J. Mol. Biol. (1985) **186**, 117–128

Rigorous Pattern-recognition Methods for DNA Sequences

Analysis of Promoter Sequences from *Escherichia coli*

David J. Galas[1], Mark Eggert[2] and Michael S. Waterman[1,2]

[1]*Department of Molecular Biology*
and [2]*Department of Mathematics*
University of Southern California
Los Angeles, Calif. 90089-1481, U.S.A.

(Received 28 September 1984, and in revised form 18 June 1985)

The basic nature of the sequence features that define a promoter sequence for *Escherichia coli* RNA polymerase have been established by a variety of biochemical and genetic methods. We have developed rigorous analytical methods for finding unknown patterns that occur imperfectly in a set of several sequences, and have used them to examine a set of bacterial promoters. The algorithm easily discovers the "consensus" sequences for the −10 and −35 regions, which are essentially identical to the results of previous analyses, but requires no prior assumptions about the common patterns. By explicitly specifying the nature of the search for consensus sequences, we give a rigorous definition to this concept that should be widely applicable. We also have provided estimates for the statistical significance of common patterns discovered in sets of sequences.

In addition to providing a rigorous basis for defining known consensus regions, we have found additional features in these promoters that may have functional significance. These added features were located on either side of the −35 region. The pattern 5′, or upstream, from the −35 region was found using the standard alphabet (A, G, C and T), but the pattern between the −10 and the −35 regions was detectable only in a sub-alphabet. Recent results relating DNA sequence to helix conformation suggest that the former (upstream) pattern may have a functional significance. Possible roles in promoter function are discussed in this light, and an observation of altered promoter function involving the upstream region is reported that appears to support the suggestion of function in at least one case.

1. Introduction

Among the functional patterns in DNA sequences that have clear biological significance, certainly the most well-studied are those that determine the expression of genetic information. Among these patterns, the most carefully examined are undoubtedly promoter sequences that specify the initiation of transcription in *Escherichia coli.* The promoters for *E. coli* RNA polymerase are known to contain two regions of partially conserved DNA sequence, that are located about 10 and 35 base-pairs from the transcription start site (Schaller *et al.*, 1975; Pribnow, 1975; Reznikoff & Abelson, 1978; Siebenlist *et al.*, 1980; Rosenberg & Court, 1979; Hawley & McClure, 1983). These "consensus" sequences are known to contain functional information that affects the activity of the promoter (Hawley & McClure, 1983; Mulligan *et al.*, 1984), but it is not known to what extent other more subtle features of the DNA sequences may also affect promoter function. Since the sequence of the promoter determines the physical structure (or the available repertoire of physical structures) of the promoter, it may be expected that the function of the promoter can be affected by features specified in other regions of the sequence, even if the largest part is determined by the −10 and −35 regions, as it seems to be (Mulligan *et al.*, 1984). Results supporting this view have been reported (Bossi & Smith, 1984).

The problem of finding unknown patterns that occur imperfectly in a set of many sequences, finding a consensus sequence without any prior assumption of the nature of the answer, is a difficult one. This is primarily because of the enormous number of alignments possible for even a modest number of sequences. We have presented a mathematical method that solves this problem in general (Waterman *et al.*, 1984). In this paper, we

0022-2836/85/210117-12 $03.00/0

describe the implementation and application of the algorithm to a biological problem of considerable interest. A fundamental aspect of any method designed to solve this problem is that it requires a careful definition of the meaning of the idea of a consensus sequence. This precise definition must then be represented in the parameters of the algorithm. Such a representation is described below.

The method has been used to re-examine *E. coli* promoters, as a test problem for the technique but, more importantly, to determine the precise nature and extent of all significant common patterns in the set of promoter sequences. The prokaryotic promoter sequences constitute an ideal test set because there are a large number of known sequences (for a recent compilation of sequences, see Hawley & McClure, 1983), the known common patterns are conserved only partially, and their positions with respect to the transcription start site are somewhat variable.

The results of the re-examination reported here demonstrate in a rigorous fashion: (1) that the most highly conserved common sequences are indeed those found at the -10 and -35 positions; and (2) that there are two additional, weaker patterns. These patterns, found in many but not all promoters, appear on either side of the -35 region. We wish to call attention to the possibility that one of these patterns, occurring in the region just upstream from the -35 region, may be implicated in promoter function by producing a conformational irregularity in the DNA.

2. Method of Analysis

(a) *Previous analyses*

There have been several previous attempts to use computer analysis to locate promoter signals. Most of these analyses depend on prior knowledge of the promoter signals, or consensus sequences. They then proceed to search for the occurrence of the patterns with a specified degree of ambiguity. An adaptation of the regular expression search algorithms (contained in the UNIX utility programs) for biological problems was described by Arbanel *et al.* (1984). Regular expression searches provide a highly flexible and useful tool to search for already known patterns, but are not useful in finding unknown patterns. In another approach, Mulligan *et al.* (1984) used the previously determined consensus alignment (Hawley & McClure, 1983) to compute a "homology" score with the *E. coli* promoters. They used this score evaluation to search for and evaluate possible promoters in DNA sequences. Staden (1984) used a weight matrix derived from the alignment presented by Hawley & McClure (1983). Again, a score is given to account for variation in the strength of the putative promoters. A discussion of problems in the use of dynamic programming (matrix) methods for these problems was given by Sadler *et al.* (1983).

The most thorough recent compilation and study of the *E. coli* promoter sequences was performed by Hawley & McClure (1983). They assumed initially that T-T-G-A-C-A in the -35 region, and T-A-T-A-A-T in the -10 region were ideal promoters and attempted to maximize the homology with these 6 base-pair patterns among the 112 listed promoter sequences. The spacing allowed between the 2 patterns in sliding the segments to find the maximum alignment was 15 to 21 base-pairs, with a preferred spacing of 17 base-pairs. In a subsequent study, Mulligan *et al.* (1984) demonstrated a rough correlation between the extent of agreement with their consensus sequence and the strength of promoter function.

Obviously, it is easier to find a pattern when it is already known than when it is unknown. However, by relying on previous work and assuming that a pattern is known, one might miss important features of the sequences. Thus, an important and general problem in sequence analysis is that of determining an unknown consensus sequence from a set of sequences of known function.

The difficulty presented by the requirement that no prior assumptions be made is substantial. A useful illustration of this point lies in considering a straightforward, though naïve, approach to the problem. If r DNA sequences are known, we write them in an initial alignment as

$$\begin{array}{lllll} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\ & & \vdots & & \\ a_{r1} & a_{r2} & a_{r3} & \ldots & a_{rn} \end{array}$$

Such an initial alignment might be made on a known biological feature, such as the start site for transcription, coding, etc. If our consensus patterns can be found directly from this alignment, the analysis is easily done. In many cases, however, as for the promoters, several of the sequences must be moved relative to the others to find the common pattern desired. The computational aspects of moving the sequences must be examined in some more detail. If, for example, we permit each sequence to remain in the initial position or move only 1 position to the left, then the total number of alignments is $2\times2\times2\ldots\times2=2^r$ for r sequences. For $r=100$, $2^r=1{\cdot}27\times10^{30}$. If the sequences were to be considered in up to 5 positions to the left, the number of alignments would be $6\times6\times6\ldots\times6=6^r$, and for $r=100$ we would have $6^{100}=6{\cdot}53\times10^{77}$. It is evident, then, that the direct approach can work on only the smallest of problems.

