Sequence of the *ebgR* Gene of *Escherichia coli:* Evidence that the EBG and LAC Operons Are Descended from a Common Ancestor[1]

Harold W. Stokes,[2] *and Barry G. Hall*

Molecular and Cell Biology Department, The University of Connecticut

The sequence of *ebgR,* the gene that encodes the EBG repressor, was determined. There is 44% DNA sequence identity between *ebgR* and *lacI,* the gene that encodes the LAC repressor. There is also 25% identity between the amino acid sequence of *lacI* and the deduced amino acid sequence of *ebgR.* The sequence of 596 bp distal to *ebgA,* the structural gene for EBG β-galactosidase, was also determined. Within that region there were two sequences, 74 and 100 bp long, that showed 46% and 50% identity, respectively, to sequences in the first 600 bp of *lacY,* the structural gene for the lactose permease. The organization and direction of transcription of the repressor and structural genes of the two operons are identical. Taken together with the homology between *ebgA* and *lacZ* (as demonstrated in the companion article in this issue), this provides strong evidence that the EBG and LAC operons are descended from a common ancestor. The map position of these two operons supports the notion that these operons diverged following a genome duplication event in an ancestor of *Escherichia coli.*

Introduction

The EBG (evolved β-galactosidase) system of *Escherichia coli* has been used extensively as a model for the study of acquisitive evolution (Hall 1982*a*). This operon consists of a repressor encoded by *ebgR,* a β-galactosidase encoded by *ebgA,* and a 79-kd protein of unknown function encoded by *ebgB* (Hall and Hartl 1975; Hall and Zuzel 1980). The wild-type repressor is insensitive to lactose as an inducer (Hall 1978), and the wild-type *ebgA* gene product is a very poor β-galactosidase (Hall 1981); thus, the wild-type operon does not permit growth on lactose. Mutations in *ebgR* and *ebgA* enhance the activity of the operon to the point where it can functionally replace the LAC operon and permit rapid growth on lactose as a sole carbon and energy source (Hall 1982*b*).

Several observations have suggested the possibility that the two operons arose via gene duplication and divergence. These observations include the virtually identical molecular weights of the polypeptides of the two β-galactosidases (Hall 1976), the similarity of the catalytic mechanisms of the two β-galactosidases (Burton and Sinnott 1983), and the overall organization of the two operons (Hall and Zuzel 1980). On the other hand, (1) *ebgA* β-galactosidase consists of six subunits, whereas *lacZ* β-galactosidase consists of four subunits; (2) there is no detectable immunological cross-

1. Key words: *ebgR* gene, EBG and LAC operons, genome duplication event.

Address for correspondence and reprints: Dr. Barry G. Hall, Molecular and Cell Biology Department, U-44, The University of Connecticut, Storrs, Connecticut 06268.

2. Current address: School of Biological Sciences, Macquarie University, North Ryde, N.S.W. 2113, Australia.

Mol. Biol. Evol. 2(6):478–483. 1985.

0737-4038/85/0206-0966$02.00

reaction between the two proteins (Campbell et al. 1973; Arraj and Campbell 1975); and (3) the two repressors fail to recognize each other's operators. To resolve this issue we have sequenced *ebgR* and also *ebgA* (Stokes et al. 1985, in this issue).

Material and Methods

Bacterial Strains and Plasmids

Bacterial strains were MG1063, SJ0R (Stokes and Hall 1984), and JM101 (Sanger et al. 1980). Plasmid pUF25 is pBR322 with the 9.6-kb EBG *Sal*I fragment from 5A1032 and encodes the allele $ebgR105^{+L}$ (Hall 1978).

Location of *ebgR*

Plasmids containing the transposon gamma-delta were constructed according to the method described by Sancar and Rupp (1979). The donor strain was MG1063 harboring pUF25 and the recipient was SJ0R. The constitutive phenotype was identified by the enhanced production of pigment on MacConkey indicator plates. Constitutive synthesis by Ebg enzyme was subsequently confirmed by whole-cell assays of O-nitrophenyl-β-galactoside hydrolysis (Hall 1980). The position of *ebgR* was then determined by restriction mapping the position of gamma-delta in plasmids producing a constitutive phenotype. Isolation and manipulation of plasmid DNA and nucleic acid sequencing was accomplished according to the method described in Stokes et al. (1985, in this issue).

