ble to obtain a large number of mutations. Moreover, in many cases the vast majority of transformants contain an inserted oligonucleotide, thus eliminating the need for a hybridization screen prior to DNA sequence analysis. Unlike other methods that produce mismatches between mutant oligonucleotides and the wild-type sequence, the oligonucleotides described here are cloned as homoduplex molecules. This avoids biases due to differential stability and preferential repair of heteroduplexes, and to screening procedures that depend on mismatch hybridization to distinguish mutants from nonmutants. Most importantly, essentially all possible mutations can be obtained without regard to their phenotypes *in vivo*. Thus, it is possible to determine directly which nucleotides are critical for a particular genetic function and which ones are unimportant.

[35] The Use of Random-Sequence Oligonucleotides for Determining Consensus Sequences

By ARNOLD R OLIPHANT and KEVIN STRUHL

Introduction

In studying the DNA sequences of various genes and organisms it has become evident that similarity in function is associated with similarity in structure. However, as genetic elements conferring similar functions do not generally have identical DNA sequences, their nucleotide requirements are described as a consensus of related sequences. A common and useful means of describing such a consensus is to construct a matrix listing the number of occurrences of all four nucleotides at each position in the consensus. The reasons for determining a consensus sequence are to increase knowledge about the function of interest and to accurately predict biological meaning and functional activity from newly acquired sequence data.

Consensus sequences are often proposed on the basis of comparing a large number of natural DNA sequences that are believed to encode a particular genetic function. Regions of DNA proposed to contain a genetic element are examined for similarities that would not be expected to occur on a random basis. However, it is difficult to show statistical significance unless the sample size is very large, the elements are localized to small regions of DNA, or the proposed sequence occurs very infrequently on a random basis. More importantly, even when statistically significant

homologies are found, the DNA sequences may not be involved in the function of interest.

Another method for defining a consensus sequence is to determine the nucleotide requirements of an individual genetic element. This is accomplished by obtaining a large number of single base pair substitutions within a region of interest and then analyzing their phenotypic effects. In order not to bias the population of mutants, many mutagenesis schemes purposefully avoid the introduction of selective pressure. However, this means that the information is limited by the number of mutations that can be sequenced and analyzed. In addition, as the DNA sequences that are examined are strongly biased by the wild-type sequence, little information is gained concerning related but different sequences that can confer the same function. Thus, although this method yields important information about the role of individual nucleotides for a specific element, many of the complexities of a consensus sequence are not addressed.

We have developed an alternative method that should be useful in defining the sequence requirements of a genetic element (Fig. 1). A collection of recombinant DNA molecules is made in which random or highly degenerate DNA replaces a genetic element of interest. A selection or

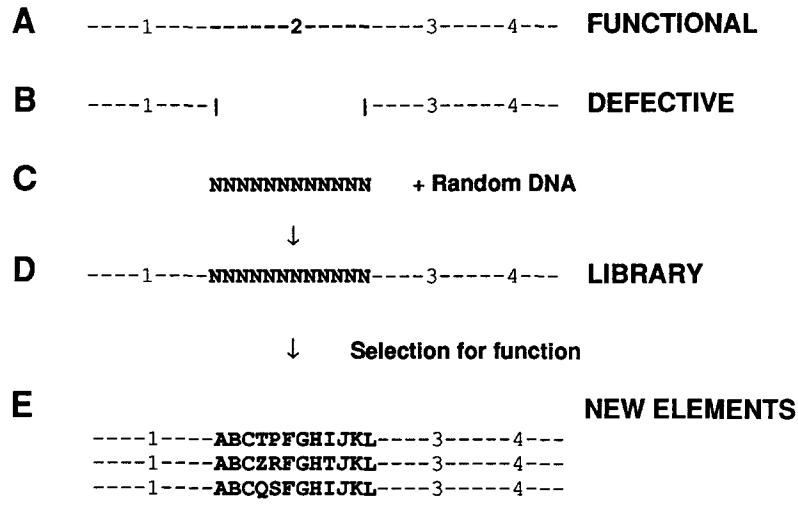

FIG. 1. General method. (A) A group of genetic elements numbered 1–4 conferring a specific function to the organism that is subject to a selection or screen. (B) A vector that is functionally defective because it lacks the genetic element of interest. (C) Double-stranded, random DNA is then substituted in place of the omitted element to form a library of hybrid molecules (D). A selection or screen is used to identify those sequences that confer the function of interest. (E) A comparison of molecules passing the selection defines the consensus for the genetic element.

screen is made to isolate from this collection those sequences that confer an equivalent function. A comparison of DNA sequences that satisfy a particular selection results in a consensus that defines the genetic element. Thus, unlike conventional mutagenesis which uncovers nonfunctional derivatives of a wild-type sequence, the method presented here uses the bias of selective pressure to select functional sequences from random DNA.

