Volume 17 Number 2 1989 Nucleic Acids Research

# Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique

Thomas D.Schneider* and Gary D.Stormo[1]

National Cancer Institute, Frederick Cancer Research Facility, Laboratory of Mathematical Biology, PO Box B, Frederick, MD 21701 and [1]Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309, USA

Received September 30, 1988; Revised and Accepted December 7, 1988

ABSTRACT

**In our previous analysis of the information at binding sites on nucleic acids, we found that most of the sites examined contain the amount of information expected from their frequency in the genome. The sequences at bacteriophage T7 promoters are an exception, because they are far more conserved (35 bits of information content) than should be necessary to distinguish them from the background of the *Escherichia coli* genome (17 bits). To determine the information actually used by the T7 RNA polymerase, promoters were chemically synthesized with many variations and those that function well in an *in vivo* assay were sequenced. Our analysis shows that the polymerase uses 18 bits of information, so the sequences at phage genomic promoters have significantly more information than the polymerase needs. The excess may represent the binding site of another protein.**

## INTRODUCTION

One of the early proteins synthesized after bacteriophage T7 infects *Escherichia coli* is a new RNA polymerase (1,2,3). This polymerase binds to a set of promoters on the T7 genome to initiate the major transcripts for middle and late genes (4). The DNA patterns at the promoters are unusual in that they are highly conserved (5,6,7). The extent of this sequence conservation can be quantitated by finding out how many bits of information are needed to describe the sequence patterns(8,9).

One bit distinguishes between two equally likely things and can specify, for example, that a purine (as opposed to a pyrimidine) is always found at a particular position in a site. Two bits specify exactly one base. When the frequencies of bases are not as simple as these examples, we can use the methods of information theory(8,10,11) to calculate the average amount of information needed to specify the pattern. We can also calculate the minimum information required to locate a site from the number of sites and the size of the genome.

Schneider *et al.*(9) found that the sequences at T7 promoters contain about 35 bits of information, approximately twice the 17 bits expected from their frequency on DNA during an infection. In other cases the information in the pattern of a binding site is just sufficient for the site to be found, so the sequences at T7's promoters have roughly twice the amount of information needed, in comparison to other binding sites. Is the excess information used by the polymerase? This paper demonstrates that the conserved sequences at phage genome promoters do indeed carry information that is unnecessary for transcription *in vivo*.

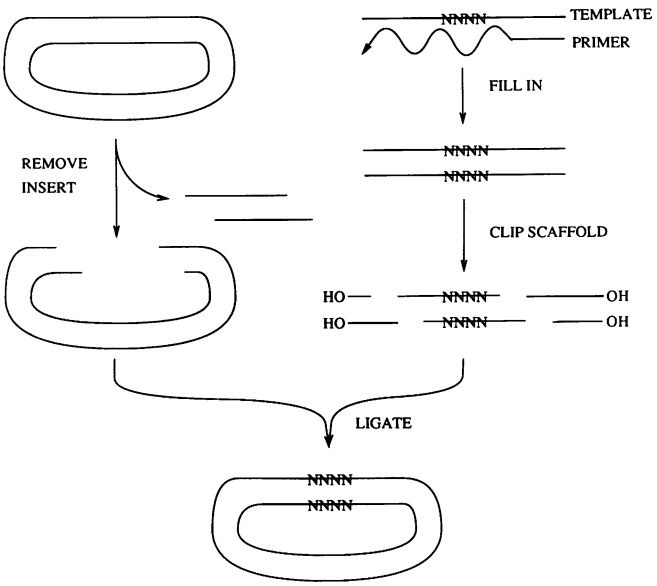


Figure 1: Synthesis and Cloning of Randomized Double-Stranded DNA.

Information that the polymerase does not use should be removable. Therefore the general plan for this experiment was to synthesize and clone a genomic T7 promoter pattern with many random mutations. A set of those that still have strong promoter function *in vivo* were then sequenced to determine the amount of information needed by the polymerase for initiating transcription. This approach is capable of rapidly characterizing the important functional components of any reasonably small nucleic-acid or protein region for which a simple assay is available.

## MATERIALS AND METHODS

### General Scheme for the Synthesis and Cloning of Randomized Double-Stranded DNA

See Figure 1. T7 promoters were chemically synthesized using the DNAs described in Figure 2. At each position to be randomized, a nucleotide mix containing all 4 bases was added to the growing DNA chains so that one cannot predict exactly which base will appear at that position in a clone. The nucleotides were in unequal ratios so that at each position the base corresponding to the wild type $\phi 10$ promoter had an 85% probability of being incorporated, while the other three bases had a 5% probability each. The synthetic single-stranded DNA fragments were converted to double-stranded 'randomized' DNA before being cloned. To allow this, a disposable 'annealing site' had been synthesized on the 3′ side of the random 'template' DNA. A primer was annealed to the template at the annealing site and Klenow polymerase was used to generate double-stranded DNA. Cleavage by two different restriction enzymes releases the randomized double-stranded DNA from the surrounding "scaffold" DNA. The small "scaffold" fragments, both of which have an end without phosphates, have never been observed in the sequences of cloned random DNAs (though no attempt was made to remove them); presumably those that ligate to the vehicle

```
                                    0
        EcoRI                --- ------ ---  HindIII
5' ATA GAATTC aaattaatacgactcactatagggaga AAGCTT CGCCGACCCGTCTG 3'
                                                 GCGGCTGGGCAGAC 5'
   T7-W template, 56 bases                       DESP primer, 14 bases
```

Figure 2: Single Stranded DNAs used to make "randomized" double-stranded DNA. Two single-stranded DNAs were synthesized on an Applied Biosystems model 380A DNA synthesizer using phosphoramidite chemistry (42), and purified by gel electrophoresis. The DESP primer anneals on the 3' side of T7-W DNA. The sequence of T7 promoter $\phi$10 is the phage consensus and served as a wild type control. The symmetric element (9) is indicated by dashes. Coordinate zero (0) is the first base transcribed; transcription proceeds to the right. Small case letters indicate that the base shown is 85% likely and the other 3 bases are 5% likely in that position. The randomized bases were made by adding one part of an equimolar mixture of the 4 bases to 4 parts of the "wild type" base. The complete "wild type" sequence should appear at a frequency of $0.85^{27} = 1\%$. In the mixture there should be an average of 4 changes per sequence.

are lost after transformation into cells. The use of a specific template/primer combination with disposable ends avoids the design difficulties and limitations of earlier methods for constructing randomized DNA (12,13,14,15,16,17).

