*Proc. Natl. Acad. Sci. USA*
Vol. 82, pp. 6414–6418, October 1985
Biochemistry

# DNA sequence of the lactose operon: The *lacA* gene and the transcriptional termination region

**(nucleotide sequence/protein sequence/thiogalactoside transacetylase/termination of transcription)**

MATTHIAS A. HEDIGER*, DAVID F. JOHNSON†, DONALD P. NIERLICH†‡, AND IRVING ZABIN*‡

*Department of Biological Chemistry, School of Medicine, †Department of Microbiology, and ‡Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90024

*Communicated by Elizabeth F. Neufeld, May 31, 1985*

**ABSTRACT** **The *lac* operon of *Escherichia coli* spans approximately 5300 base pairs and includes the *lacZ*, *lacY*, and *lacA* genes in addition to the operator, promoter, and transcription termination regions. We report here the sequence of the *lacA* gene and the region distal to it, confirming the sequence of thiogalactoside transacetylase and completing the sequence of the *lac* operon. The *lacA* gene is characterized by use of rare codons, suggesting an origin from a plasmid, transposon, or virus gene. UUG is the translation initiation codon. A preliminary examination of 3′ end of the *lac* messenger in the region distal to the *lacA* gene indicates several endpoints. A predominant one is located at the 3′ end of a G+C-rich hairpin structure, which may be involved in termination of transcription or in post-transcriptional processing. An open reading frame of 702 base pairs is present on the complementary strand downstream from *lacA*.**

Since its description more than 25 years ago, the lactose operon has been a model system of great usefulness in biology. Indeed, study of one or another aspect of the lactose operon has touched on many of the most significant questions of biology. For example, the fundamental question of the mechanisms involved in expression of genes was first studied in this system (1). The discovery of the *lac* repressor and its binding to an operator site on DNA (2, 3) was one of the first problems concerning protein–DNA interactions to be examined. Studies of β-galactosidase in relation to mutants in *lacZ* have been important in defining many aspects of gene–protein relationships (4). Studies with fragments of β-galactosidase also have served as a model system for investigating protein–protein interactions (5). The lactose permease, the product of the second structural gene, *lacY*, was the first membrane transport protein studied extensively (6). Many fundamental concepts of the transport of molecules into the cell were derived from these studies. Thiogalactoside transacetylase, which is the gene product of *lacA*, the third structural gene, has been something of a mystery. The most reasonable interpretation of its function is that it is involved in detoxification (7). As indicated by the experiments of Andrews and Lin (7), cells containing the transacetylase are able to overcome inhibition of growth by thiogalactosides under certain conditions.

It is not surprising, therefore, that structures of the components of the lactose operon have also been investigated intensively. The amino acid sequence of the *lac* repressor was determined many years ago (8), and the amino acid sequence of β-galactosidase was reported in 1977 (9). Both of these proteins were examined by classical methods of protein chemistry. When methods for determining DNA sequences became available, the DNA sequence of the control elements of the lactose operon was one of the first to be studied (10). More recently, the DNA sequence of *lacZ* was determined (11), with the amino acid sequence of β-galactosidase as confirmation (12). The DNA sequence of *lacY* also has been reported (13). Only a minimum of protein chemical techniques was necessary to complete the amino acid sequence determinations of the lactose permease. However, until now no more was known of the structure of the lactose operon. We recently have determined the amino acid sequence of thiogalactoside transacetylase, primarily by standard methods of protein chemistry (14). We present in this paper the DNA sequence of *lacA* and of a segment of approximately a thousand bases downstream containing a probable transcription termination region, essentially completing the structural determination of the lactose operon.

## MATERIALS AND METHODS

**Bacteria.** *Escherichia coli* C600 (rk⁻ mk⁻ *thi thr leuB trpB*), used for transformation, was described by Nagahari *et al.* (15). *E. coli* JM107 and the cloning vectors M13 mp18 and mp19 were obtained from Pharmacia P-L Biochemicals. *E. coli* CR63 was from the laboratory collection.