Principle of the Method

In principle, the method is applicable for any genetic element that confers a phenotype that is subject to a selection or screen. The first step in determining the consensus sequence for such a genetic element is to construct an appropriate vector (Fig. 1B). The crucial features of the vector are that (1) it lacks the genetic element of interest, (2) it contains all other sequences necessary to pass the selection, and (3) it contains unique restriction endonuclease cleavage sites that can be used to clone DNA segments at the position of the deleted genetic element. In many experiments, including the ones described here, the inserted segments are random DNA sequences that are synthesized chemically by using equal concentrations of all four nucleotide precursors during each addition step (Fig. 1C). Thus, the cloning of random DNA between the restriction sites of the vector generates a library of recombinant molecules in which the genetic element of interest has been replaced by an individual sequence from the original collection of oligonucleotides (Fig. 1D). The library is then introduced into an appropriate organism, and a selection or screen is performed to identify derivatives that confer the desired function. DNA molecules are prepared from derivatives passing the selection in order to determine the nucleotide sequences of the inserted regions (Fig. 1E). Thus, as functional elements are localized to the inserted oligonucleotides, a comparison of sequences results in the consensus.

Besides generating individual DNA sequences that satisfy a selection, this method provides information concerning the specificity of a genetic element. Such specificity is related to the frequency at which the functional element occurs and is a function of the number of positions in the element, the degree to which the nucleotide frequency at each position deviates from randomness, and the amount of function required before a sequence is said to contain a genetic element. For example, a consensus sequence composed of 10 exact nucleotides (expected frequency 10^{-6}) is 1000-fold more specific than a consensus consisting of 5 exact nucleotides (expected frequency 10^{-3}). However, as genetic elements are not usually defined by precise DNA sequences, but rather by a consensus of related sequences, it is difficult to estimate the specificity of a given element

simply by comparing natural sequences. In contrast, the method described here permits an experimental determination of the frequency of a genetic element, and hence its specificity. The specificity of an element depends strongly on the severity of the genetic selection. As the severity of the selection increases, better functioning elements are required to pass the selection. This will result in fewer molecules passing the selection, and the consensus that is derived will be one of higher specificity.

Much of this method would not be possible without the means to generate libraries of highly degenerate DNA. Although random DNA sequences can be generated by utilizing appropriate mixtures of phosphoramidite precursors at each step of the chemical synthesis, methods for cloning such oligonucleotides have not been described. Standard methods[1] are unsuitable because the extreme heterogeneity of the random DNA mixture precludes the availability of a complementary template. To overcome this limitation, we have developed single-strand ligation and mutually primed synthesis methods that facilitate the cloning of oligonucleotides with any degree of degeneracy.[2,3] Single-stranded oligonucleotides with appropriate 5′ and 3′ ends can be ligated directly into vectors containing complementary 5′ and 3′ extensions produced by restriction endonuclease cleavage. However, such a method constrains the enzyme sites which can be used for cloning the oligonucleotide. A more general and efficient method of cloning highly degenerate oligonucleotides is to convert them into double-stranded DNA by including at the 3′ side of the oligonucleotide a palindromic sequence that is recognized by a restriction endonuclease. As shown in Fig. 2B, two oligonucleotide molecules can serve as mutual primers for polymerization by the Klenow enzyme. This produces a double-stranded molecule flanked by the original 5′ restriction sites with the 3′ enzyme site in the center (Fig. 2C). This method does not require the 3′ or 5′ enzyme sites to generate specific 3′ or 5′ extensions, but it limits the 3′ restriction enzyme sequence such that it must be a palindrome.

Design of the Oligonucleotide

Oligomers were synthesized and kindly provided by Alexander Nussbaum using the phosphite triester method on an Applied Biosystems DNA Synthesizer Model 380A.[4,5] For the oligonucleotides having degenerate

[1] M. J. Zoller and M. Smith, this series, Vol. 100, p. 468.
[2] A. R. Oliphant, D. E. Hill, A. L. Nussbaum, and K. Struhl, *Gene* **44,** 177 (1986).
[3] D. E. Hill, A. R. Oliphant, and K. Struhl, this volume [34].
[4] M. D. Matteucci and M. H. Caruthers, *J. Am. Chem. Soc.* **103,** 3185 (1981).
[5] S. L. Beaucage and M. H. Caruthers, *Tetrahedron Lett.* **22,** 1859 (1981).