Ignoring the difficulties of computation for the moment, let us turn to methods that look for consensus letters in the columns of a given alignment. Several authors have computed, for the promoters, the numbers of A, T, G and C in each column (Hawley & McClure, 1983). A consensus pattern is then taken from majority letters in regions where 1 or 2 letters occur at a frequency above a predetermined level. Consider the example below, in which the 3-letter region might be part of 3 longer sequences:

Sequence 1	... A-A-T ...
Sequence 2	... A-T-A ...
Sequence 3	... T-A-A ...

A	2	2	2
T	1	1	1
G	0	0	0
C	0	0	0

By the above criterion, the consensus sequence would be A-A-A. An important feature of such an analysis is that any permutation of letters in a column leaves the column totals unchanged. Thus, in the above example the set

... A-A-A ...
... A-A-A ...
... T-T-T ...

would have the same consensus sequence. This suggests a difficulty with this sort of analysis in that it is based on letter occurrence by position instead of on the patterns themselves. Any subtlety could easily escape detection.

(b) *The present method*

The approach developed here is an implementation and application of the method of Waterman *et al.* (1984). As noted there, the method is related to work by Parzen (1962) on a method of estimating probability density functions. Waterman & Whiteman (1978) have discussed related aspects of density estimation and some of its applications. In addition, Queen *et al.* (1982) have proposed a method that has elements in common with that used here. Their method differs from ours in important ways, which limit the usefulness of their algorithm (Waterman *et al.*, 1984). Also, we have become aware of an unpublished analysis of the *E. coli* promoter sequences (T. F. Smith, personal communication) that uses the method of enumerating the occurrence of exact 4-mers. Our method, which is based on approximate occurrence, is fundamentally different.

In essence, we base our approach on the occurrence of k-letter words, the basic objects of interest. For promoter sequences, for example, a natural first choice of word size would be 6. Each pattern, or k-letter word has an associated set of neighborhoods in the set of all k-letter words. First, there is the exact k-letter word w, which occurs in the sequence. Then there are neighborhoods of word w that differ from w, by 1, 2, ... mismatches, deletions or insertions. The user of the algorithm can choose the degree and nature of the "fuzziness" of the search by specifying which neighborhoods are to be used.

Another parameter that must be set is W, the number of contiguous columns to be searched at a time, which we will call the window width. For example, $k = 6$ and $W = 10$ means that each sequence will be searched for 6-letter words in 10 contiguous base-pairs. It is also possible to think of these parameters as allowing sliding of the sequences relative to one another of up to 5 base-pairs.

Too broad a window will give statistically insignificant results, while too narrow a window will not find patterns that are misaligned by more than the window encompasses. For example, there are a large number of fuzzy T-A-T-A-A-T patterns in each entire promoter sequence if we allow up to 2 mismatches. What is statistically significant is the fact that there are a large number that are only a few base-pairs out of alignment with respect to each other when the promoters are aligned on the transcription start site. Clearly, the search must be limited in width, to limit random matches, but not so limited as to exclude misaligned patterns (see section (c), below, for quantitative estimates).

It is the specification of the word size, of the neighborhoods and of the window width that effectively define the meaning of consensus sequence here. The probability of matching random sequences increases with the number of neighbors considered and the window width specified. The number of functional, common sequences located by the algorithm also increases with the number of neighborhoods and the window width. The probability calculations determine the significance of the patterns found by the algorithm. We emphasize here that one definition of consensus sequence, made by specifying a set of parameters, may not give the same results as another definition, and therefore what is meant by a consensus must be defined carefully in each case. This important problem of specification has been ignored in most sequence analyses.

A key computation in our method can be described as follows. Let q_{w0} be the number of lines (sequences) that contain at least one exact occurrence of the word w. Then q_{w1} is the number of lines that do not have an exact occurrence of w but have a word that is 1 mismatch from w. Generally, q_{wd} is the number of lines for which the best representative of w is as a dth neighbor. Details of this computation were discussed by Waterman *et al.* (1984). Then a score for word w is found by defining:

$$S(w) = \sum_{d \leqslant d*} \lambda_d q_{wd},$$

where λ_d is the weight given to a dth neighbor. This weight determines the importance of these neighbors to the overall pattern. In the computations performed for this paper, we used a natural, but arbitrary, specification of λ_d:

$$\lambda_d = \frac{\text{Number of matches between } w \text{ and } d\text{th neighbor}}{\text{Number of bases in the pattern } w}.$$

For this scheme:

$$\lambda_{0 \text{ mismatches}} = 1,$$

$$\lambda_{d \text{ mismatches}} = \frac{k-d}{k} = 1 - \frac{d}{k},$$

$$\lambda_{d \text{ insertions/deletions}} = 1 - \frac{d}{k}.$$

Here d insertions/deletions refers to the number of letters involved.

The algorithm is then specified by the parameters:

W = window width

k = pattern or word size

d^* = neighborhood specification,

and the winning score, for the "most common" pattern, is found by:

$$S = \max_{w'} S(w').$$

The score, S, is then plotted above the alignment position representing the (right) edge of the window position for that score. Representative plots are shown in Figs 1 and 2 with the parameters indicated in the Figure legends. The running time of the program is proportional to:

$$r(n-W+1)(W-k+1)N(d^*),$$

where:

n = common sequence length,

r = number of sequences,

$N(d^*)$ = number of words in neighborhood

and:

k = pattern length.

$N(d^*)$ and 4^k can both grow rapidly, and it is practical to use word size only up to $k = 7$. For example, with $k = 6$, up to 2 mismatches gives $N(d^*) = 1+18+135 = 154$. Even so, with $k = 6$, 3 mismatches, $W = 10$ and $N(d^*) = 694$, our program runs on the SUN MC68010 computer for 59 sequences about 60 letters long in about 3 min. Notice that doubling the number of the sequences or the length of the sequences simply doubles the running time. This is because the running time is linear in n and r.

(c) *Statistical significance*

An essential addendum to the determination of the "most common" or consensus sequence according to given criteria is the determination of the statistical significance of found patterns. In the following, we assume for simplicity that all 4 letters occur at equal frequencies.

Assume that w is a given pattern of length k, having $N(d^*)$ neighbors of length k. We use:

$$\alpha = N(d^*)(W-k+1)4^{-k}$$

to approximate the probability that w or some neighbor of w occurs, on a given line (in a given sequence), with a given position of the window of width W. Thus, if the sequences are random, for each word and window position j we would expect approximate matches to w on about $(\alpha)(r)$ of r lines. The probability of a fraction $\beta > \alpha$ of lines with matches is estimated next.

Suppose we decide to look for a pattern common to some preset fraction $\beta > \alpha$ of the r sequences. For word w and window position j, the probability that at least $(\beta)(r)$ lines yield approximate matches to w is estimated as $\exp(-rH(\beta,\alpha))$, where:

$$H(\beta,\alpha) = \beta \log(\beta/\alpha) + (1-\beta)\log\left(\frac{1-\beta}{1-\alpha}\right) > 0$$

is the entropy of β relative to α (Waterman *et al.*, 1984). There are n choices for the location j of the window and, if the pattern w is unknown, there are 4^k choices for the word w. Thus our estimated significance level p for all words w is:

$$p = n\,4^k \exp(-rH(\beta,\alpha)).$$

The significance level p is an upper bound, for random data, on the probability that, in some window position, an approximate match occurs on a fraction of the sequences greater than or equal to β. When the estimate p exceeds 1, simply use 1 instead; in that case, no located matches can be said to be significant. In effect, p is indicating the noise level that corresponds to a given set of parameters. If the signal is to be detected it must be stronger than the noise, thus the practical implication that the weaker the signal is in each sequence the better we must be able to align the sequences based on other information in order to find the signal.