**Enzymes and Chemicals.** DNA polymerase I (Klenow fragment) was purchased from Boehringer Mannheim. $^{35}$S-labeled adenosine 5′-[γ-thio]triphosphate (ATP[γ-$^{35}$S]) (500 Ci/mmol) was purchased from New England Nuclear. The M13 15-base primer was purchased from Bethesda Research Laboratories. T4 DNA polymerase was purchased from New England Biolabs.

**Plasmid Construction.** The source of the DNA used for construction of the M13 clones and for sequencing was the pBR322-derived plasmid pGM8, constructed by George Murakawa (unpublished data). This plasmid posesses an *Eco*RI–*Sal* I fragment carrying the distal portion of the *lacZ* gene and the *lacY* and *lacA* genes. It was derived from pMC1396 (16) by *Eco*RI cleavage and religation of that plasmid.

**Isolation of DNA Fragments.** *Aha* III and *Sal* I fragments of plasmid pGM8 were separated by using a recently reported apparatus for preparative gel electrophoresis (17). Details of the technique will be presented elsewhere. Fragments eluted from the gel were ethanol-precipitated and used for cloning into the appropriate restriction sites of M13 mp18 or mp19 and for DNA sequencing. Some fragments were further cleaved with other restriction enzymes (*Alu* I, *Hpa* II, *Hae* III, *Taq* I, *Nru* I, *Pvu* II, *Sau*3AI, and *Bst*EII) and the subfragments were randomly cloned into M13. *Bst*EII ends were made blunt ended for cloning by end-filling with Klenow enzyme. An end-labeled *Nru* I/*Sal* I fragment was prepared for chemical sequencing and nuclease S1 mapping as follows. Plasmid pGM8 was digested with *Nru* I, and the 3′ ends were labeled with [α-$^{32}$P]dGTP and T4 DNA polymerase in the presence of dATP, dCTP, and dTTP to a specific activity of $>2 \times 10^6$ cpm/pmol of 3′ ends. The DNA was then digested with *Sal* I, and the appropriate restriction

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: bp, base pair(s).

fragment was isolated from a 0.8% agarose gel by using the apparatus for preparative gel electrophoresis.

**DNA Sequencing.** DNA sequencing using dideoxynucleotides was performed as described by Sanger *et al.* (18) with modifications for the use of $^{35}$S-labeled deoxyadenosine 5'-[$\gamma$-thio]triphosphate as described by Biggin *et al.* (19). The final concentrations for the dideoxynucleotides (ddNTPs) in the polymerization reactions were 30 $\mu$M (ddATP), 40 $\mu$M (ddCTP), 50 $\mu$M (ddGTP), and 150 $\mu$M (ddTTP). The method used for chemical sequencing was described by Maxam and Gilbert (20). The glass plates of the sequencing gels were treated by the method of Garoff and Ansorge (21).

**DNA·DNA Hybridization.** Electroblot analysis was carried out using GeneScreenPlus (New England Nuclear) as transfer membrane.

## RESULTS

**Restriction Mapping of Cleavage Sites Distal to *lacA* on Plasmid pGM8 and *E. Coli* DNA.** The *lac* genes in pGM8 and its predecessor pMC1396 have a long and relatively complex geneology. The essential feature of this is that the *lac* genes were translocated on the *E. coli* chromosome to the $\phi$80 *att* site, near *trp*, as a lysogen, and were used in the construction of *trp-lac* fusions (22). Because of this, the source of the DNA sequences distal to *lacA* on plasmid pGM8 was unknown. This question was resolved by comparing restriction maps of the region distal to *lacA* of the plasmid pGM8 and the *E. coli* chromosome. Fig. 1 shows the blot hybridization of double digests of plasmid pGM8 and *E. coli* strain CR63 DNA. The two DNAs were cleaved with either endonucleases *Nru* I and *Pvu* II or *Nru* I and *Bst*EII. The digests were separated on an agarose gel and hybridized with an end-labeled *Nru* I/*Sal* I fragment (see *Materials and Methods*). The digests of both pGM8 and *E. coli* DNA show the same size of fragments, indicating that the whole *lac* segment in plasmid pGM8, except possibly for the 18 base pairs (bp) beyond the distal *Bst*EII site, represents *E. coli* DNA adjacent to the *lac* genes.

