Sequence of the *ebgA* Gene of *Escherichia coli:* Comparison with the *lacZ* Gene[1]

Harold W. Stokes,[2] Paul W. Betts, and Barry G. Hall
Molecular and Cell Biology Department, The University of Connecticut

We have sequenced the *ebgA* (evolved β-galactosidase) gene of *Escherichia coli* K12. The sequence shows 50% nucleotide identity with the *E. coli lacZ* gene, demonstrating that the two genes are related by descent from a common ancestral gene. Comparison of the two sequences suggests that the *ebgA* gene has recently been under selection. A significant excess of identical, rather than synonymous, codons used to encode identical amino acids at the same positions in the aligned sequences implies that some form of selection is operating directly at the DNA level. This selection is independent of, and in addition to, selection based on codon usage or on function of the gene products.

Introduction

The EBG (evolved β-galactosidase) system of *Escherichia coli* has been used as a model for the detailed study of acquisitive evolution via changes in the catalytic properties of an enzyme (Hall 1983).

Wild-type EBG β-galactosidase, encoded by the *ebgA* gene, is an ineffective lactase. It has a K_m of 150 mM lactose, and a V_{max} of 620 nmoles hydrolyzed/minute/mg of enzyme, which is an insufficient level of activity to permit growth on lactose even when the protein is synthesized constitutively as 5% of the soluble protein of the cell (Hall and Hartl 1975; Hall 1981). Two classes of single point mutations (Hall 1977) dramatically improve the activity of ebg enzyme toward lactose (Hall 1976; Hall 1981). When both of these mutations are present in the same gene, the double-mutant enzyme hydrolyzes lactose considerably more efficiently than does either single-mutant class (Hall 1981). For example, the double-mutant enzyme encoded by the *ebgA205* allele has a K_m for lactose of 0.93 mM, and a V_{max} of 2,000 nmol/min/mg, i.e., V_{max}/K_m is improved 537-fold relative to the wild-type enzyme (Hall 1981). This may be compared with the K_m of 2.5 mM lactose, and a V_{max} of 32,600 nmol/min/mg exhibited by the *lacZ* β-galactosidase (Huber, Kurz, and Wallenfels 1976). The wild-type ebg enzyme does not detectably convert lactose to allolactose, whereas the double-mutant enzyme converts lactose to allolactose at ~10% of the rate at which it hydrolyzes lactose (Hall 1982). Thus, as a result of two point mutations, ebg enzyme is able to replace the *lacZ* β-galactosidase both with respect to lactose hydrolysis and induction of the lac permease gene and consequently is able to permit growth on lactose as a sole carbon and energy source. Because the *ebgA* gene can evolve to replace the function

1. Key words: *ebgA* and *lacZ* genes, nucleotide identity, identical codons.

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):469–477. 1985.

0737-4038/85/0206-0866$02.00

of the *lacZ* gene, it is of interest to determine the relationship between these two genes. We have therefore compared the sequences of these two genes.

Material and Methods

Isolation and Manipulation of Plasmid DNA

The construction and preparation of plasmid pUF2, containing the cloned *ebgA* gene, has been previously described (Stokes and Hall 1984). Purified restriction fragments of cloned DNA were prepared by electroelution from agarose gels (Maniatis et al. 1982). The DNA was then passed over a BRL NACS-52 prepac column, ethanol precipitated, resuspended in appropriate buffer, and ligated into M13 vectors MP8, MP9, MP18, and MP19 as described by Sanger et al. (1980). *Escherichia coli* strain JM101 (Sanger et al. 1980) was the host for transformation.

Construction of Recombinant M13 Clones with Bal 31 Nuclease

The appropriate restriction fragments were isolated and suspended in 300 μl of Bal 31 buffer (20 mM Tris-HCl, pH 8.1, 12 mM $MgCl_2$, 12 mM $CaCl_2$, 60 mM NaCl, 1 mM EDTA). One-half unit of Bal 31 was added, and the DNA was digested for 1–3 min. Reactions were stopped with 25 mM EDTA. Digested DNA was end repaired with Klenow fragment (Maniatis et al. 1982), and fragments were cloned into *Sma*I-cut M13 vectors.

Nucleic Acid Sequencing

DNA sequence was determined essentially by the dideoxy method of Biggin et al. (1983), except that ^{32}P was used. To resolve an ambiguity in the sequence, bp 1,793–1,894 were also sequenced, by the method of Maxam and Gilbert (1977).