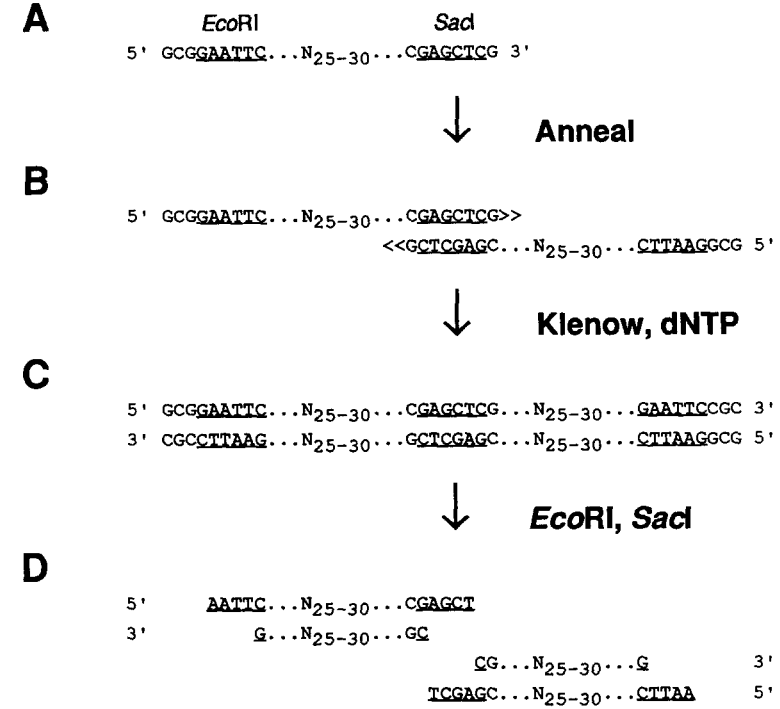

FIG. 2. Mutually primed synthesis. (A) A single-stranded oligonucleotide containing 25–30 bases of an equimolar mixture of all four nucleotide precursors bounded by *Eco*RI and *Sac*I sites; (B) two oligonucleotides annealed at their complementary 3′ ends; (C) double-stranded DNA after extension of the 3′ ends with Klenow and the four nucleoside triphosphates; (D), a pair of double-stranded DNAs suitable for cloning into any *Eco*RI–*Sac*I-cleaved vector. Sequences recognized by *Eco*RI and *Sac*I restriction endonucleases are underlined.

central portions, the procedure was modified by omitting the capping reaction after each of the central steps.[5] This modification improves the yield considerably because oligonucleotides which fail to react at a given step are not irreversibly inactivated and thus can react at subsequent steps; it also results in oligonucleotides that are more heterogeneous in length as compared to conventional syntheses. After detachment and removal of all but the 5′-dimethoxytriphenylmethyl protecting groups, the oligomers were separated from shorter congeners by HPLC chromatography on a Waters C-8 column, using a 40-min linear gradient from 0 to 25% acetonitrile of 0.1 *M* triethylammonium bicarbonate (pH 7.1). The peak containing the trityl chromogen was desalted by flash evaporation *in vacuo* at temperatures below 30° and completely deprotected by treatment with 80% aqueous acetic acid at room temperature for 20 min, followed by flash evaporation.

The sequences chosen for the 5′ and 3′ ends of the oligonucleotide are important for the success of the mutually primed synthesis. The 3′ end must contain a palindrome in order that two corresponding molecules can act as primers for each other. This reaction has been done successfully using palindromes of 6 and 8 bases. If shorter palindromes are used, the stability of the hybrid is greatly reduced. At much longer lengths the chance increases for an intramolecular folding of the palindrome to predominate over the intermolecular priming of two molecules. At the 5′ end, the sequences do not have to be palindromic, nor do they have to be cleavable by a restriction endonuclease because the mutually primed synthesis generates blunt ends suitable for cloning. However, in cases where cleavage at the 5′ site is desired, the reaction is facilitated by having 2–3 additional bases beyond the sequences that are recognized by the enzyme.

