CABIOS

Vol 6 no 2 1990
Pages 81 – 92

Identification of consensus patterns in unaligned DNA sequences known to be functionally related

Gerald Z.Hertz, George W.Hartzell,III and Gary D.Stormo*

Abstract

We have developed a method for identifying consensus patterns in a set of unaligned DNA sequences known to bind a common protein or to have some other common biochemical function. The method is based on a matrix representation of binding site patterns. Each row of the matrix represents one of the four possible bases, each column represents one of the positions of the binding site and each element is determined by the frequency the indicated base occurs at the indicated position. The goal of the method is to find the most significant matrix – i.e. the one with the lowest probability of occurring by chance – out of all the matrices that can be formed from the set of related sequences. The reliability of the method improves with the number of sequences, while the time required increases only linearly with the number of sequences. To test this method, we analysed 11 DNA sequences containing promoters regulated by the Escherichia coli LexA protein. The matrices we'found were consistent with the known consensus sequence, and could distinguish the generally accepted LexA binding sites from other DNA sequences.

Introduction

The regulation of genetic processes such as transcription and replication is frequently determined by the binding of proteins to specific regions of the DNA. Understanding these interactions requires a knowledge of both the proteins and the regions of the DNA to which they bind. In this paper, we describe an algorithm for identifying the pattern that describes the DNA sequences that are bound by a specific protein. The algorithm assumes that the binding site is a contiguous region of DNA whose essential features reside within the individual bases of the sequence.

Given a set of N sequences known to contain functionally related subsequences, such as binding sites for a common protein, our goal is to identify the pattern that describes their common sequence element. If the set of N sequences share a common L-mer (i.e. a subsequence of length L) of sufficient length to be unexpected, the answer would be that shared sequence. However, unlike the recognition sites of restriction endonucleases, the binding sites of regulatory proteins are generally not accurately described by a single consensus sequence. Some positions will be more highly conserved than others and the preference for each of the four bases can be different. A more precise representation of a binding pattern is with a $4 \times L$ matrix (Stormo *et al.*, 1982, 1986; Harr *et al.*, 1983; Staden, 1984; Stormo, 1988, 1989). The four rows correspond to the four bases (A, C, G, T), and the L columns correspond to the positions of the nucleotides within the pattern; the elements of the matrix are determined by the frequency that the indicated base occurs at the indicated position.

Working with matrices whose elements contain the number of times that the indicated base occurs at the indicated position (see Figure 1 for examples), we have developed a procedure for finding consensus patterns in unaligned DNA sequences (Stormo and Hartzell, 1989). The method searches for a matrix with a low probability of occurring by chance or, equivalently, having a high information content (Schneider *et al.*, 1986). In this paper, we describe the implementation of this procedure in more detail and demonstrate its use and robustness by analysing the LexA binding sites of *E.coli*.

System and methods

The programs were developed on a Pyramid 90x minicomputer running OSx 4.0, Pyramid's version of the UNIX operating system. The computer has 8 megabytes of memory and 900 megabytes of disk storage. All programs are written in C and may be obtained from the authors by sending magnetic tape or disks.

Algorithm

Searching for a matrix pattern

The goal of our basic algorithm is to identify a matrix that describes the common sequence pattern shared by N sequences known to contain functionally related subsequences. Before this algorithm can be used, one must first decide on an approximate width – i.e. the values of L – for the pattern. This width is determined from a knowledge of the biochemistry of the problem. For example, widths of 20 bp are typical for the binding sites of prokaryotic proteins (Schneider *et al.*, 1986).

Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309-0347, USA

**To whom reprint requests should be sent*

It is not necessary to know the precise width of the binding site; however, it can be useful to try a range of values.

To describe the algorithm, we will first describe an ideal, but impractical, strategy for finding the pattern shared by a set of functionally related sequences. For some value of L, form every possible list that contains exactly one L-mer (i.e. a subsequence of length L) from each of the N sequences. For each of these lists, determine the corresponding $4 \times L$ matrix that simply contains the number of times that each base occurs at each of the L positions. The matrix with the lowest probability of occurring by chance would describe the most novel pattern, the pattern we are presumably searching for. However, N sequences of length M will form $(M - L + 1)^N$ different matrices. If M equals 100, L equals 20 and N equals 10, the total number of matrices will be $\sim 10^{19}$· an impractical number. Our algorithm avoids this problem by saving only the most interesting matrices as it scans through the sequences.

The algorithm starts by forming a matrix for each of the L-mers in the first sequence (Figure 1B). Each of these matrices is then combined with each L-mer in the second sequence to form new matrices containing two L-mers (Figure 1C). However, for each L-mer from the first sequence, the program only saves the progeny matrix with the lowest probability of occurring by chance, i.e. the 'best' progeny matrix. In the next cycle, each saved matrix is combined with each L-mer in the third sequence to form new matrices each containing three L-mers (Figure 1D). Again, the program only saves the best progeny of each matrix from the previous cycle. This cycle is repeated until the last sequence in the set has contributed an L-mer to the saved matrices. Of the matrices saved after the last cycle, the one with the lowest probability of occurring by chance is considered to describe the consensus pattern (Figure 1E).

Since each matrix formed during any one cycle summarizes the same number of L-mers, we can use the information contents of the matrices (Schneider *et al.*, 1986) to determine their relative probabilities of occurring by chance. The formula to determine the information content of (I) a matrix is:

$$\mathrm{I} = \sum_{i=1}^{L} \sum_{b=A}^{T} \frac{N_{bi}}{N} \log_2 \frac{N_{bi}/N}{P_b} \qquad (1)$$

where b refers to the rows of the matrix (i.e. the bases A, C, G, T in alphabetical order), i refers to the columns (i.e. the positions of the nucleotides within the binding-site pattern), P_b is the genomic frequency of base b. L is the total number of columns in the matrix (i.e. the number of bases in each L-mer), N is the total number of sequences summarizes by the matrix, and N_{bi} refers to the matrix element in row b and column i (i.e. the number of L-mers containing base b at position i). For a given value of N, the higher the information content of a matrix, the lower the probability of it having occurred by chance. Since $\Sigma_{b=A}^{T} N_{bi} = N$, equation (1) can also be written as:

$$\mathrm{I} = \frac{1}{N}\left[\sum_{i=1}^{L} \sum_{b=A}^{T} N_{bi} \log_2 \frac{N_{bi}}{P_b}\right] - L \log_2 N \qquad (2)$$

Since the value of N is the same for all the matrices being compared during any one cycle, the portion of equation (2) in brackets is sufficient for determining the relative information content of each matrix.

Using the algorithm just described, the number of matrices saved after the last cycle is on the order of the number of L-mers in the first sequence. In practice, ties occur during the early cycles so that the number of matrices at the end is greater than the number of L-mers in the first sequence. In the examples presented in this paper, the final number of matrices was 2- to 3-fold greater than the number of L-mers in the first sequence.

Searching for matches to a matrix pattern

The matrix representation for consensus patterns also allows one to rate each L-mer of a DNA sequence according to how well it matches the pattern. The following is the motivation for how we do this rating. The probability of a particular matrix occurring by chance is:

$$\text{Probability} = \prod_{i=1}^{L} \left[\frac{N!}{\prod_{b=A}^{T} N_{bi}!} \prod_{b=A}^{T} P_b^{N_{bi}}\right] \qquad (3)$$

The brackets contain the probability of obtaining column i of the matrix: the product of the P_b terms represent the probability of obtaining a particular order of the bases, and the product of the factorial terms represent the number of ways N bases can be ordered when each base is present N_{bi} times.

