Volume 13 Number 6 1985 Nucleic Acids Research

Structure of the galactokinase gene of *Escherichia coli*, the last (?) gene of the *gal* operon

C. Debouck*+, A.Riccio[1]*, D.Schumperli[2]*, K.McKenney[3]*, J.Jeffers[4]+, C.Hughes[5]+ and M.Rosenberg*+
*Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205, +Smith Kline and French Laboratories, Philadelphia, PA 19101, USA, and

M.Heusterspreute, F.Brunel and J.Davison
International Institute of Cellular and Molecular Pathology, 75 Avenue Hippocrate, 1200 Brussels, Belgium

Received 30 January 1984; Accepted 28 February 1985

ABSTRACT

We present the nucleotide sequence of the galactokinase gene (galK) of Escherichia coli including its 5' and 3' flanking regions. This DNA sequence derives from the λgal8 transducing phage and is identical to the sequence present in the galK gene fusion vectors, pKO and pKG, commonly used to study transcriptional regulatory elements. We define the precise 3' junction between the bacterial and phage sequences in λgal8 and demonstrate that this junction probably results from a homologous recombination event between identical 9 bp sequences common to the gal operon and phage λ. Moreover, we examine the 300 bp region located immediately beyond galK for transcription termination function and find no gal operon terminator. Lastly, we compare the galK genes of E. coli and the yeast S. cerevisiae and find several regions of strong homology among which is a potential ATP-binding site homology shared by a variety of ATP-binding proteins including protein kinases encoded by mammalian oncogenes.

INTRODUCTION

The gal operon of E. coli (Fig. 1) is known to consist of three structurally contiguous genes which specify the enzymes required for the metabolism of galactose: galE (uridine diphosphogalactose-4-epimerase), galT (galactose-1-phosphate uridyltransferase) and galK (galactokinase). These genes are expressed from a polycistronic mRNA in the order E, T, K (1,2). The expression of the promoter distal gene of the operon, galK, is known to be coupled translationally to the galT gene immediately preceding it (3). This translational coupling results from a structural overlap between the end of the galT coding sequence and the ribosome binding region of galK (see Fig. 2). The translational coupling of galT and galK ensures the coordinate expression of these genes during the metabolism of galactose.

The product of the galK gene, galactokinase, catalizes the first reaction of galactose catabolism: galactose + ATP → galactose-1-phosphate + ADP. This reaction is readily monitored by a simple and sensitive assay that utilizes ^{14}C-galactose as substrate. In addition, by appropriate manipulation of the host cell genetic background it is possible to apply either positive or

negative genetic selection to the galK function (that is, galK expression can be made either essential or lethal to cells under the appropriate selective conditions). This versatility has prompted the use of galK as a selective marker in gene fusion vectors for the analysis of transcriptional regulatory signals in E. coli (4,5). This galK fusion vector system has been used for the isolation, characterization and mutational analysis of both promoter (pKO vectors) and terminator (pKG vectors) sequences of E. coli (see ref. 6 for a review). The use of galK as a selective marker has also been adapted to develop gene fusion vectors to study transcriptional regulatory elements in the gram-positive bacteria Streptomyces (7), yeast (8-11) and higher cell systems (5, 12-15).

In this report, we present the DNA sequence of the E. coli galK gene, including its intercistronic boundary with galT and the 300 base pair (bp) region located downstream of the gene. We define precisely the 3'boundary of the gal operon sequence as it occurs in the λgal8 transducing phage. We also examine transcription in the 300 bp region beyond galK and find no evidence for an operon terminator. Lastly, we compare the galK gene of E. coli and the yeast S. cerevisiae and find that among their homologies is included the potential ATP-binding site of the protein.

MATERIALS AND METHODS

Source of gal operon DNA

The E. coli gal operon was isolated from the λgal8 transducing phage (16) on a 5 kilobase (kb) EcoRI-SmaI restriction fragment and inserted between the EcoRI and PvuII sites of pBR322. The resulting plasmid, pKGalS, was used as a DNA source for all subsequent subclonings. The 1363 bp PvuII-HpaI DNA fragment (Fig. 1, 2; coordinates -168 to 1195), which carries the entire galK coding sequence as well as 168 bp upstream and 46 bp downstream of galK, was used to construct the pKO and pKG galK fusion vectors (4, 17). The 303 bp MboI fragment (Fig. 2; coordinates 1029 to 1331) and the 247 bp HpaI fragment (1196 to 1442) were used to construct the pDS30 and pDS50 plasmids described in this report. The entire DNA sequence presented in Figure 2 derives from fragments isolated directly or subcloned from pKGalS, and therefore represents the nucleotide sequence of the galK gene as it occurs in the λgal8 transducing phage. The nucleotide sequence of the gal operon region extending beyond the gal-λ junction in gal8 was derived from a subclone of the gal operon isolated directly from the chromosome of E. coli SA500 (18) (see text and Figure 3).