Results

Determination of the *ebgA* Gene Sequence

Restriction fragments of the segment of plasmid pUF2 that were previously shown to encompass the *ebgA* gene (Stokes and Hall 1984) were subcloned into appropriate M13 vectors. The positions and orientations of a subset of those sequenced fragments that define a contiguous *ebgA* sequence are shown in figure 1. A number of other

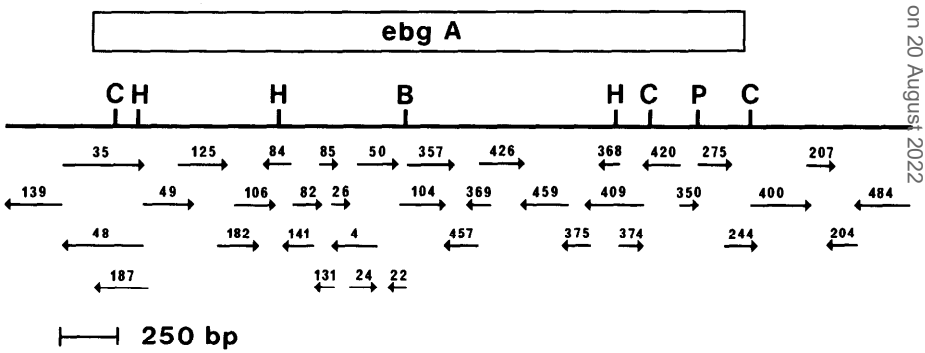

FIG. 1.—Restriction map and sequencing strategy for *ebgA*. The limits of *ebgA* are defined by the boxed region. Letters refer to restriction sites: B = *BamHI*, C = *Cla*I, H = *Hinc*II, and P = *Pst*I. Arrowed lines refer to a set of sequenced overlapping clones that define a contiguous sequence. The arrowheads indicate the strand sequenced, and the numbers are clone designations.

clones (not shown in fig. 1) were sequenced, and the majority of the sequence was confirmed by sequencing both strands with overlapping fragments.

Figure 2 shows the DNA sequence of *ebgA* and the deduced amino acid sequence. The correct reading frame was initially established by comparison with the sequences of two ebg peptides that were labeled by an active site–labeling reagent (Fowler and Smith 1983). In addition, there was only one open reading frame of sufficient length to account for the coding region of *ebgA* (~3 kb) as previously established (Stokes and Hall 1984). Table 1 shows the alignment of the *ebgA* gene with the *lacZ* gene (Kalnins et al. 1983). The two sequences were aligned by eye, with the aid of the Cornell DNA sequence analysis program (Fristensky, Lis, and Wu 1982), which was used to locate regions of significant similarity. Because the sequences were aligned by eye rather than by an algorithm that "optimizes" identity, the percentage identity shown in table 1 is a minimum estimate. The majority of the gaps used to align the two sequences were in multiples of three bases; however, two of the segments are "out of frame" with respect to each other. The alignment analyzed has 50% DNA sequence identity over 2,665 bp, and 33% amino acid identity over 850 residues.

The sequences of the two ebg peptides identified by Fowler and Smith (1983) as active-site regions differ slightly from the deduced sequences reported here. The only significant difference is that Fowler and Smith reported the labeling of a serine in *ebgA* corresponding to methionine at position 502 of *lacZ,* whereas we identify that residue as a methionine. It is unlikely that this difference arises from allelic differences, since we have sequenced this region on both strands of two additional alleles of *ebgA* (data not shown), and in each case the deduced amino acid at that position was a methionine, not a serine.

We have also sequenced the region between the termination of the adjacent *ebgR* gene and the beginning of the *ebgA* coding sequence (fig. 3). BP 91–148 of that region are 47% identical to bp 13–70 of the region between *lacI* and *lacZ* (Dickson et al. 1975), bp 149–225 are 51% identical to bp 1–77 of *lacZ,* and bp 226–269 are 52% identical to bp 149–192 of *lacZ.*

Discussion

The similarity between the *ebgA* gene and the *lacZ* gene of *Escherichia coli* leads to the conclusion that the two genes are descended from a common ancestral gene, i.e., that they are homologous. This homology is also apparent at the level of the amino acid sequences. If the *ebgA* gene were not under selection, it would be expected that the substitutions that led to 50% sequence divergence would be randomly distributed. In that case it would be unlikely that lactase activity could be restored by only a few mutations. The observation that only two mutations are required to increase the activity of wild-type EBG enzyme to the point where it approaches that of *lacZ* β-galactosidase therefore implies that the *ebgA* gene is, or has recently been, under selection.

