Determining the Evolutionary Potential of a Gene

Barry G. Hall and Harmit S. Malik

Biology Department, University of Rochester

In addition to information for current functions, the sequence of a gene includes potential information for the evolution of new functions. The wild-type *ebgA* (evolved β-galactosidase) gene of *Escherichia coli* encodes a virtually inactive β-galactosidase, but that gene has the potential to evolve sufficient activity to replace the *lacZ* gene for growth on the β-galactoside sugars lactose and lactulose. Experimental evidence, which has suggested that the evolutionary potential of Ebg enzyme is limited to two specific amino acid replacements, is limited to examining the consequences of single base-substitutions. Thirteen β-galactosidases homologous with the Ebg β-galactosidase are widely dispersed, being found in gram-negative and gram-positive eubacteria and in a eukaryote. A comparison of Ebg β-galactosidase with those 13 β-galactosidases shows that Ebg is part of an ancient clade that diverged from the paralogous *lacZ* β-galactosidase over 2 billion years ago. Ebg differs from other members of its clade at only 2 of the 15 active-site residues, and the two mutations required for full Ebg β-galactosidase activity bring Ebg into conformity with the other members of its clade. We conclude that either these are the only acceptable amino acids at those positions, or all of the single-base-substitution replacements that must arise as intermediates on the way to other acceptable amino acids are so deleterious that they constitute a deep selective valley that has not been traversed in over 2 billion years. The evolutionary potential of Ebg is thus limited to those two replacements.

Introduction

Encoded within the genome of each organism is the information for all of the functions necessary for survival and reproduction in that organism's current environment. Also present within that genome is the potential for that organism's evolution of novel functions for success in new or alternative environments. In a trivial sense, each genome's potential is infinite, because given enough additions, deletions, rearrangements, and base substitutions, any sequence can evolve into any other sequence. In reality, however, evolution is subject to a variety of constraints that limit this potential, and understanding evolutionary processes amounts to understanding those constraints.

Evolutionary biologists are usually forced to infer historical evolutionary processes by examining the present-day outcomes of those processes. That is an unsatisfactory means of understanding a dynamic process, partly because neither the historical selective constraints nor the detailed molecular functions of ancestral states are well understood.

Experimental evolution of novel enzyme functions in microbial populations provides a powerful alternative approach to understanding evolutionary processes. The typical strategy is to apply strong selection for catabolism of a novel resource to a large population of the model microorganism, and to determine from the resulting spontaneous mutants the detailed changes that have given rise to the newly selected phenotype. Even given the enormous sizes of experimental microbial populations, often in excess of 10^{11} individuals, in vivo systems are limited to examining the outcomes that can be produced by one or two mutations. Within that limitation, the array of successful mutants defines the evolutionary potential of the selected gene for the chosen novel function.

For over 25 years, the Ebg (evolved β-galactosidase) operon of *Escherichia coli*K12 has been used as a model system to study the evolution of new metabolic functions (reviewed in Hall 1990). More recently, those genetic studies have been complemented by detailed biochemical studies of catalysis mediated by the Ebg enzyme (Hall 1981; Burton and Sinnott 1983; Hall et al. 1983; Li, Osborne, and Sinnott 1983; Elliott et al. 1992; Srinivasan et al. 1993; Calugaru, Hall, and Sinnott 1995, Srinivasan, Hall, and Sinnott 1995; Calugaru et al. 1997). The operon consists of four genes: *ebgR*, which encodes a repressor that controls expression of *ebgACB* (Hall and Hartl 1975; Hall and Clarke 1977; Hall 1978*b*; Hall, Betts, and Wootton 1989); *ebgA*, which encodes the 118-kDa α subunit of Ebg enzyme (Hall, Betts, and Wootton 1989); *ebgC*, which encodes the 20-kDa β subunit of Ebg enzyme (Hall, Betts, and Wootton 1989); and *ebgB*, whose function is unknown (Hall and Zuzel 1980), but whose sequence (YGJI ECOLI gene product, GenBank/SwissProt accession number P42590) suggests that it is a transport protein. The wild-type Ebg enzyme is such a feeble β-galactosidase that even when the operon is expressed constitutively (*ebgR*$^-$) such that Ebg enzyme constitutes 5% of the soluble cell protein, Δ*lacZ* cells cannot utilize β-galactoside sugars as sole sources of carbon and energy (Hall 1982a). Studies that have examined 154 independent spontaneous mutants, 119 UV-induced mutants, 45 *mutS*-induced mutants, 52 *mutY*-induced mutants, and 99 *mutD*-induced mutants have identified only two sites in *ebgA* where mutations occur that can enhance the activity of Ebg enzyme sufficiently to permit utilization of lactose (galactosyl-1,4-β-D-glucose) or lactulose (galactosyl-1,4-β-D-fructose) (Hall 1995; Hall, Betts, and Wootton 1989). Numbering the Ebg operon sequence as in Hall, Betts, and Wootton (1989) (GenBank accession M64441), a G_{1566}→A mutation (previously called the Class I site) results in an asp_{92} (D)→asn (N) replacement. That replacement increases the activity (k_{cat}/k_m) of

Key words: Ebg, β-galactosidase, evolutionary potential.

Address for correspondence and reprints: Barry G. Hall, Biology Department, Hutchison Hall, River Campus, University of Rochester, Rochester, New York 14627. E-mail: drbh@uhura.cc.rochester.edu.