**Preparation of DNA Fragments for DNA Sequencing.** Large DNA fragments were prepared by cleaving plasmid pGM8 with *Aha* III and *Sal* I and by separating them with the new apparatus for preparative gel electrophoresis. Two fragments that contained the *lacA* gene and DNA sequences downstream from the *lacA* gene were obtained. These fragments as well as further cleavage products (*Alu* I, *Hpa* II, *Hae* III, *Taq* I, *Nru* I, *Pvu* II, *Sau*3AI, or *Bst*EII) were used for cloning and dideoxy sequencing. A diagram of the fragments is shown in Fig. 2. An end-labeled *Nru* I/*Sal* I fragment of plasmid pGM8 also was prepared and was used for chemical sequencing. This DNA sequence overlaps the sequences of the *Aha* III fragment and the *Aha* III/*Sal* I fragment (Fig. 2).

**DNA Sequence of the *lacA* Gene and the 3' End of the *Lac* Operon.** The complete DNA sequence of the *lacA* gene and the 3' end of the *lac* operon obtained from the fragments shown in Fig. 2 is presented in Fig. 3. This figure also includes the amino acid sequence of thiogalactoside transacetylase. The protein, exclusive of the initial formylmethionine, consists of 202 amino acids, and its sequence is in agreement with the results obtained by protein sequencing (14). The initiating codon for the *lacA* gene is UUG.

The sequence downstream from the *lacA* gene has been examined for palendromic sequences and for potential secondary structures of the mRNA, which may represent signals for termination of transcription. Fig. 4 shows three possible hairpin structures.

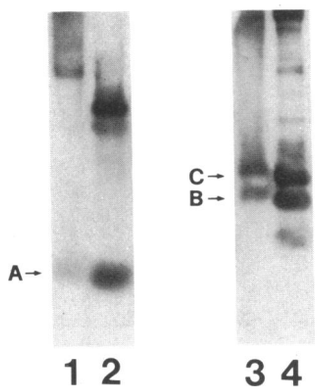


FIG. 1. Blot analysis used for the restriction site mapping of *Pvu* II (position 5217) and *Bst*EII (positions 5695 and 6211, respectively) on plasmid pGM8 and on *E. coli* CR63 DNA downstream from the *lacA* gene. Double digests with *Nru* I (position 5030) and *Pvu* II or *Bst*EII, respectively, were prepared and separated on a 1.5% agarose gel. Fragments were electroblotted to GeneScreen*Plus* and hybridized with an end-labeled *Nru* I–*Sal* I fragment. Tracks: 1, *E. coli* DNA, *Nru* I/*Pvu* II digest; 2, pGM8 DNA, *Nru* I/*Pvu* II digest; 3, *E. coli* DNA, *Nru* I/*Bst*EII digest; 4, pGM8 DNA, *Nru* I/*Bst*EII digest. The arrows indicate fragments of the same size in both pGM8 DNA and *E. coli* DNA: A, *Nru* I–*Pvu* II (187 nucleotides); B, *Bst*EII–*Bst*EII (516 nucleotides); C, *Nru* I–*Bst*EII (665 nuleotides). The heavy loading of the *E. coli* DNA lanes relative to the plasmid pGM8 lanes slightly retards the mobility of the fragments.

## DISCUSSION

We have determined the sequence of the DNA 1750 bp distal to the *lacY* gene of *E. coli*, a region that includes the *lacA* gene

