Volume 16 Number 15 1988 Nucleic Acids Research

Defining the consensus sequences of *E.coli* promoter elements by random selection

Arnold R.Oliphant and Kevin Struhl

Department of Biological Chemistry, Harvard Medical School, Boston, MA 02115, USA

Received May 13, 1988; Revised and Accepted July 6, 1988

ABSTRACT

The consensus sequence of *E.coli* promoter elements was determined by the method of random selection. A large collection of hybrid molecules was produced in which random-sequence oligonucleotides were cloned in place of a wild-type promoter element, and functional -10 and -35 *E.coli* promoter elements were obtained by a genetic selection involving the expression of a structural gene. The DNA sequences and relative levels of function for -10 and -35 elements were determined. The consensus sequences determined by this approach are very similar to those determined by comparing DNA sequences of naturally occuring *E.coli* promoters. However, no strong correlation is observed between similarity to the consensus and relative level of function. The results are considered in terms of *E.coli* promoter function and of the general applicability of the random selection method

INTRODUCTION

The relationship of genetic structure to function is especially important in the understanding and prediction of a biological phenotype. By comparing several genetic elements required for a given function, an understanding of the sequence requirements for that function can be established. Commonly found aspects of a genetic element are said to form a consensus.

Consensus sequences have been defined in two ways. One approach compares those naturally occurring DNA sequences that are believed to encode a particular genetic function. The other method is to generate many mutations of an individual genetic element. Each approach has inherent advantages, limitations and biases.

The characterization of enough wild-type elements in order to accurately define a consensus can be prohibitive. Such elements are inherently biased toward the systems that have been chosen for study and may not accurately reflect the sequence distribution found in nature. The use of wild-type elements can also be misleading because the circumstances surrounding each element are varied. If elements from different organisms or from different genetic positions within the same organism are compared, the contextual differences can have a significant effect on the ability of these elements to function. It is very difficult to assess the relative effects of these various influences.

If a consensus is defined by making many mutations of a single element, the context of each element being compared is controlled. However, the generation of those mutants can be

© IRL Press Limited, Oxford, England.

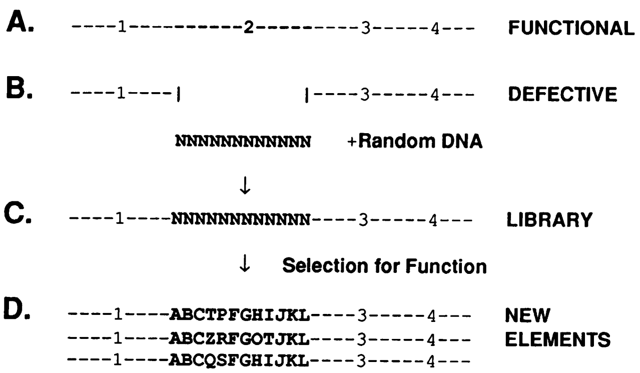

Figure 1: General method for determining a consensus sequence. (A) A group of genetic elements that can confer a specific phenotype. (B) A vector that lacks the genetic element of interest and is functionally ineffective. (C) Double-stranded, random sequence DNA is substituted in place of the omitted element to form a library of hybrid molecules. A selection or screen is used to identify those sequences that confer the function of interest. (D) A comparison of those molecules which pass the selection defines a consensus for the selected function.

time-consuming and there is no certainty that informative mutants have been generated. When mutants of a given wild-type element are compared they are also biased by the sequence of the particular element chosen for mutagenesis. Functional elements that are significantly different from the wild-type element will be overlooked.

"Random Selection" represents an alternative method that can minimize these biases in defining the sequence requirements of a genetic element (1; See Fig. 1). A collection of recombinant DNA molecules is made in which a short sequence of random DNA replaces a wild-type genetic element. A selection or screen is made to isolate from this collection those sequences that confer the function of the wild-type element. A comparison of the DNA sequences that satisfy a particular selection results in a consensus that defines the genetic element. Unlike conventional mutagenesis methods which create derivatives of a wild-type sequence, random selection uses selective pressure to choose functional elements from a population of random-sequence DNAs. Many different sequences that confer a specific function can be generated. The resulting elements are not biased by the sequence of any particular wild-type element and their function can be compared in the same organism in the same context of surrounding DNA. In conceptually related experiments, sequences sufficient for mitochondrial import or for transcriptional activation in yeast have been selected from short segments of human or *E.coli* DNA (2,3).