(d) *Extension of the methods*

The methods described here are surprisingly versatile. Many aspects are reminiscent of signal detection techniques. In the above, we have emphasized the search for the best, or most common, word. It is a simple matter to ask for the best word in several positions, so that a multiple signal can be detected. In addition, it is possible to search for the words with the worst scores (the least common patterns); that is, words that occur significantly less often than expected. These are also potential signals in DNA sequences. Another feature of the method is that it permits an alignment of the sequences on any signal that has been located, and the search can be re-done on the aligned set. This allows one to focus closely on selected portions of the sequence and to iteratively re-align the sequences according to their own common features, thus effectively amplifying the signals.

Searches can be performed with any one of the various possible alphabets. A natural sub-alphabet may be purine–pyrimidine, but one can search on any such reduced alphabet (see Results). Note that the relationship between sequence and structure of the DNA that is beginning to be elucidated (Dickerson, 1983) can be represented by sub-alphabet patterns by which common DNA structural motifs may be sought by these methods.

Finally, we wish to point out that the general method can be extended in several ways. For instance, if we would ascribe a quantitative characterization of the promoter (a quantitative indication of promoter strength like $K_b k_2$, for example (Mulligan *et al.*, 1984)), the score for a given pattern could be computed on the basis of contributions from each line made proportionately to their strengths. The consensus sequence so obtained would better characterize strong promoters than the equally weighted one.

There is no known biological significance to the scoring parameters (penalties for mismatches, insertions and deletions). Experience with pattern analysis in DNA sequences and increased biochemical insight into DNA structural constraints and DNA–protein interactions will permit refinement of the algorithm in this respect.

3. Results

Using our algorithm on a set of promoter sequences, we have found in a rigorous fashion several features that show up as common patterns, some that are known and some that have not been noted before. In addition, we are able to refine somewhat the known features of promoter sequences. Here, we describe the significant patterns we have found in the set of 59 bacterial promoters listed in Table 1.

(a) *The −10 and −35 regions*

When the program was run on a set of 59 *E. coli* promoters (compiled by Hawley & McClure, 1983), we found a strong signal at the positions expected for the known partially conserved sequences (the −10 and −35 regions) for several sets of parameters. The signals are thus seen to be strong, and robust to variations in the "definition" of consensus sequence. In Figure 1 we have plotted the weighted frequency of occurrence of the best common sequence as a function of the position of the window. For Figure 1(a), (b) and (c), the word length, window size and neighborhood definition were identical (6, 9 and 2 mismatches, respectively). For the results shown in Figure 1(a), the sequences were aligned at the transcription start sites. For Figure 1(b), however, the program aligned the sequences on the best neighborhood sequence at the top of the peak signal in the −10 region. The sequences were thus effectively lined up on the best example of what was found in the −10 region (T-A-T-A-A-T), and a new scan executed. The result of this procedure is that the signal at −10 is enhanced somewhat, while the signal at −35 is found to be about the same in magnitude. What this indicates is that the distance between the −10 and −35 signals is not correlated with the distance between the −10 signal and the transcription start site. The possibility of "self-alignment" is a powerful feature of the implemented method, in that a strong signal that appears to be weak because of misalignment of the sequences may in some cases be uncovered by re-aligning on the weaker signal, initially located by the program. In