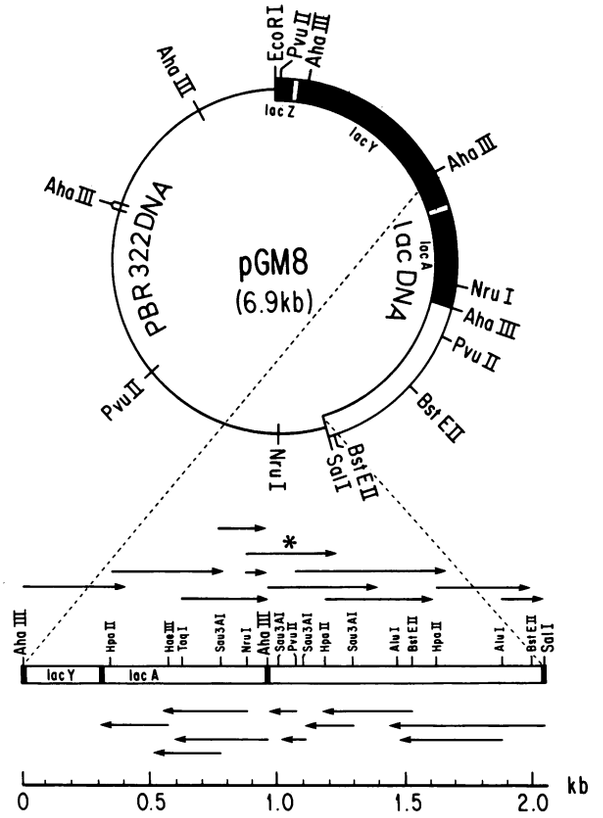


FIG. 2. Construction and restriction map of plasmid pGM8. The double segment represents the *lac* DNA segment. DNA fragments used for DNA sequencing are shown below. All fragments were used for Sanger dideoxy sequencing except an *Nru* I–*Sal* I fragment (marked with an asterisk) which was chemically sequenced as described by Maxam and Gilbert (20). For clarity some restriction cleavage sites of the enzymes *Hpa* II, *Hae* III, *Taq* I, *Sau*3A, and *Alu* I are omitted. kb, Kilobases.

