Volume 12 Number 1 1984 Nucleic Acids Research

Escherichia coli promoter sequences predict *in vitro* RNA polymerase selectivity

Martin E.Mulligan, Diane K.Hawley, Robert Entriken and William R.McClure

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

Received 22 August 1983

ABSTRACT

We describe a simple algorithm for computing a homology score for *Escherichia coli* promoters based on DNA sequence alone. The homology score was related to 31 values, measured *in vitro*, of RNA polymerase selectivity, which we define as the product $K_B k_2$, the apparent second order rate constant for open complex formation. We found that promoter strength could be predicted to within a factor of ±4.1 in $K_B k_2$ over a range of 10^4 in the same parameter. The quantitative evaluation was linked to an automated (Apple II) procedure for searching and evaluating possible promoters in DNA sequence files.

INTRODUCTION

The DNA sequences of many promoters for *Escherichia coli* RNA polymerase are now known. Hawley and McClure (1) have compiled a list of 112 promoters defined by biochemical and genetic evidence and have suggested a consensus sequence on the basis of observed homologies. This consensus sequence is similar to those proposed previously by Rosenberg and Court (2) and Siebenlist *et al.* (3) on the basis of compilations of fewer promoter sequences. The notion that promoter function is related to the agreement of the promoter sequence with the consensus sequence is supported by genetic and biochemical evidence. However, such evidence has been limited to measurements of the changes in promoter function resulting from one or two base pair changes.

Our approach is to use a computer program to locate potential promoter sites in a given DNA sequence and to evaluate these sites according to a simple rule based on the number of occurrences of different bases in different positions in the 112 promoters listed by Hawley and McClure (1). We have investigated the relationship between overall promoter sequence and promoter strength for those promoters for which reliable measurements of promoter function have been made *in vitro*. This analysis reveals a direct relationship between promoter function and the extent of agreement of the promoter sequence with the consensus.

THE PROGRAM

Our program (called TARGSEARCH), similar in design to other search programs, is written in Pascal and implemented on an Apple II or Apple II+ microcomputer.

The search for a possible promoter site within a DNA sequence occurs in two steps.

```
SEQUENCE FILE NAME <TERM>? T7A1

READING SEQUENCE FILE T7A1 FROM BLOCK 8 TO BLOCK 9:..
FILE READ, TRANSCRIBING CHARACTERS TO BASES.
..60 BASES READ

HARDCOPY <N>? Y

TYPE IN THE TARGET STRING <HELP>: P

TYPE A NUMBER FROM 1 TO 6 FOR THE PERCENTAGE OF
A TARGET TO BE CALLED A HIT, FOR EACH TARGET:
-10<4>? 3
-35<4>? 3

DO YOU WANT A LISTING OF THE LOCATIONS WHERE TARGETS ARE FOUND <Y>? Y

DEFAULT WEIGHT MAPS ARE FINAL OCCURRENCE MAPS
MAX = 332; BASELINE = 163
USE DEFAULT WEIGHTING MAPS <Y>? Y
```

```
                SEARCH OF SEQUENCE T7A1 60 BASES LONG FOR PROMOTERS

SEARCH FOR PROMOTER BLOCK TATAAT 3 OUT OF 6 RETURNS:
   8   10   11   13   15   19   20   21   26   28   33   35   38   44
SEARCH FOR PROMOTER BLOCK ATTATA 3 OUT OF 6 RETURNS:
   6    9   11   14   18   19   24   29   31   34   36   39   42
SEARCH FOR PROMOTER BLOCK TTGACA 3 OUT OF 6 RETURNS:
  15   19   20   27   36   42   46   52
SEARCH FOR PROMOTER BLOCK TGTCAA 3 OUT OF 6 RETURNS:
   3   13   16   18   19   24   26   43
```

```
DO YOU WANT A LISTING OF THE SEQUENCE WITH MARKERS <N>? Y
```

```
                        LIST OF T7A1 60 BASES LONG

      : : : : :   :: :    :*:     : :* :    * : *      *
TATCAAAAAG AGTATTGACT TAAAGTCTAA CCTATAGGAT ACTTACAGCC ATCGAGAGGG
  *  :  :  : *: * ::     : *  :  :  : :  :   :*
```

```
TYPE IN THE MINIMUM WEIGHT TO BE REPORTED <0>: 0
```

```
                          POSSIBLE PROMOTERS IN T7A1
                          FOR THE FORWARD DIRECTION

M35=AAAAGAGTATTGACTT WT= 157  SPACE(17)=14  M10=TATAGGATACTTAC WT= 117
-35 AT 15 & -10 AT 38 WEIGHT=  125 OR ----------> 740%

M35=GAGTATTGACTTAAAG WT= 107  SPACE(19)= 1  M10=ATACTTACAGCCAT WT=  89
-35 AT 19 & -10 AT 44 WEIGHT=   34 OR ----------> 201%

M35=AGTATTGACTTAAAGT WT= 122  SPACE(18)= 6  M10=ATACTTACAGCCAT WT=  89
-35 AT 20 & -10 AT 44 WEIGHT=   54 OR ----------> 320%


                          FOR THE REVERSE DIRECTION

M35=TCTCGATGGCTGTAAG WT= 105  SPACE(19)= 1  M10=AGACTTTAAGTCAA WT=  92
-35 AT 43 & -10 AT 18 WEIGHT=   35 OR ----------> 207%

M35=TCTCGATGGCTGTAAG WT= 105  SPACE(18)= 6  M10=TAGACTTTAAGTCA WT=  98
-35 AT 43 & -10 AT 19 WEIGHT=   46 OR ----------> 272%
```

```
REPEAT WITH THE SAME SEQUENCE <N>? N
REPEAT WITH A DIFFERENT SEQUENCE <N>? N
```