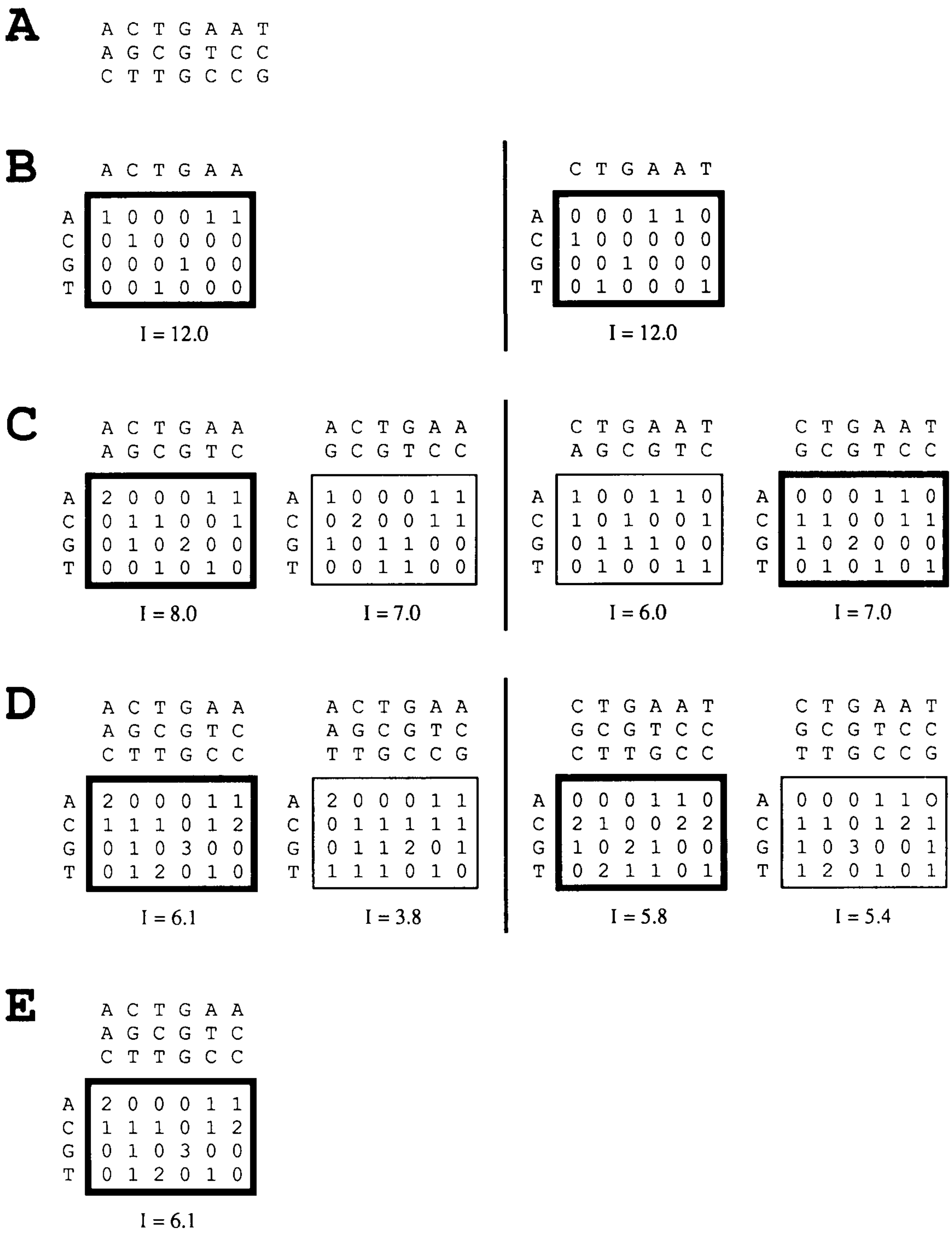
If another L-mer is added to the matrix, the probability will change by a factor of:

$$\text{Factor} = \prod_{i=1}^{L} \left[\frac{N + 1}{N_{bi} + 1} P_b\right] \qquad (4)$$

where b is the base at position i of the additional L-mer. Equation (4) can be changed into a more convenient additive measure by taking the negative of its logarithm:

$$-\log_2 (\text{Factor}) = \sum_{i=1}^{L} \left[\log_2 \frac{N_{bi} + 1}{(N + 1)P_b}\right] \qquad (5)$$

The higher an L-mer rates according to equation (5), the greater it would decrease the probability of the matrix and the more closely it is related to the pattern described by the original matrix.



A

A C T G A A T
A G C G T C C
C T T G C C G

B

A C T G A A

$$\begin{array}{c|cccccc} A & 1 & 0 & 0 & 0 & 1 & 1 \\ C & 0 & 1 & 0 & 0 & 0 & 0 \\ G & 0 & 0 & 0 & 1 & 0 & 0 \\ T & 0 & 0 & 1 & 0 & 0 & 0 \end{array}$$

I = 12.0

C T G A A T

$$\begin{array}{c|cccccc} A & 0 & 0 & 0 & 1 & 1 & 0 \\ C & 1 & 0 & 0 & 0 & 0 & 0 \\ G & 0 & 0 & 1 & 0 & 0 & 0 \\ T & 0 & 1 & 0 & 0 & 0 & 1 \end{array}$$

I = 12.0

C

A C T G A A
A G C G T C

$$\begin{array}{c|cccccc} A & 2 & 0 & 0 & 0 & 1 & 1 \\ C & 0 & 1 & 1 & 0 & 0 & 1 \\ G & 0 & 1 & 0 & 2 & 0 & 0 \\ T & 0 & 0 & 1 & 0 & 1 & 0 \end{array}$$

I = 8.0

A C T G A A
G C G T C C

$$\begin{array}{c|cccccc} A & 1 & 0 & 0 & 0 & 1 & 1 \\ C & 0 & 2 & 0 & 0 & 1 & 1 \\ G & 1 & 0 & 1 & 1 & 0 & 0 \\ T & 0 & 0 & 1 & 1 & 0 & 0 \end{array}$$

I = 7.0

C T G A A T
A G C G T C

$$\begin{array}{c|cccccc} A & 1 & 0 & 0 & 1 & 1 & 0 \\ C & 1 & 0 & 1 & 0 & 0 & 1 \\ G & 0 & 1 & 1 & 1 & 0 & 0 \\ T & 0 & 1 & 0 & 0 & 1 & 1 \end{array}$$

I = 6.0

C T G A A T
G C G T C C

$$\begin{array}{c|cccccc} A & 0 & 0 & 0 & 1 & 1 & 0 \\ C & 1 & 1 & 0 & 0 & 1 & 1 \\ G & 1 & 0 & 2 & 0 & 0 & 0 \\ T & 0 & 1 & 0 & 1 & 0 & 1 \end{array}$$

I = 7.0

D

A C T G A A
A G C G T C
C T T G C C

$$\begin{array}{c|cccccc} A & 2 & 0 & 0 & 0 & 1 & 1 \\ C & 1 & 1 & 1 & 0 & 1 & 2 \\ G & 0 & 1 & 0 & 3 & 0 & 0 \\ T & 0 & 1 & 2 & 0 & 1 & 0 \end{array}$$