Results

Locating *ebgR* within the Cloned Segment in Plasmid pUF2

The *ebgR* coding region was determined by insertional inactivation via transposition of gamma-delta into the plasmid pUF25 as described in Material and Methods. This region comprises a 0.9-kb fragment immediately adjacent to *ebgA* and corresponds to coordinates 7.5–8.4 on the pUF2 restriction map (Stokes and Hall 1984).

Determination of the *ebgR* Gene Sequence

To sequence the *ebgR* gene, we isolated the 2.6-kb *Bam-Ava* fragment of pUF2 (coordinates 5.7–8.3 on the restriction map [Stokes and Hall 1984]). Restriction fragments of this piece were subcloned into M13 sequencing vectors. Figure 1 shows a subset of those sequenced fragments that define the contiguous *ebgR* sequence. To complete the 5′ end of the *ebgR* gene, clones 538 (a *Sac*I–*Sal*I fragment) and 576 (an

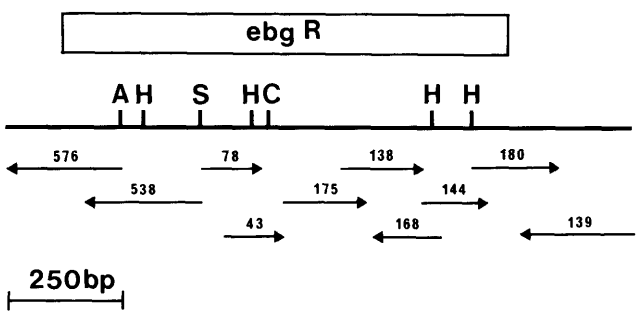

FIG. 1.—Restriction map and sequencing strategy for *ebgR*. The limits of *ebgR* are defined by the boxed region. Letters refer to restriction sites: A = *Ava*I, C = *Cla*I, H = *Hinc*II, and S = *Sac*I.