In this paper, we select functional -10 and -35 promoter elements of *E. coli* from random DNA sequences. A consensus of these selected elements is similar to the consensus derived from naturally-occurring promoter elements (4,5). This similarity demonstrates the ability to select functional elements from random-sequence DNA that are representative and descriptive of wild-type genetic elements.

MATERIALS AND METHODS

Synthesis and cloning of random sequence oligonucleotides

Oligomers were synthesized by Alexander Nussbaum using the phosphite triester method on an Applied Biosystems DNA synthesizer. The random-sequence oligonucleotide was generated by using an equal mixture of all four nucleotide precursors and by omitting the capping reaction after each of the central steps. This modification improves the yield because oligonucleotides that fail to react at a given step remain active and can react at subsequent steps; it also results in oligonucleotides that are heterogeneous in length.

Oligonucleotides were cloned after conversion to the double stranded form by mutually primed synthesis (6; see Fig. 2). Five μg of oligonucleotide was hybridized at the 3' ends at 37°C for one hour in 10 μl of 3X buffer (30 mM Tris (pH 7.5), 150 mM NaCl, 30 mM MgCl, 15 mM dithiothreitol, 0.1 mg/ml gelatin) and then cooled to room temperature and placed on ice. Deoxynucleoside triphosphates (at a concentration of 250 μM for each of the four) and 10 μCi of α-^{32}P-dATP were then added, and the reaction mixture was diluted with water to a final volume of 30 μl. Five units of Klenow enzyme were added, and after incubation at 37°C for at least one hour, the products were phenol extracted and ethanol precipitated. The DNA was resuspended and cleaved with 50 units of the restriction enzyme recognizing the original 5' sites, phenol extracted, ethanol precipitated and separated on a 12% native polyacrylamide gel. The desired product was purified from the gel, cleaved with the restriction enzyme recognizing the original 3' end of the oligonucleotide, phenol extracted, ethanol precipitated and resuspended. At this stage the DNA was double-stranded and suitable for cloning.

Vector Constructions

The vector (mp19-Sc5015, Fig. 4A) for the selection of functional -35 elements was constructed by cloning an *Eco*RI-*Xho*I fragment containing the yeast *his3* gene into an M13mp19 vector. A -10 element was then cloned between the *Eco*RI and *Sac*I sites upstream of the *his3* gene by using mutually primed synthesis to create a double-stranded version of the oligonucleotide 5'-CGCGAATTCCCATTATAGAGCTCT-3'. The construction of a vector for the selection of -10 elements (mp19-Sc5014, Fig. 3A) was done in a similar manner except that the *Eco*RI site of the original vector was deleted and a functional -35 element was inserted between the *Sal*I and *Bam*HI sites by cloning 5'-CGCGTCGACCATTCTTGACAGGATCCT-3' by mutually primed synthesis.

Libraries containing random sequence DNA were made by ligating 5 μg of each vector cut with *Bam*HI and *Sac*I and the oligonucleotide 5'-GGCGGATCC . N_{25} . CGAGCTCG-3' that had been prepared as described by mutually primed synthesis (Fig. 2). As the yield of double-stranded DNA was somewhat variable, the amount of insert to be added to a given amount of vector was determined empirically in order to optimize the ligation reaction. The ligation products were introduced into *E.coli* by standard techniques to generate libraries of 500,000 independently derived phage. After transformation the cells were grown at 37°C for 4 hours, and the resulting phage were isolated by precipitation in polyethylene glycol.

Selection for functional promoter elements

Phage libraries for the selection of functional promoter elements were used to infect *E.coli* KC5, an F^+ derivative of *hisB*463 (7) at a multiplicity of infection of 5-10. Infected cells were spread on agar plates containing glucose-M9 minimal medium with aminotriazole and incubated for two days at 37°C. Phage obtained from these colonies were cross-streaked with fresh *E.coli* KC5 cells to ensure that cell growth was phage dependent. After plaque purification, single-stranded phage DNA was prepared and subjected to DNA sequence analysis by the chain termination method (8). Relative resistance to aminotriazole was determined by patching cells on minimal media plates containing 20, 30, 40, and 50 mM aminotriazole.

