Volume 14 Number 1 1986 Nucleic Acids Research

Analysis of the occurrence of promoter-sites in DNA

Martin E.Mulligan+ and William R.McClure

Department of Biological Sciences, Carnegie-Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213, USA

Received 8 July 1985

ABSTRACT

We show that the occurrence and homology score (1) of promoter-sites in DNA depends upon the base composition of the DNA. We used simple probability theory to calculate the mean homology score expected for all promoter-sites that had a specific match in the canonical hexamers. By using the square root of this mean score as a measure of significance, we objectively classify all promoter-sites which are reported. We tested the theoretical approach in two ways. First, we used the program (PROMSEARCH)[1] to analyze ~ 150,000 base pairs of random sequence DNA with different base compositions and we found excellent agreement with the theoretical predictions. Our second test was the analysis of a number of sequences drawn from the GENBANK DNA sequence database. We have analyzed 20 bacterial and bacteriophage sequences, which consisted of at least one operon, for promoter-sites. We found no absolute preference for promoter-sites within noncoding regions. We show the results of analyzing the phages λ, T7 and fd, and the *E. coli lac* operon. The major known promoters in these sequences were all found correctly. We discuss the question of the location of a number of minor promoter-sites and show how PROMSEARCH can be used to help identify the correct location of the promoter. This approach can be applied to the search for any DNA site and should allow greater objectivity when comparing DNA sequences for meaningful subsequences.

INTRODUCTION

With the advent of rapid methods for sequencing DNA, a major goal that has arisen is the identification of control sequences, such as operators, activator binding sites and promoters, within a DNA sequence. We are especially interested in locating promoters in DNA sequences. The promoters of *Escherichia coli* are characterized by two regions of sequence homology that have been shown by genetic and biochemical criteria to be important for function (2, 3, 4). The first region (the -10 region) is located about 10 base pairs upstream from the transcription start-point. The second region (the -35 region) is located further upstream near position -35.

For simplicity, we will refer to the group of six highly conserved base pairs in these regions as the -10 hexamer (TATAAT) or the -35 hexamer (TTGACA). The distance between the two hexamers varies between 15 and 21 base pairs, with an optimal spacing of 17 base

[1]The PROMSEARCH program will be provided upon receipt of a self-addressed mailing label and a blank diskette.

pairs. We have shown previously (1) that a simple weighting algorithm can be used to evaluate promoters and that the resulting homology scores are related to an *in vitro* measure of promoter strength. Other authors (5, 6) have used similar weighting schemes to search DNA sequences for promoters although they did not attempt to relate DNA sequence to experimental data.

All of these approaches work reasonably well when searching sequences for strong promoters. However, the results become less clear if the promoter is of moderate strength or weaker. The difficulty is one of separating and assessing real promoters from the background that is observed. The simplest approach has been to use a cutoff or threshold value to achieve this separation. In our previous work, we suggested a cutoff score based on our evaluation of 112 well-defined promoters compiled by Hawley & McClure (4) and also a search of pBR322 for promoters. Staden (6) used the concept of a cutoff value and Harr *et al.* (5) set a significance level in evaluating promoter-sites. Still, these cutoff scores were necessarily subjective. A more objective way of determining levels of significance would clearly be preferable. Goad & Kanehisa (7) and Kanehisa (8) invoked a threshold when examining homologies between different DNA sequences. Recently, the statistical significance of comparing sequences has been addressed in a more quantitative manner (9, 10).

In this work, we have extended our previous methods to determine such an objective cutoff score. We base our approach on the observation that the distribution of possible promoters and their homology scores in any DNA sequence, which are found by a computer algorithm such as ours, is related to the base composition of the DNA. As a result, we have established criteria, which can be used to assign a relative significance to every promoter-site, and which are consistent for all sequences since they depend only on base composition. A similar approach has been used recently to assess fortuitous similarities when comparing two DNA sequences (9, 10).

TERMINOLOGY

In order to avoid confusion, we reserve the term "promoter" for those sites that have been characterized by biochemical and genetic criteria. We will use the term "promoter-site" or simply "site" to designate a DNA sequence that has a good degree of homology to the promoter consensus but that has not yet been proven to function as a promoter by biochemical and genetic criteria. In addition to the term promoter-site, we define the following terms. In searching for the consensus hexamer sequences, we will use the terms *specific match* and *minimum match.* A specific match refers to a match of r positions, and only r positions, out of six in the hexamer. A minimum match refers to a match of r positions and all matches better than r positions out of six in the hexamer. The *stringency* of a match descibes the value of r that is used. A match of five out of six is more stringent than a match of four out of six and so on. A promoter-site may also be described in terms of the specific and

minimum matches of its component hexamers, though now the number of specific matches that make up a minimum match is greater.

THEORY

In searching for promoter-sites, our computer algorithm first locates the positions of all -35 and -10 hexamers of a user-specified minimum match to the consensus hexamer. Following the general approach of von Hippel (11), the probability of finding a hexamer with a specific match r out of six is given by:

$$p_r = \sum_{i=1}^{k} p_i \qquad [1]$$

where p_i is the probability of finding each of the k combinations that are possible for the specific match, r. We sum k terms because the probability for each combination will depend independently on the base composition. When searching a DNA sequence for hexamers, however, our algorithm will not report sequences of a specific match, but rather, all sequences of a minimum match. Once the required match is found, the algorithm stops looking at that location; it advances to the next location and recommences the search. The theoretical probability of finding a hexamer of a given minimum match is given by

$$p'_r = \sum_{i=0}^{6-r} p_{(r+i)} \qquad [2]$$

where $p_{(r+i)}$ is the probability of finding the hexamer at a specific match $(r + i)$ out of six where r is the minimum match required.

