Nucleotide Sequence, Function, Activation, and Evolution of the Cryptic asc Operon of Escherichia coli K12

Barry G. Hall and Lei Xu
Biology Department, University of Rochester

The cryptic asc (previous called “SAC”) operon of Escherichia coli K12 has been completely sequenced. It encodes a repressor (ascG); a PTS enzyme II asc for the transport of arbutin, salicin, and cellobiose (ascF); and a phospho-β-glucosidase that hydrolyzes the sugars which are phosphorylated during transport (ascB). ascG and ascFB are transcribed from divergent promoters. The cryptic operon is activated by the insertion of IS186 into the ascG (repressor) gene. The ascFB genes are paralogous to the cryptic bgFB genes, and ascG is paralogous to galR. The duplications that gave rise to these paralogous genes are estimated to have occurred ∼320 Mya, a time that predates the divergence of E. coli and Salmonella typhimurium.

Introduction

Both laboratory strains and natural isolates of Escherichia coli possesses several sets of cryptic (silent) genes for the utilization of β-glucoside sugars (Hall and Betts 1987; Parker and Hall 1988). Cryptic genes are silent in wild-type organisms and must be activated by mutations (which may include insertion of IS or other mobile genetic elements) before they can be expressed (Hall et al. 1983). Because cryptic genes are not expected to make any positive contribution to the fitness of the organism, it is expected that they would eventually be lost because of the accumulation of inactivating mutations (Hall et al. 1983; Li 1984). Cryptic genes would thus be expected to be rare in natural population. This, however, is not the case: >90% of natural isolates of E. coli carry cryptic genes for the utilization of β-glucoside sugars (Hall and Betts 1987). The persistence of cryptic genes in the face of mutational pressure is an interesting puzzle for population biologists, and our current model is that they are retained by alternately selecting for loss and regain of function in different environments. Of the cryptic operons in E. coli, the two that are best studied are the bgI and cel operons.

The bgI operon includes three genes: bgIG, which encodes a positive regulatory protein that prevents termination of transcription in the presence of the β-glucoside sugars arbutin or salicin; bgIF, which encodes the phosphoenolpyruvate-dependent (PTS) enzyme II bgI, which (a) simultaneously transports and phosphorylates arbutin and salicin and (b) interacts with the bgIG product to prevent antitermination in the absence of substrates; and bgIB, which encodes a phospho-β-glucosidase B that hydrolyzes the phosphorylated substrates arbutin and salicin (Prasad and Schaeffer 1974; Mahadevan et al. 1987; Schnetz et al. 1987; Schnetz and Rak 1988). The bgI operon

1. Key words: cryptic genes, repressor, sugar transport, phospho-β-glucosidase, arbutin, salicin.

Address for correspondence and reprints: Barry G. Hall, Biology Department, Hutchison Hall, River Campus, University of Rochester, Rochester, New York 14627.

Mol. Biol. Evol. 9(4):688-706. 1992. © 1992 by The University of Chicago. All rights reserved. 0737-4038/92/0904-0010$02.00
Cryptic asc Operon of *E. coli* 689

can be activated either by (a) insertions of either IS1 or IS5 into a region 78–125 bp upstream of the transcription site or (b) base substitutions in the CAP-cAMP binding site (Reynolds et al. 1981, 1984). The mechanism(s) by which insertions activate the operon are not yet understood. Because transcription initiates at the same site, independent of the position of the insertion element, it is clear that insertion elements do not provide promoters necessary for transcription (Reynolds et al. 1984). It is known that insertions can increase the activity of promoters, although the mechanism by which they do so is not known (Schnetz and Rak 1988). The *bgl* operon is located at 83.8 minutes on the genetic map (Bachmann 1990) and at kb 3979.8 on the physical map (Kohara et al. 1987; Rudd et al. 1990).

The *cel* operon consists of five genes: *celA*, *celB*, and *celC*, which encode, respectively, the PTS enzyme IV*cel*, enzyme II*cel*, and enzyme III*cel*, for transport and phosphorylation of cellubiose, arbutin, and salicin (Parker and Hall 1990a, 1990b; Reizer et al. 1990) (in effect, the three functional domains that are fused as enzyme II*bgI* in the *bgl* operon are present as separate entities in the *cel* operon); *celD*, which encodes the *cel* repressor protein; and *celF*, which encodes a phospho-β-glucosidase that hydrolyzes phosphorylated cellubiose, arbutin, and salicin (Kricker and Hall 1984, 1987; Parker and Hall 1990a, 1990b). The *cel* operon can be activated by insertion of ISI, IS2, or IS5 into a region 72–180 bp upstream from the transcription start site, again by an unknown mechanism that does not involve provision of a promoter sequence (Parker and Hall 1990b). It can also be activated by base substitutions in *celD* that alter the repressor so that it can recognize cellubiose, arbutin, and salicin as inducers (Parker and Hall 1990b). The *cel* operon is located at 37.8 minutes on the genetic map (Bachmann 1990) and at kb 1835 on the physical map (Kohara et al. 1987).