If both ends of the oligonucleotide are to contain palindromic restriction enzyme sites, there is a choice as to which site will be used at the 3′ end. In such cases, the enzyme site having the highest GC content is preferred, as it will result in a stronger hybridization between the two molecules. In addition, it is desirable to minimize self-complementarity at the 5′ end because the extension reaction catalyzed by the Klenow enzyme, which lacks the 5′–3′ exonuclease, may be inhibited by hybridization at the 5′ end. This is generally accomplished by choosing nonpalindromic bases as the 2–3 additional nucleotides beyond the restriction endonuclease site. When possible, G and C bases are chosen to increase stability of the double-stranded DNA and enhance cutting efficiency. However, even with these precautions, cleavage with the enzyme recognizing the 5′ site requires a great excess of enzyme.

The DNA sequences composing the heterogeneous central portion of the oligonucleotide are chosen after consideration of a number of factors, including the length and type of consensus being studied and the types of selections available. In many cases, including those described here, the sequences of the central region are essentially random, i.e., equal frequencies for each of the four nucleotides at each position. Thus, a direct comparison of the sequences that pass the selection will be representative of the frequencies of nucleotide use at each position. In practice, however, the DNA sequences of the cloned oligonucleotide regions from the original collection will sometimes display some bias toward individual nucleotides.[6] This bias must be factored out in order to present a useful nucleotide frequency at each position.

This approach can be useful even if no sequence data are available to

[6] G. Zon, K. A. Gallo, C. J. Samson, K.-L. Shao, M. F. Summers, and R. A. Byrd, *Nucleic Acids Res.* **13**, 8181 (1985).

predict the consensus. Moreover, the consensus can be determined independently of natural sequences that normally confer the specific function. The determination of the consensus is biased only by the situation or environment from which the sequences are selected. This environment includes the surrounding DNA and the constraints imposed by the selection scheme. By varying these environmental factors, it should be possible to determine their influences on the consensus sequence.

One limitation of using random-sequence DNA is that it is poorly suited for highly specific consensus sequences. For example, if a genetic element has a specificity that is equivalent to 10 exact bases, the probability at which such a sequence occurs is 10^{-6}. The actual frequency of cloning this sequence depends on the length of the heterogeneous region of the oligonucleotide. If the central portion is the minimal length of 10 bases, the probability is indeed 10^{-6}, but if it is 20 bases, the probability is increased to 10^{-5} because there are 10 possible locations for the sequence on a given oligonucleotide molecule. However, although 1 μg of a 20-base, random-sequence oligonucleotide contains 10^{12} different molecules, the critical factor is the number of molecules that can be examined. This is limited first by the number of *Escherichia coli* transformants that can be obtained with the recombinant molecules. It can be limited further by (1) the number of molecules that can be examined by the genetic selection or screen and/or (2) the transformation efficiency of the organism in which the selection is performed if this is not *E. coli*. For most practical purposes, it is generally difficult to test more than 10^6 molecules for the property of interest. Although these considerations should not affect genetic elements such as promoters and regulatory sequences, they will probably become important in cases where the genetic element is part of a protein-coding region.

Genetic elements with a higher sequence specificity can be examined if wild-type sequences that pass the selection are known. Here the oligonucleotide is made nonrandom by using higher concentrations of a wild-type nucleotide at each step of the synthesis. Biasing the oligonucleotide toward a wild-type sequence increases the probability that a functional sequence will be generated. The degree of bias can be varied by using appropriate concentrations of the nucleotide precursors such that the frequency at which the wild-type sequence occurs can be preset. In the example described above, if the wild-type base were included at 40% and the other three at 20% of the total, then the frequency would be increased to 10^{-4}. This represents a 100-fold increase in the likelihood of finding such a wild-type sequence. Nevertheless, the high mutation frequency (60% per nucleotide) makes it unlikely that the selected sequences will be identical to the wild type.

It is worth noting that the methods described above can be modified for mutagenesis of a specific genetic element. First, an oligonucleotide is made with a heavy wild-type bias such that the majority of the constructs are functional. By using an appropriate selection or screen, nonfunctional sequences can be identified and then sequenced. This method is valuable in determining which nucleotide positions of a longer consensus sequence are most important for function. Mutants deviating from the wild type will occur only in those positions requiring significant uniformity. Second, oligonucleotides can be synthesized such that a deviation from wild type occurs on average once or twice per molecule. Upon cloning, the molecules are then sequenced and assayed individually. An accompanying paper[3] provides an example of this kind of mutagenesis applied to the *his3* regulatory site that confers inducibility during conditions of amino acid starvation.

