RNA sequence and secondary structure requirements for rho-dependent transcription termination

William D.Morgan^{1,3}, David G.Bear^{1,2}, Bronwen L.Litchman² and Peter H.von Hippel¹

¹Institute of Molecular Biology and Department of Chemistry, University of Oregon, Eugene, OR 97403, and ²Department of Cell Biology, University of New Mexico, School of Medicine, Albuquerque, NM 87131, USA

Received 19 November 1984; Accepted 19 April 1985

ABSTRACT

The interaction of *E. coli* termination factor rho with the nascent RNA transcript appears to be a central feature of the rho-dependent transcription termination process. Based on *in vitro* studies of the rho-dependent termination of the transcript initiated at the P_R promoter of bacteriophage lambda, and on earlier studies, Morgan, Bear and von Hippel (J. Biol. Chem. **258**, 9565-9574, 1983) proposed a model defining the features of a potential binding site for rho protein on transcripts subject to rho-dependent termination. This model suggested that an effective rho binding site on a nascent RNA transcript should be: (i) >70-80 nucleotide residues in length; (ii) essentially unencumbered with stable secondary structure; (iii) relatively sequence non-specific; and (iv) located within a few hundred nucleotide residues upstream of the potential rho-dependent terminus. In this paper we examine the sequences and secondary structures of several transcripts that exhibit rho-dependent termination to test this hypothesis further. Unstructured regions of approximately the expected size and location were found on all the transcripts examined. Though several short specific sequence elements were found to occur in a very similar arrangement on the λ P_R- and λ P_L-initiated transcripts of λ phage, no such elements of sequence regularity were found on any of the other rho-dependent transcripts. The results of the sequence comparisons reported here strongly support the generality of the "unstructured binding site" hypothesis for rho-dependent termination.

INTRODUCTION

Termination of transcription in the bacterium *E. coli* is a complex process involving at least two different mechanisms. Factor independent termination occurs at template sites that contain two highly conserved elements (1), including a GC-rich region of dyad symmetry, followed by a stretch of several consecutive AT base pairs. The resulting transcript thus has a stable stem-loop structure immediately followed by 4 to 8 uridine residues at the 3' terminus (for reviews see 2,3). Establishment of these features has been useful for mechanistic studies of rho-independent termination, focusing attention on the role of the dyad symmetry element in elongation pausing, and on the dissociation of ternary complexes due to the instability of stretches of rU·dA base pairs in the DNA-RNA hybrid contained within the ternary transcription complex.

It has also permitted the prediction of possible rho-independent termination sites from nucleic acid sequence data alone.

At other specific termination sites rho protein is required for release of the transcript. Although several of these termini have been identified and sequenced, a lack of obviously conserved features has revealed neither the nature of the termination signal nor a plausible rho-dependent termination mechanism (e.g. see 4). Previous work has indicated that rho-dependent termination generally occurs at multiple positions within a specific region of the template (5-10). The exact location of the 3' termini of the transcripts is determined by RNA polymerase elongation pausing sites within this region; the pausing itself is only a function of the sequence specific interactions of the polymerase with the template and the nascent RNA, and is unaffected by rho (5,8-11).

Two lines of evidence show that rho also contributes to the process of termination site selection. First, while rho-dependent termination occurs at pausing sites within a certain template region, it does not occur at pausing sites proximal to that region under the same conditions (9,10). In addition, the release of nascent RNA from ternary transcription complexes by rho shows some specificity. Earlier work had shown that transcripts of a minimal length are required for preferential release (12,13). Furthermore, the regions within which rho-mediated nascent chain release occur have been shown (for at least one template) to be identical to those at which termination occurs (14). These results together suggest that a specific segment of RNA located proximal to the 3'-terminus of the nascent transcript may be required for rho-dependent termination.

Consideration of the physical and enzymic properties of the rho protein also provides some important clues to the nature of rho-dependent termination. The protein binds to RNA as a hexamer, covering about 72-84 nucleotides (19,20). Rho also displays an ATPase activity that is activated in vitro by single-stranded cytosine-containing RNA (15-18) and has been shown to be necessary for rho function in transcription termination (22). Although poly(rC) shows the highest affinity for rho and displays the highest cofactor activity for the ATPase, polymers containing only a small proportion of cytidine residues (5-10%) also bind tightly to the protein and are effective cofactors in the ATPase reaction (16,17,19). Except for a minimal cytosine content (and/or a cytidine-containing sequence) that is presumed to be required for tight rho binding and to activate the rho ATPase, there is no evidence to indicate that rho engages in any direct recognition of specific RNA sequences.