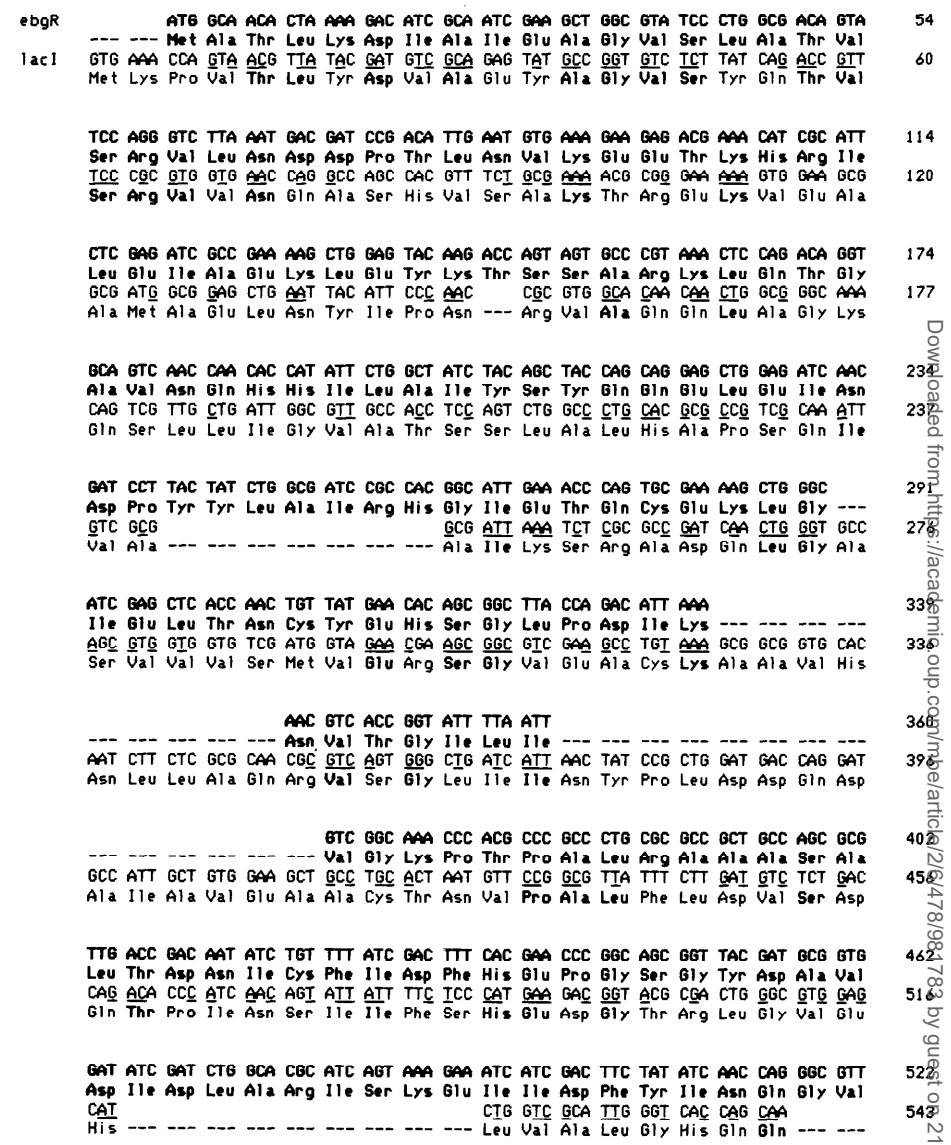

FIG. 2.—Sequence of *ebgR* and deduced amino acid sequence. The sequence of *ebgR* is shown above and aligned with the sequence of *lac*I. Dashed lines indicate gaps introduced in order to align the sequences. Bases underlined in the *lacI* sequence are identical to the aligned base in the *ebgR* sequence, and *lacI* amino acids that are boldfaced are identical to aligned *ebgR* amino acids.

*Ava*I–*Sal*I fragment) were directionally cloned into the appropriate M13 vector. A number of other clones (not shown in fig. 1) were sequenced so that the majority of the sequence was confirmed by sequencing of both strands.