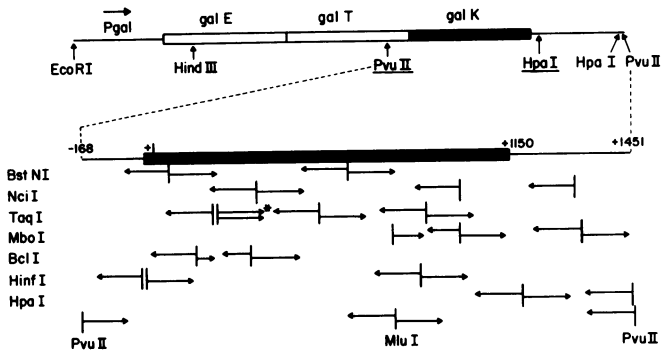


Figure 1. Schematic of the E. coli gal operon and sequencing strategy for the galK gene. The position of the major restriction sites of the gal operon is indicated. The DNA region located between the underlined PvuII and HpaI sites represents the extent of the galK insert present in the pKO and pKG galK fusion vectors (4, 17). The sequencing strategy for the galK gene is indicated on the expanded region of the map. The numbering is as in Figure 2. The position of the BstNI, NciI, BclI, HpaI, PvuII and MluI sites is shown in Figure 2. The other restriction sites have the following coordinates in Figure 2: TaqI (224, 239, 566, 914), MboI (777, 1029, 1332) and HinfI (-18, 3, 886). All DNA fragments but one (*) were labeled at their 5'end. The direction and extend of sequencing from each site is shown by the arrows.

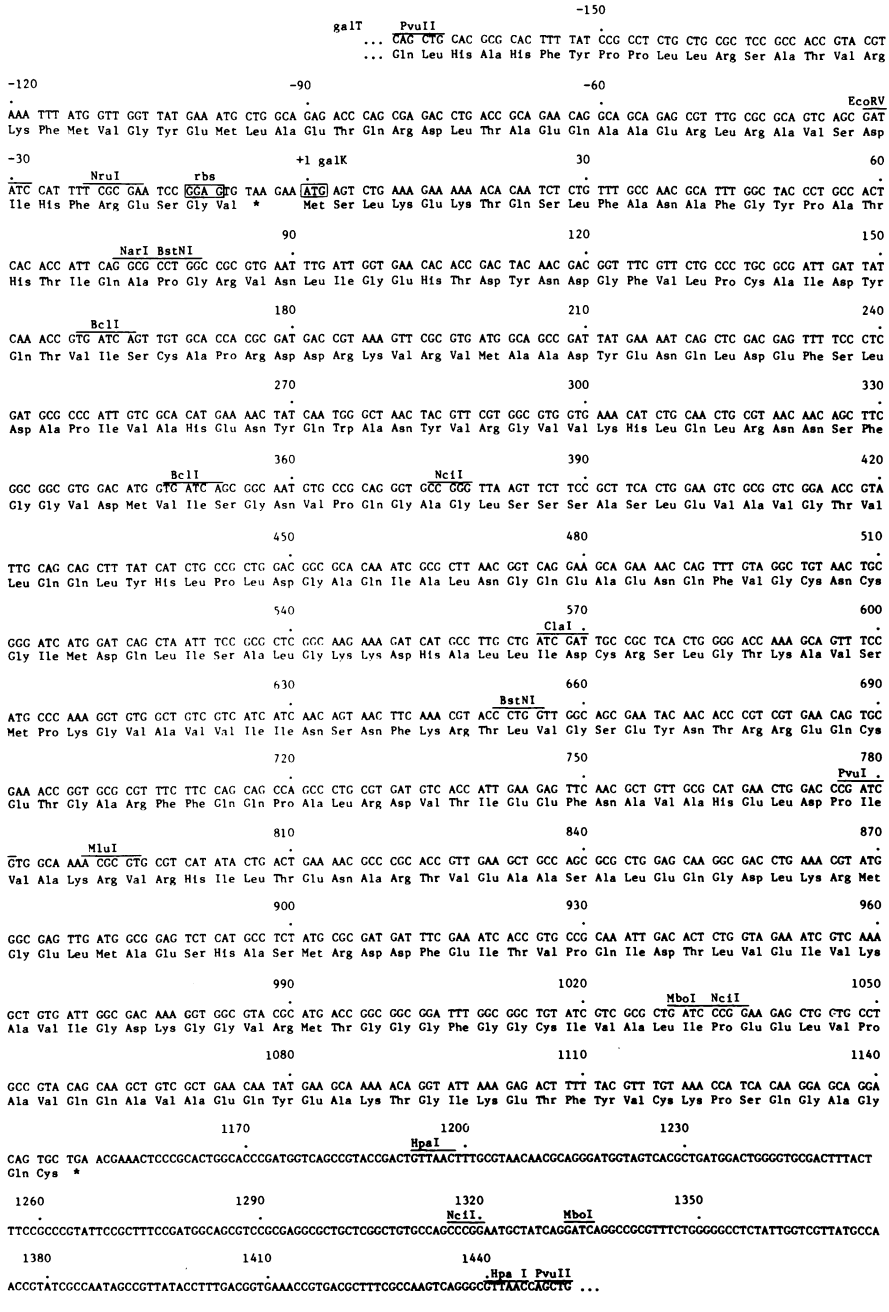
All molecular cloning techniques were performed according to standard procedures (19).

DNA sequencing

DNA sequencing was performed independently in the laboratories of M.R. and J.D. according to the chemical cleavage method of Maxam and Gilbert (20). On occasion, DNA restriction fragments used for sequencing were extracted from agarose gels by the procedure of Dretzen et al. (21). The overall sequencing strategy is outlined in Figure 1 and indicates that essentially all of the sequence was determined on both DNA strands. Nucleotide sequence data were analyzed using the computer programs of Queen and Korn (22), IntelliGenetics, Inc. or Larson and Messing (23).