Table 1
Bacterial promoters

	5' +1 3'
araBAD	TTAGCGGATCCTACCTGACGCTTTTTATCGCAACTCTCTACTGTTTCTCCATACCCGTTTTT
araC	GCAAATAATCAATGTGGACTTTTCTGCCGTGATTATAGACACTTTTGTTACGCGTTTTTGT
galP1	CTAATTTATTCCATGTCACACTTTTCGCATCTTTGTTATGCTATGGTTATTTCATACCATAAG
galP2	CACTAATTTATTCCATGTCACACTTTTCGCATCTTTGTTATGCTATGGTTATTTCATACC
lacP1	TAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGC
lacP2	TTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCG
lacI	GACACCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGT
malEFG	AGGGGCAAGGAGGATGGAAAGAGGTTGCCGTATAAAGAAACTAGAGTCCGTTTAGGTGT
malK	CAGGGGGTGGAGGATTTAAGCCATCTCCTGATGACGCATAGTCAGCCCATCATGAATG
malT	TAAAAAAACGTCATCGCTTGCATTAGAAAGGTTTCTGGCCGACCTTATAACCATTAATTACG
tnaA	TAAACAATTTCAGAATAGACAAAAACTCTGAGTGTAATAATGTAGCCTCGTGTCTTGCG
deoP1	CAGAAACGTTTTATTCGAACATCGATCTCGTCTTGTGTTAGAATTCTAACATACGGTTGC
deoP2	AATTGTGATGTGTATCGAAGTGTGTTGCGGAGTAGATGTTAGAATACTAACAAACTCGCAA
trp	TCTGAAATGAGCTGTTGACAATTAATCATCGAACTAGTTAACTAGTACGCAAGTTCACGT
trpR	TGGGGACGTCGTTACTGATCCGCACGTTTATGATATGCTATCGTACTCTTTAGCGAGTACA
aroH	GTACTAGAGAACTAGTGCATTAGCTTATTTTTTTGTTATCATGCTAACCACCCGGCGAG
trpP2	ACCGGAAGAAAACCGTGACATTTTAACACGTTTGTTACAAGGTAAAGGCGACGCCGCCC
his	ATATAAAAAAGTTCTTGCTTTCTAACGTGAAAGTGGTTTAGGTTAAAAGACATCAGTTGAA
hisA	GATCTACAAACTAATTAATAAATAGTTAATTAACGCTCATCATTGTACAATGAACTGTACAAA
leu	GTTGACATCCGTTTTTGTATCCAGTAACTCTAAAAGCATATCGCATT
ilvGEDA	CAAAAAATATCTTGTACTATTTACAAAACCTATGGTAACTCTTTAGGCATTCCTTCGA
argCBH	TTTGTTTTTCATTGTTGACACACCTCTGGTCATGATAGTATCAATATTCATGCAGTATT
thr	AAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACA
bioA	GCCTCCTCCAAAACGTGTTTTTTGTTGTTAATTCGGTGTAGACTTGTAAACCTAAATCT
bioB	TTGTCATAATCGACTTGTAAACCAAATTGAAAAGATTTAGGTTTACAAGTCTACACCGAAT
fol	CATCCTCGCACCAGTCGACGACGGTTTACGCTTTACGTATAGTGGCGACAATTTTTTTT
uvrB P1	TCCAGTATAATTTGTTGGCATAATTAAGTACGACGAGTAAAATTACATACCTGCCCGC
uvrB P2	TCAGAAATATTATGGTGATGAACTGTTTTTTTATCCAGTATAATTTGTTGGCATAATTAA
uvrB P3	ACAGTTATCCACTATTCCTGTGGATAACCATGTGTATTAGAGTTAGAAAACACGAGGCA
recA	TTTCTACAAAACACTTGATACTGTATGAGCATACAGTATAATTGCTTCAACAGAACAT
lexA	TGTGCAGTTTATGGTTCCAAAATCGCCTTTTGCTGTATATACTCACAGCATAACTGTAT
ampC	TGCTATCCTGACAGTTGTCACGCTGATTGGTGTCGTTACAATCTAACGACATCGCCAATG
lpp	CCATCAAAAAAATATTCTCAACATAAAAAACTTTGTGTAATACTTGTAACGCTACATGGA
hisJ (S.t.)	CAAGGTAGAATGCTTTGCCTTGTCGGCCTGATTAATGGCACGATAGTCGCATCGGATCTG
Pori-r	GATCGCACGATCTGTATACTTATTTGAGTAAATTAACCCACGATCCCAGCCATTCTTCTGC
Pori-l	CTGTTGTTCAGTTTTTGAGTTGTGTATAACCCCTCATTCTGATCCCAGCCATTCTTCTGC
spot 42 RNA	ATTACAAAAAGTGCTTTCTGAACTGAACAAAAAAGAGTAAAGTTAGTCGCGTAGGGTACA
M1 RNA	ATGCGCAACGCGGGGTGACAAGGGCGCGCAAACCCTCTATACTGCGCGCCGAAGCTGACC
alaS	AACGCATACGGTATTTTACCTTCCCAGTCAAGAAAACTTATCTTATTCCCACTTTTCAGT
trpS	CTACGGCGAGGCTATCGATCTCAGCCAGCCTGATGTAATTTATCAGTCTATAAATGACC
glnS	TAAAAAACTAACAGTTGTCAGCCTGTCCCGCTTATAAGATCATACGCCGTTATACGTT
tufB	ATGCAATTTTTTAGTTGCATGAACTCGCATGTCTCCATAGAATGCGCGCTACTTGATGCC
tyrT	TCTCAACGTAACACTTTACAGCGGCGCGTCATTTGATATGATGCGCCCCGCTTCCCGAT
leu1 tRNA	TCGATAATTAACTATTGACGAAAAGCTGAAAACCACTAGAATGCGCCTCCGTGGTAGCAA
supB-E	CCTTGAAAAAGAGGTTGACGCTGCAACGCTCTATACGCATAATGCGCCCCGCAACGCCGA
rrnAB P1	ATTTTAAATTTCCTCTTGTCAGGCCGGAATAACTCCCTATAATGCGCCACCACTGACACGG
rrnG P1	TTTATATTTTTCGCTTGTCAGGCCGGAATAACTCCCTATAATGCGCCACCACTGACACGG
rrnD P1	GATCAAAAAAATACTTGTGCAAAAAATTGGGATCCCTATAATGCGCCTCCGTTGAGACGA
rrnE P1	CTGCAATTTTTCTATTGCGGCCTGCGGAGAACTCCCTATAATGCGCCTCCATCGACACGG
rrnX P1	ATGCATTTTTCCGCTTGTCTTCCTGAGCCGACTCCCTATAATGCGCCTCCATCGACACGG
rrnAB P2	GCAAAAATAAATGCTTGACTCTGTAGCGGGAAGGCGTATTATGCACACCCCGCGCCGC
rrnG P2	AAGCAAAGAAATGCTTGACTCTGTAGCGGGAAGGCGTATTATGCACACCGCCGCGCCG
rrnDEX P2	CCTGAAATTCAGGGTTGACTCTGAAAGAGGAAAGCGTAATATACGCCACCTCGCGACAG
str	TCGTTGTATATTTCTTGACACCTTTTCGGCATCGCCCTAAAATTCGGCGTCCTCATAT
spc	CCGTTTATTTTTTCTACCCATATCCTTGAAGCGGTGTTATAATGCCGCGCCCTCGATA
S10	TACTAGCAATACGCTTGCGTTCGGTGGTTAAGTATGTATAATGCGCGGCTTGTCGT
rpoA	TTCGCATATTTTTCTTGCAAAGTTGGGTTGAGCTGGCTAGATTAGCCAGCCAATCTTT
rplJ	TGTAAACTAATGCCTTTACGTGGGCGGTGATTTTGTCTACAATCTTACCCCCACGTATA
rpoB	CGACTTAATATACTGCGACAGGACGTCCGTTCTGTGTAAATCGCAATGAAATGGTTTAA

Bacterial promoter sequences. This sub-set of bacterial promoters was taken from the compilation by Hawley & McClure (1983). We have added 1 or 2 bases to the 5′ end of several of these sequences where the sequence is known, taking the data from references given by Hawley & McClure (1983). The underlined letter indicates the transcription start site. The alignment is at the transcription start site (where known), following Hawley & McClure (1983).

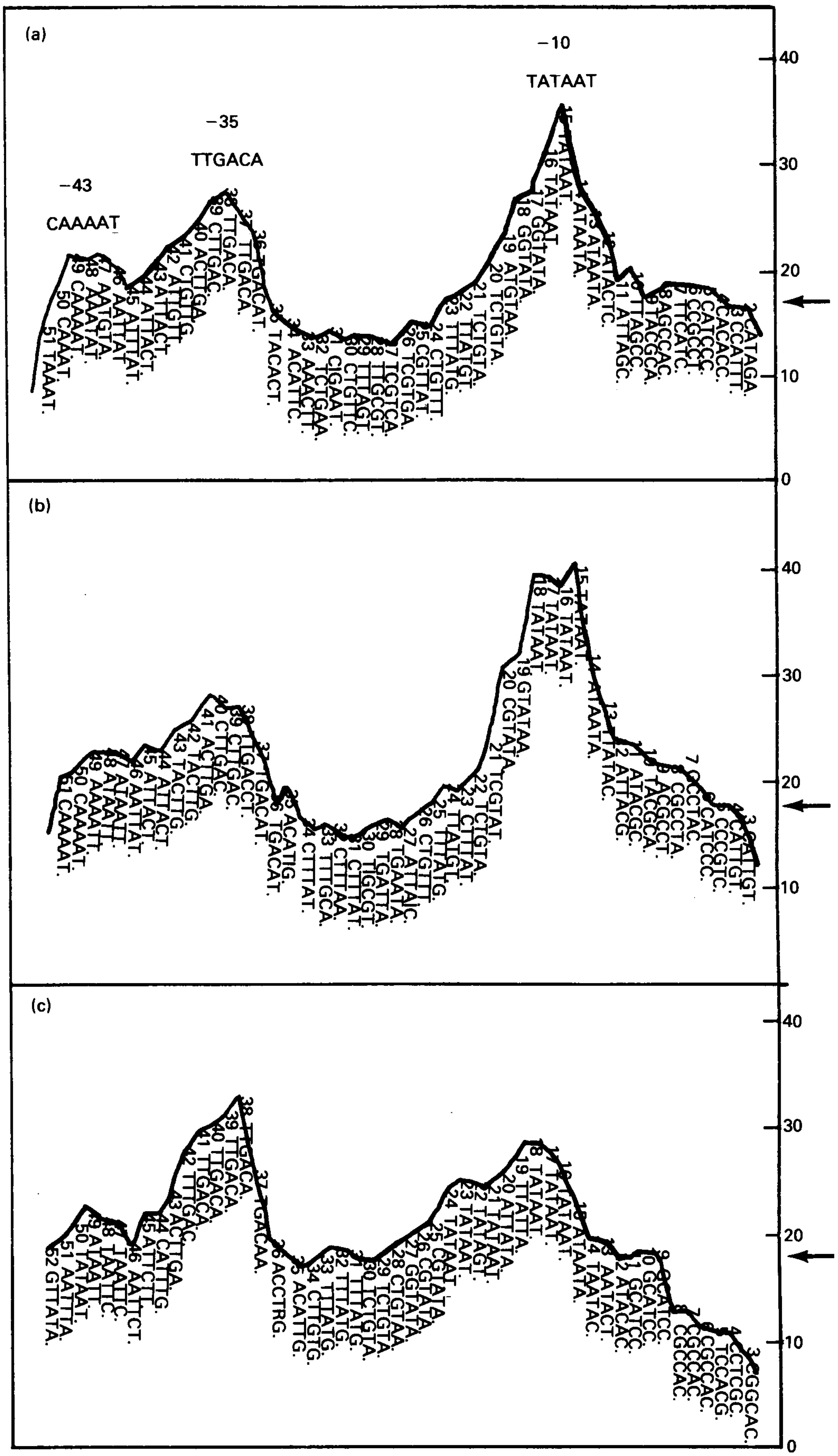
(a)
−43
CAAAAT
−35
TTGACA
−10
TATAAT
40
30
20
10
0
(b)
40
30
20
10
(c)
40
30
20
10
0

