A computer algorithm for testing potential prokaryotic terminators

V.Brendel and E.N.Trifonov

Polymer Research Department, Weizmann Institute of Science, Rehovot 76100, Israel

Received 12 September 1983; Revised 23 February 1984; Accepted 26 April 1984

ABSTRACT

The nucleotide sequences of 30 factor-independent terminators of transcription with RNA polymerase from E. coli have been compiled and analyzed. The standard features - a stretch of thymine residues and a preceding dyad symmetry - are shared by most sequences, but there are striking exceptions which indicate that these features alone are not sufficient to describe these sites. In two thirds of the sequences the 3'-half of the dyad symmetry contains the pentanucleotide CGGG(G/C) or a close derivative; about one third have TCTG or a close derivative just downstream of the termination point. The TCTG-box might be implied in termination of stringently controlled operons of E. coli. An algorithm to locate terminators in templates of known nucleotide sequence has been constructed on the basis of correlation to the distribution of dinucleotides along the aligned signal sequences. The algorithm has been tested on natural sequences of a total length of about 11,500 N. It finds all known independent terminators and only a few other sites, including some of the rho-dependent and putative terminators.

INTRODUCTION

The prokaryotic genome is arranged into transcriptional units which define the primary RNA transcripts synthesized from the DNA template by RNA polymerase. The process of transcription is tightly controlled in the cell which requires varying amounts of specific transcripts during different stages of the cell cycle or in response to environmental conditions (1). This control is exerted at the levels of initiation, elongation, and termination of transcription. The sites of transcription initiation, or promoters, vary considerably in strength, i.e. in the maximal frequency with which initiation occurs. In addition, at some sites protein factors other than RNA polymerase holoenzyme are required for initiation, or repressor molecules bound to operator sites overlapping the promoter might temporarily inactivate the promoter (2). During elongation RNA polymerase pauses at specific sites; these pauses have also been thought to serve regulatory function (3). Transcription terminators have been found in various locations within operons, serving different control functions: distal to the last gene of an operon the terminator defines the 3'-end of primary transcripts initiated from an upstream promoter; factor-dependent terminators within or between genes of a polycistronic operon may account for transcriptional polarity (4); and translationally controlled termination within the leader region of an operon causes attenuation (5).

Much interest has been directed towards the question of which features in the DNA or DNA-RNA complex at control sites are recognized as signals by the transcribing polymerase molecule. These features somehow have to be contained in the specific nucleotide sequences at the sites. Quite a number of bacterial, phage, and plasmid promoters, pausing sites, and terminators have been mapped on their respective DNA templates and their nucleotide sequences have been established. These sequences have been analyzed with respect to homologies, and consensus features have been found which might serve as signals (2,4). However, a particular site may deviate appreciably from the consensus derived from the total ensemble of all known sites with the same control function. This variability makes it difficult to find control sites in sequences that have not been subjected to transcriptional assays. So far most of such attempts have been based upon ill-defined empirical comparison with standard consensuses, often leading to ambiguous results. The general problems and obstacles in sequence-directed mapping of DNA-protein interaction sites has been reviewed by Sadler et al. (6). Recently several quantitative methods have been suggested as algorithms to locate these sites in nucleotide sequences (7-14). We report here a comparative sequence analysis of terminators and an algorithm to locate termination sites.

FACTOR-INDEPENDENT TERMINATORS

Sites at which RNA polymerase from E. coli terminates transcription fall into two distinct classes: factor-dependent and factor-independent terminators (15). Independent terminators are functionally active in *in vitro* assays in the absence of any protein factors other than RNA polymerase. Dependent terminators require the presence of rho-protein (16) or of other factors (17) for termination to occur. The dependent sites bear little homology to independent terminators and to each other, and the putative common characteristics recognized by the transcribing molecule remain obscure (18). Also terminators that have been mapped according to results from *in vivo* studies may be confused by RNA processing events. Therefore we will consider only factor-independent terminators here.

Earlier comparison of several independent terminators has revealed two common features (15) : (i) a region of dyad symmetry, rich in G·C base pairs, allowing for the formation of a stable hairpin structure in the RNA transcript; and (ii) a run of consecutive thymine residues immediately downstream of the dyad symmetry, including the point(s) of termination. The G·C-rich dyad symmetry has been implicated in slowing down the transcribing polymerase (19),whereas the relative instability of the rU·dA hybrid is thought to facilitate release of the transcript (20).

Fig.1 shows the nucleotide sequences of the 30 independent terminators available to us at the time. Dyad symmetries which potentially yield stable hairpin structures in the RNA transcript are indicated by arrows. The typical hairpin has a stem of length 8±2 base pairs and a loop size of 5±1 nucleotides. On average the stem (when present) contains 77% G·C base pairs and is stable with a free energy of $\Delta G=-18.0\pm7.3$ Kcal/mole as calculated by the rules of Tinoco et al. (21). The adjacent T-rich region on average consists of 71% thymine residues over a length of 10 nucleotides and includes the often heterogeneous point of termination about 5-8 bases downstream of the indicated dyad symmetry (Fig.1; T-rich regions are underlined and the known points of termination are marked with dashes). Individual sequences deviate considerably from this standard: bacteriophage lambda t'R1 terminating the minor rightward (6S) RNA and the plasmid R1 copA-RNA terminator exhibit unusually long hairpins, presumably reflecting structural requirements for these non-messenger RNAs; the phage ϕx174 minor terminator (ϕx T2) and the E. coli site tyrT t lack a stretch of significantly high T-content; both dyad symmetry and T-rich region are missing in the terminators of plasmid R1 RNA-III and of E. coli lacI as well as in the E. coli rho attenuator site. Transcripts with 3'-ends at the indicated rho attenuator site have been found both in rho$^+$ and in rho$^-$ cells (35), but a processing rather than a termination event leading to these observations cannot strictly be ruled out. The RNA III of plasmid R1 could be generated only from linear templates and not from superhelical templates (56); this site resembles rho-dependent terminator sequences (56). The lacI terminator maps within the lac operon control region (36). Despite possible biological peculiarities of some of the non-standard terminators it appears that the features of dyad symmetry and adjacent run of thymine residues do not suffice to unequivocally characterize the termination signal. These features may be a sufficient condition for termination to occur, but apparently they are not necessary; in other words, it may be that RNA polymerase terminates transcription at all sites bearing dyad symmetry and adjacent T-rich region, but there are other sites lacking these features where factor-independent termination does occur as well.