The studies summarized above have suggested that a long (72-84 nucleotide) stretch of cytidine-containing RNA, that is largely free of secondary structure, should provide a sufficient site for rho binding and the consequent activation of the ATPase required for rho-dependent termination. A sequence in the λP_R transcript that satisfies these requirements, located just prior to the region within which termination and transcript release occur, has been identified, and a model for the specificity of rho-dependent termination has been proposed (9,10,14,21). This model is also consistent with recent studies of rho-dependent termination in the trp operon (22).

To explore further the validity of this model we have analyzed the primary and potential secondary structures of RNA transcripts preceding several additional known and purported rho-dependent termination sites. The results of this analysis, and the implications for hypotheses of rho action in termination, are presented here.

PROCEDURES

Termination Site Sequences: We have examined essentially all defined *E. coli* and coliphage rho-dependent termination sites for which both the complete template sequence and the positions of the rho-dependent termini (either in vivo or in vitro) are known. In some cases termination is either heterogeneous, occurring at multiple positions (λP_R , trp t', ColE1), or the termini have only been mapped approximately (λP_L). Since our analysis is concerned with potential upstream sequences and structures that may be responsible for the primary interaction between rho and the transcript, the exact positions of the termini are not critical; rather we have focused on the regions upstream of the most proximal terminus, or of the approximate termination position. In several cases the involvement of these upstream sequences in rho action is indicated by genetic data (see Discussion). Thus the *S. typhimurium* hisG gene has been included in our analysis because genetic evidence suggests that a particular region of this sequence is required for strong rho-dependent polarity, although the exact termination site has not been established (23); *S. typhimurium* contains a termination factor that is molecularly and functionally similar to the *E. coli* rho factor (24). Rho-independent sites at which termination efficiency may be increased by rho were not considered. The t_{RO} termination site located 112-113 base pairs from the λP_R promoter (25) was also not included, since termination occurs at t_{RO} only at very high rho concentrations and low monovalent salt concentrations. Such conditions could lead to "artifactual" termination since excess rho (at

low salt concentrations) could "melt" the normally stable local RNA secondary structure and thus bind at a site that is normally protected from interacting with rho by secondary structure (see 9). This result points up the importance, in *in vitro* transcription studies, of comparing rho-dependent transcription termination under similar solvent conditions and protein concentrations.

For three of the sequences considered (λP_R , rho, and ColE1) the region analyzed for potential secondary structure extends from the site of initiation at the promoter to the termination site (or to the most proximal of multiple termination sites). It has not been directly shown that the attenuation observed in the rho gene leader sequence is rho-dependent, but we consider this to be likely given the other evidence showing autoregulation of rho expression (e.g. see 26). Due to the size of the transcript terminated at trp t', we have only analyzed the region between the translation termination codon of the trpA gene and the most proximal rho-dependent termination position. (We have assumed that translation of the trpA coding sequence normally prevents binding of rho within the trpA coding region.) In the λP_L transcription unit, we have considered the region between the translation termination codon of the N gene and the hypothesized location of the rho-dependent termination site, tL₁ (31). The region of the *E. coli* tyrT locus that we analyzed lies between the end of the actual tRNA^{Tyr}₁ coding sequence (127 nucleotides from the 5' end of the transcript) and the termination site; we have assumed that the noncoding leader sequence of this transcript (nucleotides 1-41) is too short to permit effective rho binding and that folding of the tRNA prevents rho binding there as well. For the hisG gene, the entire protein coding sequence was analyzed for potential rho binding sites that might result in translational polarity.

Sequence Homology and Secondary Structure Analyses: Elements of sequence homology, as well as regions of dyad symmetry leading to possible stem-loop structures in the RNA transcripts, were identified by the SEQ program (27) using a DEC 1091 computer. The SEQ program predicts the free energy of elements of potential secondary structure according to the rules of Tinoco and colleagues (28,29). G•U base pairing was allowed during the search. All dyad symmetries with a maximum loop size of 40 and a predicted negative free energy were included in the examination (see Figure 1). This criterion requires 5 or more base pairs in the stem, unless the stem is unstable due to the presence of several A•U or G•U base pairs. Less stable variants of a given structure are not shown, and some structures that overlap a much more stable dyad are omitted, although alternative possibilities are shown for some transcripts.

In addition potential structures with no limit set on maximum loop size were examined for all sequences except the hisG gene; no highly stable long-range interactions were found that would be expected to compete with the local dyad symmetries listed.

RESULTS

Secondary Structure: Sites of potential secondary structure in RNA transcript regions preceding the various rho-dependent termination sites are shown in Figure 1. From these data it appears that all of the termination sites are preceded by substantial regions of relatively weak secondary structure. The few dyad symmetries with moderate loop size within these regions are predicted to have free energies greater than (more unfavorable than) -5 kcal/mol. Various characteristics of these regions are listed in Table I.