Figure 1. Sample working session with the promoter search and evaluation program. The video display parts of the program are boxed, the hardcopy output is left clear.

After specifying which sequence is to be examined and whether a hardcopy output is desired, the program asks for a target string. (The various options are displayed if "Help" and <Return> is typed.) Typing "P" initiates the promoter search and evaluation routines. The match to the

consensus hexamers is then specified. The position of the hexamers is printed if desired. If not, the program reports only the total number of each hexamer found. The presence of hexamers in the opposite strand is detected by searching for the inverse complement in the forward direction. For this reason, the positions of the hexamers indicated by the markers ('*' for the -35 hexamer and its complement, ':' for the -10 hexamer and its complement) is the base at the 5′ end of the target in the top strand and at the 3′ end of the target in the opposite strand.

Homology scores can vary from 100% down to a minimum of -42.6%. The user can restrict the range of values to be reported, as in this example where only values above 0% are reported. In practice all known promoters have values greater than 30%.

T7 A1 is identified unambiguously in this example with a reported score of 74.0%. Note that the scores, as reported, are divided by 10 to obtain true percentages.

Initially the locations of sequences homologous to the consensus sequence of the two most highly conserved regions are identified. "TTGACA" and "TATAAT" are the specific target sequence strings for which a search is made. For both the -35 sequence string ("TTGACA") and the -10 sequence string ("TATAAT"), the number of matches to these sequences is user-specified. In practice, as indicated below, a search for 3 matches out of the 6 base pairs for both the -10 and -35 sequences was sufficient to locate 93 out of the 112 known promoters.

The second stage of the promoter search is the combination of -35 sequences with -10 sequences to form potential promoters. The restriction applied here is that only combinations which result in a spacer length of 15 to 21 base pairs is allowed. All known *E. coli* promoters have spacers in this range.

Once a potential promoter has been located, it is then evaluated according to a weighting scheme. We discuss the concept of weighting schemes and appropriate values for such weights in the Results section. In essence, each base pair within both the -35 and the -10 region is awarded a point score. These scores are added together along with a score for the spacing between the two regions. The analysis of the scores obtained is treated further below. A sample working-session with the program showing the steps involved in the location and evaluation of a promoter within a sequence is shown in Figure 1.

The program is capable of searching 4500 base pairs for a target string of up to 50 base pairs. This limit is similar to other DNA search programs (4, 5) and is necessitated by the further range of options which require memory space. In addition to searching for and evaluating promoter sequences, the program will search for a user-specified target sequence (e.g. repressor and activator binding sites, *etc.*) or for restriction enzyme sites.