The sequence identity at the DNA level (50%) exceeds that at the amino acid level (33%). This is consistent with the general observation (Riley 1984) that duplicated genes in the same genome exhibit greater nucleotide than amino acid identity. This is in contrast to the observation that homologous genes in different bacterial species exhibit greater amino acid than nucleotide similarity, and that duplicated genes within a genome tend to use identical, rather than synonymous, codons. Riley (1984) has suggested that this is because (1) duplication within a genome is more recent than the time of divergence of related bacterial genera and (2) there has been insufficient time for duplications within genomes to achieve equilibrium value for use of synonymous

```
                12                      24                      36                      48                      60                      72
ATG GCT GAC TGG GGG CAT ATT ACC GTC CCC GCC ATG TGG CAA ATG GAA GGT CAC GGC AAA CTG CAA TAT ACC
MET Ala Asp Trp Gly His Ile Thr Val Pro Ala MET Trp Gln MET Glu Gly His Gly Lys Leu Gln Tyr Thr

                84                      96                     108                     120                     132                     144
GAC GAA GGT TTT CCG TTC CCC ATC GAT GTG CCG TTT GTC CCC AGC GAT AAC CCA ACC GGT GCC TAT CAA CGT
Asp Glu Gly Phe Pro Phe Pro Ile Asp Val Pro Phe Val Pro Ser Asp Asn Pro Thr Gly Ala Tyr Gln Arg

               156                     168                     180                     192                     204                     216
ATT TTC ACC CTC AGC GAC GGC TGG CAG GGT AAA CAG ACG CTG ATT AAA TTT GAC GGC GTC GAA ACC TAT TTT
Ile Phe Thr Leu Ser Asp Gly Trp Gln Gly Lys Gln Thr Leu Ile Lys Phe Asp Gly Val Glu Thr Tyr Phe

               228                     240                     252                     264                     276                     288
GAA GTC TAT GTT AAC GGT CAG TAT GTG GGT TTC AGC AAG GGC AGT CGC CTG ACC GCA GAG TTT GAC ATC AGC
Glu Val Tyr Val Asn Gly Gln Tyr Val Gly Phe Ser Lys Gly Ser Arg Leu Thr Ala Glu Phe Asp Ile Ser

               300                     312                     324                     336                     348                     360
GCG ATG GTT AAA ACC GGC GAC AAC CTG TTG TGT GTG CGC GTG ATG CAG TGG GCG GAC TCT ACC TAC GTG GAA
Ala MET Val Lys Thr Gly Asp Asn Leu Leu Cys Val Arg Val MET Gln Trp Ala Asp Ser Thr Tyr Val Glu

               372                     384                     396                     408                     420                     432
GAC CAG GAT ATG TGG TGG TCA GCG GGG ATC TTC CGC GAT GTT TAT CTG GTC GGA AAA CAC CTA ACG CAT ATT
Asp Gln Asp MET Trp Trp Ser Ala Gly Ile Phe Arg Asp Val Tyr Leu Val Gly Lys His Leu Thr His Ile

               444                     456                     468                     480                     492                     504
AAC GAT TTC ACT GTG CGT ACC GAC TTT GAC GAA GCC TAT TGC GAT GCC ACG CTT TCC TGC GAA GTG GTG CTG
Asn Asp Phe Thr Val Arg Thr Asp Phe Asp Glu Ala Tyr Cys Asp Ala Thr Leu Ser Cys Glu Val Val Leu

               516                     528                     540                     552                     564                     576
GAA AAT CTC GCC GCC TCC CCT GTC GTC ACG ACG CTG GAA TAT ACC CTG TTT GAT GGC GAA CGC GTG GTG CAC
Glu Asn Leu Ala Ala Ser Pro Val Val Thr Thr Leu Glu Tyr Thr Leu Phe Asp Gly Glu Arg Val Val His

               588                     600                     612                     624                     636                     648
AGC AGC GCC ATT GAT CAT TTG GCA ATT GAA AAA CTG ACC AGC GCC ACG TTT GCT TTT ACT GTC GAA CAG CCG
Ser Ser Ala Ile Asp His Leu Ala Ile Glu Lys Leu Thr Ser Ala Thr Phe Ala Phe Thr Val Glu Gln Pro

               660                     672                     684                     696                     708                     720
CAG CAA TGG TCA GCA GAA TCC CCT TAT CTT TAC CAT CTG GTC ATG ACG CTG AAA GAC GCC AAC GGC AAC GTT
Gln Gln Trp Ser Ala Glu Ser Pro Tyr Leu Tyr His Leu Val MET Thr Leu Lys Asp Ala Asn Gly Asn Val

               732                     744                     756                     768                     780                     792
CTG GAA GTG GTG CCA CAA CGC GTT GGC TTC CGT GAT ATC AAA GTG CGC GAC GGT CTG TTC TGG ATC AAT AAC
Leu Glu Val Val Pro Gln Arg Val Gly Phe Arg Asp Ile Lys Val Arg Asp Gly Leu Phe Trp Ile Asn Asn

               804                     816                     828                     840                     852                     864
CGT TAT GTG ATG CTG CAC GGC GTC AAC CGT CAC GAC AAC GAT CAT CGC AAA GGC CGC GCC GTT GGA ATG GAT
Arg Tyr Val MET Leu His Gly Val Asn Arg His Asp Asn Asp His Arg Lys Gly Arg Ala Val Gly MET Asp

               876                     888                     900                     912                     924                     936
CGC GTC GAG AAA GAT CTC CAG TTG ATG AAG CAG CAC AAT ATC AAC TCC GTG CGT ACC GCT CAC TAC CCG AAC
Arg Val Glu Lys Asp Leu Gln Leu MET Lys Gln His Asn Ile Asn Ser Val Arg Thr Ala His Tyr Pro Asn

               948                     960                     972                     984                     996                    1008
GAT CCG CGT TTT TAC GAA CTG TGT GAT ATC TAC GGC CTG TTT GTG ATG GCG GAA ACC GAC GTC GAA TCG CAC
Asp Pro Arg Phe Tyr Glu Leu Cys Asp Ile Tyr Gly Leu Phe Val MET Ala Glu Thr Asp Val Glu Ser His
```

