Volume 14 Number 19 1986 Nucleic Acids Research

Nucleotide sequences of the *gal* E gene and the *gal* T gene of *E.coli*

Hans-Georg Lemaire and Benno Müller-Hill

Institut für Genetik der Universität zu Köln, Weyertal 121, D-5000 Köln 41, FRG

Received 5 August 1986; Accepted 12 September 1986

ABSTRACT

The nucleotide sequences of the gal E gene coding for UDP-galactose-4-epimerase and the gal T gene coding for galactose-1-P uridyltransferase of Escherichia coli have been determined. UDP-galactose-4-epimerase and galactose-1-P uridyltransferase are predicted to consist of 338 and 347 residues, respectively, NH_2-terminal methionines included.

INTRODUCTION

The first three reactions of galactose metabolism in E.coli are catalysed by galactokinase, galactose-1-P uridyltransferase, and UDP-galactose-4-epimerase (1). The structural genes of these enzymes, gal E, gal T and gal K, lie adjacent to one another to form the galactose operon, which is negatively controlled by a repressor (2,3). The genes of the gal operon are expressed from a polycistronic mRNA in the order E, T, K (4,5).

Here we show the nucleotide sequence of the structural genes of the gal operon. The DNA sequences of the gal K gene and 171 nucleotides preceding it have been published previously (6,7,8) and were confirmed by this work.

MATERIALS AND METHODS

Source of gal operon DNA

Starting point of this study was the plasmid pKS100 (9,10), which was constructed and kindly provided by P. Starlinger. It is a derivative of pBR322 (11) containing the wild-type galactose operon of E.coli on a 3.8 kb EcoRI / Hinc II fragment. This fragment was obtained from a partial digest with Hinc II of DNA from λpgal 8 (12), subsequently digested

© IRL Press Limited, Oxford, England.

completely with EcoRI.

DNA sequencing

The DNA sequence of the gal operon was completely determined by the dideoxy chain termination method of Sanger (13). Restriction fragments for sequencing were ligated into the appropriately linearized M13 vectors M13mp8 and M13mp9 of Messing and Vieira (14). As host was used E.coli K-12 BMH 71-18 (lac-pro) del F' pro lac I^q ZM 15 (15).
Recombinant phages were identified by the lac complementation assay of Gronenborn and Messing (16). Bacteriophage isolation and DNA extraction were carried out as described by Sanger et al. (17). The sequencing of the cloned restriction fragments was performed with [α- ^{32} P] dATP (400 Ci/mmol) using a commercially available 17-mer M13 primer. All of the synthetic primer oligomers mentioned in the legend of figure 1 were synthesized in our laboratory using a DNA synthesizer (Applied Biosystems). All molecular cloning techniques were performed according to standard procedures (18).
Three different methods were applied to accumulate sequence data: (I) the M13mp9 clone containing the 2.8 kb Hind III - Hinc II fragment of pKS100 was partially digested with Sau 3A, completely digested with Bam HI and religated. The Sau 3A concentration was adjusted to approximate one cut per molecule. In this way deletion mutants should be generated which position different regions of the gal fragment next to the priming site of the M13 vector. (II) Sequence data were accumulated by 'shotgun' cloning of the M13 clones containing the 2.8 kb Hind III - Hinc II fragment of pKS100. The enzymes used for this procedure were Sau 3A and Taq I - Hpa II, respectively. (III) Parts of the sequence where appropriate subclones were missing were deduced with the help of synthetic primer oligomers synthesized by our laboratory. The overall sequencing strategy is shown in figure 1.

Enzymes and chemicals

The enzymes and chemicals were obtained from the following sources: [α-^{32}P] dATP (400 Ci/mmol), M13mp8(9) RF-DNA and the M13 sequencing primer (17-mer) from Amersham-Buchler (Braunschweig, FRG), DNA-Polymerse I (large fragment) and the