In vitro transcriptions

Transcriptions were carried out in vitro using purified components as described previously (24). The plasmid DNA templates were first digested with the appropriate restriction enzyme (see text) in order to generate run-off transcripts of predicted size. All reaction mixtures (50 µl) contained the appropriately restricted DNA template (0.5 µg), purified E. coli RNA polymerase (2.5 µg, Enzo Biochem), the four nucleotide triphosphates (each at 150 µM) and α-^{32}P-ATP or α-^{32}P-UTP (Amersham, final specific activity of


galT
PvuII
... CAG CTG CAC GCG CAC TTT TAT CCG CCT CTG CTG CGC TCC GCC ACC GTA CGT
... Gln Leu His Ala His Phe Tyr Pro Pro Leu Leu Arg Ser Ala Thr Val Arg
-150
-120
-90
-60
EcoRV
AAA TTT ATG GTT GGT TAT GAA ATG CTG GCA GAG ACC CAG CGA GAC CTG ACC GCA GAA CAG GCA GCA GAG CGT TTG CGC GCA GTC AGC GAT
Lys Phe Met Val Gly Tyr Glu Met Leu Ala Glu Thr Gln Arg Asp Leu Thr Ala Glu Gln Ala Ala Glu Arg Leu Arg Ala Val Ser Asp
-30
+1 galK
30
60
NruI
rbs
ATC CAT TTT CGC GAA TCC GGA GTG TAA GAA ATG AGT CTG AAA GAA AAA ACA CAA TCT CTG TTT GCC AAC GCA TTT GGC TAC CCT GCC ACT
Ile His Phe Arg Glu Ser Gly Val * Met Ser Leu Lys Glu Lys Thr Gln Ser Leu Phe Ala Asn Ala Phe Gly Tyr Pro Ala Thr
90
120
150
NarI BstNI
CAC ACC ATT CAG GCG CCT GGC CGC GTG AAT TTG ATT GGT GAA CAC ACC GAC TAC AAC GAC GGT TTC GTT CTG CCC TGC GCG ATT GAT TAT
His Thr Ile Gln Ala Pro Gly Arg Val Asn Leu Ile Gly Glu His Thr Asp Tyr Asn Asp Gly Phe Val Leu Pro Cys Ala Ile Asp Tyr
180
210
240
BclI
CAA ACC GTG ATC AGT TGT GCA CCA CGC GAT GAC CGT AAA GTT CGC GTG ATG GCA GCC GAT TAT GAA AAT CAG CTC GAC GAG TTT TCC CTC
Gln Thr Val Ile Ser Cys Ala Pro Arg Asp Asp Arg Lys Val Arg Val Met Ala Ala Asp Tyr Glu Asn Gln Leu Asp Glu Phe Ser Leu
270
300
330
GAT GCG CCC ATT GTC GCA CAT GAA AAC TAT CAA TGG GCT AAC TAC GTT CGT GGC GTG GTG AAA CAT CTG CAA CTG CGT AAC AAC AGC TTC
Asp Ala Pro Ile Val Ala His Glu Asn Tyr Gln Trp Ala Asn Tyr Val Arg Gly Val Val Lys His Leu Gln Leu Arg Asn Asn Ser Phe
360
390
420
BclI
NciI
GGC GGC GTG GAC ATG GTG ATC AGC GGC AAT GTG CCG CAG GGT GCC GCG TTA AGT TCT TCC GCT TCA CTG GAA GTC GCG GTC GGA ACC GTA
Gly Gly Val Asp Met Val Ile Ser Gly Asn Val Pro Gln Gly Ala Gly Leu Ser Ser Ser Ala Ser Leu Glu Val Ala Val Gly Thr Val
450
480
510
TTG CAG CAG CTT TAT CAT CTG CCG CTG GAC GGC GCA CAA ATC GCG CTT AAC GGT CAG GAA GCA GAA AAC CAG TTT GTA GGC TGT AAC TGC
Leu Gln Gln Leu Tyr His Leu Pro Leu Asp Gly Ala Gln Ile Ala Leu Asn Gly Gln Glu Ala Glu Asn Gln Phe Val Gly Cys Asn Cys
540
570
600
ClaI
GGG ATC ATG GAT CAG CTA ATT TCC GCG CTC GGC AAG AAA GAT CAT GCC TTG CTG ATC GAT TGC CGC TCA CTG GGG ACC AAA GCA GTT TCC
Gly Ile Met Asp Gln Leu Ile Ser Ala Leu Gly Lys Lys Asp His Ala Leu Leu Ile Asp Cys Arg Ser Leu Gly Thr Lys Ala Val Ser
630
660
690
BstNI
ATG CCC AAA GGT GTG GCT GTC GTC ATC ATC AAC AGT AAC TTC AAA CGT ACC CTG GTT GGC AGC GAA TAC AAC ACC CGT CGT GAA CAG TGC
Met Pro Lys Gly Val Ala Val Val Ile Ile Asn Ser Asn Phe Lys Arg Thr Leu Val Gly Ser Glu Tyr Asn Thr Arg Arg Glu Gln Cys
720
750
780
PvuI
GAA ACC GGT GCG CGT TTC TTC CAG CAG CCA GCC CTG CGT GAT GTC ACC ATT GAA GAG TTC AAC GCT GTT GCG CAT GAA CTG GAC CGG ATC
Glu Thr Gly Ala Arg Phe Phe Gln Gln Pro Ala Leu Arg Asp Val Thr Ile Glu Glu Phe Asn Ala Val Ala His Glu Leu Asp Pro Ile
810
840
870
MluI
GTG GCA AAA CGC GTG CGT CAT ATA CTG ACT GAA AAC GCC CGC ACC GTT GAA GCT GCC AGC GCG CTG GAG CAA GGC GAC CTG AAA CGT ATG
Val Ala Lys Arg Val Arg His Ile Leu Thr Glu Asn Ala Arg Thr Val Glu Ala Ala Ser Ala Leu Glu Gln Gly Asp Leu Lys Arg Met
900
930
960
GGC GAG TTG ATG GCG GAG TCT CAT GCC TCT ATG CGC GAT GAT TTC GAA ATC ACC GTG CCG CAA ATT GAC ACT CTG GTA GAA ATC GTC AAA
Gly Glu Leu Met Ala Glu Ser His Ala Ser Met Arg Asp Asp Phe Glu Ile Thr Val Pro Gln Ile Asp Thr Leu Val Glu Ile Val Lys
990
1020
1050
MboI NciI
GCT GTG ATT GGC GAC AAA GGT GGC GTA CGC ATG ACC GGC GGC GGA TTT GGC GGC TGT ATC GTC GCG CTG ATC CCG GAA GAG CTG CTG CCT
Ala Val Ile Gly Asp Lys Gly Gly Val Arg Met Thr Gly Gly Gly Phe Gly Gly Cys Ile Val Ala Leu Ile Pro Glu Glu Leu Val Pro
1080
1110
1140
GCC GTA CAG CAA GCT GTC GCT GAA CAA TAT GAA GCA AAA ACA GGT ATT AAA GAG ACT TTT TAC GTT TGT AAA CCA TCA CAA GGA GCA GGA
Ala Val Gln Gln Ala Val Ala Glu Gln Tyr Glu Ala Lys Thr Gly Ile Lys Glu Thr Phe Tyr Val Cys Lys Pro Ser Gln Gly Ala Gly
1170
1200
1230
HpaI
CAG TGC TGA ACGAAACTCCCGCACTGGCACCCGATGGTCAGCCGTACCGACTGTTAACTTTGCGTAACAACGCAGGGATGGTAGTCACGCTGATGGACTGGGGTGCGACTTTACT
Gln Cys *
1260
1290
1320
1350
NciI.
MboI
TTCCGCCCGTATTCCGCTTTCCGATGGCAGCGTCCGCGAGGCGCTGCTCGGCTGTGCCAGCCCGGAATGCTATCAGGATCAGGCCGCGTTTCTGGGGGCCTCTATTGGTCGTTATGCCA
1380
1410
1440
.Hpa I PvuII
ACCGTATCGCCAATAGCCGTTATACCTTTGACGGTGAAACCGTGACGCTTTCGCCAAGTCAGGGCGTTAACCAGCTG ...