FIG. 2.—Sequence of *ebgA* and deduced amino acid sequence. The peptide portions corresponding to the active-site peptides of Fowler and Smith (1983) are underlined.

codons. We must demur from this position. Orthologous genes in different organisms—i.e., homologous genes that diverged following speciation—are expected to exhibit high amino acid identity arising from selection acting at the amino acid level to preserve identical functions. Thus, amino acid similarity is expected to be higher than nucleotide similarity because the degeneracy of the genetic code permits silent substitutions, predominantly at third positions of the codons. In contrast, paralogous genes—i.e., homologous genes that diverged following gene duplication within the same genome—are expected to diverge more rapidly at the amino acid level as replacements are required for divergence of function. Codon usage patterns, which may depend on the relative abundance of synonymous tRNAs within a species, impose a constraint on

```
                        1020                                1032                                1044                                1056                                1068                                1080
GGC TTT GCT AAT GTC GGC GAT ATT AGC CGT ATT ACC GAC GAT CCG CAG TGG GAA AAA GTC TAC GTC GAG CGC
Gly Phe Ala Asn Val Gly Asp Ile Ser Arg Ile Thr Asp Asp Pro Gln Trp Glu Lys Val Tyr Val Glu Arg

            1092        1104        1116        1128        1140        1152
ATT GTT CGC CAT ATC CAC GCG CAG AAA AAC CAT CCG TCG ATC ATC ATC TGG TCG CTG GGC AAT GAA TCC GGC
Ile Val Arg His Ile His Ala Gln Lys Asn His Pro Ser Ile Ile Ile Trp Ser Leu Gly Asn Glu Ser Gly

            1164        1176        1188        1200        1212        1224
TAT GGC TGT AAC ATC CGC GCG ATG TAC CAT GCG GCG AAA CGG CTG GAT GAC ACG CGA CTG GTG CAT TAC GAA
Tyr Gly Cys Asn Ile Arg Ala MET Tyr His Ala Ala Lys Arg Leu Asp Asp Thr Arg Leu Val His Tyr Glu

            1236        1248        1260        1272        1284        1296
GAA GAT CGC GAT GCT GAA GTG GTC GAT ATT ATT TCC ACC ATG TAC ACC CGC GTG CCG CTG ATG AAT GAG TTT
Glu Asp Arg Asp Ala Glu Val Val Asp Ile Ile Ser Thr MET Tyr Thr Arg Val Pro Leu MET Asn Glu Phe

            1308        1320        1332        1344        1356        1368
GGT GAA TAC CCG CAT CCG AAG CCG CGC ATC ATC TGT GAA TAT GCT CAT GCG ATG GGG AAC GGA CCG GGC GGG
Gly Glu Tyr Pro His Pro Lys Pro Arg Ile Ile Cys Glu Tyr Ala His Ala MET Gly Asn Gly Pro Gly Gly

            1380        1392        1404        1416        1428        1440
CTG ACG GAG TAC CAG AAC GTC TTC TAT AAG CAC GAT TGC ATT CAG GGT CAT TAT GTC TGG GAG TGG TGC GAC
Leu Thr Glu Tyr Gln Asn Val Phe Tyr Lys His Asp Cys Ile Gln Gly His Tyr Val Trp Glu Trp Cys Asp

            1452        1464        1476        1488        1500        1512
