71. D. L. Alkon *et al.*, *ibid.* **84**, 6948 (1987).
72. T. J. Nelson and D. L. Alkon, unpublished observations.
73. The work in H.R.'s laboratory was supported by grants from the National Institute for Arthritis, Diabetes, and Digestive Diseases (DK19813), and the National Institute of Heart and Lung (HL35849) at the National Institutes of Health and the Muscular Dystrophy Association. We thank B. Bank and J. LoTurco for preparing the figures and for sharing their recent findings with us and N. Canetti and A. DeCosta for editorial assistance.

Research Article

A Persistent Untranslated Sequence Within Bacteriophage T4 DNA Topoisomerase Gene *60*

WAI MUN HUANG, SHI-ZHOU AO, SHERWOOD CASJENS, RICHARD ORLANDI, REGINA ZEIKUS, ROBERT WEISS, DENNIS WINGE, MEI FANG

A 50-nucleotide untranslated region is shown to be present within the coding sequence of *Escherichia coli* bacteriophage T4 gene *60*, which encodes one of the subunits for its type II DNA topoisomerase. This interruption is part of the transcribed messenger RNA and appears not to be removed before translation. Thus, the usual colinearity between messenger RNA and the encoded protein sequence apparently does not exist in this case. The interruption is bracketed by a direct repeat of five base pairs. A mechanism is proposed in which folding of the untranslated region brings together codons separated by the interruption so that the elongating ribosome may skip the 50 nucleotides during translation. The alternative possibility, that the protein is efficiently translated from a very minor and undectable form of processed messenger RNA, seems unlikely, but has not been completely ruled out.

T4 DNA TOPOISOMERASE IS A PHAGE ENCODED TYPE II ATP-dependent (ATP, adenosine triphosphate) topoisomerase that is capable of changing DNA topology. The enzyme catalyzes transient double-stranded breaks in the DNA backbone through which the DNA strand is passed, resulting in the changing of DNA linking numbers in steps of two (*1*). The T4 enzyme is a complex formed by the products of three genes in the early region of the genome; gene *39* encodes the 60-kilodalton (kD) subunit (p39), gene *52* encodes the 50-kD subunit (p52), and gene *60* encodes the 18-kD subunit (p60) (*2*). These gene products are required for normal T4 DNA replication and may be involved in the initiation event of T4 chromosomal DNA replication (*2*), although the exact roles of these proteins in phage DNA metabolism and development have yet to be clearly defined. In order to obtain large quantities of the T4 protein for structural and functional analyses, we have been cloning the genes for these three T4 subunits in uninfected *Escherichia coli* cells and studying the expression products. The sequences of p39, which is the ATP-utilization subunit, and p52, which has the cutting and joining function, have been reported (*3*).

We now describe the cloning and nucleotide sequence of the T4 gene encoding the p60 DNA topoisomerase subunit. We made the surprising discovery that T4 gene *60* is a split gene with 50 nontranslated nucleotides present within the coding region of the gene. The gene can be efficiently translated even in *Escherichia coli* cells carrying clones of gene *60*, without added phage functions. Interrupted genes whose messenger RNA (mRNA) is processed by a group I self-splicing mechanism have been found in phage T4 and its related phages (*4*). However, unlike other split genes of T4, the expression of gene *60* cannot be easily accounted for by a simple removal of the interrupting sequence via self-splicing of the primary transcript. In fact, there is no indication that the interruption is removed. The sequence of the message in the neighborhood of the interruption suggests that a highly folded structure may be generated and could form a basis for an unprecedented type of posttranscriptional mRNA handling in which the translational machinery bypasses the interruption without removing it.

Cloning and expression of T4 gene *60*. T4 gene *60* is located downstream from gene *39*, another topoisomerase subunit gene, on the circular T4 genetic map (*5*). We have shown that gene *39* is located on a 3-kb Eco RI fragment of cytosine-containing T4 DNA (*3*). Using the 3-kb Eco RI fragment and the 3′-most Hind III–Eco RI subfragment as hybridization probes, we identified a 3-kb Hind III fragment as the overlapping fragment on which gene *60* is expected to reside. The 3-kb Hind III fragment was inserted into the Hind III site of pT7-5 (*6*) under the transcriptional control of a T7 RNA promoter. The recombinant plasmid, pT60-3, is capable of rescuing five mutants in gene *60* (Fig. 1) (*7*). Subcloning portions of the 3-kb insert showed that two of the gene *60* mutants, amE416 and amE1217, were rescued by the central Eco RI fragment alone, and the remaining three mutants were rescued by the neighboring 2.4-kb Eco RI–Hind III fragment (Fig. 1). This result shows that gene *60* spans the second Eco RI site.

In order to determine whether pT60-3 carries the entire coding sequence of T4 gene *60*, we put it into a host system in which T7 RNA polymerase can be provided from a second plasmid (*6*). In this expression system, T7 RNA polymerase, expressed from a bacteriophage lambda (λ) P_L promoter, and a temperature-sensitive λ

W. M. Huang, S.-Z. Ao, S. Casjens, R. Orlandi, and M. Fang are in the Department of Cellular, Viral and Molecular Biology, University of Utah Medical Center, Salt Lake City, UT 84132. R. Zeikus and R. Weiss are in the Department of Human Genetics and Howard Hughes Medical Institute, University of Utah Medical Center, Salt Lake City, UT 84132. D. Winge is in the Department of Medicine, University of Utah Medical Center, Salt Lake City, Utah 84132. S.-Z. Ao was on leave from Shanghai Institute of Biochemistry, Academia Sinica, Shanghai, China.

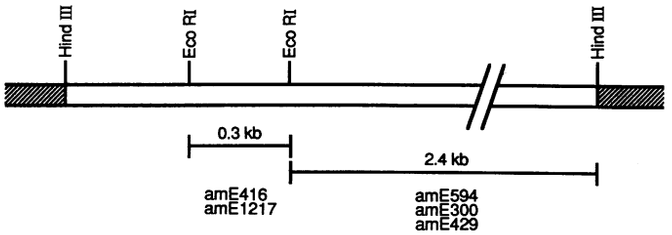

Fig. 1. Marker rescue of T4 gene *60* mutants by the recombinant plasmid pT60-3 and its derivatives. The shaded region represents the vector pT7-5 in which the 3-kb Hind III fragment from cytosine-containing T4 DNA is inserted.

repressor are cloned in a separate plasmid. At elevated temperature, T7 RNA polymerase is induced, allowing transcription and translation of the gene cloned under the control of the T7 promoter. Under induction conditions, pT60-3 overproduced a protein of approximately 18 kD, the expected monomeric molecular size of p60, as determined in SDS-polyacrylamide gels (Fig. 2). Finally, the overproduced protein was partially purified and shown to be the same as the p60 subunit of the complete topoisomerase by a Western blotting analysis in which antibody to the phage topoisomerase was used (*8*) (Fig. 2). Conversely, antibody to the 18-kD protein made from the plasmid reacted specifically with the p60 subunit of the phage enzyme (*9*). The purified cloned p60 was further shown to combine specifically with the other two purified subunits, p39 and p52, of T4 topoisomerase to reconstitute the ATP-dependent DNA relaxation activity (*9*).