If now we let p_r be the probability of finding a -35 hexamer at a specific match r and q_s be the probability of finding a -10 hexamer at a specific match, s, then the probability of finding a promoter-site with this combination of matches is given by:

$$\phi_{r,s} = p_r q_s \qquad [3]$$

Again, since the computer will report all promoter-sites of at least a match r and s in the two hexamers, the total probability is given by:

$$\phi'_{r,s} = \sum_{j=0}^{6-s} \sum_{i=0}^{6-r} p'_{(r+i)} \cdot q_{(s+j)} \qquad [4]$$

This probability is independent of the spacer length in the promoter-site. Since our algorithm allows seven spacer-lengths, the observed number of promoter-sites will be seven times that predicted by equation [4].

METHODS

Our method uses the computer program (TARGSEARCH), which we developed to search for and evaluate promoter-sites in a DNA sequence (1). For this work, we have implemented the program, written in PASCAL, on a DEC-20 mainframe computer. We have removed the ancillary search features in order to concentrate solely on promoter-site searches. We call the program PROMSEARCH to distinguish it from its predecessor. In addition, we have altered the program to handle DNA sequences in the GENBANK format (Release 13.0; October 1983). The 'SITES' specifications of the database also allows us to incorporate mutant information, at present only for single base pair changes. Other features in the 'SITES' section of the database can be used to annotate the search results with the locations of, for example, binding sites, known mRNA starts and conflicts in the DNA sequence. The GENBANK database used in this work was obtained from Bolt, Beranek, and Newman, Boston MA.

The theory of the previous section is the basis on which we set objective cutoff and significance limits. The central idea is that we can use the promoter-site probabilities in combination with the weighting table described by Mulligan *et al.* (1) to calculate the mean homology score for any composition and for any particular minimum match. The total homology score for any promoter-site is made up of three components, a weight for each of the two hexamer regions (-35 and -10) and a weight for the spacer between the -35 and -10 regions. The weight for each hexamer region can be further divided into a weight for the consensus hexamer itself and a weight for the extended regions around the hexamer.

The weight for the extended regions of each hexamer depends only on the base composition and can be calculated from the weighting table as:

$$W_{ext} = \sum_{i=1}^{n} \sum_{x=A}^{T} p_x \cdot w_{x,i} \qquad [5]$$

where p_x is the probability of finding a particular base, x, and $w_{x,i}$ is the weight assigned to that base at position i of the weight matrix. We sum these products for each base and for each position in the extended regions (The extended regions are different for the two hexamers; see Mulligan *et al.*, (1)). The weight for the consensus hexamer depends upon the specific match and upon the probability of finding each particular combination that makes up the specific match. The weight, w_i, for each combination is the sume of two terms: a score for the consensus bases in the combination plus a mean score for the nonconsensus bases in the remaining positions of the hexamer. The mean hexamer weight is given by:

$$\bar{W}_{hex} = \frac{\sum_{i=1}^{k} w_i \cdot p_i}{\sum_{i=1}^{k} p_i} \qquad [6]$$

Since all seven allowed spacer-lengths are equally probable, the mean spacer weight is simply the sum of all possible spacer weights divided by seven. For any specific match, the mean promoter-site weight is the sum of weights for the extended regions with the mean hexamer weights and the mean spacer weight. For any minimum match, the mean promoter-site weight is easily obtained from the mean promoter-site weight of the component specific matches and their probablities. Finally, the mean promoter-site weight is converted into a mean homology score.

The mean promoter-site weight is used to set the cutoff point for the homology scores that will be reported. We assume, for simplicity, that the standard deviation of promoter-site weights about the mean weight is given by the square root of the mean weight. We then set the cutoff at the 95% significance level or at a weight which represents the mean plus 1.645 times the standard deviation. We have extended this concept so that we can attach an approximate significance to each promoter-site that is reported by assigning each site to a class as follows: Class 1, above the 99% (2.33 times the deviation added to the mean) significance level; Class 2, above the 99.5% (3.09 times the deviation added to the mean) significance level; Class 3, above the 99.995% (3.89 times the deviation added to the mean) significance level. In all cases the cutoff and significance level weights are then converted into homology scores.

RESULTS

The dependence of hexamer and promoter probability on base composition is shown in Figure 1 for a minimum match of three out of six. As might be expected from the different consensus sequences of the hexamers, the -10 hexamer is more sensitive to the AT composition than is the -35 hexamer. Both have an equal probability of occurrence at 50% AT. Figure 1 also shows the theoretical curve for promoter-site occurrences with a minimum match of three out of six, and of a single spacer-length. At any minimum match, the mean homology score depends upon the base composition, as shown in Figure 2. Consequently, in addition to the tendency of AT-rich regions to have more promoter-sites (Figure 1), these promoter-sites will also tend to have higher scores. Clearly, the cutoff and significance levels will be higher if the AT content is higher. These graphs can be used to calculate the cutoff score and significance levels manually by following the procedure outlined in *Theory*. However, the cutoff and significance levels that are calculated within the PROMSEARCH program do not require that A and T be present in equimolar amounts (and similarly for G and C), on the DNA strand analyzed.

We have generated a number of files of randomly generated DNA sequence, ranging in size from 100 to 10,000 base pairs and in composition from 40% to 60% AT. The analysis of these random sequences confirmed the preceding theoretical ideas. In all cases, the random sequences had a base composition close to that desired. The mean score and the cutoff score depended only on the base composition and were close to that predicted. At a minimum match of three out of six, sequences of 40% AT had a mean observed cutoff score of 36.83

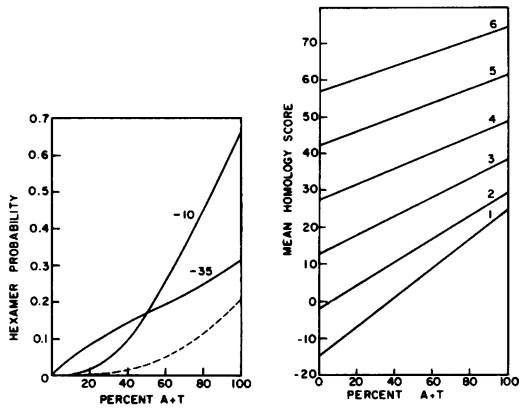

Figure 1. (left) The probability of finding hexamers and promoter-sites as a function of base composition. The probability per base pair of finding a -10 or a -35 hexamer at a match of at least three out of six is plotted *versus* % AT composition. The dashed line shows the probability of finding a promoter-site at a minimum match of three out of six. In searching DNA, the observed occurrence of promoter-sites will be seven times that shown here because of the 7 spacer-lengths that are allowed in defining a promoter-site. In this calculation, equimolar amounts of A and T were used.