The single-stranded regions identified in these analyses are variable in length. Three are shorter than the unstructured sequence of the λP_R transcript, suggesting that the latter may be longer than necessary. (This may reflect the possibility that rho binding to the initial part of this region of the λP_R transcript is prevented in vivo by translation of the cro gene.) Our rho binding model (10,14) suggests a total unstructured site length of 72-84 residues, corresponding to the postulated binding site size of a rho hexamer. (We note the formal possibility that this total unstructured length requirement could be discontinuous; i.e., it could be interrupted by one or more short regions of secondary structure that would then be "looped out" in binding to the rho hexamer.) The unstructured sequence on the rho transcript (as well as the transcript coded by SV40 (5), which shows rho-dependent termination in vitro, but is not included in Table I), is approximately 80 nucleotide residues in length and is thus consistent with the "unstructured binding site" model. The potential single-stranded region preceding the termination site of the tyrT transcript is shorter than expected on this basis, suggesting that a shorter stretch (~60 nucleotides) of RNA may be sufficient for rho binding under some conditions, or that an alternate secondary structure of the tyrT RNA may exist in which a longer unstructured sequence (~90 nucleotide residues) is exposed (see Table I). It does seem likely that rho binds to this region of the tyrT transcript, rather than elsewhere, since this sequence directly precedes the termination locus and there is no other extensive single-stranded region in the proximal portion of the transcript.

The base compositions of the predicted single-stranded regions are given in Table I. All these sequences have a significant proportion of cytidine residues. All are AU-rich, and all except the sequence from colE1 contain more cytidine than guanosine. The overall potential (weakly structured) single-stranded stretch within the colE1 sequence is unusually long (see footnote 1, Table I); thus it is possible that rho interacts with only a portion of it, so that the overall base composition of the sequence (showing low cytidine content) does not accurately reflect the rho binding site. The other sites, which are normally functional in vivo, share two common compositional characteristics: an unusually low proportion of guanosine residues and a high adenosine content.

Elements of Specific Sequence: We approached the hunt for elements of specific sequence in rho-dependent transcripts as follows. Reasoning that the λP_R and λP_L sequences may be most closely related evolutionarily, we used the SEQ program to examine the regions upstream of the 3' termini of these

A λP_R

```

10   20   30   40   50
AUGUACUAGS GAGGUUUUUU GGACACACCG AUAAACCCUGA AAGAUUUUUC
60   70   80   90   100
AAUCCGCCUUU GGGCAAAACA AGACAGCUUA AGAUUCUCCG GUUUAUCAA
110  120  130  140  150
GCGCGAUACA CAAGGCCAUU CAUCCAGGCC GAAAGAUUUU UUUUAUUAUA
160  170  180  190  200
AAACCGCAUG GAAGCGUUUA UGCGGAAGAG GUAAAGCCU UCCGAGUAA
210  220  230  240  250
CAAAAUAACA ACAGCAUAAA UAACCCCGCU CUUACACAUU CCAGCCUGCA
260  270  280  290  300
AAAGGCCAU CAAGUAACAC CACACCUAGG GUGUAUGCAU UUAUUUGCAU
310
ACAUUACAUAC AA

```

λP_L

```

10   20   30   40   50
UAGUUAUUAU GAUAUUGGCC UUGCCUUUAU CUCAUAUAUA UAUUGACAUU
60   70   80   90   100
AGCUUGGCAC ACUAUUCAGUC CAGUAAUAUA CCUACAUAAGG GAUAUAUAUA
110  120  130  140  150
UGCUUUCUAA UCCAUUCGGA AAAAGUUUUG UUCAACACAC CAAGCUCAAU
160  170  180  190  200
CAACUCACUA AUGUAUGGGA AUGUUUUUGA UGUAAACACA UAUCUCCUUG
210  220  230  240  250
CUCAUAUAAG GCGCGGCAC AAAACCAUAG AUUUGCUUUC UGUAAUGUUU
260  270  280  290  300
UGAAUAUUG AUUCGACUUU AUCCUUUUUC AUUUUAUAGC GUUUUUAUAG
310  320  330  340  350
CUUAUAUAGC UUAUAUGUGG CCGUGGCAAU AGCUGAUAUA CGAUUGACAU
360  370  380  390  400
UAUUUUGUAG CGAAAAUGCA AAGCAAGAAG CGAAUAUAGC CCACACAGCA
410  420  430  440  450
GGAAUACCGA UUCUUCUAAU AACAAUAUUA GGCACAUUAU CUUGGCGUUA
460  470  480  490  500
AAGCAGAAGU CCAACCCAGA UAAAGAUCAU AUACAUGUUU CUUCCAGAG
510  520  530  540
GUUCAUUCU GAACACUCU CCGAGAAUAA CGAGUUGAUC