Evidently the polymerase recognizes secondary properties of the nucleo-

SITE	SEQUENCE	NORM.CORR.SUM
E trp att	AAAGCAATCAGATACCCAGCCCGCCTAATGAGCGGGCTTTTTTTTGAACAAAATTAG	10.18
E tyrT t	TCACTTTCAAAAGCCCCGGAATTCTCAAACGAATCCGCAATCAAATATTCTGCCCAA	2.54
E his att	ACGCATGAGAAAGCCCCCGGAAGATCACCTTCCGGGGGCTTTTTATAATTAGCGCGG	7.11
E trp t	ACGCGCAGTTAATCCCACAGCCGCCAGTTCCGCTGGCGGCATTTTAACTTTCTTTAA	5.91
E phe att	GCGAAGACGAACAATAAAGGCCTCCCAAATCGGGGGGCCTTTTTTATTGATAACAAA	7.12
E thr att	GAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGT	7.78
E rrnC	CGCCCCTGCCAGAAATCATCCTTAGCGAAAGCTAAGGATTTTTTTTATCTGAAATAA	7.91
E ilvGEDA att	CTTAACGAACTAAGACCCCCGCACCGAAAGGTCCGGGGGTTTTTTTTGACCTTAAAA	9.74
E rrnF(G)	TCCGCCACTTATTAAGAAGCCTCGAGTTAACGCTCGAGGTTTTTTTTCGTCTGTATA	5.18
E fraB	ACGCATCGCCAATGTAAATCCGGCCCGCCTATGGCGGGCCGTTTTGTATGGAAACCA	8.18
E leu att	ACGCAGTCAAACAAAAAACCCGCGCCATTGCGCGGGTTTTTTTATGCCCGAAGCGAG	7.41
E rho att	TCACAACATTAAGTTCGAGATTTACCCCAAGTTTAAGAACTCACACTACTATGAATC	2.37
E lacI	GCAACGCAATTAATGTAAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTA	3.76
E spot42-RNA	GAATATTTTAGCCGCCCCAGTCAGTAATGACTGGGGCGTTTTTTATTGGGCGAAAGA	7.29
E supB-E	TACCCCAGCCACATTAAAAAAGCTCGCTTCGGCGAGCTTTTTGCTTTTCTGCGTTCA	4.34
S trp att	AATCAGCCAAACGATACCCGGCCCGCCTGTTAAGCCGTGCGTTTTTTGAACAAAAATA	8.62
S his att	ACGCATGAGAAAGCCCCCGGAAGATCATCTTCCGGGGGCTTTTTTTTTGGCGCGCGA	7.71
S leu att	TCAGCTCGAAGTCAAACAAAACCCGCGCCGTTGCGCGGGTTTTTTTATGCCTGACGC	5.94
λ T'R1	CGCAGGTAATAGTTAGAGCCTGCATAACGGTTTCGGGATTTTTTATATCTGCACAAC	7.06
λ t0	ATCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGGTGAGAATC	6.98
λ nutL●	CCAAAGCCTTCTGCTTTGAATGCTGCCCTTCTTCAGGGCTTAATTTTTAAGAGCGTC	5.89
λ tI	AGAGCATTAGCGCAAGGTGATTTTTGTCTTCTTGCGCTAATTTTTTGTCATCAAACC	4.74
λ tL3	GTCTACTCCGTTACAAAGCGAGGCTGGGTATTTCCCGGCCTTTCTGTTATCCGAAAT	7.55
φx T4	CCCAATTGTATGTTTTCATGCCTCCAAATCTTGGAGGCTTTTTTATGGTTCGTTCTT	5.14
φx T2	GGTGGTCAACAATTTTAATTGCAGGGGCTTCGGCCCCTTACTTGAGGATAAATTATG	3.88
fd	TGATAAACCGATACAATTAAAGGCTCCTTTTGGAGCCTTTTTTTTTGGAGATTTTCA	5.13
T7	AGAAAATGTAATCACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATAAGGAG	6.18
CE1 RNA-I	TGATCCGGCAAACAAACCACCGTTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAG	6.28
R1 copA-RNA	TTTCGTACTCGCAAAAGTTGAAGAAGATTATCGGGGTTTTTGCTTTTCTGGCTCCTG	5.84
R1 RNA-III	TTAAAAATTTACAGGCGATGCAATGATTCAAACACGTAATCAATATCTGCAGTTTAT	4.69
	. -40 . -30 . -20 . -10 . -1+ +5 +10	

Fig.1. Nucleotide sequences of thirty prokayotic factor-independent terminators. Only the DNA strand that is colinear with the transcript is shown. The exact points of termination, if known, are marked by dashes. Arrows indicate dyad symmetries that potentially yield stable hairpin structures in the RNA transcript. Underlined are T-rich regions immediately downstream of the dyad symmetries, and the TCTG consensus (dashed lines). Nucleotides are numbered from -45 to +12 with negative and positive numbers generally referring to non-transcribed and transcribed portions of the template, respectively. Also shown for each sequence is its normalized correlation sum with the terminator dinucleotide distribution matrix (see text). References for the sites are as follows: E trp att (22), E tyrT t (23,24) E his att (25), E trp t (26), E phe att (27), E thr att (28), E rrnC (29), E ilvGEDA att (30), E rrnF(G) (31), E frdB (32,33), E leu att (34), E rho att (35), E lacI (36,37), E spot42-RNA (38), E supB-E (39), S trp att (22), S his att (40), S leu att (41), λ T'R1 (42,43), λ t0 (44,43), λ nutL● (45; the ● - sign refers to rightward transcriptional orientation), λ tI (46), λ tL3 (47), φx T4 (48-50), φx T2 (48-50), fd (51,52), T7 (53), CE1 RNA-I (54), R1 copA-RNA (55), R1 RNA-III (56).

tide sequence at the control site rather than an invariable primary sequence by itself. At the terminator these properties presumably include the potential of forming a stable hairpin in the RNA transcript and the average base composition over certain segments of the sequence. Other structural and physical properties of DNA of probable importance in specific DNA-protein interaction are the helical twist angles (57,58) and the assumed wedge angles between adjacent base pairs (59,60), and the local thermostability (61,62). Since these properties are basically functions of base-stacking interactions we decided to analyze the dinucleotide- rather than the mononucleotide distribution at the termination sites. A similar analysis of promoters and preliminary results for terminators have been published elsewhere (63,13).

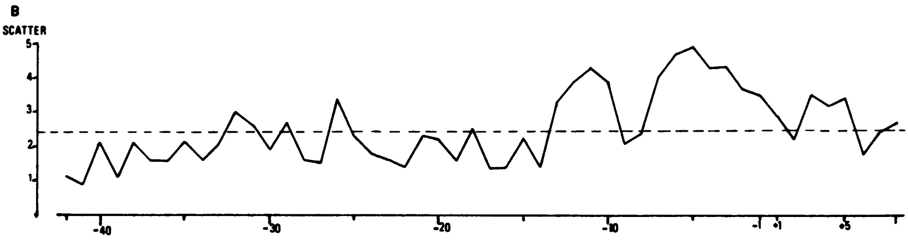
To evaluate the dinucleotide distribution the terminator sequences were initially aligned with respect to the prominent, central, or likely termination points as in Fig.1. The dinucleotide at the termination point was designated -1,+1; continued positive numbering to the right and negative numbering to the left indicate non-transcribed and transcribed portions of the template, respectively. In each of 50 dinucleotide positions along the sequences stretching from -42 to +9 (which is about the maximal region covered by the polymerase at a terminator (2)) the abundance of any particular dinucleotide was calculated. The result was obtained in form of a 16*50 matrix (m_{ij}) (not shown) in which rows correspond to different dinucleotides and columns to positions, whereby the column sums $\sum_i m_{ij}$ equal the number of sequences analyzed, here 30. To improve the matching of the sequences this initial alignment was slightly changed in the following way, taking into consideration similarity of the whole sequences rather than just the uncertain termination point: For each sequence its similarity with the matrix of the remaining 29 sequences was determined in 7 different frames corresponding to nucleotides -45 to +6, -44 to +7, ..., -39 to +12, respectively. As a measure of similarity we used the sum of matrix elements (normalized by row means and variances) as read from the dinucleotide sequence in each particular frame (see below). A new matrix was built from the sequences aligned in the frame which gave maximal similarity. This alignment procedure was repeated in the same way with the new matrix. After a few cycles this iteration came to an end in that the frames giving maximal similarity did not shift anymore. The final matrix is shown in Fig.2A. Other iteration procedures may be used; in all cases the matrices obtained (not shown) retain the prominent characteristics discussed below.

