The DNA sequence of the phage lambda genome between P_L and the gene bet

Kurt Ineichen, John C.W. Shepherd and Thomas A. Bickle

Microbiology Department, Biozentrum, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland

Received 17 July 1981

ABSTRACT

We have determined 3,400 base pairs of DNA sequence from the phage \( \lambda \) genome which starts to the right of P_L and runs to the left into the gene bet. The sequence thus includes the genes, N, ral, Eal0, cIII, kil and gam, as well as the transcription terminators TL1 and TL2. One surprising feature of the sequence is the presence in the region expected to be occupied by ral of a long open reading frame that, if it is expressed, would have to be transcribed from left to right, or counter to transcription from P_L.

INTRODUCTION

The region of the \( \lambda \) genome to the left of gene N has been reported to code for more gene products than could be accounted for by the space available in the DNA unless some of the genes were overlapping (for a recent review see ref. 1). Our interest in this region of the \( \lambda \) genome stems from the fact that it contains the gene ral which codes for a function that helps \( \lambda \) to overcome to some extent the full effects of host coded restriction systems\(^2,3\). For these reasons, we have determined the DNA sequence of the region in question. The sequence that we present here begins to the right of P_L at about map unit 74.0 and continues to the left till about map unit 67; it should include the genes N, ral, Eal0, cIII, kil and gam\(^1\). Some 700 base pairs of this sequence have been also determined in other laboratories, comprising the P_L region and the 5' end of the message transcribed from it\(^4,5\) and gene N\(^6\). Our results are in excellent agreement with these authors.
MATERIALS AND METHODS

The λcl857g7 strain that was used for the sequencing was brought to Basle in 1972 by R. Yuan from the laboratory of M. Meselson in Harvard University. Phages were prepared following heat induction of lysogens and were purified by two cycles of CsCl step gradient centrifugation. DNA was purified from the phage particles by phenol extraction and the phenol was removed by extensive dialysis.

All DNA sequencing was done with phage DNA; no cloned derivatives were used. In general, a large DNA fragment from the region of interest was isolated (for example: a BamHI/BglII fragment (see Fig. 1)) by agarose gel electrophoresis and purified following dissolution of the gel in saturated KI by chromatography on hydroxylapatite. The fragment was then digested with a fine cutting restriction enzyme, the 3' ends were labelled with the appropriate α-32P-labelled deoxynucleotide (Amersham) and the Klenow fragment of DNA polymerase I (Boehringer), the DNA was re-cut with another restriction enzyme and the labelled fragments separated by polyacrylamide gel electrophoresis. DNA was eluted from slices of polyacrylamide gel as described in ref. 10, except that the elution buffer was phenol saturated Tris/EDTA. DNA sequencing was done with the method of Maxam and Gilbert, and the products were separated on 8% or 17.4% polyacrylamide gels 0.5 x 350 x 430 mm in size.

RESULTS AND DISCUSSION

Sequencing strategy - The region of the λ genome that we have sequenced together with the sequencing strategy is shown in Figure 1. Eighty four percent of the sequence has been determined from both strands of the DNA and all of the secondary restriction sites that were used for sequencing except the BglII site have been sequenced across, so that we are sure that we have missed no small DNA fragments. The presence of 5-meC residues, which would be resistant to cleavage in the sequencing reactions, was monitored by digestion with the restriction enzyme BstNI, which cleaves at the sites of methylation in E. coli K12. Some
Figure 1. Sequencing strategy. The top line represents the phage λ genome, the region that we have sequenced is shown expanded in the centre of the figure and the sequencing strategy is shown at the bottom. A listing of restriction enzyme recognition sites in this region is available from the authors upon request.
Figure 2. The DNA sequence of the region of the \( \lambda \) genome defined in Figure 1.
regions of the DNA gave particular difficulties in interpreta-
tion when sequence was read from one strand. These difficulties,
which occurred in regions with the potential to form G + C-rich
secondary structure\textsuperscript{14}, could all be resolved from the sequence
derived from the other strand. We believe that the sequence which
we have determined, and which is shown in Figure 2, most likely
contains no errors.