In addition to these cryptic operons, *E. coli* also possesses a constitutively expressed gene, *bglA*, for a phospho-β-glucosidase that is specific for phosphorylated arbutin (Prasad et al. 1973). That enzyme does not permit arbutin utilization, because, unless one of the cryptic operons is activated, arbutin cannot be transported and phosphorylated. The *bglA* gene is located at kb 3056.1 on the physical map (Kohara et al. 1987; Junne et al. 1990), corresponding to ~62.3 minutes on the genetic map (Rudd et al. 1990).

As a parallel to the expressed *bglA* gene, *E. coli* K12 also possesses a cryptic *arbT* gene, which, when activated, permits the transport and phosphorylation of arbutin but not of other β-glucoside sugars (Kricker and Hall 1987). Neither the physical nor the genetic location of *arbT* is known.

Three years ago we reported the existence of a fourth cryptic-gene system for β-glucoside utilization in *E. coli* (Parker and Hall 1988), a system which we designated "SAC." That system has now been renamed "asc," to avoid confusion with the *sac* genes for sucrose utilization in *Bacillus subtilis* (Hall et al. 1991). The *asc* operon was discovered by selecting for cellubiose utilization in a Δ*bgI* Δ*cel* strain of *E. coli* K12, strain LP100 (Parker and Hall 1988). The first-step spontaneous mutant, strain LP101, utilized arbutin well but utilized salicin poorly and failed to utilize cellubiose. A second round of selection, applied to strain LP101, yielded the spontaneous mutant strain LP102, which grew well on both arbutin and salicin but remained unable to grow on cellubiose. A third round of selection, applied to strain LP102, yielded strain LP103, which grew well on both arbutin and salicin and which utilized cellubiose on plates but failed to utilize cellubiose in liquid medium. Strain LP103 exhibited a weak cellubiose-positive phenotype (pink colonies) on MacConkey cellubiose medium. The
asc operon was cloned from strain LP103, as a 9-kb HindIII-EcoRI fragment, into the high-copy-number vector pBlu + (plasmid pUF718). When transformed into strain LP100, that plasmid conferred a strong arbutin-, salicin-, cellobiose-positive phenotype (red colonies on MacConkey plates) (Parker and Hall 1988; in that paper it was erroneously reported that a subclone of pUF718, plasmid pUF721, also confers a strong arbutin-, salicin-, and cellobiose-positive phenotype; in fact, for reasons discussed below, pUF721 only confers a strong arbutin-positive phenotype). The asc operon is located at kb 2845 on the physical map (Kohara et al. 1987), corresponding to 58.3 minutes on the genetic map (Hall et al. 1991).

A detailed study of the asc operon was undertaken in the hope that it might shed more light on both the evolutionary basis of retention of cryptic genes in natural populations and the mechanisms by which they might be alternately silenced and activated. Here we report the complete nucleotide sequence of the asc operon, discuss the roles of each of the gene products, and report the mechanism by which the cryptic operon is activated.

Material and Methods

Culture Media and Conditions

Minimal medium consisted of 423 mg sodium citrate/liter, 100 mg MgSO₄ · 7H₂O/liter, 1 g (NH₄)₂SO₄/liter, 540 μg FeCl₃/liter, 1 mg thiamine/liter, 3 g KH₂PO₄/liter, 7 g K₂HPO₄/liter, and 1 g carbon source/liter. When required, amino acids were added to a concentration of 100 mg/liter.

The rich, complete medium was LB (Luria Broth) (Miller 1972). To select for retention of plasmids by plasmid-bearing strains we added ampicillin, to a concentration of 100 mg/liter. MacConkey indicator plates (Difco) were used to distinguish arbutin- or salicin-fermenting strains from those that did not ferment the sugars.

E. coli Strains and Plasmids

Strains and plasmids are listed in table 1.

DNA Sequencing

DNA was sequenced by the dideoxy method (Sanger et al. 1980) using deoxyadenosine 5'-[α-³⁵S]thio)triphosphate and by using a modified T7 DNA polymerase sequencing kit (Pharmacia) according to the manufacturer's instructions, except that the primers were hybridized to template DNA by boiling the template-primer mixture and then were quick-frozen in dry ice for 5 min, and the mixture was allowed to warm slowly to room temperature. Oligonucleotides were purchased from Oligo, Etc. of Ridgefield, Conn.