RESULTS

The method of random selection is applicable to any genetic element that confers a phenotype that is subject to a selection or screen. Random-sequence oligonucleotides can be generated by using equal mixtures of the four nucleotide precursors at each step of the chemical synthesis. However, standard methods of cloning these oligonucleotides (9,10) are unsuitable

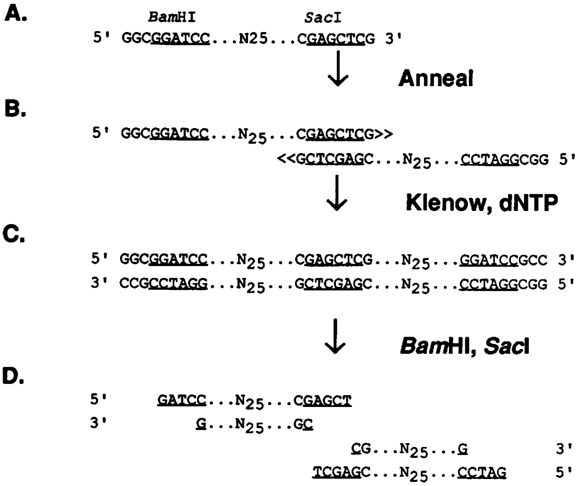
A.
BamHI
SacI
5' GGCGGATCC...N25...CGAGCTCG 3'
Anneal
B.
5' GGCGGATCC...N25...CGAGCTCG>>
<<GCTCGAGC...N25...CCTAGGCGG 5'
Klenow, dNTP
C.
5' GGCGGATCC...N25...CGAGCTCG...N25...GGATCCGCC 3'
3' CCGCCTAGG...N25...GCTCGAGC...N25...CCTAGGCGG 5'
BamHI, SacI
D.
5' GATCC...N25...CGAGCT
3' G...N25...GC
CG...N25...G 3'
TCGAGC...N25...CCTAG 5'