\textbf{Coding capacity} - The proteins larger than 60 amino acids
that could theoretically be coded by this sequence in all six
frames (assuming an AUG start codon and the usual stop codons)
are shown in Figure 3. Two of these reading frames (those marked ?
in Figure 3) are unlikely to code for proteins. The amino acid
compositions that they predict bear no resemblance to those that
have been found for real proteins (for example, the amount of
serine in both of them would be over 12 moles %). They will
not be further considered here. The method described by
Shepherd\textsuperscript{15} when applied to the whole length of the genes pre-
dicts the correct reading frames for all of them, suggesting that
they are indeed protein-coding.

The rightmost of these reading frames has already been
shown by Franklin and Bennett\textsuperscript{6} to code for gene N and will not
be further commented here except to remark that there is not a
single nucleotide difference between their sequence and ours.
In the following paragraphs we discuss the reasons for
positioning the other genes shown in Figure 3.

There is some confusion in the literature regarding the
positions of genes in the left end of our sequence. According to
the last authoritative review of the \(\lambda\) genetic map\textsuperscript{1}, the gene order
should be the amino terminal of \textit{bet}, \textit{gam}, \textit{cIII}, \textit{Eal0} and \textit{ral} with
a sixth gene, \textit{kil}, mapping between \textit{gam} and \textit{cIII} and possibly overlapping
with both of them. The extreme left of the sequence is
occupied by an un terminated reading frame coding for 102 amino
acids which must represent the amino terminus of \textit{bet}. The \textit{gam}
gene product has been reported to have a molecular weight of
16,500\textsuperscript{16} and that of \textit{kil} 16,000\textsuperscript{17}. That they are different
genes (rather than two phenotypes of the same gene) is indicated
Figure 3. The genes in the DNA sequence of Figure 2. All open reading frames capable of coding for proteins longer than sixty amino acids are shown. For reasons that are discussed in the text, we do not believe that the two open reading frames marked ? code for proteins. The exact positions of these genes in the coordinates of Fig. 2 are: bet, from 308 to the left beyond our sequence, kil 610-317, gam 730-317, cIII 954-688, Eal0 1402-1037, ral 1785-1588, N 2858-2538 and git 1995-2521. The predicted amino acid sequences are shown in the lower part of the Figure, the c terminal part of gam which corresponds to kil is underlined.
by the fact that nonsense mutations have been isolated that affect each phenotype separately. We find only one open reading frame in this region that could code for a protein of this size (16,300) which we assign to gene gam. No other protein approaching this size could be coded in any other reading frame even allowing for unusual start codons. We note, however, that an internal methionine codon in the gam gene is preceded by a good Shine and Dalgarno sequence (5 bases long), which might be used for the production of an 11,600 molecular weight protein. A similar case of proteins overlapping in the same frame already exists for the C and Nu3 genes of λ (J. Shaw and H. Murialdo, quoted in (1)). We note also that the determination of the molecular weight of the kil product was done using phage that carry the bio substitution, which removes the carboxyl end of the gam gene. It is possible that the kil gene product identified in ref. 17 was initiated at the internal methionine codon and that translation continued into RNA coded by the bio substitution to produce the 16,000 molecular protein. The fact that gam\textsuperscript{−} kil\textsuperscript{+} nonsense mutations can be isolated could be explained if these mutations lay in the portion of the gam gene that precedes the kil gene. It is more difficult to explain the properties of the gam\textsuperscript{−} kil\textsuperscript{−} ochre mutant that has been described.

The gene mapping immediately to the right of kil is cIII which produces a protein whose molecular weight has never been determined. In this region we find an open reading frame that could code for a protein of 10,100 molecular weight preceded by a good Shine and Dalgarno sequence (7 bases long). The carboxyl end of this gene would overlap with the amino terminal end of gam to the extent of 14 codons, again fulfilling the prediction that some overlapping of genes in this region must occur. The only other gene product in this region for which a molecular weight has been reported is Eal0, which in different publications has been said to have a molecular weight of 10,000\textsuperscript{19}, 12,500\textsuperscript{20} and 16,000\textsuperscript{1}. We find an open reading frame in the region where Eal0 is to be expected, that is, to the right of cIII preceded by a Shine and Dalgarno sequence 4 bases long, which
would produce a protein of molecular weight 13,800.