Mol. Biol. Evol. 15(8):1055–1061. 1998

Table 1
Protein Sequences Used in this Study

Organism and Enzyme	Accession Number	Classification
Actinobacillus pleuropneumoniae β-galactosidase	1655857	Eubacteria/Proteobacteria (gamma subdivision)/Pasteurellaceae
Arthrobacter sp. β-galactosidase	2127398	Eubacteria/Firmicutes/Actinomycetes
Clostridium acetobutylicum β-galactosidase	1352076	Eubacteria/Firmicutes/low GC gram positive/Clostridiaceae
Enterobacter cloacae β-galactosidase	1091877	Eubacteria/Proteobacteria (gamma subdivision)/Enterobacteriaceae
Escherichia coli β-glucuronidase	584839	Eubacteria/Proteobacteria (gamma subdivision)/Enterobacteriaceae
Escherichia coli lacZ β-galactosidase	2623984	Eubacteria/Proteobacteria (gamma subdivision)/Enterobacteriaceae
Escherichia coli Ebg β-galactosidase	114935	Eubacteria/Proteobacteria (gamma subdivision)/Enterobacteriaceae
Homo sapiens β-glucuronidase	114963	Eukaryotae/Metazoa/Vertebrata/Primates
Klebsiella pneumoniae β-galactosidase	114941	Eubacteria/Proteobacteria (gamma subdivision)/Enterobacteriaceae
Kluyveromyces lactis β-galactosidase	399112	Eukaryotae/Fungi/Ascomycota/Saccharomycetaceae
Lactobacillus delbreuckii bulgaricus	114943	Eubacteria/Firmicutes/low GC gram positive/Lactobacillaceae
Lactobacillus sake β-galactosidase	1223762/1223763[a]	Eubacteria/Firmicutes/low GC gram positive/Lactobacillaceae
Lactococcus lactis β-galactosidase	1556406	Eubacteria/Firmicutes/low GC gram positive/Streptococcaceae
Staphylococcus xylosus β-galactosidase	2462706	Eubacteria/Firmicutes/low GC gram positive/Bacillaceae
Streptococcus salivarius thermophilus β-galactosidase	153688	Eubacteria/Firmicutes/low GC gram positive/Streptococcaceae
Thermotoga maritima β-galactosidase	473272	Eubacteria/Thermotogales

[a] The alignment utilized both the large and small subunits of the *Lactobacillus sake* β-galactosidase, which consists of two subunits as the result of splitting the gene. For purposes of this alignment, the N-terminus of the small subunit was joined to the C-terminus of the large subunit.

Ebg enzyme toward lactose 46-fold (Hall 1981) and allows the mutant to utilize lactose. However, because that mutation increases activity toward lactulose only 1.4-fold, those mutants cannot utilize lactulose. An alternate $G_{4223}\rightarrow$(T or C) mutation (previously called the Class II site) results in a trp_{977} (W)→cys (C) replacement. That substitution increases the activity of Ebg enzyme 11-fold toward lactose, increases the activity 49-fold toward lactulose (Hall 1981), and allows those mutants to utilize both lactose and lactulose.

Because all of the spontaneous and induced single-step mutants selected for growth on lactose or lactulose have one of the above mutations, it was previously suggested that the evolutionary potential of the *ebgA* gene to promote effective β-galactosidase activity is limited to those two changes (Hall 1995).

Because the experimental system is limited to investigating the consequences of single base-substitutions, it was not possible to determine whether multiple base-substitutions might yield equal or better improvements in catalytic efficiency. If such multiple changes are possible, then the limitations of the experimental system might have given a false impression of a very limited evolutionary potential for Ebg enzyme.

One approach to resolving this issue would be to use *in vitro* site-directed mutagenesis, in which multiple base-substitutions are introduced to individual codons, or in which saturation mutagenesis is applied to a series of two or three adjacent codons to generate all possible combinations of amino acid substitutions in candidate regions. The latter approach has been used very successfully to define the set of amino acid replacements that can greatly improve the activity of the TEM-1 β-lactamase (Huang et al. 1996). However, the results of site-directed mutagenesis experiments cannot provide the necessary information about ''real'' evolutionary potential. Even if some multiple mutations were found to give equal or better activity than the single mutations already identified, we would not know if those changes could evolve naturally, i.e., if they were within the natural evolutionary potential of the genome. Because multiple mutations require that the organism traverse one or more intermediate steps, evolutionary potential is limited by the fitness associated with those intermediate steps. For some possible protein changes, it may well be the case that ''you can't get there from here'' because the intervening single-step mutations are too deleterious.

Because neither in vivo nor in vitro experimental evolution reveals whether the observed evolutionary potential for β-galactosidase activity is actually limited to those two amino acid replacements, we have returned to comparisons of the outcomes of nature's 3.8-billion-year experiment with life to shed light on the issue.