Figure 2: Mutually primed synthesis. (A) the oligonucleotide containing 25 bases of a mixture of the four nucleotides bounded by *Bam*HI and *Sac*I sites; (B) two oligonucleotides annealed at their complementary 3' ends; (C) double-stranded DNA after extension of the 3' ends with Klenow and the four nucleotide triphosphates; (D) a pair of double-stranded DNAs suitable for cloning into a *Bam*HI-*Sac*I cleaved vector.

```
A.   GTCGACCATTCTTGACAGGATCCCCGGGTACCGAGCTC--his3--
     SalI        -35   BamHI           SacI

B.   GGCGGATCC...N25...CGAGCTCG
        BamHI           SacI

C.  ---GTCGACCATTCTTGACAGGATCC+       +CGAGCTC---
M63  17  1.27 -------TGCCGAGGGCC CATACT
M11  18  0.87 ------GCCCTGTTCTAA TATTCT G
M41  17  1.03 -------TTCAACCGACG TACTTT GTCATTTT
M23  17  0.60 -------CAAACTTCTAT CACAGT AGAGGTT
M53  16  0.13 --------CCGTATAGGC GACACT ATAGTCGTCA
M31  17  1.14 -------CACCTAGGCAT TAGTCT CCCATTTT
M13  16  1.44 --------GCTTTGCTGC TATATT GCGGGGTCT
M46  19 -0.50 -----CCCCGGCCGCTCG TAGGCT AT
M21  17  1.25 -------CCATGACCCTG TAACTT CCTCCTCA
M42  18  0.65 ------CACTGTTATGGC TAATCT GCCCCCT
124  17  0.64 -------AGGTTTTCACC CATTCT ATAGCTCT
M26  17  0.87 -------CCTTCACCAGC TACCGT GGACCGAA
116  19  0.03 -----GACAAATAAGGCA TACACT ACGGTA
110  18 -0.48 ------CCCTATATGTGA TAGATC ACGATTT
125  17  1.03 -------CGTGGCTGTTT TACTTT CATATTTA
103  17  0.12 -------AACATTCATGT CAGCTT GTCTTGCT
100  17  1.05 -------CACATACAGCG AATAAT CTCCAACT
M64  17  0.41 -------CTATGAGGTGG CAATCT GAGAATTT
M62  17  1.05 -------CCTAGCACGCG CAAACT ACTCCTAT
94   17  0.95 -------CGTAACCTCTT TAGTGT GGTACCCA
93   17  1.19 -------CATTCCAGACA TAGGCT GTACCATT
112  17  2.14 -------TGGGCGGCTCG TATATT GTGAC
M15  18  0.87 ------GTACGTGGATAC TATTCT TGCGTAT
114  17  0.28 -------GAAACATGTGT TCAATT CACCATCT
98   17  1.14 -------GCGAACCTGAC TAGTCT CACAATAT
M25  17  1.77 -------TCACTTCACTC TAGACT TTGTGAC
91   17 -0.52 -------CCATCTTACCG TAAGTA TTAACGCT
123  17 -0.20 -------CGGGTAAGGGT TACAAG GTACCAAA
M24  17  1.54 -------TCTTGCGTGGT TATCCT GGGGAATT
129  17  0.82 -------CAACCGCGAGA CAATAT TATTTTCT
113  17  0.16 -------CCCTGCCGTGA CACTCT G
130  16 -0.93 --------ACCAGAAGCT GTTATT AGAGGTGCA
106  17  1.21 -------GGTGACTCATC TAAGGT CATAGAT
95   18  0.51 ------ATCGCAATGTTA CATATT GCGCTTT
102  16  0.85 --------ATTTCCTACG TACGAT AGACATCAA
111  18 -0.16 ------GTCCACTCAGGG CACCAT TACCCT
99   17  0.57 -------ACAAAGACGGG TCTACT ACAAAAGTT
131  17  1.24 -------ATCGCAATAGG GACAAT CTACTCC
128  18  0.44 ------CTTCAATAAGTC TAGTCT ACTCGAC
115  15 -0.54 ---------AGTTTTCCG TGTAAT AGAGGACCC
92   18  0.79 ------CATAGCTTGGTC GAAAAT CATTTAA
44   17  0.65 -------GCTATCACAGG CAGAGT CTGACCATT
45   17  1.06 -------ACATGCCCCAC TACCCT TCAATAAT
132  18 -0.21 ------CAGGCGTGCTGT TTAACT ACTGCGTA
63   16  1.02 --------ATAGCTCAAT TAACAT ATCCCGTCT
96   16  0.36 --------AGACTCCTGA TACCCT TAAGCACTT
M12  17 -0.18 -------TCTACCACTGG CATAAA AACAGTCT
127  16  0.92 --------GCATCGCGTT TATGCT TAGAAGTAT
126  17 -0.03 -------ATCCATTCAAG CTAAAT GAGCGAAA
49   17  0.21 -------AGATTTTCCAC CAGTCT GTTCCAC
```

Figure 3: Experiment for selection of the -10 element. (A) Sequence of vector mp19-Sc5014 which includes a functional -35 element. (B) The oligonucleotide used for mutually primed synthesis and cloning into mp19-Sc5014. (C) Identification number, spacing of elements, liklihood scores, and the sequences of each element, aligned by the six conserved bases of the presumptived -10 sites. They are divided into four groups according to the cell's resistance to aminotriazole. The top group is resistant up to 50 mM, and the remaining groups up to 40, 30 and 20 mM respectively. Liklihood scores representing the extent of similarity to the wild-type consensus matrix and spacing of elements were calculated as described by Staden (12) except that only the 6 most conserved positions of the element were used and all nucleotide frequencies were divided by 0.25 such that random sequences will score an average of 0. Positive scores indicate more similarity to the consensus matrix and negative scores indicate less similarity.

because the heterogeneity of random-sequence DNA does not allow the generation of a complementary template. To clone random-sequence oligonucleotides with the high efficiency necessary for large libraries, the method of mutually primed synthesis has been developed (6). In this paper, we cloned oligonucleotides containing 25 base pairs of random-sequence DNA flanked by *Bam*HI and *Sac*I sites as described in Fig. 2.