RESULTS

Weighting of Potential Promoters

We have developed a computer program that finds and evaluates potential promoter sequences. The evaluation of any potential promoter generated in the program depends on a set of numbers which are derived from the distribution of bases in known promoters as indicated by Hawley and McClure (Fig. 3 of ref. 1).

For the purpose of weighting the -35 region, we define an extended 16 base pair -35 region that includes 9 base pairs upstream of the canonical -35 region hexamer and a single base

pair downstream. This allows inclusion of the weakly conserved A at position -45. Similarly we define a 14 base pair -10 region that includes 5 base pairs upstream of the -10 region hexamer and 3 bases downstream. This includes the weakly-conserved T at -18 and the weakly-conserved TG at -16 and -15 respectively. The sample print-out shown in Figure 1 shows these extended regions for the T7 A1 promoter.

The problem now is to assign a value to each base at each position in these extended -35 and -10 regions. Our basic approach has been to consider the occurrence of each base with respect to its random expectation of occurrence in the 112 known promoters. We assign a value related to the number of standard deviations (assuming Poisson statistics) away from the observed occurrence. Equal occurrence of all four bases is 112/4 = 28. One standard deviation = $\sqrt{28}$ (5.3). For example the so-called invariant T at -8 occurs in 108 out of 112 promoters. Accordingly, a T at this position is assigned a value of 20 (108 ÷ 5.3).

The overall weight also includes a score for the spacing between the -35 and -10 region hexamers. We have assigned scores to the seven different allowed spacings also on the basis of their occurrence among the 112 known promoters. A spacing of 17 base pairs, which occurs in 56 out of 112 promoters is given a score of 14 (56 ÷ $\sqrt{16}$). Similarly spacings of 16 and 18 are given a score of 6; spacings of 15, 19, 20 and 21 are given a score of 1.

Fig. 2 lists the scores for both regions and depicts in bar-chart form the relative values for each base at each position. For each promoter a homology score was calculated as follows:

$$\text{Homology Score} = 100 \left(\frac{\text{sum of base pair scores + spacing score - baseline score}}{\text{maximum score - baseline score}} \right) \quad (1)$$

We have chosen to subtract the random occurrence score as a baseline score. The score for base pairs which occur at a random frequency is effectively countered by subtraction of this baseline score. In essence, the major contributors to the homology score for any promoter are those base pairs which occur at a higher or lower frequency than random.

Relating Promoter Homology Score to RNA Polymerase Selectivity

Promoter strength is properly defined as the rate at which RNA chains are initiated at the promoter. We have shown elsewhere that initiation frequency can be characterized for a number of wild-type and mutant promoters (8, 9). The rate constant for open complex formation, under a number of simplifying assumptions, is:

$$k_{obs} = K_B k_2 [R]/(1 + K_B[R]) \quad (2)$$

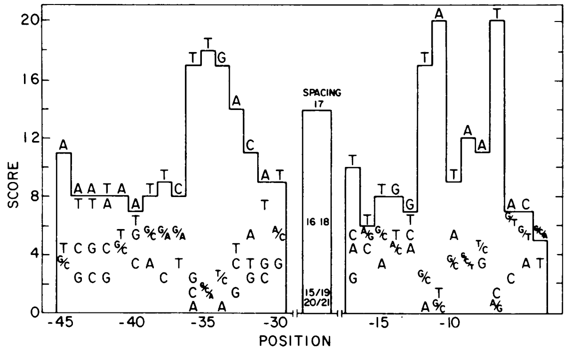
In this paper we wish to consider the enzyme selectivity for promoters, which we define as the product $K_B k_2$. The product is the second-order rate constant for the formation of active 'open-complexes' between the enzyme and the promoter and it is analogous to the term V_{max}/K_m encountered in steady-state enzyme kinetics. We used enzyme selectivity as a measure of promoter strength for two reasons. First, we found no correlation between either K_B or k_2 and homology score. Second, by using enzyme selectivity we avoid the need to define a standard enzyme concentration, which would be necessary in using k_{obs}. We note that in the limit of [RNA polymerase] $\ll 1/K_B$ then k_{obs} and enzyme selectivity become identical.