Figure 1(c) we see the result of aligning the sequences of the −35 signal from the peak in Figure 1(a). In this case, the signal at −35 is enhanced, while the −10 signal is somewhat diminished. This effect indicates that there is a greater variation in the spacing between the −10 and the −35 than between the −10 and the transcription start site.

The principal result, then, for this set of parameters is that the well-known regions at −10 and −35 are found easily by the algorithm. Note, however, that there is also a third, weaker peak upstream from the −35 region that is evident in Figure 1(a),(b) and (c) but particularly in (a). We wish to postpone discussion of this new signal for the moment, and consider first the nature of the "consensus sequences" in the −10 and −35 regions.

The sequences at the two principal peaks of the plots in Figure 1 indicate the consensus sequences by the criteria specified in the Figure legend. At −10, the (6-letter) consensus for all criteria used here is T-A-T-A-A-T, but at −35 the consensus may vary slightly as the criteria are altered: the two principal consensus sequences are C-T-T-G-A-C and T-T-G-A-C-A. This result is fully consistent with the compilation and alignment done by Hawley & McClure (1983). As an illustration of the importance of the parameters of window width and number of mismatches on our ability to see the signals, we have varied these parameters and plotted the results in Figure 2.

We have found also that including insertions and deletions in the allowed neighborhoods did not alter the results: the consensus sequence or the strength of the signal in the −10 or −35 regions. The algorithm thus allows us to determine that sequences related to each other by insertions and deletions do not contribute to the consensus signals in these regions. The precise position of each base-pair within the signals is apparently important. This is the first analysis in which this question could be addressed.

By varying the parameters of the scan, we can learn something further about the common sequences and sub-sequences. For example, the most common exact, two-letter word in a three base-pair window appears at the −10 region (T-A), while the most common three-letter word (with a 7 base-pair window) appears at the −35 region (T-T-G). The T-T-G in the −35 region is the most strongly conserved three-letter word in the set of promoters.

Because a signal could be encoded easily in the sequence using an alphabet other than the standard one (A,G,C,T), the promoters were scanned for signals using a number of sub-alphabets. When we examined the −10 and −35 regions using all 14 sub-alphabets of the standard alphabets (A or G = R, C or T = Y, etc.). We found nothing striking in these regions that was not a reflection of the consensus sequences in the standard alphabet. Note, however, that a careful quantitative analysis of these signals must be done to determine whether any additional signal in the sub-alphabet may be superimposed on the signal in the standard alphabet. This question will be addressed in a later paper. This was not the case, however, in another region where a signal appears to be encoded in a sub-alphabet (see below).

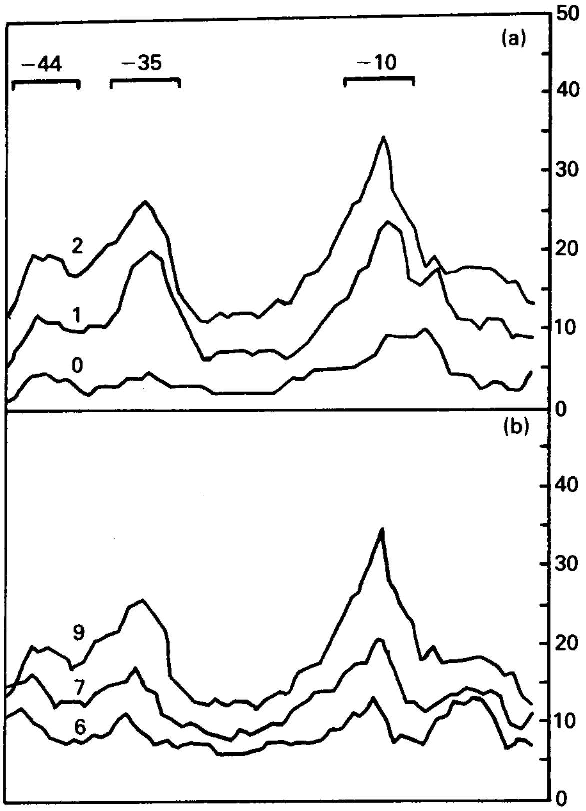

Figure 2. Parameter variation. (a) and (b) The word length $k = 6$. For (a) the window width is held constant, $W = 9$, while the number of mismatches is varied. Plots are shown for 0, 1 and 2. In (b) the number of mismatches is held constant at 2, and the window width is varied. Plots are shown for 6, 7 and 9.

The average spacing between the −10 and −35 signals can be determined by examining the distance between the peaks in Figure 1, for example. The average spacing thus determined is 17 base-pairs (23 minus the 6-letter word length). This is the same result obtained by previous analyses (Hawley & McClure, 1983), and by experimental determination (Stephano & Gralla, 1982; Aoyama *et al.*, 1983).

Figure 1. Plots of scores, S, for highest scoring sequence as a function of window position. The parameters are word length $k = 6$, window width $W = 9$, and number of mismatches = 2 (no insertions or deletions permitted). (a) The sequences are aligned on the transcription start site, as shown in Table 1; (b) the sequences are aligned on the −10 peak; (c) on the −35 peak, as described in the text. The expected value for random sequences of identical composition is marked. Sequence hyphens are omitted for clarity in the Figures.

(b) *The −44 region*

As noted above, we found a distinct signal (in the standard alphabet) in the region of −44. This signal, shown in Figure 1, appears to be the next most strongly conserved six-letter word. When we reduced the word size, this signal remained relatively strong. The sequence that defines the consensus varies to some extent with the parameters (more so than for the −10 and −35 signals), but always includes the triplet A-A-A or A-A-T. A hierarchy of common sequences obtained by increasing the word size reads as follows (A-T, A-A-A or A-A-T, A-A-T-T, C-A-A-A-A-T or A-T-A-A-T-T. It is notable that a shadow of this consensus sequence is not found in all promoter sequences. We find a much wider distribution of relatedness (or degree of match) to a consensus in this region than for the −10 or −35 regions, with some promoters exhibiting no trace of the most common sequence. We suggest in the Discussion that the sequence may have a function (perhaps in altering DNA conformation) that requires its presence in only some promoters. McClure and co-workers have noticed the presence of a weak signal in this region, which is reflected in the conservation of an A at position −45 (Hawley & McClure, 1983). The striking A + T-richness 5′ to the −35 region also has been noted for several promoters, in particular those for ribosomal RNA genes (Brosius *et al.*, 1981).