Cloning Oligonucleotides by Mutually Primed Synthesis[2]

From 2 to 5 μg of oligonucleotide is hybridized at 37° for at least 1 hr in 10 μl of 3× buffer [30 m*M* Tris (pH 7.5), 150 m*M* NaCl, 30 m*M* MgCl, 15 m*M* dithiothreitol, and 0.1 mg/ml gelatin] and is then cooled slowly to room temperature and placed on ice. Deoxynucleotide triphosphates (to a final concentration of 250 μM for each of the four) and 10 μCi of [α-^{32}P]dATP are then added, and the reaction mixture is diluted with water to a final volume of 30 μl. The reaction is initiated by the addition of 5 units of Klenow enzyme, and after incubation at 37° for at least 1 hr the products are phenol extracted and ethanol precipitated. The DNA is resuspended and cleaved with 50 units of the restriction enzyme recognizing the original 5′ end sites. In cases where the 5′ end of the oligonucleotide does not contain such a site, this step is eliminated. After the cleavage reaction, the mixture is phenol extracted, ethanol precipitated, and resuspended in 20 μl of TE (10 m*M* Tris and 1 m*M* EDTA). After electrophoretic separation on a native polyacrylamide gel (6–15%, depending on the size of the oligonucleotide), the desired product is identified by autoradiography, cut out, and extracted overnight by elution at room temperature in 0.5 *M* ammonium acetate and 1 m*M* EDTA, and is finally concentrated by ethanol precipitation. After resuspension in TE, the DNA is cleaved with the restriction enzyme recognizing the central site (originally at the 3′ end of the oligonucleotide), phenol extracted, ethanol precipitated, and resuspended. It should be noted that because of the possibility of denaturation of short oligonucleotides, temperatures below 37° are maintained throughout the procedure.

For cloning purposes, the vector DNA is cleaved with the appropriate

restriction endonucleases, combined with the insert, and ligated with T4 DNA ligase by conventional procedures. As the yield is somewhat variable, the amount of insert to be added to a given amount of vector is determined empirically in order to optimize the ligation reaction.

Determining the Consensus Sequence for an *E. coli* Promoter

As a test of the method, we decided to apply the ideas in such a way that the results could be compared to a previously defined consensus sequence. The −10 and −35 elements of the *E. coli* promoter were chosen because a simple selection system was available and because a consensus sequence has been proposed by comparing numerous wild-type promoter sequences and by analyzing mutations in several individual promoters. The canonical −10 and −35 elements are defined, respectively, by the sequences TATAAT and TTGACA, where the underlined nucleotides indicate those that are most conserved.[7] In addition to the sequence specificity, the distance between the elements is important. The optimal distance is 17 bp, although 16 and 18 bp are acceptable distances.

The basis of the selection scheme is that the expression of a particular structural gene depends on a functional promoter. The structural gene used in these experiments is the yeast *his3* gene, which encodes the histidine biosynthetic enzyme imidazoleglycerol phosphate dehydratase. Although *his3* derives from a eukaryotic organism, its expression permits *E. coli hisB*463 strains, which lack the analogous bacterial enzyme, to grow in the absence of histidine.[8] However, it can be expressed in *E. coli* only if appropriate promoter sequences are located upstream of the structural gene.[9] Thus, by inserting random DNA segments just upstream of the *his3* structural gene, sequences containing a functional promoter can be selected by their ability to permit cells to grow in medium lacking histidine.

Random DNA segments were cloned into an M13 vector at restriction endonuclease cleavage sites located 34 nucleotides upstream of the intact *his3* structural gene.[10] The use of an M13 vector allowed for rapid sequencing of the resulting hybrids.[11] The ligation products were introduced into JM101[11] by standard techniques, and the resulting cells were grown at 37° in broth for 4 hr. The resulting phages were isolated by precipitation in

[7] D. K. Hawley and W. R. McClure, *Nucleic Acids Res.* **11,** 2237 (1983).
[8] K. Struhl, J. R. Cameron, and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* **73,** 1471 (1976).
[9] K. Struhl, D. T. Stinchcomb, and R. W. Davis, *J. Mol. Biol.* **136,** 291 (1980).
[10] K. Struhl, *Nucleic Acids Res.* **13,** 8587 (1985).
[11] J. Messing, this series, Vol. 101, p. 20.

polyethylene glycol and were then infected into *E. coli* KC5 (*hisB*463 F^+) at a multiplicity of 5 to 10 phages per cell. Infected cells were spread on agar plates containing glucose–M9 minimal medium and were incubated for 2 days at 37°. His^+ colonies were observed when the cells were infected with the library of random-sequence DNA but not with the vector. Phages obtained from these colonies were cross-streaked with fresh *E. coli* KC5 cells to assure that cell growth was dependent on infecting phage. After plaque purification, single-stranded phage DNA was prepared and subjected to DNA sequence analysis by the chain termination method.[12] This experiment was performed in three different ways, and the results obtained will be described separately.