```

E. coli trp t'

```

10   20   30   40   50
UAUCCCCACA CCGCCAGAU CCGUGGCGG CAUUUUAAAU UUCUUUAUUA
60   70   80   90   100
AAACCGGAAA AUUCUUAAU UCAUUUAUA UUUUAUUUU UACCCUUGC
110  120  130  140  150
CUUACCCCGG UCGAUUCGUG ACUUACGUCA UUUUUCCGCC CAACAGUAUU
160  170  180  190  200
AUAACAACAC AAUUUAUACC CGACACAUAA CACCAUAUAA AUCAUAUAU
210  220  230  240  250
UUCUUUAUUG CACUUUAUCC UGAGGUUAUU CUUAUUUAUU CCAGAUUGU
260  270  280  290  300
CCUUAUAUA UAUUUUCCU CUUAUUUUC GUUCCGCUUA AUUUUGACUA

```

E. coli tyrT

```

10   20   30   40   50
UCACUUUACA AAGUCCUGCA ACUUCGCCAC GAACCGCAA UUAUAUAUUC
60   70   80   90   100
UGCCCAUUGG GCGAAGAUAG AGAAGCUUGC ACCAAGUUUC GACUCGAGGG
110  120  130  140  150
CCAGCGAAGG ACCUUUGCGG GAGGCAAGCA CCGAAGGCGC GAACCGGACA
160  170  180  190  200
GCGCUAGUA AUCCUUCCCG CACCAACAUU ACUUUCAAUA GUCCCUAGAC
210  220  230  240  250
UCUCAACGCA ADCCGCAUUC AAAUAUUUUC CCGAUGCGGG GAAGGAUAGG
260
AAGCUUCCAC

```

B *S. typhimurium* *hisG*

ColE1

E. coli *rho*

Figure 1: Analysis of potential RNA secondary structures that precede rho-dependent termination sites. References that established the sequences and the presence of rho-dependent termination are given in Table I. Numbering system for sequences: λP_R : 1 = P_R promoter mRNA start. *E. coli* trp t': 1 = first nucleotide of the *trpA* gene translation termination codon. *E. coli* tyrT: 1 = nucleotide 127 of the transcript, and is the first nucleotide following the tRNA coding sequence. *E. coli* rho: 1 = mRNA start. *S. typhimurium* *hisG*: 1 = first nucleotide of the translation initiation codon, 900 = third nucleotide of the translation termination codon. λP_L : 1 = N gene translation termination codon (position 35039 of Sanger et al. (45)). ColE1: 1 = I_1 promoter mRNA start (-74 of Yamada et al. (47)). Termination occurs at or after the following positions, according to this numbering system: λP_R , > 290; trp t', > 257; tyrT, > 225; rho, 250-252; λP_L , 472-479; ColE1, 426. Arrows below sequences indicate dyad symmetries (potential RNA secondary structure); vertical arrows indicate most proximal termination positions; brackets indicate regions that appear to be involved in rho-dependent termination, from deletion mutation analysis.

TABLE I.
Potential Single-Stranded Rho Binding Sites

Sequence	Evidence for Rho-Dependent Termination	Length (in nucleotide residues) and Position of Potential Single-Stranded RNA Regions Preceding the Termination Site ¹	Base Composition of Potential Single-Stranded Region (Number of Residues of each type)			
			A	C	G	U
λ P _R (47)	<u>in vitro</u> transcription (7,9,10,30) <u>in vivo</u> transcription (48)	86 (195-280)	38	24	9	14
<u>E. coli</u> trp t' (6,42)	<u>in vitro</u> transcription (6) <u>in vivo</u> transcription (49) genetics (6,41)	123 (100-222)	42	33	11	34
<u>E. coli</u> tyrT (34)	<u>in vitro</u> transcription (34) <u>in vivo</u> transcription (39)	68-89 ² (134 or 155 to 223)	27(22)	30(25)	17(11)	13(13)
<u>E. coli</u> rho (50)	<u>in vivo</u> transcription (50) genetics (23)	81 (170-250) 101 (370-470)	28 26	20 28	12 19	21 28
<u>S. typhimurium</u> - hisG (51)	genetics (38,31) <u>in vitro</u> transcription (31)	125 (75-199) 73 (363-435)	43 30	25 17	17 13	40 13
λ P _L (45)	<u>in vivo</u> transcription, <u>in vitro</u> transcription (35)	67 (141-207)	16	11	25	15
ColE1 ³ (46)						

¹See Figure 1 for explanation of numbering systems. The positions of relatively unstable dyad symmetries that occur within these potential single-stranded regions are indicated below: λ P_R sequence: 244-248/254-258 (nutR). E. coli trp t' sequence: 223-232/243-253. ColE1 sequence: 208-235/257-281, 281-303/331-352, 337-357/370-387, 325-336/343-354, 357-379/391-411.