As shown in the next section, locating possible promoters within DNA sequences is a relatively simple task. Several procedures can do just that (4, 5). The problem is to decide which of the putative sites is most likely to function. In this section we show that the promoter score evaluation can assist in that task. As a test of our evaluation procedure we have correlated 31 values of enzyme selectivity for promoters with the calculated homology score for each of these promoters. The listing of these values is given in Table 1.

Table 1: Promoters Used for Sequence-Enzyme Selectivity Correlation

Promoter	Homology Score	log $K_B k_2$	
1 TAC18	74.0	7.76	(6)
2 T7 A1	74.0	7.40	(7)
3 T7 A3	72.8	7.22	(7)
4 T7 A2	73.4	7.20	(8)
5 λ PR	58.6	7.13	(9)
6 *lac* UV5	59.2	6.94	(10)
7 Tn*10* Pout	56.2	6.71	(14)
8 Tn*10* Pin HH104	52.7	6.55	(14)
9 *lac* UV5-L305	51.5	6.36	(10)
10 *lac* P^S-Δ1	59.8	6.30	(10)
11 *lac* UV5	59.2	6.30	(11)
12 T7 D	63.9	6.29	(8)
13 Tn*10* Pin	52.1	6.18	(14)
14 T7 C	58.6	6.00	(7)
15 T7 D	63.9	6.00	(7)
16 λ PRM up-1	54.4	5.95	(12)
17 *lac* UV5-L241	58.0	5.94	(10)
18 *lac* UV5-L157	55.6	5.90	(10)
19 *lac* P^S-Δ3	55.0	5.54	(10)
20 *lac* P^S	55.0	5.41	(10)
21 *lac* P^S	55.0	5.11	(11)
22 λ PR *x*3	50.9	5.09	(9)
23 *lac* P^+	49.7	4.95	(11)
24 *lac* P^S-L305	47.3	4.83	(10)
25 λ PRM	49.7	4.71	(13)
26 λ PRM E93	50.9	4.55	(13)
27 *lac* P^S-L241	53.8	4.40	(10)
28 λ PRM	49.7	4.14	(12)
29 λ PRM E37	47.3	4.08	(13)
30 λ PRM E104	47.9	4.00	(13)
31 λ PRM 116	39.6	3.85	(13)

Values of log $K_B k_2$ are listed in descending order of magnitude. The corresponding value of the homology score is also shown. TAC18 is our designation for a hybrid *trp-lac* promoter with a spacing of 18 base pairs (6). Citations refer to the source of data. In all cases, the $K_B k_2$ values have been corrected for the fraction of active enzyme as reported by the authors.



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

A	4	5	3	4	4	0	20	5	12	11	0	7	4	6
C	5	4	5	4	5	2	0	3	3	4	1	2	7	6
G	2	5	5	8	7	2	0	3	3	3	0	6	5	6
T	10	6	8	5	6	17	1	9	3	4	20	6	5	4

Figure 2. Table of statistical weights for evaluating *E. coli* promoter homology scores. The histogram displays the scores (tabulated in the lower panel) for each base at each of the positions used in the evaluation of the -35 region, spacer, and -10 region. The individual scores were calculated from the occurrences of each base reported by Hawley and McClure (1): Score = occurrence $\div \sqrt{28}$. The total homology score is obtained for each promoter according to equation (1).

The correlation between enzyme selectivity and homology score is shown by the plot of log $K_B k_2$ *versus* homology score (Figure 3). We used log $K_B k_2$ because we assume that selectivity is related to a sum of free energy terms in the binding and isomerization steps and that these terms depend on the sum of contributions from the different DNA sequences included in the homology score calculation. The correlation shown in Figure 3 is consistent with these assumptions. The least squares fit for all of the data yielded a correlation coefficient of 0.83. Note that the range of selectivity spans about 10^4. The linear relationship obtained is

$$\log K_B k_2 = 0.109(\text{Homology score}) - 0.363 \qquad (3)$$

The broken lines in Figure 3 correspond to regions of the plot that include ±1 and ±2 standard deviations from the least squares fit. The root mean square deviation for all 31 entries is a factor of ±4.1 in $K_B k_2$.