I = 6.1

A C T G A A
A G C G T C
T T G C C G

$$\begin{array}{c|cccccc} A & 2 & 0 & 0 & 0 & 1 & 1 \\ C & 0 & 1 & 1 & 1 & 1 & 1 \\ G & 0 & 1 & 1 & 2 & 0 & 1 \\ T & 1 & 1 & 1 & 0 & 1 & 0 \end{array}$$

I = 3.8

C T G A A T
G C G T C C
C T T G C C

$$\begin{array}{c|cccccc} A & 0 & 0 & 0 & 1 & 1 & 0 \\ C & 2 & 1 & 0 & 0 & 2 & 2 \\ G & 1 & 0 & 2 & 1 & 0 & 0 \\ T & 0 & 2 & 1 & 1 & 0 & 1 \end{array}$$

I = 5.8

C T G A A T
G C G T C C
T T G C C G

$$\begin{array}{c|cccccc} A & 0 & 0 & 0 & 1 & 1 & 0 \\ C & 1 & 1 & 0 & 1 & 2 & 1 \\ G & 1 & 0 & 3 & 0 & 0 & 1 \\ T & 1 & 2 & 0 & 1 & 0 & 1 \end{array}$$

I = 5.4

E

A C T G A A
A G C G T C
C T T G C C

$$\begin{array}{c|cccccc} A & 2 & 0 & 0 & 0 & 1 & 1 \\ C & 1 & 1 & 1 & 0 & 1 & 2 \\ G & 0 & 1 & 0 & 3 & 0 & 0 \\ T & 0 & 1 & 2 & 0 & 1 & 0 \end{array}$$

I = 6.1

Fig. 1. A diagram of the algorithm for finding matrix patterns. In this example, the width of the matrix being sought has been set to six positions. (**A**) A list of the three sequences being analysed. Since each sequence contains seven bases, each sequence will contain two 6-mers. (**B**) The two matrices formed during the first cycle of the algorithm. These matrices correspond to the first and second 6-mers of the first sequence, respectively. 'I' is the information content of the corresponding matrix. The higher the value of I, the more desirable the matrix (**C**) The four matrices formed during the second cycle of the algorithm. Each matrix from **B** was combined with each 6-mer in the second sequence to form all possible progeny matrices. The two pairs of matrices are the progeny of the first and second 6-mers of the first sequence, respectively. For each parental matrix in B, only the progeny matrix with the highest value of I is saved for the next cycle of the program—the matrices in heavy boxes are the progeny to be saved. (**D**) The four matrices formed during the third cycle of the program. Each saved matrix from C was combined with each 6-mer in the third sequence to form all possible progeny matrices. All other details are the same as in **C**. (**E**) The best matrix found by the algorithm.

The bracketed portion of equation (5) can be used to transform the original matrix into a 4 × L specificity matrix (e.g. see Figure 5B) (Stormo, 1988, 1989):

$$(\text{element})_{bi} \text{ of specificity matrix } = \log_2 \frac{N_{bi} + 1}{(N + 1)P_b} \qquad (6)$$

Each nucleotide of an L-mer has a corresponding matrix element whose row matches the nucleotide's base and whose column matches the nucleotide's position. Rating an L-mer with equation (5) is identical to summing the matrix elements corresponding to the nucleotides of the L-mer. When we rate an L-mer against a matrix pattern, we always sum the elements of the corresponding specificity matrix rather than explicitly using equation (5).

Implementation

The basic algorithm described in the previous section has been implemented in a program called CONSENSUS. The required input for this program is a list of files, each containing one of the sequences of interest; the sequences are processed by the program in the order presented in this list. In addition, three other parameters can be changed. The genomic frequency of each base can be changed from a default value of 0.25, the size of the L-mers can be changed from a default value of 10 and the list of L-mers for each sequence can include the L-mers of the complementary sequence. This last option is useful in case the orientation of the binding sites is unknown.

The CONSENSUS program has previously been demonstrated to accurately identify the known consensus pattern for the *E.coli* CRP protein (Stormo and Hartzell, 1989). To further demonstrate the robustness of our algorithm, we now present the results of testing the program on the bacterial promoters that are regulated by the *E.coli* LexA protein. LexA represses the transcription of the genes involved in the SOS response, an inducible system for repair of bacterial DNA damage (Walker, 1984, 1985; Peterson *et al.*, 1988). LexA represses transcription of the SOS genes—including itself—by binding to DNA near the RNA start site. The 20-bp symmetrical sequence TACTGTATATATATACAGTA is its consensus binding site (Walker, 1984).

The CONSENSUS program finds a matrix consistent with the known LexA consensus sequence

From the GenBank database, we collected the DNA sequences of 16 bacterial promoters thought to be regulated by LexA (Table I). In all these sequences, the locations of the putative LexA binding sites had been determined either by footprinting or by homology with the consensus sequence. For our analysis with the CONSENSUS program, we decided to search for patterns within the 200 bases flanking the start sites of the RNAs (i.e. the 100 bases preceding and the 99 bases following the start sites). This is a region almost twice the size of the one used in our previous test of the program (Stormo and Hartzell, 1989). We were not able to find the flanking 200 bases for three of the genes, and the LexA regulation of two of the promoters is considered questionable; therefore, only the first 11 genes listed in Table I qualified for our analysis.

Table I. The LexA-regulated promoters analysed in this paper

Gene	Number of LexA binding sites	Means of identification[a]	References
cloacin DF13	1	homology	van den Elzen *et al.*, 1983
colicin E1	2	protection	Yamada *et al.*, 1982 Ebina *et al.*, 1983
colicin Ia	2	homology[b]	Mankovich *et al.*, 1986
colicin Ib	2	homology[b]	Varley and Boulnois, 1984
recA	1	protection	Horii *et al.*, 1980 Sancar *et al.*, 1980
recN	2	protection	Rostas *et al.*, 1987
sulA	1	homology	Beck and Bremer, 1980 Cole, 1983
umuDC	2	protection[c]	Kitagawa *et al.*, 1985
uvrA	1	protection	Sancar *et al.*, 1982a
uvrB	1	protection	Sancar *et al.*, 1982b
uvrD	1	protection	Easton and Kushner, 1983 Finch and Emmerson, 1983
colicin A[d]	2	homology	Morlon *et al.*, 1983
lexA[d]	2	protection	Brent and Ptashne, 1981 Little *et al.*, 1981 Markham *et al.*, 1981
mucAB[d]	1	homology	Perry *et al.*, 1985
himA[e]	1	homology[g]	Miller *et al.*, 1981 Miller, 1984 Mechulam *et al.*, 1985
uvrC[f]	1	homology[g]	van Sluis *et al.*, 1983 Forster and Strike, 1985 Sharma *et al.*, 1986

[a] Indicates whether the LexA binding site was identified by DNase protection or simply by homology with a consensus sequence.
[b] Downstream LexA binding site is questionable.
[c] Upstream LexA binding site is not protected from DNase and is identified only by homology.
[d] Less than 100 bases before the transcriptional initiation site were obtainable
[e] Questionable whether the promoter has been correctly located.
[f] Questionable whether the promoter is regulated by LexA.
[g] LexA binding site is questionable; other equally questionable sites have been proposed, but they are > 100 bp upstream of the transcriptional initiation site.