Double-stranded DNA templates for sequencing were prepared by a variety of methods. In some cases plasmid DNA was sequenced using synthetic oligonucleotide primers corresponding either to vector sequences that flanked the insert or, when portions of the insert sequence had already been determined, to sequences within the insert. In other cases, portions of the insert were amplified by the polymerase chain reaction (PCR) as described elsewhere (Hall et al. 1989), and the amplified DNA was sequenced directly.

The bulk of the sequence was obtained by sequencing plasmid pUF740, pUF741, or pUF742 (fig. 1) by transposon-facilitated DNA sequencing (Strathmann et al. 1991). Transposon γδ (Tn1000) was introduced at random sites into the plasmids, and the positions of the insertions were determined by PCR mapping. The sequences
Table 1
*E. coli* Strains and Plasmids

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genotype</th>
<th>Source or Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP100</td>
<td>F- rpsL, trp, his, argG, metA or B, ara, leu, lacZΔ4680, lacY (Δbgl-pho)201, celΔ100</td>
<td>Parker and Hall 1988</td>
</tr>
<tr>
<td>LP100R</td>
<td>recA56 slt300::Tn10 derivative of LP100</td>
<td>Parker and Hall 1988</td>
</tr>
<tr>
<td>LP101</td>
<td>ascG1::IS186 mutant of LP100</td>
<td>Parker and Hall 1988</td>
</tr>
<tr>
<td>LP102</td>
<td>asc mutant of LP101, improved salicin utilization</td>
<td>Parker and Hall 1988</td>
</tr>
<tr>
<td>LP103</td>
<td>asc mutant of LP102, cellulose utilization</td>
<td>Parker and Hall 1988</td>
</tr>
<tr>
<td>JC10241</td>
<td>thr300, srl300::Tn10, relA1, ilv318, spoT1, thi1, rpsE300, lambda-, Hfr (PO45)</td>
<td>Csonka and Clark 1980</td>
</tr>
<tr>
<td>DPWC</td>
<td>supE, ΔrecA, srl::Tn10 (tet sensitive), F+</td>
<td>Strathmann et al. 1991</td>
</tr>
<tr>
<td>JGM</td>
<td>F- Tn5 seq1</td>
<td>Strathmann et al. 1991</td>
</tr>
<tr>
<td>MK91243</td>
<td>F- rpsL, trp, his, argG, metA or B, ara, leu, lacZΔ4680, lacY (Δbgl-pho)201, celR2+, arblT1*</td>
<td>Kricker and Hall 1987</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pBlu+</td>
<td>High-copy-number cloning vector</td>
</tr>
<tr>
<td>pUF718</td>
<td>Complete asc operon of strain LP103 and flanking regions cloned into pBlu+</td>
</tr>
<tr>
<td>pUF721</td>
<td>ascGF and 5' portion of ascB cloned into pBlu+</td>
</tr>
<tr>
<td>pMOB</td>
<td>1.8-kb minivector for γδ-based sequencing</td>
</tr>
<tr>
<td>pUF740</td>
<td>ascGF and 5' portion of ascB cloned into pMOB</td>
</tr>
<tr>
<td>pUF741</td>
<td>ascF and 5' portion of ascB cloned into pMOB</td>
</tr>
<tr>
<td>pUF742</td>
<td>3' Portion of ascF and 5' portion of ascB cloned into pMOB</td>
</tr>
<tr>
<td>pUF745</td>
<td>3' Portion of ascB cloned into pMOB</td>
</tr>
</tbody>
</table>

flanking the insertions were then sequenced using oligonucleotide primers corresponding to the unique portions of the two ends of γδ. With the exception of bases 1–39, the entire sequence shown in figure 2 was determined on both strands.

Computer Analyses of DNA and Protein Sequences

Sequences were analyzed with the Sequence Analysis Software Package, version 7.0, by Genetics Computer (GCG Package). Alignments were conducted according to the Needleman and Wunsch method (1970) using the GAP program of the GCG package. Significance of alignments was determined using the Randomizations option of the GAP program. Promoter sequences were identified by M. C. O’Neill using a neural network program (O’Neill 1991). The GenBank data bases of nucleic acid and protein sequences were searched with the FASTA (Pearson and Lipman 1988) program. Divergence at silent sites was estimated using a program provided by W.-H. Li (Li et al. 1985).

Results

Figure 1 shows a diagram of the asc operon structure in strain LP103 and indicates the extents of the cloned inserts in the plasmids used in this study. Figure 2 shows the complete nucleotide sequence of the operon.