There is only one short open reading frame to the right of Ea10 in the transcript from P, that could code for ral. This would code for a small protein with a molecular weight of 7,600. Again, it has a reasonable Shine and Dalgarno sequence (4 base pairs).

An unexpected finding given the genetic crowding of this region was that between the carboxyl terminus of the N gene and the beginning of this reading frame are some 750 base pairs of sequence which contain no coding information in the transcript from P. Inspection of the sequence shows, however, that this entire region consists of a single open reading frame, the largest in the sequence, in the other strand of the DNA; that is, such that it would have to be transcribed from left to right. This is the only significant open reading frame in this strand of the DNA. If it were used, protein synthesis would most likely initiate at the second methionine codon (the first is not preceded by a Shine and Dalgarno homology while the second is) to produce a protein of 21,000 molecular weight whose termination codon would be contiguous with that of the N gene. Is this really a gene in the sense that it is transcribed and translated into protein? The probability of finding an open reading frame of this length in 3,400 base pairs of sequence by chance is low but finite (p = 0.017). This "gene" is unlikely to be transcribed early after infection because transcription from P would block any polymerase coming through from the left. It could represent a late gene, either transcribed from its own promoter or, in the extreme case, from the normal late promoter P'. The gene is unlikely to represent ral because: a) ral has been reported to be an early function and b) ral lies to the left of 21 junction in λ-imm21 hybrid phage and this junction lies in the middle of the new gene.

 Provisionally, we would like to suggest that the new gene be called git for gene involved in tro because it is a distinct possibility that it might be the second component (Ea10) in the Tro phenotype described by Georgiou et al. Briefly, this is a
phenotype exhibited by its cro" lysogens at high temperature. Although they are induced, these cells cease growth and most macromolecular synthesis and fail to produce phage. This phenotype was partially alleviated in phages carrying bio substitutions that destroy Ea10, leading these authors to suggest that this gene was one component of the phenotype, and completely abolished when the substitution extended into the carboxyl termina end of gene N, leading to the suggestion that N was a second component of the phenotype. We would like to suggest that an alternate explanation might be that, because the cro" phage fail to shut of transcription from P, the git gene that we describe here could never be transcribed^21 and that the second component of the Trop phenotype might well be the git product rather than the N gene product (the N gene product made by the bio substitution is active, although less stable than the wild type^23). In other words, rather than being due to an overproduction of Ea10 and N, the Trop phenotype would be due to an overproduction of Ea10 and an underproduction of git. Experiments are currently underway to test this hypothesis.

Transcriptional terminators - Two transcriptional terminators, T_L1 and T_L2', have been reported to exist in the region of the λ genome that we have sequenced^. Genes beyond these terminators are not expressed unless the anti-termination factor, N, is produced. The first of these, T_L1', lies downstream from N and is a rho-dependent terminator which has been mapped to 71.1 map units by transcriptional mapping^25 or about 0.7 units upstream by an estimate of the length of the terminated message^26. Oppenheim^27 has also deduced the presence of a terminator in this region from consideration of the kinetics of synthesis of the gene products of the operon. The second is reported to be rho-independent (M. Gottesman; A. Das & S. Adhya, quoted in ref. 1) and to lie in, or close to, gene cIII^25. Less strong evidence exists for the presence of yet a third terminator, called T_L0', which would map between the promoter and the first gene in the operon, N^25.

All of these results need to be re-evaluated in the light of
the recent finding that RNA polymerase often pauses at specific sites during RNA synthesis without terminating transcription\textsuperscript{28,29}. Such a pause would mimic a termination event in experiments that rely on quantitation of the amount of RNA, or of protein, produced from different parts of the operon\textsuperscript{25,27}. In this respect, only T\textsubscript{L1} has been definitely shown to be a terminator because the terminated message has been isolated\textsuperscript{26}.

