DNA Sequence Analysis of Artificially Evolved *ebg* Enzyme and *ebg* Repressor Genes

Barry G. Hall,*,† Paul W. Betts* and John C. Wootton‡

**Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut 06268, †Department of Biology, University of Rochester, Rochester, New York 14627, and ‡Department of Genetics, University of Leed, Leeds LS2 9JT, England*

Manuscript received April 27, 1989
Accepted for publication September 7, 1989

ABSTRACT

The *ebg* system has been used as a model to study the artificial selection of new catalytic functions of enzymes and of inducer specificities of repressors. A series of mutant enzymes with altered catalytic specificities were previously characterized biochemically as were the changes in inducer specificities of mutant, but fully functional, repressors. The wild type *ebg* operon has been sequenced, and the sequence differences of the mutant enzymes and repressors have been determined. We now report that, contrary to our previous understanding, *ebg* enzyme contains 180-kD α-subunits and 20-kD β-subunits, both of which are required for full activity. Mutations that dramatically affect substrate specificity and catalytic efficiency lie in two distinct regions, both well outside of the active site region. Mutations that affect inducer specificity of the *ebg* repressor lie within predicted sugar binding domains. Comparisons of the *ebg* β-galactosidase and repressor with homologous proteins of the *Escherichia coli* and *Klebsiella pneumoniae* lac operons, and with the galactose operon repressor, suggest that the *ebg* and *lac* operons diverged prior to the divergence of *E. coli* from Klebsiella. One case of a triple substitution as the consequence of a single event is reported, and the implications of that observation for mechanisms of spontaneous mutagenesis are discussed.

THE *ebg* (evolved β-galactosidase) system of *Escherichia coli* provides a model for studying the details of acquisitive evolution via changes in the catalytic properties of enzymes and accompanying changes in the properties of regulatory elements (HALL 1983).

The *ebg* operon is located on the opposite side of the chromosome from the *lac* operon (HALL and HARTL 1974). The wild-type *ebg* operon does not permit utilization of lactose or other β-galactoside sugars, however a series of mutations in the regulatory and structural genes of the *ebg* operon allow *ebg* to replace the *lacZ* β-galactoside for growth on lactose (HALL 1982a). The wild-type *ebg* β-galactosidase is an ineffective lactase that will not hydrolyze β-galactoside sugars effectively enough for growth even when the operon is expressed constitutively at a level such that *ebg* β-galactosidase constitutes 5% of the soluble protein of the cell. A series of spontaneous mutations in the structural gene for *ebg* β-galactosidase can increase the catalytic efficiency of that enzyme. When efficiency is expressed as V_{max}/K_m, Class I mutations increase efficiency for lactose 40-fold, but do not significantly affect the efficiency with which lactulose (galactosyl-β-1,4-fructose) is hydrolyzed (HALL 1981). Class II mutations increase the efficiency of lactose hydrolysis only tenfold, but they increase the efficiency of lactulose hydrolysis 48-fold (HALL 1981). As a consequence *ebg* constitutive class I strains can grow on lactose, but not lactulose; and constitutive class II strains grow well on both sugars.

When both class I and class II mutations are present in the same *ebg* gene, either as the result of sequential spontaneous mutations or as the consequence of a recombination between a class I and a class II strain, the gene is designated as class IV. Genetic analysis showed that there was about 1% recombination between the class I and class II sites, and it was estimated that the two sites were about 1000 bp apart within the gene (HALL and ZUZEL 1980b). Class IV *ebg* β-galactosidase is dramatically different from both wild type enzyme and from class I and class II enzymes in several respects. First, with respect to the wild-type enzyme, the efficiency of lactose hydrolysis is increased 450-fold, and the efficiency of lactulose hydrolysis is increased 140-fold (HALL 1981). Second, the efficiency with which galactosyl-β-1,4-arabinose (gal-ara) is hydrolyzed is increased 300-fold (HALL 1981), a level sufficient to permit class IV, but not wild type, class I, or class II, strains to grow on Gal-Ara (HALL 1978a). Third, class IV enzyme exhibits detectable activity toward lactobionic acid (galactosyl-β-1,4-gluconic acid), an activity that is undetectable in purified wild type, class I or class II enzymes (HALL 1981). That activity of class IV enzyme is insufficient

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

for growth (HALL 1978a), but it does create the potential for lactobionate utilization as the result of additional mutations. A third mutation in the *ebg* β-galactosidase gene increases the efficiency of lactobionate hydrolysis 18-fold, and results in a class V strain that can grow on lactobionate (HALL 1978a, 1981). Fourth, class IV *ebg* β-galactosidase (but not wild-type class I or class II enzymes) exhibits transgalactosylation activity, an activity that permits class IV *ebg* enzyme to synthesize allolactose (galactosyl-β-1,6-glucose) from lactose (galactosyl-β-1,4-glucose).

An inefficient β-galactosidase is not the only barrier that prevents products of the *ebg* operon from hydrolyzing β-galactoside sugars rapidly enough to permit growth. Synthesis of *ebg* enzyme is subject to regulation by the *ebg* repressor, the product of the *ebgR* gene (HALL and HARTL 1974, 1975). The wild-type *ebg* repressor is not very sensitive to lactose as an inducer, permitting only 100-fold induction of the operon. Even with the most efficient *ebg* enzyme, that level of expression is insufficient for growth on lactose (HALL and CLARKE 1977). Furthermore, the wild-type repressor is completely insensitive to lactulose and Gal-Ara as inducers (HALL and CLARKE 1977). A deliberate effort was made to select *ebgR* mutations which permitted lactulose to act as an effective inducer. The mutants that were obtained all resulted in repressor that was 20–40-fold more sensitive to lactulose, ten-fold more sensitive to lactose, and 50–100-fold more sensitive to Gal-Ara than was wild-type repressor. The basal level of synthesis permitted by the mutant repressors was indistinguishable from wild type, thus it appeared that the mutations affected only sugar binding, not DNA binding itself (HALL 1978b).

Sequencing of the active site peptides of *ebg* enzyme showed that the *ebg* and *lacZ* proteins are homologous (FOWLER and SMITH 1983). That conclusion was confirmed by DNA sequencing of one allele of the *ebg* operon (STOKES, BETTS and HALL 1985; STOKES and HALL 1985), and it was shown that both *ebg* enzyme and *ebg* repressor are, respectively, related to the *lacZ* encoded β-galactosidase and the *lacI* encoded repressor of the *lac* operon.

Several questions about the *ebg* system have remained unresolved: Where are the class I, class II and class V sites located in the *ebg* enzyme gene? How dispersed are the sites? Do any of these sites coincide with the active site of the enzyme? What is the nature and locations of the mutations in *ebgR* that modify the inducer specificities of the repressor?

In this communication we identify the sites of several of the mutations in questions by direct DNA sequencing. In the course of this study we correct some errors in the previously published sequences (STOKES, BETTS and HALL 1985; STOKES and HALL 1985), and we identify a previously unknown *ebg* gene that specifies a second subunit of *ebg* enzyme.

MATERIALS AND METHODS

Strains and plasmids: All strains (Table 1) are *E. coli* K12. All plasmids in Table 1 were constructed by ligating either a *Sal*I fragment (pUF2, pUF4, pUF5, pUF7, pUF16, pUF17 and pUF26) or a *Sal*I-*Hin*dIII fragment (pUF8, pUF9) of genomic DNA from the listed strain into similarly digested plasmid pBR322. All plasmids except pUF8 carry active *ebg* alleles, and were isolated by selection for growth on lactose as previously described (STOKES, BETTS and HALL 1985). Plasmid pUF8 was isolated by colony hybridization to an *ebgA* specific probe.

Media and growth conditions have been previously described (HALL and HARTL 1974, 1975; SANGER *et al.* 1980; STOKES, BETTS and HALL 1985; STOKES and HALL 1985).

PCR (polymerase chain reaction) amplification of genomic DNA: Amplications were carried out using the Taq DNA polymerase and the GeneAmp Kit produced by Perkin Elmer-Cetus, and reactions were carried out in a Thermal Cycler produced by the same company. DNA for sequencing was produced in two stages. In the first stage the reaction included 100 ng of genomic *E. coli* DNA as template, and two 20-base oligonucleotide primers with nucleotide triphosphates, Taq DNA polymerase and buffers as provided in the GeneAmp kit in a total volume of 100 μl. The first stage reaction consisted of 15 cycles in which DNA was denatured at 94° for 45 seconds, annealed at 55° for 10 sec, and polymerized at 72° for 2 min. The second stage used 2 μl of the first stage reaction as template, and only one of the two oligonucleotide primers. The reaction was carried out for 25 cycles in which DNA was denatured at 94° for 30 sec, annealed at 55° for 5 sec, and polymerized at 72° for 2 min, otherwise conditions were identical to those of the first stage. The amplified DNA was purified over Quiagen-5 tips (Quiagen, Inc.) according to the manufacturer's instructions to remove any remaining primer. Purified samples were ethanol precipitated and resuspended in 30 μl. An aliquot of 6 μl of that preparation was used in a sequencing reaction.