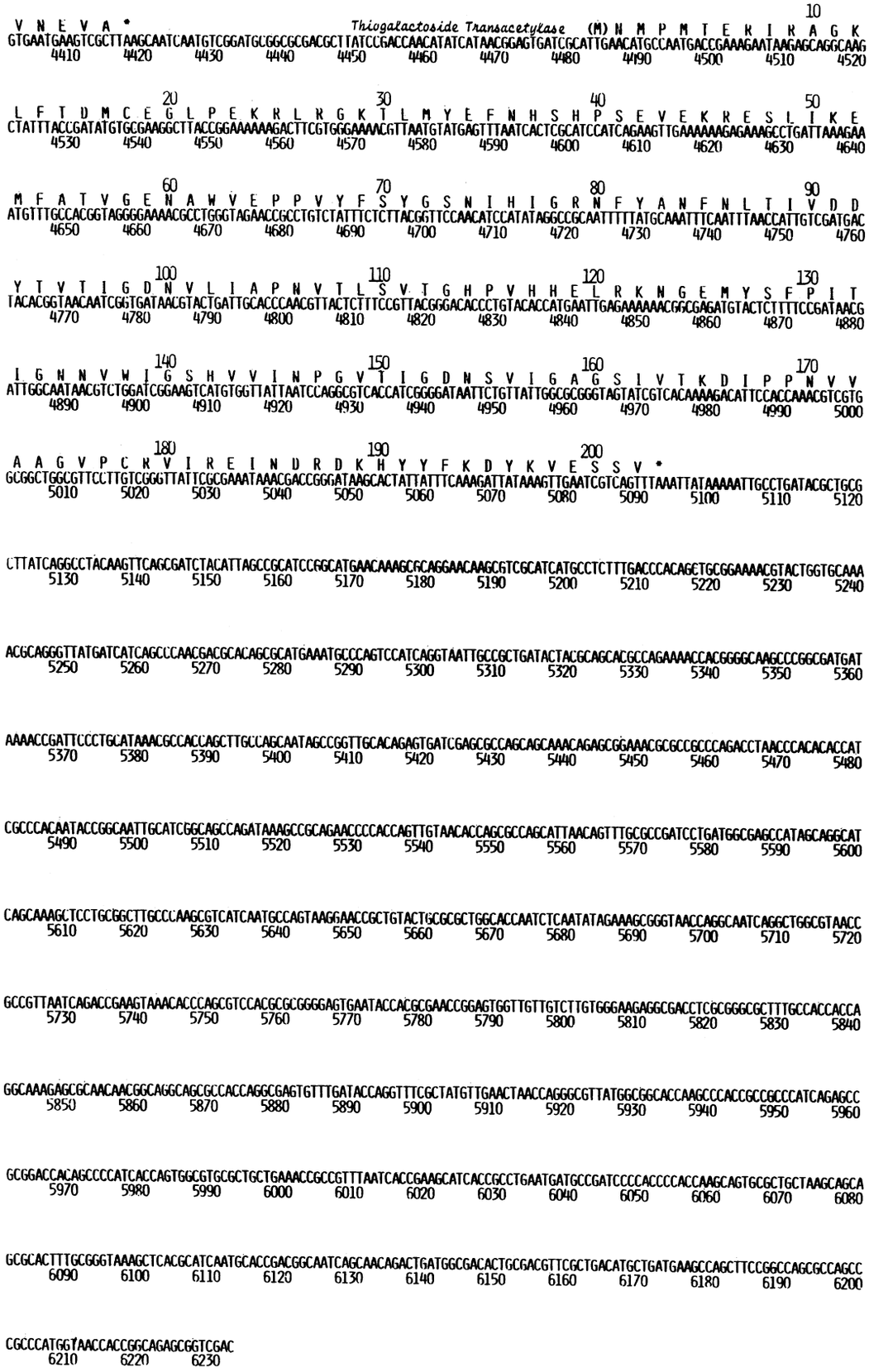


FIG. 3. Nucleotide sequence of the *lacA* gene and the 3′ end of the *lac* operon. The numbers correspond to the nucleotide positions in the Lac mRNA. The translational products of the nucleotide sequence, the end of lactose permease and all of thiogalactoside transacetylase, are shown.

and the region of transcriptional termination. This sequence completes that of the lactose operon, approximately 5300 bp, including the *lac* operator–promotor, the *lacZ*, *lacY*, and *lacA* genes, and the termination region. The sequence also confirms the sequence of thiogalactoside transacetylase, the *lacA* gene product, which was recently reported (14). The

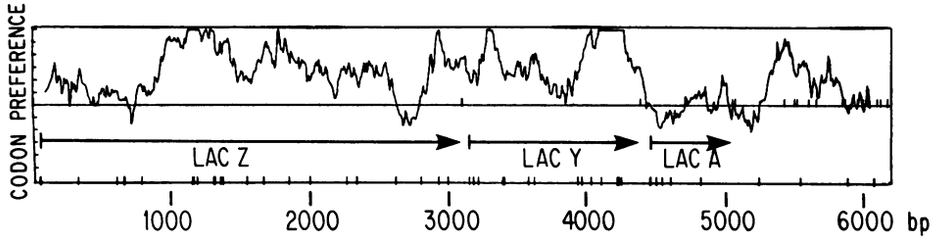


FIG. 4. Codon preference plot of the *lac* genes. Stop codons are marked on the medial horizontal line, and methionine codons are marked on the lower horizontal line.

DNA sequence was determined primarily by the Sanger dideoxy technique on DNA cloned into M13 bacteriophage; both strands of the DNA were sequenced. The strategy was based on the production of large DNA fragments, allowing the sequence to be deduced from relatively long overlapping segments. The sequence was confirmed by sequencing smaller subfragments. Isolation of fragments was facilitated by use of the new apparatus for preparative gel electrophoresis.

The *lacA* gene includes 203 codons (609 bp), including the initiation codon. As indicated previously by Büchel *et al.* (13), who sequenced the *lacY* gene and a short adjacent region of *lacA*, the initiation codon is UUG. We have confirmed this sequence with our independently isolated clones. Although UUG normally codes for leucine, determination of the amino acid sequence of the transacetylase indicates that it codes here for methionine (formylmethionine). UUG is unusual as an initiation codon; it is found only in three other genes of *E. coli* of the approximately 300 sequences that have been tabulated (23, 24). The UUG codon is preceded in the *lacYA* intergenic space by a good Shine–Dalgarno sequence, as shown in Fig. 3 (bases 4466–4478) (25).