In Table 2 we have listed the promoters, from the set of 59 bacterial promoters used here (Hawley & McClure, 1983), in which the signal sequence is present in the −40 to −50 region. In A we list the promoters and exhibit the signal we find there as it appears. In B we have listed those for which the signals appear in a reverse orientation. Since this signal may be implicated as a conformation switch or determinant, we reasoned that it is possible that the orientation is irrelevant and we took note of signals in reverse orientation in this Table. With 25 promoters listed in A and an additional seven in B (two promoters have a signal in *both* orientations, deoPl, rrnABPl) we note that roughly half of the original set of 59 exhibit some form of the signal.

If we align the set of promoter sequences on the representations of the consensus found at −44 (left-most peak in Fig. 1(a)), we can somewhat enhance the signal in this peak. The signal, however, is not significantly strengthened over that evident in the unaligned set (not shown). This is because only half of the sequences have a representative of the consensus in the −44 region (Table 2A). When we search for consensus signals with only this sub-set of sequences, we obtain the plot shown in Figure 3, in which the peak is now greatly enhanced. For this set, the −44 signal is comparable in importance to the other two consensus signals. We consider the possible significance of the −44 signal in the Discussion.

(c) *The transcription start site*

Since the sample sequences are aligned according to the position of the transcription start site, the analysis of any consensus pattern immediately adjacent to the site is straightforward. We examined the promoters for many different parameters and found that the transcription start site gives the largest signal at a word size of 2 (exact match) and a window size of 3. The best sequence in this case is C-A. However, when we shift to the purine–pyrimidine alphabet, the transcription start gives the strongest signal of the entire promoter (not shown). It would appear that the consensus sequence at the start site may better be represented as Y-R-Y than as C-A-T.

(d) *The −23 region*

It is clear from the plots in Figure 1 that in the segment between the −35 and −10 regions there is

Table 2

Promoters and signals

A. *Promoters with a signal in the −44 region*

lpp	CAAAAAAAT
malT	TAAAAAAAC
his	TAAAAAAG
glnS	TAAAAAAC
spoT 42 RNA	CAAAAG
rrnAB P2	CAAAAT AAAT
supB-E	GAAAAAG
thr	AAT TAAAT
recA	TA CAAAC
bioA	CAAAAC
trp P2	GAA GAAAC
hisA	TA CAAAC AAT
araC	CAAAT AAT
rplJ	TAAAC TAAT
tnaA	TAAAC AAT
rrnAB Pl	TAAAT
rrnD Pl	CAAA GAAAT
deo Pl	CA GAAAC
trp	CA GAAAT
uvrB P2	CA GAAAT
rrnG P2	CAAA GAAAT
rrnDEX P2	GAAAT
hisJ	TAGAAT
uvrB Pl	TATAAT
bio B	CATAAT

B. *Promoters with a signal in reverse orientation*

spc	GAAAAAAT AAAC
tufB	TAAAAAAT
rpoA	GAAAAAT
rrbG Pl	GAAAAAT
rrnE Pl	GAAAAAT
rrnX Pl	GAAAAAT
thr	TAAAAT
deo Pl*	TAAAAC
rrnAB Pl*	TAAAAT
lexA	TAAAC
gal P2	TAAAT

A: Promoters with a signal (in direct orientation) in the −44 region. These 25 promoters were selected from the 59 in Table 1 having a strong representation of the signal 5′ from the −35 region. They are arranged in descending order with respect to the length of the poly(A) stretch and the flanking pyrimidines.

B: Promoters with a signal in reverse orientation selected by the same criterion for the opposite strand. Those marked with an asterisk also have a signal in direct orientation.

Hyphens have been omitted from the sequences for clarity.

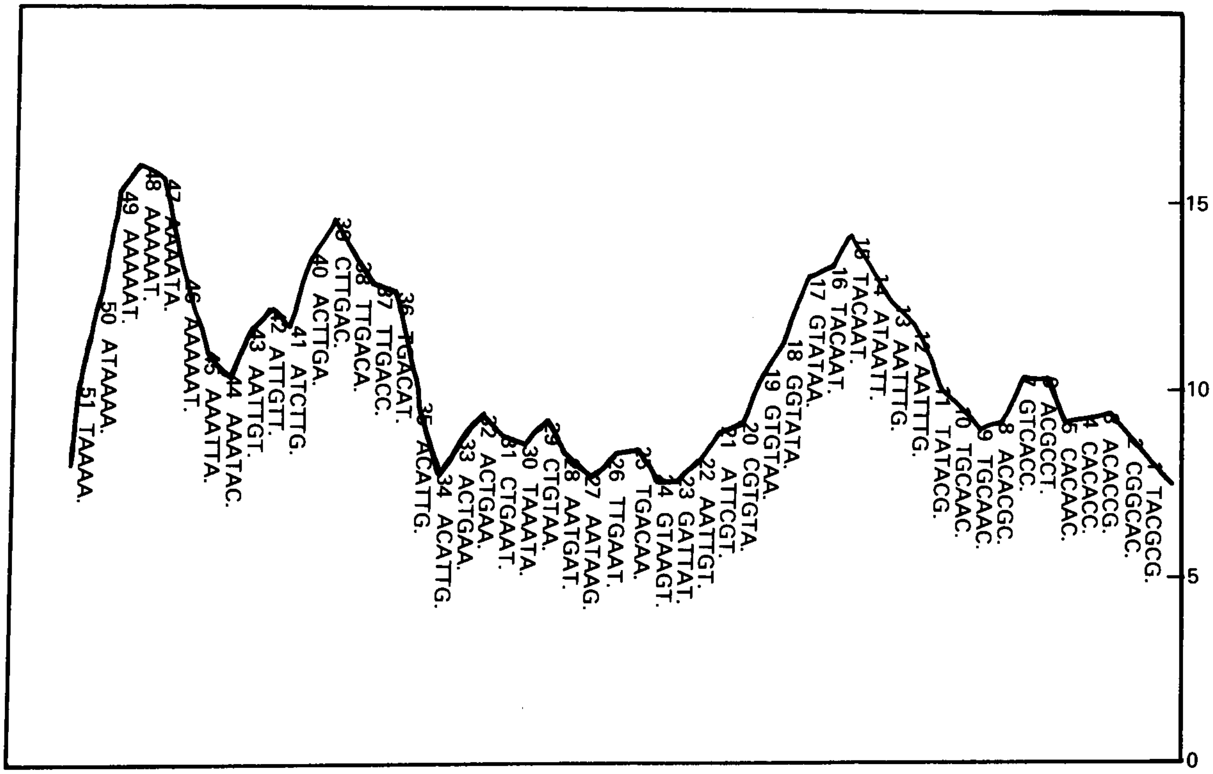

Figure 3. Plot of scores for set of promoters for Table 2A, aligned on the transcription start site. Window width $W = 9$, word length = 6, and 2 mismatches were permitted (same parameters as for Fig. 1).

no significant signal in the standard alphabet. In Figure 4 we show a plot of the results of searches using the sub-alphabet (purine, C, T). With the identity of A to G it is, of course, necessary to sharply reduce the window width, and require more stringent matches to detect signals above the effected random matches. When this is done (in the case of Fig. 4(b), a window width of 7, and a 3-base word length), we see a signal appear directly between the −35 and −10 regions. These latter signals are still seen in the sub-alphabet. The signal is not as strong as either of these two, but it is evident that a T-R-R sequence is found at a significant frequency. We have no way of knowing if this feature has any function in the promoters in which it appears, but the detection of a signal here that cannot be seen in the standard alphabet does demonstrate the versatility of this method as an analytical tool.