Figure 2 shows the DNA sequence of *ebgR* and the deduced amino acid sequence. This sequence is the only open reading frame in this region in either orientation that is longer than 200 bp. The DNA and amino acid sequences of *lacI* (Farabough 1978) are shown below the *ebgR* sequences. The two sequences were aligned by eye with the aid of the Cornell DNA Sequencing Program (Fristensky, Lis, and Wu 1982),

```
AAT CGT ATT GGT TTT ATT GGC GGT GAA GAT GAG CCT GGC AAG GCG GAT ATT CGT GAG GTC    582
Asn Arg Ile Gly Phe Ile Gly Gly Glu Asp Glu Pro Gly Lys Ala Asp Ile Arg Glu Val
        ATC GCG CTG TTA GCG GGC CCA TTA AGT TCT GTC TCG GCG CGT CTG CGT CTG GCT    597
--- --- Leu Leu Ala Gly Pro Leu Ser Ser Val Ser Ala Arg Leu Arg Leu Ala

GCC TTT GCG GAA TAT GGC CGA CTG AAA CAA GTG GTA CGC GAA GAG GAT ATC TGG CGC GGC    642
Ala Phe Ala Glu Tyr Gly Arg Leu Lys Gln Val Val Arg Glu Glu Asp Ile Trp Arg Gly
GGC TGG CAT AAA TAT CTC ACT CGC AAT CAA ATT CAG CCG ATA GCG GAA     CGG GAA GGC    654
Gly Trp His Lys Tyr Leu Thr Arg Asn Gln Ile Gln Pro Ile Ala Glu --- Arg Glu Gly

GGT TTT TCC AGT TCG TCG GGT TAT GAA CTG GCA AAA CAA ATG CTG GCG CGG GAA GAC TAT    702
Gly Phe Ser Ser Ser Ser Gly Tyr Glu Leu Ala Lys Gln Met Leu Ala Arg Glu Asp Tyr
GAC TGG AGT GCC ATG TCC GGT TTT CAA CAA ACC ATG CAA ATG CTG AAT GAG GGC ATC GTT    714
Asp Trp Ser Ala Met Ser Gly Phe Gln Gln Thr Met Gln Met Leu Asn Glu Gly Ile Val

CCG AAG GCA CTG TTT GTT GCT TCC GAT TCC ATT GCT ATC GGC GTA CTG CGG GCA ATT CAT    762
Pro Lys Ala Leu Phe Val Ala Ser Asp Ser Ile Ala Ile Gly Val Leu Arg Ala Ile His
CCG ACT GCG ATG CTG GTT GCC AAC GAT CAG ATG GCG CTG GGC GCA ATG CGC GCC ATT ACC    774
Pro Thr Ala Met Leu Val Ala Asn Asp Gln Met Ala Leu Gly Ala Met Arg Ala Ile Thr

GAA CGA GGC CTG AAC ATC CCA CAG GAT ATT TCG CTT ATC AGC GTT AAC GAT ATC CCC ACC    822
Gly Arg Gly Leu Asn Ile Pro Gln Asp Ile Ser Leu Ile Ser Val Asn Asp Ile Pro Thr
GAG TCC GGG CTG CGC GTT GGT GCG GAT ATC TCG GTA GTG GGA TAC GAC GAT ACC GAA GAC    834
Glu Ser Gly Leu Arg Val Gly Ala Asp Ile Ser Val Val Gly Tyr Asp Asp Thr Glu Asp

GCG CGA TTT ACC TTT CCG CCG CTC TCC ACC GTG CGC ATC CAT TCC GAG ATG ATG GGA AGT    882
Ala Arg Phe Thr Phe Pro Pro Leu Ser Thr Val Arg Ile His Ser Glu Met Met Gly Ser
AGC TCA TGT TAT ATC CCG CCG TCA ACC ACC ATC AAA CAG GAT TTT CGC CTG CTG GGG CAA    894
Ser Ser Cys Tyr Ile Pro Pro Ser Thr Thr Ile Lys Gln Asp Phe Arg Leu Leu Gly Gln

CAG GGC GTT AAC CTG GTG TAT GAA         AAA GCC CGC GAT GGT CGC GCG CTG CCG CTG    936
Gln Gly Val Asn Leu Val Tyr Glu --- --- Lys Ala Arg Asp Gly Arg Ala Leu Pro Leu
ACC AGC GTG GAC CGC TTG CTG CAA CTC TCT CAG GGC CAG GCG GTG AAG GGC AAT CAG CTG    954
Thr Ser Val Asp Arg Leu Leu Gln Leu Ser Gln Gly Gln Ala Val Lys Gly Asn Gln Leu

TTA GTC TTC GTT CCC AGC AAA TTA AAA CTG CGC GGC ACG ACC CGT TAA                   984
Leu Val Phe Val Pro Ser Lys Leu Lys Leu Arg Gly Thr Thr Arg
TTG CCC GTC TCA CTG GTG AAA AGA AAA ACC ACC CTG GCG CCC AAT ACG CAA ACC GCC TCT  1014
Leu Pro Val Ser Leu Val Lys Arg Lys Thr Thr Leu Ala Pro Asn Thr Gln Thr Ala Ser

CCC CGC GCG TTG GCC GAT TCA TTA ATG CAG CTG GCA CGA CAG GTT TCC CGA CTG GAA AGC  1074
Pro Arg Ala Leu Ala Asp Ser Leu Met Gln Leu Ala Arg Gln Val Ser Arg Leu Glu Ser

GGG CAG TGA
Gly Gln
```

FIG. 2 (*Continued*)

which was used to locate regions of significant identity. For the alignment shown, excluding the "gap" regions, there is 44% DNA sequence identity over 915 bp and 25% amino acid identity over 305 amino acids. Because the sequences were aligned by eye rather than by an algorithm that seeks an "optimal" alignment, the percentage of identities is a minimal estimate.