As an initial application of the random selection method, the -10 and -35 elements of the *E. coli* promoter were chosen because the results could be compared to the many wild-type examples known (4,5) and because of the availability of a simple genetic selection for obtaining functional elements. Interpretation of wild-type promoters requires the positioning of both the -10 and -35 elements and an analysis of their relative contributions to overall promoter strength. When different wild-type promoters are compared, expression levels can not be attributed solely to the -10 or -35 sequences as both will vary. To avoid these complications functional -35 and -10 sequences were included in two different vectors (Figs. 3A,4A) allowing an independant study of the two components of the promoter. Selecting for only one of the elements at a time greatly simplifies the interpretation of the experiment. The sequence and position of the element under selection can be studied without sequence or position variation of the other.

In each vector, the double-stranded version of the random-sequence oligonucleotide was inserted in an M13 vector between *Bam*HI and *Sac*I sites upstream of the *his3* gene. The use of an M13 vector allows for rapid sequencing of the resulting hybrids. A collection of 500,000 phages was then produced from each vector by transformation of *E.coli.* In order to assess the base composition of the inserted oligonucleotide, seventeen of these clones were sequenced without selection, yielding a total of 87 G's, 103 C's, 105 T's, 93 A's and showing no significant deviation from a random distribution of nucleotides.

The genetic selection requires the expression of the yeast *his3* gene which encodes the enzyme imidazolglycerolphosphate dehydrogenase. Although *his3* is derived from a eukaryote, its expression permits *E.coli* containing the *his*B463 mutation to grow in the absence of histidine (7). To obtain functional elements, KC5 (*his*B463 F^+) cells were infected with the M13 phage library containing the *his3* gene and potential promoter sequences and selected for growth on minimal media containing aminotriazole, a competitive inhibitor of IGP dehyratase. Phage were isolated from the bacteria that passed the selection, plaque purified and sequenced. As expected, the frequency of phages passing the selection decreased as the selective pressure was increased by using higher concentrations of aminotriazole. Under the most stringent conditions (50 mM aminotriazole), approximately 10^{-4} of the phage were able to pass the selection.

As the degree of aminotriazole resistance is related to the level of *his3* expression (11), it is possible to rank the promoter elements according to their relative levels of function. For this purpose, phage were again infected into *E.coli* KC5 and the cells assayed for growth in various concentrations of aminotriazole. The sequences of these clones are presented in Figs. 3C and 4C and ordered according to the cells' relative aminotriazole resistance. The central six nucleotides