Figure 2. (right) Dependence of mean homology score on base composition. The calculated mean homology scores are plotted against base composition (%A+T). Each line represents a minimum match, which is the same for both -35 and -10 regions. The match is indicated next to each line. The mean homology score at 50% AT is 5 (at a minimum match of 1), 14(2), 25.6(3), 38.3(4), 51.9(5) and 65.9(6).

(predicted to be 36.81), at 50% it was 39.35 (39.57) and at 60% AT it was 42.22 (42.32). When the random DNA sequences of length 10,000 base pairs and composition 50% AT were analyzed for each specific match of promoter-site, we found that the mean number of sites found, at a particular homology score, followed a normal distribution. This was true for all matches that we were able to analyze. The overall distribution of promoter-sites for a minimum match of three out of six in both the -35 and -10 regions is shown in Figure 3. The close fit of the data to the theoretical curve reflects the close fit of the data for each specific match to its normal curve. The cutoff point (marked **a** in the Figure) is calculated based on a normal distribution about the overall mean score, as are the other significance level boundaries **b**, **c** and **d**. By setting this cutoff point, we calculate that we will find 73.5% of all the promoter-sites of every possible match that would be expected to score above the cutoff. Using a minimum match of 2 out of 6, we calculate that we would find 97% of the sites. The corresponding percentages at other base compositions are similar.

In the remainder of this section, we present the results of the analysis of a number of DNA sequences. We discuss the analysis of the GENBANK database in general and we give four representative examples from it. We also discuss the presence and the functional

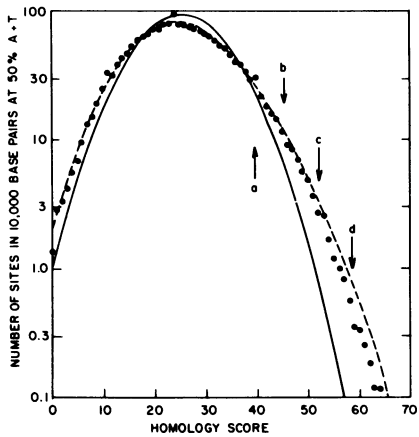

Figure 3. Distribution of promoter-site homology scores. The distribution of promoter-site homology scores found in DNA of random sequence and 50% A+T is shown. Ten sequences of 10,000 base pairs were searched at a minimum match of three out of six. The mean number of sites found at each particular homology score in all 10 sequences, which were analyzed in both directions, was calculated. A rough data-smoothing procedure was used as follows: the value at each homology score was added to the values for the homology scores above and below. The mean of the sum is plotted. The distribution of sites follows the theoretical curve (dashed line), which is the sum of the 16 component normal curves. The solid line represents a normal curve drawn at the overall mean homology score that was calculated for these sites. **a** is the cutoff score (39.6) that is calculated from the mean homology score and its associated normal curve. Class 1 sites fall between the scores marked **b** (45.4) and **c** (51.9); Class 2 sites between **c** and **d** (58.7); Class 3 sites have scores greater than **d**.

significance of promoter-sites that are found in addition to known promoters. Finally, we show how the probability of finding promoters through mutation can be calculated and how promoters, which are created through DNA rearrangement can be assessed.

A. Searching the GENBANK database. We have used the methods outlined above to analyze seven complete genome sequences from the phage subsection of the database and 13 *E. coli* operon sequences, all of which were longer than 1500 base pairs. We can ask two questions in searching these sequences for promoter-sites. Where are the high-scoring sites? And, are these sites significant or could they be due simply to a favorable base composition and hence count as false positives? The answer to the first question is found in a straightforward search of the DNA sequence (discussed below). The results of our analysis of random sequences of different lengths and compositions allows us to answer the second question. We have seen that the probability of finding a site depends upon the base composition. Hence, we can calculate how many sites we would expect to observe. We can also calculate how many sites of each significance class we would expect. We have performed these calculations for the 20 sequences mentioned above. We subdivided each sequence into three categories, namely, coding regions, noncoding regions in the forward direction and