Construction and Cloning of Specifically Randomized Double-Stranded DNA

This is a general scheme for cloning specifically randomized DNAs that gives a high yield of independent clones.

270 pmol of DESP primer (Figure 2) were mixed with 240 pmol of T7-W template in nick translation buffer(18), heated to 70°C for 10 minutes and slowly cooled to room temperature. 37.5 units of Klenow DNA polymerase (Boehringer Mannheim) were added for a final volume of 180$\mu$l containing 333$\mu$M dNTP's and the reaction was incubated at 37°C for 30 minutes. The mixture was phenol extracted, ethanol precipitated, resuspended and digested with both *Eco*RI and *Hin*dIII (New England Biolabs) in high salt buffer(19). The fragments were phenol extracted, ethanol precipitated and resuspended in 20 mM tris, 1 mM EDTA (TE). Pilot constructions were monitored by labeling with $^{32}$P and displaying on an 8% sequencing gel.

The 1904 base pair *Eco*RI - *Hin*dIII fragment from the b region of $\lambda$ (20) (New England Biolabs $\lambda$ *Hin*dIII sizing standard) was cloned into pKC7 (21,18) to create plasmid pTS36, which is a ColE1 plasmid with kanamycin and ampicillin resistance genes. 500 ml of the bacterial strain MC1061(22) containing pTS36 was grown, and the plasmid was isolated(23) and cesium purified(18). The DNA was cut with *Eco*RI and *Hin*dIII and electrophoresed on a horizontal 1% agarose gel. The 1904bp fragment allowed complete separation of the single- from the double- digest products so that after cloning all transformants contained randomized DNA. The pure DNA piece containing $amp^R$ $kan^R$ and ColE1 markers flanked by *Eco*RI and *Hin*dIII sites was excised and electroeluted from the gel fragments (2.5 h, 2.5 watt) in TBE (10x TBE is 121.1g Tris-base, 60.5g boric acid, 7.44g $Na_2EDTA$ per liter, pH ~8.3) in an ISCO electroelution cup, then ethanol precipitated and resuspended in TE. 1.6 pmol of vector DNA was ligated to 169 pmol of the double-stranded random DNAs with 2000 units of T4 DNA ligase (New England Biolabs) in 300$\mu$l ligation buffer (18) overnight at 17.5°C, then precipitated with 0.4 volumes of $NH_4Ac$ and 2 volumes of NaCl-$H_2O$ saturated isopropanol and resuspended in 10$\mu$l TE to reduce the volume. The DNA was transformed into 1 ml of freshly prepared competent MC1061 cells(24). The

cells were allowed to recover in 20 ml H Broth(25) by shaking at 37°C for 1 hour. A pilot experiment had shown that at this time the cells are transformed, but have not yet begun to grow exponentially. Therefore by titering the culture at 1 hour, we determined that there were 6100 independent clones of T7-W promoters. After titering, 100 ml of H Broth containing kanamycin (50 $\mu$g/ml) and ampicillin (50 $\mu$g/ml) was added and the culture was shaken overnight at 37°C. Plasmid DNA was isolated(23) from the overnight culture and retransformed into BL21/DE3 (26) to obtain 70,000 independent clones (each clone was transferred 11 $\pm$ 3 times). The two step procedure was used because BL21/DE3 is about 50 fold less transformable than MC1061.

Screening for Functional T7 Promoter Variants

In BL21/DE3, the gene for T7 RNA polymerase is under control of the *lac*UV5 promoter, and may be induced by adding IPTG (isopropyl-$\beta$-D-thiogalactopyranoside) to the media. When a T7 promoter is present on a high copy number plasmid in BL21/DE3, IPTG induction kills the cell, probably because the polymerases can transcribe entirely around each copy of the plasmid(27), depleting ribonucleoside triphosphates(3,26). In this work, cell death was used as an assay for the presence of strong T7 promoters(26). Wild type T7 promoter-containing plasmids transform BL21/DE3 as well as (90 $\pm$ 10%) plasmids without a promoter when the cells are grown in non-inducing conditions.

$\beta$-lactamase from the ampicillin resistance gene in pKC7 destroys ampicillin, allowing bacteria that have lost the plasmid to grow next to those that still have it. Therefore carbenicillin (Sigma), which reduces this effect(28), was used as a replacement for ampicillin. Kanamycin, which is not destroyed, was also used. High drug and IPTG concentrations were used to reduce the number of these 'satellite' bacteria in toothpick stabs and to fully induce the T7 promoters. Bacterial colonies of BL21/DE3 carrying wild type promoters cannot grow when toothpicked to EHA plates(25) containing 10mM IPTG(29), 0.5 mg/ml kanamycin(18) and 2 mg/ml carbenicillin (KICAR plates), but can grow on plates that contain only kanamycin and carbenicillin (CARK plates).

200 clones were screened by toothpicking to KICAR and CARK plates. After incubation for two days at 37°C the clones were scored for the absence of growth on KICAR and growth on CARK. Approximately 30% of the clones screened by this method could not grow on KICAR. These were streak-purified, stored and titered for growth on CARK and KICAR.

Sequencing of T7 Promoter Variants

Double-stranded plasmid DNA was prepared for each IPTG sensitive clone and sequenced using the two primers shown in Figure 3.

The dideoxy termination method was used for DNA sequencing (30,31), with the following modifications. 10 ml of cells grown overnight were pelleted, resuspended by vortexing in 100$\mu$l lysozyme buffer(18), and transferred to a microfuge tube containing 100$\mu$l of 40 to 50 mg/ml lysozyme in lysozyme buffer. The DNA was extracted(23,18) and resuspended in 20$\mu$l TE. The DNA was denatured by mixing 7.9$\mu$l into 2.1$\mu$l of 2M NaOH and incubating for 5 minutes at room temperature. 75$\mu$l of 95% ethanol and 9$\mu$l 1M NaOAc pH 4.5 were added and the DNA was pelleted in a cold-room microfuge for 15 minutes. The pellet was carefully rinsed twice by adding 200$\mu$l of 80% ethanol, and spinning 5 minutes. The pellet was dried. 1.5$\mu$l (0.75 pmol) of 5'$^{32}$P labeled primer was diluted in 7.5$\mu$l RT buffer (50mM Tris-Cl pH 8.6, 60mM NaCl, 6mM $Mg(OAc)_2$, 10mM DTT) and this was used to resuspend the DNA pellet. The annealed primer and DNA were immediately placed on ice and held there until 2$\mu$l was aliquoted to each of 4 tubes containing 2$\mu$l of AMV reverse

```
                                                    0
upstream primer-->          EcoRI            --- ------ ---  HindIII
5' GTATCACGAGGCCCTttcgtcttcaa gaattc aaattaatacgactcactatagggaga aagctt cacgctgccgcaagcactcagggcgcaagggctgcta 3'
3' catagtgctccgggaaagcagaagtt cttaag tttaattatgctgagtgatatccctct ttcgaa gtgcgacggcgttcgtgagtcccGCGCGTTCCCGACGAT 5'
                                                                                    <-downstream primer
```