about 10 Ci mmole^{-1}). Purified rho protein (0.5 µg) was added where indicated. Reactions were carried out at 37°C for 20 min, the labeled RNA products were then resolved on 4% polyacrylamide slab gels containing 8 M urea and autoradiographed as previously described (24).

Galactokinase assay

E. coli strain N100 containing the various recombinant plasmids was grown to logarithmic phase (OD_{650} = 0.6) in M56 medium with fructose as the carbon source. Galactokinase (EC 2.7.1.6) was assayed as described (4). Each value presented in Table 2 represents the average of 3 independent determinations.

RESULTS AND DISCUSSION

Structure of the E. coli galK gene

The complete nucleotide sequence of the galK gene is shown in Figure 2. The galK gene consists of 1149 bp, encoding a 382 amino acid protein of 44,000 dalton predicted molecular weight. The amino acid sequence of galK deduced from its nucleotide sequence is consistent with published analyses of the size, amino-terminal sequence and amino acid composition of the protein purified from E. coli (25,26). The sequence of the 19 amino-terminal residues of galactokinase determined by Schlesinger et al. (25) agrees perfectly with our predicted sequence. In addition, the amino acid composition of galK determined by Wilson and Hogness (26) is consistent with the composition derived from the DNA sequence, except for some discrepancy in the numbers of arginine, serine and threonine residues. Presumably, these differences represent difficulties inherent in the precise determination of amino acid composition. Both studies showed that the amino-terminal residue of the purified protein is serine, which corresponds to the second codon of the gene. This indicates that the f-Met is removed to form the mature protein in E. coli. The codon utilization of the galK gene is shown in Table 1. It reveals no strong codon biases such as those characterizing strongly or weakly

Figure 2. Complete nucleotide sequence of the E. coli galK gene and flanking regions. The sequence shown covers the carboxy-terminal end of the galT gene, the entire galK gene and about 300 bp downstream of galK up to the gal-lambda junction in the λgal8 transducing phage (see text and Figure 3). The nucleotide sequence of the coding strand of the DNA is given with the 5' to 3' reading from left to right. The nucleotide positions are numbered with +1 corresponding to the A of the ATG initiation codon of galK. The predicted amino acid sequence is shown below the DNA sequence and a dot is placed over every 10th codon of the galK gene. The ribosome binding site (rbs) and the ATG initiation codon of galK are boxed. The termination codon of both the galT and galK genes are indicated by *. Restriction sites referred to in the text are marked.