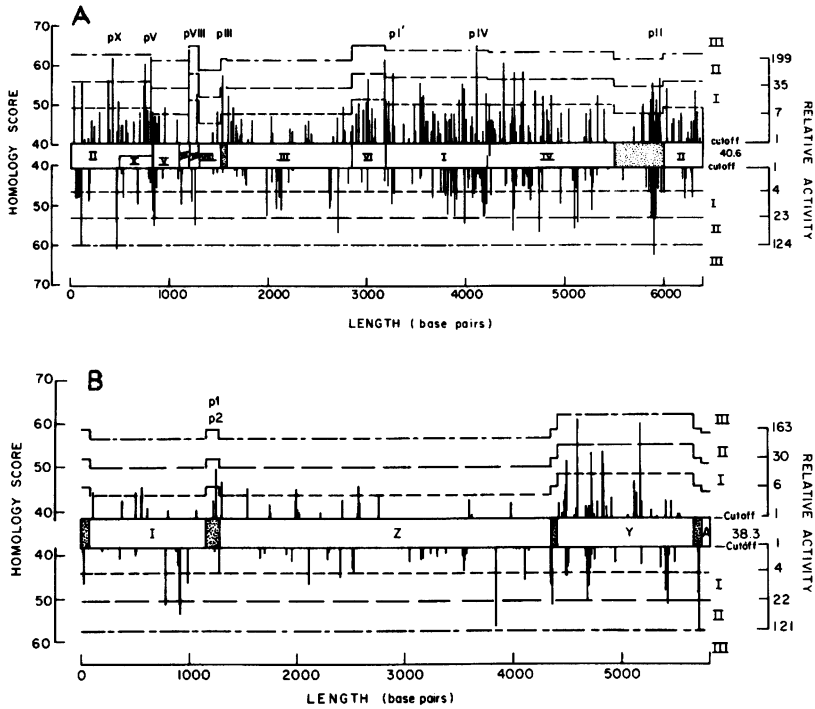

Figure 4. The distribution of promoter-sites in bacteriophage fd and in the *E. coli lac* operon. A: Promoter-sites in bacteriophage fd. **B:** Promoter-sites in *lac*. Only those sites which have homology scores greater than the cutoff score are shown. The cutoff score was 40.6% for fd and for *lac* it was 38.3%. The gene boundaries are indicated. The ordinate on the left indicates the homology score scale. One vertical line is drawn for each promoter-site as defined in the text except in cases where two or more sites overlap, in which case only the highest scoring site is shown. Sites above the genome are for the forward direction; those below are for the reverse direction. The known promoters that we found are marked (*e.g.* pIV for the promoter in front of gene IV in fd). The significance levels are drawn for each coding region and for the noncoding regions. The ordinate on the right indicates relative activity based on the correlation of homology score with activity (1). The values are marked to correspond to the significance levels of the noncoding regions in both cases. This representation is analogous to that used by Staden (6).

noncoding regions in the reverse direction. For the purpose of positioning a promoter-site uniquely, we define the location of a promoter-site to be a point 8 base pairs downstream from the 3' end of the -10 region hexamer. This position corresponds approximately with the start-point of transcription for known promoters. The results of these calculations are described in detail by Mulligan (12), but the main result was that all sequences had the expected number of promoter-sites of all classes in all categories. There was no tendency for coding regions to have fewer sites than expected.

The distribution of promoter-sites on bacteriophage fd is shown in Fig. 4a. At first

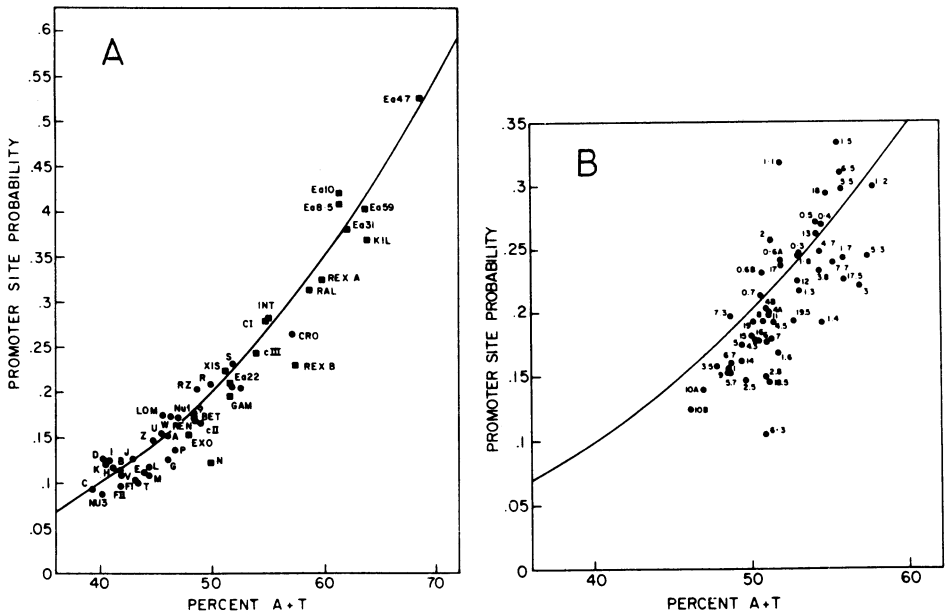

Figure 5. Promoter-sites in coding regions of λ and T7. A: The number of promoter-sites, with a minimum match of three out of six in both hexamers, found in each of the 49 known coding regions of phage λ are plotted as a function of their base composition (%AT). The solid line is the theoretical curve, calculated as described in the text, for promoter-sites with a minimum match of three out of six in both -35 and -10 hexamers. Coding sequences in the forward direction (rightward) are indicated (•), those in the reverse direction are indicated (▪). 28 coding regions fall below the line, while 21 are above it. **B:** The number of promoter-sites found in each of the 53 known coding regions of bacteriophage T7 are plotted as a function of their base composition. The solid line is the same as that in panel **A.**

sight, this representation suggests a plethora of promoters in fd. However, there is a logarithmic relationship between activity and homology score (1), and since most of the sites fall between the cutoff and the first (95%) significance level, they are probably functionally not relevant. The number of sites found in fd is consistent with its AT-rich base composition. The two class 3 promoters are the fd VIII and fd IV promoters (13). Of the remaining nine promoters listed by Schaller *et al.*, (13) we find five successfully. Of all the class 2 and class 3 sites shown in the figure, seven are known promoters while the remaining six are not.

The distribution of promoter-sites in the *lac* operon is shown in Fig. 4b. In contrast to the fd sequence there appears to be a dearth of promoters in the operon. Again, however, this distribution is just what is expected from the lower AT content of the *lac* operon (47% AT). The usefulness of setting the significance levels for each coding region can be seen by comparing the Y gene levels with those for the Z gene.

We have analyzed the DNA sequence of bacteriophage λ for possible promoter-sites. The

Table 1. The highest scoring promoter-sites in bacteriophage λ.