Figure 3: Sequence surrounding the cloning region in the pTS385 ($\phi$10) promoter. The two primers are shown capitalized as part of the sequence, with arrows to indicate the direction of reading by dideoxy sequencing. The upstream primer was #1204 from New England Biolabs(43) and the downstream primer, KC7, was synthesized.

transcriptase (1.2 units, Life Sciences), a ddNTP (0.24 mM) and the 4 dNTPs (0.45 mM each) in RT. Reactions were at 48°C for at least 30 minutes. 8% acrylamide gels were used to analyze the sequence. Certain sequences from the KC7 primer have 'bands-in-all-lanes' artifacts in the region 2 to 4 of the promoter, while primer #1204 gives minimal artifacts. The restriction sites used for cloning were uniformly correct, so it was not necessary to screen for their presence.

Since only some clones were sequenced, it is important to show that the chemical synthesis was approximately as planned. DNA from the entire mixture of clones in BL21/DE3 was sequenced with both primers and compared to the wild type sequence pTS385. This experiment confirmed that the randomized positions of T7-W DNA did have bases other than the predominant base (data not shown).

Calculating Information in Randomized Sequences

The experimental protocol described above generates many variations of a binding site. Because these variations can have more than one mutation, the amount of information we can gather per sequence is larger than can be obtained from single mutations. From these data we must determine how much information the polymerase used to select the functional sites. Our first simplifying assumption is that the effects of bases are independent. In other analyses we have shown this to be a reasonable first approximation (32,17). The essential idea, then, is to treat the polymerase as a simple chemical 'black box', which has a binding constant for each base at each position in the sequence. The chemical synthesis generates a set of 'input' frequencies where one base is predominant, namely 0.85, 0.05, 0.05, 0.05. The polymerase then selects binding sites and we determine the sequences that function. Thus we have two tables that define the experiment: an input table (0.85 etc.) and an output table (left hand side of Figure 5) and we want to determine the amount of information that the polymerase uses to select promoter sequences from the equiprobable-base sequences it is exposed to in the cell. To do this, we must recast the output table *as if* the experiment had been done with equiprobable input frequencies. This 'normalizes' the table.

Let the frequencies of bases $B(= A, C, G, T)$ at positions $L$ in the sequences input to the experiment be $f_i(B, L)$, and let $f_o(B, L)$ be the corresponding output frequencies from the experiment.

We can model how the polymerase recognizes sites by a simple linear transformation:

$$f_o(B, L) = \rho(B, L) f_i(B, L). \tag{1}$$

The $\rho(B, L)$ are relative binding constants in the sense that $\rho(A, L)/\rho(C, L)$ is the ratio of the binding constants $K(A, L)/K(C, L)$ for bases $A$ and $C$. (This is shown in the section on 'Black Box Model' below.) Having determined the $\rho(B, L)$ in one experiment, we expect that these ratios will be the same in another experiment (indicated by the prime symbol) in

which the polymerase and assay conditions are the same, but the input frequencies differ:

$$f'_o(B,L) = \frac{f'_i(B,L)\rho(B,L)}{\sum_{B=A}^{T}(f'_i(B,L)\rho(B,L))}. \quad (2)$$

(Division by the sum assures us that $\sum_{B=A}^{T} f'_o(B,L) = 1.$)

The experiment gives us an amount of information

$$I(L) = \sum_{B=A}^{T} f_o(B,L) log_2(f_o(B,L)/f_i(B,L)) \quad (3)$$

(33), while the information we would get from an experiment with different input frequencies, $f'_i(B,L)$ is:

$$I'(L) = \sum_{B=A}^{T} f'_o(B,L) log_2(f'_o(B,L)/f'_i(B,L)). \quad (4)$$

Procedure for Normalizing an Experiment

The normalizing procedure is carried out as follows. First determine the input table, $f_i(B,L)$. In this experiment we usually assume that these frequencies are (0.85, 0.05, 0.05, 0.05), although by increasing the sequencing effort or by mass sequencing (assuming no artifacts) they could be determined directly. The output frequencies, $f_o(B,L)$ are found from the table (Figure 5) by dividing by the number of bases at each position. The next step is to determine the binding constant ratio table, using equation [1]:

$$\rho(B,L) = \frac{f_o(B,L)}{f_i(B,L)}. \quad (5)$$

Now, we want to normalize to equiprobable inputs, so we set $f'_i(B,L) = \frac{1}{4}$, and from equation [2] we find the normalized output frequencies, $f'_o(B,L)$. We therefore have the two tables necessary to calculate the normalized experimental information from equation [4].

Chemical 'Black Box' Model of Recognizers

This derivation shows that $\rho(B_1)/\rho(B_2)$ is the ratio of binding constants $K(B_1)/K(B_2)$(34). Assume that the enzyme, $E$ recognizes each base, $B$ independently from all other bases and that the degree of binding is determined by chemical constants. Then at a particular position in the site,

$$E + B \rightleftharpoons EB, \quad (6)$$

and we have 4 binding constants to the 4 bases:

$$K(B) = \frac{[EB]}{[E][B]}. \quad (7)$$

The input and output frequencies are simply:

$$f_i(B) = \frac{[B]}{\sum[B]} \quad \text{and} \quad f_o(B) = \frac{[EB]}{\sum[EB]}. \quad (8)$$

Then the ratio $\rho(B)$ is

$$\rho(B) = \frac{f_o(B)}{f_i(B)} = \frac{[EB]}{\sum[EB]} \cdot \frac{\sum[B]}{[B]} = \frac{\sum[B]}{\sum[EB]} \cdot K(B) \cdot [E] \quad (9)$$

so

$$\rho(B_1)/\rho(B_2) = K(B_1)/K(B_2). \quad (10)$$

Standard Deviation of Information

Because of the normalization procedure, we do not know how to calculate a standard deviation for the information, and repeating the entire experiment a sufficient number of times is far too laborious. Fortunately, an upper bound can be determined empirically. The data set consists of 53 sequences, which we assume form an unbiased sample of the functional promoters. We further assume that the randomized synthesis was unbiased, as suggested by mass sequencing. Under these circumstances, subsets of the 53 sequences should give us an idea of how the information measure varies with sample size.