In order to determine the signal qualities of different positions and

A

		-40					-35					-30					-25					-20					-15					-10					-5					-1+1				+5				
AA	2	4	1	5	7	4	5	7	4	6	**10**	3	4	**8**	5	3	2	3	2	0	1	1	4	6	3	2	0	2	1	2	1	0	1	0	1	1	2	0	1	2	1	0	0	3	1	1	2	2	**8**	**11**
AC	3	1	1	0	3	1	1	2	2	2	4	4	1	0	1	2	1	0	4	0	2	0	0	3	2	0	0	1	0	0	1	1	1	1	0	0	0	1	0	2	1	0	1	1	2	3	0	1	1	1
AG	4	0	1	2	4	1	1	5	3	1	0	**7**	3	2	0	4	1	1	2	4	2	0	2	0	2	1	1	2	2	1	1	4	1	0	0	0	1	0	0	0	0	0	0	0	2	0	1	2	1	3
AT	1	4	6	1	1	5	2	1	0	4	0	4	0	0	5	0	3	1	0	0	3	1	0	1	**7**	2	1	1	1	0	0	0	0	0	3	1	1	2	0	0	1	**7**	3	5	1	1	2	2	0	3
CA	4	3	3	4	0	0	5	1	3	6	3	3	2	2	1	1	0	3	1	5	0	3	3	2	2	0	0	1	1	0	1	1	0	2	0	1	1	1	1	1	0	1	0	0	1	2	1	2	3	1
CC	2	3	3	0	1	2	1	1	1	2	3	4	6	**8**	5	1	3	7	5	4	3	7	2	2	2	1	0	0	5	1	1	0	2	4	3	2	0	0	0	0	1	0	0	0	1	3	2	0	2	1
CG	2	3	0	1	0	2	0	4	0	0	2	0	1	1	5	4	0	1	1	5	3	2	1	1	2	3	2	0	3	**10**	**7**	3	0	3	2	0	0	0	0	0	0	1	1	1	0	0	1	3	3	3
CT	1	0	3	0	0	3	0	0	1	0	0	0	0	1	0	1	1	5	2	3	1	1	6	1	1	4	2	3	2	0	0	0	0	4	5	3	2	0	2	1	3	1	0	1	0	**9**	2	0	0	1
GA	2	1	3	3	0	3	1	0	4	1	1	1	3	1	2	2	2	1	1	2	1	1	1	3	0	0	2	1	0	0	3	2	0	2	0	0	0	0	0	0	0	0	1	3	2	1	3	5	2	2
GC	2	4	0	0	2	0	1	2	4	2	0	1	2	3	1	0	**11**	2	3	2	5	3	2	0	1	1	2	4	4	5	2	1	**10**	5	3	0	0	0	1	2	1	0	0	1	1	0	2	6	3	0
GG	1	2	0	1	0	0	1	0	0	0	0	0	2	1	0	5	2	0	0	0	3	1	0	1	2	3	2	2	2	3	**13**	**16**	**13**	6	0	0	0	1	0	0	0	0	0	2	1	4	1	1	0	0
GT	0	2	2	0	5	1	0	3	3	0	0	0	1	1	1	0	0	1	0	0	2	3	2	0	2	2	3	0	1	0	0	2	0	3	6	2	0	0	0	1	1	0	2	2	1	1	0	1	1	2
TA	1	1	1	3	4	2	4	1	2	1	4	1	1	0	1	1	1	1	0	1	0	1	2	3	0	0	4	0	1	1	0	0	0	0	1	2	0	0	2	0	**6**	3	**8**	0	1	1	1	1	5	1
TC	2	1	1	1	1	3	3	0	1	2	0	0	3	0	0	1	1	0	5	1	3	2	2	2	3	2	2	6	2	3	0	0	0	0	0	1	1	2	1	0	0	1	1	0	**10**	0	1	1	0	0
TG	2	0	3	3	0	0	3	2	0	0	0	1	0	0	2	2	1	2	1	2	0	2	1	3	0	2	2	3	1	4	0	0	2	0	0	0	0	0	3	2	0	2	**7**	2	3	2	**10**	0	0	0
TT	1	1	2	6	2	3	2	1	2	3	3	1	1	2	1	3	1	2	3	1	1	2	2	2	1	7	7	4	4	0	0	0	0	0	6	**17**	**22**	**23**	**19**	**19**	**15**	**14**	6	9	3	2	1	3	1	1

B



C

C G G G $\substack{G\\C}$ T T T T T T T T $\substack{A\\G}$ T C T G

C G G $\substack{G\\C}$ T T T T T T A T $\substack{A\\G}$

Fig.2. (A). Dinucleotide distribution matrix for thirty prokaryotic factor-independent terminators. Columns correspond to positions along the aligned sequences and are numbered as in Fig.1. The most prominent elements are boxed. The vertical dashed lines are inserted for better readability only. (B). Scatter distribution. S is the logarithm of the maximal product of squared standardized deviations from row and column means of the matrix in each dinucleotide position. The dashed line corresponds to the log 256 threshold (see text). (C) Continuous sequences that can be read from the regions of prominent matrix elements in positions -13 to -10, -7 to +1, and +3 to +5, respectively.

the significance of particular elements the matrix was normalized with respect to row and column means and variances. The row variances were estimated by the row means $\overline{m_{i.}}$. The column means $\overline{m_{.j}}$ were calculated discarding minimal and maximal column elements, and the column variances were estimated by binomial variances $\overline{m_{.j}}$ * $(1-\overline{m_{.j}})/16$. Fig.2B gives the logarithm of the maximal product of squared standardized deviations from row and column averages plotted for each column as a measure of the scatter of the distribution. High scatter would indicate that in this position certain dinucleotides occur preferentially as compared to others. Matrix elements that exceed a threshold of log 256 upon standardization (corresponding to the equivalent of deviations from row and column means by more than four standard deviations) are boxed in Fig.2A.