Location	-35	spacer	-10	Score	Class	Name or Region
		Forward				
23764	ATAACATTATGTTTTT	17	TATCCTATAATCTG	75.6	3	*Ea47*
44588	GGCATGATATTGACTT	17	TTGGGTAAATTTGA	68.6	3	PR'
23267	ATTATTTTATTGTCAT	18	TAAATGACAATTTG	62.6	3	*Ea47*
25277	GGTAAATAACTGACCT	17	TATTCTATATTGTT	62.0	3	*Ea31*
27744	TTATATCATTTTACGT	18	TTTTTTATACTAAG	62.0	3	*att*
36444	AGATGGCCTTTTTCTG	17	TCTGTTAAAATATC	59.1	2	*rexA*
38024	CCGTGCGTGTTGACTA	17	GCGGTGATAATGGT	58.6	2	PR
		Reverse				
24581	CAGAATTTATTGAAGC	18	TATTATAGATTTGA	62.0	2	*Ea31*
23703	ATATGACATTTGGTAT	17	TATTGAAAAATGGA	60.9	2	*Ea47*
26690	AAATGTTTTTTTCCTT	17	TTGACTACTATGTT	60.4	2	*Ea59*
23393	AATTATTACATGCCTT	16	TATGGCAGAATGTA	59.1	2	*Ea47*
22970	ATATATTTTTTGGCGT	18	AGAGCCAAAATAAC	58.6	2	*Ea47*
38675	CTGCCGAAGTTGAGTA	17	TTTGTCATAATGAC	58.6	2	Po
35581	CTGGCGGTGTTGACAT	17	GCGGTCATACTGAG	58.0	2	PL

Only promoter sites with spacers of 16, 17 or 18 base pairs are listed. The location of a site is considered to be 8 base pairs downstream from the end of the conserved -10 region hexamer. The significance class of each site is indicated. These depended on the region of the genome in which the site occurred and were assigned as explained in the text. The four major non-activated promoters of λ are named. The region in which the other promoter-sites are found is listed. The region in which a promoter-site is listed is not strand-specific. The *Ea47*, *Ea31* and *Ea59* genes are transcribed from right to left (*i.e.* in the reverse direction) on the λ genome.

observed promoter-site probability is shown in Figure 5a for all the genes of λ as a function of their AT composition. Although the theoretical curve assumes equimolar amounts of A and T, it can be seen that most of the coding regions are positioned quite close to the curve with no apparent selection against the occurrence of promoter-sites in coding regions. The distribution of promoter-sites in λ does follow the AT content of the different regions of λ. Thus, there are more sites found in the *b* region (an AT-rich region) than elsewhere in the genome.

Table 1 lists the major promoter-sites that we find in λ. Of the 14 highest scoring sites

in λ, only 4 are known to be promoters. As might be expected, the weak positively-controlled promoters of λ (PRM, PE and PI) cannot be distinguished from the background sites. Of the rest, all except one occur in the *b* region. This region, although not necessary for phage development, is known to be transcriptionally active and to contain RNA polymerase binding sites (14, 15, 16). We do not know if any of our sites correspond to these latter sites. For example, we do not know if the high-scoring rightward promoter site at 23764 has any physiological significance. Such a high-scoring site is unlikely to occur in λ on a random basis but it has a poor match to the -35 hexamer. Rosenvold *et al.* (17) identified a transcript, Pb_L, in the *b* region, which has been positioned to start at 23231 in the leftward direction (18). This position does not score well in our analysis, only 42.5%. There are also two other sites, one on either side of Pb_L, which score 55.0% (at 23221) and 50.9% (at 23234). These sites may be responsible for some of the other transcripts in this region (17). We also find a discrepancy between our results and previous ones concerning the location of P_{lit}, the minor leftward promoter in the *rex* region. There is some disagreement concerning the position of this promoter. A strong RNA polymerase binding site was first identified in the *rex* region by Pirotta *et al.* (19). The position of this site was fixed at 36271 in the λ sequence (20, 21) However, this site does not score above our cutoff value of 39.5%. Recently, P_{lit} has been positioned at 36322 (18). This site scores 41.9%. We find a site at 36307 with a score of 51.4%. We also find a site at 36459, which scores 56.2%. A transcript from the latter site would be 655 bases long, which is consistent with the early estimates of the length of this transcript (~600 bases) (22).

The distribution of promoter-sites in T7 is somewhat different from that in λ. Figure 5b shows how the ratio of observed sites to calculated sites varies with the AT composition for each of the coding regions of T7. The distribution of points about the theoretical line is clearly assymetric for T7. Seven of the ten early genes fall above the line. Of the 43 middle and late genes, 34 fall below the line. There appears to be a definite tendency, especially for the later coding regions of T7 to be deficient in promoter-sites. There are not as many very high-scoring promoter-sites in T7 as in λ and apart from the three major early T7 promoters, there are more promoter-sites on the leftward or non-transcribed strand. Table 2 lists the highest-scoring sites which we found in T7. Of 13 promoter-sites, six are the well-known T7 promoters, A1, A2, A3, B, C and D. We do not know if any of the other sites have any physiological relevance. The promoter-site at 5703 was also identified by Dunn & Studier (23). We do not find any of the other sites proposed by them. However, the sites in the forward direction are consistent with other evidence of RNA polymerase binding and transcription sites (24, 25, 26). The E promoter has been located at position 36928 by Dunn & Studier (23). That promoter-site does not appear above our cutoff score (40.1%). However, we do find a site at 36835 (score = 55.0), which is between genes 18 and 18.5, that is consistent with earlier evidence for the location of this promoter (27, 28). This site has recently been shown to be the correct location of the T7E promoter (29)

Table 2. The highest scoring promoter-sites in bacteriophage T7.