In all the examples presented in this paper, the L-mers of the complementary strand were included in the analysis and the genomic frequencies of the bases were left at 0.25 since that is the approximate value for *E.coli*. Each run of the program took ~8 min of cpu time on a Pyramid 90x or ~0.5 min on a MIPS M-2000. In our initial experiment, the width of the matrices was set to 20 bases. Since each sequence was

A

I = 27.241

A	1	9	0	0	0	0	9	0	10	2	6	4	11	4	6	0	11	0	3	4
C	0	0	11	0	0	0	0	1	0	0	1	3	0	0	4	11	0	0	0	0
G	0	2	0	0	11	2	2	0	1	0	3	0	0	0	0	0	0	9	1	1
T	10	0	0	11	0	9	0	10	0	9	1	4	0	7	1	0	0	2	7	6
	T	A	C	T	G	T	A	T	A	T	A	n	A	T	A	C	A	G	T	t

B

I = 27.239

A	6	7	1	9	0	0	0	0	11	0	10	1	6	4	11	4	5	0	11	0
C	0	1	0	0	11	0	0	0	0	1	0	1	1	3	0	2	5	11	0	0
G	0	2	0	1	0	0	11	1	0	0	1	0	2	0	0	0	0	0	0	9
T	5	1	10	1	0	11	0	10	0	10	0	9	2	4	0	5	1	0	0	2
	a/t	a	T	A	C	T	G	T	A	T	A	T	A	n	A	T/a	A/c	C	A	G

C

I = 27.162

A	1	9	0	0	0	0	11	0	10	2	6	4	11	4	5	0	11	0	2	4
C	0	0	11	0	0	0	0	1	0	0	1	3	0	2	5	11	0	0	1	0
G	0	1	0	0	11	1	0	0	1	0	3	0	0	0	0	0	0	9	1	1
T	10	1	0	11	0	10	0	10	0	9	1	4	0	5	1	0	0	2	7	6
	T	A	C	T	G	T	A	T	A	T	A	n	A	T/a	A/c	C	A	G	T	t

D

I = 27.091

A	7	7	1	10	0	0	0	0	11	0	10	1	6	3	11	4	5	0	11	0
C	0	1	0	0	11	0	0	0	0	2	0	1	1	3	0	2	5	11	0	1
G	0	2	0	0	0	0	11	1	0	0	1	0	2	0	0	0	0	0	0	8
T	4	1	10	1	0	11	0	10	0	9	0	9	2	5	0	5	1	0	0	2
	a	a	T	A	C	T	G	T	A	T	A	T	A	T	A	T/a	A/c	C	A	G

Fig. 2. The top 20-position matrices found by the CONSENSUS program after analysing the first 11 LexA-regulated genes listed in Table I. The input sequences were analysed in five different orders. Each matrix corresponds to a different ordering; matrix B was identified twice. 'I' is the information content of a matrix. The sequence below each matrix is a consensus sequence derived from the matrix; the bases in capitals match the TACTGTATATATATACAGTA consensus binding site for the LexA protein.

considered to include its complement, each sequence contained 362 20-mers; thus, $362^{11} \approx 10^{28}$ different matrices could possibly be formed.

Since the program is sensitive to the order of the input sequences, we ran the program with five randomized orders of the 11 input sequences. The top matrices for all five runs were consistent with the generally recognized LexA binding site (Figure 2); however, they were not all identical. This variability is presumably because each binding site is somewhat symmetrical and because some fragments contain more than one binding site. As a result, each fragment contains at least two subsequences that are homologous with each strand of every other LexA binding site.

The LexA binding site is generally considered to be a 20-bp symmetrical site centered on a highly conserved 16-bp core. However, two of the top matrices were not centered on this core (Figure 2B and 2D). These two matrices contained the positions corresponding to the core, but the four additional positions all fell to one side of the 16 bases. These results indicate that the width of the binding site is between 16 and 24 bp. A search for a 22-base matrix identified the site expected from the previous results, a site corresponding to the 20-bp symmetrical sequence plus two additional base pairs on one side of the symmetrical sequence (Figure 3A). This 22-base matrix had a significantly higher information content (Appendix to Schneider *et al.*, 1986) than the 20-base matrix. However, the best 24-base matrices identified by the program were still not centered on the 16-bp core (Figure 3B) and did not contain significantly more information than the best 22-base matrix.

These results suggest that the LexA binding site may be 22 bases wide. However, the program has a choice of which way to orient the largely symmetrical binding sites; therefore, the

A

I = 28.426

A	6	7	1	9	0	0	0	0	11	0	10	1	6	4	11	4	5	0	11	0	2	4
C	0	1	0	0	11	0	0	0	0	1	0	1	1	3	0	2	5	11	0	0	1	0
G	0	2	0	1	0	0	11	1	0	0	1	0	2	0	0	0	0	0	0	9	1	1
T	5	1	10	1	0	11	0	10	0	10	0	9	2	4	0	5	1	0	0	2	7	6
	a/t	a	T	A	C	T	G	T	A	T	A	T	A	n	A	T/a	A/c	C	A	G	T	t

B

I = 28.926

A	3	6	5	7	4	9	0	0	0	1	7	0	6	2	6	0	11	1	7	0	11	0	3	7
C	0	0	0	0	0	1	11	0	0	0	0	2	0	0	1	2	0	1	4	11	0	1	0	0
G	3	0	1	4	0	1	0	0	11	2	1	0	2	0	3	0	0	0	0	0	0	8	0	1
T	5	5	5	0	7	0	0	11	0	8	3	9	3	9	1	9	0	9	0	0	0	2	8	3
	t	a/t	a/t	a	T	A	C	T	G	T	A	T	A	T	A	T	A	T	A	C	A	G	T	A

I = 28.926

A	6	5	7	4	9	0	0	0	1	7	0	6	2	6	0	11	1	7	0	11	0	3	7	3
C	0	0	0	0	1	11	0	0	0	0	2	0	0	1	2	0	1	4	11	0	1	0	0	0
G	0	1	4	0	1	0	0	11	2	1	0	2	0	3	0	0	0	0	0	0	8	0	1	3
T	5	5	0	7	0	0	11	0	8	3	9	3	9	1	9	0	9	0	0	0	2	8	3	5
	a/t	a/t	a	T	A	C	T	G	T	A	T	A	T	A	T	A	T	A	C	A	G	T	A	t

Fig. 3. The top 22- and 24-position matrices found by the CONSENSUS program after analysing the first 11 LexA-regulated genes listed in Table I. The matrices are presented as in Figure 2. (**A**) The top 22-position matrix found after analysing the input sequences in five different orders, this matrix was identified twice. (**B**) The top 24-position matrices found after analysing the input sequences in five different orders.

program might end up overstating or understating the importance of some positions in an effort to maximize the information content of the resulting matrix. This phenomenon can also account for why the 20-base consensus sequence derived from the matrices are not completely symmetrical (Figures 2 and 3A).