DNA sequencing: Cloned *ebg* operons were sequenced by subcloning fragments into either plasmid pBlu$^+$ or plasmid pBlu$^-$ (Stratagene, Inc.), and sequencing the double stranded DNA by a modification of the dideoxy method (SANGER *et al.* 1980). Other alleles were partially sequenced as indicated in Table 3 from single stranded DNA produced by PCR amplification.

Computer analysis of the *ebgR* region: The ISIS integrated data and software resource of protein sequence and structure (AKRIGG *et al.* 1988) was used for sequence similarity searches of the OWL composite protein sequence database and as a guide to the multiple alignment shown in Figure 4. Pattern discriminator matrices based upon critical residue information from crystal structures were used to predict potential DNA and sugar binding regions of the *ebgR* encoded repressor using the PATSCAN program of ISIS. Strongly positive results were obtained with matrices DNADJOW (helix-turn-helix DNA binding motif), SUGAR1SR, SUGAR2SR, and SUGAR3 (three sugar binding regions of periplasmic chemoreceptor proteins) from the Feature Library of ISIS. These pattern matches were used to guide small manual adjustments to automated sequence alignments of *ebgR* repressor with homologous repressors and sugar binding proteins.

TABLE 1

Strains and plasmids

Strain	Relevant genotype	Reference	Corresponding plasmid
DS4680A	Wild type	HALL and HARTL (1974)	pUF 8
A2	*ebgR2⁻ ebgA2*	HALL and HARTL (1974)	pUF 7
A4	*ebgR2⁻ ebgA4*	HALL and HARTL (1974)	pUF 9
A272	*ebgR2⁻ ebgA198*	HALL (1978a)	pUF 17
A23	*ebgR2⁻ ebgA134*	HALL (1978a)	None
A27	*ebgR2⁻ ebgA138*	HALL (1978a)	None
5A1	*ebgR⁺ ebgA51*	HALL and CLARKE (1977)	None
RT512	*ebgR2⁻ ebgA108*	HALL and ZUZEL (1980b)	pUF 5
5A1032	*ebgR105^{+L} ebgA109*	HALL (1982a)	pUF 26
SJ 60	*ebgR52⁻ ebgA205*	ROLSETH, FRIED and HALL (1980)	pUF 2
R42	*ebgR1⁻ ebgA143*	HALL and ZUZEL (1980b)	pUF 16
SJ48	*ebgR7⁻ ebgA168*	HALL (1980)	pUF 4
5A101	*ebgR103^{+L} ebgA51*	HALL (1978b)	None
5A102	*ebgR104^{+L} ebgA51*	HALL (1978b)	None
5A103	*ebgR105^{+L} ebgA51*	HALL (1978b)	None
5A104	*ebgR106^{+L} ebgA51*	HALL (1978b)	None
5A105	*ebgR107^{+L} ebgA51*	HALL (1978b)	None
5A106	*ebgR108^{+L} ebgA51*	HALL (1978b)	None
5A107	*ebgR109^{+L} ebgA51*	HALL (1978b)	None
5A108	*ebgR110^{+L} ebgA51*	HALL (1978b)	None
5A109	*ebgR111^{+L} ebgA51*	HALL (1978b)	None

RESULTS AND DISCUSSION

Structure of the *ebg* operon: The sequence of the wild type *ebg* operon is shown in Figure 1. The sequence of *ebgR* is identical to that previously reported (STOKES and HALL 1985), however, due to a book keeping error, that sequence was previously identified as that of the *ebgR105^{+L}* allele.

The location and gene order of the *ebg* operon was previously determined by classical genetic mapping (HALL and HARTL 1975). We have now compared the restriction map of the *ebg* operon, as deduced from the DNA sequence, with the restriction map of the whole *E. coli* chromosome (KOHARA, AKIYAMA and ISONO 1987). The previously reported gene order (*tolC-ebgR-ebgA-argG*) is confirmed. However, the restriction map places the *ebg* operon between kilobase 3278 and 3283 of the Kohara map of the *E. coli* chromosome. This corresponds to 67.5 min on the genetic map, rather than 66 min where it was originally mapped.

We had previously reported that *ebg* enzyme was encoded by a single gene, *ebgA* (STOKES, BETTS and HALL 1985). We now find that the enzyme consists of two subunits, encoded by adjacent genes *ebgA* and *ebgC*. Note that *ebgC* shares a 4-bp overlap with the end of *ebgA*. Evidence presented below supports the hypothesis that the product of the *ebgC* gene forms part of the active *ebg* enzyme. The *ebgA* gene begins 201 bp (67 amino acids) upstream of the previously reported site. The *ebg* operon includes at least one additional gene, *ebgB*, that encodes a protein of MW 68,000 (HALL and ZUZEL 1980a), and that is located distal to *ebgC*.

A stem-loop that probably functions as a terminator for the *ebgR* mRNA encompasses bp 1120–1155.

The transcription regulation region includes bp 1167–1266. The *ebg* operon is subject to catabolite repression (B. G. HALL, unpublished results). The region from 1167–1182 is a good candidate for the cyclic AMP receptor (CAP) protein binding site. It is 50% identical with the CAP protein binding site of the *lac* operon, and it resembles the consensus CAP binding region in that which begins with TGTGA and contains a less well conserved inverted repeat 6 bp downstream from this sequence (DE CROMBRUGGHE, BUSBY and BUC 1984). The putative *ebg* CAP region begins with CGTGA and ends with the partly conserved inverted repeat TAAAG. The potential −10 region at bp 1247–1252 matches the canonical consensus sequence at 4 bases, and the potential −35 region at bp 1217–1222 matches 4 out of 6 bp of the canonical consensus sequence. The rather poor fit of the putative *ebg* promoter to the canonical consensus, and the sub-optimal spacing between the −35 and the −10 regions of the promoter are also typical of CAP-dependent promoters (DE CROMBRUGGHE, BUSBY and BUC 1984). The region including bp 1247–1266 is palindromic and is a good candidate for the repressor binding site, however no operator mutants are available to rigorously define that region.

At the translation level, a ribosomal binding site for *ebgA* is present at bp 1283–1289, and for *ebgC* at bp 4374–4378.