Location	-35	spacer	-10	Score	Class	Name or Region
		Forward				
498	AAAAGAGTATTGACTT	17	TATAGGATACTTAC	73.9	3	A1
626	AAACAGGTATTGACAA	18	TGCAGTAAGATACA	73.3	3	A2
750	ACAAAACGGTTGACAA	17	CACGGTACGATGTA	72.8	3	A3
5703	AAACTATGGTTGACAC	16	TGTGATGTACTGGC	64.4	3	Gene 1
1514	GATTATCACTTTACTT	17	ATGTATATGCTTAC	59.1	2	B
18616	CTTTAAGATTTAACTC	17	TTTATTATGTTAAC	59.1	2	Gene 6.5
414	CTTTAATCATTGTCTT	17	CTCACTATAAGGAG	58.6	2	Left arm
3114	ATAAGCAACTTGACGC	17	GCTGATAGTCTTAT	58.6	2	C
		Reverse				
24630	AATCGGTTGTTGAACT	17	TTGGTGACAATCCA	66.2	3	Gene 11
223	AGATAGGCGTTGACTT	17	AGGTGTAGGCTTTA	63.8	3	D
32980	ATGTCACCATTGACAC	17	GCTGGCATGATGCG	63.5	3	Gene 16
39581	CCTTAGGACTTGACTC	17	AGTGGTGTGATGCA	61.4	2	Right arm
16527	ATAGGGTAATAGACAG	16	TTTGGTAAATTTGT	60.4	2	Gene 5.3

See the legend to Table 1 for an explanation of the features of this Table. The known *E. coli* RNA polymerase promoters are indicated.

B. Promoter-sites in addition to known promoters. In the foregoing sections we have reported that the number of promoter-sites found in DNA can be predicted from the base composition, and that the major promoters are correctly identified. However the predicted presence of promoter-sites in addition to those that are known to function as promoters demands further examination. Three non-exclusive explanations could account for these extra sites. First, these predicted promoter-sites might actually function as promoters, but this function may ordinarily be masked by the presence of a stronger promoter nearby. For example, the *trp* operon is expressed from a strong promoter (H.S. = 61.5) under derepressing conditions. However, comparison of the ratio of enzymes under derepressing and repressing conditions suggested the presence of a low-level internal constitutive promoter (P2) with 3% of the activity of the strong derepressed promoter (30). The internal promoter has been located (31) and it had an homology score of 53.8 (1). These scores are consistent with the observed promoter strength allowing for differences in translational efficiency and the error in calculating homology score. This was the likeliest candidate that was found by PROMSEARCH in the region known to contain P2. A second, similar example, is the internal promoter (P2)

of the *his* operon. This promoter has been mapped genetically (32) and located by DNA sequencing (33). Under derepressing conditions P2 showed 15% the activity of P1. The homology score for P1 and P2 were 61.5 and 47.9 respectively; these scores are consistent with the observed expression, again allowing for difference in translational efficiency and the error in calculating homology score. However, the available evidence on the expression of the *lac* operon indicates that there are no significant internal promoters in the *z* or *y* genes. The ratio of β-galactosidase to transacetylase enzyme activity remains constant under repressed and induced conditions (34). We do not find any promoter-sites in the *z* gene that score above 50%, but there are six such sites in the *y* gene. If any of these sites promote transcription we conclude that they do not result in expression of transacetylase. A different example of possible roles for extra promoter-sites concerns the many sites found on the opposite (noncoding) DNA strand (which are also just as expected). Antisense RNA may play a regulatory role in the expression of some operons and some of the sites on the opposite strand may act as promoters for anti-sense RNA. Some of the extra promoter-sites on the noncoding strand may represent unknown but functionally important promoters for antisense RNA transcription. For example, the small antisense RNA promoter (Psar) in the bacteriophage P22 *imm1* region was located initially with the PROMSEARCH program (Liao, S.-M. and McClure, in preparation).

A second explanation for extra promoter-sites is that they may function as promoters but that any RNA that is transcribed may be efficiently terminated through the action of the termination factor ρ and be degraded without having any significant function. Molecules of RNA polymerase involved in the synthesis of such RNA would behave as if they were bound to nonspecific sites. There is no data about the amount of transcription from nonspecific DNA *in vivo* at present.

The third explanation for the detection of extra promoter-sites in DNA may simply be due to deficiencies in homology score evaluation. For example, the weighting scheme used by Staden (6), gave virtually the same correlation as we have reported earlier (1) suggesting that it is the lack of experimental evidence that is limiting at present. If so, then we expect that future experimental evidence will provide better weights for the evaluation procedure corresponding to the contributions of DNA sequence to promoter function. When the new weighting schemes are formulated, the same theory, random sequence tests and search algorithms that we have used here can be used to evaluate these new models.

C. Promoters created by DNA rearrangements. The number of mutations that result in the creation of new promoters, whether internal or otherwise is rather small. Hawley and McClure (4) documented only four such examples (λ*cin*, *bio*P98, *lac*P115, λc17). The average homology score of these sites before mutation was 49.5; the single base pair changes increased the average score to 60.8. The probability of any site being mutated by a single base pair change into a site that is stronger by 10 homology score points is very small and can be

calculated using the principles documented here. We consider only sites that are Class 1 or better so that a single base pair mutation could create a promoter with a moderate homology score. We consider only the most conserved positions in the -35 or -10 hexamer since only mutations in these positions will result in a 10 point change in homology score. Then for 50% AT and a minimum match of three out of six, we calculate that the probability of finding a site that can be mutated into a promoter is 0.0041 per base. For a mutation frequency of 10^{-8}, then this probability corresponds to a frequency of creating promoters of 4×10^{-11} per base pair. For example, in the *lac* region, (ECOLAC) ~ 5000 bp, we calculate that there should be 22 sites where such a mutation could be observed and the range should be 14 to 29 based on the analysis of random sequence DNA. By inspection of the PROMSEARCH output of the ECOLAC sequence, we found 18 examples in one direction and 21 in the other.

A second method of creating promoters is through *in vivo* or *in vitro* fusions. Hawley and McClure listed two such examples: λL57 was formed *in vivo* in the construction of a λ transducing phage; IS2 I-II was the result of an insertion of IS2 that created a promoter. We cannot predict the probability of finding these promoters but we can examine the DNA sequence in the vicinity of such fusions for the presence of promoter-sites. Fortuitous promoters can then be properly handled in experimental situations. We discuss some examples below.