There are 53 subsets in which one sequence has been removed. The information of each of these subsets can be calculated, and it is 18.4±0.3 bits. There are many more combinations possible for smaller subsets, so we can only sample a few of them. (However, the smallest subset has only 1 sequence. In this case the information is 54±0 bits because no variation is possible, and there are 27 randomized positions with 2 bits per base. This is a form of the small sampling error described in the appendix of Schneider *et al* (9). That is, smaller samples will tend to give larger information measures. Conversely, if more then 53 sequences were included, the information measure would tend to be smaller than 18 bits for this reason.)

The largest variation of the subsets occurs with those containing about 13 sequences. In a sample of 530 subsets, these have 34±2.2 bits. Subsets containing fewer than 13 elements have less variation than ±2 bits because of the small sampling effect. (A single sequence has no variation whatsoever.) Although the small sampling effect plays less of a role for larger sets, those sets containing more than 13 sequences also have less variation than ±2 bits. This is because—according to the central limit theorem—the $f_o(B, L)$ frequencies approach the population frequencies for large samples. Thus if the set of 53 sequences is an unbiased sample then the standard deviation of samples this size should not be larger than 2 bits. By using ±2 bits as our bound, we overestimate the variation.

This variation is due to sample size, and does not account for our uncertainty in the input distribution, which could have been different from (0.85, 0.05, 0.05, 0.05). However, if we use (0.91, 0.03, 0.03, 0.03), we obtain 18.3 ± 2.3 and (0.70, 0.10, 0.10, 0.10) gives 21.7±2.4, so even a drastic correction to the input frequencies would not significantly alter our conclusions.

Positions that are completely conserved in the experiment are normalized to 2 bits. If these positions are in fact not conserved in the parent distribution (*i.e.* the collection of all functional promoters), then the information measured should be below 2 bits. This effect also means that we tend to overestimate the amount of information.

## RESULTS

We synthesized and cloned 6100 T7 promoter variants. 200 of these were tested for their function *in vivo*, and those strong enough to kill the cells were purified and sequenced. Of the 58 clones sequenced, 5 had multiple promoter inserts and 3 were wild type. The expected number of wild type clones from the collection of 200 screened clones, assuming 85% probability of the wild type base at each position, is 2.5 ($= 200 \times 0.85^{27}$). The sequences of the single-insert, non-wild type promoters are shown in Figure 4. The functional sequences (including wild type) are tabulated in Figure 5.

We applied the $\chi^2$ test(35) to several hypotheses about the experimental input frequen-

```
------------------                  ++
22222222111111111--------- +++++++++11
765432109876543210987654321012345678901
GAATTCaaattaatacgactcactatagggagaAAGCTT
 EcoRI           --- ------ --- HindIII  PLASMID PRIMER
      c..a.t.........c......c..          pTS368  both
1     c..c.......................          pTS361  upstream
1     c..c.......................          pTS362  both
      c...a.........c......c.            pTS351  both
      c.............c...........          pTS380  upstream
      g...a.....................          pTS352  upstream
      g.....c...................          pTS356  upstream
      g......a.........g........          pTS360  upstream
      g.........................          pTS386  upstream
      t..................g...tt.          pTS358  both
      .c....t........t..........          pTS339  downstream
      .c..............t........          pTS363  downstream
      .g.......................          pTS378  upstream
      .t...g..t................t          pTS341  downstream
      ..ga..........g.......acc.          pTS348  downstream
      ..gg..................cc..c          pTS349  downstream
2     ..g.......a.......c.g......          pTS338  both
2     ..g.......a.......c.g......          pTS347  both
3     ..g.......a.........c..c...          pTS337  downstream
3     ..g.......a.........c..c...          pTS353  downstream
4     ..g.......a.........t......          pTS364  downstream
4     ..g.......a.........t......          pTS382  upstream
      ..g..............a.........          pTS357  upstream
      ..g..............c.......t.          pTS367  downstream
      ..g...................c...          pTS336  downstream
     x..g.......................          pTS344  downstream
      ...ag.....................          pTS333  downstream
      ...g................c.t....          pTS350  downstream
      ....a......t..............          pTS388  upstream
      ....c...t..t..............          pTS355  both
      .....g...................a.          pTS372  upstream
      .....g...................c.          pTS387  upstream
      .....t..............a..x..          pTS384  upstream
      .......a............g......          pTS342  both
      ........c.........g.....c..          pTS332  upstream
      ........g.........ca....g..          pTS373  upstream
      ........t.........t......a.          pTS331  upstream
      ........t..............t.t.          pTS379  upstream
      ..........ag..........a.g..          pTS365  both
      ...........c......g.......t          pTS376  upstream
      ...........c......t.....c..          pTS377  upstream
      ...........t.......a...a...          pTS345  both
      ...............g.....a....c          pTS346  downstream
      ................t...........          pTS369  both
      ..................c.........          pTS371  downstream
      ..................g..g....c.          pTS335  upstream
      ...................g......tc          pTS381  upstream
      .....................c....a.          pTS340  both
      .....................t.....          pTS343  downstream
      ........................g             pTS354  both
```

Figure 4: Sequences of Strong T7 promoters.
The coordinate numbers at the top are to be read vertically. Below this is the sequence of the $\phi 10$ promoter, surrounded by the two restriction sites used for cloning. The location of the symmetry element is indicated by '-'. On both a wild type clone and the mixture of 6100 T7-W clones, transcription starts mostly at base zero (dideoxy sequencing of RNA using primer KC7 showed an intense band at this position, data not shown). The sequences of 50 strong promoters isolated from the T7-W DNA mixture are shown. The sequences of three wild type sequences (pTS359-downstream, pTS370-downstream, pTS385-upstream) are not included. Numbers on the left-hand side indicate duplicate sequences. Regions not randomized are left blank. Changes from the $\phi 10$ sequence are indicated by the differing base. An 'x' means that that base was not determined. Each sequence is followed by the plasmid name and the primer(s) that were used to obtain the sequence. The sequences are sorted alphabetically.