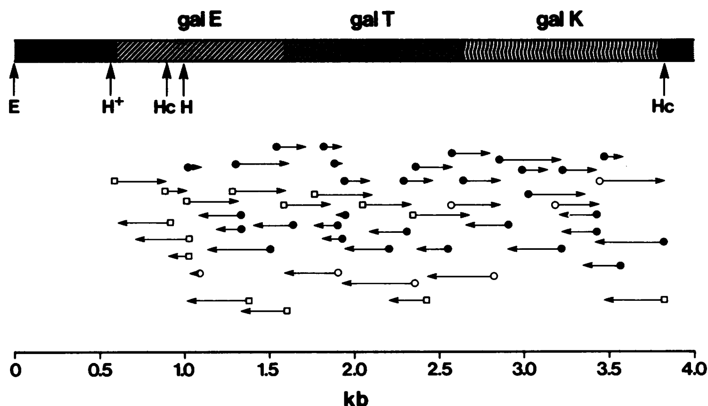

Figure 1. DNA sequencing strategy. The top line represents the 3.8 kb EcoRI - HincII fragment of the gal operon of pKS100. Position of the major restriction sites are indicated: E, EcoRI; Hc, HincII; H, H$^+$, HindIII. These restriction sites were used to generate five DNA fragments: a 0,97 kb EcoRI-HincII fragment, a 1,07 kb EcoRI-HindIII fragment, a 0,1 kb HincII-HindIII fragment, a 2.9 kb HincII-HincII fragment and a 2.8 kb HindIII-HincII fragment.
A 0.32 kb HindIII-HincII (H$^+$-Hc) fragment was obtained from the plasmid pLF001 (10). This plasmid differs from the parent pKS100 by the absence of the single HindIII site in gal E and a single point mutation (G/C to A/T) which causes the generation of a unique HindIII (H$^+$) site located within the untranslated 'leader' sequence of the gal operon. All the fragments were cloned into the M13 vectors M13mp8 and M13mp9 of Messing and Vieira (14). Below the thick line the strategy for sequencing for M13 clones is shown and indicates that all of the sequence was determined on both strands. The arrows are marked with circles, filled circles and squares, corresponding to the three different methods used to accumulate sequencing data as described in the text. o : Sau 3A (partial) / BamHI ; ● : Sau 3A and TaqI/HpaII, □ : fragments and synthetic primer oligomers.