DNA sequence of gene *60*. The 3-kb Hind III fragment and its subfragments were recloned into M13 derivatives, and sequences were determined by dideoxynucleotide sequencing (*10*) (Fig. 3B). The sequence of 1070 nucleotides starting from the 5′ Hind III site to a Dra I site is given in Fig. 3A. The beginning of an open reading frame starting from position 382 was identical to the NH_2-terminal seven amino acids of the p60 subunit of phage-induced enzyme (*2*). This establishes the location of the 5′ end of gene *60*. However, an in-frame TAG termination codon was present at position 520, which would only allow the synthesis of a 46–amino acid protein. A second open reading frame, which terminates at position 911, was found farther downstream. The combination of the two reading frames would, in theory, encode an 18-kD protein. In order to confirm that the sequence established by single-stranded M13 sequencing is indeed present in the plasmid pT60-3, which overproduces the 18-kD p60 protein, we prepared synthetic oligodeoxynucleotide primers; selected regions, including both strands of the interruption, were again sequenced with pT60-3 as the double-stranded template (Fig. 3B). The sequences determined in this way were identical to those obtained with the M13 derivatives as templates for the sequencing reactions. Since the above genetic data suggest that the coding region of the gene extends beyond the second Eco RI site (position 637), which is located 3′ to the interruption, it is apparent that some special mechanism is required to allow the expression of the split gene.

The gene *60* interruption is present in the T4 genome and its message. In order to determine whether the interruption found in the cloned T4 gene *60* is also present in the T4 genome, we sequenced the T4 gene *60* mRNA using AMV reverse transcriptase and a synthetic oligodeoxynucleotide primer (23 nucleotides) that hybridizes near the 3′ end of the gene (from position 677 to 655 of Fig. 3A) in primer-extension dideoxynucleotide sequencing analysis. We synthesized complementary DNA (cDNA) from mRNA isolated from T4 phage–infected cells. The resulting sequence (Fig. 4), which is complementary to the sequence in Fig. 3A, is identical to the sequence of the cloned gene. It starts in the 3′ region, includes the interruption and continues into the 5′ portion of the gene. A very prominent stop site for the reverse transcriptase is seen between 521 and 522. This position is within the interruption. The strong stop is also present in a reaction containing no dideoxynucleotide chain terminator (Fig. 4, lane O). Furthermore, the autoradiogram of the sequencing gel (Fig. 4) did not show any faint shadow bands, even after long exposure. If mRNA species having a different 5′ region and a common 3′ region were both present, cDNA sequencing with a primer that is complementary to the common region would give multiple, superimposed sequences depending on the relative abundance of the two mRNA species. The absence of another readable cDNA sequence across this region suggests that the isolated primary transcript is "stable" at 48°C (the temperature of cDNA synthesis), and that no processed message is detected in T4-infected cells. In a similar analysis, we examined the cellular transcript of T4 gene *60* made in a plasmid-containing cell under conditions where p60 is overproduced (Fig. 2); again no processed message was detected (*11, 12*). We estimate that if RNA processing were occurring at a 5 percent level it would have been detected.

Amino acid and protein sequencing analyses. In order to define the exact length of the interruption, we sought to partially determine the amino acid sequence of p60. The alignment of the independently derived protein sequence with the predicted amino acid sequence based on the DNA information should provide the location of the interruption. The p60 made from pT60-3 was purified from SDS-polyacrylamide gels and subjected to NH_2-terminal protein sequence analysis (Applied BioSystems model 470A and Beckman model 390D sequenators) (*13*). The first 49 amino acids were determined (Fig. 5). The protein sequence showed that Gly^{46}, encoded by the codon preceding the first stop codon, was followed by Leu-Gly-Ser. This result reveals that 50 nucleotides starting with the UAG stop codon were not translated. In order to confirm the protein sequence, we used two additional sequencing strategies.

During the purification of the cloned p60, it was found that 0.5*M* NaCl (or a higher concentration) was needed to stabilize the protein. A proteolytically cleaved peptide (approximately 15 kD)

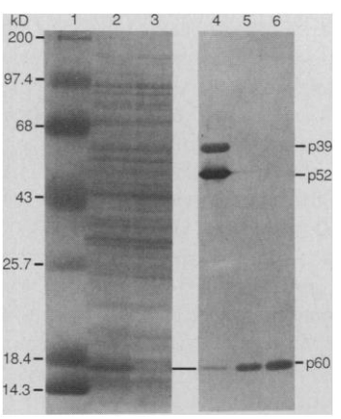

Fig. 2. Expression and properties of p60. Plasmid pT60-3 was used in a coupled T7 RNA polymerase-promoter system in which T7 RNA polymerase, made from a separate plasmid, is expressed from λ P_L promoter and regulated by a temperature-sensitive λ repressor gene (*6*). *Escherichia coli* cells harboring both plasmids were grown in M9 medium at 30°C to an optical density at 595 nm of 0.8. The culture was shifted to 42°C for 45 minutes, then held at 37°C for 2 hours, and finally harvested. Cells were lysed by a solution of 5 percent SDS, 3 percent 2-mercaptoethanol, 60 m*M* tris-HCl (*p*H 6.8), and 10 percent glycerol; the lysate was boiled for 3 minutes and loaded on a 10 percent SDS-polyacrylamide gel. The gel was stained with Coomassie blue. (Lane 1) Stained protein standards (BRL) of the indicated molecular size, (lane 2) induced culture, and (lane 3) uninduced culture. (Lanes 4, 5, and 6) In a separate experiment, p60 was partially purified from an induced culture (*9*), analyzed by Western blot analysis with antibody to purified T4 topoisomerase (*8*), and stained with horseradish peroxidase (Bio-Rad): (lane 4) purified T4 topoisomerase; (lanes 5 and 6) 1 and 2 μg of p60, respectively. The positions of the three T4 topoisomerase subunits are marked on the right.

was recovered if the protein preparation was stored in 0.2*M* NaCl. Western blot analysis indicated that the shorter protein was recognized antigenically as p60-related (*11*). The shorter protein was isolated from SDS gels, and its amino acid sequence was determined. The first 22 amino acids had been removed, and the shortened p60 starts at Arg^{23}. Amino acid sequencing starting from this point again showed that Gly^{46} was followed by Leu-Gly-Ser and continued on to Phe^{58} (Fig. 5). In the second approach, purified p60 was subjected to chemical cleavage by cyanogen bromide and subsequent trypsin digestion. Because of the large numbers of Lys (16) and Arg (7) residues present in p60, as shown by its amino acid composition (Table 1) and the predicted open reading frames, the digestion products should consist mostly of small peptides less than ten amino acids long (Fig. 5). The expected junction peptide between the two open reading frames lies within the largest expected peptide formed as a result of cleavage occurring at Met^{31} and Arg^{69} (Fig. 5); [the peptide bond Met^{39}-Thr^{40} is not expected to be cleaved efficiently (*13*)]. The doubly cleaved product was subjected to gas phase protein sequencing analysis. The first ten cycles of analysis yielded mixtures of amino acids, as would be expected from a mixture of peptides. From cycle 11 to cycle 27 clear amino acid determinations were made in that the shorter peptides no longer contributed to the sequencing output. In each of these cycles, two major amino acids were identified. From alignment with the predicted amino acid sequence, we deduced that two, superimposed sequences gave rise to the sequenator output. They derive from positions 42 to 58 and 82 to 98 (Fig. 5). The first peptide beginning at Ala^{42} is the result of cyanogen bromide cleavage at Met^{31}, yielding Ala^{42} at the 11th cycle. The sequence again shows that the amino acids that follow Gly^{46} are Leu-Gly-Ser- and so on. The second sequence is the result of trypsin cleavage at Arg^{71}, with Val^{82} in the 11th position. This peptide is apparently the product of partial tryptic digestion since some lysine residues were not hydrolyzed. The protein sequence across the interruption has been determined by three different measurements; they support the conclusion that 50 nucleotides, from 520 to 569 (Fig. 3A), are not utilized in the translation of the final protein product, and all the stop codons present in the middle of the coding sequence are therefore bypassed.