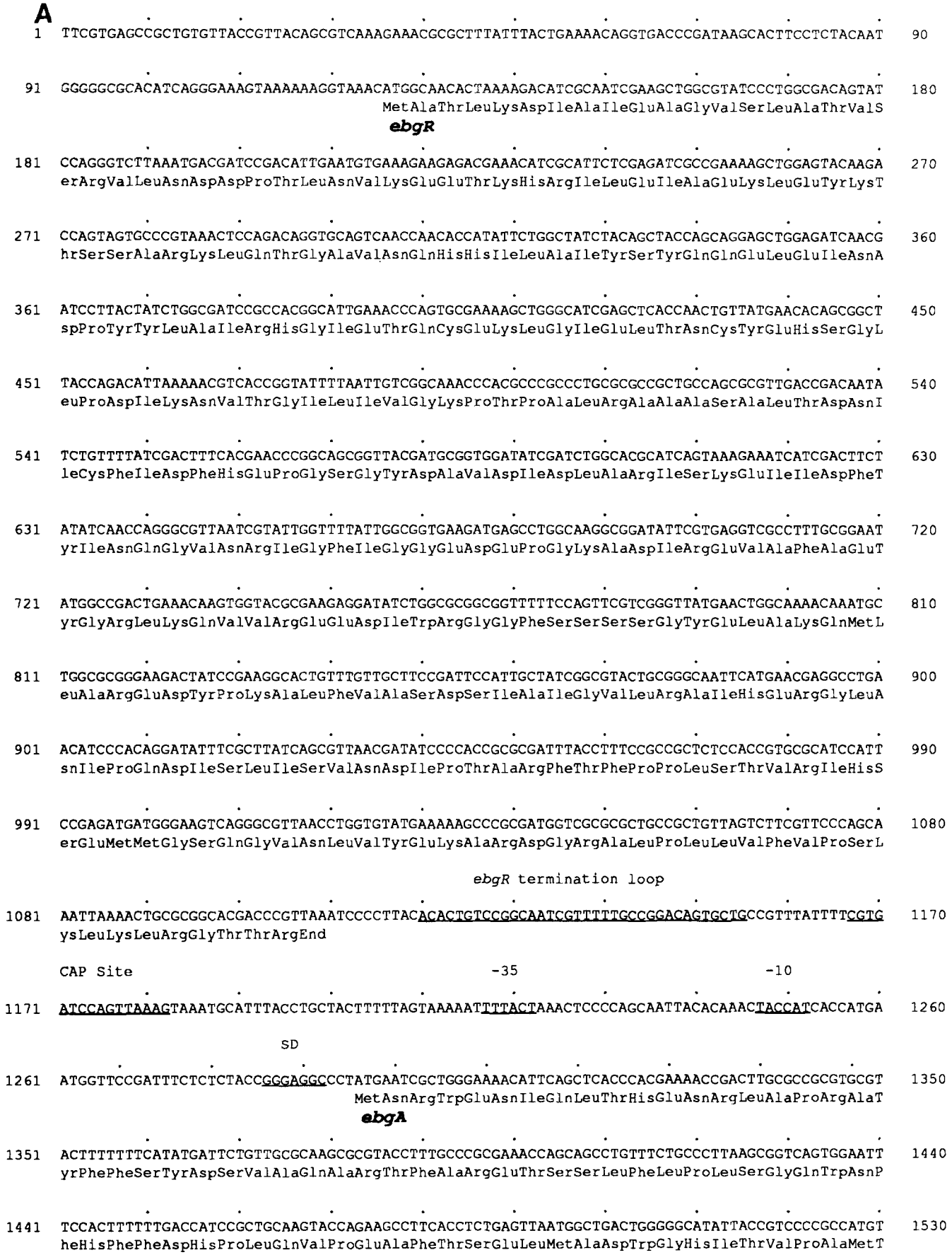

FIGURE 1.—Sequence of the wild-type *ebg* operon of *E. coli* K12. CAP, −35 and −10 refer to the probable cyclic AMP receptor protein binding site, and to the probable −35 and −10 regions of the promoter, respectively. SD indicates probable ribosome binding sites. Note the overlap between the end of the *ebgA* and the start of the *ebgC* coding regions.

B

```
1531  GGCAAATGGAAGGTCACGGCAAACTGCAATATACCGACGAAGGTTTTCCGTTCCCCATCGATGTGCCGTTTGTCCCCAGCGATAACCCAA  1620
      rpGlnMetGluGlyHisGlyLysLeuGlnTyrThrAspGluGlyPheProPheProIleAspValProPheValProSerAspAsnProT

1621  CCGGTGCCTATCAACGTATTTTCACCCTCAGCGACGGCTGGCAGGGTAAACAGACGCTGATTAAATTTGACGGCGTCGAAACCTATTTTG  1710
      hrGlyAlaTyrGlnArgIlePheThrLeuSerAspGlyTrpGlnGlyLysGlnThrLeuIleLysPheAspGlyValGluThrTyrPheG

1711  AAGTCTATGTTAACGGTCAGTATGTGGGTTTCAGCAAGGGCAGTCGCCTGACCGCAGAGTTTGACATCAGCGCGATGGTTAAAACCGGCG  1800
      luValTyrValAsnGlyGlnTyrValGlyPheSerLysGlySerArgLeuThrAlaGluPheAspIleSerAlaMetValLysThrGlyA

1801  ACAACCTGTTGTGTGTGCGCGTGATGCAGTGGGCGGACTCTACCTACGTGGAAGACCAGGATATGTGGTGGTCAGCGGGGATCTTCCGCG  1890
      spAsnLeuLeuCysValArgValMetGlnTrpAlaAspSerThrTyrValGluAspGlnAspMetTrpTrpSerAlaGlyIlePheArgA

1891  ATGTTTATCTGGTCGGAAAACACCTAACGCATATTAACGATTTCACTGTGCGTACCGACTTTGACGAAGCCTATTGCGATGCCACGCTTT  1980
      spValTyrLeuValGlyLysHisLeuThrHisIleAsnAspPheThrValArgThrAspPheAspGluAlaTyrCysAspAlaThrLeuS

1981  CCTGCGAAGTGGTGCTGGAAAATCTCGCCGCCTCCCCTGTCGTCACGACGCTGGAATATACCCTGTTTGATGGCGAACGCGTGGTGCACA  2070
      erCysGluValValLeuGluAsnLeuAlaAlaSerProValValThrThrLeuGluTyrThrLeuPheAspGlyGluArgValValHisS

2071  GCAGCGCCATTGATCATTTGGCAATTGAAAAACTGACCAGCGCCACGTTTGCTTTTACTGTCGAACAGCCGCAGCAATGGTCAGCAGAAT  2160
      erSerAlaIleAspHisLeuAlaIleGluLysLeuThrSerAlaThrPheAlaPheThrValGluGlnProGlnGlnTrpSerAlaGluS

2161  CCCCTTATCTTTACCATCTGGTCATGACGCTGAAAGACGCCAACGGCAACGTTCTGGAAGTGGTGCCACAACGCGTTGGCTTCCGTGATA  2250
      erProTyrLeuTyrHisLeuValMetThrLeuLysAspAlaAsnGlyAsnValLeuGluValValProGlnArgValGlyPheArgAspI

2251  TCAAAGTGCGCGACGGTCTGTTCTGGATCAATAACCGTTATGTGATGCTGCACGGCGTCAACCGTCACGACAACGATCATCGCAAAGGCC  2340
      leLysValArgAspGlyLeuPheTrpIleAsnAsnArgTyrValMetLeuHisGlyValAsnArgHisAspAsnAspHisArgLysGlyA

2341  GCGCCGTTGGAATGGATCGCGTCGAGAAAGATCTCCAGTTGATGAAGCAGCACAATATCAACTCCGTGCGTACCGCTCACTACCCGAACG  2430
      rgAlaValGlyMetAspArgValGluLysAspLeuGlnLeuMetLysGlnHisAsnIleAsnSerValArgThrAlaHisTyrProAsnA

2431  ATCCGCGTTTTTACGAACTGTGTGATATCTACGGCCTGTTTGTGATGGCGGAAACCGACGTCGAATCGCACGGCTTTGCTAATGTCGGCG  2520
      spProArgPheTyrGluLeuCysAspIleTyrGlyLeuPheValMetAlaGluThrAspValGluSerHisGlyPheAlaAsnValGlyA

2521  ATATTAGCCGTATTACCGACGATCCGCAGTGGGAAAAAGTCTACGTCGAGCGCATTGTTCGCCATATCCACGCGCAGAAAAACCATCCGT  2610
      spIleSerArgIleThrAspAspProGlnTrpGluLysValTyrValGluArgIleValArgHisIleHisAlaGlnLysAsnHisProS

2611  CGATCATCATCTGGTCGCTGGGCAATGAATCCGGCTATGGCTGTAACATCCGCGCGATGTACCATGCGGCGAAACGGCTGGATGACACGC  2700
      erIleIleIleTrpSerLeuGlyAsnGluSerGlyTyrGlyCysAsnIleArgAlaMetTyrHisAlaAlaLysArgLeuAspAspThrA

2701  GACTGGTGCATTACGAAGAAGATCGCGATGCTGAAGTGGTCGATATTATTTCCACCATGTACACCCGCGTGCCGCTGATGAATGAGTTTG  2790
      rgLeuValHisTyrGluGluAspArgAspAlaGluValValAspIleIleSerThrMetTyrThrArgValProLeuMetAsnGluPheG

2791  GTGAATACCCGCATCCGAAGCCGCGCATCATCTGTGAATATGCTCATGCGATGGGGAACGGACCGGGCGGGCTGACGGAGTACCAGAACG  2880
      lyGluTyrProHisProLysProArgIleIleCysGluTyrAlaHisAlaMetGlyAsnGlyProGlyGlyLeuThrGluTyrGlnAsnV

2881  TCTTCTATAAGCACGATTGCATTCAGGGTCATTATGTCTGGGAGTGGTGCGACCACGGGATCCAGGCACAGGACGACCACGGCAATGTCT  2970
      alPheTyrLysHisAspCysIleGlnGlyHisTyrValTrpGluTrpCysAspHisGlyIleGlnAlaGlnAspAspHisGlyAsnValT

2971  GGTATAAATTCGGCGGCGACTACGGCGACTATCCCAACAACTATAACTTCTGTCTTGATGGTTTGATCTATTCCGATCAGACGCCGGGAC  3060
      rpTyrLysPheGlyGlyAspTyrGlyAspTyrProAsnAsnTyrAsnPheCysLeuAspGlyLeuIleTyrSerAspGlnThrProGlyP
```