restriction endonucleases EcoRI, HindIII, Sau3A from Boehringer (Mannheim, FRG), restriction endonucleases TaqI, HpaII, HincII, and agarose and urea from BRL (Neu-Isenburg, FRG), T4-DNA-Ligase from New England Biolabs (Bad Schwalbach, FRG), nucleotides from PL Biochemicals (Milwaukee, Wisc., USA), 'Trizma Base', dithiothreitol, EDTA, Brij 58, polyethylene glycol 6000, ethidium bromide, sodium deoxycholate from Sigma Chemie (München, FRG), substances for polyacrylamide gel electrophoresis from Serva Feinbiochemie (Heidelberg, FRG),

```
0001 ATGAGAGTTCTGGTTACCGGTGGTAGCGGTTACATTGGAAGTCATACCTGTGTGCAATTACTGCAAAACGGTCATGATGTCATCATTCTT 0090
     MetArgValLeuValThrGlyGlySerGlyTyrIleGlySerHisThrCysValGlnLeuLeuGlnAsnGlyHisAspValIleIleLeu

0091 GATAACCTCTGTAACAGTAAGCGCAGCGTACTGCCTGTTATCGAGCGTTTAGGCGGCAAACATCCAACGTTTGTTGAAGGCGATATTCGT 0180
     AspAsnLeuCysAsnSerLysArgSerValLeuProValIleGluArgLeuGlyGlyLysHisProThrPheValGluGlyAspIleArg

0181 AACGAAGCGTTGATGACCGAGATCCTGCACGATCACGCTATCGACACCGTGATCCACTTCGCCGGGCTGAAAGCCGTGGGCGAATCGGTA 0270
     AsnGluAlaLeuMetThrGluIleLeuHisAspHisAlaIleAspThrValIleHisPheAlaGlyLeuLysAlaValGlyGluSerVal

                                   HincII
0271 CAAAAACCGCTGGAATATTACGACAACAATGTCAACGGCACTCTGCGCCTGATTAGCGCCATGCGCGCCGCTAACGTCAAAAACTTTATT 0360
     GlnLysProLeuGluTyrTyrAspAsnAsnValAsnGlyThrLeuArgLeuIleSerAlaMetArgAlaAlaAsnValLysAsnPheIle

                                                       HindIII
0361 TTTAGCTCCTCCGCCACCGTTTATGGCGATCAGCCCAAAATTCCATACGTTGAAAGCTTCCCGACCGGCACACCGCAAAGCCCTTACGGC 0450
     PheSerSerSerAlaThrValTyrGlyAspGlnProLysIleProTyrValGluSerPheProThrGlyThrProGlnSerProTyrGly

0451 AAAAGCAAGCTGATGGTGGAACAGATCCTCACCGATCTGCAAAAAGCCCAGCCGGACTGGAGCATTGCCCTGCTGCGCTACTTCAACCCG 0540
     LysSerLysLeuMetValGluGlnIleLeuThrAspLeuGlnLysAlaGlnProAspTrpSerIleAlaLeuLeuArgTyrPheAsnPro

0541 GTTGGCGCGCATCCGTCGGGCGATATGGGCGAAGATCCGCAAGGCATTCCGAATAACCTGATGCCATACATCGCCCAGGTTGCTGTAGGC 0630
     ValGlyAlaHisProSerGlyAspMetGlyGluAspProGlnGlyIleProAsnAsnLeuMetProTyrIleAlaGlnValAlaValGly

0631 CGTCGCGACTCGCTGGCGATTTTTGGTAACGATTATCCGACCGAAGATGGTACTGGCGTACGCGATTACATCCACGTAATGGATCTGGCG 0720
     ArgArgAspSerLeuAlaIlePheGlyAsnAspTyrProThrGluAspGlyThrGlyValArgAspTyrIleHisValMetAspLeuAla

0721 GACGGTCACGTCGTGGCGATGGAAAAACTGGCGAACAAGCCAGGCGTACACATCTACAACCTCGGCGCTGGCGTAGGCAACAGCGTGCTG 0810
     AspGlyHisValValAlaMetGluLysLeuAlaAsnLysProGlyValHisIleTyrAsnLeuGlyAlaGlyValGlyAsnSerValLeu

0811 GACGTGGTTAATGCCTTCAGCAAAGCCTGCGGCAAACCGGTTAATTATCATTTTGCACCGCGTCGCGAGGGCGCCTTCCGGCCTACTGGG 0900
     AspValValAsnAlaPheSerLysAlaCysGlyLysProValAsnTyrHisPheAlaProArgArgGluGlyAlaPheArgProThrGly

0901 CGGACGCCAGCAAAGCCGACCGTGAACTTGAACTGGCGCGTAACGCGCACACTCGATGAAATGGCGCAGGACACCTGGCACTGGCAGTCA 0990
     ArgThrProAlaLysProThrValAsnLeuAsnTrpArgValThrArgThrLeuAspGluMetAlaGlnAspThrTrpHisTrpGlnSer

0991 CGCCATCCACAGGGATATCCCGATTAAGGAACGACCATGACGCAATTTAATCCCGTTGATCATCCACATCGCCGCTACAACCCGCTCACC 1080
     ArgHisProGlnGlyTyrProAsp *         MetThrGlnPheAsnProValAspHisProHisArgArgTyrAsnProLeuThr

1081 GGGCAATGGATTCTGGTTTCACCGCACCGCCTAAGCCCCTGGCAGGGGGCGCAGGAAACGCCAGCCAAACAGGTGTTACCTGCGCACGAT 1170
     GlyGlnTrpIleLeuValSerProHisArgLeuSerProTrpGlnGlyAlaGlnGluThrProAlaLysGlnValLeuProAlaHisAsp

1171 CCAGATTGCTTCCTCTGCGCAGGTAATGTGCGGGTGACAGGCGATAAAAACCCCGATTACACCGGGACTTACGTTTTCACTAATGACTTT 1260
     ProAspCysPheLeuCysAlaGlyAsnValArgValThrGlyAspLysAsnProAspTyrThrGlyThrTyrValPheThrAsnAspPhe

1261 GCGGCTTTGATGTCTGACACGCCAGATGCGCCAGAAAGTCACGATCCGCTGATGCGTTGCCAGAGCGCGCGCGGCACCAGCCGGGTGATC 1350
     AlaAlaLeuMetSerAspThrProAspAlaProGluSerHisAspProLeuMetArgCysGlnSerAlaArgGlyThrSerArgValIle

1351 TGCTTTTCACCGGATCACAGTAAAACGCTGCCAGAGCTCAGCGTTGCAGCATTGACGGAAATCGTCAAAACCTGGCAGGAGCAAACCGCA 1440
     CysPheSerProAspHisSerLysThrLeuProGluLeuSerValAlaAlaLeuThrGluIleValLysThrTrpGlnGluGlnThrAla

1441 GAACTGGGGAAAACGTACCCATGGGTGCAGGTTTTTGAAAACAAAGGCGCGGCGATGGGCTGCTCTAACCCGCATCCGCACGGTCAGATT 1530
     GluLeuGlyLysThrTyrProTrpValGlnValPheGluAsnLysGlyAlaAlaMetGlyCysSerAsnProHisProHisGlyGlnIle

1531 TGGGCAAATAGCTTCCTGCCTAACGAAGCTGAGCGCGAAGACCGCCTGCAAAAAGAATATTTTGCCGAACAGAAATCACCAATGCTGGTG 1620
     TrpAlaAsnSerPheLeuProAsnGluAlaGluArgGluAspArgLeuGlnLysGluTyrPheAlaGluGlnLysSerProMetLeuVal

1621 GATTATGTTCAGCGCGAGCTGGCAGACGGTAGCCGTACCGTTGTCGAAACCGAACACTGGTTAGCCGTCGTGCCTTACTGGGCTGCCTGG 1710
     AspTyrValGlnArgGluLeuAlaAspGlySerArgThrValValGluThrGluHisTrpLeuAlaValValProTyrTrpAlaAlaTrp

1711 CCGTTCGAAACGCTACTGCTGCCCAAAGCCCACGTTTTACGGATCACCGATTTGACCGACGCCCAGCGCAGCGATCTGGCGCTGGCGTTG 1800
     ProPheGluThrLeuLeuLeuProLysAlaHisValLeuArgIleThrAspLeuThrAspAlaGlnArgSerAspLeuAlaLeuAlaLeu

1801 AAAAAGCTGACCAGTCGTTATGACAACCTCTTCCAGTGCTCCTTCCCCTACTCTATGGGCTGGCACGGCGCGCCATTTAATGGCGAAGAG 1890
     LysLysLeuThrSerArgTyrAspAsnLeuPheGlnCysSerPheProTyrSerMetGlyTrpHisGlyAlaProPheAsnGlyGluGlu

1891 AATCAACACTGGCAGCTGCACGCGCACTTTTATCCGCCTCTGCTGCGCTCCGCCACCGTACGTAAATTTATGGTTGGTTATGAAATGCTG 1980
     AsnGlnHisTrpGlnLeuHisAlaHisPheTyrProProLeuLeuArgSerAlaThrValArgLysPheMetValGlyTyrGluMetLeu

1981 GCAGAGACCCAGCGAGACCTGACCGCAGAACAGGCAGCAGAGCGTTTGCGCGCAGTCAGCGACATCCATTTTCGCGAATCCGGAGTGTAA 2070
     AlaGluThrGlnArgAspLeuThrAlaGluGlnAlaAlaGluArgLeuArgAlaValSerAspIleHisPheArgGluSerGlyVal *
```