Table 1. Codon Usage in galK

	U		C		A		G		
U	PHE	6/13	SER	4/19	TYR	5/10	CYS	4/9	U
	PHE	7/13	SER	4/19	TYR	5/10	CYS	5/9	C
	LEU	1/30	SER	3/19	END	0/0	END	1/1	A
	LEU	4/30	SER	0/19	END	0/0	TRP	1/1	G
C	LEU	2/30	PRO	3/14	HIS	7/9	ARG	10/18	U
	LEU	3/30	PRO	3/14	HIS	2/9	ARG	8/18	C
	LEU	1/30	PRO	3/14	GLN	10/23	ARG	0/18	A
	LEU	19/30	PRO	5/14	GLN	13/23	ARG	0/18	G
A	ILE	9/22	THR	4/18	ASN	3/17	SER	4/19	U
	ILE	12/22	THR	12/18	ASN	14/17	SER	4/19	C
	ILE	1/22	THR	2/18	LYS	15/16	ARG	0/18	A
	MET	9/9	THR	0/18	LYS	1/16	ARG	0/18	G
G	VAL	8/36	ALA	8/39	ASP	10/20	GLY	8/33	U
	VAL	9/36	ALA	10/39	ASP	10/20	GLY	18/33	C
	VAL	5/36	ALA	10/39	GLU	19/26	GLY	4/33	A
	VAL	14/36	ALA	11/39	GLU	7/26	GLY	3/33	G

expressed genes (27), with the exception of asparagine and glycine codons which are biased toward strongly and weakly expressed genes, respectively.

The galK DNA sequence contains several unique restriction sites such as ClaI, MluI and NarI (Fig. 2). On phage and/or plasmid vectors which carry and express galK these restriction sites provide cloning sites for introducing foreign DNA segments by insertional inactivation of galK. For example, we (M.H., F.B., & J.D., unpublished) have selected for insertion of DNA fragments at the MluI site in the pKG plasmid vector system (6) by inactivation of galK expression.

Our DNA sequence also includes the 171 nucleotides immediately preceding galK, which has been published previously (3,4). This sequence comprises the 3'end of the galT gene and indicates that the galT termination codon,TAA, is separated from the galK ATG initiation codon by only three nucleotides. The sequence encoding the carboxy-terminus of galT overlaps with sequences involved in ribosome binding (rbs) and translation initiation (ATG) of the galK gene. In fact, the tetranucleotide sequence GGAG, presumably part of the galK ribosome binding site, lies entirely within the galT coding sequence. This overlap has been shown previously to be responsible for the translational coupling between the galT and galK genes (3).

The λgal8 transducing phage is generated by homologous recombination

The nucleotide sequence of the galK gene and flanking regions presented in this report derives from the λgal8 transducing phage (see Materials and

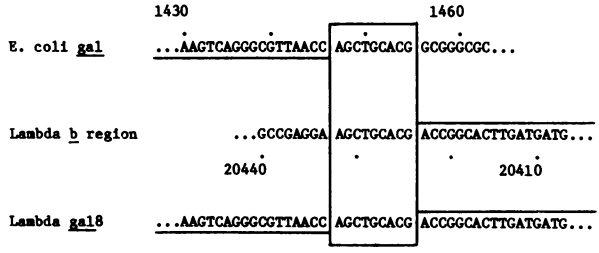


Figure 3. E. coli/lambda junction in the λgal8 tranducing phage. The DNA sequence of the E. coli/lambda junction in λgal8 is shown. Also shown are the sequences of the corresponding regions of the E. coli gal operon and the b region of bacteriophage λ(numbered according to Figure 2 and ref. 32, respectively). The part of the λgal8 sequence that is in common with the gal operon is underlined, and that part which is in common with the λb region is overlined. A 9 bp homology (boxed) is shared by all three sequences. The λgal8 transducing phage was created by an homologous recombination event that occurred between these identical 9 bp sequences in the gal and b regions.

Methods). Presumably, the distal junction between the bacterial gal operon DNA and phage DNA is formed by sequences located downstream of galK and in the b region of λ(16). This junction point was positioned by comparing our gal sequence to the known sequence of the λb region (Fig. 3). The junction occurs about 300 nucleotides beyond the end of galK within the adjacent HpaI/PvuII sites at coordinate 1440 of the sequence shown in Figure 2.

In order to gain additional insight into how this λgal8 junction was formed, we isolated and characterized this same region of the gal operon directly from the E. coli genome. A 7 kb EcoRI-BglII fragment carrying the entire gal operon (3 kb) flanked by 1 kb upstream and 3 kb downstream, was cloned out of E. coli strain SA500 (18). By restriction analysis, we were able to locate the same adjacent HpaI/PvuII sites as those found in λgal8 near the gal-λ junction. We determined the nucleotide sequence around this region and compared this sequence to that of the lambda b region and of λgal8 (Fig. 3). The data indicate that the E. coli gal operon and the b region of λ share an identical 9 bp sequence which is found precisely at the gal/lambda junction in the λgal8 transducing phage. Apparently, λgal8 arose not by an illegitimate recombinational event as originally postulated (28), but rather by an homologous recombination within this 9 bp homology. Several other cases have been reported, in which large spontaneous deletions result from recombinational events occurring between short stretches of sequence homology (6, 29- 31). These small regions of homology seem to be critical (although not exclusive, see ref. 31) elements in the generation of large deletions in E.