Plasmid pUF721 (fig. 1) allows strain LP100R (Δbgl Δcel recA) to utilize arbutin. Sequencing of the left end of the pUF721 insert revealed that the insert began at bp
666 (the PstI site) of insertion element IS186. A primer complementary to bp 103–120 of IS186 was used to sequence the region to the left of the IS186 insertion, using pUF718 DNA as a template. The other strand was sequenced using a primer complementary to bp 1–18 of the asc operon sequence. Insertion of IS186 caused an 11-bp duplication of the insertion site, and the site of the IS186 insertion is arbitrarily positioned at the center of that duplication, at bp 190 (fig. 2). The asc operon includes three large open reading frames (fig. 2), which have been named in keeping with the names of the functionally similar genes in the bgl operon.

The ascG Gene

ascG is encoded on the lower (complementary) strand, includes bp 126–1130, is preceded by a ribosome-binding site (bp 1137–1142), and is disrupted by IS186 in strain LP103 (see discussion on the mechanism of activation, below). ascG encodes a 334-amino-acid peptide with a molecular weight of 36,787. The ascG gene product is identified as a repressor protein, on several grounds. First, a Fasta search of the SwissProt protein data base reveals that the top 10 in the list of similar proteins are all repressor proteins. The alignment of the asc repressor with the most similar of those repressors, the galactose repressor (galR gene product), is shown in figure 3. There is 31.6% sequence identity, and 54.8% sequence similarity, between the ascG and galR repressors. The quality score for the alignment shown is 211.5, while the score for alignment of the ascG repressor with a randomly shuffled galR repressor sequence (100 randomizations) is 105.6 ± 4.0 (mean ± standard deviation). Second, the ascG repressor exhibits a DNA-binding helix-turn-helix motif (Kelly and Yanofsky 1985) (fig. 2) that includes identities with 14 of the 17 consensus residues in that region for the ascG repressor and 12 homologous repressor proteins. Third, disruption of ascG by IS186 results in semiconstitutive expression of the asc operon in strain LP101 (Parker and Hall 1988) (see discussion of activation of the operon, below).
Fig. 2.—Sequence of the asc operon of strain LP103. Structural features (Shine-Dalgarno ribosome-binding sites, promoter sequences) were determined by pattern similarity, not experimentally.
The ascF Gene

ascF includes bp 1390–2847, is preceded by a ribosome-binding site (bp 1377–1382), and encodes a 485-amino-acid peptide with a molecular weight of 51,030. The gene product is identified as a PTS enzyme II for β-glucoside sugars, enzyme II_asc. That identification is based on both experimental and pattern-similarity criteria.

First, strain LP100R carrying pUF721, pUF740, or pUF741 is arbutin positive, but salicin negative and cellobiose negative. All three plasmids carry the entire ascF gene, but only a 5' fragment of the ascB (phospho-β-glucosidase; see discussion below)
gene. If enzyme II<sub>asc</sub> transports and phosphorylates arbutin, then the phosphorylated arbutin can be hydrolyzed by the constitutively synthesized phospho-β-glucosidase A protein that is specific for arbutin. Those plasmids render the cell sensitive to growth inhibition by salicin or cellobiose in glucose minimal medium, which is consistent with the transport and accumulation of those phosphorylated sugars. Plasmid pUF742, which carries only the 3′ portion of asc<sub>F</sub> and the 5′ portion of the asc<sub>B</sub>, neither allows growth on arbutin nor renders the cell sensitive to salicin or cellobiose inhibition.

Second, the four most closely related proteins identified in a Fasta search of the SwissProt data base were PTS enzymes II. The most closely related of these was the PTS enzyme II<sub>bg</sub> of Escherichia coli. Figure 4 shows that alignment of the asc<sub>F</sub> and bg<sub>F</sub> gene products. There is 30.9% sequence identity, and 55.4% sequence similarity, between the two proteins. The quality score for the alignment shown is 295, which is 23.8 standard deviations above the score for the alignment of the asc<sub>F</sub> product versus the randomly shuffled bg<sub>F</sub> product (quality score = 180.8 ± 4.8 for 100 random shuffles).