Figure 2. Complete nucleotide sequences of the *gal* E and the *gal* T genes and their deduced protein sequences. Bases 1027 and 2070 correspond to bases 1 and 1044 of *gal* T, respectively. The initiation codons of both genes are underlined, the termination codons are indicated by an asterisk. The ribosome binding site (SD) of *gal* T is boxed. The restriction sites for *Hinc*II (301) and *Hind*III (414) are marked.

isopropyl-ß-D-thiogalactopyranoside and 5-bromo-4-chloro-3-indolyl-ß-D-galactoside from Bachem Fine Chemicals (Torrance, Calif., USA). All other chemicals used were of analytical reagent grade.

Table 1. Amino acid composition of UDP-galactose-4-epimerase and galactose-1-P-uridyltransferase

Amino acid	Residues per polypeptide derived from		Predicted from nucleotide sequence	
	Epimerase	Transferase	Epimerase	Transferase
Asx	46	39	41	32
Thr	16	21	18	23
Ser	20	14	18	20
Glx	32	42	26	42
Pro	23	24	23	26
Gly	33	20	31	16
Ala	29	38	24	32
Val	31	24	30	22
Met	9	7	9	8
Ile	19	8	18	6
Leu	30	31	26	31
Tyr	14	10	13	11
Phe	10	15	10	15
Lys	15	14	14	12
His	10	13	13	15
Arg	13	17	17	20
Trp	6	11	4	10
Cys	4		3	6
Total	360	348 + Cys-residues	338	347

RESULTS AND DISCUSSION

Figure 2 shows the DNA sequence of the *gal* E and *gal* T genes and the protein sequences deduced from the open reading frames. The proteins encoded by the 1017 bp *gal E* and the 1044 bp *gal* T gene consist of 338 and 347 amino acids, respectively. The data are in good agreement with the published size and amino acid composition of the UDP-galactose-4-epimerase and the galactose-1-P uridyltransferase, shown in Table 1 (19).