We conclude that the simple occurrence weights in Figure 2 can be used to predict RNA polymerase selectivity to within a factor of about ±4. We emphasize that the correlation in Figure 3 employed objective weights derived from DNA sequence alone together with published enzyme selectivity values from several laboratories. We consider some of the limitations of this procedure in *Discussion*. We are, however, encouraged by the fact that this rather primitive approach can predict RNA polymerase selectivity to within ±4 over as wide a range as 10^4.

Searching for Known Promoters

We have searched the DNA sequence of the known promoters compiled by Hawley and

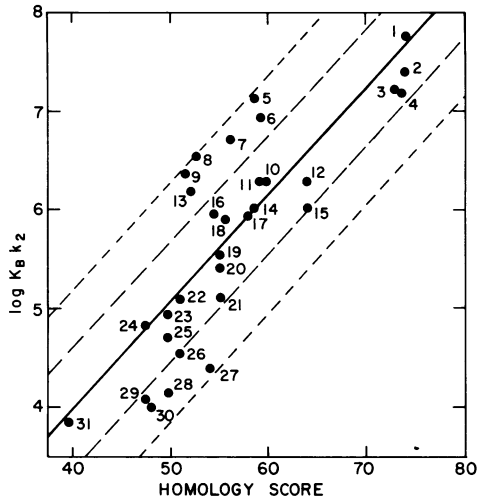


Figure 3. Correlation between log $K_B k_2$ and the homology score calculated for each promoter listed in Table 1. The solid line (linear least squares) has a slope of 0.1086 per homology score % and an intercept of -0.3634 with a correlation coefficient of 0.83. The dashed lines are drawn one (long dashes) and two (short dashes) standard deviations from the best fit line. 20 promoters fall within one standard deviation and eleven were between one and two standard deviations. The value of log $K_B k_2$ at the maximum homology score is 10.5 which would correspond to a value for $K_B k_2$ of 3.15×10^{10} $M^{-1}s^{-1}$.

McClure as an additional test of the search and evaluation procedure. An initial search for promoters was successful for 93 out of 112 promoters, using a required match of 3 out of 6 for both the -10 and -35 regions. Twelve additional promoters were found by requiring only 2 matches in the -35 region and four were found by requiring only 2 matches in the -10 region. λ PRE has only 1 out of 6 base pairs in the -35 region. Two promoters (LEU and IS2 I-II) could not be evaluated by the program because there is insufficient sequence data to define extended -35 regions for them.

The homology scores obtained for the promoters compiled by Hawley and McClure are listed in Table 2 in descending order. These scores cover the same range as the homology scores obtained for the limited set listed in Table 1. 88 out of 110 promoters were found unambiguously. That is to say, these promoter sequences were found and evaluated with a homology score significantly greater than any other possibility in the 60 base pair region used for the search. The remaining 22 are considered ambiguous for one or both of two reasons. First, 15 promoters scored lower than 45, a score which we feel can be set as a lower limit for effective promoters. Although somewhat arbitrary, the score is reasonable based on an examination of the typical scores found in these searches, particularly when the scores of other potential sites within promoters are considered, and also on an examination of the scores found in