The enzyme II<sub>asc</sub>, containing 485 amino acids, is only 78% as long as is enzyme II<sub>bg</sub>, which contains 625 amino acids. The sugar-specific proteins of the PTS phosphotransferase transport system include three functional domains that may be present either on a single polypeptide or on two separate polypeptides, in which case the two proteins are designated as an enzyme II-enzyme III pair and sum to an average of ~635 ± 13 amino acids (Saier et al. 1988; Saier and Reizer 1990). A phosphate group
is transferred from phosphoenolpyruvate through two system-wide proteins—PTS-enzyme I and HPr—to a histidine in phosphorylation site I, which is located in the domain that itself is located either in the C-terminal portion of full-length enzyme II’s or in enzyme III of enzyme II–enzyme III pairs (Saier et al. 1988). From there the phosphate is transferred to a phosphorylation site II in the enzyme II and is finally transferred to the sugar itself (Saier et al. 1988). The asc operon does not include an enzyme III (we have sequenced 287 bases beyond those shown in figure 2 and have found no open reading frames that could encode a portion of an enzyme III), and the asc phosphotransferase system thus lacks a site for the first phosphorylation step. Enzyme IIasc, however, does include phosphorylation site II (fig. 4).

The most likely explanation for the ability of enzyme IIasc to function without
an enzyme III asc is that it uses an enzyme III that is part of another sugar-transport system. This is the case for the plasmid-encoded enzyme II sucrose of a clinical isolate of *Salmonella typhimurium*, which functions in conjunction with enzyme III glucose (Lengler et al. 1982). In that regard, after enzyme II bgI, the closest relative of enzyme II asc is the 455-amino-acid protein, enzyme II sucrose of *Salmonella*, which shares 27.2% sequence identity with enzyme II asc. 

The ascB Gene

*ascB* includes bp 2856–4280, is preceded by a ribosome-binding site (bp 2847–2852), and encodes a 477-amino-acid protein with a molecular weight of 53,883. The gene product is identified as a phospho-β-glucosidase that can hydrolyze salicin and cellobiose and that can probably hydrolyze arbutin. That identification is based on both experimental and pattern-similarity criteria.

First, strain LP100R carrying the entire asc operon on plasmid pUF718 (fig. 1) is strongly arbutin positive, salicin positive, and cellobiose positive, whereas LP100R carrying plasmid pUF721, pUF740, or pUF741, which carry only the 5' end of asc (fig. 1), is only positive on arbutin (which is hydrolyzed by the constitutively expressed phospho-β-glucosidase A).

Second, the eight most closely related proteins in the SwisProt protein data base are all glycosidases, of which the most similar is the bgI-encoding phospho-β-glucosidase B. Alignment of the ascB- and the bgI-encoding phospho-β-glucosidases (fig. 5) shows that the proteins share 53.9% sequence identity and 70.5% sequence similarity. The quality score for that alignment is 438.6, 81 standard deviations above that of the randomly shuffled alignment (137.2 ± 3.7 for 100 shuffles).

We cannot firmly conclude that the ascB-encoding phospho-β-glucosidase hydrolyzes phosphorylated arbutin, because we do not have available a strain that fails to synthesize the bgI-encoding phospho-β-glucosidase A. We believe that it is likely to be active toward arbutin because (a) all phospho-β-glucosidases from the Enterobacteriaceae that have been described so far act on arbutin and (b) plasmid pUF718, which has an intact ascB gene, produces much more intensely red (positive) 1-d-old colonies on MacConkey arbutin plates than does plasmid pUF721, in which the ascB gene is disrupted.

DNA Sequence Homologies and Estimated Time at Which Gene Duplication Occurred

Given the homologies between the ascF and ascB proteins and between the bgI and bgI proteins, and the homology between the ascG- and galR-encoded repressors, it is not surprising that a Fasta search of the GenBank data base found that the two DNA sequences most closely related to the asc operon were the bgI operon and the galR sequences. The DNA sequences of the above coding regions were compared after introduction of gaps corresponding to the gaps in the aligned protein sequences by setting the gap and gap-length penalties so high that no additional gaps were introduced (table 2). The quality scores for those DNA alignments were all >24 standard deviations above the scores for randomized alignments, showing that the DNA sequences are clearly paralogous. These comparisons can shed some light on the time when the duplications of the ancestral genes occurred. Sharp (1991) has recently compared 67 homologous (orthologous) genes in *E. coli* and *S. typhimurium*. For those genes the average identity at the DNA level is 84.4%, but the range is 72.5%–99%. The observation that none of the gene pairs share >58% DNA sequence identity suggests that the
duplications preceded the divergence of E. coli from S. typhimurium. That conclusion, however, is complicated by the fact that these genes are paralogous; that is, their products carry out different functions, and the genes could have diverged rapidly as a result of having been subjected to diversifying selection. In the E. coli-versus-S. typhimurium comparison the most diverse of the homologous (orthologous) proteins shared 76.8% sequence identity (Sharp 1991), while the most similar of the paralogous proteins share only 58.6% identity. A more meaningful conclusion can be drawn from the divergence at synonymous sites, those sites in which substitutions do not result in amino acid substitutions. The aligned paralogous DNA sequences were analyzed by the method of Li et al. (1985). The number of synonymous substitutions per synonymous site ($K_s$) and the number of amino acid replacement substitutions per replacement site ($K_a$), corrected for multiple substitutions, are shown in Table 2. For the ascG-versus-galR comparison, the $K_s$ value is meaningless: the synonymous sites
Table 2
Similarities and Divergences between Paralogues