The COOH-terminal amino acids of p60 were identified by analyzing the digestion products of purified p60 by amino acid analysis after the protein was digested by carboxypeptidase A. In a time course analysis Ser was released first, and then Met (*11*). The Met-Ser at position 158–159 (Fig. 5) occurs only once in the predicted p60 sequence. The predicted COOH-terminal Gln was not found in the analysis. Other experiments suggest that it was modified.

The analysis of the DNA sequence and partial protein sequence indicates that p60 consists of 160 amino acids. The calculated peptide molecular size of p60 based on the DNA sequence is 18,632

A

```
                                                28                                                80
                       AGCTTCCTGGAATTAAAGGCATGACATGCTTTAAAAATGAAAAGGGCGAAGTTTTTCGTCTTCATGTTAATGATCCACGC

                                               135                                               187
ATTAAAACTGAAAAATTAGTTGGTGTTGGAAATACTGTTGCTGCTACTGCAAAAGCAGCTGAACTTGAAAAAGCGAAACCATGGTATAATAAATCAGCGACAAATCC

                                               242                                               294
AGAGGCTGTAAAATTAATACCAAATTTATATGAATGGTACGTCACGAAATATGACCCCGATCATTATAAACGAACGGGAGTTGCTAAATGGAAATCAGTTAATAATA

                                               349                                               396
TAACTGTGAATTCAAAATTATTCGGCCGAGCATTTAATGAATTTAAACGCGGTTGGATTCCTGATGAAAAGTTCTATGAGGTGTATA ATG AAA TTT GTA AAA
                                                                                         MET Lys Phe Val Lys

                                               438                                               477
ATT GAT TCT TCT AGC GTT GAT ATG AAA AAA TAT AAA TTG CAG AAC AAT GTT CGT CGT TCT ATT AAA TCC TCT TCA ATG AAC
Ile Asp Ser Ser Ser Val Asp Met Lys Lys Tyr Lys Leu Gln Asn Asn Val Arg Arg Ser Ile Lys Ser Ser Ser Met Asn

                                               519                                               569
TAT GCG AAT GTC GCT ATT ATG ACA GAC GCA GAT CAC GAT GGA TAGCCTTCGGGCTATCTATAGAAATACCTCATAATTAAGAGATTATTGGA
Tyr Ala Asn Val Ala Ile Met Thr Asp Ala Asp His Asp Gly

                                               611                                               650
TTA GGT TCT ATT TAT CCT TCT CTG CTC GGA TTT TTT AGT AAT TGG CCA GAA TTG TTT GAG CAA GGA CGA ATT CGC TTT GTC
Leu Gly Ser Ile Tyr Pro Ser Leu Leu Gly Phe Phe Ser Asn Trp Pro Glu Leu Phe Glu Gln Gly Arg Ile Arg Phe Val

                                               692                                               731
AAA ACT CCT GTA ATC ATC GCT CAG GTC GGT AAA AAA CAA GAA TGG TTT TAT ACA GTC GCT GAA TAT GAG AGT GCC AAA GAT
Lys Thr Pro Val Ile Ile Ala Gln Val Gly Lys Lys Gln Glu Trp Phe Tyr Thr Val Ala Glu Tyr Glu Ser Ala Lys Asp

                                               773                                               812
GCT CTA CCT AAA CAT AGC ATC CGT TAT ATT AAG GGA CTT GGC TCT TTG GAA AAA TCT GAA TAT CGT GAG ATG ATT CAA AAC
Ala Leu Pro Lys His Ser Ile Arg Tyr Ile Lys Gly Leu Gly Ser Leu Glu Lys Ser Glu Tyr Arg Glu Met Ile Gln Asn

                                               854                                               893
CCA GTA TAT GAT GTT GTT AAA CTT CCT GAG AAC TGG AAA GAG CTT TTT GAA ATG CTC ATG GGA GAT AAT GCT GAC CTT CGT
Pro Val Tyr Asp Val Val Lys Leu Pro Glu Asn Trp Lys Glu Leu Phe Glu Met Leu Met Gly Asp Asn Ala Asp Leu Arg

                                               941                                               993
AAA GAA TGG ATG AGC CAG TAG TTTACTTTACCACAAGGATGTGGTATAATTAATTGGGCAAATGAGGATATTGAAATGAAATCATATAAAGTAAATTTAG
Lys Glu Trp Met Ser Gln END

                                              1048                1070
AACTTTTTGATAAAGCAGTTCATCGAGAATATAGAATCATTCAACGCTTTTTCGATATGGGAGAAGCCGAAGAATTT
```

B

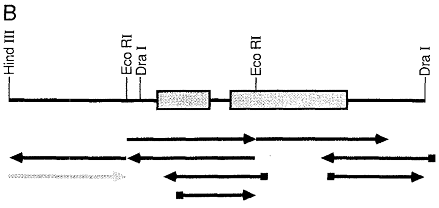

Fig. 3. (**A**) The nucleotide (DNA) sequence and the predicted amino acid sequence of T4 gene *60*. The Hind III–Dra I fragment is 1070 nucleotides in length. The beginning seven amino acids of the first open reading frame are identical to those at the NH_2-terminal of the p60 protein subunit isolated from the phage topoisomerase. This establishes that the ATG, starting at position 382, is the initiation codon of the protein. The shaded area from position 520 to 569 designates the interrupted region with the presence of stop codons in all three frames indicated by one, two, or three underline bars (only one of several possible precise interruption endpoint locations is shown). A second open reading frame is found starting from position 570 (based on protein sequencing, see Fig. 5) and extending to position 911. (**B**) The sequencing strategy was as follows. The fragment that contained gene *60* was analyzed by dideoxynucleotide sequencing (*10*) with single-stranded M13 derivatives as templates (simple arrows), by primer extension from selected regions within the insert with plasmid pT60-3 DNA as template (*25*) (arrows starting with a square), and by the chemical modification method of Maxam and Gilbert (*26*) (stippled arrow). Approximately 85 percent of the entire region was sequenced from both strands, and the remaining region was unambiguously sequenced twice with two independent isolates from a subcloning experiment. The boxed areas in the top line represent the coding regions of gene *60*.