The sequence of the *lacA* gene has been compared with those of the *lacI*, *lacZ*, and *lacY* genes. According to the gene evolution theory of Horowitz (26), the genes of a metabolic pathway may have evolved by a systematic duplication of genes. However, with the computer program DIAGON (27), only short and scattered homologies were found. A similar result was obtained when comparisons were carried out with the amino acid sequences of the Lac proteins.

The computer program SRCHN (28) has been used to compare the sequence of *lacA* with other sequences. The Genbank files of bacterial sequences distributed by Bolt, Beranec, and Newman (Cambridge, MA), dated February 1985, were used. One remarkable homology was found: 94 nucleotides on the noncoding DNA strand in the center of *lacA* (positions 4781–4688) are identical to a putative promoter region (bases 193–286) preceding the *E. coli* gene *proC*, which codes for pyrroline-5-carboxylate reductase (29). It is not evident what such a homology might represent. Since it is too extensive to be likely to have occurred by chance, we conclude that there is an error in the sequence determination of the *proC* promoter region. A segment of *lacA* DNA may have been incorporated by mistake into the *proC* region prior to or during the original construction of the *proC* sequencing clones. This interpretation seems reasonable because there is no substantiating protein sequence information, the *proC* and *lacA* genes are close (30), and the original construction of

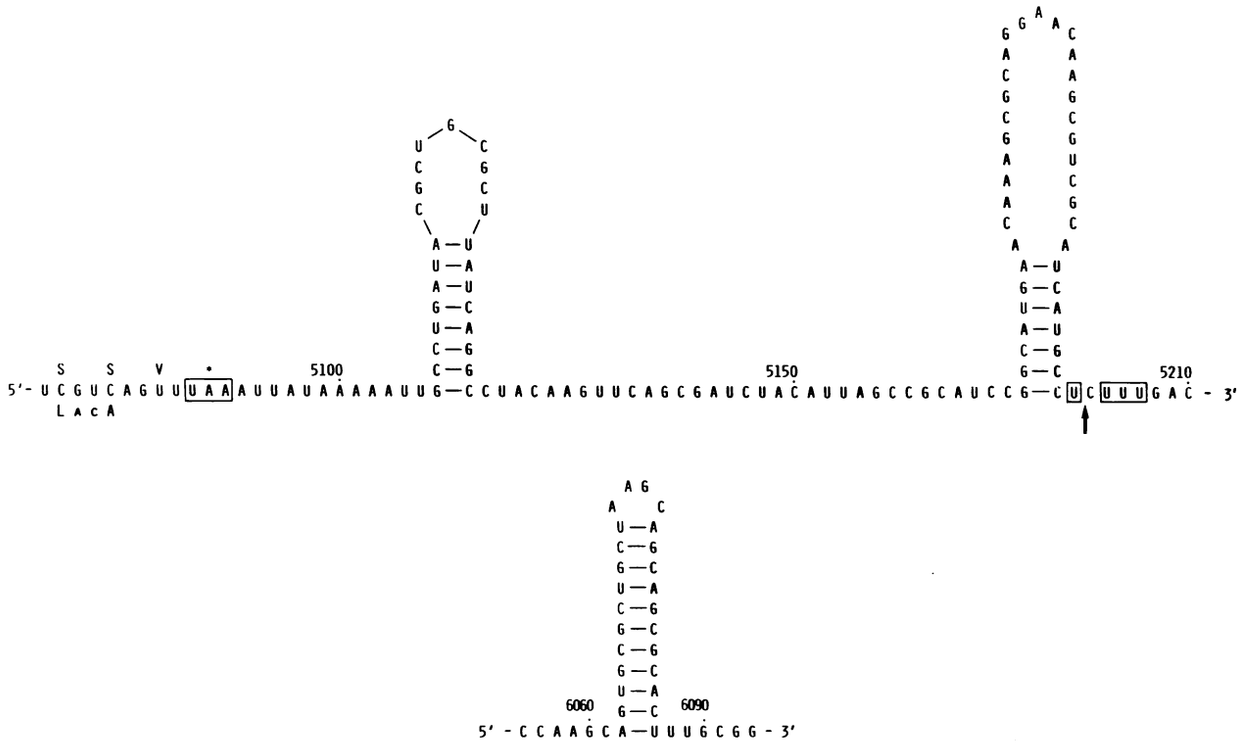


FIG. 5. Possible hairpin structures of the Lac mRNA downstream from the *lacA* gene obtained by secondary structure predictions. The arrow points to position 5203.

phage λp *phoA*$^+$, used for generating sequencing clones, involved a *lac* operon fusion (31, 32).