CAC GGG ATC CAG GCA CAG GAC GAC CAC GGC AAT GTC TGG TAT AAA TTC GGC GGC GAC TAC GGC GAC TAT CCC
His Gly Ile Gln Ala Gln Asp Asp His Gly Asn Val Trp Tyr Lys Phe Gly Gly Asp Tyr Gly Asp Tyr Pro

            1524        1536        1548        1560        1572        1584
AAC AAC TAT AAC TTC TGT CTT GAT GGT TTG ATC TAT TCC GAT CAG ACG CCG GGA CCG GGC CTG AAA GAG TAC
Asn Asn Tyr Asn Phe Cys Leu Asp Gly Leu Ile Tyr Ser Asp Gln Thr Pro Gly Pro Gly Leu Lys Glu Tyr

            1596        1608        1620        1632        1644        1656
AAA CAG GTT ATC GCG CCG GTA AAA ATC CAC GCG CGG GAT CTG ACT CGC GGC GAG TTG AAA GTC GAA AAT AAA
Lys Gln Val Ile Ala Pro Val Lys Ile His Ala Arg Asp Leu Thr Arg Gly Glu Leu Lys Val Glu Asn Lys

            1668        1680        1692        1704        1716        1728
CTG TGG TTT ACC ACG CTT GAT GAC TAC ACC CTG CAC GCA GAG GTG CGC GCC GAA GGT GAA AGC CTC GCG ACG
Leu Trp Phe Thr Thr Leu Asp Asp Tyr Thr Leu His Ala Glu Val Arg Ala Glu Gly Glu Ser Leu Ala Thr

            1740        1752        1764        1776        1788        1800
CAG CAG ATT AAA CTG CCG GAC GTT GCG CCG AAC AGC GAA GCC CCC TTG CAG ATC ACG TGC CGC AGC TGG ACG
Gln Gln Ile Lys Leu Pro Asp Val Ala Pro Asn Ser Glu Ala Pro Leu Gln Ile Thr Cys Arg Ser Trp Thr

            1812        1824        1836        1848        1860        1872
CCC GCG AAG CGT TCC CTC AAC ATT ACG GTG ACC AAA GAT TCC CGC ACC CGC TAC AGC GAA GCC GGA CAC CCT
Pro Ala Lys Arg Ser Leu Asn Ile Thr Val Thr Lys Asp Ser Arg Thr Arg Tyr Ser Glu Ala Gly His Pro

            1884        1896        1908        1920        1932        1944
ATC GCC ACT TAT CAG TTC CCG CTG AAG GAA AAC ACC GCG CAG CCA GTG CCT TTC GCA CCA AAT AAA TGC GCG
Ile Ala Thr Tyr Gln Phe Pro Leu Lys Glu Asn Thr Ala Gln Pro Val Pro Phe Ala Pro Asn Lys Cys Ala

            1956        1968        1980        1992        2004        2016
TCC GTG ACG CTG GAA GAC GAT CGT TTG AGC TGC ACC GTT CGC GGC TAC AAC TTC GCG ATC ACC TTC TCA AAA
Ser Val Thr Leu Glu Asp Asp Arg Leu Ser Cys Thr Val Arg Gly Tyr Asn Phe Ala Ile Thr Phe Ser Lys
```

FIG. 2 (*Continued*)

the use of synonymous as opposed to identical codons. Thus, within a species, codon usage alone may be expected to conserve nucleotide identity to a degree greater than it does amino acid identity even over very long periods of time. We suggest (Stokes and Hall 1985, in this issue) that *ebgA* and *lacZ* diverged following genome, rather than simple gene, duplication. Given the similarity of the *E. coli* and *Salmonella typhimurium* genetic maps (Riley and Anilionis 1978), this implies that the duplication event was very ancient and preceded the divergence of *E. coli* and *S. typhimurium*. Thus there should have been sufficient time for equilibrium to have been reached.