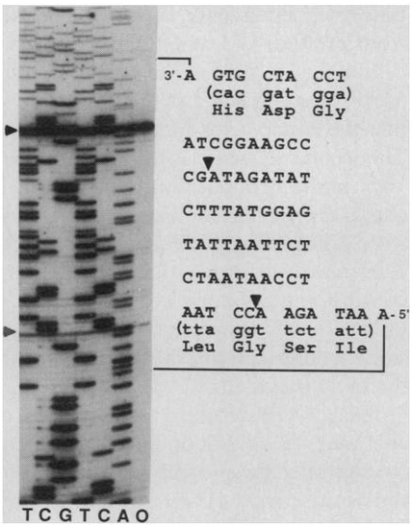

Fig. 4. Analysis of T4 mRNA by dideoxynucleotide sequencing and primer extension. Total T4 mRNA was isolated from T4 am4314 (defective in T4 DNA polymerase)-infected culture. The cells were infected at a multiplicity of 10 for 37°C for 15 minutes. They were then lysed with a mixture containing 2 percent SDS, 2.5 m*M* EDTA (*p*H 8), and boiled for 2 minutes. The preparation was centrifuged through a solution containing 5.7*M* CsCl and 0.1*M* EDTA (*p*H 8) (*27*); mRNA in good yield was recovered from the pellet. The precipitate was resuspended in water, treated with ribonuclease-free deoxyribonuclease (Promega), extracted with phenol and chloroform twice, and finally precipitated with ethanol. As primer, a 5′ end-labeled 23-base oligodeoxynucleotide, complementary to nucleotides 655 to 677 (Fig. 3A), was used to hybridize to the RNA. The AMV reverse transcriptase was then used to extend the primer (20 minutes at 48°C) in the presence of all four dNTP's (deoxynucleotide triphosphates, 0.6 m*M* each), and one ddNTP (dideoxynucleotide triphosphate, 0.2 m*M*). The dideoxynucleotide triphosphate added in the reaction mixture is indicated by A, G, C, and T, and O indicates no ddNTP added. The cDNA products were analyzed by electrophoresis in a 5 percent polyacrylamide gel containing 8*M* urea. The sequence of the indicated region is given on the right side of the autoradiograph. The complement of the coding sequence is shown. The codons for the amino acids are given in lower case within brackets below their complement. The filled and stippled triangles mark the positions of the major and minor reverse transcriptase stops, respectively.

daltons. These values are consistent with the independently determined amino acid composition analysis of the purified protein shown in Table 1. They are obviously different from those derived from the predicted 46–amino acid peptide encoded by the first open reading frame.

Gene *60* message synthesized in vitro does not self-splice. There are precedents in the T4 system for the existence of split genes. Transcripts of these genes are processed by a group I self-splicing mechanism (*14*). This is the case for the T4 thymidylate synthase gene and the gene for the small subunit of ribonucleotide reductase (*4, 15*). The cDNA sequencing analyses of the gene *60* message from the T4 infected cells and plasmid-containing cells suggest that splicing may not be the mechanism used to bypass the interruption present in the primary transcript of the gene (Fig. 4), unless the spliced mRNA represents a small fraction of the total message. Nonetheless, the possibility of processing of the RNA by a group I self-splicing reaction was further explored with a more sensitive assay in which RNA was synthesized from the cloned gene with T7 RNA polymerase. We used RNA's shorter than the full gene length so that all the expected fragments generated by self-splicing would be conveniently displayed in one high-resolution polyacrylamide gel.

In order to shorten the transcripts, we removed sequences unrelated to T4 gene *60* from pT60-3 (Fig. 1), and brought gene *60* closer to the T7 promoter. The 3-kb Hind III fragment containing gene *60* was digested with Dra I (which cuts at positions 339 and 1070 of Fig. 3A), and the resulting 741-bp fragment was cloned into the Sma I site of vector pT7-5. The resulting plasmid, pT60-32, contains all information necessary for p60 production (*11*), because a similar level of p60 protein was induced in pT60-32 as in the parent pT60-3 plasmid in the protein expression system described in Fig. 2.

In order to determine whether the truncated RNA's used below contain all of the information required to bypass the interruption, we inserted the β-galactosidase gene in-frame into gene *60* (3′ to the interruption) with a plasmid constructed by Weiss *et al.* (*16*). If the interruption is not bypassed, no full-length β-galactosidase can be made because the fused protein will terminate prematurely when it encounters the first termination codon in the interruption. The "reporter" system provided preliminary data suggesting that expression of β-galactosidase does not require the full-length gene *60*. In fact, the 335-bp Eco RI fragment (from nucleotides 303 to 638) is sufficient to allow efficient expression (*12*). In this case, the β-galactosidase activity is presumably due to translation initiation from the gene *60* AUG. This information suggests that sequences downstream from the Eco RI site at position 638 can be deleted without influencing the ability of the gene to effect the bypass event.

We next examined a runoff transcript of gene *60* which terminates at the Dde I site at 671 (3′ to the 638 Eco RI site and is thus expected to contain all the information required for bypass) for possible self-splicing. The Dra I–digested pT60-32 DNA supports the synthesis of a single transcript 359 nucleotides long (Fig. 6E, lane 2). This is the expected length for the primary transcript. If RNA processing had occurred, the mature mRNA would be expected to be 309 nucleotides long after the release of the 50-nucleotide interruption. As RNA size markers, we used RNA similarly transcribed from Bal I–digested pT60-32 and pT60-del 12 DNA as well as Dra I–digested pT60-del 12 and pT60-del 6 DNA (Fig. 6). Their primary transcript sizes are 303, 324, 48, and 432 nucleotides, respectively. Two of these transcripts (Fig. 6E, lanes 1 and 4) might also be potentially capable of splicing to yield shorter RNA species since they also contain the interruption. All RNA products appear as single-species transcripts with no sign of processing due to self-excision or ligation (or both). Previously characterized group I self-splicing reactions are apparently more efficient at higher temperatures (*4, 14*). Transcription reactions involving the gene *60* interruption gave identical results whether the reaction was carried out at 30° or 42°C. We also incubated the purified primary transcripts, prepared at either 30° or 42°C, at 45° or 60°C with guanosine triphosphate (GTP) and $MgCl_2$ at various concentrations, conditions where RNA self-splicing would be encouraged (*17*). Under these conditions again no new RNA species was found, showing no sign of self-splicing of the transcripts. Since in vitro group II self-splicing reactions require only Mg^{2+} ions as cofactor (*18*), we may extrapolate our inability to effect self-splicing to cover group II reactions. Examination of the DNA sequence in the region of the interruption further confirms such a contention, since there is no apparent sequence homology between the gene *60* sequence and the consensus sequences established for group I or group II self-splicing reactions (*14, 18, 19*). We believe autocatalyzed splicing is unlikely to be the mechanism of avoiding the untranslated 50 nucleotides in the synthesis of full-length p60.

Cell-free synthesis of p60. In order to better define the conditions under which the protein can be synthesized, we investigated the ability of the T4 gene *60*–containing plasmid DNA to direct in vitro protein synthesis. Both pT60-3 and its derivative pT60-32 (Fig. 6) are capable of directing the synthesis of an 18-kD protein (p60) in a coupled *E. coli* cell-free protein synthesizing system (Fig. 7). The addition of T7 RNA polymerase prior to the S-30 protein-synthesizing extract greatly stimulated the production of this protein (*11*), confirming that it is under the control of the T7 promoter. A second protein of approximately 7 kD was also made from the

pT60-3 plasmid. Its synthesis can be attributed to a region downstream of gene *60* since it was absent in the pT60-32–directed reaction products.