FIGURE 1B

C

```
3061 CGGGCCTGAAAGAGTACAAACAGGTTATCGCGCCGGTAAAAATCCACGCGCGGGATCTGACTCGCGGCGAGTTGAAAGTCGAAAATAAAC 3150
     roGlyLeuLysGluTyrLysGlnValIleAlaProValLysIleHisAlaArgAspLeuThrArgGlyGluLeuLysValGluAsnLysL

3151 TGTGGTTTACCACGCTTGATGACTACACCCTGCACGCAGAGGTGCGCGCCGAAGGTGAAAGCCTCGCGACGCAGCAGATTAAACTGCCGG 3240
     euTrpPheThrThrLeuAspAspTyrThrLeuHisAlaGluValArgAlaGluGlyGluSerLeuAlaThrGlnGlnIleLysLeuProA

3241 ACGTTGCGCCGAACAGCGAAGCCCCCTTGCAGATCACGCTGCCGCAGCTGGACGCCCGCGAAGCGTTCCTCAACATTACGGTGACCAAAG 3330
     spValAlaProAsnSerGluAlaProLeuGlnIleThrLeuProGlnLeuAspAlaArgGluAlaPheLeuAsnIleThrValThrLysA

3331 ATTCCCGCACCCGCTACAGCGAAGCCGGACACCCTATCGCCACTTATCAGTTCCCGCTGAAGGAAAACACCGCGCAGCCAGTGCCTTTCG 3420
     spSerArgThrArgTyrSerGluAlaGlyHisProIleAlaThrTyrGlnPheProLeuLysGluAsnThrAlaGlnProValProPheA

3421 CACCAAATAATGCGCGTCCGCTGACGCTGGAAGACGATCGTTTGAGCTGCACCGTTCGCGGCTACAACTTCGCGATCACCTTCTCAAAAA 3510
     laProAsnAsnAlaArgProLeuThrLeuGluAspAspArgLeuSerCysThrValArgGlyTyrAsnPheAlaIleThrPheSerLysM

3511 TGAGTGGCAAACCGACATCCTGGCAGGTGAATGGCGAATCGCTGCTGACTCGCGAGCCAAAGATCAACTTCTTCAAGCCGATGATGATCG 3600
     etSerGlyLysProThrSerTrpGlnValAsnGlyGluSerLeuLeuThrArgGluProLysIleAsnPhePheLysProMetMetIleA

3601 ACAACCACAAGCAGGAGTACGAAGGGCTGTGGCAACCGAATCATTTGCAGATCATGCAGGAACATCTGCGCGACTTTGCCGTAGAACAGA 3690
     spAsnHisLysGlnGluTyrGluGlyLeuTrpGlnProAsnHisLeuGlnIleMetGlnGluHisLeuArgAspPheAlaValGluGlnS

3691 GCGATGGTGAAGTGCTGATCATCAGCCGCACAGTTATTGCCCCGCCGGTGTTTGACTTCGGGATGCGCTGCACCTACATCTGGCGCATCG 3780
     erAspGlyGluValLeuIleIleSerArgThrValIleAlaProProValPheAspPheGlyMetArgCysThrTyrIleTrpArgIleA

3781 CTGCCGATGGCCAGGTTAACGTGGCGCTTTCCGGCGAGCGTTACGGCGACTATCCGCACATCATTCCGTGCATCGGTTTCACCATGGGAA 3870
     laAlaAspGlyGlnValAsnValAlaLeuSerGlyGluArgTyrGlyAspTyrProHisIleIleProCysIleGlyPheThrMetGlyI

3871 TTAACGGCGAATACGATCAGGTGGCGTATTACGGTCGTGGACCGGGCGAAAACTACGCCGACAGCCAGCAGGCTAACATCATCGATATCT 3960
     leAsnGlyGluTyrAspGlnValAlaTyrTyrGlyArgGlyProGlyGluAsnTyrAlaAspSerGlnGlnAlaAsnIleIleAspIleT

3961 GGCGCCAAGCCGTCGATGCCATGTTCGAGAACTATCCCTTCCCGCAGAACAACGGTAACCGTCAGCATGTCCGCTGGACGGCACTGACTA 4050
     rpArgGlnAlaValAspAlaMetPheGluAsnTyrProPheProGlnAsnAsnGlyAsnArgGlnHisValArgTrpThrAlaLeuThrA

4051 ACCGCCACGGTAACGGTCTGCTGGTGGTTCCGCAGCGCCCAATTAACTTCAGCGCCTGGCACTATACCCAGGAAAACATCCACGCTGCCC 4140
     snArgHisGlyAsnGlyLeuLeuValValProGlnArgProIleAsnPheSerAlaTrpHisTyrThrGlnGluAsnIleHisAlaAlaG

4141 AGCACTGTAACGAGCTGCAGCGCAGTGATGACATCACCCTGAACCTCGATCACCAGCTGCTTGGCCTCGGCTCCAACTCCTGGGGCAGCG 4230
     lnHisCysAsnGluLeuGlnArgSerAspAspIleThrLeuAsnLeuAspHisGlnLeuLeuGlyLeuGlySerAsnSerTrpGlySerG

4231 AGGTGCTGGACTCCTGGCGCGTCTGGTTCCGTGACTTCAGCTACGGCTTTACGTTGCTGCCGGTTTCTGGCGGAGAAGCTACCGCGCAAA 4320
     luValLeuAspSerTrpArgValTrpPheArgAspPheSerTyrGlyPheThrLeuLeuProValSerGlyGlyGluAlaThrAlaGlnS

                                                    SD

4321 GCCTGGCGTCGTATGAGTTCGGCGCAGGGTTCTTTTCCACGAATTTGCACACGGAGAATAAGCAATGAGGATCATCGATAACTTAGAACA 4410
     erLeuAlaSerTyrGluPheGlyAlaGlyPhePheSerThrAsnLeuHisThrGluAsnLysGlnEnd  IleIleAspAsnLeuGluGl
                  Start of ebgC overlaps end of ebgA      MetArg

4411 GTTCCGCCAGATTTACGCCTCTGGCAAGAAGTGGCAACGCTGCGTTGAAGCGATTGAAAATATCGACAACATTCAGCCTGGCGTCGCCCA 4500
     nPheArgGlnIleTyrAlaSerGlyLysLysTrpGlnArgCysValGluAlaIleGluAsnIleAspAsnIleGlnProGlyValAlaHi

4501 CTCCATCGGTGACTCATTGACTTACCGCGTGGAGACAGACTCCGCGACCGATGCGCTATTTACCGGGCATCGACGCTATTTTGAAGTGCA 4590
     sSerIleGlyAspSerLeuThrTyrArgValGluThrAspSerAlaThrAspAlaLeuPheThrGlyHisArgArgTyrPheGluValHi
```

FIGURE 1C

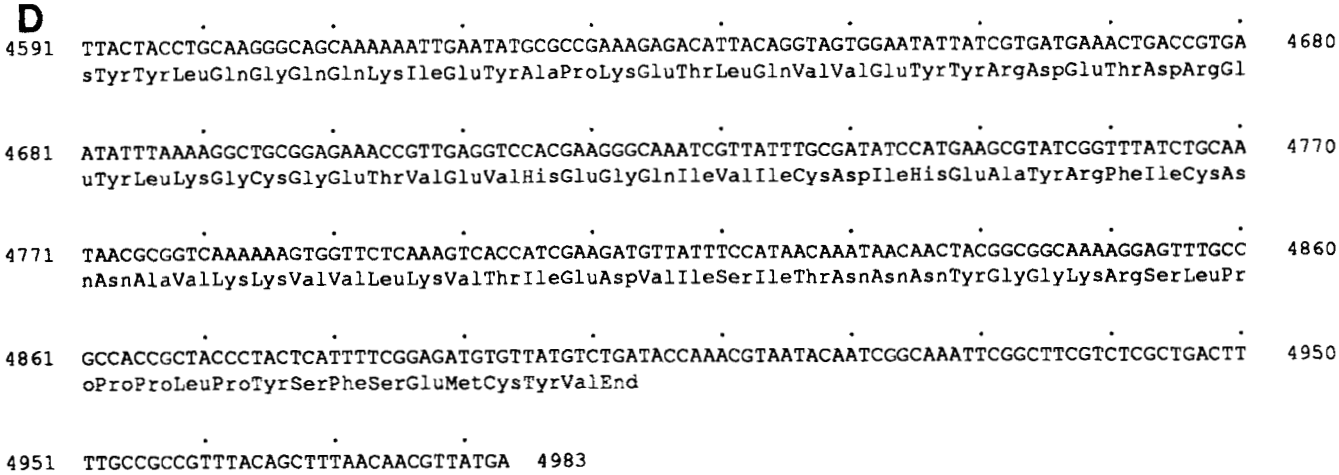

FIGURE 1D

TABLE 2

Properties of *ebg* operon gene products

Gene	Span	Product	MW (calculated)	MW (SDS-PAGE)
ebgR	bp 126–1109	Repressor	36,169	Not determined
ebgA	bp 1293–4388	β-Galactosidase, α subunit	117,927	120,000[a]
ebgC	bp 4385–4903	β-Galactosidase, β subunit	19,917	22,000
ebgB	After bp 4979	Unknown		79,000[b]

[a] From HALL (1976).
[b] From HALL and ZUZEL (1980a).