As is the case for other duplicated genes in *E. coli* (Riley 1984), the use of identical codons by *lacZ* and *ebgA* (151 identical codons) exceeds the use of synonymous codons (118 synonymous codons) (these values exclude methionine and tryptophan, which

```
          2028                2040                2052                2064                2076                2088
ATG AGT GGC AAA CCG ACA TCC TGG CAG GTG AAT GGC GAA TCG CTG CTG ACT CGC GAG CCA AAG ATC AAC TTC
MET Ser Gly Lys Pro Thr Ser Trp Gln Val Asn Gly Glu Ser Leu Leu Thr Arg Glu Pro Lys Ile Asn Phe

          2100                2112                2124                2136                2148                2160
TTC AAG CCG ATG ATG ATC GAC AAC CAC AAG CAG GAG TAC GAA GGG CTG TGG CAA CCG AAT CAT TTG CAG ATC
Phe Lys Pro MET MET Ile Asp Asn His Lys Gln Glu Tyr Glu Gly Leu Trp Gln Pro Asn His Leu Gln Ile

          2172                2184                2196                2208                2220                2232
ATG CAG GAA CAT CTG CGC GAC TTT GCC GTA GAA CAG AGC GAT GGT GAA GTG CTG ATC ATC AGC CGC ACA GTT
MET Gln Glu His Leu Arg Asp Phe Ala Val Glu Gln Ser Asp Gly Glu Val Leu Ile Ile Ser Arg Thr Val

          2244                2256                2268                2280                2292                2304
ATT GCC CCG CCG GTG TTT GAC TTC GGG ATG CGC TGC ACC TAC ATC TGG CGC ATC GCT GCC GAT GGC CAG GTT
Ile Ala Pro Pro Val Phe Asp Phe Gly MET Arg Cys Thr Tyr Ile Trp Arg Ile Ala Ala Asp Gly Gln Val

          2316                2328                2340                2352                2364                2376
AAC GTG GCG CTT TCC GGC GAG CGT TAC GGC GAC TAT CCG CAC ATC ATT CCG TGC ATC GGT TTC ACC ATG GGA
Asn Val Ala Leu Ser Gly Glu Arg Tyr Gly Asp Tyr Pro His Ile Ile Pro Cys Ile Gly Phe Thr MET Gly

          2388                2400                2412                2424                2436                2448
ATT AAC GGC GAA TAC GAT CAG GTG GCG TAT TAC GGT CGT GGA CCG GGC GAA AAC TAC GCC GAC AGC CAG CAG
Ile Asn Gly Glu Tyr Asp Gln Val Ala Tyr Tyr Gly Arg Gly Pro Gly Glu Asn Tyr Ala Asp Ser Gln Gln

          2460                2472                2484                2496                2508                2520
GCT AAC ATC ATC GAT ATC TGG CGC CAA GCC GTC GAT GCC ATG TTC GAG AAC TAT CCC TTC CCG CAG AAC AAC
Ala Asn Ile Ile Asp Ile Trp Arg Gln Ala Val Asp Ala MET Phe Glu Asn Tyr Pro Phe Pro Gln Asn Asn

          2532                2544                2556                2568                2580                2592
GGT AAC CGT CAG CAT GTC CGC TGG ACG GCA CTG ACT AAC CGC CAC GGT AAC GGT CTG CTG GTG GTT CCG CAG
Gly Asn Arg Gln His Val Arg Trp Thr Ala Leu Thr Asn Arg His Gly Asn Gly Leu Leu Val Val Pro Gln

          2604                2616                2628                2640                2652                2664
CGC CCA ATT AAC TTC AGC GCC TGG CAC TAT ACC CAG GAA AAC ATC CAC GCT GCC CAG CAC TGT AAC GAG CTG
Arg Pro Ile Asn Phe Ser Ala Trp His Tyr Thr Gln Glu Asn Ile His Ala Ala Gln His Cys Asn Glu Leu

          2676                2688                2700                2712                2724                2736
CAG CGC AGT GAT GAC ATC ACC CTA GGA ACC TCG ATC ACC AGC TGC TTG GCC TCG GCT CCA ACT CCT GGG GCA
Gln Arg Ser Asp Asp Ile Thr Leu Gly Thr Ser Ile Thr Ser Cys Leu Ala Ser Ala Pro Thr Pro Gly Ala

          2748                2760                2772                2784                2796                2808
GCG AGG TGC TGG ACT CCT GGC GCG TCT GGT TCC GTG ACT TCA GCT ACG GCT TTA CGT TGC TGC CGG TTT CTG
Ala Arg Cys Trp Thr Pro Gly Ala Ser Gly Ser Val Thr Ser Ala Thr Ala Leu Arg Cys Cys Arg Phe Leu

          2820                2832                2844                2856                2868                2880
GCG GAG AAG CTA CCG CGC AAA GCC TGG CGT CGT ATG AGT TCG GCG CAG GGT TCT TTT CCA CGA ATT TGC ACA
Ala Glu Lys Leu Pro Arg Lys Ala Trp Arg Arg MET Ser Ser Ala Gln Gly Ser Phe Pro Arg Ile Cys Thr

          2892
CGG AGA ATA AGC TAA
Arg Arg Ile Ser
```