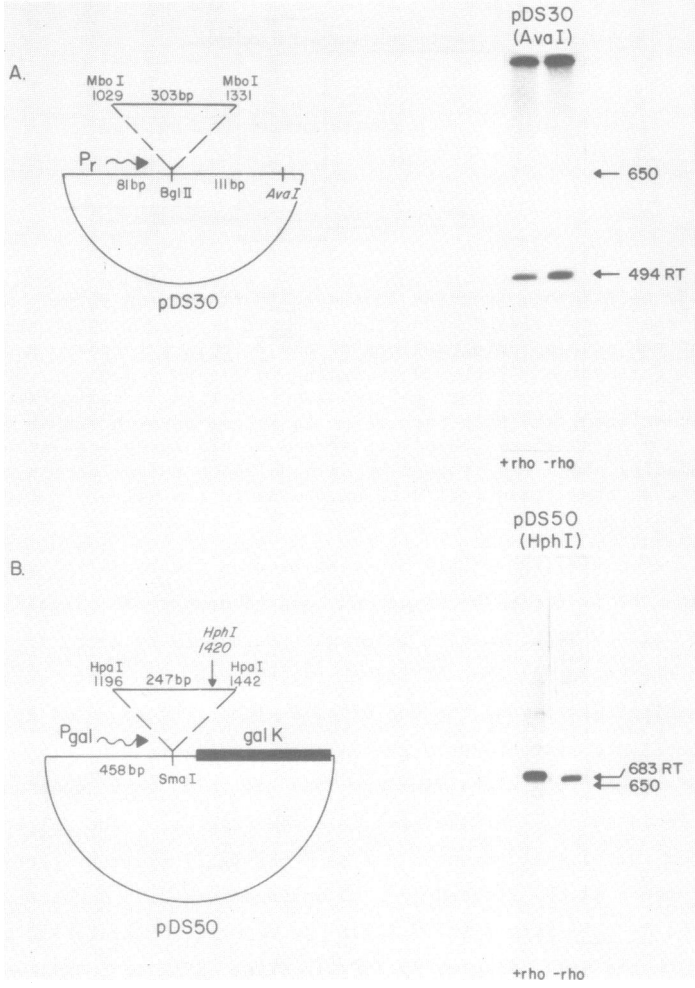


Figure 4: Analysis of transcription termination in vitro. (A) Termination function of the 303 bp MboI fragment (coordinates 1029-1331, Fig. 2). The fragment was inserted at the BglII site downstream of the P_R promoter in plasmid 1B1 to generate pDS30. (Plasmid 1B1 consists of a 1.6 kb HindIII fragment from cI857cro$^-$ r32 inserted at the HindIII site of pBR322, K. M. & M. R., unpublished). pDS30 was cleaved with AvaI and transcribed in vitro as described in Materials and Methods. The transcription reaction contained factor rho where indicated. The ^{32}P-labeled RNA products were separated on polyacrylamide gel and autoradiographed. The positions of the read-through transcript (RT) and of a transcript of known size are indicated by the arrows. (B) Termination function of the 247 bp HpaI fragment (coordinates 1196-1442, Fig 2). The fragment was inserted at the SmaI site downstream of the Pgal promoter in plasmid pKG1800 to generate pDS50. pDS50 was then cleaved with HphI and used as template for in vitro transcription reactions as in (A).

Table 2. Analysis of transcription termination in vivo

Vector	% galK Activity(a)
pKG 1800	100
pDS 50	88
pKG t_o	4

(a) pKG1800 was given the value 100.

coli, and may also be critical to the formation of tranducing phages.

Absence of a transcription terminator immediately beyond galK

GalK is the third and last known gene of the gal operon. Hence, we expected to find the transcription termination signal of the gal operon immediately downstream of the galK coding sequence. We examined the nucleotide sequence of the 300 bp region beyond the end of galK for a possible transcription terminator. No sequences were found resembling either rho-independent or rho-dependent termination signals (33), or the complex terminator at the end of the trp operon (34). However, since factor-dependent transcription terminators do not always exhibit a consistent set of common structural features, sequence comparisons alone are not sufficient to rule out the occurrence of a signal. Thus, we examined this 300 bp region for termination function both in vitro and in vivo.

We first examined two separate but overlapping segments from the 300 bp region of interest for transcription termination function in vitro, both in the presence and absence of rho factor. One of the segments, a 303 bp MboI restriction fragment, includes the carboxy-terminal end of the galK coding sequence and 183 nucleotides beyond galK (Fig. 2, coordinates 1029 to 1331). The other segment, a 247 bp HpaI fragment (coordinates 1196 to 1442), extends from 48 to 294 bp beyond galK and thus overlaps the MboI fragment by 135 nucleotides. Each fragment was cloned downstream from a well-characterized promoter signal known to function in vitro (see Fig. 4 and legend). The resulting plasmids (pDS30 and PDS50, respectively) were cleaved at a restriction site positioned beyond or near the end of the insert (AvaI and HphI, respectively). These linearized plasmids served as DNA templates for in vitro run-off transcription experiments (see Materials and Methods). The results indicate that transcription initiated at the promoter traversed the gal operon DNA inserts without terminating and gave rise to discrete, run-off RNAs. No evidence for termination was seen either in the absence or presence of rho factor. Thus, at least in vitro this 300 bp region does not appear to contain a terminator.