The matrix generated by the CONSENSUS program can be used to find LexA binding sites

We next decided to score each 22-mer of each sequence according to how well it matched the 22-base matrix identified by the CONSENSUS program [see equation (5) in the Algorithm section]. As negative controls, we first scored the primate virus SV40 and the bacterial plasmid pBR322. The highest score was 20.9 and <0.06% of the 22-mers scored above 15. Given this background, we decided to examine the 22-mers from the LexA-regulated DNAs only if they scored above 20. This strategy correctly identified all 18 LexA binding sites expected (Table II), i.e. there were no false negatives. All the recognized sites scored above 24 in at least one of their strands and each promoter region had at least one site that scored above 26. If one excludes the sites that are adjacent to another site and, thus, have the possibility of cooperative binding (Brent and Ptashne, 1981; Ebina *et al.*, 1983), the remaining sites all scored above 28 in at least one strand.

Several questionable sites were not identified, namely, the second sites in the colicin Ia (score of 11.8) and Ib (score of 12.8) promoter regions, the first site in the *umuDC* operator (score of 11.1), and the sites in the himA (score of −0.1) and uvrC (score of 3.8) genes; thus, justifying the ambiguity in the literature regarding these sites (Table I) (van Sluis *et al.*, 1983; Miller, 1984; Varley and Boulnois, 1984; Forster and Strike, 1985; Kitagawa *et al.*, 1985; Mechulam *et al.*, 1985; Mankovich *et al.*, 1986; Sharma *et al.*, 1986). The sequences for *himA* and *uvrC* had not been included in the original training set because the presence of a promoter (Miller, 1984) and the regulation by LexA (Forster and Strike, 1985), respectively, are in question. However, the proposed sites in these two genes still had very low scores even when rated by a matrix in whose training set they had been included (data not shown).

Finally, two of the sites scoring above 20 were not recognized LexA binding sites. The 'false' positive site in the *lexA* promoter region had a score of 20.4, which is below the top score of the negative controls (20.9 in SV40) and probably should not be considered a binding site. However, the other 'false' positive site, which is 18 bases downstream of the second LexA binding site in the *recN* promoter region, had a score of 23.2. Although this score is less than those obtained by the recognized LexA binding sites, given this site's relatively high score and its placement overlapping another LexA binding site, i.e. cooperative

Table II. Sites scoring above 20 according to the 22-base matrix

Gene	Location of RNA start site[a]	Location of potential binding site	Score of potential binding site[b]	Notes
cloacin DF13	101, 103	95	29.1	
		97	34.5	c
colicin E1	101	95	31.3	
		112	28.1	c
colicin Ia	101	97	31.6	d
colicin Ib	101	97	30.1	d
recA	101	69	26.4	
		71	28.9	c
recN	101	69	33.5	d
		91	26.6	
		71	23.3	c
		111	23.2	c,e
sulA	101	83	29.3	
		85	21.3	c
umuDC	101	89	33.2	
		91	28.2	c
uvrA	101	58	30.8	
uvrB	101	69	23.1	
		71	30.7	c
uvrD	101	100	28 2	
		102	23.7	c
colicin A	37	32	29.0	d
		48	25 2	c
lexA	74	53	24.0	
		74	26.4	
		14	20.4	c,e
		55	21.4	c
mucAB	59	47	30.6	d
		49	26.2	c
himA	101	none		d
uvrC	101	none		

[a] The bases are numbered in the direction of transcription.

[b] The score was determined as described in the Algorithm section.

[c] Score was obtained from the complementary strand, therefore, the location will be two bases downstream of the corresponding site on the 5′ to 3′ strand because of the asymmetry of the matrix.

[d] The RNA initiation site was approximated by the location of the −10 consensus sequence.

[e] Not a recognized LexA-binding site.

binding is possible, it may have biological or evolutionary significance.

The matrix obtained from aligning the LexA binding sites identifies a 22-base consensus pattern

To further understand how the information in the LexA binding sites is distributed, we aligned the binding sites and determined the information content for each position over a region of 90 bases around the point of alignment (Figure 4B). For each LexA binding site, we chose the DNA strand that rated highest when scored by the 22-base matrix identified by the CONSENSUS program (Table II). Not surprisingly, the consensus pattern found by alignment of the binding sites was essentially the same 22-base matrix found previously (Figure 4A versus positions 31–52 of Figure 4B).

In an attempt to further understand the consensus pattern, we did the alignment after limiting ourselves to only one copy of genes that appeared to be very closely related. This meant that we excluded the binding sites for all the colicin genes except for colicin E1, and we excluded the *mucAB* operon since it is a plasmid analog of the *umuDC* operon (Perry *et al.*, 1985). The consensus pattern identified by this alignment (positions 31–52 of Figure 4C) was essentially the same as when all the binding sites were included. The information content of most positions increased slightly, but that was expected because background levels of information increase when the number of sequences is decreased (Appendix to Schneider *et al.*, 1986).

Is the LexA binding site best described by a symmetrical or an asymmetrical matrix?

Although the consensus sequence for the LexA binding site is symmetrical, the sequence derived from the 22-base matrix is not completely symmetrical. As mentioned previously, if we choose between both orientations of each binding site to maximize the information content of the whole matrix, the information content of partially conserved positions may be overstated or understated. In addition, when functionally symmetrical binding sites are aligned, the resulting matrix would not be expected to be perfectly symmetrical due to the finite number of binding sites. Since half of the consensus site will almost always appear to have more information than the other half, the asymmetry we observed may have no biochemical significance.

To help determine the significance of the information asymmetry, we analysed the LexA binding sites under the assumption that the matrix should be symmetrical. This was done by using both DNA strands when aligning binding sites. Under these conditions, only the 20-base consensus sequence appeared to be important (positions 33–52 of Figure 4D). We also used the 20-position matrix resulting from this alignment (Figure 5A) to search for LexA binding sites (Table III) as described previously for the 22-position matrix of Figure 3A (Table II). The highest rated 20-mer in either SV40 or pBR322 had a score of 15.0, and <0.04% of their 20-mers scored above 10; therefore, we decided to examine the 20-mers that scored above 15.

Since the symmetrical matrix had a lower information content than the matrix previously used for scoring (22.1 for the symmetrical matrix in Figure 5A versus 28.4 for the matrix in Figure 3A), the scores shown in Table III tended to be lower than those in Table II. The symmetrical matrix detected all of the generally accepted LexA binding sites. In addition, it detected the additional site in the *recN* promoter region, although this site was the lowest rated site above the cutoff. Once again,