Materials and Methods

The BLAST program (Altschul et al. 1990) was used to search the non-redundant protein database for sequences related to that of the EbgA protein. Fourteen β-galactosidases were identified in which the probability of a match by chance alone was $<10^{-90}$. Table 1 lists those sequences and associated accession numbers. The

```
                           42       52  74       84  88     96  101    108       130
 1 E.coli ebg              LPLSGQWNFHF  ITVPAMWQMEG  LQYTDEGFP  VPFVP  NPTGAY    IKFDGVETYFEVYVNGQYVGFSKGSRLTAEFD
 2 K.lactis lacZ           ESLNGPWAFAL  ISVPSHWELQE  PIYTNVQYP  IPNPP  NPTGVY    LRFEGVDNCYELYVNGQYVGFNKGSRNGAEFD
 3 Arthrobacter sp. lacZ   PTAPGTPGAGS  LAVPSHWVLAE  PIYTNVQYP  PPFVP  NPTGDY    LRFDGVESRYKVWVNGVEIGVGSGSRLAQEFD
 4 A.pleuropneumoniae lacZ TLLNGQWDFNY  IPVPANWQNHG  HHYTNINYP  PPFVP  NPCGVY    LNFEGVDSCLFVYVNKQFVGYSQISHNTSEFD
 5 S.xylosus lacZ          TLLNGEWYFQY  LTVPSVWNYLG  IQYLNTQYP  PPYVP  NPCGHY    LNFEGVDSAFYVWINNEFIGYSQISHAISEFD
 6 T.maritima lacZ         ISLNGNWRFLF  IEVPSNWEMKG  PIYTNVVYP  PPFVP  NPTGVY    LHFEGVRSFFYLWVNGKKIGFRQRQLHARRIQ
 7 C.acetobylicum lacZ     QNLNGKWRFSY  IEVPGHIQLQG  CQYINTMYP  PPHIS  NPVGSY    ISFQGVETAFYVWVNGEFVGYSEDTFTPSEFD
 8 S.salivarius lacZ       QSLNGKWKIHY  INVPGHLELQG  PQYVNTQYP  PPQVP  NAVASY    ISFQGVATSIFVWVNGNFVGYSEDSFTPSEFE
 9 L.delbrueckii lacZ      QSLDGDWLIDY  VKVPGNLELQG  PQYVNVQYP  PPQIP  NPLASY    LKFDGAATAIYVWLNGHFVGYGEDSFTPSEFM
10 L.sake lacZ             QSLNGTWQFHY  ITVPQHIELAG  LHYINTMYP  PAFST  NPVGSY    IRFEGVEQAMYVWLNGQFIGYAEDSFTPSEFD
11 E.coli lacZ             RSLNGEWRFAW  VVVPSNWQMHG  PIYTNVTYP  PPFVP  NPTGCY    IIFDGVNSAFHLWCNGRWVGYGQDSRLPSEFD
12 E.cloacae lacZ          QTLNGLWRFSY  MPVPSNWQMQG  PIYTNVTYP  PPFVP  NPTGCY    IIFDGVNSAFHLWCNGQWIGYSQDSRLPAEFD
13 K.pneumoniae lacZ       RQLDGSGSSLT  TPVPSNWQMEG  PIYTNVRYP  PPRVP  NPTGCY    IIFDGVNSAFHLWCNGVWVGYSQDSRLPAAFD
14 L.lactis lacZ           QSLNGLWNFDH  IIVPSNWQIEF  PIYTNVTYP  PPYVP  NPVGAY    LTFEGVGSAFHFWLNGEYGGYSEDSRLPAEFD
                               *          **           * .  .*         *    *      * *       . *    *

                                          203  286   294  315                       343  360              379
 1 ISAMVKTGDNLLCVRVMQWADSTYVEDQDMWWSAGIFRDVYL  WSAESPYLY  VGFRDIKVRDGLFWINNRYVMLHGVNRHD  DLQLMKQHNINSVRTAHYPN
 2 IQKYVSEGENLVVVKVFKWSDSTYIEDQDQWWLSGIYRDVSL  WTAENPTLY  VGFRQVELKDGNITVNGKDILFRGVNRHD  DLILMKKFNINAVRNSHYPN
 3 VSEALRPGKNLLVVRVHQWSAASYLEDQDQWWLPGIFRDVKL  WSAEVPRLY  LGFRTVKIVGDQFLVNGRKVIFHGVNRHE  DLALMKRFNVNAIRTSHYPP
 4 VSDYLQAGTNHLTVVVLKWCDGSYLEDQDKFRMSGIFRDVYL  WNAEKPQLY  IGFRKVEIKDGILLFNQQPIKFKGVNRHD  DLQLMKQHNINAIRTAHYPN
 5 ISNFVKQGENNIEVLVLKYSDGTYLEDQDMFRHSGIFRDVYI  WSTENPVLY  VGIREVAIQNNQFYINGQSIKIRGTNYHD  DLELMKQGNFNAIRTAHYPK
 6 THRCSKTREESDHVEVLKWSDGSYLEDQDMWWFAGIYRDVYL  WSAETPHLY  FGFRKIEIKDGTLLFNGKPLYIKGVNRHE  DIKLMKQHNINTVRTSHYPN
 7 ITDYLREGENKLAVEVYKRSSASWIEDQDFWRFSGIFRDVYL  WSAEEPNLY  IGFRHFEMKDKIMCLKWKRIIFKGVNRHE  DIKFLKQHNINAVRTSHYPN
 8 ISDYLVEGDNKLAVAVYRYSTASWLEDQDFWRLYGIFRDVYL  WSAESPKLY  VGFRRFEIKDKLMLLNGKRIVFKGVNRHE  DIKVMKQHNINAVRTSHYPN
 9 VTKFLKKENNRLAVALYKYSSASWLEDQDFWRMSGLFRSVTL  WSAEKPNLY  VGFRNFELKDGIMYLNGQRIVFKGANRHE  DIKTMKRSNINAVRCSHYPN
10 LTPYLKETDNCLAVEVHKRSSAAFIEDQDFFRFFGIFRDVKL  WSNQTPNLY  FGFRKIEIKDKVMLLNGKRLVINGVNRHE  DIACMQRNHINAVRTSHYPD
11 LSAFLRAGENRLAVMVLRWSDGSYLEDQDMWRMSGIFRDVSL  WSAEIPNLY  VGFREVRIENGLLLLNGKPLLIRGVNRHE  DILLMKQNNFNAVRCSHYPN
12 LSAALRPGQNRLAVMVLRWCDGSYLEDQDMWRMSGIFRDVTL  WSAETPELY  VGFRRVEISNGLLKLNGKPLLIRGVNRHE  DIETMKQHSFNAVRCSHYPN
13 LSPFLRPGDNRLCVMVMRWSAGSWLEDQDMWRMSGIFRSVWL  WSAETPNCY  IGFRRIEIADGLLRLNGKPLLIRGVNRHE  DILLMKQNNFNAVRCSHYPN
14 ISNLAKEGQNCLKVLVFRWSKGTYFEDQDMWRMSGIFRSVNL  WSDEIPYLY  IGIRKIAIEKGQLKINGKALLVRGVNKHE  DIKLMKEHNFNAVRCSHYPN