²E. coli tyrT sequence: alternate secondary structure can form at the positions 110-119/124-133 ($\Delta G = -7.5$ kcal/mol) or 111-121/144-155 ($\Delta G = -18$ kcal/mol). The second is more stable, but the first might form previously by sequential folding, extending the single-stranded region from 68 to 89 nucleotide residues. In this case, however, an additional weak structure might also form at position 137-144/164-170. The first set of base composition data is for the 68 nucleotide sequence, and the second set is for the 89 nucleotide sequence.

³The ColE1 sequence analyzed is the PstI deletion mutant of Ebina and Nakazawa (35), which contains nucleotides -74 to +148 and +147 to +1624 of the sequence determined by Yamada et al. (46). Termination at the same site has been demonstrated both in vivo and in vitro, although only the in vitro experiments directly showed involvement of rho.

Figure 2: Comparison of regions of the transcripts initiating at the P_R and P_L promoters of phage λ that show homologous short sequence-specific elements. These elements are boxed and labeled I through V. The λ_{P_R} sequence shown here runs between residues 229 and 312 of that represented in Figure 1, while that shown for λ_{P_L} runs between residues 103 and 154 in Figure 1.

transcripts for segments of sequence homology. Five such regions were found, and are designated by boxes as elements I-V in Figure 2. Element I differs by one base in the two transcripts; elements II-V are exact. Furthermore these segments occur in the same order in both transcripts, and, to some extent, with comparable spacings.

This set of homologous elements occurs in the λP_R transcript just upstream of the t_{R1} site, which has been identified as the major termination site of this transcript in vivo (48). For the λP_L transcript, this series of homologous elements are located approximately 320 bases upstream of the proposed λt_{L1} termination site (31). The first three homologous elements are contained within the Box A sequence and stem-loop structure of the nutR antitermination site (30,32,33). However these elements of the P_L transcript are not part of the region genetically identified as nutL. In both transcripts the majority of the sequence is predicted to be single-stranded, and not to be engaged in base pairing with other adjacent sequences.

We searched for elements II-V in the remainder of the transcripts; the results we obtained are summarized below and in Table II. The Box A sequence (element I) was difficult to search for, since it is known to vary by several bases in other known Box A sequences and has generally been identified only in the context of other genetically defined nut sites. Element II is not found in any of the other transcripts. The others were found as follows:

Element III -- G(A)AAAAG AND C(A)AAAAG: A frequently encountered sequence element in the rho-dependent transcript sequences is G(A)AAAAG (or C(A)AAAAG). We find GAAAAG in single-stranded regions of the col E1 and rho leader transcripts. In addition CAAAAG is found in a predicted

TABLE II
The Occurrence of λP_R - and λP_L -Detected Elements of Specific Sequence in
Other Rho-Dependent Transcripts

Transcript	Element ¹	Region and Predicted Local "Strandedness" of the Transcript
<u>E. coli</u> trp t'	III(GAAAAA)	200 nucleotide residues upstream of the first termination site; in a double-stranded region.
	V(CAAACAAA)	101 nucleotide residues upstream of the first termination site; in a single-stranded region.
<u>E. coli</u> tRNA ^{tyr}	III(CAAAAG)	38 nucleotide residues upstream of 3' end; in a single-stranded region.
	IV(CACCACC)	53 nucleotide residues upstream of 3' end; in a single-stranded region.
	VCAAUCAA	At the 3' end in a single-stranded region
<u>E. coli</u> rho	III(GAAAAG)	98 nucleotide residues upstream from the proposed end of the proposed leader transcript; in a stem-loop structure.
	IVCACC	Crosses the proposed end of the transcript; in a single-stranded region.
Col E1	III(GAAAAG)	236 nucleotide residues upstream of the proposed 3' end of transcript; in a single-stranded region.
	V(CAAACAAA)	2 nucleotide residues downstream from the proposed 3' end of the transcript; in a single-stranded region immediately following a stem-loop.

¹Sequences in parentheses represent one base deviations from the consensus elements shown in Figure 2.

single-stranded region of the tRNA^{tyr} transcript; CAAAACG and GAAGAAG are found in predicted single-stranded regions of the hisG transcript, and GAAAAA is found in a predicted double-stranded region of trp t'.

Element IV -- CAC(C)ACC: The sequence CACCACC, in addition to being found in the λP_R and P_L transcripts, is also found in the rho leader transcript. CACCACC is found in the tRNA^{tyr} transcript.

Element V -- CAAUCAA and CAAACAAA: The sequence CAAUCAA is found in both the λP_L and the λP_R transcripts. It has also been found at the end of

the tRNA^{tyr} transcript (34). It has been proposed by several groups that such a sequence, or other loosely related sequences, might be common to all rho-dependent transcripts. In the absence of exact homologies, it is difficult to determine if these sequences are significant, or whether they arise solely as a consequence of the cytosine richness of the transcript termini. We have searched for such sequences in the other members of our group of transcripts; the CAAUCAA sequence was not found in any of them. However, we did find the sequence CAAACAAA in both the colE1 and trp t' transcripts. In the colE1 transcript, this sequence is found adjacent to the region predicted to be the 3' terminus of one of the major transcripts (35). However, in the trp t' terminator transcript this sequence is approximately 80 bases upstream from the first sites of a multiple-site termination region (6). We do find a sequence CCUCAA at the 3' end of the first major termination site of trp t', immediately following a predicted stem-loop structure; this sequence is also found in a predicted single-stranded region of the hisG transcript.