A number of workers have shown that the codon usage in a gene may be correlated with the relative abundance of the gene product and with the origin of the gene (33–38). In order to evaluate the codon usage of *lacA*, a codon preference plot of the *lac* genes as described by MacLachlan *et al.* (38) was prepared (Fig. 4). The figure shows a relatively low codon preference for *lacA*. The index for "optimal codon usage" calculated as described by Ikemura is 56% for *lacA* (35). This value is lower than the value for *lacI* (63%). Such low values have been found to be characteristic of weakly expressed genes, such as repressor, and of plasmid and transposon genes (35, 38), indicating that *lacA* may have its origin from such a source.

An open reading frame of 702 bp is present downstream from *lacA* on the complementary strand (position 5900–5199). It contains a potential translation initiation site (an AUG start codon at position 5600 and a possible Shine–Dalgarno sequence at position 5615–5603) such that it could encode a polypeptide of 134 amino acids.

The region downstream from *lacA* has been analyzed for possible hairpin structures of the Lac mRNA, which may be involved in termination of transcription or in a post-transcriptional processing. Three predominant hairpin structures of this region are shown in Fig. 5. Experiments on the mapping of the Lac messenger endpoints will be presented elsewhere. Results obtained so far indicate that there are several endpoints, including a prominent one at residue 5203 located at the 3′ end of a potential hairpin structure (Fig. 5).

It has been known for many years that there is a natural polarity of *lac* expression. This is manifest by the fact that *lacA* is expressed much more weakly than *lacZ* (39). Several features may contribute to this control: the *lacA* gene makes use of rare codons and translation of the gene is initiated with the unusual UUG codon. Two other features of the *lac* operon that may influence the polarity are a palindromic sequence (a *rep* sequence) in the *lacYA* intergenic space (40–42), and a probable rho-independent transcription attenuator in the *lacZY* intergenic space (40, 43).

The *lac* genes have served as models and tools for investigating many questions in biology. The results presented here, completing the DNA sequence determination of these genes, may allow the lactose operon to be even more useful in the future.

We thank George Murakawa for plasmid construction, Roger Staden for use of the computer programs, Robert Andersen and Larry Simpson for help with the computer analysis and sequencing procedures, and Audree Fowler for her comments and interest in the project. This work was supported by Public Health Service Grant AI-04181 from the National Institute of Allergy and Infectious Diseases.