Table 2: Homology Scores of Known Promoters

str	79.9	*argCBH*	61.5	**ampC**	53.8
recA	74.6	*his*	61.5	*bioB*	53.8
P22 *ant*	74.0	ϕX A	61.5	*trp* P2	53.8 *
rrnAB P2	74.0	*rrnE* P1	61.5	Pori-1	53.3
T7 A1	74.0	*trp*	61.5	*fol*	52.7
T7 A2	73.4	ColE1 P2	60.9 *	pBR322 *bla*	52.7
T7 A3	72.8	RSF RNA I	60.9	spot 42 RNA	52.1
434 PR	72.2	*spc*	60.9	Tn*10* Pin	52.1
rrnG P2	72.2	λ L57	60.4	*lexA*	51.5
uvrB P1	71.0	pBR322 primer	60.4	*araC*	51.5
rrnAB P1	70.4	**Tn*10* *tetR***	60.4	*trpR*	51.5
rrnG P1	70.4	***uvrB* P2**	60.4 *	*ilvGEDA*	51.8
				deo P2	50.3
R100 RNA I	69.8	fd X	59.8 *		
rpoA	69.8	*gal* P2	59.8	*lac* P1	49.7
bio P98	69.2	ColE1 P1	59.2	λ PRM	49.7
λ c17	68.6	P22 *mnt*	59.2	λ PI	49.1
λ PR'	68.6	λ Po	58.6	*uvrB* P3	49.1
ϕX D	67.5	λ PR	58.6	*alaS*	48.5
Tn*10* *tetA*	67.5	pBR322 P1	58.6	*lac* P115	48.5
R100 RNA II	66.9	RSF primer	58.6	pBR322 P4	47.3
rrn DEX P2	66.9	S10	58.6	*hisJ*	46.7
supB-E	66.9	T7 C	58.6	P22 PRM	46.7
tyrT	66.3	λ PL	58.0	Tn*5* IR	46.2
rrnX P1	65.7	Tn*5* *neo*	58.0		
rrnD P1	65.1	M1 RNA	57.4	*malT*	43.8
		Tn*10* Pout	57.4	*araBAD*	41.4
λ *cin*	64.5	*aroH*	56.8	*gal* P1	40.8 *
pBR322 RNA I	64.5	CloDF 1	57.1	*bioA*	40.2 *
thr	64.5	*deo* P1	57.1		
tufB	64.5	ϕX B	57.1	*hisA*	39.1 *
fd VIII	63.9	434 PRM	56.5	R100 RNA III	37.9
pBR322 *tet*	63.9	$\gamma\delta$ *tnpA*	55.0	*lacI*	38.2 *
rplJ	63.9	*tnaA*	55.0	*trpS*	37.9
T7 D	63.9	$\gamma\delta$ *tnp R*	54.4	*lac* P2	36.1 *
leu 1 tRNA	63.3	*lpp*	54.4 *	*malEFG*	34.3 *
P22 PR	63.3			*rpoB*	33.1 *
R1 RNA II	62.7			*cat*	32.5
glnS	62.1			λ PRE	32.5 *
				malK	32.0
				Pori-r	31.4 *

The promoters listed by Hawley and McClure are tabulated according to their homology score. 88 promoters out of these 110 were found unambiguously. Those promoters (e.g. fd X) marked with an asterisk(*) are promoters for which alternate alignment possibilities were indicated within 5 points of that alignment listed by Hawley and McClure. Two promoters could not be evaluated properly (see text).

a search of pBR322 which is described in the next section. Second, 14 promoters presented an ambiguity with respect to alignment. In these promoters, the search revealed a number of potential promoter alignments with a comparable score to that obtained by the alignment of Hawley and McClure.

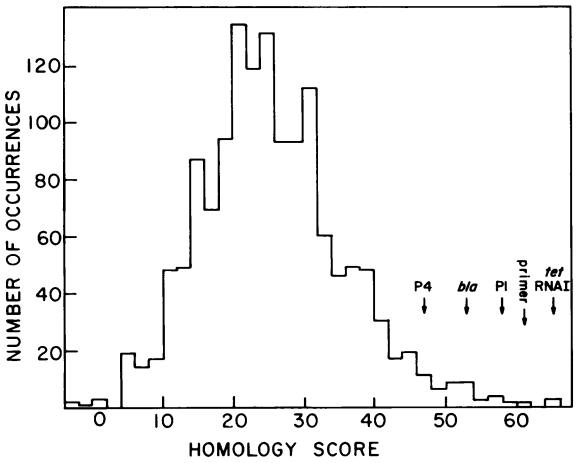


Figure 4. Frequency distribution of possible promoters in pBR322. pBR322 was searched in both directions for all possible promoters having a match of three or greater to both the -10 and -35 consensus hexamers. A total of 1396 possible promoters were found. The mean homology score was 25.0%. Each bar represents the frequency of a double weight class (e.g. 10% and 11%). The locations of the six known promoters of pBR322 in this distribution are indicated.