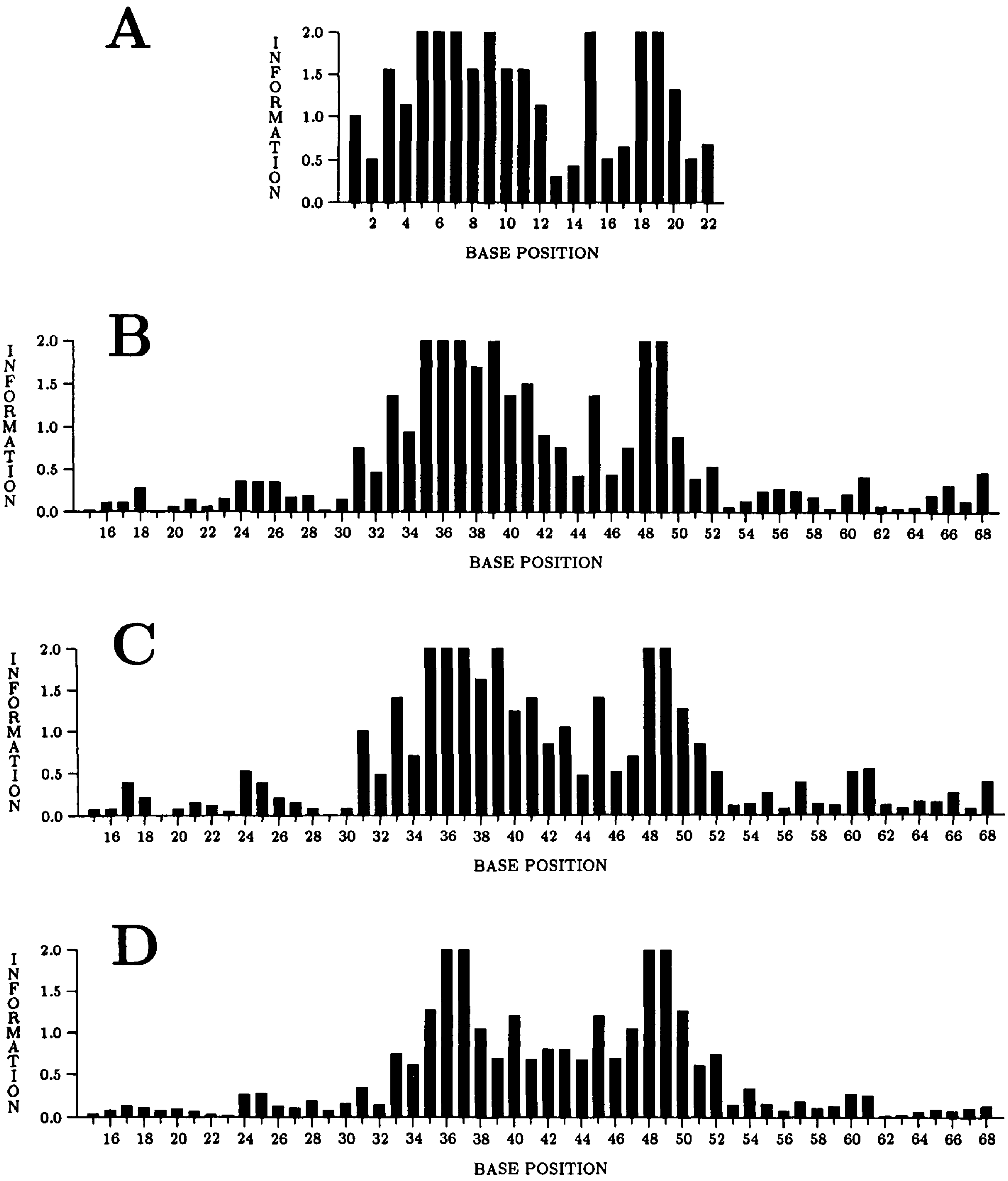


Fig. 4. Bar graph of the information content at each position of the LexA binding site The informations content was calculated from equation (1) in the Algorithm section. (**A**) The information content of each position of the matrix in Figure 3A. (**B**) The information content of each position of the matrix obtained by aligning the 19 LexA binding sites identified in Table II (includes the 'extra' *recN* site, but excludes the 'extra' *lexA* site). (**C**) The information content of each position of the matrix obtained by aligning the 14 LexA binding sites remaining after limiting the analysis to only one representative of groups of genes that are closely related. (**D**) The information content of each position of the matrix obtained by aligning both strands of each of the 19 binding sites.

A

I = 22.125

A	12	26	5	0	0	1	24	3	24	5	28	7	32	8	28	0	38	0	3	23
C	3	5	32	0	0	0	0	3	0	4	1	7	0	6	9	38	0	1	4	0
G	0	4	1	0	38	9	6	0	7	1	4	0	3	0	0	0	0	32	5	3
T	23	3	0	38	0	28	8	32	7	28	5	24	3	24	1	0	0	5	26	12
	T	A	C	T	G	T	A	T	A	T	A	T	A	T	A	C	A	G	T	A

B

A	0.42	1.47	-0.70	-3.28	-3.28	-2.28	1.36	-1.28	1.36	-0.70	1.57	-0.28	1.76	-0.11	1.57	-3.28	2.00	-3.28	-1.28	1.30
C	-1.28	-0.70	1.76	-3.28	-3.28	-3.28	-3.28	-1.28	-3.28	-0.96	-2.28	-0.28	-3.28	-0.47	0.04	2.00	-3.28	-2.28	-0.96	-3.28
G	-3.28	-0.96	-2.28	-3.28	2.00	0.04	-0.47	-3.28	-0.28	-2.28	-0.96	-3.28	-1.28	-3.28	-3.28	-3.28	-3.28	1.76	-0.70	-1.28
T	1.30	-1.28	-3.28	2.00	-3.28	1.57	-0.11	1.76	-0.28	1.57	-0.70	1.36	-1.28	1.36	-2.28	-3.28	-3.28	-0.70	1.47	0.42

Fig. 5. (**A**) The 20-position symmetrical matrix corresponding to the peak of information in Figure 4D, i.e. positions 33–52 The matrix is presented as in Figure 2. (**B**) The specificity matrix used for scoring 20-mers in Table III. This matrix was formed by transforming the elements of the matrix in A with equation (6) as described in the Algorithm section.

the lowest rated sites—those with scores at or below 21—were adjacent to other binding sites. These results are consistent with the symmetrical matrix being an accurate representation of the LexA binding site. We also determined how much each half of a binding site contributed to the site's total score (column 5 versus column 4 of Table III). These results indicate that a high scoring half binding site can compensate for a low scoring half (e.g. *uvrA* in Table III).

The analysis of half binding sites was also used to analyse the proposed, low-scoring binding sites in the *himA* and *uvrC* genes. The most noted feature of the consensus sequence for the LexA binding site are the highly conserved trimers CTG and CAG, which are separated by a 10-bp spacer. The suggested LexA binding sites in the *himA* and *uvrC* operons satisfy this minimal criteria only if the spacing is allowed to be 8 bp for *himA* (Mechulam *et al.*, 1985) and 11 bp for *uvrC* (van Sluis *et al.*, 1983; Sharma *et al.*, 1986). However, if these sites are really LexA binding sites, we would still expect the corresponding half binding sites to have relatively high scores when rated by the symmetrical matrix. We found that the half binding sites of the proposed sites are still very poor (2.2/−1.1 for *himA* and 0.3/3.0 for *uvrC*). These results further indicate that the LexA binding sites proposed for *himA* and *uvrC* are not correct.

The CONSENSUS program correctly identifies LexA binding sites even when random DNA sequences are included in its input

One question about the CONSENSUS program is whether it is robust enough to work even if some of the input sequences do not contain a binding site. Therefore, we introduced two random sequences into the list of sequences being analysed

Table III. Sites scoring above 15 according to the 20-base symmetrical matrix

Gene	Location of RNA start site[a]	Location of potential binding site	Score of potential binding site[b]	Score of half binding sites[c]	Notes
cloacin DF13	101, 103	97	30.5	14.3/16.2	
colicin E1	101	97	22 7	13.7/8.9	
		112	19 6	6 5/13 1	
colicin Ia	101	99	22.2	16 2/6.0	d
colicin Ib	101	99	22.2	16.2/6 0	d
recA	101	71	24 2	12.2/12.0	
recN	101	71	26.8	16.2/10.6	d
		93	20.6	10.6/10.0	
		111	19.1	9.1/10.1	e
sulA	101	85	23.8	13.1/10.7	
umuDC	101	91	29.2	16.2/13.0	
uvrA	101	60	21 5	16 2/5.3	
uvrB	101	71	24.5	12.2/12.4	
uvrD	101	102	22.0	12.5/9.5	
colicin A	37	34	23.3	16.2/7 2	d
		48	21.0	5.7/15.3	
lexA	74	55	20.9	13.7/7.2	
		76	21 7	15.3/6.4	
mucAB	59	49	27.7	13.9/13.8	d
himA	101	none			d
uvrC	101	none			

[a] The bases are numbered in the direction of transcription.