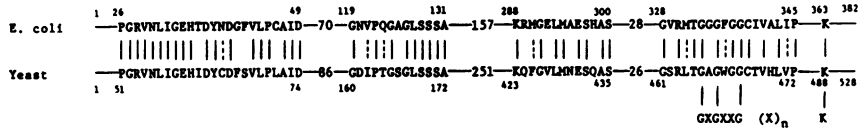


Figure 5. Comparison of E. coli and yeast galactokinase genes. The sequence (in the one-letter amino acid code) of the galactokinase proteins from E. coli (galK) and S. cerevisiae (GAL1) are aligned to reveal 4 domains of extensive homology. Perfect matches are shown by a vertical straight line, whereas a dotted line indicates amino acids of identical polarity. The position of the relevant amino acid residues is shown in smaller figures above or under the sequence. The number of amino acid residues between homologous domains is also indicated. The $GXGXXG(X)_nK$ domain (where X is an undefined amino acid) is highly conserved among protein kinases, with n=15 for v-src, 16 for v-abl, 20 for v-erbB,....(37, 38). In the case of galK (E. coli) n=24, whereas for GAL1 (yeast) n=22.

We also examined the 247 bp HpaI fragment for transcription termination function in vivo, since the pDS50 plasmid derives from pKG1800, a galK fusion vector used for the detection of termination signals. In this vector, the 247 bp distal gal operon segment is inserted between the promoter and the galK gene and its effect on galK expression can be monitored. The results, shown in Table 2, indicate that the HpaI segment has little, if any, termination function in vivo, especially when compared to a control fragment containing an authentic terminator. The HpaI fragment only reduced galK expression by about 10 percent, which corresponds to the commonly observed minimal polar effect produced in the pKG1800 vector system by fragments which do not contain terminators. In contrast, the DNA segment carrying the authentic terminator t_o of phage λ (33,35), reduced galK expression by more than 95 percent.

We conclude that there is probably no transcription terminator within the 300 bp region beyond the galK gene. Thus, the end of the gal operon remains undefined and our results suggest the possible existence of yet another function (e.g. gene) positioned beyond galK and under Pgal transcriptional regulation. An undefined open reading frame is found in the 300 bp sequence beyond galK, starting at coordinate 1218 or 1233 and extending through the end of our sequence, but no typical ribosome binding site precedes the potential initiation codons. Alternatively, the operon may simply contain an unusually long untranslated trailer segment, particularly uncommon in E. coli. We point out, however, that our conclusions are somewhat tentative at this time, since it is possible that we were unable to detect an operon terminator within the 300 bp region beyond galK. Although we examined two extensively overlapping fragments, we cannot exclude the possibility that in some way we interfered

with termination function, perhaps by placing these segments in a "foreign context" which does not allow proper function. Moreover, our in vitro transcriptions were carried out in the presence or absence of rho factor, and it is possible that some other factor(s) which was missing in our system, is required for effective in vitro termination. In vivo, we could only examine termination function of the 247 bp HpaI fragment. Perhaps sequences located upstream of this segment are required or again, improper context affected function. In any case, now that we have cloned the gal operon directly from the E. coli genome on a 7 kb fragment which extends more than 3 kb beyond galK, it should be possible to map in vivo transcripts in the region distal to galK. These studies are currently in progress. Clearly, if the 300 bp region immediately distal to galK contains a terminator, then it may be both structurally and functionally distinct from all previously characterized termination signals.

Comparison of E. coli and yeast galactokinase genes; potential ATP-binding site.

Most organisms, including E. coli and yeast, metabolize galactose using an analagous catabolic pathway in which the first reaction is catalyzed by the enzyme galactokinase. The galactokinase gene of the yeast S. cerevisiae has been cloned and its DNA sequence determined (36). We compared the structure of the E. coli galactokinase gene (galK) with that of yeast (GAL1) (Fig. 5; also see ref. 36 for another comparison). Although the gene products have different sizes (382 amino acid residues for galK versus 528 for GAL1), we found 4 regions of strong structural homology: two separate amino- terminal domains and two equally spaced carboxy-terminal domains. These regions may comprise, at least in part, the galactose and ATP-binding sites which may be conserved among these enzymes. In support of this contention, we noted that the carboxy-terminal conserved domain of both galactokinases contains the sequence gly-X-gly-X-X-gly-$(X)_n$-lys (Fig. 5), a motif which is thought to constitute part of the ATP-binding site of many proteins (37-39). This identical motif has been found in many oncogene proteins (v-src, v-abl, v-erbB,...) and a variety of other proteins (EGF receptor, mammalian cAMP-K, cell division control protein CDC28) which exhibit tyrosine kinase or more general kinase activities. The preservation of this ATP-binding hallmark among proteins as different as the galactokinases and the protein kinases, and from organisms as distantly related as bacteria, yeast, and man, indicates that this domain structure has been placed under powerful structural restraints.

ACKNOWLEDGEMENTS:

We would like to thank M.-F. Pilaete and V. Ha Thi for excellent technical assistance, S. Gamble for her expertise in computer analysis, L. Hampton and N. Smith for typing and editing the manuscript, and M. Hughes and B. Foy for help with the figures.