FIG. 2 (*Continued*)

use only a single codon each). Codon usage by *ebgA* (data not shown) exhibits no significant differences from codon usage by *lacZ* (Kalnins et al. 1983). We therefore pooled the two codon usage tables to generate an "average" codon usage for the two sequences, from which we calculated for each amino acid the probability of identical codons being used where the same amino acid occurred at a given position (table 2). Table 2 shows a χ^2-test of the hypothesis that the proportion of identical codons used is due to chance alone. The value of χ^2 calculated in table 2 is significant at the 2.5% level. Since only identical amino acids were considered, it is not possible that selection at the protein level can account for the observed excess of identical codons. Since codon usage was taken into account in this calculation and since there is only a 2.5% likelihood that chance alone accounts for the excess, it is likely that selection is operating directly at the DNA level. By this we mean that some form of selection seems to be operating on the DNA sequence itself, rather than on the products of the DNA sequence. It would therefore appear that there are at least two constraints that would

Table 1
Alignment of *ebgA* and *lacZ* Sequences

Segment No.	Base Pairs Aligned		% DNA Identity	No. of Identical Amino Acids	No. of Identical Codons
	ebgA	*lacZ*			
1	1–168	232–399	54.2	21	10
2	169–546	403–780	52.1	54	29
3	547–1,026	793–1,272	54.2	56	35
4	1,036–1,248	1,273–1,485	55.4	32	21
5	1,249–1,320	1,489–1,560	47.2	9	8
6	1,321–1,602	1,597–1,878	55.3	41	23
7	1,612–1,716	1,879–1,983	38.1	8	5
8	1,717–1,758	2,020–2,061	57.1	4	3
9	1,759–1,800	2,065–2,106	40.8	4	2
10	1,837–1,944	2,107–2,214	47.2	13	9
11	1,945–1,983	2,242–2,280	46.2	2	2
12	1,984–2,054	2,327–2,397	47.9	...	...
13	2,068–2,151	2,398–2,481	41.7	5	2
14	2,191–2,280	2,482–2,571	41.1	3	2
15	2,290–2,388	2,572–2,670	34.3	4	3
16	2,389–2,511	2,674–2,796	48.0	15	8
17	2,518–2,571	2,797–2,850	46.3	2	1
18	2,587–2,646	2,851–2,910	43.3	3	1
19	2,650–2,685	2,911–2,946	50.0	2	1
20	2,686–2,729	2,951–2,994	45.5	...	...
21	2,794–2,841	2,995–3,042	50.0	2	1
22	2,845–2,871	3,043–3,069	40.7	0	0
23	2,872–2,892	Terminated			

NOTE.—All gaps are in multiples of three except those surrounding segments 12 and 20. Those segments are out of reading frame with respect to each other; thus, identical amino acids and codons are not considered.

maintain greater DNA than amino acid identity as duplicated genes diverge. The first is codon usage, and the second is selection at the DNA level. Although the basis of selection at the DNA level is certainly not clear, Lipman and Wilbur (1983) have

```
        10         20         30         40         50         60         70
ATCCCCTTAC ACACTGTCCG GCAATCGTTT TTGCCGGACA GTGCTGCCGT TTATTTTCGT GATCCAGTTA

        80         90        100        110        120        130        140
AAGTAAATGC ATTTACCTGC TACTTTTTAG TAAAAATTTT ACTAAACTCC CCAGCAATTA CACAAACTAC

       150        160        170        180        190        200        210
CATCACCATG AATGGTTCCG ATTTCTCTCT ACCGGGAGGC CCTATGAACT CGCTGGGAAA ACATTCAGCT

       220        230        240        250        260        270        280
CACCCACGAA AACCGACTTG CGCCGCGTGC GTACTTTTTT TCACTATGAT TCTGTTGCGC AACGGCGTAC

       290        300        310        320        330        340        350
CTTTGCCCGC GAAACCAGCA GCCTGTTTCT GCCCTTAAGC GGTCAGTGGA ATTTCCACTT TTTTGACCAT

       360        370        380
CCGCTGCAAG TACCAGAAGC CTTCACCTCT GAGTTA
```

FIG. 3.—Sequence of the region immediately preceding *ebgA*. This sequence begins immediately after the stop codon of *ebgR* (Stokes and Hall 1985, in this issue). The segments showing similarity to the 5' region of *lacZ* are underlined once, and the segment showing similarity to the control region preceding *lacZ* is underlined twice. A possible Pribnow box is enclosed in a box.