Fig.2B reveals a nonuniform scatter distribution. Three major signal regions map from positions -13 to -10, -7 to +1, and +3 to +5. As can be seen from Fig.2A the corresponding prominent elements can be aligned to consensus sequences: CGGG(G/C) or CGG(G/C), TTTTTTTT(A/G) or TTTTTTAT(A/G), and TCTG, respectively (Fig.2C). The -13 to -10 region coincides with the 3'-half of the standard dyad symmetry. It appears that the high G·C-content of the dyad symmetry is unevenly distributed in favour of the consensus sequence, as has been suggested from an early analysis of the first few published terminator sequences (64). In fact, 8 sequences match the consensus exactly and another 12 deviate in just one position (from Fig.1). Of 30 random sequences at most 1 would be expected to deviate from the consensus in not more than one position (for random sequences constrained to contain only Cs and Gs the expected numbers of sequences mismatching the consensus in none or only one position would be about 2 and 8, respectively). Termination relief mutants mapping in the CGGG(G/C) region further emphasize that the particular sequence rather than just the C+G content is important for the termination signal (28,65). The run of thymine residues is shared by all but five sequences (see above). The TCTG sequence is found in seven sequences and with allowance for one mismatch in another five (Fig.1). This common element was reported before for three of the sequences (66). The other prominent matrix elements found in the left half of the matrix partly map in the 5'-halves of the dyad symmetries and may be reflections of this constraint, as is the broad G·C-rich band between -30 and -20. Interestingly, the distance between the cluster of prominent elements around -30 and the CGGG(G/C) region is about the same as the distance between the two contact sites of RNA polymerase with promoters (67). It is not known whether these contacts are also implied in the termination event.

AN ALGORITHM TO LOCATE TERMINATORS

If the dinucleotide distribution matrix correctly reflects the termination signal then the similarity of a particular sequence to the matrix would give indication as to whether this sequence might have termination activity. The matrix could then serve as a probe to locate terminators on any given template. Similar ideas have been implied in the construction of algorithms to

```
        -40                 -35                 -30                 -25                 -20                 -15                 -10                     -5            -1+1              +5

AA  o   o   o   o   1.  o   o   1.  o   1.  1.  o   o   1.  o   o   o   o   o   o   o   o   o   1.  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   1.  1.  AA
AC  o   o   o   o   o   o   o   o   o   o   .40 .57 o   o   o   o   o   o   .80 o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   AC
AG  o   o   o   o   o   o   o   .71 o   o   o   1.  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   .25 o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   AG
AT  o   o   1.  o   o   1.  o   o   o   o   o   o   o   o   1.  o   o   o   o   o   o   o   o   o   1.  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   .50 o   .56 o   o   o   o   o   o   AT
CA  o   o   o   o   o   o   1.  o   o   1.  o   o   o   o   o   o   o   o   o   1.  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   CA
CC  o   o   o   o   o   o   o   o   o   o   o   o   1.  1.  1.  o   o   1.  1.  o   o   1.  o   o   o   o   o   o   1.  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   CC
CG  o   o   o   o   o   o   o   o   o   o   o   o   o   o   1.  o   o   o   o   1.  o   o   o   o   o   o   o   o   o   1.  .54 o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   CG
CT  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   .71 o   o   o   o   1.  o   o   o   o   o   o   o   o   o   o   o   .83 o   o   o   o   o   o   o   o   o   o   1.  o   o   o   o   CT
GA  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   .83 o   o   GA
GC  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   1.  o   o   o   1.  o   o   o   o   o   o   o   o   .50 o   o   .77 .83 o   o   o   o   o   o   o   o   o   o   o   o   o   1.  o   o   GC
GG  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   1.  o   o   o   o   o   o   o   o   o   o   o   o   o   o   1.  1.  1.  1.  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   GG
GT  o   o   o   o   .71 o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   1.  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   GT
TA  o   o   o   o   o   o   o   o   o   o   .40 o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o       o   1.  o   o   o   o   o   .63 o   TA
TC  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   1.  o   o   o   o   o   o   o   o   1.  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   1.  o   o   o   o   o   TC
TG  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   .40 o   o   o   o   o   o   o   o   o   o   o   o   .88 o   o   o   1.  o   o   o   TG
TT  o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   o   1.  1.  o   o   o   o   o   o   o   o   1.  1.  1.  1.  1.  1.  1.  o   1.  o   o   o   o   o   o   TT

  G   A   T   T   C   C   A   A   A   A   C   C   C   C   G   C   C   G   G   C   G   C   A   A   A   C   C   G   G   G   C   G   G   G   G   T   T   T   T   T   C   G   T   T   T   A   A   G   C   A   C
    o   o   o   o   o   o   o   1.  o   o   o   o   1.  o   o   o   o   o   o   1.  1.  o   o   1.  o   o   o   o   o   .50 .54 1.  1.  1.  1.  1.  1.  1.  1.  o   o   o   o   1.  o   o   o   1.  o   o      (= 16.04)

normalized correlation sum = (16.04 - 4.0)/2.12 =  5.68
```

Fig.3. Example of calculation of the correlation sum of a sequence (here the E ilvB attenuator (Fig.6)). The signal matrix (top) was obtained from the original dinucleotide distribution matrix after discarding low frequency elements and normalizing by dividing the remaining entries by the maximal column elements. The contributions to the correlation sum, as read from the matrix, are shown below the matrix together with the aligned sequence. The final figure of the normalized correlation sum is obtained by subtracting 4.0, the mean correlation sum for random sequences, and dividing by 2.12, the standard deviation for random sequences.

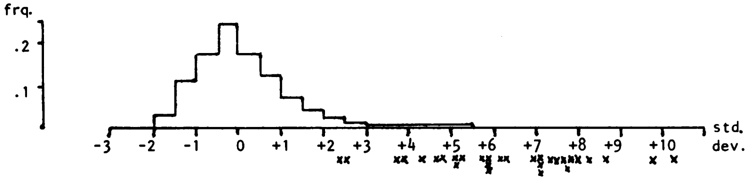


Fig.4. Distribution of the normalized correlation sum for 10,000 random sites (histogram) and the 30 known independent terminators (crosses). The correlation sum measures similarity of a sequence to the dinucleotide distribution matrix Fig.2A (see text). Mean and standard deviation are 4.0 and 2.12 for the random sites and 17.3 and 4.02 for the terminators, respectively. Results are shown normalized with respect to the random site mean and standard deviation.

find ribosome binding sites, promoters, and nucleosomes (11-14).

To quantitatively define similarity of a given sequence of length 51 nucleotides to the matrix we calculate a correlation sum as follows: First we discarded all low frequency elements of the matrix which contribute hardly, if at all, to the signal pattern. More precisely, we replaced by zeros all matrix elements that yield a product of squared standardized deviations from row and column averages less than 16 (corresponding to the equivalent of deviations from row and column means by less than two standard deviations); this improved the signal-to-noise ratio (see below). The remaining entries were normalized by dividing them by their respective maximal column elements. The correlation sum for a sequence is then obtained by adding up matrix elements as read from the sequence of dinucleotides. An example of these calculations is shown in Fig.3.