            .   * . . .  .. **** .   *..* * .  *  . *  *   * *   .         . .   * * *.  *.  ..    *..* .***
 A E.coli β-glucuronidase                                 VGIRSVAVKGEQFLINHKPFYFTGFGRHE  DHALMDWIGANSYRTSHYPY
 B Human β-glucuronidase                                  VGIRTVAVTKSQFLINGKPFYFHGVNKHE  DFNLLRWLGANAFRTSHYPY
                                                          269                       297  314              333

   384               404    435             453    464       476    484   492    506           522
 1 YELCDIYGLFVMAETDVESHG    QKNHPSIIIWSLGNESGYG    KRLDDTRLVHYEE    DIISTMYTR    KPRIICEYAHAMGNGPG
 2 YDLFDKLGFWVIDEADLETHG    DVNHPSIIIWSLGNEACYG    KQLDPTRLVHYEG    DIFSFMYPT    KPLILCEYGHAMGNGPG
 3 LDLADELGFWVILECDLETHG    DKNHASIVMWSLGNESGTG    HARDLSRPVHYEG    DVYSRMYSS    RPFILCEYVHAMGNGPG
 4 SELCDQYGFYLIGESDVESHG    DKNRTSIIIWSLGNEAGYG    KQRDKSRLVHYES    DFYSEMYGS    KPFVLCEYSHAMGNSNG
 5 YEMTDQYGFYVMSEADIETHG    LKNYSSIVSWSLGNESGFG    KSLDNTRPIHYEG    DMISRMYPS    KPFILCEYAHAMGNSPG
 6 YDLCDYFGLYVIDEANIESHG    DKNHPSIIFWSLGNEAGDG    KKRDNTRLIHYEG    DVFSLMYPK    KPFIMCEYAHAMGNSVG
 7 YRLCDEYGIYLIDETNLESHG    DKNHPSVLIWSCGNESYAG    RKKDPSRLVHYEG    RHESRMYAK    KPYISCEYMHSMGNSTG
 8 YELCDEYGLYVIDEANLETHG    DKNHASVIIWSCGNESYAG    RSVDNTRPVHYEG    DIESRMYAK    KPYISCEYMHTMGNSGG
 9 YRLCDKYGLYVIDEANLESHG    DKNHASILIWSLGNESYAG    RKADPTRVQHYEG    QIESRMYAP    KPFISVEYAHAMGNSVG
10 YNGCDQAGIYMMAETNLESHG    FKNHVSILFWSLGNESYAG    KQQDPTRLVHYEG    DVESRMYAT    KPFILCEYMHDMGNSLG
11 YTLCDRYGLYVVDEANIETHG    DRNHPSVIIWSLGNESGHG    KSVDPSRPVQYEG    DIICPMYAR    RPLILCEYAHAMGNSLG
12 YQLCDRYGLYVVDEANIETHG    DRNHPSIIIWSLGNESGHG    KTTDPTRPVQYEG    DIVCPMYAR    RPLILCEYAHAMGNSFG
13 YELCNRYGLYVVDEANIETHG    NRNHPCIIIWSLGNESGGG    KRNDPSRPVQYEG    DIICPMYAR    RPLILCEYAHAMGNSLG
14 YELCDEYGLYVMDEANIETHG    DRNHPSIIIWSLGNESGYG    KSFDSSRPVHYEG    DIICPMYAR    RPLILCEYAHDMGNSLG
       .  * ... *...*.**      *   ... ** ***.  *    .  * .*  .**        . **      .* .  ** * ***  *
 A LDWADEHGIVVIDETAAVGFN    DKNHPSVVMWSIANEPDTR    RKLDPTRPITCVN
 B MQMCDRYGIVVIDECPGVGLA    DKNHPAVVMWSVANEPASH    KSLDPSRPVTFVS
   338               358    399             417    433       445

   538     548     562                             596 869      880   901  908   970     979
 1 QGHYVWEWCDH     KFGGDYGDYPNNYNFCLDGLIYSDQTPGPGLKEYK YYGRGPGENYAD   YPFPQNNG   LGLGSN-SWGS
 2 QGGFIWEWANH     AYGGDFKEEVHDGVFIMDGLCNSEHNPTPGLVEYK WLGRGPGESYPD   YDYPQENG   HGVGSE-ACGP
 3 HGGFVWEWRDH     AYGGDFDEVIHDGNFVMDGMILSDSTPTPGLFEYK WFGAGPRESYPD   YARPQETG   HGLGSR-ACGP
 4 CGGFVWEWCDH     GYGGDFGESPHDGNFCMDGLVSPDRIPHSNLLELK YFGYGEQESYVD   YVKPQENG   SGIGSN-SCGP
 5 IGGFVWEWCDH     RYGGDFGEKLHDGNFCVDGIVFPNRVPHEGYYEFK YYGKGPFSSYQD   HIRPQETG   SGIGSN-SCGP
 6 HGGCIWDWVDQ     AYGGDFGDTPNDGNFCINGVVLPDRTPEPELYEVK WYGRGPHETYWD   YVRPQETG   MGLGGDDSWGA
 7 QGGFIWDYGDQ     AYGGDFTDRPTDYNFSGNGLIYADRTISPKAQEVK WYGMGPEENYID   ????????   ???????????
 8 QGGFIWDFIDQ     SYGGDWHDRPSDYEFCGNGIVFADRTLTPKLQTVK YYGYGAEESYRD   YLMPQESG   MGVGGDDTWGA
 9 QGGFIWDWIDQ     LYGGDFDDRPTDYEFCGNGLVFADRTESPKLANVK YYGLGPNESYPD   YLRPQETG   MGVGGDDSWGQ
10 QGGFIWDFIDQ     RYGGDFDDRPSDYEFSGDGLVFATRDEKPAMQEVR YQGLS-GETYPD   YLVPQDCG   RGVGGIDSWGS
11 QGGFVWDWVDQ     AYGGDFGDTPNDRQFCMNGLVFADRTPHPALTEAK WLGLGPQENYPD   YVFPSENG   MGIGGDDSWSP
12 QGGFVWDWVDQ     AYGGDFGDKPNDRQFCLNGLVFPDRTPHPALYEAH WLGLGPHENYPD   YIFPTENG   MGVGGDDSWSP
13 QGGFIWDWADQ     AYGGDFGDKPNDRQFCMNGLVFPDRTPHPSLVEAK WLGLGPHENYPD   YIFPTENG   MGVGGDDSWTP
14 QGGFIWDWVDQ     AYGGDFGDKPNDRQFSLNGLVFPNRQAKPALREAK YFGLGPDENYPD   YIFPSENG   MGVGGDDSWSP
    *  .*.. ..      .***. .   .  *  .*.              . . *      * *      * * *    *.*   .
```