Other Possible Sequence Homologies: The regions of the transcripts that seem likely to represent rho binding sites (indicated as potential single stranded regions or by deletion analyses, as described below) were also examined pairwise in a general search for homology. No significant homologies, other than those listed above, were detected.

DISCUSSION

The presence of extensive stretches of unstructured sequences, or of sequences containing relatively weak potential secondary structure, in rho-dependent transcripts located proximal to regions where rho-dependent termination occurs, is consistent with the model that binding of a rho hexamer to a long stretch of single-stranded RNA is required for subsequent termination of transcription. The following detailed observations further support this hypothesis for specific transcripts.

(1) For the λ P_R transcript, the "unstructured" sequence directly precedes the boundary that separates the region of the template in which termination and release occur from the proximal region in which they do not; the effect of inosine substitution in the nascent RNA on both termination and release is also consistent with this interpretation (9,10,14). The interval between nucleotides 19 and 216 codes for the cro gene product and is efficiently translated. Therefore it is unlikely that rho binding here is involved in termination at tr₁ (termination site [2] of the five discrete rho-dependent termination sites located by Morgan et al. (9,10), for the P_R-initiated

transcript). Cloning of a 400 base pair fragment containing the t_{R1} termination site has demonstrated that the first 127 nucleotides of the P_R transcript are not essential for termination (36). Another fact that suggests that rho does not require sequences within the cro gene is the difference between λ and phage 434 DNA sequences: these phages have different cro gene sequences, but their sequences in the intercistronic cro-cII region (including t_{R1}), are almost identical (37).

(ii) The λP_L transcript contains two AU-rich potential single-stranded regions between the N gene and the termination site t_{L1} . The more proximal site is longer (124 nucleotides); however the other site (74 nucleotides) might also be sufficient. Genetic evidence suggests that the 393 base pair ninL99 deletion removes a site required for rho action at downstream termination sites (38). This deletion removes all of the proximal single-stranded region, and a portion of the distal one; thus the genetic evidence appears to be in good agreement with the analysis of potential rho binding sites contained in this paper.

(iii) This model also provides an explanation for the site specificity of rho-dependent termination for tyrT. In this transcript a 178 base pair sequence beginning within the tRNA coding sequence is repeated almost exactly 3 times. A rho-dependent termination site occurs within the second repeat, and at the corresponding position within the third repeat, but not within the first repeat (34,39). There are only a few differences between the nucleotide sequence in the repeats. However we note that a potential unstructured rho binding region occurs prior to the termination site in the second and third sites, but not in the first repeat. Therefore, the requirement for such a potential single-stranded region can explain the absence of termination within the first repeat.

(iv) Genetic evidence for both the trp t' and hisG sequences suggests that specific regions (defined by deletion mutations) are required for or are involved in rho-dependent termination and polarity (23,40,41,42). There is excellent agreement between the predicted single-stranded regions where rho might bind and the regions defined by genetic evidence for both the trp t' and hisG sequences. The region of weak RNA secondary structure found preceding the first of the multiple termination positions of the trp t' site stretches between nucleotides 97 and 255 (numbered from the translation termination codon of the trpA gene, see Figure 1; see also ref. 22). Deletions affecting rho-dependent termination also overlap the region between nucleotides 97 (deletion RT38, nucleotides 97-280) and 255 (deletion RT610, nucleotides 2-255).

(v) Within the hisG gene, the region between 370 and 500 base pairs from the AUG translation initiation codon seems to be important for rho-dependent polarity (23,40). The only extensive predicted single-stranded RNA region within the hisG gene occurs from nucleotides 370 to 470 (see Figure 1). A stable dyad symmetry is located distal to this segment, and this may induce RNA polymerase pausing that would lead to strong rho-dependent termination when the mRNA is untranscribed.

All these results and observations are consistent with the "unstructured binding site" hypothesis for rho action.

It should be clear that there are limitations to such sequence-based calculations to define potential sites of rho binding on nascent RNA transcripts. These include the difficulty of predicting RNA secondary structure. Thus free energy "rules" based on studies of short model oligonucleotides may not be valid for all sequences or for larger structures. It is also not easy to choose the most likely possibility from sets of alternate overlapping structures. Nascent transcripts may also fold sequentially, perhaps producing transient structures that differ from the most stable arrangements (43). Such factors could affect the extent and amount of internal secondary structure within potentially single-stranded regions of the various transcripts.