A. GGATCCCCGGGTACCGAGCTCTATAATGGGAATTCCAAAAAT--*his3*--
*Bam*HI *Sac*I -10 *Eco*RI

B. GGCGGATCC...N_{25}...CGAGCTCG
*Bam*HI *Sac*I

```
C. ---GGATCC+     +CGAGCTCTATAATGGGAATTC---
M62  16  1.38 ---------ACACAATACT TTGACG GCAGGCGTC
M54  17  0.93 ----------TACCCTGCC TTTCCT TTAACCCATT
M51  15  1.11 --------GTTGATATCTC TTGACT TTTTTCAG
M42  17  1.52 ----------GGTCCTTTC TTTACT GTTCTGACAT
M41  19  0.27 -------------CACGTC TTGTCA ACTCCATTGTTC
M36  16  0.89 ---------GAAACCAGAA TTGCAA AAGCCCTCA
M14  17  1.22 ----------AGTCCTTCG TTGCTT CTCATACACA
M11  18  0.11 ---------CCCCTTTAAC TTGCAC GCCTTTCATTC
M61  18  0.89 -------------CACGTC TTGCAA CTCCATTGTTC
M43  17  1.22 -----------GCCGGTTA TTGCTT TCAATGCGCC
M26  16  0.53 ---------TACCATCTAT TTGCTT CTTTGCCTG
M21  17  0.92 ----------TTTTACCAC TTGTTG CTCCACAATA
91   17  1.41 ----------TCCTGGAAG TTGAAC ATTTTCTGTC
90   17  0.37 ----------GTCTAGCAT TGTACT TTGCATCCCG
68   17  1.38 ----------GGGTTATCC TCGACA GGTATTTTTA
66   16  0.47 ---------ACCGATCTCA TTGATC AAAACGCTA
M44  19  0.84 ------------GGCTCCC TTGACA AAACATATTTCC
M13  18 -0.12 ---------------TGGC TTGCTC TGACACTGATT
M12  17  1.65 ----------TACTGGGTC TTCACA CAGCACTCTC
98   16  1.25 --------GGACCGGCCAT TTGCCA TTTGTGTTC
M34  16  1.12 ---------GCCTAGAAGC TTGCCT ACAGTCTAA
87   18  1.48 -----------CTCCAAGG TTGAAA CATGAGTATAC
81   18  0.23 -----------GTAACTTC TAGACG GATCTCGTCTC
77   16  0.69 ---------ACACCCCGGC TGGACA ACTAACGAC
86   17  1.15 -----------CGCCACCC TTGTAG ACTTGCGAGG
69   17  2.08 ----------TCGTACCCA TTGACG CTATAGCAAT
M22  18  0.11 -----------AGGAAGTA TTGCAC CTAATCTGACC
76   17  0.50 ----------AGTAAATTA TGTACA CTCCTTCCCG
75   17  0.27 ----------ACGCCCACT TTACCT AAGGGCTGCA
74   18  0.82 -----------GTGGCTGT TTTACT TTACCCTTGAT
100  16  0.23 ----------ATCTGACTA TTCATT TTCCAGCCA
88   18  1.84 -----------CTCTCACT TTGACA GTGCTGTTGAC
64   18  0.55 ------------TCGCGAC TTGTTT AATAGCACGCT
M35  18 -0.23 -----------ATCACGTA TTGTGC ATAACAACCAA
80   17  1.65 ----------AAGAAATCA TTCACA GGCCATTCAA
33   17  1.26 -----------TCTCTTAC TCGACT CTCCCCCGCC
105  17  1.38 ----------CTAGTGGTA TTGTTA CTAGGGGCTA
102  17  0.81 -----------TCAGGTGA TTGCAC GTATCCTATC
M25  18  0.82 -----------GACAGTTA TTTACT CCGCTTATCAT
83   18  0.56 -----------GATACTAG TCGACT AATACAGGAGT
82   17  1.13 -----------GACTTTGC TTGGAA ACCATCAAAT
104  17 -0.22 -----------CAACCCCC TATATT TATTCGGCTT
73   19  0.07 --------------GACCC TTGACC TGTTCCCCATAA
36   17  0.50 ------------GTCAAAC TATACA CACTTAGGCG
85   17  1.26 ----------ATCGGAACA TGGACT GAATTCCGCT
89   17  1.29 ----------GCTATAGGC TTTAAA ACCTGCCGAT
67   16  0.49 ---------TCTTTTTGGC TTTACG ACCCTTCAC
M32  16  0.04 ---------TCTTCGGCCC TCGCCT GGCTCTCTC
84   16  1.12 --------GCGTGTATACA TTGATT CCCCTTACA
72   17  0.95 ----------CGATGGATC TTTTCT TACCGGCTGG
71   18  0.42 -------------TGCCTG TTGTGT TACTCATTTCG
70   18  1.07 ------------CGCCGTT TTGACC AAACTTGCTGC
99   17  0.93 ---------AGCTGATCCT TTTATT TTAAACGACA
96   18  0.47 ------------CGCCGGT TTGATC ATATGACTACT
79   18  0.66 ---------------GCCG TTGGCT AAACCCTAAAA
19   17  1.48 -----------GTACCCAT TTGCCG CTATACCAGC
```

Figure 4: Experiment for selection of the -35 element. (A) Sequence of vector mp19-Sc5015 which includes a functional -10 element. (B) The oligonucleotide used for mutually primed synthesis and cloning into mp19-Sc5015. (C) Identification number, spacing of elements, and the sequences of each element, aligned by the six conserved bases of the presumptive -35 sites. The top group is resistant up to 50 mM aminotriazole, and the remaining groups up to 40, 30 and 20 mM respectively. Liklihood scores representing the extent of similarity to the wild-type consensus matrix and spacing of elements were calculated as described by Staden (12) except that only the 6 most conserved positions of the element were used and all nucleotide frequencies were divided by 0.25 such that random sequences will score an average of 0. Positive scores indicate more similarity to the consensus matrix and negative scores indicate less similarity.

A.

			-35							-10				
G	1	4	42	2	2	8	...	3	1	11	7	6	2	G
C	0	4	3	15	32	11	...	15	2	14	8	24	1	C
A	0	3	1	33	11	16	...	1	47	13	21	12	2	A
T	57	47	12	8	13	23	...	33	2	14	16	10	47	T

B.