FIG. 1.—Anchor sites used in the alignment of β-galactosidases. The alignment of the most conserved stretches among the different β-galactosidases and β-glucuronidases is shown. The numbers on top of each segment are based on the EbgA protein numbering, with the beginning and end of each anchor site indicated. Numbers at the bottom refer to the *E. coli* β-glucuronidase numbering. Previously identified active-site residues are highlighted in boldface. The *Lactobacillus sake lacZ* indicated here is a combination of the large and small subunits (table 1). "*" and "." indicate identical and similar amino acid residues, respectively.

14 sequences were initially aligned by the CLUSTAL W 1.6 program (Thompson, Higgins, and Gibson 1994) using the BLOSUM similarity matrix, a multiple-alignment gap opening penalty of 10, and a gap extension penalty of 0.05 (default parameters). That initial alignment permitted the identification of 19 strongly conserved regions that we term "anchor sites" (fig. 1). Each of the regions between those sites, together with the flanking anchor sites, were then independently aligned using a multiple-alignment gap opening penalty of 20

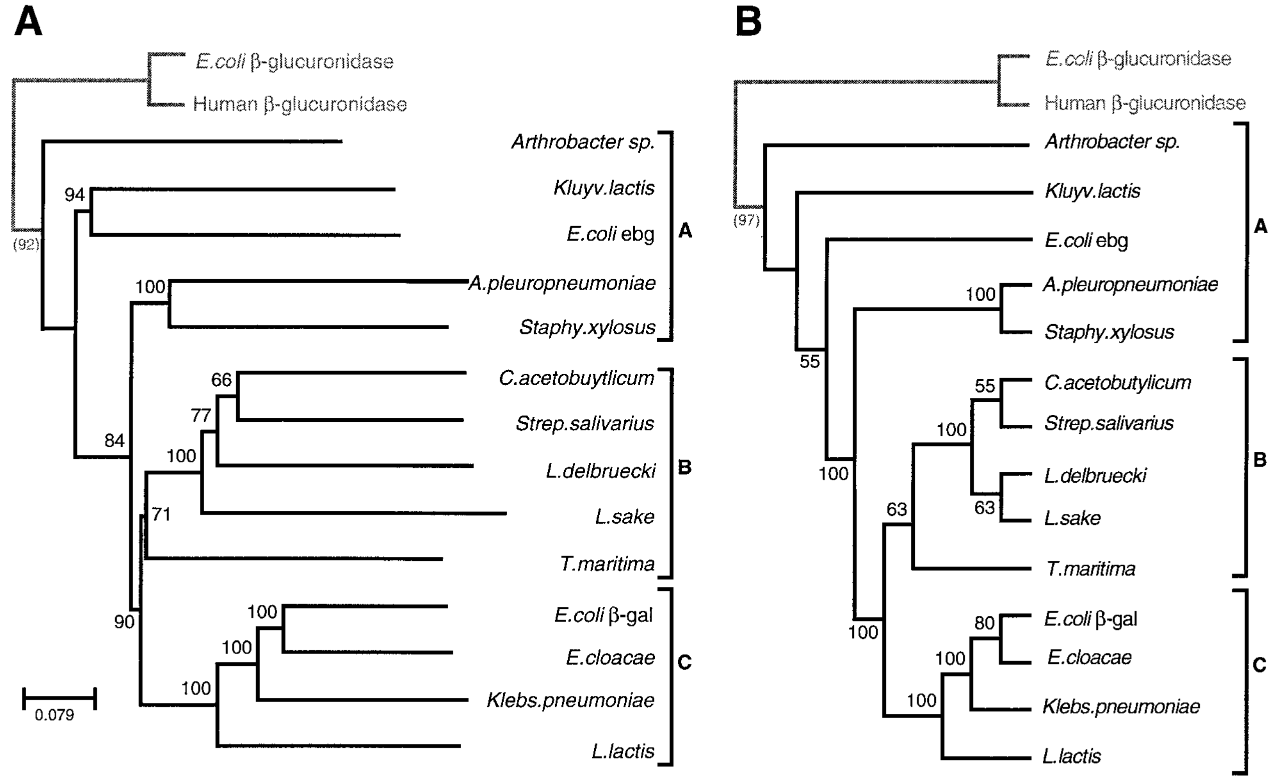


FIG. 2.—Phylogeny of the β-galactosidases. *A,* Distance tree based on the Neighbor-Joining method. Bootstrap values are indicated as percentages based on 1,000 replicates. *B,* Maximum-parsimony tree for the β-galactosidases, with bootstrap values indicated. In both *A* and *B,* the β-glucuronidases are used to conceptually root the trees as described in the text. Actual bootstrap values are based on analysis of entire β-galactosidase sequences. By both methods, the β-galactosidases split into three distinct clades. The branches of the β-glucuronidases are shown in gray to indicate that the rooting is conceptual and that the lengths of the β-glucuronidase branches are unrelated to actual distances.

and a gap extension penalty of 0.5. The objective of this approach was to ensure an optimal alignment of the intervening sequences between the conserved ''anchor sites.'' The inclusion of the anchor sites ensures that the ends of each of those regions remain aligned. Finally, the individually aligned segments were reassembled with the aligned anchor sites. The resulting final alignment has been deposited with the EMBL under accession number DS32829.

To provide an outgroup for phylogenetic tree construction a human β-glucuronidase was aligned with an *E. coli* β-glucuronidase (table 1), and the β-glucuronidase alignment was aligned to the β-galactosidase alignment using the ''profile alignment'' option of CLUSTAL W. That option retains all of the gaps within each alignment, but permits the introduction of new gaps between the alignments. Only a segment of β-glucuronidase is homologous to the β-galactosidases. Thus, to maximize phylogenetic resolution, we use that homologous segment of the β-glucuronidases only to root the tree.