However (see also Introduction), physical and enzymic findings also support the "unstructured binding site" hypothesis of rho action. Thus the presence of extensive secondary structure in RNA has been shown to reduce the effectiveness of RNA as an ATPase activator of rho (18). In addition, rho has been shown to be a single-stranded RNA binding protein, in that it can "melt" at least some weak RNA secondary structure. Thus rho can fully saturate the random copolymer poly(rC,rI), under conditions in which the free polynucleotide shows appreciable base-paired secondary structure (D.G. Bear, unpublished results). Our inability to detect significant common recognition elements of specific sequences within the transcripts examined is also compatible with the effectiveness of a variety of cytidine-containing random synthetic polynucleotides in binding to rho (J.A. McSwiggen, unpublished results) and in activating its ATPase (15-18). Furthermore, the hexameric nature of the active form of rho (44) suggests that any specific recognition sequence should be multiply tandemly repeated; no such repeats are evident in the sequences analyzed here.

It is difficult to assess the significance of the various homologous sequence elements that we have picked out of the rho-dependent transcripts and summarized in Figure 2 and Table II. The fact that they occur in the same sequential arrangement in the two λ transcripts suggests either a functional

significance in λ (perhaps in antitermination?) or the conservation of an ancestral sequence homology between these two transcription units. These short sequence elements do not occur in the same arrangement in the other rho-dependent transcripts, but appear as relatively random and isolated elements in diverse locations relative to the rho-dependent termini. We believe their occurrence within the non-lambda transcripts to be fortuitous, rather than indicative of a functional role. Thus this search has provided no evidence for the occurrence of specific RNA sequences in potential rho binding sites.

We conclude, based on the present evidence, that the principle criterion for a potential rho binding site is, indeed, an unstructured sequence of appropriate length and that, as described elsewhere (10,21), the specificity of rho-dependent transcription termination depends on the availability of such a site on the nascent transcript, coupled with the occurrence of a sequence specific rho-independent pause of appropriate duration in the rate of transcription of RNA polymerase through specific sequences of the DNA template. Binding studies that strongly support the specific sequence predictions of the unstructured rho binding site hypothesis for the λP_R transcript (see 10,14) are currently in progress (J.P. Richardson and coworkers, and D.G. Bear and coworkers, private communications).

ACKNOWLEDGEMENTS

This research was supported in part by USPHS Research Grants GM-15792 and GM-29158 (to PHvH) and GM-32055 (to DGB). WDM was a pre-doctoral trainee of USPHS Training Grant GM-00715 and USPHS Institutional National Research Service Award GM-07759. At Oregon DGB was supported in part by USPHS Postdoctoral Fellowship GM-06676. Part of this work has been presented by WDM to the Graduate School of the University of Oregon in partial fulfillment of the requirements for the Ph.D. degree in Chemistry.

³Present address: Department of Biochemistry, University of California, Berkeley, CA 94720, USA

REFERENCES

1. Rosenberg, M. and Court, D. (1979) *Ann. Rev. Genet.* 13, 319-353.
2. Platt, T. (1981) *Cell* 24, 10-23.
3. Platt, T. and Bear, D.G. (1984) in *Gene Function in Prokaryotes*, Beckwith, J. and Davies, J., Eds., pp. 123-161, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
4. Lau, L.F., Roberts, J.W., Wu, R., Georges, F. and Narang, S.A. (1984) *Nucl. Acids Res.* 12, 1287-1299.
5. Reisbig, R and Hearst, J.E. (1981) *Biochemistry* 20, 1907-1918.