Sequence Distal to *ebgA*

We have also determined the sequence of a 596-bp segment beginning at the first base after the stop codon of *ebgA* (see fig. 3). BP 178–251 of that segment are 46% identical with bp 244–317 of the *lacY* gene, and bp 497–596 are 50% identical with bp 491–590 of the *lacY* gene (Buchel et al. 1980).

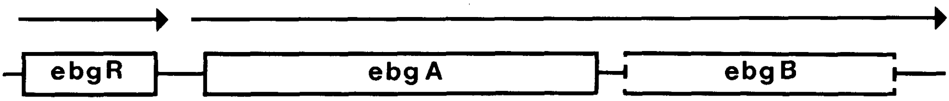

FIG. 3.—Organization of the EBG operon. The figure is to scale. The lengths of *ebgR, ebgA,* and the sequence between them were established by DNA sequencing. The precise start of *ebgB* is unknown; however, the length of the gene is inferred from the known molecular weight of the polypeptide (Hall and Zuzel 1980). In addition it is known that *ebgB* is distal to, and cotranscribed with, *ebgA* (Hall and Zuzel 1980). Arrows indicate direction of transcription as determined from DNA sequencing.

Discussion

The most notable aspect of the *ebgR* DNA sequence is its striking similarity with the sequence of *lacI* (Farabough 1978). Overall, and on a segment-by-segment comparison, the amino acid identity (25%) is considerably less than the nucleotide identity (44%). The single exception to this rule occurs in the region of bp 10–69, where the nucleotide identity of 55% is only slightly above average but the amino acid identity is 65%, about 2.5 times the average amino acid identity. This suggests that this region of the protein is under stronger selection than is the remainder. This region precisely corresponds to the region of *lacI* that is responsible for DNA binding by the lac repressor (Miller et al. 1979). DNA binding regions are strongly conserved among a large number of regulatory proteins (Kelley and Yanofsky 1985). We therefore infer that this region is the operator binding region of the EBG repressor.

The sequence similarity between *ebgR* and *lacI* leads us to conclude that the two genes are homologous, that is, descended from a common ancestral gene. Furthermore, the similarity between the 5′ end of *lacY* and the region immediately distal to *ebgA* leads to the conclusion that the entire operon is homologous to the lac operon. This conclusion is supported by the sequence similarity of *ebgA* and *lacZ* (Stokes et al. 1985, in this issue) and by the observation that the two operons are organized and transcribed in the same way (fig. 3).

We have long been intrigued by Riley's hypothesis that the genome of *Escherichia coli* evolved by at least one, and perhaps two, genome duplications (Riley and Anilionis 1978). The proposed genome duplications would generate a circularly permuted chromosome with duplicated genes lying opposite one another. The EBG operon is located at 66 min on the *E. coli* map (Bachmann 1983), whereas the LAC operon is at 8 min, which is roughly, but not exactly, opposite EBG. However, the EBG operon is transcribed in a clockwise direction on the map, whereas the LAC operon is transcribed counterclockwise, indicating than an inversion of one operon with respect to the other must have occurred. Such an inversion could well account for the displacement of the two operons from the expected 180° positions. We consider these observations as being supportive of the Riley hypothesis and suggest that the two operons may have arisen via genome, rather than simple gene, duplication.

Acknowledgments

We are grateful to Carol Crowther for assistance with sequence interpretation and for preparation of figures and to Susan Hall for assistance in preparation of the manuscript. We are particularly grateful to J. Lis for the gift of the Cornell DNA sequencing program. We are also grateful to W. Fitch for suggesting an improved "eyeball" alignment of the *ebgR* and *lacI* sequences. This work was supported by National Institutes of Health grant AI 14766. B.G.H. is the recipient of Research

Career Development Award 1 KO4 AI00366 from the National Institute of Allergy and Infectious diseases.