In order to evaluate whether the correlation sum might serve to separate terminators from non-terminators we calculated the correlation sum for all possible frames on a random sequence of length 10,000 N as well as for the ensemble of terminators. The result is shown in Fig.4. The histogram represents the distribution of the correlation sums for the random sequence. The distribution has been normalized with respect to the sample mean of 4.0 and the sample standard deviation of 2.12. The crosses indicate the values for the terminators (see also Fig.1). These values were obtained by taking the terminator under consideration out of the matrix (i.e. by subtracting 1 from each matrix element corresponding to the equivalently positioned terminator dinucleotide), correlating it with the remaining matrix as before, and adjusting the value received by a factor of 30/29. The thirty terminator correlation sums were distributed with a mean of 17.3 and a standard deviation of 4.02. Normalized by the random sequence mean and standard deviation they have an average correlation sum of 6.27. None of the random sequence sites had a correlation sum as high as that, and only .5% of the random sites had a correlation sum greater than 3.0. The random sequence had been generated with dinucleotide probabilities fixed as in E. coli sequences (estimated from ECRPOL, see below). Random sequences with dinucleotide probabilities fixed as in phages ϕx174 or lambda gave very similar results (data not shown). We also tried other correlation functions (data not shown); the one described above gave best separation between terminators and random sequences.

The matrix was used to screen natural sequences for terminators. A normalized correlation sum of 3.0 was defined as threshold, i.e. we would consider sites giving a higher correlation sum to be likely candidates for terminators. By Fig.4, two of the thirty terminators of our collection would have been missed this way, but on the same hand the calculations with random sequences indicated that only few sites would be predicted wrongly. The first sequence analyzed was a 3,400 bp long segment of the genome of bacteriophage lambda (68) as contained in the EMBL Data Library (Release 1; the EMBL name is LAMBDA). This sequence covers the early part of the leftward operon and includes the terminators ti, tL1, tL2a, tL2b, tL2c, and tL2d, all of which except for tL1 have been located exactly (68-71; Fig.5A1). None of these terminators was included in our collection: tL1, tL2c, and tL2d are rho-dependent terminators, ti has not been assayed *in vitro*, and the two in-

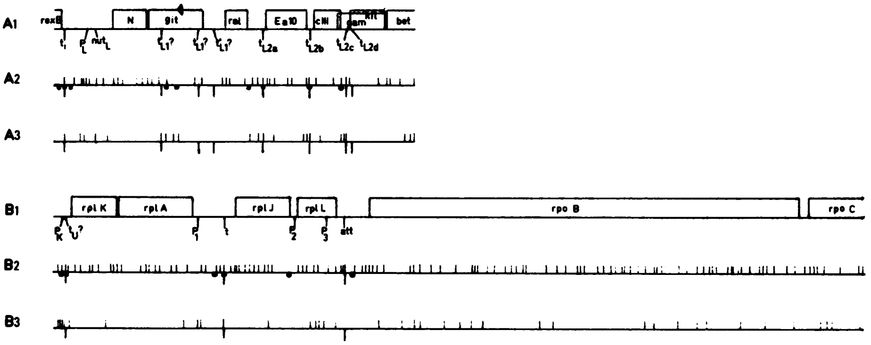
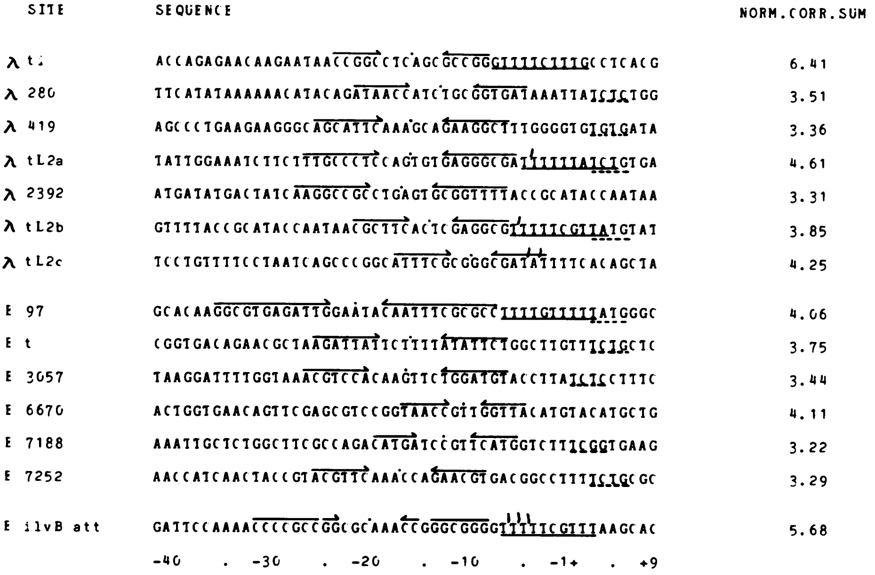


Fig.5. (A1 & B1) Genetic maps of LAMBDA and ECRPOL, respectively. Transcriptional orientation is from left to right except for the git gene which is in opposite orientation. The maps are drawn approximately to scale. (A2 & B2) Dyad symmetries (upward whiskers) and T-rich regions (dots) in LAMBDA and ECRPOL, respectively, as defined in the text and determined by the sequence analysis program of Sege et al. (75). Downward whiskers correspond to the known or putative terminators. (A3 & B3) Sites of high correlation to the terminator matrix (as defined in the text) on LAMBDA and ECRPOL, respectively (upward whiskers). Downward whiskers correspond to the known or putative terminators.

dependent sites tL2a and tL2b were published after we had completed the initial calculations leading to the present matrix. Secondly we analyzed the cluster of genes coding for ribosomal proteins and the β- and β'-subunits of RNA polymerase at about 88 min. of the E. coli genetic map (EMBL library name is ECRPOL; length 7,604 bp). This sequence contains one site known to function as an attenuator in vivo and two other putative terminators (72,73; Fig.5B1). As a third example we will discuss results on the recently published control sequence of the E. coli ilvB operon (370 bp) which contains an attenuator (74).

The results of our analysis are summarized in Fig5. On average every one to two hundredth site yielded a normalized correlation sum greater than 3.0 (23 sites in LAMBDA (Fig.5A3), 39 sites in ECRPOL (Fig.5B3), and 3 sites in ilvB (not shown)). Amongst the sites found by the matrix are on LAMBDA the three independent terminators ti (normalized correlation sum 6.41), tL2a (4.61), and tL2b (3.85), the rho-dependent terminator tL2c (4.25), and the first of the suggested sites (68) for tL1 (3.08); on ECRPOL the putative (72) terminators t"U" (3.05) and t (3.75); and on ilvB the attenuator. Not found are either of the other two locations suggested (69) for tL1 nor the rho-dependent terminator tL2d on LAMBDA, nor the attenuator in ECRPOL. We also determined the number of hairpins (stem containing at least 5 Watson-Crick base pairs, ratio of matches to length greater than .75, no loopouts allowed, loop size 3-7 N) and the number of T-rich regions (at least 7 Ts within a stretch of length 10) using the sequence analysis program of Sege et al. (75; Fig5 A2&B2). Hairpins occur with a frequency of about 1 per 50 nucleotides in LAMBDA and ECRPOL. T-rich regions are much less frequent (9 sites in LAMBDA and 6 in ECRPOL). For comparison we also analyzed two random sequences with dinucleotide frequencies as in LAMBDA or ECRPOL, respectively. It appears that the features occur with frequencies independent of the particular dinucleotide composition (except for the T-rich regions which are more frequent in the random sequence equivalent to LAMBDA which has 6.9% TT dinucleotides as compared to 4.9% in ECRPOL). The natural sequences seem to be rich in hairpins as compared to the random sequences (about 1 hairpin per 80 nucleotides). Similarity to the matrix is apparently also on random sequences more selective than the potential for formation of a stable RNA hairpin (the frequency of sites with normalized correlation sum greater than 3.0 is about 1 per 200 nucleotides).