6. Wu, A., Christie, G.E. and Platt, T. (1981) *Proc. Natl. Acad. Sci. USA* 78, 2913-2917.
7. Lau, L.F., Roberts, J.W. and Wu, R. (1982) *Proc. Natl. Acad. Sci. USA* 79, 6171-6175.
8. Kassavetis, G. and Chamberlin, M.J. (1981) *J. Biol. Chem.* 256, 2777-2786.
9. Morgan, W.D., Bear, D.G. and von Hippel, P.H. (1983) *J. Biol. Chem.* 258, 9553-9564.
10. Morgan, W.D., Bear, D.G. and von Hippel, P.H. (1983) *J. Biol. Chem.* 258, 9565-9574.
11. Lau, L.F., Wu, R. and Roberts, J.W. (1983) *J. Biol. Chem.* 258, 9291-9297.
12. Richardson, J.P. and Conaway, R. (1980) *Biochemistry* 19, 4293-4299.
13. Shigesada, K. and Wu, C.-W. (1980) *Nucl. Acids. Res.* 8, 3355-3369.
14. Morgan, W.D., Bear, D.G. and von Hippel, P.H. (1984) *J. Biol. Chem.* 259, 8664-8671.
15. Lowery-Goldhammer, C. and Richardson, J.P. (1974) *Proc. Natl. Acad. Sci. USA* 71, 2003-2007.
16. Lowery, C. and Richardson, J.P. (1977) *J. Biol. Chem.* 252, 1375-1380.
17. Lowery, C. and Richardson, J.P. (1977) *J. Biol. Chem.* 252, 1381-1385.
18. Richardson, J.P. and Macy, M.R. (1981) *Biochemistry* 20, 1133-1139.
19. Galluppi, G.R. and Richardson, J.P. (1980) *J. Mol. Biol.* 138, 513-539.
20. Bear, D.G., McSwiggen, J.A., Singer, J.D., Feinstein, D.M., Morgan, W.D., Griffith, J.D. and von Hippel, P.H. (manuscript in preparation).
21. von Hippel, P.H., Bear, D.G., Morgan, W.D. and McSwiggen, J.A. (1984) *Ann. Rev. Biochem.* 53, 389-446.
22. Sharp, J.A. and Platt, T. (1984) *J. Biol. Chem.* 259, 2268-2273.
23. Ciampi, M.S., Schmid, M.B. and Roth, J.R. (1982) *Proc. Natl. Acad. Sci. USA* 79, 5016-5020.
24. Housley, P.R., Leavitt, A.D. and Whitfield, H.J. Jr. (1981) *J. Bacteriology* 147, 13-24.
25. Calva, E. and Burgess, R.R. (1980) *J. Biol. Chem.* 255, 11017-11022.
26. Kung, H-F., Bekesi, E., Guterman, S.K., Gray, J.E., Traub, L. and Calhoun, D.H. (1984) *Mol. Gen. Genet.* 193, 210-213.
27. Brutlag, D.L., Clayton, J., Friedland, P. and Kedes, L.H. (1982) *Nucl. Acids Res.* 10, 279-294.
28. Tinoco, I. Jr., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) *Nature New Biology* 246, 40-41.
29. Borer, P.N., Dengler, B., Tinoco, I. Jr. and Uhlenbeck, O.C. (1974) *J. Mol. Biol.* 86, 843-853.
30. Rosenberg, M., Court, D., Shimatake, H., Brady, C. and Wulff, D. (1978) *Nature* 272, 414-423.
31. Drahos, D. and Szybalski, W. (1981) *Gene* 16, 261-274.
32. Somasekhar, G., Drahos, D., Salstrom, J.S. and Szybalski, W. (1982) *Gene* 20, 477-480.
33. Olson, E.R., Flamm, E.L. and Friedman, D.I. (1982) *Cell* 31, 61-70.
34. Kupper, H., Sekiya, T., Rosenberg, M., Egan, J. and Landy, A. (1978) *Nature* 272, 423-428.
35. Ebina, Y. and Nakazawa, A. (1983) *J. Biol. Chem.* 258, 7072-7078.
36. Luk, K.-C. and Szybalski, W. (1982) *Gene* 20, 127-134.
37. Grosschedl, R. and Schwarz, E. (1979) *Nucl. Acids Res.* 6, 867-881.
38. Salstrom, J.S., Flandt, M. and Szybalski, W. (1979) *Mol. Gen. Genet.* 168, 211-230.
39. Rossi, J., Egan, J., Hudson, L. and Landy, A. (1981) *Cell* 26, 305-314.
40. Ciampi, M.S. Personal communication.
41. Guarente, L.P., Beckwith, J., Wu, A.M. and Platt, T. (1979) *J. Mol. Biol.* 132, 189-197.
42. Wu, A.M., Chapman, A.B., Platt, T., Guarente, L.P. and Beckwith, J. (1980) *Cell* 19, 829-836.

43. Nussinov, R. and Tinoco, I. Jr. (1981) *J. Mol. Biol.* 151, 519-533.
44. Finger, L.R. and Richardson, J.P. (1982) *J. Mol. Biol.* 156, 203-219.
45. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982) *J. Mol. Biol.* 162, 729-773.
46. Yamada, M., Ebina, Y., Miyata, T., Nakazawa, T. and Nakazawa, A. (1982) *Proc. Natl. Acad. Sci. USA* 2827-2831.
47. Schwarz, E., Scherer, G., Hobom, G. and Kossel, H. (1978) *Nature* 272, 410-414.
48. Court, D., Brady, C., Rosenberg, M., Wulff, D.L., Behr, M., Mahoney, M. and Izumi, S. (1980) *J. Mol. Biol.* 138, 231-254.
49. Wu, A.M. and Platt, T. (1978) *Proc. Natl. Acad. Sci. USA* 75, 5442-5446.
50. Brown, S., Albrechtsen, B., Pederson, S. and Klemm, P. (1982) *J. Mol. Biol.* 162, 283-298.
51. Barnes, W.M. Unpublished results.