4. Discussion

In this study we have sought to demonstrate the usefulness of the approach we have taken to the general problem of DNA sequence pattern recognition. It is natural to turn to the problem of promoter recognition to work out our methods because there are many data available, and previous workers have provided important insights into the problem. The fact that we have demonstrated, with no prior assumptions, that the algorithm finds the known −10 and −35 consensus sequences is not unexpected, and certainly reassuring. Moreover, we have been able to explore the structure of the patterns inherent in the set of *E. coli* promoters and to detect the presence of other, weaker signals that may have some function.

It is important to note that promoters vary in strength (frequency of transcription initiations), over a very wide range. These variations must be encoded, in some way, in the sequence of the DNA itself. The way in which the strength is specified by the sequence has been a subject of a number of studies (see, for example, Rosenberg & Court, 1979; Siebenlist *et al.*, 1980; Mandeki & Reznikoff, 1982; McClure *et al.*, 1983; Youderian *et al.*, 1982). Biochemical and genetic studies have suggested that the degree of match of the −10 and −35 regions with the consensus sequence is indicative of the promoter strength. Mulligan *et al.* (1984) have now shown that a homology score of a given promoter with the consensus, or most common promoter sequence (at each single-letter position) does correlate roughly with the promoter strength, confirming prior suggestions, but there is much left unexplained.

The signal we find upstream from the −35 region is significantly weaker than the −10 and −35 signals of the promoter set we used, and has several notable features. It is found to be strongly represented in a sub-set of the promoters, weakly represented in some, while some of the sequences have no detectable shadow of the signal. This is in contrast to the −10 and −35 signals that are present in some form in the vast majority of the

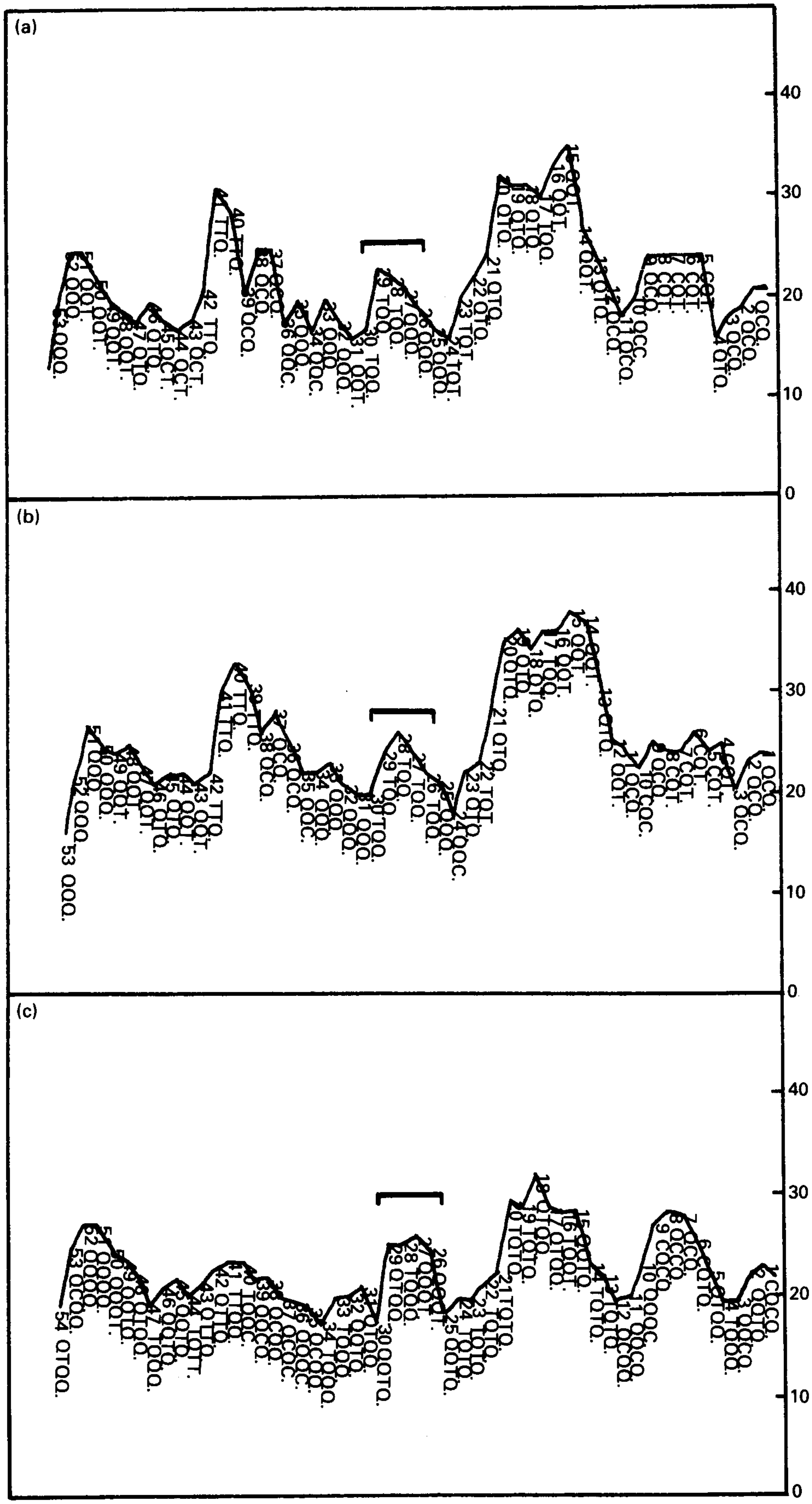

Figure 4. Plot of scores for the full set using the sub-alphabet (C, T, purine). The signal discussed in the text at −23 is indicated by the bracket. For (a), $W = 6$, $k = 3$, no mismatches. For (b), $W = 7$, $k = 3$, no mismatches. For (c), $W = 5$, $k = 4$, and 1 mismatch was permitted. Q represents a purine in this Figure.

promoters. The variation may have a functional correlate, as suggested below.

One of the most strongly represented sequences in this signal is Py-A-A-A-A-Py (with a poly(A) stretch of varying length), a sequence that may assume an unusual conformation. Arnott and coworkers have suggested that double helices formed of poly(A)·poly(T) exhibit a conformation quite distinct from *B*-form DNA because of differences in the conformation of the furanose rings on either side of the double helix (Selsing *et al.*, 1979; Arnott *et al.*, 1983). They call this form heteronomous (H) DNA. Wu & Crothers (1984) have implicated the sequence C-$A_{(5-6)}$-T as the essential part of a pattern that causes bending of the DNA. It is entirely possible that the function of the sequences contributing to the signal upstream from the −35 region is related to their ability to alter the conformation of the DNA in this region. In support of some function for this sequence, it has been found that the CRP protein (CAP) alters the conformation of the *lac* control region when it binds at its site just upstream from the −35 region (Wu & Crothers, 1984; Kolb *et al.*, 1983). These workers postulate that the conformation switch involves bending of the DNA molecule at this site.

If the conformational change induced by CRP protein is important to the enhancement of promoter function that is induced by CRP binding, then there is an obvious hypothesis for the function of the upstream signal. These sequences may function in the same way in the absence of bound CRP protein that the DNA in the same region in the *lac* promoter does when the CRP protein is bound. The bend (or other conformational change) would be switched on and off by CRP binding in the *lac* promoter, enhancing RNA polymerase binding (for a review, see de Crombrugghe *et al.*, 1984), while in other promoters the DNA sequence itself would provide the conformation state necessary to increase promoter activity. Alternatively, the conformational state could affect other promoter sequences 3′ from itself in different ways.