15	16	17	18	19
2	18	59	26	5

Figure 5: Matrix for the consensus of *E.coli* promoter elements. The number of times each base occurs in each position of the -35 and -10 sequence elements selected from random DNA (A), and the number of those elements separated by a spacing of from 15 to 19 base pairs (B).

representing the genetic element in Figs. 3 and 4 were selected by the method of Staden (12).

A compilation of genetic elements can be represented as a matrix listing the number of occurrences of each nucleotide in each position of the element (4,5). The frequency of each nucleotide can then be related to the probability of that nucleotide's occurrence in a wild-type element. This relationship can be a useful tool in predicting the presence and activity of wild-type elements (12-14). A matrix of nucleotide use for these selected elements is shown in Fig. 5. For each derivative in Figs. 3 and 4, the similarity of the selected sequence to the consensus matrix is presented in terms of liklihood scores (12).

If random selection is a valid method for determining consensus sequences, the frequency of nucleotide use in a particular position of the selected elements is expected to be similar to the frequency found in wild-type elements. The most frequently occurring nucleotides in wild-type *E.coli* promoters are TTGACA for the -35 element and TATAAT for the -10 element (underlined positions are most highly conserved) with an optimal spacing of 17 nucleotides between the elements (4,5). All of the most highly conserved positions in the consensus for both of the randomly selected elements show the most frequent nucleotides to be the same as those from wild-type matricies. The last position of the -35 sequence and the fifth position of the -10 sequence do not agree with the most preferred wild-type nucleotide. However, in these cases the second most common nucleotide matches the most common wild-type base. No significant sequence preference was observed outside of the six bases of each element presented in the matrix. Experiments for the selection of both elements generate information about the spacing requirements between the elements. All the selected elements are located 15-19 base pairs away from the other element (Fig. 5B). As expected the predominant spacing is 17 base pairs, correlating with the spacing for optimal expression in wild-type promoters. Thus, the major features of *E.coli* promoters can be discerned by the random selection method.

The *E.coli* promoter elements generated by random selection yield a poor correlation between aminotriazole resistance and similarity to the consensus matrix (Figs. 3 and 4). Some highly functional elements have poor matches to the matrix, and some good matches to the matrix are poorly functional. One trivial explanation for the poor correlation is that aminotriazole resistance is not a direct measurement of the level of transcription. Particularly in the comparisons of the selected -10 elements, differences near and at the initiation site could affect level of transcription or the stability or translatability of the RNA. This potential artifact is unlikely to affect comparisons of the selected -35 elements because the RNA initiation sites in these promoters are likely to be the same due the common -10 element and downstream sequences. Alternative causes for the poor correlation may be that elements were selected that are functioning by different mechanisms, that neighboring sequences are having a varied effect on expression, or that cooperativity exists between segments of the element.

Although a correlation between *in vitro* open complex formation and similarity to the *E.coli* promoter consensus has been established for some elements (14), our results agree with previous experiments indicating that the relationship between promoter sequence and *in vivo* function seems to be more complicated (15,16). The number and types of interactions between σ factors or other proteins involved in the initiation of transcription and the sequences of the promoter elements allow for many potential rate-limiting steps. It has been suggested that the consensus sequences for the -10 and -35 elements might only be involved in part of the overall initiation mechanism and that individual promoters may have different rate-limiting steps (15,16)

DISCUSSION

A rapid method has been presented for defining the sequence requirements of a genetic element. The consensus derived by random selection of *E.coli* promoters is in close agreement with that obtained by wild-type sequence comparisons and mutational analysis.

In comparison to the standard approaches for determining consensus sequences, random selection offers several advantages. Using mutually primed synthesis, large libraries of degenerate DNA can be created with considerable ease allowing for the representation of many examples of functional genetic elements. The relative quality of these selected elements can be assessed in a consistent experimental situation. Random selection can be used even if no sequence data are available to predict the consensus, and the sequences generated are independent of any particular naturally occurring element. They are biased only by the situation or environment from which the sequences were selected. This environment includes the surrounding DNA, the organism, and the conditions in which the organism exists. In contrast, traditional methods require considerable effort to identify an equal number of wild-type elements or to create a large number of mutants of a particular element. Moreover, wild-type elements are embedded in varied sequence contexts which have been subjected to evolutionary selection pressures that may be unrelated to the function of interest.