Search of pBR322 for Promoters

A search of pBR322 for possible promoters revealed all six of the known promoters. The search, with the restriction of 3 matches out of 6 in both the -35 and -10 region, revealed a wide spectrum of possible promoters ranging from ~0% to 65%. The overall distribution is shown in Fig. 4. This distribution of homology scores contrasts with that seen for the known promoters which ranged in score from 30% to 80%. The six known promoters in pBR322 serve to calibrate the distribution since RNA I is a relatively high scoring promoter (64.5%) while P4 has a poor score (47.3%). Clearly, considering our results with known *E. coli* promoters, the vast majority of possible sites in pBR322 will not be promoters. We base this conclusion on the correlation of Figure 3, where it is shown that promoter function increases by a factor of 10 for each increase of 10 points in homology score. For example, a promoter with a score of 65 is predicted to function at a level one hundred times that of a promoter with a score of 45. In this view, the large number of sites found in the pBR322 search (average score = 25) are functionally insignificant. However, in the range of possibilities of score greater than that of P4, there are 26 candidates in addition to the six known promoters. Some of these may prove, upon further investigation, to function *in vitro* as promoters.

DISCUSSION

The work reported here indicates that a DNA sequence known to contain a promoter, can be analyzed semi-quantitatively to locate the promoter and to obtain an estimate of its strength *in*

vitro. Our evaluation scheme rests on a number of assumptions which ultimately must be tested. The major assumption that we make is that the DNA sequence alone is sufficient to specify promoter strength. We assume that a promoter will consist of two distinct regions separated by a spacer. All known *E. coli* promoters follow this pattern. The search for these two hexamers, while stringent, nevertheless represents an initial selection of potential candidates. By varying the degree of match required, this stringency can be as varied as desired. In practice, 3 matches out of 6 found 83% of known promoters.

An important assumption in the derivation of the table of weights for the different base pairs is that the consensus sequence is the most favorable one. Evidence in support of this assumption comes from a consideration of promoter mutations. Hawley and McClure (1) recorded 98 promoter mutations. With a few special exceptions the down mutations decreased homology to the consensus sequence and the up mutations increased the homology to the consensus sequence. Most of the exceptions to the above rule involved changes in nonconsensus base pairs to other nonconsensus base pairs suggesting a base pair hierarchy at these positions.

In using the weight tables, we assume that the weight for the overall -10 or -35 region is given by the sum of the weights for the individual base pairs. If we postulate that, to a first approximation, the free energy of interaction of the enzyme and a -10 or -35 region is the sum of the free energies of the individual interactions, and if our weights are related to these individual free energies in some fashion, this assumption is plausible. By a similar argument, we assume that the overall homology score is a linear combination of the three terms, -35 region, spacer and -10 region. Until these postulates are tested critically, we are unable to include terms representing conformational equilibria in the DNA and enzyme.

In assigning values to the different base pairs at different positions, we have assumed that the observed occurrence of a base pair at a certain position is directly related to its functional significance. We have used the deviation from the statistical mean as the weighting unit. However, our use of a Poisson distribution to calculate the weight is not essential for the method but allows a more convenient way of using different weight tables.

Despite these caveats, we have obtained a result demonstrating, for the first time, a direct relationship between DNA sequence and RNA polymerase selectivity. We do not wish to draw any quantitative conclusion outside the range of experimental data. The experimental value of the current relationship (Figure 3) lies in the predictions it makes about the relative values of $K_B k_2$ of one promoter with respect to another within this range. We expect that measurements on additional promoters will test the simple relationship between $K_B k_2$ and homology score that is suggested by Figure 3, and will also lead to improvements in the formulation of homology score.

There are two important limitations in our analysis. First, some promoters (*e.g. lac* and λ PRM and their mutant derivatives) are over-represented in the experimental data. As a result, the homology scores at several positions, including some within the hexamers, have not been tested in our correlation. This means that the correlation of Figure 3 is merely consistent with the assumptions identified above. A second limitation is the experimental uncertainties in the selectivity data. Standard deviations for $K_B k_2$ have not been reported, however, $K_B k_2$ values for

four of the promoters listed in Table 1 were determined in two different laboratories. The average difference in $K_B k_2$ for these pairs of determinations was 2.8. Thus, the correlation between selectivity and homology score cannot be expected to be better than about ±3 in $K_B k_2$ with the current data.