There are, of course, alternative explanations for the functions of these sequences. They could act as protein-binding sites, for example, or enhance the binding of RNA polymerase directly by an unknown mechanism. The hypothesis entertained above has several direct implications, however. It is unlikely that the precise position of a conformation-switching sequence is important. Like the CRP binding sites, the sequences do not seem to be aligned precisely. In addition to the CRP-dependent promoters, and those with a strong signal sequence upstream from the −35 region, there should be a third class, in which the absence of a signal sequence may be important in tuning the promoter strength to its required limited level of expression. It is interesting to note that the positions of the signals at the −35 and −44 regions fall about one turn of the helix apart on average.

The most important implications of the above discussion probably lie in the experiments that are suggested. It should be possible to test systematically the effects of the −44 sequence on the strength of the promoter and its interaction with the −35 and −10 regions. The hypothesis implied by the comparison with CRP protein action is that it will have its primary effect on the binding constant K_b, and thus interact primarily with the −35 region. One specific hypothesis suggested by the work of Malan *et al.* (cited by de Crombrugghe *et al.* (1984)) is that a substitution of a sequence like Py-A-A-A-A-A-Py upstream from the −35 region in the wild-type *lac* promoter will increase its strength and render it less dependent on cAMP-CRP. At present there is no direct evidence for a function of this signal.

It is interesting to note that the sequence C-A-A-A-T, or a close homologue, has been found as a common feature of elements involved in the expression of mouse and immunoglobulin genes, in the simian virus 40 enhancer (Falkner & Zachau, 1984), and in several other sequences involved in control of gene expression (Mattaj *et al.*, 1985).

At present, while there is no direct evidence of function for the −44 region, the region upstream from −35 is implicated in at least one case. Bossi & Smith (1984) have shown that a conformationally unusual sequence in a region much further upstream of a tRNA promoter (at −70) has a distinct effect on the strength of the promoter. We expect that there is a variety of alterations of DNA structure in and near promoter sequences that may affect the strength of transcription. It appears that the −44 signal is a representative sequence that is present in a large enough fraction of *E. coli* promoters to be seen as a consensus signal in the set used here. We have run the program on the set of phage promoters and plasmid promoters listed by Hawley & McClure (1983). The plasmid promoters exhibited essentially the same signals as the set used here (not shown), but the phage promoters as a group showed little evidence of the −44 signal or the signal at −23 (not shown).

There are clearly other patterns of sequence signals that can be detected in selected sub-sets of the promoters. Analyses of the sequence structure of promoters that are stringently controlled, for example, has revealed other patterns (Travers, 1984). In particular, a "discriminator" sequence downstream from the −10 region has been noted. Travers and others have also pointed out that more complex patterns involving multiple occurrences of −35 and −10 regions that may be part of overlapping promoters are present in some cases (Travers *et al.*, 1983). The analysis of these more complex situations is approachable using the tools we have developed.

We are grateful for support by grants from the System Development Foundation (to M.S.W.) and the NIH (to D.J.G., GM19036).

References

Arbanel, R. M., Wienecke, P. R., Mansfield, E., Jaffe, D. A. & Brutlag, D. L. (1984). *Nucl. Acids Res.* **12**, 263–280.

Aoyama, T., Takanami, M., Ohtsuka, E., Taniyama, Y., Marumoto, R., Sato, H. & Ikehara, M. (1983). *Nucl. Acids Res.* **11**, 5855–5868.

Arnott, S., Chandrasekaran, R. S., Hall, I. H., Purgjaner, L. C., Walker, J. K. & Ulang, M. (1983). *Cold Spring Harbor Symp. Quant. Biol.* **47**, 53–65.

Bossi, L. & Smith, D. M. (1984). *Cell,* **39**, 643–652.

Brosius, J., Dull, T. J., Slaeter, D. D. & Noller, H. F. (1981). *J. Mol. Biol.* **148**, 107–127.

de Crombrugghe, B., Busby, S. & Buc, H. (1984). *Science,* **224**, 831–838.

Dickerson, R. (1983). *J. Mol. Biol.* **166**, 419–441.

Falkner, F. G. & Zachau, H. G. (1984). *Nature (London),* **310**, 71–74.

Hawley, D. K. & McClure, W. R. (1983). *Nucl. Acids Res.* **11**, 2237–2255.

Kolb, A., Spassky, A., Chapon, C., Blazy, B. & Buc, H. (1983). *Nucl. Acids Res.* **11**, 7833–7852.

Mandecki, W. & Reznikoff, W. S. (1982). *Nucl. Acids Res.* **10**, 903–912.

Mattaj, I. W., Lienard, S., Jirincy, J. & de Robertis, E. M. (1985). *Nature (London),* **316**, 163–167.

McClure, W. R., Hawley, D. K., Youderian, P. & Susskind, M. M. (1983). *Cold Spring Harbor Symp. Quant. Biol.* **47**, 477–481.

Mulligan, M. E., Hawley, D. K., Entriken, R. & McClure, W. R. (1984). *Nucl. Acids Res.* **12**, 789–800.

Parzen, E. (1962). *Ann. Math. Statist.* **33**, 1065–1076.

Pribnow, D. (1975). *J. Mol. Biol.* **99**, 419–443.

Queen, C. M., Wegman, N. & Korn, L. T. (1982). *Nucl. Acids Res.* **10**, 449–456.

Reznikoff, W. S. & Abelson, J. N. (1978). In *The Operon* (Miller, J. H. & Reznikoff, W. S., eds), pp. 221–244, Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Rosenberg, M. & Court, D. (1979). *Annu. Rev. Genet.* **13**, 319–353.

Sadler, J. R., Waterman, M. S. & Smith, T. F. (1983). *Nucl. Acids Res.* **11**, 2221–2231.

Schaller, H., Gray, C. & Herman, K. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 737–741.

Selsing, E., Wells, R., Alden, C. & Arnott, S. (1979). *J. Biol. Chem.* **254**, 5417–5422.

Siebenlist, U., Simpson, R. B. & Gilbert, W. (1980). *Cell,* **20**, 269–281.

Staden, R. (1984). *Nucl. Acids Res.* **12**, 505–519.

Stephano, J. E. & Gralla, J. D. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 1069–1072.

Travers, A. A. (1984). *Nucl. Acids Res.* **12**, 2605–2618.

Travers, A. A., Lamond, A. I., Mace, H. A. F. & Berman, M. L. (1983). *Cell,* **35**, 265–273.

Waterman, M. S. & Whiteman, D. E. (1978). *Int. J. Math. Educ. Sci. Technol.* **9**, 127–137.

Waterman, M. S., Galas, D. & Arratia, R. (1984). *Bull. Math. Biol.* **46**, 515–527.

Wu, H. M. & Crothers, D. (1984). *Nature (London),* **308**, 509–513.

Youderian, P., Bouvier, S. & Susskind, M. M. (1982). *Cell,* **30**, 843–849.

Note added in proof. It has recently been reported that the heteronomous DNA model for poly(dA)·poly(dT) is inconsistent with solution nuclear magnetic resonance and Raman studies (Sarma *et al.* (1985) *J. Bimol. Struct. Dyn.* **2**, 1057–1084). While the precise conformation of these stretches of sequence is thus not at all clear, there is much evidence that it is an unusual one, or can take on an unusual conformation under certain conditions.

Edited by M. Gottesman