[b] The score was determined as described in the Algorithm section.

[c] The score was split into two components representing the first and second 10 bp of the binding site, respectively.

[d] The RNA initiation site was approximated by the location of the −10 consensus sequence.

[e] Not a recognized LexA-binding site.

A

I = 25.115

A	8	7	2	10	0	0	1	0	12	0	11	1	6	5	13	5	5	1	12	0	2	5
C	0	2	0	1	12	0	0	0	1	1	0	1	1	3	0	2	5	11	0	0	1	0
G	0	3	0	1	1	0	12	1	0	0	2	0	3	0	0	0	1	0	0	11	1	2
T	5	1	11	1	0	13	0	12	0	12	0	11	3	5	0	6	2	1	1	2	9	6
	a	a	T	A	C	T	G	T	A	T	A	T	A	T/a	A	T/a	A/c	C	A	G	T	t/A

B

I = 25.108

A	7	8	3	10	0	0	1	0	12	0	12	1	7	6	12	5	5	0	13	1	2	4
C	0	1	0	1	13	0	0	0	1	1	0	1	1	3	1	2	6	11	0	0	1	1
G	0	3	0	1	0	1	12	1	0	0	1	0	2	0	0	0	0	1	0	10	1	2
T	6	1	10	1	0	12	0	12	0	12	0	11	3	4	0	6	2	1	0	2	9	6
	a/t	a	T	A	C	T	G	T	A	T	A	T	A	a	A	T/a	c/A	C	A	G	T	t

C

I = 25.013

A	2	7	8	3	10	0	0	1	0	12	0	12	1	7	6	12	5	5	0	13	1	2
C	3	0	1	0	1	13	0	0	0	1	1	0	1	1	3	1	2	6	11	0	0	1
G	2	0	3	0	1	0	1	12	1	0	0	1	0	2	0	0	0	0	1	0	10	1
T	6	6	1	10	1	0	12	0	12	0	12	0	11	3	4	0	6	2	1	0	2	9
	t	a/t	a	T	A	C	T	G	T	A	T	A	T	A	a	A	T/a	c/A	C	A	G	T

D

I = 24.623

A	8	6	9	3	12	0	1	0	0	7	0	7	0	7	0	12	1	8	0	12	0	3
C	1	0	0	0	0	11	0	1	0	0	3	0	2	1	3	0	2	4	13	1	2	0
G	0	0	4	2	1	0	0	12	3	1	1	1	0	2	0	1	0	0	0	0	9	2
T	4	7	0	8	0	2	12	0	10	5	9	5	11	3	10	0	10	1	0	0	2	8
	a	t/a	a	T	A	C	T	G	T	A	T	A	T	A	T	A	T	A	C	A	G	T

Fig. 6. The top 22-position matrices found by the CONSENSUS program after analysing the first 11 LexA-regulated genes listed in Table I along with two random sequences. The matrices are presented as in Figure 2 Each matrix corresponds to a different ordering of the input sequences.

(Figure 6). As expected, the resulting consensus matrices had a reduced information content relative to the previous results (Figure 3A). However, the correct consensus pattern was identified even when the random sequences were the first two in the list (Figure 6D). Initially we were surprised by this result; however, upon further consideration, we realized that this result is expected.

If one has M random matrices and M L-mers, then, on average, each L-mer will be the top L-mer for one of the matrices. By Poisson statistics, this means that an L-mer has a $(1-e^{-1}) = 63\%$ chance of being the top L-mer for at least one matrix. However, for each LexA binding site, there are five 20-mers (or seven 22-mers) containing the 16-base core that has the most information. Therefore, there is a $(1-e^{-5}) = 99.3\%$ chance that at least one 20-mer containing the 16-base core will be the top 20-mer for one of the matrices. Thus, our algorithm can handle data even if it contains a couple of 'bad' sequences.

Discussion

We have developed a method for identifying consensus patterns in sets of biologically related, but unaligned, DNA sequences. The method describes sequence patterns with matrices and seeks the matrix with the highest information content. It builds up the pattern by sequentially comparing additional sequences to the saved matrices and identifying, for each saved matrix, which portion of the new sequence will most increase the information content of the matrix. The memory required for the program is independent of the number of sequences and is linearly dependent on the length of the first sequence. The time required is dependent linearly on the number of sequences and on the square of their lengths.

This method can accurately determine the patterns of the binding sites for the *E.coli* LexA (this paper) and CRP (Stormo and Hartzell, 1989) proteins. In both these examples, the order in which the sequences are presented to the program is not critical. The program is also robust enough to tolerate some sequences that do not contain binding sites. In addition, we have demonstrated how the matrices determined by our method can be used to identify binding sites.

Our analysis suggests that, in the *recN* promoter region, there may be a LexA binding site that has not been previously noticed. It also indicates that weaker binding sites tend to be adjacent to other LexA binding sites, perhaps to facilitate cooperative binding. In this paper, we have presented various matrices for describing the LexA binding site; however, we believe that the 20-base symmetrical matrix (Figure 5) is the best. This is because (i) the matrix is exactly symmetrical, consistent with LexA being a dimer (Brent and Ptashne, 1981), (ii) it includes all the known binding sites, and (iii) it cleanly distinguishes the binding sites from other sequences (Table III).

We have recently modified the CONSENSUS program to work with an arbitrary alphabet, instead of being limited to the four nucleotides. Thus, the program can now also be used with protein sequences. In recent tests, we have found that the program succeeds in identifying many known protein motifs (unpublished data). We have also developed a related algorithm that can construct matrices containing more than one binding site from each DNA sequence (G.Hertz and G.Stormo, in preparation).

We are currently refining the algorithm and are starting to use it to analyse sets of sequences whose consensus patterns are not already known. We are also investigating alternative methods for limiting the number of matrices to save after each cycle of the algorithm. One approach we plan to test is to save the best matrices, e.g. the best 1000 matrices, without regard to their parentage (see Bacon and Anderson, 1986, for an application of this approach).

As the amount of sequence data increases, simple ways to identify consensus patterns with a small amount of biological or biochemical data will become increasingly important. Our algorithm can be an important tool for achieving this. For example, if one knows a few DNA fragments that are bound by a common protein, our program can potentially determine what the protein's recognition site is without the need for making mutants or doing footprints. In conclusion, we have developed a simple method for identifying consensus patterns in functionally related DNA sequences. This method should be a valuable tool for helping to analyse the enormous amount of sequence data that is currently being generated.

Acknowledgements

This work was supported by Public Health Service grants GM-28755 and GM-41766 from the National Institutes of Health.