Properties of the *ebg* gene products, as deduced from the DNA sequence, are compared with some actual properties in Table 2.

Evidence that the *ebgC* gene product is part of the active *ebg* enzyme: Purified preparations of *ebg* enzyme that were used in a previous study (HALL 1981) occasionally contained a contaminating protein of MW ~18,000 as judged by SDS-PAGE gels stained with Coomassie brilliant blue. A similar contaminating band was detected by M. SINNOTT (personal communication). Because the intensity of that band was quite variable relative to the 120,000 MW *ebgA* encoded band, it was assumed that the band represented an unrelated protein that sometimes copurified with *ebg* enzyme. Reexamination of some purified preparations that had been stored at −70° since 1980 showed that the "contaminating" 22,000 MW band was present at about the same concentration as the 120,000 MW band when proteins were detected by silver staining. That band can not be detected in these preparations by staining with Coomassie brilliant blue.

A pair of plasmids was constructed to determine the role of the *ebgC* gene product in *ebg* enzyme activity. One plasmid, pUF856, contained both the *ebgA* and *ebgC* genes (bp 1187–5385) from the class II allele *ebgA51*, while the otherwise identical plasmid, pUF854, carried only the *ebgA* gene (bp 1187–4395) from that strain. Crude enzyme extracts were prepared from the *ebg* deletion strain SJ84R harboring each of the plasmids, and the K_m and V_{max} of the *ebg* enzyme from those extracts was determined in triplicate. The enzyme encoded by pUF856 exhibited a K_m of 0.48 ± 0.06 mM *O*-nitrophenyl-β-galactoside (ONPG) and a V_{max} of 5,300 ± 300 nmol/min; while that from plasmid pUF854 (lacking the *ebgC* peptide) exhibited a K_m of 0.96 ± 0.12 mM ONPG and a V_{max} of 92.2 ± 3.8 nmol/min. The K_m for the enzyme encoded by pUF856 was in good agreement with that reported for purified class II *ebg* enzymes, 0.56 ± 0.04 mM ONPG (HALL 1981). The absence of the *ebgC* peptide thus reduces the *ebg* enzyme activity toward ONPG by about 50-fold. *In vivo*, plasmid pUF856 confers a strong lactose positive phenotype on MacConkey plates, and produces intensely blue colonies on XGAL plates. In contrast, plasmid pUF854 confers a lactose negative phenotype on MacConkey plates, and produces pale blue colonies on XGAL plates. It seems reasonable to conclude the *ebgC* gene product is required for full activity of *ebg* enzyme. The *ebg* enzyme thus consists of two subunits: the α-subunit, encoded by *ebgA*, and the β-subunit, encoded by *ebgC*.

The MW of the native *ebg* enzyme was originally reported as 720,000 on the basis of sedimentation equilibrium measurements (HALL 1976), indicating a hexameric structure. That value was based on an estimate, rather than a measurement, of the partial

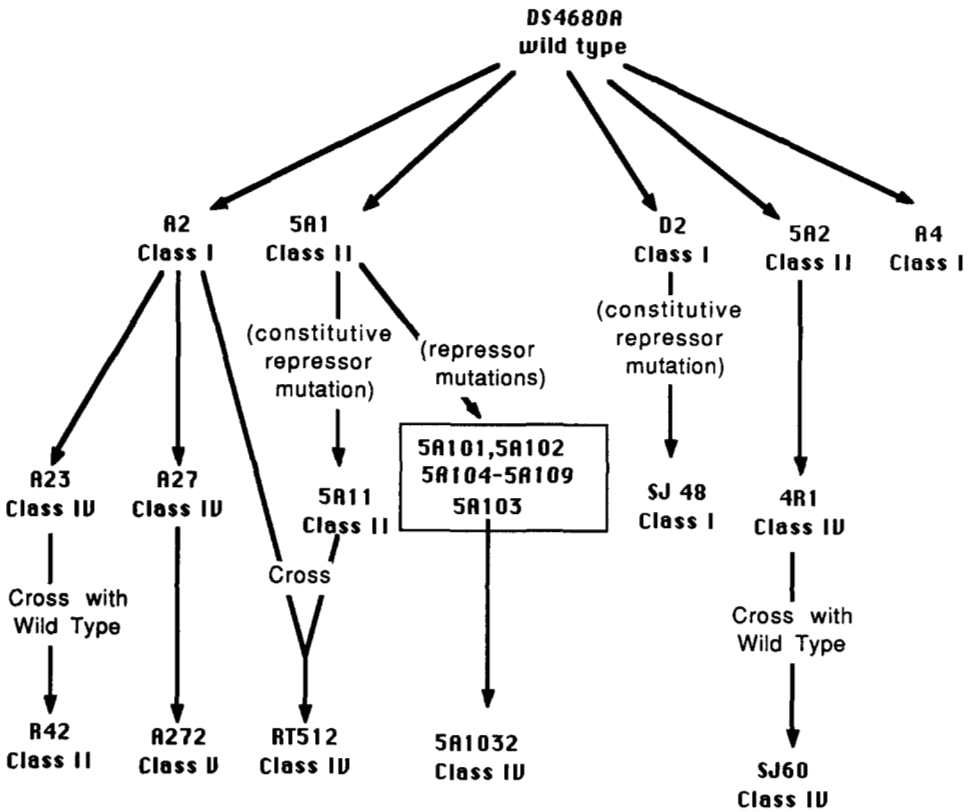

FIGURE 2.—Phylogeny of *ebg* mutants. The class of enzyme synthesized (see text) is shown below each strain name.

specific volume of *ebg* enzyme. Assuming that the native structure involves an equal number of α and β, as would appear to be the case from relative silver staining intensities, an $\alpha 6$-$\beta 6$ structure would yield a native MW 827,000 and an $\alpha 4$-$\beta 4$ would yield 551,400; neither of which is easily reconciled with the previously estimated 720,000.

Analysis of mutations involved in the evolution of new functions

Figure 2 shows the phylogeny of the alleles that are discussed in this study. In order to confirm the identities of cloned mutant alleles, the regions bearing the mutations were amplified by PCR from genomic DNA of the originating strain. In all cases the cloned genes were identical with the genes in the original strains.

Mutations in *ebg* enzyme structural genes: The *ebgA* and *ebgC* genes from the wild-type and five mutant strains have been sequenced in their entireties. Five additional mutant alleles were partially sequenced, either by PCR or by direct sequencing from plasmid DNA. The mutation at the class I site is identical in four independent alleles, however two versions of the class II site were detected among the five independent class II mutations examined (Table 3). Properties of enzymes encoded by the different sequences are presented in Table 4.

The *ebg* enzymes were labeled with the active site directed inhibitor 4-nitrophenyl-β-D-galactopyranosylmethyltriazine (FOWLER and SMITH 1983). Wild-type enzyme and class I enzyme specified by the *ebgA2* allele were preferentially labeled in a peptide encoded by bp 2703–2762, while the class II enzyme encoded by the *ebgA52* allele was preferentially labeled in a peptide encoded by bp 2625–2660. It is notable that none of the mutations which alter the catalytic specificities of *ebg* enzyme fall within this active site region.

Although it is clear that the β-subunit plays an important role in *ebg* enzyme activity, it seems unlikely that the mutation in *ebgC*, and the second mutation in the class II site of *ebgA*, significantly affect the activities of *ebg* enzyme. Properties of the purified enzymes encoded by the class IV alleles listed in Table 3 have been determined (HALL 1981), and there is no indication that the enzymes with the mutant β-subunit are significantly different from those with the wild type β-subunit.