DISCUSSION

The probability of finding a promoter-site in DNA using an algorithm such as ours depends only on the composition of the DNA sequence under consideration and the stringency that is imposed on the search. We have shown that the PROMSEARCH program correctly finds the number and distribution of promoter-sites within random DNA sequences. The theory and the random sequence trials form the firm basis upon which we have evaluated the pattern of of promoter-sites on a broader level within bacterial DNA sequences. We set cutoff and significance levels based on the overall distribution of promoter-site scores. The choice of a particular cutoff is still necessarily somewhat arbitrary; we have chosen the 95% significance level assuming that the distribution of homology scores is a normal one. Any other level could be selected for the cutoff and for the significance levels. However, by choosing levels that depend upon the distribution, we can be consistent in our screening of all sequences.

As currently implemented, PROMSEARCH will use the base composition of the sequence being analyzed in setting cutoff scores and significance levels and classify all promoter-sites accordingly. If, however, information is provided about the extent of coding and noncoding regions within a sequence (on the 'SITES' field of the database), then promoter-sites will be classified according to the base compositions of the individual coding and non-coding regions. The effect of analyzing promoter-sites in this way can be seen in Figure 4. However, it is arguable that the cutoff score and significance levels should be uniformly set at the scores

corresponding to 50% AT since there is only one enzyme in *E. coli* responsible for transcription of the promoters that form the basis of our analysis and weighting table. The global base composition is probably most important for evaluating function. However, the local base composition is probably more useful for evaluating the statistical significance of promoter-sites and for deciding which sites warrant further examination for evidence of functional activity. It must be emphasized that our use of base-composition normalization bears no implications for the mechanism of promoter recognition by RNA polymerase. As long as consensus sequences form the basis for the search for recognition sequences in DNA, there will always be a finite probability of finding a recognition sequence, which depends only on the base composition of the DNA. Since the consensus sequence for *E. coli* promoters is not an absolute one but allows for partial homologies, the probability of finding such sequences is greatly increased. The purpose of this paper is to show how to account for that probability in looking for promoters.

At present, 50% of the promoter-sites reported by PROMSEARCH are well documented promoters. The results shown for λ, T7, *lac* and fd are representative of the promoter-site localization and evaluation found for the other *E. coli* sequences examined. Either the extra sites that are found function as promoters or they do not (*i.e.* they are false positives). We have presented a number of possible functions for these sites in *Results*. We currently believe that some, but not most, of the extra promoter-sites will turn out to be promoters. The rest will be false positives. The presence of false positives in our results is not due to our use of base composition to assign significance to each site, but is due to the deficiencies in the weighting table that assigns the homology score to each site. We hope that future results will provide a greater understanding of the relationship between promoter sequence and promoter strength and lead to better weighting tables that will eliminate the problem of false positives. As the correlation is improved, we will be more certain of the predictive ability of the program. Conversely, if the correlation worsens then the assumptions that lie behind the weighting table will need to be revised. If PROMSEARCH consistently failed to detect certain promoters (i.e. false negatives) then the inclusion of *in vitro* data would worsen +the correlation. At present, we do not know of any examples of a non-activated promoter that is not found by PROMSEARCH.

An important conclusion of our results is that great care is required whenever a functional role is assigned on the basis of DNA sequence analysis. DNA sequence homology may exist that is not always readily apparent but which is so when quantified sytematically. For example, the early conclusions (36) that the -35 region was dispensible as long as a good -10 region is present can now be reevaluated. These workers found that a variety of bacteriophage fd fragments ligated upstream of the fd VIII promoter -10 region (TATAAT) were active in binding RNA polymerase. In drawing this conclusion, they did not detect the -35 region homologies that were present as a result of the ligation. We have evaluated these sequences and find that they have good homology scores in the range 50-70, and the binding

ability is both understandable and expected. Recently, Gragerov *et al.*, (35) have found that *tet* gene expression was detected when replacement -35 regions were cloned into a plasmid that contained the *tet* gene -10 region but which had been inactivated by removal of the *tet* gene -35 region. The functional importance of the replacement -35 region was confirmed by the isolation of a mutation in the region. By contrast, DNA homologies, especially those of shorter sequences at low stringency, should not be overemphasized until some evidence of function is obtained. For example, the presence of sequences that partially resemble -10 or -35 hexamers in the vicinity of selected promoters (37) cannot be construed as secondary RNA polymerase binding sites until the random probability of finding them has been taken into account and unless there is firm biochemical or genetic evidence for such a role.

An additional rule for assessing promoter-sites can be found from the compilation of Hawley and McClure (4). Although the -35 and -10 regions are conserved hexamers, nevertheless there are three positions within each hexamer that are most highly conserved. In the 112 compiled promoters, all have the consensus base in at least two of the three positions of the -10 region (TA---T), and all have the consensus base in at least one of the three positions of the -35 region (TTG---). We suggest that, based on current evidence, promoter-sites that do not obey this rule be excluded from consideration as likely promoters. We conclude that promoters contain contributions from both the -35 and the -10 regions and that both regions are required for function. The evidence at present strongly suggests that a -35 region or a -10 region alone is insufficient for promoter activity.

Although we have combined this analysis to extend our earlier program, the significance levels of any promoter-site of particular interest may be determined manually using the data of Figure 2. From the mean homology score, the significance levels are easily calculated, and any site can be classified. The method for calculating significance levels manually is as follows. First, the base composition must be defined. Either the base composition of the sequence being searched or a defined base composition can be used (*e.g.* 50% A+T for *E. coli*). Then, using the data of Figure 2, and for a particular minimum match, the mean homology score can be computed. This score must be converted back to a weight (multiply by 1.69 and add 163). The square root of this weight is the standard deviation and this value is converted into an homology score (divide by 1.69). Cutoff values and significance levels can then be set by reference to tables for the normal distribution or as desired. The values that we have used are listed in *Methods*.