Proposed mechanism of translation. Our data suggest that the interruption in the DNA and in the message of T4 gene *60* cannot be explained by conventional RNA processing. Although spliced message was not detected by our analyses, it may still be possible that a functional message, devoid of the interruption, may exist in vivo, but is present at a very low level. The processing of this RNA would have to be the result of a novel form of RNA splicing, since there is no obvious sequence homology between the gene *60* interruption and other known introns (*14*). Furthermore, it seems improbable that a low abundance message could be responsible for the accumulation of p60 to 5 to 10 percent of total cellular protein (Fig. 2). Inspection of the DNA sequence shows that it is unlikely that any conventional protein synthesis initiation site (*20*) could be used to initiate translation in the second open reading frame. Also, p60 is readily synthesized in a cell-free system without special conditions (Fig. 7). We believe that models in which the two halves of gene *60* are translated independently and covalently joined later are also very unlikely.

A novel mechanism of translational control that involves the synthesis of a protein molecule from overlapping reading frames on a message has recently been described. The coupling of these reading frames requires programmed +1 or −1 ribosomal frameshifts (*16, 21*). This mode of gene expression has been described both in prokaryotes and eukaryotes with varying efficiencies. In *Escherichia coli*, ribosomal frameshifts from −2 to +6 nucleotides allow bypass of stop codons (*16*). In the case of T4 gene *60*, an in-frame stop

Table 1. Amino acid composition of T4 p60.

	Analysis*		I†		II‡	
	Residues	%	Residues	%	Residues	%
Ala	8.9	5.7	3	6.5	8	5.0
Arg	7.9	5.1	2	4.3	7	4.4
Asx	15.3	9.8	9	19.6	17	10.6
Cys	ND		0	0	0	0
Glx	18.2	11.7	1	2.2	18	11.3
Gly	8.9	5.7	1	2.2	8	5.0
His	2.1	1.4	1	2.2	2	1.3
Ile	8.9	5.7	3	6.5	10	6.3
Leu	13.8	8.9	1	2.2	12	7.5
Lys	14.4	9.2	6	13.0	16	10.0
Met	10.7	4.6	4	8.7	8	5.0
Phe	7.1	4.6	1	2.2	7	4.4
Pro	6.6	4.3	0	0	6	3.8
Ser	13.2	8.5	7	15.2	15	9.4
Thr	3.3	2.1	1	2.2	3	1.9
Trp	ND		0	0	4	2.5
Tyr	7.6	4.9	2	4.3	8	5.0
Val	12.2	7.8	4	8.7	11	6.9
Total			46		160	

*SDS-gel–purified material was hydrolyzed in 6*N* HCl for 24, 48, and 72 hours. The values were obtained by extrapolation to zero time. Values from three independent experiments agree to within 10 percent of the reported values. †Calculation based on the NH_2-terminal 46 amino acids. ‡Calculation based on the full-length p60.

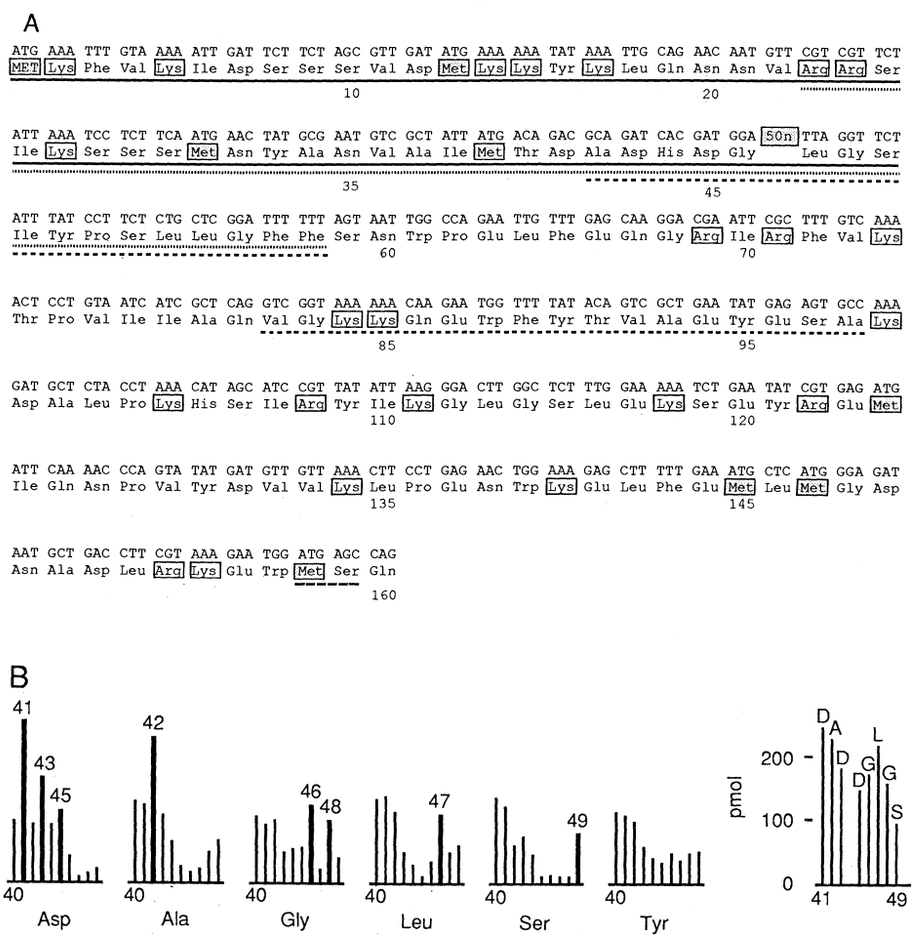

Fig. 5. Protein sequencing of p60. (**A**) Partially purified p60 was separated on 10 percent SDS-polyacrylamide gels, stained with Coomassie blue. The protein bands corresponding to p60 and its cleaved derivative were excised, extracted with 1 percent SDS, and concentrated by 25 percent trichloroacetic acid precipitation. Alternatively, SDS gel–purified p60 was subjected to cleavage by cyanogen bromide (in 70 percent formic acid and 0.17 m*M* tryptophan acting as a scavenger reagent) for 48 hours at room temperature; lyophilized product was then resuspended in 0.1*M* *N*-ethylmorpholine acetate (*p*H 8.5). Trypsin (TPCK, Cooper Biomedical) was added at a ratio of substrate to enzyme of 20:1, added in two portions and incubated for 3 hours at 37°C. These preparations were subjected to sequential NH_2-terminal analysis (Applied BioSystems 470A or Beckman 390D sequenator). The amino acid sequence obtained for each run is indicated as follows (complete p60, ——; shortened p60, ||||||||||; cyanogen bromide–trypsin cleaved, - - - -). The COOH-terminus of p60 was analyzed by treatment of the SDS gel–purified p60 with carboxypeptidase A (Sigma) at a ratio of substrate to enzyme of 25:1 at room temperature in 0.2*M* *N*-ethylmorpholine acetate (*p*H 8.5) and 0.1 percent SDS for times ranging from 30 minutes to 5 hours. The digested products were loaded either directly or after 20 percent acetic acid treatment (to precipitate residual undigested peptides) on a Beckman amino acid analyzer (results indicated by – – –). The positions of Met, Lys, and Arg are boxed for easy identification. The label, 50n, designates the location of the interruption. (**B**) An example of NH_2-terminal sequencing analysis of p60 from cycles 40 to 49. Each panel shows the relative yield of phenylthiohydantoin (PTH)-amino acids analyzed during the ten cycles beginning with number 40. The number above the vertical line represents the major PTH-amino acids recovered from that cycle. The last panel on the right is the summary of the sequencing output. Cycle 44, which is predicted to yield His according to the DNA sequence, gave a weak signal, and no assignment was made in this run. Other shorter runs which began from internal amino acids were used to make the assignment at this position. Ser was identified as dehydroalanine.