| | | Experiment T7-W | | | | Phage Promoters | | | |
|---|---|---|---|---|---|---|---|---|---|
| L | $\phi$10 | A | C | G | T | A | C | G | T |
| -21 | a | 43 | 5 | 4 | 1 | 8 | 3 | 4 | 2 |
| -20 | a | 49 | 2 | 1 | 1 | 12 | 0 | 2 | 3 |
| -19 | a | 41 | 0 | 12 | 0 | 10 | 1 | 3 | 3 |
| -18 | t | 3 | 2 | 2 | 46 | 3 | 1 | 3 | 10 |
| -17 | t | 3 | 1 | 1 | 48 | 0 | 3 | 0 | 14 |
| -16 | a | 48 | 0 | 3 | 2 | 16 | 0 | 0 | 1 |
| -15 | a | 51 | 1 | 0 | 1 | 17 | 0 | 0 | 0 |
| -14 | t | 2 | 0 | 0 | 51 | 0 | 0 | 0 | 17 |
| -13 | a | 47 | 1 | 1 | 4 | 14 | 1 | 0 | 2 |
| -12 | c | 0 | 53 | 0 | 0 | 0 | 16 | 1 | 0 |
| -11 | g | 7 | 0 | 46 | 0 | 2 | 0 | 15 | 0 |
| -10 | a | 47 | 2 | 1 | 3 | 17 | 0 | 0 | 0 |
| -9 | **c** | 0 | 53 | 0 | 0 | 0 | 17 | 0 | 0 |
| -8 | **t** | 0 | 0 | 0 | 53 | 0 | 0 | 0 | 17 |
| -7 | **c** | 0 | 53 | 0 | 0 | 0 | 17 | 0 | 0 |
| -6 | a | 48 | 1 | 2 | 2 | 17 | 0 | 0 | 0 |
| -5 | **c** | 0 | 53 | 0 | 0 | 0 | 16 | 1 | 0 |
| -4 | **t** | 1 | 3 | 1 | 48 | 0 | 0 | 0 | 17 |
| -3 | **a** | 42 | 4 | 4 | 3 | 17 | 0 | 0 | 0 |
| -2 | **t** | 2 | 0 | 0 | 51 | 5 | 0 | 0 | 12 |
| -1 | **a** | 42 | 4 | 5 | 2 | 16 | 0 | 0 | 1 |
| 0 | **g** | 2 | 0 | 50 | 1 | 2 | 0 | 15 | 0 |
| 1 | g | 1 | 1 | 50 | 1 | 2 | 0 | 15 | 0 |
| 2 | **g** | 2 | 4 | 46 | 1 | 5 | 0 | 12 | 0 |
| 3 | **a** | 45 | 4 | 2 | 1 | 10 | 0 | 7 | 0 |
| 4 | **g** | 3 | 4 | 42 | 4 | 5 | 1 | 11 | 0 |
| 5 | a | 47 | 3 | 1 | 2 | 13 | 2 | 0 | 2 |

Figure 5: Tabulated Results for Strong T7 Promoters.
The numbers of each kind of base at each position (L) are given for both the experiment described here and for sequences at promoters in wild type T7 phage(4). The symmetry element is indicated by boxed letters.

cies $f_i(B, L)$ to explain the raw experimental output frequencies $f_o(B, L)$ in Figure 5. The simplest hypothesis is that the expected input frequencies (0.85, 0.05, 0.05, 0.05) explain the observed $f_o(B, L)$ data (completely saturated model). This hypothesis can easily be rejected ($\chi^2 = 156$, degrees of freedom (df) $= 81$, $p = 1.3 \times 10^{-6}$), which shows that the data contain significant patterns. The worst possible hypothesis to explain the data, without supposing selection by the polymerase, is that the $f_o(B, L)$ data can be explained by a set of 4 chemical synthesis bottles, each with its own proportions, given by $f_o(B, L)$. This hypothesis is also rejected ($\chi^2 = 104$, df $= 69$, $p = 4.3 \times 10^{-3}$). These tests suggest that the sequences selected by their activity as functional promoters differ significantly from the input sequences, even though the input sequences were biased toward the "consensus sequence" of phage promoters.

Position -21 was included in the randomization because it carries little or no information in the phage promoters. It acts as a control for the randomization. By the binomial distribution ($n = 53$, $p = 0.85$) we expect $45.1 \pm 2.6$ A's. 43 A's were observed. This result and the mass sequencing of the 6100 clones indicate that the sample of active promoters comes from essentially all possible sequences in the region. In contrast, positions -12, -9, -8, -7, and -5 carry a large amount of phage promoter information (as judged by the high frequency of one base at each of these positions), and little variation from this pattern