Our method of locating and evaluating promoter-sites is suitable, in principle for any target sequence in DNA, especially those in which partial homologies can have significant functional importance. The analysis of promoter-sites is assisted greatly by two independent correlates of DNA sequence with function: i) based on mutational evidence, the consensus sequence is likely to represent maximal function (4); ii) *in vitro* selectivity correlates with DNA sequence (1). The calculation of a meaningful homology score relies on both of these

important conclusions. The information available for most other DNA sites (eg. operators) is currently insufficient to allow comparable analysis.

ACKNOWLEDGEMENTS

We thank Kathryn Galligan for her assistance in preparing the manuscript and M. Roederer and J. Karohl for helpful discussions. Our research on RNA polymerase is supported by the N.I.H. (GM 30375). Development of the PROMSEARCH program was supported, in part by the N.S.F. (PCM 8140433).

+Present address: Department of Molecular Genetics and Cell Biology, University of Chicago, 920 E. 58th St., Chicago, IL 60637, USA

REFERENCES

1. Mulligan, M.E., Hawley, D.K., Entriken, R. & McClure, W.R. (1984) *Nucleic Acids Res.* **12**, 789-800.
2. Rosenberg, M. & Court, D. (1979) *Ann. Rev. Genet.* **13**, 319-353.
3. Siebenlist, U., Simpson, R. & Gilbert, W. (1980) *Cell* **20**, 269-281.
4. Hawley, D.K. & McClure, W.R. (1983) *Nucleic Acids Res.* **11**, 2237-2255.
5. Harr, R., Häggström, M. & Gustaffson, P. (1983) *Nucleic Acids Res.* **11**, 2943-2957.
6. Staden, R. (1984) *Nucleic Acids Res.* **12**, 505-519.
7. Goad, W.B. & Kanehisa, M.I. (1982) *Nucleic Acids Res.* **10**, 247-263.
8. Kanehisa, M. (1984) *Nucleic Acids Res.* **12**, 203-213.
9. Reich, J.G., Drabsch, H. & Däumler, A. (1984) *Nucleic Acids Res.* **12**, 5529-5543.
10. Smith, T.F., Waterman, M.S., & Burks, C. (1985) *Nucleic Acids Res.* **13**, 645-656.
11. von Hippel, P.H. (1979) in *Biological Regulation and Development*, vol. 1, pp. 279-347, Goldberger, R.F. (ed.) Plenum Press, NY.
12. Mulligan, M.E. (1984) *Ph.D. Thesis*, Harvard University, Cambridge, Mass., USA.
13. Schaller, H., Beck, E. & Takanami, M. (1978) in *Single-Stranded DNA Phages*, Denhardt, D.T., Dressler, D. and Ray, D.S. (eds.) pp 139-163, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
14. Botchan, P. (1976) *J. Mol. Biol.* **105**, 161-176.
15. Vollenweider, H.J. & Szybalski, W. (1978) *J. Mol. Biol.* **123**, 485-498.
16. Kravchenko, V.V., Vassilanko, S.K. & Grachev, M.A. (1979) *Gene* **7**, 181-195.
17. Rosenvold, E.C., Calva, E., Burgess, R.R. & Szybalski, W. (1980) *Virology* **107**, 476-487.
18. Daniels, D.L., Schroeder, J.L., Szybalski, W., Sanger, F. and Blattner, F.R. (1983) in *Lambda II*, pp. 469-676, Hendrix, R.W., Roberts, J.W., Stahl, F.W. & Weisberg, R.A. (eds.) Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
19. Pirotta, V., Ineichen, K. & Walz, A. (1980) *Mol. Gen. Genet.* **180**, 369-376.
20. Landsmann, J., Kröger, M. & Hobom, G. (1982) *Gene* **20**, 11-24.
21. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. & Petersen, G.B. (1982) *J. Mol. Biol.* **162**, 729-773.
22. Hayes, S. & Szybalski, W. (1973) *Mol. Gen. Genet.* **126**, 275-290.
23. Dunn, J.J. & Studier, F.W. (1983) *J. Mol. Biol.* **166**, 477-535.
24. Minkley, E.G. & Pribnow, D. (1973) *J. Mol. Biol.* **77**, 255-277.
25. Studier, F.W. & Rosenberg, A.H. (1981) *J. Mol. Biol.* **153**, 503-525.
26. Hsieh, T.-S. & Wang, J.C. (1976) *Biochemistry* **15**, 5776-5783.
27. Delius, H., Westphal, H. & Axelrod, A. (1973) *J. Mol. Biol.* **74**, 677-687.
28. Stahl, S.J. & Chamberlin, M.J. (1977) *J. Mol. Biol.* **77**, 577-601.

29. Prosen, D.E. & Cech, C.L. (1985) *Biochemistry* **24**, 2219–2227.
30. Jackson, E.N. & Yanofsky, C. (1972) *J. Mol. Biol.* **69**, 307–313.
31. Horowitz, H. & Platt, T. (1982) *J. Mol. Biol.* **156**, 257–267.
32. Schmid, M.B. & Roth, J.R. (1983) *J. Bacteriol.* **153**, 1114–1119.
33. Grisola, V., Riccio, A. & Bruni, C.B. (1983) *J. Bacteriol.* **155**, 1288–1296.
34. Hopkins, J.D. (1974) *J. Mol. Biol.* **87**, 715–724.
35. Gragerov, A.I., Smirnov, O.Y., Mekhedov, S.L., Nikiforov, V.G., Chuvpilo, V.G. and Korobko, V.G. (1984) *FEBS Letters*, **172**, 64–66.
36. Okamoto, T., Sugimoto, K., Sugisaki, H. & Takanami, M. (1977) *Nucleic Acids Res.* **4**, 2213–2272.
37. Travers, A.A. (1984) *Nucleic Acids Res.* **12**, 2605–2618.
38. Kalnins, A., Otto, K., Rüther, U. & Müller-Hill, B. (1983) *EMBO J.* **2**, 593–597.