<table>
<thead>
<tr>
<th>Gene Pair</th>
<th>Amino Acid Identity (%)</th>
<th>DNA Identity (%)</th>
<th>CAIa</th>
<th>KsA</th>
<th>Ks</th>
</tr>
</thead>
<tbody>
<tr>
<td>ascG vs. galR</td>
<td>31.6</td>
<td>44.3</td>
<td>0.275 vs. 0.328</td>
<td>0.76</td>
<td>100.0b</td>
</tr>
<tr>
<td>ascF vs. bglF</td>
<td>30.9</td>
<td>46.0</td>
<td>0.330 vs. 0.273</td>
<td>0.75</td>
<td>2.69</td>
</tr>
<tr>
<td>ascB vs. bglB</td>
<td>53.9</td>
<td>58.0</td>
<td>0.382 vs. 0.277</td>
<td>0.41</td>
<td>2.54</td>
</tr>
</tbody>
</table>

* Codon adaptation index, first vs. second member of pair.
* Silent sites are saturated for substitutions.

have been saturated with repeated substitutions. For the remaining two paralogous comparisons, Ks values are higher than the highest Ks observed in the comparison of orthologous genes in *E. coli* and *S. typhimurium* (trpA; Ks = 1.77). Even though they do not result in amino acid replacements, synonymous substitutions are not necessarily neutral, because there exists a codon-usage bias that varies with the level of expression of the gene and that is inversely correlated with the rate of synonymous substitutions (Sharp and Li 1987b, 1987c). The codon adaptation index (CAI) is a measure of codon bias (Sharp and Li 1987a). The CAI for the trpA gene is 0.34 in *E. coli* and 0.32 in *S. typhimurium* (Sharp 1991), values that are not very different from the CAIs of the paralogous genes in table 2. On the basis of synonymous substitution rates, it therefore appears that ascFB diverged from bglFB before *E. coli* and *S. typhimurium* diverged and that ascG and galR diverged earlier still. For 28 *E. coli* and *S. typhimurium* genes with CAIs of 0.275 and 0.382, the value for Ks is 1.15 ± 0.069 (mean ± standard error). *Escherichia coli* and *S. typhimurium* are estimated to have diverged 120-160 Mya (Ochman and Wilson 1987). The mean Ks for ascFB versus bglFB is 2.62, suggesting that they diverged ~320 Mya.

Promoters

Promoter sequences were identified by using a neural network program (O'Neill 1991). Only two highly probable promoters, each with a 17-bp spacing between the −35 and −10 regions, were identified within the region between the ascG and ascFB genes. In keeping with convention (O'Neill 1991), the regions between −50 and +8 are underlined in figure 2.

Activation of the asc Operon

Because the active asc operon in strain LP103 contains IS186 inserted into the 3′ end of the ascG (repressor) gene, it is reasonable to ask whether the insertion event was responsible for the initial activation of asc in strain LP101. To determine whether strains LP100 (wild-type asc operon), LP101 (activated mutant of LP100), or LP102 (descendant of LP101, parent of LP103) contained an insertion into ascG, we amplified bp 29–400 of the asc operon by using suitable primers, and the sizes of the amplified fragments were determined by agarose gel electrophoresis. Strain LP100 yielded a fragment of ~370 bp, indicating that it did not contain an insertion within the region, while strains LP101, LP102, and LP103 yielded a 1.7-kb fragment, indicating that each contained IS186 within the 3′ region of ascG. It is reasonable to conclude that insertional inactivation of the asc operon repressor was responsible for activation of the operon.

To determine whether the asc operon could be activated by other mutations, we
plated strain LP100 onto MacConkey arbutin and MacConkey salicin media, and the plates were incubated at 30°C in humidified chambers. After 7 d, red papillae began to appear on the colonies on MacConkey arbutin plates, and 10 Arb⁺ mutants were isolated for analysis. In contrast to the first-step asc mutant—strain LP101—one of the new mutants were even slightly positive on MacConkey salicin medium. Each mutant was transduced with bacteriophage P1 grown on strain JC10241, which carries Tn10::srl. All 1,935 of the tetracycline-resistant transductants were srl (sorbitol negative), but all remained arbutin positive. There is 50% cotransduction between srl and asc (Hall et al. 1991); thus the failure to find cotransduction between srl and the arbutin-utilization phenotype indicated that the new Arb⁺ mutants were not asc mutants. When the arbT (arbutin-utilizing) strain MK91243 was transduced with the same P1 lysate from strain JC10241, all 286 tetracycline-resistant transductants were srl, and all were arbutin positive; thus it seems likely that the new Arb⁺ mutants were arbT.