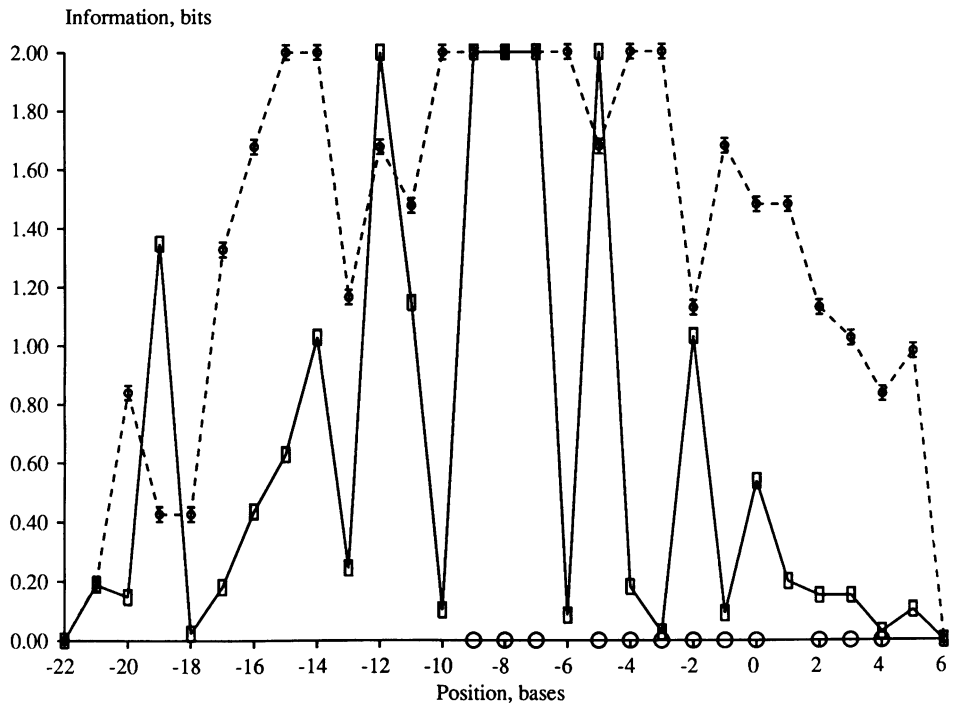


Figure 6: Information at T7 Promoters.
The abscissa is the position along the sequence, in bases, while the ordinate shows the amount of information at each position in bits. Dashed lines and Φ: information at the patterns of phage genome promoters; solid lines and □ : normalized information in strong T7 promoters in the T7-W experiment ($I'(L)$). The symmetry element is indicated by circles on the abscissa.

was found in this experiment, nor in 10 sequences of an earlier pilot experiment (data not shown). Similar results on the important regions of the promoter were obtained *in vitro* by Chapman and Burgess (36) and Milligan *et al.*(37).

We can compare these *in vivo* results with the *in vitro* data of Chapman and Burgess (36) to roughly determine the sensitivity of the cell death assay. Chapman and Burgess' mutations fall into two classes, those with less than 50% of wild type activity (*in vitro* on supercoiled DNA at low salt) at positions -9, -8 and -7 and those with 70% or more of wild type activity clustered in positions -6 to 0. This corresponds neatly with our results, since no promoter variants were observed in positions -9, -8 and -7 (even though they were present in the original cloning population) while many variants were found in -6 to 0. This indicates that T7 promoters with 50% or more activity are detected by the cell death assay. Furthermore, a single T at position -5 creates a promoter with a phenotype of tiny colonies rather than cell death (data not shown). This promoter is 30% as active as wild type promoters(36) and this phenotype was eliminated during the promoter screen because of its slight growth when induced. Therefore we can be confident that the set of promoters shown in Figure 4 are all stronger than 30% of wild type activity, and that most

of them are stronger than 50% of wild type activity.

The normalized information curve, $I'(L)$, is shown in Figure 6. The 'area' under this curve is $18 \pm 2$ bits per sequence.

## DISCUSSION

### Overview of the Experiment

In a number of genetic systems in *E. coli*, we previously found(9) that the information—measured in bits per site—needed to describe the patterns in DNA or RNA at a binding site is reasonably well predicted from the minimum information required to locate the sites in the genome. Thus binding sites have just about enough information for them to be located. At the same time, we also found that T7 promoters present a strong exception to this rule: their sequences apparently have twice as much information as one would predict is necessary. The issue at stake, then, is whether information theory can make sensible predictions of the information in binding sites.

A simple explanation of these data is that the polymerase does not use all of the information present at the phage genome binding sites. If this were true, we should be able to generate T7 promoter variants that still function strongly even though half of the genomic information has been removed.

We therefore designed a cloning scheme which would generate many T7 promoter variants (Figure 1). The method is completely general and has allowed us to create as many as 24,000 independent clones. Although this is 20 times less than the number required to fully saturate the site ($27^4 = 5 \times 10^5$) it is far more than the number of individuals we are ever likely to sequence. The only design requirement is that the randomized region must be between two unique restriction sites on a piece of DNA short enough to be synthesized. We are not limited by the rate of collecting the sequence data since new techniques make rapid sequencing of hundreds of variant clones a standard practice. There has been, however, a conceptual limitation: how can we analyse mutations that have many changes per sequence? To avoid this, many labs generate random clones with only one or two changes per sequence. This severely limits the rate of gathering data.

We have developed several methods for analysis of sequence data to eliminate this difficulty(9,32). Ideally, we would generate randomized DNA in which the random bases are equiprobable, and then select functional sites. This would generate the largest possible amount of experimental information per sequence. Unfortunately, in the case of T7 promoters, a functioning promoter always killed the cells. (Even when the *Eco*RI to *Hin*dIII fragment of pBR322 that contains the -35 of the *tet* promoter(14) was replaced by a functional T7 promoter, direct selection for tetracyclin resistance was still not possible.) We therefore had to rely on a screen. Unfortunately, equiprobable random DNA only rarely contains functional promoters (even if the promoter only has 18 bits). Therefore we biased the input frequencies of the experiment toward a known functional sequence, $\phi 10$, so that 1% of the sequences would be wild type and so detectable by the screen. Although it made this experiment feasible, a bias does limit the amount of information that can be obtained from each sequence, and it makes analysis more difficult. We therefore devised the 'normalization' procedure to be able to answer the question: 'What distribution of output frequencies would we expect the polymerase to give if the input frequencies had been equiprobable?' This method has the advantage that the degree of randomization may be chosen to suit experimental constraints.

Information At T7 Promoters

We will first look at the data without the normalization. Position -3 is particularly interesting (Figure 5). Although the phage genome always has an A there, the table reveals that any other base is acceptable for strong promoter function. The frequency of bases at this position is close to both the experimental plan (0.85, 0.05, 0.05, 0.05) and to the frequencies of position -21, which was synthesized from the same bottle of mixed nucleotides, and for which the phage genomic patterns show little preference. Position -3 is a position where the phage sequence is completely conserved, but the conserved base is not necessary for T7 polymerase to function strongly.

We may analyze the rest of the data by using a more precise concept of information(8). One bit of information is sufficient to make a choice between two equally likely things, while two bits corresponds to a choice of 1 in 4. At position -3, the phage genome always has an A, so we represent that position as having 2 bits of information. Likewise, a position that has only purines would contain 1 bit of information, and one with equally likely bases would contain no information. We must extend this measure somewhat to account for any frequency of bases (11,9) and to represent the information that we learn from an experiment(33), but the concept is the same.

The data of Figure 5 are analyzed in Figure 6. At position -3 we graph 0 bits for the information that the polymerase needs according to this experiment and 2 bits for the information at the phage genome promoters. Position -3 is only one of several in which the two information curves diverge. We can estimate the total information by adding together the information from each position. The 'area' under the phage curve is $35.4 \pm 0.7$ bits(9) while the 'area' under the T7-W curve is $18 \pm 2$ bits. This experiment reveals that T7 polymerase only uses half of the information presented to it by the phage genome to determine a strong promoter as defined by our *in vivo* assay. That is, the polymerase uses just about the amount of information required (in theory) to locate the sites(9).