Evidently any of the three features taken by itself is unsatisfactory as a terminator algorithm in that either many sites are probably predicted wrongly or some known terminators are missed. The combination of hairpin + adjacent T-rich region is only found at the terminators ti, tL2a, and tL2b of LAMBDA, at the putative t"U" terminator as well as at an upstream site overlapping the promoter P_K in ECRPOL, and at the ilvB attenuator. These results are compatible with the notion that the two features taken together form a sufficient condition for termination. Fig.6 shows all sites with a normalized correlation sum greater than 3.0 and an overlapping dyad symmetry with its 3'-end mapping 34-40 nucleotides downstream to the 5'-end of the site (as for the known terminators, Fig.1). In LAMBDA there are 7 sites with these properties, including ti, tL2a, tL2b, and tL2c. ti, tL2a, and tL2b match the CGGG(G/C) consensus in four positions; tL2c gives a perfect match. tL2a also has the TCTG sequence, tL2b the derivative TATG. tL2c, a terminator which is rho-dependent in tandem with the downstream tL2d site lacks the T-rich region. The rho-dependent terminators tL2d and tL1 are not found. The other three LAMBDA sites shown in Fig.5 map proximal to gene N and immediately distal of Ea10. It is not known whether they are active in vivo. The algorithm finds 6 sites on ECRPOL. The first of these overlaps with the PK promoter and could be the putative t"U" terminator. The second site maps between rplA



SITE	SEQUENCE	NORM.CORR.SUM
λ ti	ACCAGAGAACAAGAATAACCGGCCTCAGCGCCGGGTTTTCTTTGCCTCACG	6.41
λ 280	TTCATATAAAAAACATACAGATAACCATCTGCGGTGATAAATTATCTCTGG	3.51
λ 419	AGCCCTGAAGAAGGGCAGCATTCAAAGCAGAAGGCTTTGGGGTGTGTGATA	3.36
λ tL2a	TATTGGAAATCTTCTTTGCCCTCCAGTGTGAGGGCGATTTTTTATCTGTGA	4.61
λ 2392	ATGATATGACTATCAAGGCCGCCTGAGTGCGGTTTTACCGCATACCAATAA	3.31
λ tL2b	GTTTTACCGCATACCAATAACGCTTCACTCGAGGCGTTTTTCGTTATGTAT	3.85
λ tL2c	TCCTGTTTTCCTAATCAGCCCGGCATTTCGCGGGCGATATTTTCACAGCTA	4.25
E 97	GCACAAGGCGTGAGATTGGAATACAATTTCGCGCCTTTTGTTTTTATGGGC	4.06
E t	CGGTGACAGAACGCTAAGATTATTCTTTTATATTCTGGCTTGTTTCTGCTC	3.75
E 3057	TAAGGATTTTGGTAAACGTCCACAAGTTCTGGATGTACCTTATCTCCTTTC	3.44
E 6670	ACTGGTGAACAGTTCGAGCGTCCGGTAACCGTTGGTTACATGTACATGCTG	4.11
E 7188	AAATTGCTCTGGCTTCGCCAGACATGATCCGTTCATGGTCTTTCGGTGAAG	3.22
E 7252	AACCATCAACTACCGTACGTTCAAACCAGAACGTGACGGCCTTTTCTGCGC	3.29
E ilvB att	GATTCCAAAACCCCGCCGGCGCAAACCGGGCGGGGTTTTTCGTTTAAGCAC	5.68
	-40 . -30 . -20 . -10 . -1+ . +9	

Fig.6. Sites on LAMBDA, ECRPOL, and the ilvB control region predicted by the terminator algorithm (see text). The notation refers to Fig.5 (numbers indicate the positions of the 3'-ends of the sites calculated from the left ends of the respective templates). Features are marked as in Fig.1.

and rplL (t; Fig5.2A) where preliminary in vivo results indicated termination or processing. The other four sites at both ends of rpoB and at the beginning of rpoC are not known as terminators. The in vivo attenuation site between rplL and rpoB is not found by the algorithm. This site has a 11 bp long perfect dyad symmetry but the adjacent sequence, TTTTGCGCTG, is not T-rich by our criteria. On ilvB the algorithm only predicts the known attenuator.

The combined criteria of high similarity to the terminator matrix and presence of dyad symmetry give best results in terms of neither missing known terminators nor predicting many sites probably wrongly. We propose search for sites satisfying these criteria as an algorithm to locate terminators.

DISCUSSION

RNA polymerase recognizes specific sites on the DNA template where to initiate and terminate transcription and where to pause during elongation. Many of these sites have been sequenced (2,76). Although some sequence homology is shared between functionally homologous sites, the exact nature of the signal recognized by the polymerase has remained unclear. Concurrently no reliable algorithms have been formulated that would allow to locate promoters, terminators, or pausing sites according to the nucleotide sequence of the template.

Transcription terminators are known to occur in various positions within operons, serving different control functions and displaying several distinct patterns of dependencies on additional protein factors (15,18). Factor-independent termination sites generally contain a G·C-rich dyad symmetry followed by a run of consecutive thymine residues (4). As there are striking exceptions to this general scheme (Fig.1) these properties may be sufficient conditions for termination but apparently are not necessary.

Unlike the translational code or the code for restriction enzyme cutting sites where error-proofness is of vital importance for the cell, transcriptional control sites presumably should not be "all-or-nothing"- signals. The required frequency of initiation will vary from one promoter to the other, and even at a particular site at different times in the cell cycle or according to environmental conditions (1). Readthrough at terminators allows for the regulation of the ratio of terminator-upstream messenger RNAs to terminator-downstream messengers (5). These modulations of signal strength could be accounted for if the signal were composed of additive elements such that their sum would determine signal strength. Indeed, the model seems to hold

as far as the two contact sites of RNA polymerase in the promoter are con-
ᶜerned: it appears that a bad match to the ideal sequence in the '-35 region' can be compensated by a good match in the '-10 region', and vice versa (13). In terminators a dyad symmetry potentially yielding a stable RNA hairpin in conjunction with a run of thymine residues may be sufficient to surn ss the signal threshold. In the absence of either feature other sequence properties or additional protein factors may compensate this lack. This principle of distributional recognition has been described in detail elsewhere (13).