[1]Present address: Istituto di Patologia Generale II, Facolta Di Medicina, Naples, Italy
[2]Present address: Institut fur Molekularbiologie II, Universitat Zurich, Winterthurerstrasse 266A, CH-8057 Zurich, Switzerland
[3]Present address: Laboratory of Molecular Genetics, NINCDS, NIH, Bethesda, MD 20205, USA
[4]Present address: Ouachita Baptist University, Department of Chemistry, OBU Box 748, Arkadelphia, AR 71923, USA
[5]Present address: University of Cambridge, Department of Pathology, Cambridge, CB2 1QP, UK

REFERENCES

1. Michaelis, G. and Starlinger, P. (1967) Mol. Gen. Genet. 100, 210-215
2. Adhya, S. and Shapiro, J. A. (1968) Genetics 62, 231-247.
3. Schumperli, D., McKenney, K., Sobieski, D. A. and Rosenberg, M. (1982) Cell 30, 865-871
4. 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.
5. Rosenberg, M., McKenney, K. and Schumperli, D. (1982) in Promoters: Structure and Function, Chamberlin, M. and Rodriguez, R. L. Eds. pp. 387-406, Praeger, New York.
6. Rosenberg, M., Chepelinsky, A. B. and McKenney, M. (1983) Science 222, 734-739.
7. Brawner, M., Auerbach, J.I., Rosenberg, M. and Taylor, D. P. (1985) Gene, submitted.
8. Zitomer, R.S., Rymond, B.C., Schumperli, D. and Rosenberg, M. (1983) in Gene Expression - UCLA Symposia on Molecular and Cellular Biology, Hamer, D.H. and Rosenberg, M. Eds. Vol. 8, pp. 523-541, Liss, New York.
9. Rymond, B.C., Zitomer, R.S., Schumperli, D. and Rosenberg, M. (1983) Gene 25, 249-262.
10. Butt, T.R., Sternberg, E.J., Gorman, J.A., Clark, P., Hamer, D., Rosenberg, M. and Crooke, S.T. (1984) Proc. Natl. Acad. Sci. USA 81, 3332-3336.
11. Heusterspreute, M., Ha Thi, V. and Davison, J. (1984) DNA 3, 377-386.
12. Schumperli, D., Howard, B. H. and Rosenberg, M. (1982) Proc. Natl. Acad. Sci. USA 79, 257-261.
13. Berg, P.E., Yu, J-K., Popovic, Z., Schumperli, D., Johansen, H., Rosenberg, M. and Anderson, W. F. (1983) Mol. Cell. Biol. 3, 1246-1254.
14. Johansen, H., Reff, M., Schumperli, D. and Rosenberg, M. (1984) in Gene Expression, Alfred Benzon Symposium 19, Clark, B.F.C. and Petersen, H.V. Eds. pp 413-429, Munksgaard, Copenhagen.
15. Johansen, H., Schumperli, D and Rosenberg, M. (1984) Proc. Natl. Acad. Sci. 81, 7698-7702.
16. Feiss, M., Adhya, S. and Court, D.L. (1972) Genetics 71, 189-206.
17. McKenney, K. (1982) Ph.D. thesis, Johns Hopkins University.
18. Nakanishi, S., Adhya, S., Gottesman, M. and Pastan, I. (1974) Cell 3, 39-46.

19. Maniatis, T., Fritsch, E.F. and Sambrook, J. Eds. (1982) Molecular Cloning, Cold Spring Harbor Laboratories, Cold Spring Harbor, New York
20. Maxam, A.M. and Gilbert, W. (1980) Meth. Enzymol. 65, 489-560.
21. Dretzen, G., Bellard, M., Sassone-Corsi, P. and Chambon, P. (1981) Anal. Biochem. 112, 295-298
22. Queen, C.L. and Korn, L.J. (1980) Meth. Enzymol. 65, 595-609.
23. Larson, R. and Messing, J. (1982) Nucl. Acids Res. 10, 39-49
24. Rosenberg, M., Court, D., Shimatake, H., Brady, C. and Wulff, D.L. (1978) Nature 272, 414-423.
25. Schlesinger, D.H., Schell, M. A. and Wilson, D.B. (1977) FEBS Letters 83, 45-47.
26. Wilson, D.B. and Hogness, D.S. (1969) J. Biol. Chem. 244, 2137-2142.
27. Grosjean, H. and Fiers, W. (1982) Gene 18, 199-209.
28. Franklin, N.C. (1971) in the Bacteriophage Lambda, Hershey, A.D. Ed. Cold Spring Harbor Laboratories, Cold Spring Harbor, New York.
29. Albertini, A.M., Hofer, M., Calos, M. P. and Miller, J.H. (1982) Cell 29, 319-328.
30. Dambly-Chaudiere, C., Gottesman, M., Debouck, C. and Adhya, S. (1983) J. Molec. Appl. Genetics 2, 45-56.
31. Brunel, F., Heusterspreute, M., Merchez, M., Ha Thi, V., Pilaete, M.-F. and Davison, J. (1983) Plasmid 9, 201-214.
32. Daniels, D.L., Schroeder, J.L., Szybalski, W., Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., Petersen, G.B. and Blattner, F.R. (1983) in Lambda II, Hendrix, R. W., Roberts, J. W., Stahl, F. W. and Weisberg, R.A. Eds. Cold Spring Harbor Laboratories, Cold Spring Harbor, New York.
33. Rosenberg, M. and Court, D. (1979) Ann. Rev. Genetics 13, 319-353.
34. Wu, A.M., Christie, G.E. and Platt, T. (1981) Proc. Natl. Acad. Sci. 78, 2913-2917.
35. Dahlberg, J.E. and Blattner, F. R. (1973) Fed. Proc. 32, 664.
36. Citron, B.A. and Donelson, J.E. (1984) J. Bact. 158, 269-278.
37. Privalsky, M. L., Ralston, R. and Bishop, J. M. (1984) Proc. Natl. Acad. Sci. USA 81, 704-707.
38. Sternberg, M.J.E. and Taylor, W.R. (1984) FEBS Letters 175, 387-392.
39. Kamps, M.P., Taylor, S.S. and Sefton, B.M. (1984) Nature 310, 589-592.