The final alignment was used to construct a distance tree (fig. 2*A*) by the Neighbor-Joining method (Saitou and Nei 1987) as implemented by CLUSTAL W 1.6 with 1,000 bootstrap replications and to construct a parsimony tree (fig. 2*B*) with PAUP 3.1.1 (Swofford 1993) using the heuristic search method of branch swapping with stepwise addition of closest neighbors with 100 bootstrap replications.

Results and Discussion

By far, the most thoroughly studied of these β-galactosidases is that encoded by the *lacZ* gene of *E. coli.* That enzyme has now been crystallized, and the 15 active site residues identified by X-ray crystallography (Jacobson et al. 1994) are highlighted by boldface in figure 1. Twelve of these residues are conserved across all β-galactosidases. Phylogenetic analyses indicate that the β-galactosidases fall into three distinct clades, by both maximum-parsimony and Neighbor-Joining methods (fig. 2). Clade A includes representatives from *E. coli* (ebgA), *Arthrobacter* sp., *Actinobacillus pleuropneumoniae, Staphylococcus xylosus,* and *Kluyveromyces lactis.* Clade B includes *Thermotoga maritima, Lactobacillus sake, Lactobacillus delbrueckii, Clostridium acetobutylicum,* and *Streptococcus salivarius (subspecies thermophilus),* while clade C includes β-galactosidases from *E. coli* (lacZ), *Lactococcus lacti*s, *Klebsiella pneumoniae,* and *Enterobacter cloacae.* The observation that several amino acid replacements in active-site-containing anchor sites occurred along the branch that leads to clade BC supports the existence of clade A either as a real clade or as an arbitrary group that includes everything that is not clade BC.

Resolving the apparent trichotomy of the three clades of β-galactosidases to assign the correct rooting of the trees was achieved by utilizing the β-glucuronidases, a related family of sugar-hydrolyzing enzymes. Amino acids 269–445 of the β-glucuronidases are homologous to residues 315–476 of the β-galactosidases. The inherent assumption is of related ancestry, a premise justified by a high degree of sequence conservation. For example, four of the six active-site residues are conserved over this homologous stretch (fig. 1). Phylogenies based on that region of homology clearly show that the β-glucuronidases form an outgroup with respect to all of the β-galactosidases. The possibility that β-glucuronidases are merely a recently derived subclade of the β-galactosidases (and thereby not an appropriate outgroup) is belied by the presence of very divergent β-glucuronidases in *E. coli* and mammals. Rooting the tree conceptually on the β-glucuronidases gives us a rooted tree of the 14 β-galactosidases, shown in figure 2. While it may appear that the addition of the β-glucuronidases ''breaks up'' clade A, it should be noted that this is merely a lack of the ability to resolve clade A. The β-galactosidases emerge as an ancient clade of enzymes, with their origin well before the inception of mammals, the major extant source of lactose. Clearly, the suggestion emerges that this ancient class of enzymes may have historically been preserved to hydrolyze a β-galactoside other than lactose. The *lacZ* and *ebgA* genes of *E. coli* are paralogs that resulted from a very ancient gene duplication event. Their divergence dates back close to the root of all β-galactosidases, predates the 2.2-billion-year-old divergence of gram-positive and gram-negative *Eubacteria* (Feng, Cho, and Doolittle 1997), and may even predate the divergence of *Eubacteria* and *Eukaryota,* between 3 and 4 billion years ago (Feng, Cho, and Doolittle 1997). While a rigorous maximum-likelihood analysis might slightly modify branching orders that are weakly supported by distance and parsimony, it is both unlikely and without precedent that maximum-likelihood would give contrary results to well-supported nodes. Thus, a time-consuming and resource-intensive maximum-likelihood analysis would not alter the conclusions of this work.

Within the conserved anchor regions that include the active-site residues (see *Materials and Methods*), the sequences of the enzymes from clade A share several features that further support the notion that they belong to a the same clade (summarized in fig. 3). Thus, for the *ebgA* enzyme, cys$_{977}$ (C) is acceptable only within the context of clade A, with the trp (W)$_{977}$→cys (C) mutation bringing *ebgA* into close conformity with the ancestral clade to which it clearly belongs. The asp (D)$_{92}$→asn (N) replacement, on the other hand, brings ebg into conformity with all the other enzymes. Given that uniformity, it might well be supposed that an asparagine at that position is an absolute requirement for effective activity, but that is not the case. The Ebg trp$_{977}$→cys mutant enzyme, with sufficient activity to permit growth on lactose and lactulose, retains the aspartic acid (D) at position 92.

At position 977, most enzymes have a tryptophan; only clade A enzymes have a cysteine at that site. Ebg is an exception to the clade A enzymes at position 977; it has a tryptophan at that site. The trp$_{977}$→cys substitution, however, brings Ebg enzyme into conformity with the other members of that group.