Experiment 1

The M13-based vector mp19-Sc5002 was constructed such that random sequence DNA could be inserted between *Eco*RI and *Sac*I sites immediately upstream of the yeast *his3* structural gene (Fig. 3A). The oligonucleotide shown in Fig. 3B, which contained 25–30 bases of random-sequence DNA, was subjected to mutually primed synthesis (Fig. 2) and cloned into the vector. The recombinant molecules were infected into KC5, and His^+ phages were isolated.

The DNA sequences of the inserted region from five phages which produced His^+ colonies are shown in Fig. 3C. In two of these derivatives, Sc5005 and Sc5007, the −10 element is presumably defined by the nucleotides TA from the random DNA followed by the *Eco*RI site; this generates the sequence TAGAAT which closely resembles the consensus. The presumptive −35 sequences, TTGCGC and TTGAGA, show good agreement to the consensus and are spaced 17 bp upstream of the −10 element, the optimal distance. The other three derivatives are not as easily interpreted. The −10 element of Sc5004 may be defined by the CA preceding the *Eco*RI site. In Sc5003 and Sc5006, the −35 sequences are probably parts of the *Sac*I site, being TCGATC and TCGCAT, respectively. In these cases, the probable −10 elements are the sequences TATTTT and TATACT, which are located in the region of random DNA.

These results indicate two major problems of this experimental design. First, the observation that the promoter elements were often composed in part by the restriction endonuclease cleavage sites suggests that the length of the inserted oligonucleotides was not adequate to allow for both the −10 and −35 elements to be present in the random sequence DNA.

[12] F. Sanger, A. R. Coulson, B. G. Barrell, A. J. Smith, and B. A. Roe, *J. Mol. Biol.* **143**, 161 (1980).

A **mp19-Sc5002**

GTCGACTCTAGAGGATCCCCGGGTACCGAGCTCGAATTCCAAAAAT--***his3***--
SalI *XhoI* *BamHI* *KpnI* *SacI* *EcoRI*

B

GCGGAATTC...N_{25-30}...CGAGCTCG
EcoRI *SacI*

C **SELECTION FOR -35 & -10**

Sc5003	GAGCTC GATCAAGGCAACATATTCGTGTATTTTAGG GAATTC
Sc5004	GAGCTA GGACATTAGGAAATCTGTGAAATATCTGGCA GAATTC
Sc5005	GAGCTC GTTGCGCAGACTTACGGGAAGGATTA GAATTC
Sc5006	GAGCTC GCATTTGTTAGAAGCCGGGGCTATACTACTG GAATTC
Sc5007	GAGCTC ATTGAGATGAGACAGCGGTGGTGATA GAATTC

FIG. 3. Experiment 1. (A) DNA sequence of the relevant region in vector mp19-Sc5002. (B) The oligonucleotide used for mutually primed synthesis and subsequent cloning into the vector. (C) DNA sequences of the oligonucleotide regions of five phages that produced His^+ colonies. Presumptive −10 and −35 elements are underlined.

Second, the sequences of Sc5005 and Sc5007 clearly indicate that the *Eco*RI site can provide most of the nucleotides of the genetic element. In this example, the frequency of having the dinucleotide TA precede the *Eco*RI site is 1/16. In this subset of molecules, the probability of creating a functional promoter is artifactually high because all that is necessary is the selection of a functional −35 element from random DNA. In contrast, the remaining molecules require selection of two elements from random DNA, an event of lower probability. Thus, although this design provides useful information concerning the −35 element, the determination of a consensus for the entire promoter is not possible.

Experiment 2

To overcome the problems presented above, the vector mp19-Sc5008 was constructed by treating *Eco*RI-cleaved mp19-Sc5002 DNA with S_1 nuclease and reclosing the blunt ends (Fig. 4A). In addition, a longer oligonucleotide was constructed with about 40–50 bases of random-sequence DNA flanked by *Bam*HI and *Sac*I sites (Fig. 4B). After mutually primed synthesis, a library was generated and phages containing functional promoters were isolated.