Table 2
Use of Identical and Synonymous Codons in the *ebgA* and *lacZ* Sequences

	Identical Codons			Synonymous Codons		
Amino Acid	Observed	Expected	χ^2 [a]	Observed	Expected	χ^2 [a]
Ala	7	4.114	2.025	7	9.856	0.828
Arg	7	5.656	0.319	7	8.344	0.216
Asn	11	9.911	0.120	6	7.089	0.167
Asp	11	10.300	0.048	9	9.700	0.051
Cys	1	1.728	0.307	2	1.272	0.408
Gln	8	7.381	0.052	3	3.689	0.129
Glu	13	9.856	1.003	3	6.144	1.609
Gly	16	10.819	2.481	15	20.181	1.330
His	5	3.500	0.643	2	3.500	0.643
Ile	8	5.445	1.199	3	5.555	1.175
Leu	12	8.970	1.024	11	14.030	0.654
Lys	4	2.428	1.018	0	1.572	1.572
Phe	11	7.500	1.633	4	7.500	1.633
Pro	13	7.880	3.327	7	12.120	2.163
Ser	4	1.910	2.287	4	6.880	1.206
Thr	5	6.192	0.229	13	11.808	0.120
Tyr	4	7.530	1.655	11	7.470	1.668
Val	11	12.826	0.260	11	9.170	0.365

Note.—Observed = the number of identical or synonymous codons used; expected = the number of identical or synonymous codons expected on the basis of the codon usage for the two sequences. χ^2 = (observed − expected)^2/expected.
[a] The sum of the χ^2 values = 31.788, which is significant at the 2.5% level for 17 degrees of freedom.

pointed out that the choice of the degenerate third base in a codon exhibits statistical dependence on its nearest neighbors on each side. This also suggests the existence of some sort of selection operating directly on DNA sequences rather than on their products.

Acknowledgments

We are grateful to Carol Crowther for assistance with sequence interpretation and for preparation of figures and to Bob Donnelly for help with Maxam and Gilbert sequencing. We are particularly grateful to J. Lis for the gift of the Cornell DNA sequencing program. This work was supported by National Institutes of Health grant AI 14766. B.G.H. is the recipient of Research Career Development Award 1 K04 AI00366 from the National Institute of Allergy and Infectious Diseases.

LITERATURE CITED

Biggin, M. D., T. J. Gibson, and G. F. Hong. 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80**:3963–3965.

Dickson, R. C., J. Ableson, W. M. Barnes, and W. S. Reznikoff. 1975. Genetic regulation: the lac control region. Science **187**:27–35.

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

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^0) and evolved (ebg$^+$) enzymes. J. Mol. Biol. **107**:71–84.

———. 1977. Number of mutations required to evolve a new lactase function in *Escherichia coli.* J. Bacteriol. **129**:540–543.

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

———. 1982. Transgalactosylation activity of the ebg β-galactosidase synthesizes allolactose from lactose. J. Bacteriol. **150**:132–140.

———. 1983. Evolution of new metabolic functions in laboratory organisms. Pp. 234–257 *in* M. NEI and R. KOEHN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

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

HUBER, R. E., G. KURZ, and K. WALLENFELS. 1976. A quantitation of the factors which affect the hydrolase and transgalactosylase activities of β-galactosidase (*E. coli*) on lactose. Biochemistry **15**:1994–2001.

KALNINS, A., K. OTTO, U. RUTHER, and B. MULLER-HILL. 1983. Sequence of the *lacZ* gene of *Escherichia coli.* EMBO J. **2**:593–597.

LIPMAN D. J., and W. J. WILBUR. 1983. Contextual constraints on synonymous codon choice. J. Mol. Biol. **163**:363–376.

MANIATIS, T., E. F. FRITSCH, and J. SAMBROOK. 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratories, Cold Spring Harbor, New York.

MAXAM, A. M., and W. GILBERT. 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA **74**:560–564.

RILEY, M. 1984. Arrangement and rearrangement of bacterial genomes. Pp. 285–315 *in* R. P. MORTLOCK, ed. Microorganisms as model systems for studying evolution. Plenum, New York.

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

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.

———. 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 (in this issue).

WALTER M. FITCH, reviewing editor

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