The ability to obtain functional elements depends on the frequency at which the element occurs in the oligonucleotide population, the size of the library, and the number of molecules that can be examined by the genetic selection or screen. The use of random-sequence DNA is limited to those situations where the frequency of generating a functional element is sufficently high.

However, if the genetic element being studied has a high degree of specificity, functional elements can be obtained if the nucleotide frequencies in the oligonucleotide are biased toward a wild-type sequence that is known to confer the function under study.

Although the generation of functional elements can now be done with minimal effort, the design of the experiment and interpretation of the results are not trivial (1). The compilation of many sequences into a consensus has inherent problems that in some cases can be minimized by the random selection method. One problem involves the effects of neighboring or overlapping sequences that may have an effect on the functional element under selection. Such sequences could directly contribute to the function of an element, or represent other elements that might interact with the element of interest. Random selection allows one to analyze elements in a consistent sequence context such that potential effects of neighboring DNA sequences are more uniform; however, the potential still exists for an effect from a neighboring sequence or for varied effects from overlapping elements in the selected region.

The current definition of a consensus for a genetic element may include sequences that should be classified separately. Distinct proteins, such as σ factors in *E.coli*, may interact with a given element and each of these interactions may require different sequences for optimal function. If the sequences that are used by each mechanism could be defined, and a consensus for each mechanism established, the ability to predict function from sequence would be greatly enhanced. The influence of the environment on a genetic element could be a useful tool if multiple mechanisms are involved in expressing the same function. Using random selection, different elements might be segregated by altering conditions such as temperature, growth media or cell line during the selection to favor a particular mechanism and identify the dependence of a consensus on environmental influences.

The diversity that is created by cloning segments of degenerate oligonucleotides is of a magnitude and type that is not found in nature or by using other conventional mutagenesis techniques. This diversity and the selection from that diversity mimic the evolutionary process but may in some cases have a greater power to generate novel genetic function. A related aspect of this approach is the independence of the generated elements from the evolutionary history of a wild-type element. This history may have imposed constraints on the sequences of wild-type elements that are not directly required for function. The random selection approach allows one to study elements that are optimized for function in the current conditions of the organism.

ACKNOWLEDGEMENTS

We thank Alexander Nussbaum for synthesizing the oligonucleotides used in this work. This work was supported by a grant from the National Institute of Health (GM30186).

REFERENCES

1. Oliphant, A.R & Struhl, K. (1987) *Methods Enzymol.* **155,** 568-582.
2. Kaiser. C.A., Preuss, D., Grisafi, P. & Botstein. (1987) *Science* **235,** 312-317.

3. Ma, J. & Ptashne, M. (1987) *Cell* **51,** 113-139.
4. Hawley D.K. & McClure, W.R. (1983) *Nucl. Acids Res.* **11,** 2237-2255.
5. Harley, C.B. & Reynolds, R.P. (1987) *Nucl. Acids Res.* **15,** 2343-2361.
6. Oliphant, A.R, Nussbaum, A.L. & Struhl, K. (1986) *Gene* **44,** 177-183
7. Struhl, K., Cameron, J.R. & Davis, R.W. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 1471-1475.
8. Sanger, F., Coulson, A.R., Barell, B.G., Smith, A.J. & Roe, BA. (1980) *J. Mol. Biol.* **143,** 161-178.
9. Zoller, M.J. & Smith, M. (1983) *Meth. Enzymol.* **100,** 468-500.
10. Hutchison, C.A., Nordeen, S.K., Vogt, K. & Edgell, M.H. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 710-714.
11. Brennan, M.B. & Struhl, K. (1980). *J. Mol. Biol.* **136,** 333-338.
12. Staden, R. (1984) *Nucl. Acids Res.* **12,** 505-519.
13. Mulligan, M.E., Hawley, D.K., Entriken, R. & McClure, W.R. (1984) *Nucl. Acids Res.* **12,** 789-800.
14. Mulligan, M.E. & McClure, W.R. (1986) *Nucl. Acids Res.* **14,** 109.
15. Deuschle, U., Kammerer, W., Gentz, R. & Bujard, H. (1986) *EMBO J.* **5,** 2987-2994.
16. Brunner, M. & Bujard, H. (1987) *EMBO J.* **6,** 3139-3144.