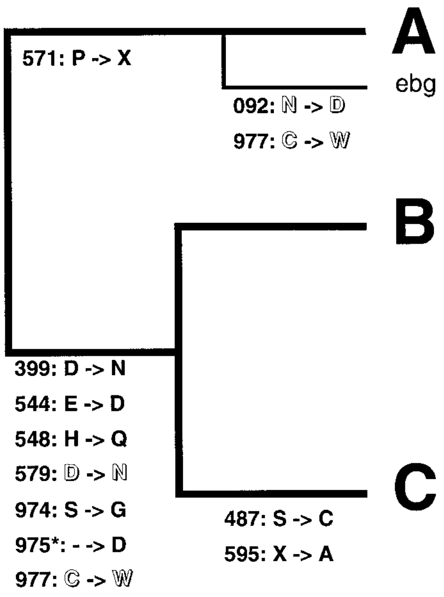
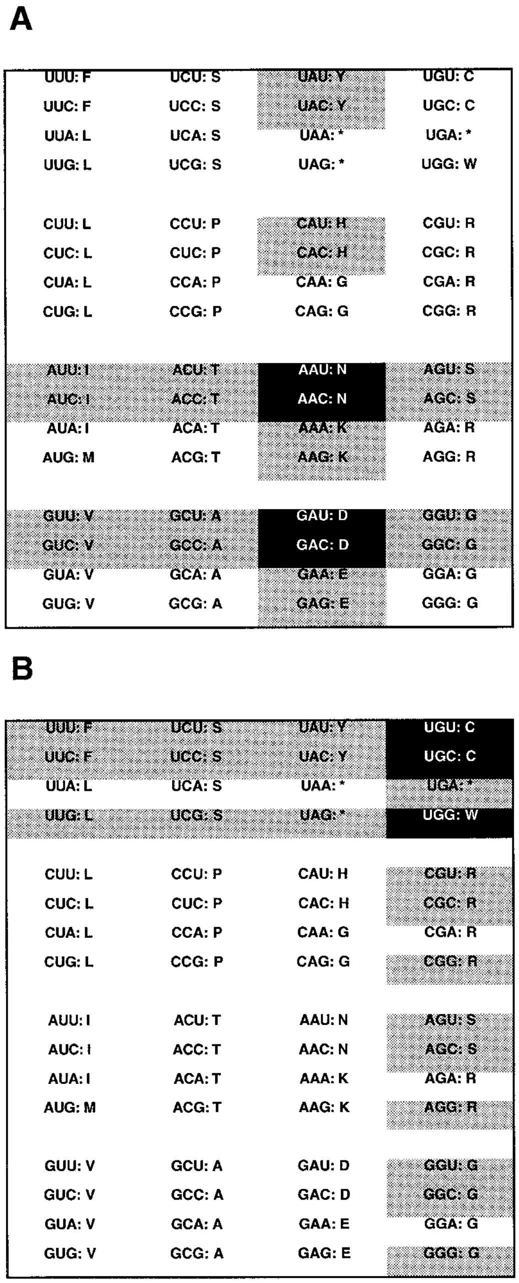


FIG. 3.—Some amino acid replacements that are common to a lineage. A schematic tree of the β-galactosidases is shown, with clade A as the ancestral clade. Amino acid replacements are within anchor sites that include active-site residues. Active-site residues are highlighted by outlining. Note that the C$_{977}$→W substitution that occurred along the branch to clade BC also occurred in Ebg. Note that position 975 is a gap in all members of clade A, while it is a conserved aspartic acid (D) in all members of clade BC.

The asp$_{92}$→asn + trp$_{977}$→cys double mutant Ebg enzyme is identical to the other clade A enzymes at all 15 active site residues. That double-mutant enzyme (designated Class IV [Hall 1978a]) differs profoundly from either of the single-mutant enzymes. Its activity toward lactose is 466 times greater than that of the wild-type enzyme, with most of that increase resulting from a dramatic drop in the k_m (Hall 1981). The double-mutant enzyme hydrolyzes galactosyl-β-D-arabinose well enough to permit growth on that sugar, while neither of the single-mutant enzymes permits such growth (Hall 1978*a*, 1981). The double mutant enzyme, unlike either of the single-mutant enzymes, converts lactose into an inducer of the *lac* operon (Rolseth, Fried, and Hall 1980; Hall 1982*b*).

Mutagenesis with UV light and with several mutator alleles of different specificity showed that the failure to find spontaneous Ebg mutations other than asp$_{92}$→asn and trp$_{977}$→cys was not the result of a mutational bias imposed by the spontaneous mutagenesis machinery of *E. coli* (Hall 1995), but that study could not determine whether other substitutions, such as asp$_{92}$(D)→gln (G), for instance, might have produced equally improved enzyme. Figure 4 shows the amino acid replacements that could result from each of the possible single-base substitutions at positions 92 and 977. The experimental evidence shows that none of those replacements allows effective lactose or lactulose hydrolysis. The phylogenetic evidence indicates either that asn$_{92}$ and cys/trp$_{977}$ are the only acceptable amino acids at those positions, or that all of the single base-substi-



A

UUU: F	UCU: S	UAU: Y	UGU: C
UUC: F	UCC: S	UAC: Y	UGC: C
UUA: L	UCA: S	UAA: *	UGA: *
UUG: L	UCG: S	UAG: *	UGG: W
CUU: L	CCU: P	CAU: H	CGU: R
CUC: L	CUC: P	CAC: H	CGC: R
CUA: L	CCA: P	CAA: G	CGA: R
CUG: L	CCG: P	CAG: G	CGG: R
AUU: I	ACU: T	AAU: N	AGU: S
AUC: I	ACC: T	AAC: N	AGC: S
AUA: I	ACA: T	AAA: K	AGA: R
AUG: M	ACG: T	AAG: K	AGG: R
GUU: V	GCU: A	GAU: D	GGU: G
GUC: V	GCC: A	GAC: D	GGC: G
GUA: V	GCA: A	GAA: E	GGA: G
GUG: V	GCG: A	GAG: E	GGG: G

B

UUU: F	UCU: S	UAU: Y	UGU: C
UUC: F	UCC: S	UAC: Y	UGC: C
UUA: L	UCA: S	UAA: *	UGA: *
UUG: L	UCG: S	UAG: *	UGG: W
CUU: L	CCU: P	CAU: H	CGU: R
CUC: L	CUC: P	CAC: H	CGC: R
CUA: L	CCA: P	CAA: G	CGA: R
CUG: L	CCG: P	CAG: G	CGG: R
AUU: I	ACU: T	AAU: N	AGU: S
AUC: I	ACC: T	AAC: N	AGC: S
AUA: I	ACA: T	AAA: K	AGA: R
AUG: M	ACG: T	AAG: K	AGG: R
GUU: V	GCU: A	GAU: D	GGU: G
GUC: V	GCC: A	GAC: D	GGC: G
GUA: V	GCA: A	GAA: E	GGA: G
GUG: V	GCG: A	GAG: E	GGG: G

FIG. 4.—Evolutionary potential at two active sites of the *ebgA* gene. *A,* Residue 92 of EbgA. *B,* Residue 977 of EbgA. The darkly shaded boxes represent amino acids found at those positions in all β-galactosidases. The lighter shading indicates single-base substitutions which are never observed in experimental studies of Ebg evolution and thus must be deleterious to β-galactosidase activity. The unshaded residues are codon changes which can be reached only after traversing the deleterious single-site mutations.

tution replacements are so deleterious that they constitute a deep selective valley that has not been traversed in over 2 billion years. In either case, the natural evolutionary potential of Ebg enzyme with respect to lactose and lactulose hydrolysis is limited to those two replacements.