The amino terminal sequence of the *gal* T gene product corresponds to the extreme 5' DNA sequence of the *gal* T gene

(20). Bases 877 to 1044 of *gal* T have been determined previously (6,7). Bases 964 to 1017 of *gal* *E*, 1 to 56 and 131 to 180 of *gal* T, as well as the region between *gal* E and *gal* T have all been determined previously (20). All this data could be confirmed by this study except for bp 131 of *gal* T, which is an 'A' in the study of Grindley (20) and a 'T' in our study. An 'A' at this site would have resulted in an ochre codon.

There are 9 bases between the *gal* E termination codon, TAA, and the *gal* T initiation codon, ATG. The ribosome binding site is part of this intervening sequence and involves the third base of the *gal* E termination codon.

We looked for sequence homology between the three *gal* enzymes using computer programms of the University of Wisconsin Genetics Computer Group, and found no significant homologies. It remains to be seen whether X-ray analysis will show similar tertiary structures indicating after all a common origin as proposed by Horowitz (21).

ACKNOWLEDGEMENTS

We would like to thank U. Stadelmann for excellent technical assistance, and Ramzija Suljic, Claudia Quirini, and Claudia Henkel for typing and editing the manuscript. We thank Fonds der Chemie for Support.

REFERENCES

1. Kalckar, H.M. (1958), Advan. Enzymol. 20, 111
2. Buttin, G. (1963), J. Mol. Biol. 7, 183-205
3. v.Wilcken-Bergmann, B. and Müller-Hill, B. (1982), PNAS 79, 2427-2431
4. Michaelis, G. and Starlinger, P. (1967) Mol.Gen.Genet. 100, 210-215
5. Adhya, S. and Shapiro, J.A. (1968), Genetic 62, 321-247
6. McKenney, K., Shimatake, H., Court, D., Schmeissner, U., Brady, C. and Rosenberg, M. (1981), in Gene Amplification and Analysis, Chirikjian, J.S. and Papas, T.S., Eds. Vol II, pp. 383-415, Elsevier/North-Holland, New York
7. Schümperli, D., Mckenney, K., Sobieski, D.A. and Rosenberg, M. (1982), Cell 30, 865-871.
8. Debouck, C., Riccio, A., Schümperli, D. McKenney, K., Jeffers, J., Hughes, C. and Rosenberg, M. (1985), Nucl. Acids Research, 13, 1841-1853
9. Trinks, K., Habermann, P., Beyreuther, K., Starlinger, P. and Ehring, R. (1981), Mol.Gen.Genet. 182, 183-188
10. Fritz, H.J., Bicknäse, H., Gleumes, B., Heibach, C., Rosahl, S. and Ehring, R. (1983), EMBO Journal, Vol. 2, No. 12, pp. 2129-2135

11. Bolivar, F., Rodriguez, R.L., Greene, P. J., Betlach, M.C., Heyneker, H.L. and Boyer, H.W. (1977) Gene, 2, 95-113
12. Echols, H. (1970), J. Mol. Biol., 47, 575-583
13. Sanger, F., Nicklen, S., Coulson, A.R. (1977), PNAS 74, 5463-5467
14. Messing, J., Vieira, J. (1982), Gene 19, 269-276
15. Messing, J., Gronenborn, B., Müller-Hill, B. and Hofschneider, P.H. (1977), PNAS 74, 3642-3646
16. Gronenborn, B., Messing, J. (1978), Nature 272, 375-377.
17. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. (1980), J. Mol. Biol., 143, 161-178.
18. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982), J.Eds, Molecular Cloning, Cold Spring Harbor Laboratories, ColdSpring Harbor, New York.
19. Wilson, D.B. and Hogness, D.S. (1969), J. Biol. Chemistry, 244, 2137-2142.
20. Grindley, N.D.F., (1978), Cell 13, 419-426.
21. Horowitz, N.H., (1965), in: Evolving genes and proteins, ed Bryson, V., Vogel, H., pp 15-23, Academic, New York.