Alternative Hypotheses

T7 promoters have a dyad-symmetric element, ctc-ctatag-gag (38), that represents about 16.4 bits of information. Previously we proposed that this element might explain the excess information(9). Two lines of evidence now show that the leftmost 4 symmetry element positions (ctc-c) are important for promoter function. First, no clones with changes in these positions were found in this experiment nor in 10 sequences of a pilot experiment (which had a DNA synthesis that randomized only the symmetry region). Second, *in vitro* experiments(36,37) show that if these bases are altered, transcription is substantially reduced. Conversely, Figure 6 shows that there is excess information to the left of the symmetry element, as indicated by the divergence of the curves. So a large part of the symmetry element is required for efficient transcription, and some information outside of the symmetry is not required for transcription. Therefore the promoter region is not neatly divided into two independent domains by the symmetry element, as we had proposed, and the symmetry element does not explain the excess information.

The result that only 18 bits are used by the T7 polymerase in specific positions eliminates several other alternative explanations.

First, we might suppose that all of the pattern at the promoters is used to help the polymerase initiate. If this were true, then if we pick up mutations that are slightly less active than wild type (yet still functional by our assay), they should be scattered evenly throughout the site. However, by our assay (and those of others(39,36,37,40)) we can distinguish between positions that are very important to activity and those that are not,

even though both kinds are equally conserved in the phage.

Second, the patterns in the phage genome do not have excess information to allow the polymerase to be more accurate(9) or active, since that information is evidently not used by the polymerase.

Third, the two domain model of Chapman and Burgess(36) is insufficient to explain the results. This kind of model claims that the extra information is used for some specific function—such as opening the DNA—once the polymerase has bound to the site. Once again, the results described here eliminate this possibility since the polymerase functions well without the extra information. In any case, we would expect catalytic use of the excess to be unnecessary because the only requirement is to open the DNA and begin transcription. Given that the polymerase is already at the site, the opening could be done by binding non-specifically to the DNA backbone. Information that might be required for opening (in terms of sequence pattern) should be near the first base of the transcript. This information is already included in the 18 bits because the *in vivo* assay demands not only binding but also transcription. It may be appropriate to ignore this information, in which case less information than 18 bits would be needed to locate the sites.

Although our experiment did not show variation at positions -12 and -5, we know from our own and other results(41,39,40) that these two positions can be changed with little effect on transcription. We believe that this is an artifact of the toothpicking method of screening (see Results), and that changes in these positions weaken but do not destroy the function. If sites with changes in -12 or -5 are to be considered functional—which would be reasonable since some variation is allowed by the phage at these positions—then the total information we measure for the promoters would be *less* than 18 bits and the argument in favor of excess information would be strengthened.

The reader should be wary of a common but incorrect conclusion, namely that 'T7 promoters are over specified'. Our results indicate exactly the opposite. The confusion lies in what one defines as a promoter. We take 'promoter' to be those patterns in DNA that are required for initiation of transcription. Yet, traditionally the entire pattern at the transcriptional initiation points in T7 has been *assumed* to be the promoter. The results described in this paper show that this assumption is incorrect. That is, the promoter alone has just about as much information as one would predict(9) is needed to find the sites. So T7 promoters are *not* "over-specified". Apparently there is another pattern at genomic T7 promoters beyond that required for transcription. This distinction requires that we pay careful attention to our language. There is extra information *at* the phage promoter sites that is not required for transcription initiation, but there is *not* extra information *in* T7 promoters, because they contain just the amount needed to perform their function. Thus it would be incorrect to substitute 'in' for 'at' in the title of this paper.

We assumed that the polymerase interacts independently with each base in the site. In some cases this is a good first approximation(32,17), however further experiments will be required to determine if there are correlations between the bases. This will only require sequencing and characterization of more clones.

The role of the excess information in T7 physiology remains unknown. As discussed above, models that require the polymerase to use the excess directly are apparently ruled out. Some other effect has created the pattern. Since many particular models are possible(5), we describe only two of them here. First, the excess could be maintained by a recombination mechanism that exchanges information between sites, as do the $\lambda$ *att*P and *E. coli att*B sites or the repetitious sequences of higher organisms. A strong argument

against such a mechanism is that it may also tend to cause a high level of deletion between the repeats, which would be detrimental to T7. If this mechanism exists, it must be very important to T7.

The most likely explanation for the excess information is that one or more proteins bind at the same genomic locations that T7 polymerase uses for initiation. If only a single other protein binds, then to account for the excess we must also postulate that the two recognizers do not share information from the DNA patterns(9). (When we measure how much pattern there is at the genomic promoters while presuming that there are no other sites, we will naively think that there is an excess, whereas we have merely double counted.) If more than two proteins bind, one of them could be binding to the symmetric element. Thus the excess information could represent regulatory protein binding sites that are used under some conditions or at certain times during an infection. The putative protein(s) are as yet unidentified.

Summary

The method of random cloning and information analysis has the advantage that one can rapidly and precisely characterize the entire binding domain of a DNA binding protein. The method is also directly applicable to the study of any region of DNA—including protein coding regions—for which one or more functional assays are available. Data from quantitative assays can be analyzed by related methods(32).

This experiment shows that only a portion—roughly half—of the information at the sites of genomic bacteriophage T7 promoters is required for active transcription by T7 polymerase under the *in vivo* conditions used to assay activity. The rest of the information could represent the binding sites of one or more proteins(9), or it could have other unknown functions(5).

ACKNOWLEDGEMENTS

We thank W. Studier for BL21/DE3 and 4107(26), M. Casadaban for MC1061, B. Weiss for pKC7, J. Binkley for a careful DNA synthesis, A. Pelletier for suggesting the use of carbenicillin, C. Tuerk and D. McPheeters for help with DNA and RNA sequencing, C. E. Lawrence and G. W. Alvord for statistical advice, P. Lemkin for help with gel scanning, L. Sinclair, M. Nawroz and T. Hollingsworth for technical help, and A. Barber, D. Court, A. Konopka, D. Lipman, J. Maizel, E. Max, H. Nash, P. Rogan, J. Ruckman, K. Rudd, J. Strathern, J. Spouge, and R. Weisberg for critically reading the manuscript. We give special thanks to L. Gold who supported and guided this project from its inception. This work was supported in part by NIH grant GM28755.