The thermal stabilities of various *ebg* enzymes were determined (HALL 1981), and no significant correlation was found between the rate of thermal inactivation and the number of mutations assumed to be present in each enzyme (classes I and II were assumed to have 1 mutation, class IV 2 mutations, and class V 3 mutations). We have reexamined the data in light of the sequencing data, and there is a highly significant ($P < 0.01$) effect of the number of amino acid substitutions on thermal stability, with about a 13.6% in-

TABLE 3

Mutations in *ebg* enzyme genes

Strain	Genotype	Enzyme class	Sequence changes		Comment
			ebgA	*ebgC*	
DS4680A	Wild type	0	None	None	Completely sequenced
A4	*ebgA4*	I	1566 G → A	None	Completely sequenced
A2	*ebgA2*	I	1566 G → A	None	Completely sequenced
SJ48	*ebgA168*	I	1566 G → A	None	Class I and II sites sequenced
5A11	*ebgA51*	II	4223 G → T	None	Class I and II sites sequenced
5A2	*ebgA52*	II	4223 G → A	None	Class I and II sites sequenced
R42	*ebgA143*	II	4223 G → T	None	Recombinant, carries class II site of A23; class I and II sites sequenced
A23	*ebgA134*	IV	1566 G → A 4223 G → T	None	Class I and II sites sequenced
A27	*ebgA138*	IV	1566 G → A 4223 G → A 4227 A → G	4749 A → G	Class I and II sites sequenced
RT512	*ebgA108*	IV	1566 G → A 4223 G → T	None	Recombinant, carries class I site of A2 and class II site of 5A11; completely sequenced
SJ60	*ebgA205*	IV	1566 G → A 4223 G → A 4227 A → G	4749 A → G	Completely sequenced
A272	*ebgA198*	V	1566 G → A 1569 G → A 4223 G → A 4227 A → G	4749 A → G	Completely sequenced

crease in the rate of decay with each additional substitution (Figure 3).

Mutations in the *ebg* repressor gene: The *ebg* operon is subject to negative control by the repressor encoded by *ebgR* (HALL 1978b; HALL and CLARKE 1977; HALL and HARTL 1975). Three kinds of regulatory mutations have been reported, $ebgR^-$, $ebgR^{+U}$ and $ebgR^{+L}$.

Three independent $ebgR^-$ (constitutive *ebg* enzyme synthesis) alleles were sequenced (Table 5). The two spontaneous mutations, *ebgR2* and *ebgR52*, were single base insertion frame shifts, and the EMS induced *ebgR1* mutation involved two adjacent substitutions that resulted in a nonsense codon.

The existence of $ebgR^{+U}$ alleles was deduced from the observation that regulated ebg^+ (lactose utilizing) mutants synthesized four times as much *ebg* enzyme *protein* upon induction with lactose as did wild-type (unevolved) strains. One such strain, A4, synthesized class I *ebg* enzyme and was defined as $ebgR4^{+U}$ *ebgA4*. A cross between wild type and a constitutive derivative of strain A4 generated a recombinant that synthesized class I enzyme at the level expected of wild type strains, indicating that the difference in level of gene expression was not a function of the *ebgA* allele present. It was concluded that such strains possessed mutant repressor alleles that were more sensitive than wild type to lactose induction (HALL and CLARKE 1977). Repeated sequencing of both strands of the $ebgR4^{+U}$ allele failed to detect any differences from the wild-type sequence (Table 5). We have reexamined the recombinant strain from the earlier study, and confirmed that it does synthesize class I enzyme at the wild-type level, thus strain A4 did indeed have at least two mutations that distinguished it from wild type. We conclude that the second mutation must be distal to *ebgA*, and that the recombinant strain was a double recombinant. The designation $ebgR^{+U}$ is thus inappropriate.

The wild-type repressor is not inducible by either lactulose or Gal-Ara, and a deliberate effort was made to isolate spontaneous lactulose inducible *ebgR* mutants (HALL 1978b). Nine lactulose inducible ($ebgR^{+L}$) mutants were isolated. Because those mutations affected only the specificity of induction, and not the basal level of synthesis, it was expected that the mutations would occur in the sugar binding domain of the repressor. Residues likely to be involved in sugar binding were predicted from database searching and sequence alignment (Figure 4) of *ebgR* with other repressors (*lacI*, *galR*, and *cytR*), and with sugar binding regions of periplasmic chemoreceptor proteins.

TABLE 4

Properties of mutant *ebg* enzymes

Allele	Enzyme class	Amino acid substitutions	Substrate	K_m[a] (mM)	V_{max}[a] (nmol/min/mg)	Comment[b]
Wild type	0	None	Lactose	150	620	Can not synthesize allolactose from lactose
			Lactulose	180	270	
			Gal-Ara	64	52	
			Lactobionate	No detectable activity		
ebgA2	I	Asp-92 → Asn	Lactose	22	4200	Can not synthesize allolactose from lactose
			Lactulose	57	5113	
			Gal-Ara	24	340	
			Lactobionate	No detectable activity		
ebgA52	II	Trp-977 → Cys	Lactose	72	2700	Can not synthesize allolactose from lactose
			Lactulose	34	2200	
			Gal-Ara	34	460	
			Lactobionate	No detectable activity		
ebgA134	IV	Asp-92 → Asn	Lactose	0.82	1600	Efficiently synthesizes allolactose from lactose
		Trp-977 → Cys	Lactulose	6.2	470	
			Gal-Ara	2.8	940	
			Lactobionate	15	67	
ebgA138, C138	IV	Asp-92 → Asn	Lactose	0.89	1710	Efficiently synthesizes allolactose from lactose
		Trp–977 → Cys	Lactulose	10.7	480	
		Ser-979 → Gly	Gal-Ara	4.3	840	
	(β-subunit)	Glu-122 → Gly	Lactobionate	9	74	
ebgA198	V	Asp-92 → Asn	Lactose	0.69	590	
		Glu-93 → Lys	Lactulose	6.5	215	
		Trp-977 → Cys	Gal-Ara	4.96	349	
		Ser-979 → Gly	Lactobionate	3.0	370	
	(β-subunit)	Glu-122 → Gly				

[a] Data from HALL (1981).
[b] Data from HALL (1982b).

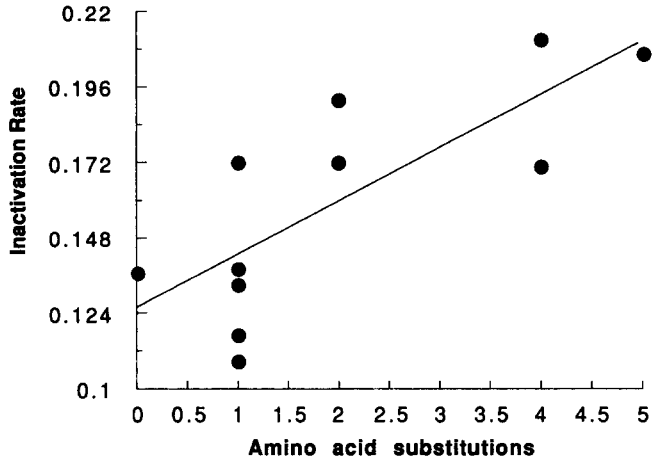

FIGURE 3.—Thermal stabilities of *ebg* enzymes. Abscissa is the number of amino acid replacements in the enzyme when compared with the wild type sequence, ordinate is the first order decay constant in min^{-1}, measured at 50°.

From the high resolution crystal structures of the L-arabinose binding protein (ABP) (QUIOCHO and VYAS 1984), and its structural homolog the D-galactose/D-glucose binding protein (GBP) (VYAS, VYAS and QUIOCHO 1988), critical sugar binding residues were used to construct pattern discriminators and to identify the best matching segments of the repressors. This analysis, summarized in Figure 4, confirmed and extended the homology between *lacI*, *galR*, and the sugar binding protein that had previously been noted (MULLER-HILL 1983; SARIS *et al.* 1984).

Three different substitutions were found in the nine *ebgR*$^{+L}$ alleles. The sequence of the entire *ebgR* gene was determined for one allele, *ebgR105*$^{+L}$, and the region around the site of the mutation in *ebgR105*$^{+L}$ was sequenced in the remaining eight alleles. The *ebgR103* and *ebgR106* alleles have identical substitutions, *ebgR105* is unique, and *ebgR104* plus *ebgR107* through *ebgR111* have identical substitutions (Table 5). All three amino acid substitutions fell within a predicted sugar binding region. The three amino acid substitutions, Asp-190 to Gly, Ala-195 to Thr, and Phe-196 to Cys, are close to the predicted primary sugar contacting residue Arg-192. The corresponding critical arginines (Figure 4), Arg-151 of ABP and Arg-158 of GBP, each donate two hydrogen bonds to oxygen atoms of the bound sugars (QUIOCHO *et al.* 1987; VYAS, VYAS and QUIOCHO 1988). Presumably residues 188 to 196 of *ebgR* are similarly crucial for the specificity of molecular recognition of lactose, possibly by affecting the orientation of Arg-192 through secondary contacts or through the polypeptide conformation.