LITERATURE CITED

ARRAJ, J. A., and J. A. CAMPBELL. 1975. Isolation and characterization of the newly evolved ebg β-galactosidase of *Escherichia coli.* J. Bacteriol. **124**:849–856.

BACHMANN, B. J. 1983. Linkage map of *Escherichia coli* K12, edition 7. Microbiol. Rev. **47**: 180–230.

BUCHEL, D. E., B. GRONENBORN, and B. MULLER-HILL. 1980. Sequence of the lactose permease gene. Nature **283**:541–545.

BURTON, J., and M. L. SINNOTT. 1983. Catalytic consequences of experimental evolution. I. Catalysis by the wild type second β-galactosidase (ebg°) of *Escherichia coli:* a comparison with the *lacZ* enzyme. J. Chem. Soc. Perkin Trans. **2**:359–364.

CAMPBELL, J. A., J. LENGYEL, and J. LANGRIDGE. 1973. Evolution of a second gene for β-galactosidase in *Escherichia coli.* Proc. Natl. Acad. Sci. USA **70**:1841–1845.

FARABOUGH, P. G. 1978. Sequence of the *lac*I gene. Nature 274:765–769.

FRISTENSKY, B., J. LIS, and R. WU. 1982. Cornell DNA sequence analysis program. Nucleic Acids Res. **10**:6451–6463.

HALL, B. G. 1976. Experimental evolution of a new enzymatic function: kinetic analysis of the ancestral (ebg°) and evolved (ebg$^+$) enzymes. J. Mol. Biol. **107**:71–84.

———. 1978. Regulation of newly evolved enzymes. IV. Directed evolution of the ebg repressor. Genetics **90**:673–691.

———. 1980. On the evolution of new metabolic functions in diploid organisms. Genetics **96**: 1007–1017.

———. 1981. Changes in the substrate specificities of an enzyme during directed evolution of new functions. Biochemistry **20**:4042–4049.

———. 1982*a*. Evolution on a petri dish: a model system for the study of acquisitive evolution. Evol. Biol. **15**:85–150.

———. 1982*b*. Evolution of a regulated operon in the laboratory. Genetics **101**:335–344.

HALL, B. G., and D. L. HARTL. 1975. Regulation of newly evolved enzymes. II. The ebg repressor. Genetics **81**:427–435.

HALL, B. G., and T. ZUZEL. 1980. The ebg operon consists of at least two genes. J. Bacteriol. **144**:1208–1211.

KELLEY, R. L., and C. YANOFSKY. 1985. Mutational studies with the trp repressor of *Escherichia coli* support the helix-turn-helix model of repressor recognition of operator DNA. Proc. Natl. Acad. Sci. USA **82**:483–487.

MILLER, J. H., C. COULONDRE, M. HOFER, U. SCHMEISSNER, H. SOMMER, and A. SCHMITZ. 1979. Genetic studies of the Lac repressor. IX. Generation of altered proteins by the supression of nonsense mutations. J. Mol. Biol. **131**:191–222.

RILEY, M., and A. ANILIONIS. 1978. Evolution of the bacterial genome. Annu. Rev. Microbiol. **32**:519–560.

SANCAR, A., and D. RUPP. 1979. Cloning of *uvrA, lexC,* and *ssb* genes of *Escherichia coli.* Biochem. Biophys. Res. Commun. **90**:123–129.

SANGER, F., A. R. COULSON, B. G. BARRELL, A. J. H. SMITH, and B. A. ROE. 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. J. Mol. Biol. **143**:161–178.

STOKES, H. W., and B. G. HALL. 1984. Topological repression of gene activity by a transposable element. Proc. Natl. Acad. Sci. USA **81**:6115–6119.

STOKES, H. W., P. W. BETTS, and B. G. HALL. 1985. Sequence of the *ebgA* gene of *Escherichia coli:* comparison with the *lacZ* gene. Mol. Biol. Evol. **2**:469–477. (in this issue).

WALTER M. FITCH, reviewing editor

Received May 13, 1985; revision received June 27, 1985.