*To whom correspondence should be addressed

REFERENCES

1. Chamberlin, M., McGrath, J., and Waskell, L. (1970) *Nature* **228**, 227–231.
2. Davanloo, P., Rosenberg, A. H., Dunn, J. J., and Studier, F. W. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2035–2039.
3. Tabor, S. and Richardson, C. C. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1074–1078.
4. Dunn, J. J. and Studier, F. W. (1983) *J. Mol. Biol.* **166**, 477–535.
5. Rosa, M. D. (1979) *Cell* **16**, 815–825.

6. Jolliffe, L. K., Carter, A. D., and McAllister, W. T. (1982) *Nature* **299**, 653–656.
7. Clift, B., Haussler, D., McConnell, R., Schneider, T. D., and Stormo, G. D. (1986) *Nucleic Acids Research* **14**, 141–158.
8. Pierce, J. R. (1980) *An Introduction to Information Theory: Symbols, Signals and Noise*, Dover Publications, Inc., New York., second edition.
9. Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986) *J. Mol. Biol.* **188**, 415–431.
10. Shannon, C. E. (1948) *Bell System Tech. J.* **27**, 379–423, 623–656.
11. Shannon, C. and Weaver, W. (1949) *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.
12. Hui, A., Hayflick, J., Dinkelspiel, K., and de Boer, H. A. (1984) *EMBO J.* **3**, 623–629.
13. Childs, J., Villanueba, K., Barrick, D., Schneider, T. D., Stormo, G. D., Gold, L., Leitner, M., and Caruthers, M. (1985) Ribosome binding site sequences and function, In Calendar, R. and Gold, L., editors, *Sequence Specificity in Transcription and Translation, UCLA Symposia on Molecular and Cellular Biology,* vol. 30, pp. 341–350, Alan R. Liss, Inc, New York.
14. Horwitz, M. S. Z. and Loeb, L. A. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7405–7409.
15. Hill, D. E., Oliphant, A. R., and Struhl, K. (1987) *Methods in Enzymology* **155**, 558–568.
16. Oliphant, A. R. and Struhl, K. (1987) *Methods in Enzymology* **155**, 568–582.
17. Barrick, D., Villanueba, K., Childs, J., Kalil, R., Schneider, T. D., Lawrence, C. E., Gold, L., and Stormo, G. D., Quantitative analysis of ribosome binding sites in *E. coli.*, in preparation for *J. Mol. Biol.*
18. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning, A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
19. Davis, R. W., Botstein, D., and Roth, J. R. (1980) *Advanced Bacterial Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
20. Daniels, D. L., Schroeder, J. L., Szybalski, W., Sanger, F., and Blattner, F. R. (1983) A molecular map of coliphage lambda, In *LAMBDA II*, pp. 469–517, Cold Spring Harbor Laboratory.
21. Rao, R. N. and Rogers, S. G. (1979) *Gene* **7**, 79–82.
22. Casadaban, M. J. and Cohen, S. N. (1980) *J. Mol. Biol.* **138**, 179–207.
23. Birnboim, H. C. and Doly, J. (1979) *Nucleic Acids Research* **7**, 1513–1523.
24. Dagert, M. and Ehrlich, S. D. (1979) *Gene* **6**, 23–28.
25. Steinberg, C. M. and Edgar, R. S. (1962) *Genetics* **47**, 187–208.
26. Studier, F. W. and Moffatt, B. A. (1986) *J. Mol. Biol.* **189**, 113–130.
27. McAllister, W. T., Morris, C., Rosenberg, A. H., and Studier, F. W. (1981) *J. Mol. Biol.* **153**, 527–544.
28. Gale, E. F., Cundliffe, E., Reynolds, P. E., Richmond, M. H., and Waring, M. J. (1981) *The Molecular Basis of Antibiotic Action*, John Wiley & Sons, Ltd., London, second edition.
29. Magasanik, B. (1970) Glucose effects: inducer exclusion and repression, In Beckwith, J. R. and Zipser, D., editors, *The Lactose Operon*, pp. 189–219, Cold Spring Harbor Laboratory.
30. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
31. Chen, E. Y. and Seeburg, P. H. (1985) *DNA* **4**, 165–170.
32. Stormo, G. D., Schneider, T. D., and Gold, L. (1986) *Nucleic Acids Research* **14**, 6661–6679.
33. Hobson, A. (1971) *Concepts in Statistical Mechanics*, Gordon and Breach Science Publishers, New York.
34. Stormo, G. D. (1988) *Ann. Rev. Biophys. Biophys. Chem.* **17**, 241–263.
35. Fienberg, S. E. (1980) *The Analysis of Cross Classified Catagorical Data*, The MIT Press, Cambridge, Mass., second edition.

36. Chapman, K. A. and Burgess, R. R. (1987) *Nucleic Acids Research* **15**, 5413–5432.
37. Milligan, J. F., Groebe, D. R., Witherell, G. W., and Uhlenbeck, O. C. (1987) *Nucleic Acids Research* **15**, 8783–8798.
38. Oakley, J. L., Strothkamp, R. E., Sarris, A. H., and Coleman, J. E. (1979) *Biochem* **18**, 528–537.
39. Morris, C. E., McGraw, N. J., Joho, K., Brown, J. E., Klement, J. F., Ling, M. L., and McAllister, W. T. (1987) Mechanisms of promoter recognition by the bacteriphage T3 and T7 RNA polymerases, In Reznikoff, W. S., Gross, C. A., Burgess, R. R., M. T. Record, J., Dahlberg, J. E., and Wickens, M. P., editors, *RNA Polymerase and the Regulation of Transcription*, pp. 47–58, Elsevier, New York.
40. Chapman, K. A., Gunderson, S. I., Anello, M., Wells, R. D., and Burgess, R. R. (1988) *Nucleic Acids Research* **16**, 4511–4524.
41. Osterman, H. L. and Coleman, J. E. (1981) *Biochem.* **20**, 4884–4892.
42. Beaucage, S. L. and Caruthers, M. H. (1981) *Tetrahedron Letters* **22**, 1859–1862.
43. Wallace, R. B., Johnson, M. J., Suggs, S. V., Miyoshi, K., Bhatt, R., and Itakura, K. (1981) *Gene* **16**, 21–26.