TABLE 5

Mutations in *ebgR*

Allele	Mutation	Amino acid replacement	Relative level of *ebg* operon expression with inducer present[a]			
			None	Lactose	Lactulose	Gal-Ara
WT	None	None	1.0	103	11	8
ebgR103^{+L}, *ebgR106*$^{+L}$	bp 695 T → G	Asp-190 → Glu	0.9	268	207	229
ebgR105^{+L}	bp 708 G → A	Ala-195 → Thr	1.0	821	450	753
ebgR104^{+L}, *ebgR107*$^{+L}$, *ebgR108*$^{+L}$, *ebgR109*$^{+L}$, *ebgR110*$^{+L}$, *ebgR111*$^{+L}$	bp712 T → G	Phe-196 → Cys	1.1	711	300	768
ebgR1^{-}	bp 760 & 761 GG → AA	Trp-212 → ochre	2200			
ebgR2^{-}	A or T inserted[b] after bp 1021	Frameshift at Leu-299	2100			
ebgR52^{-}	T inserted after bp 1021	Frameshift at Leu-299	2300			

[a] Data from HALL (1978b) and HALL and CLARKE (1977).
[b] Determined by the creation of a *Sty*I restriction site and the loss of a *Bst*NI restriction site, not directly by DNA sequencing.

The three amino acid replacements in *ebgR*$^{+L}$ alleles broaden the specificity of inducer recognition to include lactulose and Gal-Ara (Table 5). This implies that the sugar contacts made by the 188–196 region of the *ebgR* encoded repressor, probably involving direct bonds of Arg-192, are to the glucopyranose moiety of lactose, and, in the mutant alleles, to the fructofuranose and arabinopyranose moieties of lactulose and Gal-Ara. The three sugars are identical in the galactopyranose moiety. It is not possible to make more precise stereochemical predictions about the bonding of Arg-192 to these three sugars, because there are differences between the ABP and GBP in the roles of the corresponding arginines, resulting from opposite orientations of the bound sugars. Arg-151 of ABP donates hydrogen bonds to O-4 and O-5 of L-arabinose, whereas Arg-158 of GBP bonds O-1 and O-2 of D-glucose (QUIOCHO *et al.*, 1987; VYAS, VYAS and QUIOCHO 1988). An alternative interpretation is that the 188 to 196 regions of the wild-type *ebgR* repressor might exclude lactulose and Gal-Ara by steric hindrance, whereas a looser conformation of this region in the repressors encoded by mutant *ebgR*$^{+L}$ alleles might permit a broader range of sugar analogs to bind and to act as inducers.

Homology with other genes

All comparisons are based upon alignments of the deduced amino acid sequences using the UWGCG GAP program with a gap penalty of 5.0 and length penalty of 0.3. The *ebg* β-galactosidase was aligned with the *lacZ* β-galactosidase of *E. coli* and with the *lacZ* β-galactosidase of *K. pneumoniae*; and the *ebg* repressor protein was aligned with the *lacI* and *galR* encoded repressors of *E. coli*, and with the *lacI* repressor of *K. pneumoniae* (Table 6).

No homology could be detected between the *ebgC* encoded β subunit sequence and that of either *lacZ* or *lacY* (permease) of *E. coli*.

Both the β-galactosidase and the repressor genes support a picture in which *ebg* diverged from *lac* prior to the time that Klebsiella diverged from *E. coli*. It is very surprising that in two comparisons, the *lac* operon of Klebsiella vs the *lac* operon of *E. coli*, and the *ebg* operon vs the *lac* operon of *E. coli*, the repressor genes have diverged much more than have the β-galactosidase genes. If most of the replacements are neutral, then the repressors and β-galactosidases should have diverged to about the same extent. With respect to selective replacements, the β-galactosidases would be expected to have diverged more than the repressors, since the β-galactosidase of *E. coli* is the most electrophoretically variable proteins tested in the ECOR collection of natural isolates of *E. coli*, with an allelic diversity that is 2.4 times the mean allelic diversity for 35 enzymes studied (SELANDER, CAUGANT and WHITTAM 1987).

Multiple spontaneous mutations

We have two examples of alleles that differ from their immediate ancestral alleles by 3-bp substitutions, *ebgAC134* (strain A27) and *ebgAC205* (strain SJ60). Because all of the mutations reported here, except *ebgR1*, were spontaneous this raises the issue of multiple substitutions as the consequence of single events.

The *ebgA205* allele (strain SJ60) was selected during the course of serial transfer experiments, alternating between lactose and lactulose as carbon sources (ROLSETH, FRIED and HALL 1980). It is therefore quite possible that the three mutations occurred sequentially.

The *ebgAC138* allele (strain A27), in contrast, was isolated as a papilla on the surface of an aged strain A2 colony (allele *ebgA2*) growing on a MacConkey

FIGURE 4.—Alignment of the *ebg* repressor sequence with homologous repressors and sugar binding proteins. Abbreviations for proteins are (NBRF or SWISSPROT database codes in parentheses): repressors: EC EbgR: *ebg* repressor of *E. coli*; EC *LacI*: *lac* repressor of *E. coli* (RPECL, LACR$KLEPN); KP *LacI*: *lac* repressor of *K. pneumoniae*; EC GalR: *gal* repressor of *E. coli* (RPECG); EC CytR: repressor for *deo* operon, *udp* and *cdd* genes of *E. coli* (RPECCT); C-terminal domains of periplasmic chemoreceptor/binding proteins: EC ABP: arabinose binding protein of *E. coli* (JGECA); EC GBP: galactose/glucose binding protein of *E. coli* (JGECG); EC RBP: ribose binding protein of *E. coli* (JGECR). The DNA binding helix-turn-helix motif of the repressors is boxed. Symbols below the ABP sequence indicate critical sugar binding residues deduced from crystal structures of ABP and GBP, which make either direct hydrogen bonds (large open arrows) or indirect hydrogen bonds via other side chains or water (small closed arrows). Arrows above the EC EbgR repressor sequence indicate amino acids substituted in lactulose inducible mutants.

lactulose plate. Under those conditions there was strong selection for lactulose utilization, however the observation that only the Trp → Cys substitution at amino acid 977 is required to generate a class II site (alleles *ebgA51*, *ebgA52* and *ebgA134*) suggests that of the three substitutions, only the G → T substitution at bp 4223 was required to produce the advantageous phenotype. Other cases of multisite spontaneous mutations have been reported (GOLDING and GLICKMAN 1985; HAMPSEY *et al.* 1988) and discussed (DRAKE, GLICKMAN and RIPLEY 1983; GASC, SICARD and CLAVERYS 1989; GOLDING 1987), but in some (but not all) of these cases the sites were close to each other and involved in probable stem-loop structures that could

TABLE 6

Protein homologies

	ebg repressor *E. coli*	*lac* repressor *E. coli*	*gal* repressor *E. coli*
lac repressor *E. coli*	24.9%		
gal repressor *E. coli*	23.3%	24.4%	
lac repressor *K. pneumoniae*	20.1%	39.8%	26.4%
	ebg β-galactosidase *E. coli*	*lacZ* β-galactosidase *E. coli*	
lacZ β-galactosidase	33.7%		
lacZ β-galactosidase *K. pneumoniae*	31.4%	61.0%[a]	

[a] From BUVINGER and RILEY (1985).

permit one mutation to engender a second mutation as the result of templating errors. The sites of the *ebg* multiple mutations do not lend themselves such an interpretation.

Since the probability that those three substitutions would have occurred simultaneously and independently is about 10^{-26}, it does not seem likely that they occurred simultaneously and independently. There are three explanations for the three mutations that distinguish the *ebgA138* allele of strain A27 from its immediate parent: (1) sequencing errors, (2) sequential mutations, and (3) simultaneous mutations that were not independent of each other, but that were the consequence of some common initiating event.

We reject the sequencing error explanation on the grounds that the appropriate sequence from strain A27 was repeatedly amplified and sequenced, and that the sequence changes were clear.