After 5 wk incubation, papillae began to appear on the colonies on MacConkey salicin plates, and eight of those were isolated as pink or red colonies when they were restreaked onto MacConkey salicin plates. Of the eight, only four strains were genetically stable. The remainder segregated Arb⁻ Sal⁻ colonies at such high rates that 15%–55% of the cells in apparently Sal⁺ colonies on MacConkey salicin medium proved to be Arb⁻ Sal⁻ when the colonies were suspended in buffer and replated onto the same medium. During growth in liquid media, segregation was so rapid that <1% of the cells remained Arb⁺ Sal⁺. PCR amplification of the 29–400-bp region of the asc operon by using DNA isolated from cultures of those unstable strains gave mixed results: two strains showed no evidence of IS186 within that region, suggesting either that the original activation was by some other mechanism or that precise excision had restored the integrity of the ascG gene; and the other two strains contained an insertion into that region. Sequencing the amplified DNA showed that, in those strains, IS186 had inserted into the identical site and in the same orientation as shown in figures 1 and 2. In those cases, loss of the Arb⁺ Sal⁺ phenotype involved some lesion other than loss of the IS186 insertion.

All four of the stable Arb⁺ Sal⁺ strains proved to have IS186 inserted into the same site and in the same orientation as did strain LP103. Even in “stable” strains, however, the genotype is not completely stable. A strain LP103 colony consisting of 2.5 × 10⁸ cells contained ~2.5% Arb⁻ Sal⁻ cells, which proved to have lost IS186 from within the 29–400-bp region of the asc operon. The finding that excision of IS186 can accompany loss of the asc Arb⁺ Sal⁺ phenotype supports the notion that the two unstable strains in which IS186 could not be detected within the 29–400-bp region of the asc operon had originally contained the insertion but had lost it during growth of the culture.

Discussion

Clearly, ascFB is paralogous (homologous via a genomic duplication) to bglFB, and ascG is paralogous to galR. The duplications that led to these divergent genes are ancient, on the order of 320 Myr old, and predate the divergence of Escherichia coli from Salmonella. (However, because of the tremendous multihit corrections required, the exact time of divergence cannot be taken very seriously.) The asc and bgl operons appear to have been assembled from different “modules,” because, although their phospho-β-glucosidase and transport genes are homologous, their regulatory genes are unrelated. A similar phenomenon was observed for the cryptic cet operon of E. coli:
its glycosidase and regulatory genes (celDF) are homologous to the mellibiose genes melR and melA, while its enzyme III_{cat} transport gene is homologous to a gene from *Staphylococcus aureus* (Parker and Hall 1990a).

Given the ancient divergence of ascFB from bglFB, it is reasonable to ask how these genes could have persisted in a silent state for so long without being eliminated by random inactivating mutations. The fact that both operons can be activated by simple insertions of mobile elements indicates that inactivating mutations have not accumulated. One possibility is that both genes were recently silenced and that their "crypticity" is something of a laboratory artifact. This appears unlikely. None of the 72 natural isolates in the ECOR (Ochman and Selander 1984) collection can utilize either arbutin or salicin, but 66 of those strains produced spontaneous mutants capable of utilizing one or both of those sugars (Hall and Betts 1987), indicating that they carried β-glucoside utilization genes in a cryptic state. It appears equally unlikely that both genes could have (a) remained silent (and thus not subject to purifying selection) for over 100 Myr (~10^{10} generations) and (b) escaped inactivation. Mutations that inactivate the gal operon occur at the rate of 2.6 × 10^{-5}/cell division (B. G. Hall, unpublished data); thus, if inactivating mutations were truly neutral over that entire period, both operons would have been eliminated long ago. This is by no means the only case of a gene that appears to have been cryptic for a 100 Myr. One of the major criteria that is used to distinguish *Salmonella* from *E. coli* is the latter's inability to utilize citrate as a carbon source, an inability that arises from the lack of a functional citrate transport system in *E. coli*. Nevertheless, *E. coli* carries a cryptic citrate-transport system that can be activated by spontaneous mutation (Hall 1982).