Unfortunately, for four out of five examples, the −35 sequence is probably provided entirely by the flanking DNA. Derivatives Sc5010–Sc5013 utilize the TAGAGG from the polylinker as a −35 sequence with

A **mp19-Sc5008**

GTCGACTCTAGAGGATCCCCGGGTACCGAGCTCCAAAAATGAGC--***his3***---
SalI *XhoI* *BamHI* *KpnI* *SacI*

B

GGCGGATCC...N_{50}...CGAGCTCG
BamHI *SacI*

C **SELECTION FOR -35 & -10**

Sc5009	TAGAGGATCC GTTGTTCCCAGATTGCCCGGCCAATATTACATGTAAGATAGGGTT	CGAGCTC
Sc5010	TAGAGGATCC TGAAGTTTACCAGTAGAATAAGCCTGGCAGACGCAGTACTATTAAATTCA	CGAGCTC
Sc5011	TAGAGGATCC ATTTGGGGTCGTGTAGACTAAAGGCTCTGGGCTCTGGCTGGATATGGT	CGAGCTC
Sc5012	TAGAGGATCC TCAACCGGGGGACGTAGAGTGGCCGGCCGGCCTTTACGTTGTGGTAAGGTAG	CGAGCTC
Sc5013	TAGAGGATCC AACGCCCTCTGTACTATAATCTGTGCCTCTAAAGACA	CGAGCTC

FIG. 4. Experiment 2. (A) DNA sequence of the relevant region in vector mp19-Sc5008. (B) The oligonucleotide used for mutually primed synthesis and subsequent cloning into the vector. (C) DNA sequences of the oligonucleotide and flanking regions of five phages that passed the selection. Presumptive −10 and −35 elements are underlined.

TAGAAT, TAGACT, TAGAGT, and TATAAT functioning as the respective −10 sequences. In all of these cases the spacing is 17 or 18 nucleotides. The TAGAAT of Sc5010 may serve as both a −10 and a −35 element because 17 bp downstream from this sequence, there is another potential −10 element, TACTAT; this could create two potential mRNA start sites. Sc5009 probably contains both promoter elements within the random DNA sequences between the *Bam*HI and *Sac*I sites. The sequence TAAGAT is a potential −10 element, and the sequences TTGTTC and TTGGCC are two potential −35 elements.

Thus, despite the experimental modifications, it is clear that the problem of the flanking sequences has not been eliminated. As only one out of the five phages represents a success of the experiment, a consensus sequence for the entire promoter cannot be derived from such data. The reason for the persistence of this problem is essentially as described above. Even when the flanking DNA does not appear to contain a genetic element, certain sequences within the region are likely to be more homologous to the consensus than a random sequence located at the same position. If this is the case, the genetic selection will be biased toward derivatives containing the preferable sequence as one element rather than derivatives in which both elements occur in the random DNA region. The degree of bias will depend on a number of factors, including the particular sequences involved, the length of the random DNA, and the nature of the consensus sequence.

A **mp19-Sc5015**

GGATCCCCGGGTACCGAGCTC<u>TATAAT</u>GGGAATTCCAAAAAT--***his3***--
BamHI *SacI* -10 *EcoRI*

B

GGCGGATCC...N_{25}...CGAGCTCG
BamHI *SacI*

C **SELECTION FOR -35**

Sc5016	GGATCC	GTACCCAT	<u>TTGCCG</u>	CTATACCAGC	CGAGCTC
Sc5017	GGATCC	TCGTACCCA	<u>TTGACG</u>	CTATAGCAAT	CGAGCTC
Sc5018	GGATCC	TGCCTG	<u>TTGTGT</u>	TACTCATTTCG	CGAGCTC
Sc5019	GGATCC	AAGGAAATCA	<u>TTCACA</u>	GGCCATTCAA	CGAGCTC
Sc5020	GGATCC	TCCTGGAAG	<u>TTGAAC</u>	ACTTTCTGTC	CGAGCTC

D **mp19-Sc5014**

GTCGACCATTC<u>TTGACA</u>GGATCCCCGGGTACCGAGCTC--***his3***--
SalI -35 *BamHI* *SacI*

E

GGCGGATCC...N_{25}...CGAGCTCG
BamHI *SacI*

F **SELECTION FOR -10**

Sc5021	GGATCC	CATTCCAGACA	<u>TAGGCT</u>	GTACCATT	CGAGCTC
Sc5022	GGATCC	GGTGACTCATC	<u>TAAGGT</u>	CATAGAT	CGAGCTC
Sc5023	GGATCC	TGGGCGGCTCG	<u>TATATT</u>	GTGAC	CGAGCTC
Sc5024	GGATCC	CGTGGCTGTTT	<u>TACTTT</u>	CATATTTA	CGAGCTC
Sc5025	GGATCC	ACATGCCCCAC	<u>TACCCT</u>	TCAATAAT	CGAGCTC