Pause sites have been reported to be regions in which the newly synthesised RNA has the capacity to form strongly hydrogen bonded stem and loop structures; if these structures are followed by a stretch of uridyl residues in the RNA, they become rho-independent terminators\textsuperscript{30}. We have searched our sequence for such stem and loop structures and the results are shown in Figure 4. A curious regularity emerges in the distribution of these sequences. Each of the genes in the left hand operon, as well as \textit{git}, is preceded by such a region of potential secondary structure at a distance of 40 to 60 basepairs before the initiating ATG measured from the centre of the loop. Of these only three, those in front of \textit{EalO}, \textit{kil} and \textit{cIII}, resemble known rho-independent terminators in having a long stretch of thymidyl residues 3' to the stem and loop structure. (For reasons that will be discussed later, we believe that the region in front of \textit{kil} is most probably TL2). The region in front of the \textit{cIII} gene is especially interesting in that it has the capacity to form alternative and mutually exclusive secondary structure as shown in Figure 4 somewhat reminiscent of the structure proposed for the attenuator of the tryptophan operon\textsuperscript{31}. We note also that the \textit{nutL} site situated before the \textit{N} gene also has the potential of forming a stem and loop (Fig. 4) and might act as a pause site to account for the TL0 postulated by Salstrom and Szybalski\textsuperscript{25}.

Spotting rho-dependent terminators from sequence data is more difficult than spotting rho-independent ones because the sequence of fewer of them has been determined and they do not display many obvious common features in their secondary structure. However, as pointed out in ref. 30, a common feature found in the rho-dependent terminators so far sequenced, that is, \(\lambda\text{t}_{R1}\textsuperscript{32,33}, \lambda\text{t}_{R0}\textsuperscript{34} and the
Figure 4. Potential stem and loop structures. All of these potential structures are 40-60 base pairs 5' to the indicated genes. The coordinates are those of the first base in the stem from the 5' side.
tRNA\textsubscript{\textsc{tyr}} terminator\textsuperscript{35} is the presence of the sequence 5'-CAATCAA-3' (or for λ\textsubscript{T}R\textsubscript{0} the sub-sequence 5'-ATCAA-3') just preceding the stop site. Only one such sequence can be found in the region of the λ genome presented here, beginning at position 2391. Immediately downstream from this is a sequence with 17 out of 23 base pairs homology with the \textit{E. coli} trp attenuator; the significance of this is unknown. These homologies are shown below.

\begin{verbatim}
tRNA\textsubscript{\textsc{tyr}} CAACCGAATCCGCAATCAA
tR\textsubscript{1} TTTGCATACATTCAATCAA
tR\textsubscript{0} TATATCAAAGCGCGATCAA
 tL\textsubscript{1} ACACACCAAGCTCAATCAACTCACTAATGTATGGGAATTGTTTTGA
 trp att CTAATGAGCGGGCTTTTTTTTGA
\end{verbatim}

We would like to suggest that this region is the rho-dependent terminator, T\textsubscript{\textsc{L}1}. Its position fits very well with that expected for the size of the terminated transcript\textsuperscript{26} although it is somewhat upstream of the position predicted from a kinetic analysis\textsuperscript{25}. This could be explained by transcriptional pausing.

Our search for homologies with known terminators found two others in the sequence. One of these is a 19 out of 26 base pair homology with trp t' just down stream from the git gene in a likely position to terminate a message for that gene.

\begin{verbatim}
trp t' AATATATTTTCCCCCTATCTCTGT...3'
git 2547 AACTGATTTTCCCCTATTTTCTGGC...3'
\end{verbatim}

The second is just upstream from the stem-and-loop structure preceding the kil gene and is a 21 out of 31 base pair homology with trp t.

\begin{verbatim}
trp t CAATTTTACCGCGGTGGCCTTGACCGCGCGAC...5'
kil 632 CACTTTTATACCGGGGGGCTTCTACGGC...5'
\end{verbatim}

The significance of these homologies with known terminators is difficult to evaluate. However, taken together with their positions and previous knowledge of terminators in this region, we would like to suggest that the above homology represents T\textsubscript{\textsc{L}2}.
ACKNOWLEDGEMENTS

This work was supported by grants from the Swiss National Science Foundation.

REFERENCES