Ebg arose from a functional clade A β-galactosidase. The ''natural'' function of Ebg enzyme is unknown, but, given the observed degree of sequence and open reading frame conservation, it is unlikely that Ebg performs no function and is therefore not under selection. The only two active-site residues that differentiate Ebg from the other Clade A enzymes are asp_{92} and trp_{977}. Neither the asn_{92}→asp nor the cys_{977}→trp substitution would have eliminated effective β-galactosidase activity of Ebg's ancestor, but those two substitutions together certainly did so. Not only is the wild-type Ebg virtually inactive toward lactose and lactulose, but its activity toward eight other β-galactosides is so low that it is not, in any meaningful sense, a functional β-galactosidase. All of the β-galactosidases in this study have been identified on the basis of their lactase activity. Since it is clear that β-galactosidases arose long before lactose was being produced by mammals, it might be the case that some lactose-negative organisms retain homologs of these β-galactosidases. It will be interesting to see if, as more sequences accumulate in the databases, non-lactase proteins that belong to clade A begin to appear. Despite the strong conservation of proteins in the ancient clade A, it is clear that those proteins do not provide a function that is universally required, because no clade A homologs, as detected by BLAST, are present in any of the five other genomes that have been completely sequenced.

Acknowledgments

This study was supported by grant #NP-932 from the American Cancer Society. We are grateful to Simon Easteal for an essential suggestion about the fitness of possible multiple base-substitution mutants.

LITERATURE CITED

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

BURTON, J., and M. L. SINNOTT. 1983. Catalytic consequences of experimental evolution. Part 1. Catalysis by the wild-type second β-galactosidase (ebg°) of *Escherichia coli*: a comparison with the *lacZ* enzyme. J. Chem. Soc. Perkin Trans. II 359–364.

CALUGARU, S. V., B. G. HALL, and M. L. SINNOTT. 1995. Catalysis by the large subunit of the second β-galactosidase of *Escherichia coli* in the absence of the small subunit. Biochem. J. **312**:281–286.

CALUGARU, S. V., S. KRISHNAN, C. J. CHANY II, B. G. HALL, and M. L. SINNOTT. 1997. Larger increases in sensitivity to paracatalytic inactivation than in catalytic competence during experimental evolution of the second β galactosidase of *Escherichia coli.* Biochem. J. **325**:117–121.

ELLIOTT, A. C., S. K. M. L. SINNOTT, P. J. SMITH, J. BOMMUSWAMY, Z. GUO, B. G. HALL, and Y. ZHANG. 1992. The catalytic consequences of experimental evolution. Studies on the subunit structure of second (ebg) β-galactosidase of *Escherichia coli*, and on catalysis by ebg^{ab}, an experimental evolvant containing two amino acid substitutions. Biochem. J. **282**:155–164.

FENG, D.-F., G. CHO, and R. F. DOOLITTLE. 1997. Determining the divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. USA **94**:13028–13033.

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

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

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

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

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

———. 1990. Directed evolution of a bacterial operon. BioEssays **12**:551–558.

———. 1995. Evolutionary potential of the ebgA gene. Mol. Biol. Evol. **12**:514–517.

HALL, B. G., P. W. BETTS, and J. C. WOOTTON. 1989. DNA sequence analysis of artificially evolved *ebg* enzyme and *ebg* repressor genes. Genetics **123**:635–648.

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. 1975. Regulation of newly evolved enzymes. II. The ebg repressor. Genetics **81**:427–435.

HALL, B. G., M. MURRAY, S. OSBORNE, and M. L. SINNOTT. 1983. The catalytic consequences of experimental evolution. Part III. Construction of reaction profiles for hydrolysis of lactose by ebg^o, ebg^a, and ebg^b enzymes via measurements of the enzyme-catalyzed exchange of galactose-1-^{18}O by ^{13}C NMR spectroscopy. J. Chem. Soc. Perkin Trans. II 1595–1598.

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

HUANG, W., J. PETROSINO, M. SIRSCH, P. S. SHENKIN, and T. PALZKILL. 1996. Amino acid sequence determinants of β-lactamase structure and activity. J. Mol. Biol. **258**:688–703.

JACOBSON, R. H., X. J. ZHANG, R. F. DUBOSE, and B. W. MATTHEWS. 1994. Three-dimensional structure of beta-galactosidase from E. coli. Nature **369**:761–766.

LI, B. F. L., S. OSBORNE, and M. L. SINNOTT. 1983. Catalytic consequences of experimental evolution. Part 2. Rate-limiting degalactosylation in the hydrolysis of Aryl β-D-galactopyranosides by the experimental evolvants ebg^a and ebg^b. J. Chem. Soc. Perkin Trans. II 365–369.

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.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructring phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SRINIVASAN, K., B. G. HALL, and M. L. SINNOTT. 1995. The catalytic consequences of experimental evolution. Catalysis by a ‘‘third generation’’ evolvant of the second β-galactosidase of *Escherichia coli,* Ebg^{abcde} and Ebg^{abcd}, a ‘‘second generation’’ evolvant containing two supposedly ‘‘kinetically silent’’ mutations. Biochem. J. **312**:971–977.

SRINIVASAN, K., A. KONSTANTINDIS, M. L. SINNOTT, and B. G. HALL. 1993. Large changes of transition state structure during experimental evolution of an enzyme. Biochem. J. **291**:15–17.

SWOFFORD, D. L. 1993. PAUP: phylogenetic analysis using parsimony. Illinois Natural History Survey, Champaign.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

SIMON EASTEAL, reviewing editor

Accepted May 11, 1998