codon is found. However, about 50 nucleotides after the stop codon, a much larger region than in previously characterized frameshifting events, must be avoided. This would correspond to a giant leap for the ribosome. Thus, a different type of mechanism may be entertained. If translation is to bypass the interruption without excising it, the two codons bracketing the interruption might be expected to be brought into close proximity. Therefore, the RNA sequence in the region of the interruption was examined for possible secondary structure. A possible hairpin structure with the sequence CUUCGG forming the loop and a 10-bp stem (including one G-U pair) can be drawn which has a calculated value of $\Delta G = -15.1$ kcal/mol (*22*) (Fig. 8A). The stability of the stem loop is consistent with the observation that, when the mRNA from this region was used for cDNA synthesis by AMV reverse transcriptase, the enzyme frequently stopped at the third base 3′ from the closing base pair of the loop (Fig. 4). Similar AMV reverse transcriptase stop sites at similar locations relative to CUUCGG stem loops have been observed (*23*). This suggests that some type of secondary structure may exist in the mRNA. An unusual abundance of stable CUUCGG stem loops in the T4 genome in intergenic regions has also been noted (*23*).

The protein sequence suggests that the ends of the 50-nucleotide interruption in T4 gene *60* ***must*** occur within the repeated 5-nucleotide sequence, UGGAU, which is present at both ends of the interruption. For ease of discussion, we assume that the ends lie between codons 46 and 47 (Figs. 3 and 4). If translation is to bypass the interruption without excising it, these two codons, namely GGA for Gly[46] and UUA for Leu[47], might be expected to be closely juxtaposed in the three-dimensional structure of the gene *60* message. Such a highly folded structure, an alternative to that shown in Fig. 8A, is suggested in Fig. 8B. If it or a similar structure is stable in solution, it might allow the formation of a functionally uninterrupted

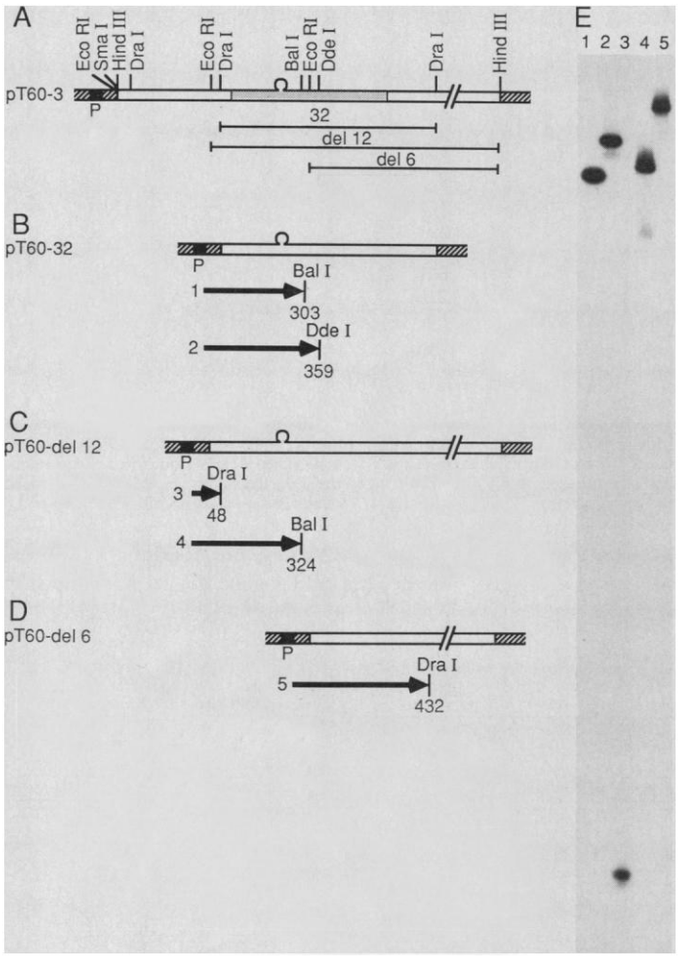

Fig. 6. Incubation of in vitro synthesized RNA under self-splicing promoting conditions. (**A**) Partial restriction map of the plasmid pT60-3 that contains T4 gene *60*. The vector sequences are striped and a filled box P marks the T7 promoter. The shaded area and loop mark the coding region of gene *60* and the position of the interruption, respectively. Regions of the 3-kb Hind III fragment used in subsequent construction of deletions are indicated below the map. (**B**) Derivative plasmid pT60-32 and its runoff transcripts. The 1070-bp Dra I fragment 32 [marked in (A)] was excised and reinserted into the Sma I site of pT7-5 to generate pT60-32. The resulting plasmid was cut with Bal I or Dde I and transcribed with T7 RNA polymerase generating transcripts 1 and 2, respectively. (**C**) Derivative plasmid pT60-del 12 and its runoff transcripts. pT7-del 12 was generated by deleting the first Eco RI fragment of pT60-3; therefore, it carries as insert the del 12 fragment [marked in (A)]. The resulting plasmid was cut with Dra I or Bal I, and transcribed with T7 RNA polymerase to generate transcripts 3 and 4, respectively. (**D**) Derivative plasmid pT60-del 6 and its runoff transcript. pT60-del 6 was generated by deleting two Eco RI fragments from pT60-3; therefore, it contains the del 6 fragment (A) as insert. The T7 RNA polymerase transcript 5 was generated by cutting the plasmid with Dra I. (**E**) Analysis of in vitro RNA. Electrophoresis was done in a 5 percent polyacrylamide gel containing 7*M* urea. Transcripts were prepared as described in (B), (C), and (D). Transcription reactions contained 40 m*M* tris-HCl (*p*H 7.5), 6 m*M* $MgCl_2$, 2 m*M* spermidine, 10 m*M* NaCl, 10 m*M* dithiothreitol, 20 units of RNasin, 0.5 m*M* each of the triphosphates ATP, UTP, and GTP, and 50 μCi of [^{32}P]CTP. Incubation was at 30°C or 42°C for 10, 30, or 45 minutes. Similar results were obtained under all tested combinations of these conditions. Lanes 1 to 5 are 42°C products derived from reactions whose templates and expected lengths of the runoff transcripts are similarly numbered and indicated in (B), (C), and (D).