1. Jacob, F. & Monod, J. (1961) *J. Mol. Biol.* **3,** 318–356.
2. Pardee, A., Jacob, F. & Monod, J. (1959) *J. Mol. Biol.* **1,** 165–178.
3. Gilbert, W. & Müller-Hill, B. (1967) *Proc. Natl. Acad. Sci. USA* **58,** 2415–2419.
4. Fowler, A. V. & Zabin, I. (1966) *Science* **154,** 1027–1029.
5. Zabin, I. (1982) *Mol. Cell. Biochem.* **49,** 87–96.
6. Kennedy, E. P. (1970) in *The Lactose Operon*, eds. Beckwith J. R. & Zipser, D. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 49–92.
7. Andrews, K. S. & Lin, E. C. C. (1976) *J. Bacteriol.* **128,** 510–513.
8. Bereuther, K., Adler, K., Geisler, N. & Klemm, A. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 3576–3580.
9. Fowler, A. V. & Zabin, I. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 1507–1510.
10. Gilbert, W. & Maxam, A. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 3581–3584.
11. Kalnins, A., Otto, K., Rüther, U. & Müller-Hill, B. (1983) *EMBO J.* **2,** 593–597.
12. Fowler, A. V. & Zabin, I. (1978) *J. Biol. Chem.* **253,** 5521–5525.
13. Büchel, D. E., Gronenborn, B. & Müller-Hill, B. (1980) *Nature (London)* **283,** 541–545.
14. Fowler, A. V., Hediger, M. A., Musso, R. E. & Zabin, I. (1985) *Biochimie* **67,** 101–108.
15. Nagahari, K., Tanaka, T., Hishinuma, F., Kuroda, M. & Sakaguchi, K. (1977) *Gene* **1,** 141–152.
16. Casadaban, M. J., Chou, J. & Cohen, S. N. (1980) *J. Bacteriol.* **143,** 971–980.
17. Hediger, M. A. (1984) *Anal. Biochem.* **142,** 445–454.
18. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143,** 161–178.
19. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3963–3965.
20. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
21. Garoff, H. & Ansorge, W. (1981) *Anal. Biochem.* **115,** 450–457.
22. Miller, J. H., Reznikoff, W. S., Silverstone, A. E., Ippen, K., Signer, E. R. & Beckwith, J. R. (1970) *J. Bacteriol.* **104,** 1273–1279.
23. Gren, E. J. (1984) *Biochimie* **66,** 1–29.
24. Alba, H., Mori, K., Tanaka, M., Ooi, T., Roy, A. & Danchin, A. (1984) *Nucleic Acids Res.* **12,** 9427–9440.
25. Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S. & Stormo, G. (1981) *Annu. Rev. Microbiol.* **35,** 365–403.
26. Horowitz, N. H. (1965) in *Evolving Genes and Proteins*, eds. Bryson, V. & Vogel, H. J. (Academic, New York), pp. 15–23.
27. Staden, R. (1982) *Nucleic Acids Res.* **10,** 2951–2961.
28. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 726–730.
29. Deutch, A. H., Smith, C. J., Rushlow, K. E. & Kretschmer, P. J. (1982) *Nucleic Acids Res.* **10,** 7701–7714.
30. Hadley, R. G., Hu, M., Timmons, M., Yun, K. & Deonier, R. C. (1983) *Gene* **22,** 281–287.
31. Berg, P. E. (1981) *J. Bacteriol.* **146,** 660–667.
32. Sarthy, A., Michaelis, S. & Beckwith, J. (1981) *J. Bacteriol.* **145,** 288–292.
33. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. (1981) *Nucleic Acids Res.* **9,** r43–r74.
34. Grosjean, H. & Fiers, W. (1982) *Gene* **18,** 199–209.
35. Ikemura, T. (1981) *J. Mol. Biol.* **151,** 389–409.
36. Ikemura, T. (1982) *J. Mol. Biol.* **158,** 573–597.
37. Gribskov, M., Devereux, J. & Burgess, R. R. (1984) *Nucleic Acids Res.* **12,** 539–549.
38. McLachlan, A. D., Staden, R. & Boswell, D. R. (1984) *Nucleic Acids Res.* **12,** 9567–9575.
39. Zabin, I. & Fowler, A. (1970) in *The Lactose Operon*, eds. Beckwith, J. R. & Zipser, D. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 27–47.
40. Nierlich, D. P., Kwan, C., Murakawa, G. J., Mahoney, P. A., Ung, A. W. & Caprioglio, D. (1985) in *The Molecular Biology of Bacterial Growth*, eds. Schaechter, M., Neidhardt, F. C., Ingraham, J. L. & Kjeddgaard, N. O. (Jones and Bartlett, Boston), pp. 185–193.
41. Gilson, E., Clément, J.-M., Brutlag, D. & Hofnung, M. (1984) *EMBO J.* **3,** 1417–1421.
42. Stern, M. J., Ames, G. F.-L., Smith, N. H., Robinson, E. C. & Higgins, C. F. (1984) *Cell* **37,** 1015–1025.
43. Kwan, C., Murakawa, G. J., Mahoney, P. A. & Nierlich, D. P. (1984) *Annu. Am. Soc. Microbiol.* **84,** 115.