This analysis suggests two guidelines for locating a promoter within a DNA sequence. First, the promoter should have a relatively high overall score. In our scheme a cut-off score of 45% seems appropriate. By suggesting this cut-off score we do not wish to include or exclude categorically any potential site from being a promoter. For example, most of the promoters listed in Table 2, which have poor homology scores (<45%), are known to require activators for maximal expression. We believe that a continuum of selectivity is a fundamental property of RNA polymerase-promoter interactions. By virtue of the correlation, which we observe between homology score and enzyme selectivity we are in a better position to assess possible promoters in this continuum. We see, for example, that *lacI* is unlikely to be a strong promoter. Similarly, we exclude most of the possible sites in pBR322. Second, the promoter should be the highest scoring potential promoter within the region under consideration. The presence of additional potential sites of comparable score adds to the uncertainty of location and alignment.

A significant difference between our analysis and that of Harr, *et al.* (15) is that we have correlated promoter sequence with function. Also, we have included regions of DNA sequence in addition to the consensus hexamers in evaluating promoter homology score.

In conclusion, using an objective set of weighting values to describe the contribution of each base pair at each position of a promoter sequence, we have found that the degree to which a DNA sequence resembles the consensus of promoter sequences can be related to RNA polymerase selectivity. We have also suggested criteria that should be useful for predicting the location of promoters. But, we emphasize that this analysis only suggests possible promoter sites. The location of a promoter can only be established by biochemical (5' end determination) or genetic (mutations) evidence. However, by identifying and evaluating possible promoter sites, the expectation is that such evidence would be obtained in a more informed and efficient manner.

ACKNOWLEDGEMENTS

We thank Kathryn Galligan for her assistance in preparing this manuscript. We are grateful to Carol Cech for communicating results prior to publication and to Peter von Hippel for comments on this manuscript. Research on RNA polymerase in this laboratory is supported by the N.I.H. (GM 30375). The development of the TARGSEARCH Program was supported by the N.S.F. (PCM 8140433).

REFERENCES

1. Hawley, D. K. and McClure, W. R. (1983) *Nucleic Acids Res.* 11, 2237-2255.
2. Rosenberg, M. and Court, D. (1979) *Ann. Rev. Genet.* 13, 319-353.
3. Siebenlist, U., Simpson, R. B. and Gilbert, W. (1980) *Cell* 20, 269-281.
4. Fristensky, B., Lis, J. and Wu, R. (1982) *Nucleic Acids Res.* 10, 6451-6463.
5. Larson, R. and Messing, J. (1982) *Nucleic Acids Res.* 10, 39-49.

6. Mulligan, M. E., Brosius, J. and McClure, W. R., in preparation.
7. Dayton, C. J., Prosen, D. E., Parker, K. L. and Cech, C. L. (1983) *J. Biol. Chem.* (in press).
8. McClure, W. R. (1980) *Proc. Natl. Acad. Sci.* 77, 5634-5638.
9. Hawley, D. K. and McClure, W. R. (1980) *Proc. Natl. Acad. Sci.* 77, 6381-6385.
10. Stefano, J. E. and Gralla, J. D. (1982) *J. Biol. Chem.* 257, 13924-13929.
11. Malan, T. P., Kolb, A., Buc, H. and McClure, W. R., (1984) *J. Mol. Biol.* (in press).
12. Hawley, D. K. and McClure, W. R. (1982) *J. Mol. Biol.* 157, 493-525.
13. Shih, M-C. and Gussin, G. N. (1983) *Proc. Natl. Acad. Sci.* 80, 496-500.
14. Simons, R.W. Hoopes, B.C., McClure, W.R. and Kleckner, N. (1983) *Cell* 34, 673-682.
15. Harr, R., Häggström, M. and Gustafsson, P. (1983) *Nucleic Acids Res.* 11, 2943-2957.