References

Bacon,D.J. and Anderson,W.F. (1986) Multiple sequence alignment *J Mol. Biol.*, **191**, 153–161

Beck,E. and Bremer,E. (1980) Nucleotide sequence of the gene *ompA* coding the outer membrane protein II* of *Escherichia coli* K-12. *Nucleic Acids Res.*, **8**, 3011–3024.

Brent,R. and Ptashne,M. (1981) Mechanism of action of the *lexA* gene product. *Proc. Natl. Acad. Sci. USA*, **78**, 4204–4208.

Cole,S.T. (1983) Characteristics of the promoter for the LexA regulated *sulA* gene of *Escherichia coli*. *Mol. Gen. Genet.*, **189**, 400–404.

Easton,A.M. and Kushner,S.R. (1983) Transcription of the *uvrD* gene of *Escherichia coli* is controlled by the *lexA* repressor and by attenuation. *Nucleic Acids Res.*, **11**, 8625–8640.

Ebina,Y., Takahara,Y., Kishi,F., Nakazawa,A. and Brent,R. (1983) LexA protein is a repressor of the colicin E1 gene. *J. Biol. Chem.*, **258**, 13258–13261.

Finch,P. and Emmerson,P.T. (1983) Nucleotide sequence of the regulatory region of the *uvrD* gene of *Escherichia coli*. *Gene*, **25**, 317–323.

Forster,J.W. and Strike,P. (1985) Organization and control of the *Escherichia coli uvrC* gene. *Gene*, **35**, 71–82.

Harr,R., Häggström,M. and Gustafsson,P. (1983) Search algorithm for pattern match analysis of nucleic acid sequences. *Nucleic Acids Res.*, **11**, 2943–2957.

Horii,T., Ogawa,T. and Ogawa,H. (1980) Organization of the *recA* gene of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **77**, 313–317.

Kitagawa,Y., Akaboshi,E., Shinagawa,H., Horii,T., Ogawa,H. and Kato,T. (1985) Structural analysis of the *umu* operon required for inducible mutagenesis in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **82**, 4336–4340.

Little,J.W., Mount,D.W. and Yanisch-Perron,C R. (1981) Purified *lexA* protein is a repressor of the *recA* and *lexA* genes. *Proc. Natl. Acad. Sci. USA*, **78**, 4199–4203.

Mankovich,J.A., Hsu,C -H. and Konisky,J. (1986) DNA and amino acid sequence analysis of structural and immunity genes of colicins Ia and Ib. *J. Bacteriol.*, **168**, 228–236.

Markham,B.E., Little,J.W. and Mount,D.W. (1981) Nucleotide sequence of the *lexA* gene of *Escherichia coli* K-12. *Nucleic Acids Res.*, **9**, 4149–4161.

Mechulam,Y., Fayat,G. and Blanquet,S. (1985) Sequence of the *Escherichia coli pheST* operon and identification of the *himA* gene. *J. Bacteriol.*, **163**, 787–791.

Miller,H.I. (1984) Primary structure of the *himA* gene of *Escherichia coli*. homology with DNA-binding protein HU and association with the phenylalanyl-tRNA synthetase operon. *Cold Spring Harbor Symp Quant. Biol.*, **49**, 691–698.

Miller,H.I., Kirk,M. and Echols,H (1981) SOS induction and autoregulation of the *himA* gene for site-specific recombination in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **78**, 6754–6758.

Morlon,J., Lloubès,R., Varenne,S., Chartier,M. and Lazdunski,C. (1983) Complete nucleotide sequence of the structural gene for colicin A, a gene translated at non-uniform rate. *J. Mol Biol.*, **170**, 271–285.

Perry,K.L , Elledge,S.J., Mitchell,B B., Marsh,L. and Walker,G.C. (1985) *umuDC* and *mucAB* operons whose products are required for UV light- and chemical-induced mutagenesis. UmuD, MucA, and LexA proteins share homology. *Proc. Natl. Acad. Sci. USA*, **82**, 4331–4335.

Peterson,K.R., Ossanna,N., Thliveris,A.T., Ennis,D.G. and Mount,D W. (1988) Derepression of specific genes promotes DNA repair and mutagenesis in *Escherichia coli*. *J. Bacteriol.*, **170**, 1–4.

Rostas,K., Morton,S.J., Picksley,S.M. and Lloyd,R.G. (1987) Nucleotide sequence and LexA regulation of the *Escherichia coli recN* gene. *Nucleic Acids Res.*, **15**, 5041–5049.

Sancar,A., Stachelek,C., Konigsberg,W and Rupp,W.D (1980) Sequences of the *recA* gene and protein. *Proc. Natl. Acad. Sci. USA*, **77**, 2611–2615.

Sancar,A , Sancar,G.B., Rupp,W.D., Little,J.W. and Mount,D W. (1982a) LexA protein inhibits transcription of the *E.coli uvrA* gene *in vitro*. *Nature*, **298**, 96–98.

Sancar,G.B , Sancar,A., Little,J.W. and Rupp,W.D. (1982b) The *uvrB* gene of *Escherichia coli* has both *lexA*-repressed and *lexA*-independent promoters. *Cell*, **28**, 523–530.

Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.

Sharma,S., Stark,T.F., Beattie,W.G. and Moses,R.E. (1986) Multiple control elements for the *uvrC* gene unit of *Escherichia coli*. *Nucleic Acids Res.*, **14**, 2301–2318.

Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic 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. (1989) Consensus patterns in DNA. In Doolittle,R.F. (ed.), *Methods in Enzymology: Computer Analysis of Protein and Nucleic Acid Sequences*. Academic Press, NY, in press.

Stormo,G.D. and Hartzell,G.W.,III (1989) Identifying protein binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA*, **86**, 1183–1187.

Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A (1982) Use of the 'perceptron' algorithm to distinguish translational initiation sites in *E.coli*. *Nucleic Acids Res.*, **10**, 2997–3011

Stormo,G.D., Schneider,T.D. and Gold,L. (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity *Nucleic Acids Res.*, **14**, 6661–6679.

van den Elzen,P.J.M , Hakkaart,M.J.J., van Putten,A.J., Walters,H.H.B., Veltkamp,E. and Nijkamp,H.J.J. (1983) Structure and regulation of gene expression of a Clo DF13 plasmid DNA region involved in plasmid segregation and incompatibility. *Nucleic Acids Res.*, **11**, 8791–8808.

van Sluis,C.A., Moolenaar,G.F. and Backendorf,C. (1983) Regulation of the *uvrC* gene of *Escherichia coli* K12: localization and characterization of a damage-inducible promoter. *EMBO J.*, **2**, 2313–2318.

Varley,J.M. and Boulnois,G.J. (1984) Analysis of cloned colicin Ib gene: complete nucleotide sequence and implication for regulation of expression *Nucleic Acids Res.*, **12**, 6727–6739.

Walker,G.C. (1984) Mutagenesis and inducible responses to deoxyribonucleic acid damage in *Escherichia coli*. *Microbiol. Rev.*, **48**, 60–93.

Walker,G.C. (1985) Inducible DNA repair systems. *Annu. Rev. Biochem.*, **54**, 425–457.

Yamada,M., Ebina,Y., Miyata,T., Nakazawa,T. and Nakazawa,A. (1982) Nucleotide sequence of the structural gene for colicin E1 and predicted structure of the protein *Proc. Natl Acad. Sci. USA*, **79**, 2827–2831.

Received on November 6, 1989; accepted on December 20, 1989