In search for signal components of terminators we have analyzed the dinucleotide distribution at termination sites. The matrix of dinucleotides versus positions along the aligned sequences (Fig.2A) reveals three major signal regions (Fig.2B) corresponding to the consensus sequences CGGG(G/C) or CGG(G/C), TTTTTTTT(A/G) or TTTTTTAT(A/G), and TCTG, respectively. The first of these coincides with the 3'-half of the standard dyad symmetry. TCTG maps in the non-transcribed portion downstream of the termination point. The pro-
nortion of terminators containing these consensuses or close derivatives (Fig.1) is well beyond what would be expected on a random sequence basis. Notably, TCTG is present in both E. coli tyrT t, which lacks the thymine rich region, and in the plasmid R1 RNA-III terminator, which lacks both dyad symmetry and thymine rich region; the E. coli rho attenuator, which also lacks both standard features, contains the derivative TATG. If the TCTG consensus is implied in the termination signal these would be further examples for compensation in distributional recognition. Besides of in E rrnC, E rrnF(G), E supB-E, and E tyrT t (Fig. 1) of our initial collection we later (subsequent to our calculations) found another five sites in E. coli bearing TCTG: the ribosomal RNA operon rrnD terminator (77,78), the in vivo terminators E rpmG t (79) and E lpp t (80,81), the putative terminator E rrnB T1 (66), and the site E t beyond rplA (Fig.6). These sites terminate transcripts coding for either ribosomal RNAs, transfer RNAs, or ribosomal proteins (except for the llp transcript which codes for the lipoprotein of the outer membrane, which is probably the most abundant protein in the E. coli cell (81)). This might suggest a role for TCTG in stringent control (82). Other consensuses that had been suggested from preliminary calculations (63,13) are not confirmed by the current matrix. This may be due to statistical inaccuracies resulting from the small ensemble of sequences available to us at that time or due to previous inclusion of factor-dependent sites which have not been considered here.

We have developed an algorithm to locate independent terminators on the basis of similarity of a sequence to the dinucleotide distribution matrix of Fig.1A. Similarity of a sequence to the matrix is measured by a correlation sum to which each dinucleotide of the sequence contributes in proportion to the relative frequency of occurrence of this dinucleotide in same position in known terminators as read from the matrix. A sequence with high similarity to the matrix, i.e. with high correlation sum, would be a likely candidate for termination activity. We checked this notion on random and natural sequences (Figs. 4&5). The known terminators indeed generally give a correlation sum much higher than the average for random sequences (Fig.4). Two of the thirty terminators give exceptionally low correlation sums (Fig.1). This suggests that high correlation to the matrix is not an absolutely necessary property of terminators either, though it seems to be more general than either dyad symmetry or T-rich region.

Best results on the natural sequences were obtained by combining the criteria of high correlation sum with an overlapping dyad symmetry upstream of the proposed termination point. This way the algorithm correctly finds all known independent terminators, one rho-dependent site, and three sites known or likely to function *in vivo* (Fig.6). About the same amount of sites are found for which termination activity has not been shown or suggested. One rho-dependent site and two *in vivo* terminators are missed. Further experimental evidence will be required to evaluate the performance of the algorithm.

Limitations of the algorithm described lie partly in the fact that the number of sequenced terminators is still relatively small. A growing ensemble should improve the signal-to-noise ratio in the dinucleotide distribution matrix. Also the current collection may be biased in favour of canonical terminators. The matrix includes directly the standard feature of a T-rich region covering the termination point(s) and indirectly the preceding standard dyad symmetry. The length and stability of the hairpin structure potentially formed in the RNA transcript is a feature that has to be taken into account separately. It is not clear as yet what the relative contribution to the signal is of each of these features. Some answers to this question are provided by mutant analyses (5), transcriptional assays with base analogues (5), heteroduplex analysis (83), and construction of synthetic sites (15). A refined algorithm will presumably have to include quantitative measures of all signal components weighted by their relative significance.

Factor-dependent terminators do not share obvious homologies with independent sites or with each other (18). We are currently investigating whether they have common patterns in their dinucleotide distribution and how they match with the independent terminator matrix.

ACKNOWLEDGEMENTS

We would like to thank E. Pashnina for taking part in earlier stages of the work and EMBO for a short-term fellowship to V.B.. V.B. is currently supported by a MINERVA fellowship.

REFERENCES

(1) Pribnow, D. (1979) in: Biological Regulation and Development, Vol. I, Goldberger,R.F.(ed.), Plenum Press, NY, pp. 219-277
(2) Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13, 319-353
(3) Winkler, M.E. and Yanofsky, C. (1981) Biochem. 20, 3738-3744
(4) Adhya, S. and Gottesman, M. (1978) Ann. Rev. Bioch. 47, 967-996
(5) Yanofsky, C. (1981) Nature 289, 751-758
(6) Sadler, J.R., Waterman, M.S., and Smith, T.F. (1983) Nucl. Acids Res. 11, 2221-2231
(7) Smith, T.F., Waterman, M.S., and Sadler, J.R. (1983) Nucl. Acids Res. 11, 2205-2220
(8) McMahon, J.E. and Tinoco, I.,Jr. (1978) Nature 271, 275-277
(9) Scherer, G.E.F., Walkinshaw, M.D., and Arnott, S. (1978) Nucl. Acids Res. 5, 3759-3773
(10) Otsuka, J. and Kunisawa, T. (1982) J. Theor. Biol. 97, 415-436
(11) Stormo, G.D., Schneider, T.D., Gold, L., and Ehrenfeucht, A. (1982) Nucl. Acids Res. 10, 2997-3011
(12) Harr, R., Häggström, M., and Gustafsson, P. (1983) Nucl. Acids Res. 11, 2943-2957
(13) Trifonov, E.N. (1983) Cold Spring Harb. Symp. Quant. Biol. 47, 271-278
(14) Mengeritsky, G. and Trifonov, E.N. (1983) Nucl. Acids Res. 11, 3833-3851
(15) Platt, T. (1981) Cell 24, 10-23
(16) Roberts, J.W. (1969) Nature 224, 1168-1174
(17) Kingston, R.E. and Chamberlin, M.J. (1981) Cell 27, 523-531
(18) Holmes, W.M., Platt, T., and Rosenberg, M. (1983) Cell 32, 1029-1032
(19) Gilbert, W. (1976) in: RNA Polymerase, Losick, R. and Chamberlin, M. (eds.), Cold Spring Harbor Laboratory, pp. 193-205
(20) Martin, F.H. and Tinoco, I.,Jr. (1980) Nucl. Acids Res. 8, 2295-2299
(21) Tinoco, J.,Jr., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. & Gralla, J. (1973) Nature New Biol. 246, 40-41
(22) Lee, F., Bertrand, K., Bennett, G., and Yanofsky,C. (1978) J. Mol. Biol. 121, 193-217
(23) Küpper, H., Sekiya, T., Rosenberg, M., Egan, J., and Landy, A. (1978) Nature 272, 423-428
(24) Rossi, J., Egan, J., Hudson, L., and Landy, A. (1981) Cell 26, 305-314
(25) Frunzio, R., Bruni, C.B., and Blasi, F. (1981) Proc. Natl. Acad. Sci. USA 78, 2767-2771
(26) Wu, A.M., Christie, G.E., and Platt, T. (1981) Proc. Natl. Acad. Sci. USA 78, 2913-2917
(27) Zurawski, G., Brown, K., Killingly, D., and Yanofsky, C. (1978) Proc. Natl. Acad. Sci. USA 75, 4271-4275
(28) Gardner, J.F. (1982) J. Biol. Chem. 257, 3896-3904
(29) Young, R.A. (1979) J. Biol. Chem. 254, 12725-12731
(30) Nargang, F.E., Subrahmanyam, C.S., and Umbarger, H.E. (1980) Proc. Natl. Acad. Sci. USA 77, 1823-1827
(31) Sekiya, T., Mori, M., Takahashi, N., and Nishimura, S. (1980) Nucl. Acids Res. 8, 3809-3827
(32) Jaurin, B., Grundstrôm, T., Edlund, T., and Normark, S. (1981) Nature 290, 221-225
(33) Grundström, T. and Jaurin, B. (1982) Proc. Natl. Acad. Sci. USA 79, 1111-1115
(34) Wessler, S.R. and Calvo, J.M. (1981) J. Mol. Biol. 149, 579-597
(35) Brown, S., Albrechtsen, B., Pedersen, S., and Klemm, P. (1982) J. Mol. Biol. 162, 283-298
(36) Horowitz, H. and Platt, T. (1982) J. Biol. Chem. 257, 11740-11746