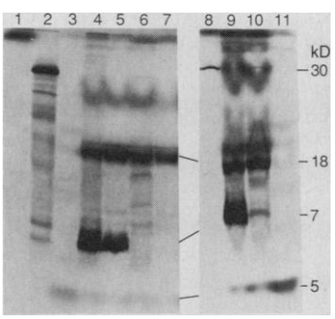

Fig. 7. Cell-free protein synthesis of p60. Plasmids pT60-3 and pT60-32 were used to synthesize [^{35}S]methionine labeled proteins with a commercially available DNA-directed coupled transcription-translation system derived from *E. coli* (Amersham). The protocol and reagents provided by the supplier were used except for the inclusion of a 25-minute preliminary incubation step, in which T7 RNA polymerase and its incubation buffer (described in Fig. 6E with 0.5 m*M* CTP replacing [^{32}P]CTP) were mixed with DNA at 30°C. Protein synthesis was continued at 37°C for 60 minutes. The products were analyzed in 13 percent SDS gels. (lane 1) No DNA. (lane 2) pAT153, a β-lactamase–producing plasmid provided by the supplier as a control. (lane 3) RNA, prepared from Bal I–digested pT60-32 (Fig. 6E, lane 1), was used in the translation reaction. (lanes 4 to 7) Products from reactions with a preliminary incubation to optimize RNA production from T7 promoter. (lanes 4 and 5) pT60-3 directed reactions; (lanes 6 and 7) pT60-32 directed reactions. In a separate experiment, shorter SDS gel running time and longer autoradiographic exposure were used to highlight the reaction products. (lane 8) pT7-5 the vector for the gene *60*–containing plasmids. (lane 9) Same as lane 5. (lane 10) Same as lane 7. (lane 11) Same as lane 3 except that three times more RNA than in lane 3 was used in the translation reaction. Purified p60 and stained protein markers were used as size markers to identify the 18-kD product. A small protein of about 5 kD was also observed from the gene *60*–containing plasmids (lanes 4 through 7 and 9, 10), but was absent from the reactions directed by the pT7-5 vector and other controls (lanes 1, 2, and 8). The molecular size of this protein is consistent with its being made from the NH_2-terminal open reading frame of the interrupted gene *60* (from nucleotide 382 to 519 coding for 46 amino acids); it was also synthesized in reactions in which isolated RNA transcript prepared from Bal I digested pT60-32 (Fig. 6E, lane 1) was used (lanes 3 and 11). According to its DNA sequence, the only peptide with as many as 46 amino acids that this transcript can encode in any frame, is the portion of gene *60* that is 5′ to the interruption. Further experiments are needed for absolute confirmation.

message on which ribosomes can bypass the interruption without requiring its removal. According to this hypothesis, it is possible that the secondary structure of the interruption rather than its specific sequence is important in allowing the translation to proceed across the interruption. It is also likely that interactions among sequences within the interruption and the adjoining regions may be required.

Because alternative pairing can occur anywhere along the duplicated UGGAU bases with identical results in the final protein sequence, we note that, in addition to the structure proposed in Fig. 8B, the proposal cannot distinguish whether coding information for the third U of Asp^{45}, for any part of the GGA of Gly^{46}, or for the first U of Leu^{47} derives from the UGGAU sequence at the 5′ or the 3′ end of the interruption. For all of these possibilities, the length of the interruption is identical and the different potential interruption sequences are specific circular permutations of one another; another variation might be the length of the CUUCGG stem in the suggested folded structure.

An 8-nucleotide region starting from position 545 in the interruption is complementary to the Shine-Dalgarno region of gene *60* (positions 369 to 376). The significance of this potential pairing is unclear. We have further explored this question by inserting the 5′ portion (including the interruption) of gene *60* into the middle of derivatives of the *lacZ* gene, which were constructed to use either its endogenous ribosome-binding site or that of *E. coli* lipoprotein (*16*). These experiments show that new Shine-Dalgarno sequences that share no complementarity with the T4 gene *60* interruption allow efficient expression of β-galactosidase (*12*). They suggest that pairing between the interruption and the translation initiation sequence may not be required for bypassing the interruption during translation.

β-Galactosidase activity was measured in constructs in which 5′ coding portions of gene *60* were inserted into the coding region of the *lacZ* gene. The levels of enzymatic activity are comparable in fused genes with and without the interrupted region (Table 2, lines 2 and 3). These values are similar to those obtained for *lacZ* constructs in which non-T4 DNA in-frame insertions are placed at a comparable positions (line 1) (*16*). These experiments suggest that expression from the interrupted gene is not due to the preferential translation of a very small subset of mRNA which may have been processed before translation, since the rate of translation initiation in all these constructs is expected to be similar. Although we have not completely ruled out the possibility that gene expression from mRNA's carrying the special gene *60* interruption involves splicing, we believe the latter possibility is therefore unlikely.

The interrupted message is unique. The 50-nucleotide interruption within the apparently functional message of T4 gene *60* appears to be the first and at present the only example of its kind. In the closely related phage T2, the interruption is absent. Although the T2 and T6 DNA topoisomerases are functionally and immunologically related to the T4 enzyme, they have different subunit structures (*8*). The T2 and T6 enzymes are defined by two genes, whereas the T4 enzyme complex is formed by products of three genes (*2*); in T2, sequences equivalent to T4 genes *39* and *60* are joined to form a gene that encodes the larger T2 p39. These structural arrangements and the similarity between T2 and T4 enzymes at the sequence level have been determined (*24*). In T2, the gene *39-60* message consists of one continuous open reading frame without interruption. Moreover, the protein sequence at the p60-interrupted region is completely conserved; the T4 junction amino acids Asp^{45}-Gly^{46}-Leu^{47} are encoded by adjacent codons in T2 DNA and mRNA. Although we do not know the evolutionary relation between phages T2 and T4, in addition to making gene *60* a separate gene in T4 with its own translation initiation signals, T4 also has the interruption in the gene documented in this article. Such an interruption can be expected to provide the cellular translational machinery with an added level of potential regulation. The discovery of the unusual decoding of the T4 gene *60* message may provide a new dimension in our appreciation of the translational machinery.

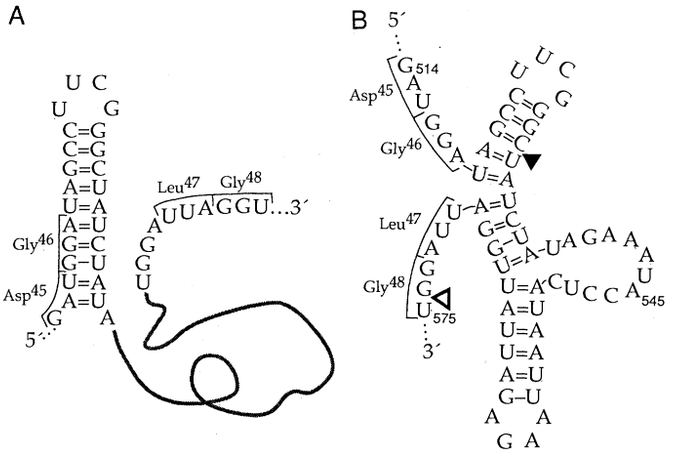

Fig. 8. Proposed secondary structures of T4 gene *60* mRNA in the interrupted region. The filled and unfilled triangles mark the positions of the major and minor AMV reverse transcriptase stops, respectively. Nucleotide numbers (indicated at the lower right-hand corners of a nucleotide) are defined according to Fig. 3A. An 8-base region, starting from position 545, is complementary to the gene *60* translation initiation region (Fig. 3A, position 369 to 376). The significance of this potential pairing remains to be determined.

