom L.R.C. also is indebted for valuable input on model development. G.D.S. was supported by NIH grants GM-28755 and HG-00249.

References

Berg, O. G. & von Hippel, P. H. (1987). Selection of DNA bindings sites by regulatory proteins. *J. Mol. Biol.* **193**, 723–750.

Beyer, W. H. (1988). *Handbook of Tables for Probability and Statistics*, 2nd edit., CRC Press, Boca Raton, FL.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA.

Deuschle, U., Kammerer, W., Reinerg, G. & Hermann, B. (1986). Promoters of *Escherichia coli* – A hierarchy of in vivo strength indicates alternate structures. *EMBO J.* **5**, 2987–2994.

Edwards, A. W. F. (1972). *Likelihood*, Cambridge University Press, London.

Galas, D. J., Waterman, M. S. & Eggert, M. (1985). Rigorous pattern-recognition methods for DNA sequences: analysis of promoter sequences from *E. coli*. *J. Mol. Biol.* **186**, 117–128.

Harley, C. B. & Reynolds, R. P. (1987). Analysis of *E. coli* promoter sequences. *Nucl. Acids Res.* **15**, 2343–2361.

Hawley, D. K. & McClure, W. R. (1983). Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucl. Acids Res.* **11**, 2237–2255.

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

Kendall, M. & Stuart, A. (1977). *The Advanced Theory of Statistics*, vol. 1, Macmillan, New York.

Koudelka, G. B., Harrison, S. C. & Ptashne, M. (1987). Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. *Nature (London)* **326**, 886–888.

Lawrence, C. E. & Reilly, A. A. (1990*a*). An Expectation Maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Struct. Funct. Genet.* **7**, 41–51.

Lawrence, C. E. & Reilly, A. A. (1990*b*). Misaligned (shuffled) data analysis with application to gene regulation. Technical Report 3, University of Albany School of Public Health Technical Report Series.

Little R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley and Son, New York.

Mengeritsky, G. & Smith, T. F. (1987). Recognition of characteristic patterns in sets of functionally equivalent DNA sequences. *Comput. Appl. Biosci.* **3**, 223–227.

Mulligan, M. E. & McClure, W. R. (1986). Analysis of the occurrence of promoter-sites in DNA. *Nucl. Acids Res.* **14**, 109–126.

O'Neill, M. C. (1989*a*). *Escherichia coli* promoters: I. Consensus as it relates to spacing class, specificity, repeat substructure, and three-dimensional organization. *J. Biol. Chem.* **264**, 5522–5530.

O'Neill, M. C. (1989*b*). *Escherichia coli* promoters: II. A spacing class-dependent promoter search protocol. *J. Biol. Chem.* **264**, 5531–5534.

O'Neill, M. C. (1989*c*). Consensus methods for finding and ranking DNA binding sites: application to *Escherichia coli* promoters. *J. Mol. Biol.* **207**, 301–310.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 2444–2448.

Pribnow, D. (1975). Bacteriophage T7 early promoters: nucleotide sequences of two RNA polymerase binding sites. *J. Mol. Biol.* **99**, 419–443.

Rosenberg, M. D. & Court, D. L. (1979). Regulatory sequences involved in the promotion and termination of RNA transcription. *Annu. Rev. Genet.* **13**, 319–353.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucl. Acids Res.* **12**, 505–519.

Stormo, G. D. (1988). Computer methods for analyzing sequence recognition of nucleic acids. *Annu. Rev. Biophys. Biophys. Chem.* **17**, 241–263.

Stormo, G. D. (1990*a*). Consensus patterns in DNA. *Methods Enzymol.* **183**, 211–221.

Stormo, G. D. (1990*b*). Identifying regulatory sites from DNA sequence data. In *Biological Structure, Dynamics, Interactions & Expression, Proc. 6th Conv. Biomol. Stereodynam.* (Sarma, R. & Sarma, M., eds), pp. 103–111, Adenine Press, New York.

Stormo, G. D. & Hartzell, G. W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 1183–1187.

Edited by F. E. Cohen