(37) Cone, K.C., Sellitti, M.A., and Steege, D.A. (1983) J. Biol. Chem. 258, 11296-11304

(38) Joyce, C.M. and Grindley, N.D.F. (1982) J. Bact. 152, 1211-1219

(39) Nakajima, N., Ozeki, H., and Shimura, Y. (1982) J. Biol. Chem. 257, 11113-11120

(40) Barnes, W.M. (1978) Proc. Natl. Acad. Sci. USA 75, 4281-4285

(41) Gemmill, R.M., Wessler, S.R., Keller, E.B., and Calvo, J.M. (1979) Proc. Natl. Acad. Sci. USA 76, 4941-4945

(42) Lebowitz, P., Weissman, S.M., and Radding, C.M. (1971) J. Biol. Chem. 246, 5120-5139

(43) Rosenberg, M., DeCrombrugghe, B., and Musso, R. (1976) Proc. Natl. Acad. Sci. USA 73, 717-721

(44) Dahlberg, J.E. and Blattner, F.R. (1973) in: Virus Research, Fox, C.F. and Robinson, W.S. (eds.), Academic Press, NY and London, pp. 533-543

(45) Luk, K.-C. (1982) Mol. Gen. Genet. 187, 320-325

(46) Luk, K.-C., Dobrzanski, P., and Szybalksi, W. (1982) Gene 17, 259-262

(47) Luk, K.-C. and Szybalski, W. (1982) Gene 17, 247-258

(48) Axelrod, N. (1976) J. Mol. Biol. 108, 753-770

(49) Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes,J.C., Hutchison III, C.A., Slocombe, P.M. & Smith, M. (1977) Nature 265, 687-695

(50) J. Drouin, unpublished; referred to in Godson, G.N., Barrell, B.G., Staden, R., and Fiddes, J.C. (1978) Nature 276, 236-247

(51) Sugimoto, K., Sugisaki, H., Okamoto, T., and Takanami, M. (1977) J. Mol. Biol. 111, 487-507

(52) Gentz, R., Langner, A., Chang, A.C.Y., Cohen, S.N., and Bujard, H. (1981) Proc. Natl. Acad. Sci. USA 78, 4936-4940

(53) Dunn, J.J. and Studier, F.W. (1980) Nucl. Acids Res. 8, 2119-2132

(54) Morita, M. and Oka, A. (1979) Eur. J. Bioch. 97, 435-443

(55) Stougaard, P., Molin, S., and Nordström, K. (1981) Proc. Natl. Acad. Sci. USA 78, 6008-6012

(56) Rosen, J., Ryder, T., Ohtsubo, H., and Ohtsubo, E. (1981) Nature 290, 794-797

(57) Kabsch, W., Sander, C., and Trifonov, E.N. (1982) Nucl. Acids Res. 10, 1097-1104

(58) Calladine, C.R. (1982) J. Mol. Biol. 161, 343-352

(59) Trifonov, E.N. and Sussman, J.L. (1980) Proc. Natl. Acad. Sci. USA 77, 3816-3820

(60) Trifonov, E.N. (1980) Nucl. Acids Res. 8, 4041-4053

(61) Gotoh, O. and Tagashira, Y. (1981) Biopolymers 20, 1033-1042

(62) Moreau, J., Marcaud, L., Maschat, F., Kejzlarova-Lepesant, J., Lepesant, J.-A. & Scherrer, K. (1982) Nature 295, 260-262

(63) Brendel, V., Pashnina, E.P., and Trifonov, E.N. (1982) Hoppe-Seyler's Z. Physiol. Chem. 363, 923 (abstract)

(64) Korn, L.J., Queen, C.L., and Wegman, M.N. (1977) Proc. Natl. Acad. Sci. USA 74, 4401-4405

(65) Stauffer, G.V., Zurawski, G., and Yanofsky, C. (1978) Proc. Natl. Acad. Sci. USA 75, 4833-4837

(66) Brosius, J., Dull, T.J., Sleeter, D.D., and Noller, H.F. (1981) J. Mol. Biol. 148, 107-127

(67) Siebenlist, U., Simpson, R.B., and Gilbert, W. (1980) Cell 20, 269-281

(68) Ineichen, K., Shepherd, J.C.W., and Bickle, T.A. (1981) Nucl. Acids Res. 9, 4639-4653

(69) Drahos, D. and Szybalski, W. (1981) Gene 16, 261-274

(70) Landsmann, J., Kröger, M., and Hobom, G. (1982) Gene 20, 11-24

(71) Luk, K.-C. and Szybalski, W. (1983) Virology 125, 403-418

(72) Post, L.E., Strycharz, G.D., Nomura, M., Lewis, H., and Dennis, P.P. (1979) Proc. Natl. Acad. Sci. USA 76, 1697-1701

(73) Barry, G., Squires, C., and Squires, C.L. (1980) Proc. Natl. Acad. Sci. USA 77, 3331-3335

(74) Friden, P., Newman, T., and Freundlich, M. (1982) Proc. Natl. Acad. Sci. USA 79, 6156-6160

(75) Sege, R.D., Söll, D., Ruddle, F.H., and Queen, C. (1981) Nucl. Acids Res. 9, 437-444

(76) Hawley, D.K. and McClure, W.R. (1983) Nucl. Acids Res. 11, 2237-2255

(77) Duester, G.L. and Holmes, W.M. (1980) Nucl. Acids Res. 8, 3793-3807

(78) Holmes, M., Elford, R., and Duester, G., unpublished; referred to in Postle, K. and Good, R.F. (1983) Proc. Natl. Acad. Sci. USA 80, 5235 5239

(79) Lee, J.S., An, G., Friesen, J.D., and Isono, K. (1981) Mol. Gen. Genet. 184, 218-223

(80) Nakamura, K. and Inouye, M. (1979) Cell 18, 1109-1117

(81) Pirtle, R.M., Pirtle, I.L., and Inouye, M. (1980) J. Biol. Chem. 255, 199-209

(82) Gallant, J.A. (1979) Ann. Rev. Genet. 13, 393-415

(83) Ryan, T. and Chamberlin, M.J. (1983) J. Biol. Chem. 258, 4690-4693