We regard the sequential mutations explanation as implausible on several grounds. In order to account for the isolation of a strain resulting from three sequential mutations from a colony consisting of about 10^9 cells, it must be posited that *each* mutation conferred a growth advantage sufficient to provide a large enough population for the succeeding mutation to occur in the background of the first mutation. Since all of the alleles tested that permit growth on lactulose carry the α-subunit Trp-977 → Cys substitution, that substitution is probably required for lactulose utilization. The question, then, is whether the additional substitutions, α-subunit Ser-979 → Gly and β-subunit Glu-122 → Gly confer a sufficient selective advantage on lactulose. One potential advantage might be increased enzyme stability, but we see a 13.6% *decrease* in stability with each additional substitution (Figure 3). Another possibility is increased enzyme activity, but the specificity (V_{max}/K_m) for lactulose of class IV enzyme with the extra substitutions is only 44.9, compared with 75.8 for class IV enzyme without the extra substitutions. Furthermore, colonies of strains A23 and A27 are indistinguishable on MacConkey lactulose, and the two strains grow at the same rates in lactulose minimal medium. Finally, given a mutation rate of 2×10^{-9} per cell division (HALL 1977), in order to account for sequential selection of three mutations the first step mutant (presumably the α-subunit Trp-977 → Cys substitution) must grow to at least 10^8 cells, thus forming a large, visible papilla. At that stage there is about a 0.2 probability of the second mutation occurring, following which the double mutant must grow to 10^8 cells, forming a second large, visible papilla on the surface of the first. Finally, this must happen once more to generate the third mutation. No papillae-on-papillae were observed when strain A27 was isolated. Since the single substitution mutant grows very well on lactulose, *viz.* strain A23, the second mutation would have to provide a tremendous growth advantage to permit it to reach the required frequency. We see no evidence that even the triple mutant, *viz.* strain A27, provides such an advantage.

The third explanation, simultaneous mutations that were not independent of each other, but that were the consequence of some common initiating event, seems most likely. The *ebgAC138* allele was selected under conditions virtually identical to those that were reported to produce multiple spontaneous mutations that involved an insertion sequence (HALL 1988). In that study, and in an earlier study by CAIRNS, OVERBAUGH and MILLER (1988), it was suggested that some adaptive mutations occur as specific responses to environmental challenges. The mechanisms by which microorganisms may be able to target potentially advantageous genes for mutation are unknown, but at this time the phenomenon appears to be most easily demonstrated in cells within aged, nutritionally depleted, colonies. One possibility is that, under these biologically stressful conditions, transcription is mutagenic. Were this the case, then the *ebg* operon of strain A2 would be particularly vulnerable since it is transcribed at a very high rate [approximately 3% of the total protein synthesized is *ebg* enzyme in that strain (HALL and CLARKE 1977)]. While it is obviously highly speculative, it does not seem completely implausible that mutagenic transcription could be the common initiating event in the occurrence of the three simultaneous substitutions of *ebgAC138*.

This study was supported by U.S. Public Health Service grants AI-14766 and GM-37110 to B.G.H. The ISIS database and software (J.C.W.) was developed under grant GR-D-28881 of the SERC Protein Engineering Initiative.

LITERATURE CITED

AKRIGG, D., A. J. BLEASBY, N. I. M. DIX, J. B. FINDLAY, A. C. T. NORTH, P. D. SMITY, J. C. WOOTTON, T. DLUNDELL, S. P.

GARDNER, F. HAYES, C. ISLAM, M. J. E. STERNBERG, J. M. THORNTON, I. J. TICKLE and P. MURRAY-RUST, 1988 A protein sequence/structure database. Nature **335:** 745–746.

BUVINGER, W. E., and M. RILEY, 1985 Nucleotide sequence of *Klebsiella pneumoniae* lac genes. J. Bacteriol. **163:** 850–857.

CAIRNS, J., J. OVERBAUGH and S. MILLER, 1988 The origin of mutants. Nature **335:** 142–145.

DE CROMBRUGGHE, B., S. BUSBY and H. BUC, 1984 Cyclic AMP receptor protein: role in transcription activation. Science **224:** 831–838.

DRAKE, J. W., B. W. GLICKMAN and L. S. RIPLEY, 1983 Updating the theory of mutation. Am. Sci. **71:** 621–630.

FOWLER, A. V., and P. J. SMITH, 1983 The active site regions of *lacZ* and *ebg* β-galactosidase are homologous. J. Biol. Chem. **258:** 10204–10207.

GASC, A.-M., A.-M. SICARD and J.-P. CLAVERYS, 1989 Repair of single- and multiple-substitution mismatches during recombination in *Streptomyces pneumoniae*. Genetics **121:** 29–36.

GOLDING, G. B., 1987 Multiple substitutions create biased estimates of divergence times and small increases in the variance to mean ratio. Heredity **58:** 331–339.

GOLDING, G. B., and B. W. GLICKMAN, 1985 Sequence-directed mutagenesis: evidence from a phylogenetic history of human α-interferon genes. Proc. Natl. Acad. Sci. USA **82:** 8577–8581.

HALL, B. G., 1976 Experimental evolution of a new enzymatic function. Kinetic analysis of the ancestral (ebg^0) and evolved (ebg^+) enzymes. J. Mol. Biol. **107:** 71–84.

HALL, B. G., 1977 The number of mutations required to evolve a new lactase function in *Escherichia coli*. J. Bacteriol. **129:** 540–543.

HALL, B. G., 1978a Experimental evolution of a new enzymatic function. II. Evolution of multiple functions for EBG enzyme in *E. coli*. Genetics **89:** 453–465.

HALL, B. G., 1978b Regulation of newly evolved enzymes. IV. Directed evolution of the ebg repressor. Genetics **90:** 673–691.

HALL, B. G., 1980 On the evolution of new metabolic functions in diploid organisms. Genetics **96:** 1007–1017.

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

HALL, B. G., 1982a Evolution of a regulated operon in the laboratory. Genetics **101:** 335–344.

HALL, B. G., 1982b Transgalactosylation activity of ebg β-galactosidase synthesizes allelactose from lactose. J. Bacteriol. **150:** 132–140.

HALL, B. G., 1983 Evolution of new metabolic functions in laboratory organisms, pp. 234–257 in *Evolution of Genes and Proteins*, edited by M. NEI and R. KOEHN. Sinauer Associates, Sunderland, Mass.

HALL, B. G., 1988 Adaptive evolution that requires multiple spontaneous mutations. I. Mutations involving and insertion sequence. Genetics **120:** 887–897.

HALL, B. G., and N. D. CLARKE, 1977 Regulation of newly evolved enzymes. III. Evolution of the ebg repressor during selection for enhanced lactase activity. Genetics **85:** 193–201.

HALL, B. G., and D. L. HARTL, 1974 Regulation of newly evolved enzymes. I. Selection of a novel lactase regulated by lactose in Escherichia coli. Genetics **76:** 391–400.

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, 1980a The ebg operon consists of at least two genes. J. Bacteriol. **144:** 1208–1211.

HALL, B. G., and T. ZUZEL, 1980b Evolution of a new enzymatic function by recombination within a gene. Proc. Natl. Acad. Sci. USA **77:** 3529–3533.

HAMPSEY, D. M., J. F. ERNST, J. W. STEWART and F. SHERMAN, 1988 Multiple base-pair mutations in yeast. J. Mol. Biol. **201:** 471–486.

KOHARA, Y., K. AKIYAMA and K. ISONO, 1987 The physical map of the whole *E. coli* chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. Cell **50:** 495–508.

MULLER-HILL, B., 1983 Sequence homology between Lac and Gal repressors and three sugar-binding proteins. Nature **302:** 103–104.

QUIOCHO, F. A., AND N. K. VYAS, 1984 Novel stereospecificity of the L-arabinose binding protein. Nature **310:** 381–386.

QUIOCHO, F. A., N. K. VYAS, J. S. SACK and M. N. VYAS, 1987 Atomic protein structures reveal basic features of binding of sugars and ionic substrates and calcium ions. Cold Spring Harbor Symp. Quant. Biol. **52:** 453–463.

ROLSETH, S. J., V. A. FRIED and B. G. HALL, 1980 A mutant ebg enzyme that converts lactose into an inducer of the *lac* operon. J. Bacteriol. **142:** 1036–1039.

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.

SARIS, C. F., N. K. VYAS, F. A. QUIOCHO and S. K. MATTHEWS, 1984 Predicted structure of the sugar binding site of the *lac* repressor. Nature **310:** 429–430.

SELANDER, R. K., D. A. CAUGANT and T. S. WHITTAM, 1987 Genetic structure and variation in natural populations of *Escherichia coli*, pp. 1625–1648 in *Escherichia coli and Salmonella typhimurium*, edited by F. C. NEIDHARDT. American Society for Microbiology, Washington, D.C.

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.

STOKES, H. W., and B. G. HALL, 1985 Sequence of the *ebgR* gene of *Escherichia coli*: evidence that the EBG and LAC operons are descended from a common ancestor. Mol. Biol. Evol. **2:** 478–483.

VYAS, N. K., M. N. VYAS and F. A. QUIOCHO, 1988 Sugar and signal transducer binding sites of the *Escherichia coli* galactose chemoreceptor protein. Science **242:** 1290–1295.

Communicating editor: J. R. ROTH