Table 2. β-Galactosidase activity from *lacZ*–gene *60* fusion plasmids.

	Construct*	Gene *60* sequence†	β-Galactosidase activity‡
1	ATG AAA AGC AAT TCA	None	10,000 ± 100
2	ATG AAA AGC T AAT TCA	383–641 (interruption) is at 520–569)	7,000 ± 400
3	ATG AAA AGC TTG AAT TCA	567–641	10,600 ± 900

*Thick lines represent the NH_2-terminal coding sequence of *lacZ* derivatives that are controlled by an *E. coli* lipoprotein ribosome binding site (*16*). The dashed line represents the insertion of a 21-nt synthetic oligonucleotide to form plasmid 3p901 (*16*), which is used to provide the basal level of enzymatic activity. Thin lines represent T4 gene *60* sequence and the interruption is depicted by a loop. Underlined codons indicate the boundaries of the invariant regions. †The coordinates of Fig. 3A are used. ‡β-Galactosidase activity was measured in whole cell lysates (*28*). The values are averages of three measurements each of two independent transformants of each construction; error limits represent the ranges of values obtained.

REFERENCES AND NOTES

1. J. Wang, *Annu. Rev. Biochem.* **54**, 665 (1985).
2. G. Stetler, G. King, W. M. Huang, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 3737 (1979); L. Liu, C. Liu, B. Alberts, *Nature (London)* **281**, 465 (1979).
3. W. M. Huang, *Nucleic Acids Res.* **14**, 7751 (1986); *ibid.*, p. 7379.
4. F. K. Chu, G. F. Maley, D. West, M. Belfort, F. Maley, *Cell* **45**, 157 (1986); K. Ehrenman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 5875 (1986); J. Pedersen-Lane and M. Belfort, *Science* **237**, 182 (1987).
5. E. Kutter and W. Ruger, in *Bacteriophage T4*, C. Mathews, E. Kutter, G. Mosig, P. Berget, Eds. (American Society for Microbiology, Washington, DC, 1983), pp. 277–290.
6. S. Tabor and C. C. Richardson, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1074 (1985); pT7-5 is a derivative of pT7-1, and it was provided by S. Tabor.
7. S. Mufti and H. Bernstein, *J. Virol.* **14**, 860 (1974).
8. W. M. Huang, L. Wei, S. Casjens, *J. Biol. Chem.* **260**, 8973 (1985).

9. W. M. Huang and M. Fang, in preparation.
10. J. Messing and J. Vieira, *Gene* **19**, 269 (1982); M. Biggin, T. Gibson, G. Hong, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3963 (1983).
11. W. M. Huang and M. Fang, unpublished data.
12. R. B. Weiss and W. M. Huang, unpublished data.
13. R. Hewick, M. W. Hunkapiller, L. Hood, W. J. Dreyer, *J. Biol. Chem.* **256**, 7990 (1981); G. Allen, in *Sequencing of Proteins and Peptides* (Elsevier, Amsterdam, 1981).
14. T. R. Cech, *Science* **236**, 1532 (1987); *Cell* **44**, 207 (1986).
15. J. Gott, D. Shub, M. Belfort, *Cell* **47**, 81 (1986).
16. R. B. Weiss, D. M. Dunn, J. Atkins, R. F. Gesteland, *Cold Spring Harbor Symp. Quant. Biol.*, in press.
17. ^{32}P-labeled RNA products (Fig. 6) were treated with ribonuclease-free deoxyribonuclease, extracted with phenol, and precipitated with alcohol. The purified materials were used in a GTP-directed self-splicing reaction. Guanosine nucleotide involvement is one of the diagnostic features characteristic of group I self-splicing reaction. The reaction mixture contained 40 m*M* tris-HCl (*p*H 7.5), 0.5 m*M* GTP, and 10 m*M* $MgCl_2$. Incubation was at 45°C for 15 minutes or at 60°C for 10 minutes. These products were analyzed by electrophoresis in 5 percent polyacrylamide gels containing 7*M* urea. Alternatively, the in vitro RNA synthetic reactions, which were done at 45°C initially, were shifted to 60°C for 10 minutes, and the products were similarly analyzed. In the temperature shift experiments, the concentrations of GTP and $MgCl_2$ added at the beginning of the reactions were 0.5 m*M* and 6 m*M*, respectively. These experiments were designed to investigate the appearance of new RNA species as a function of temperature. It is expected that self-splicing of primary transcripts, if it occurs, would be incomplete at low temperature (*14*); therefore, more than one species of RNA would be found after self-splicing reactions. The absence of new RNA species after incubations at temperatures from 30° to 60°C as well as the fact that RNA transcripts shown in Fig. 6E have the expected length are taken as evidence that no significant RNA processing has occurred under these conditions.
18. A. Jacquier and F. Michel, *Cell* **50**, 17 (1987); C. L. Peebles *et al.*, *ibid.* **44**, 213 (1986); R. Van der Veen *et al.*, *ibid.*, p. 225.
19. In group I reactions, a U is invariably found at the 3′ end of the first exon, and a G is universal at the 3′ end of the intron. As is discussed below, the 5′ junction of the gene *60* interruption can occur at any point along the five nucleotides UGGAU surrounding the Gly^{46} codon (underlined). Thus, if the interruption were an intron, one of these U's could possibly serve as a 5′ junction at the end of the first open reading frame, but the nucleotide preceding the 3′ junction cannot be a G; it is a U in accordance with the protein sequence. In contrast, a G could serve as the sequence at the 3′ end of the interruption, then the last nucleotide of the first open reading frame has to be a G and could not be a U. By this analysis the group I self-splicing consensus cannot be satisfied.
20. G. Stormo, in *Maximizing Gene Expression*, W. Reznikoff and L. Gold, Eds. (Butterworth, Stoneham, MA, 1986), pp. 195–224.
21. T. Jacks and H. Varmus, *Science* **230**, 1237 (1985); T. Jacks, K. Townsley, H. Varmus, J. Majors, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4298 (1987); W. J. Craigen and C. T. Caskey, *Cell* **50**, 1 (1987).
22. I. Tinoco *et al., Nature (London) New Biol.* **246**, 40 (1973).
23. C. Tuerk *et al., Proc. Natl. Acad. Sci. U.S.A.*, in press. Curiously, CUUCGG stem loops are found near the beginning of the two other T4 topoisomerase genes, *39* and *52*, where they may be part of transcription terminators (*3*).
24. W. M. Huang and A. Gibson, in preparation.
25. R. Zagursky, K. Baumeister, N. Lomax, M. Berman, *Gene Anal. Tech.* **2**, 89 (1985).
26. A. M. Maxam and W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560 (1977).
27. V. Glisin, R. Crkvenjakov, C. Byus, *Biochemistry* **13**, 2633 (1974).
28. J. H. Miller, in *Experiments in Molecular Genetics* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1972), p. 352.
29. We thank Drs. Roger Hendrix, John Atkins, William Gray, and Ray Gesteland for stimulating discussions. Supported by NIH grant GM 21960 (W.M.H.) and a fellowship from the Helen Hay Whitney Foundation (R.W.). We thank D. Dunn for technical assistance.

21 October 1987; accepted 15 January 1988