FIG. 5. Experiment 3. Sequences of the −35 element (top) and −10 element (bottom) are determined separately. For determining the −35 element, the sequence of the vector mp19-Sc5015 (A), which includes a functional −10 element (underlined), the inserted oligonucleotide (B), and five sequences (C) containing functional −35 elements (underlined) are indicated. For determining the −10 element, the sequence of the vector mp19-Sc5014 (D), which includes a functional −35 element (underlined), the inserted oligonucleotide (E), and five sequences (F) containing −10 elements (underlined) are shown.

Experiment 3

Because of these experiences, we decided to study the two components of the *E. coli* promoter separately. Two new vectors were made by using mutually primed synthesis to insert nondegenerate oligonucleotides containing canonical −35 and −10 elements upsteam of the *his3* gene to generate vectors mp19-Sc5014 and mp19-Sc5015 (Fig. 5A,D). Thus, as each vector already contained one element in a defined position, the selection was only for the other element. In this approach, a shorter insert was used in order to ensure that there was space only for the element of interest. A 25-base, random-sequence oligonucleotide (Fig. 5B and E) was subjected to mutually primed synthesis and the products were inserted into the two vectors; the resulting molecules were subjected to the genetic selection.

The results of these experiments are shown in Fig. 5C and F. In the experiment to determine the consensus for the −10 element, all five examples contain sequences within the random portion of the oligonucleotide that strongly resemble the canonical element. Moreover, these −10 elements are located 17 bp away from the −35 element that was fixed in the vector. Similarly, the experiment designed to determine the consensus for the −35 element produced sequences homologous to the canonical element. In four out of five cases, the spacing was 17 bp, and in the remaining case, the spacing was 18 bp.

From these results, it is clear that this approach greatly simplifies the interpretation of the experiment. Since one of the elements is fixed both in position and sequence, interpretation requires determining the position of only one of the elements and its position is greatly restricted by the spacing restraints of a functional promoter. Moreover, the sequence of one element can be studied independently of sequence effects due to the other. Thus, although this method requires that twice the number of phage must be sequenced to analyze the same number of promoters, the quality of the information gained is significantly better. A more complete analysis of experiment 3 will be presented elsewhere.

Comments

The three variations described above have important implications concerning the design of further experiments on other genetic elements. As the methods for this type of experiment are not technically difficult, success is very dependent on proper planning. Our ability to assess the mistakes made and to redesign the experiment was largely due to prior knowledge of the consensus sequences involved. Clearly, a compromise

must be made between the amount of information that can be obtained and the ability to interpret that information. In the first attempts, efforts were made to gain information about many nucleotides of the promoter region in one experiment. The data generated, however, were too complex and vague to be practically useful. The unexpected influences of the surrounding DNA, the spacing of the two elements with respect to each other, and the large regions of random DNA created too many uncontrollable variables. By constraining one of the elements both in position and sequence, the interpretation of the data was made much easier. Thus, a useful guideline for future experiments is to analyze one element at a time.

There are many other genetic functions whose consensus sequences can be determined by this method. The limitations are that (1) a selection or screen must be available for the function under investigation and (2) appropriate restriction endonuclease sites must be available in order to insert the degenerate DNA. The selection or screen need not be carried out *in vivo*. In cases involved in defining the sequences recognized by DNA-binding proteins or by DNA-modifying enzymes, the selection might be done strictly by biochemical means. Where appropriate restriction endonuclease sites can be found or engineered, these methods are applicable to protein-coding regions; here, the problem of localizing the functional element within the degenerate region will not be a factor.

In interpreting the results, one potential complicating factor is worth mentioning. If more than one type of sequence can qualify under the selection, it may be difficult to identify the separate types from all of the sequences obtained. In some cases, these might be discerned by altering the conditions of the experiment to selectively favor one of the sequence types. In general, however, a well-designed experiment should permit the determination of a consensus sequence.

[36] Computer Programs for Analyzing DNA and Protein Sequences

By F. I. LEWITTER and W. P. RINDONE

Introduction

Techniques developed over the last 15 years have enabled scientists to rapidly determine the nucleotide sequences of a large variety of nucleic acid molecules. These sequences, in turn, have enabled the inference of