Several years ago we (Hall et al. 1983) proposed that cryptic genes could be maintained by selection acting in opposite directions in alternating environments. We suggested that functional alleles would be advantageous in one, rarely encountered environment, while both cryptic and nonfunctional alleles (those that had been permanently inactivated) would be advantageous in another, commonly encountered environment. That model was supported by a theoretical study showing that cryptic and functional alleles can be retained without having any advantage over nonfunctional alleles, if selection for the active allele occurs occasionally for a substantially long time (Li 1984). The cryptic alleles would be retained, for instance, if selection favored the functional (active) allele for 200 generations and favored the silent allele for 250,000 generations. On the other hand, if selection for the active allele occurred only for 100 of the 250,000 generations, then both the functional and cryptic alleles would be eliminated after only a few cycles between the environments. Cryptic genes, then, are genes that have the peculiar property that it is disadvantageous to express them in normal environments.

Experimental support for that model requires demonstrating (1) that silent alleles of some gene are actually favored in some environment and are disfavored in another environment, (2) that cryptic genes do accumulate mutations when they are silent and that some alleles are occasionally eliminated as the result of the mutations that accumulate, and (3) that cryptic genes can be activated by easily reversible means that will permit cycling between the active and cryptic states. The first point was demonstrated by showing that, in a mixed-resource environment where both cellobiose and glycerol were present, the cryptic *cel* (cellobiose-utilization) operon was strongly favored over active *cel* alleles but that, in a single-resource environment where only cellobiose was present, the active alleles were strongly favored (Hall et al. 1986). The second point was supported both by the finding that recently activated alleles are quite
Cryptic asc Operon of E. coli

temperature sensitive (Hall and Faunce 1987) and by a survey that showed that the
bgl operon was deleted in 29% of the strains in two collections of natural isolates of
E. coli (Hall 1988).

The work presented here supports the third point. The asc operon can be activated
by insertion of IS186 into ascG, but that activation is readily reversible by precise
excision of IS186. The finding that asc can be activated by insertion sequences is
consistent with the findings that both the bgl and cel operons can be activated by
insertion sequences (Reynolds et al. 1981, 1984; Parker and Hall 1990b).

The simplest model to explain the normally cryptic state of the asc operon is that
the ascG-encoded repressor is insensitive to β-glucosides as inducers and that the
operon is only expressed after inactivation of the repressor. It is not particularly sur-
prising to find that the asc operon can be activated by insertion of IS186 into the
repressor gene, ascG. It is surprising, however, to discover that all seven of the insertions
were into exactly the same site and were in the same orientation. If simple repressor
inactivation is all that is required to activate the operon, why does it take 4–5 wk for
asc mutants to appear as papillae? Why is the repressor not inactivated by other
mutations, such as missense mutations, frameshifts, deletions, and insertions of other
mobile elements? Since all of the insertions appear to be identical, why are some
mutants, such as LP101-103, quite stable while others are highly unstable? It is possible,
although we have no evidence for it, that the stable mutants have accumulated ad-
ditional mutations in IS186 that reduce its excision frequency. Simple repressor in-
activation is also inconsistent with the observation that expression of the asc operon
is regulated: in strain LP103, arbutin induces asc expression about fourfold above the
basal level, and salicin induces it about twofold above the basal level (Parker and Hall
1988). We have not yet identified the mutations that distinguish strain LP102 from
LP101 and LP103 from LP102. Additional experimental work is required (1) to dis-
tinguish insertion of IS186 from other mutations that might inactivate the putative
ascG-encoded repressor—and that thus might activate the operon—and (2) to examine
the roles of other mutations, both in ascFB and in IS186, in determining the stability
of the active state of the asc operon.

These results support the notion that cryptic genes are not simply functionless
genes that are at an early stage of becoming pseudogenes, but instead that they arise
repeatedly in response to environmental changes. It is tempting to think of cryptic
genes as some sort of genomic hedge against future challenges, but that sort of antic-
ipatory evolution appears extremely unlikely. The silencing and activation of cryptic
genres can better be viewed as a form of genetic (as opposed to physiological) regulation
of gene expression for those genes whose products have the peculiar physiological
property that they normally are dangerous but occasionally are very useful. Although
they have been studied most extensively by model systems involving β-glucoside uti-
ization, cryptic genes as a class are probably related only by the peculiar physiological
properties of their products, rather than by ancestral homology. The cel operon, for
instance, appears to be completely unrelated to either bgl or asc.

The repeated cycles of silencing and activation of some genes leads to a picture
in which these genes could have a significant effect on the population dynamics of E.
coli. During the long periods when the genes are silent, they are expected to accumulate
mutations, because the genes are not subject to purifying selection. When the envi-
ronment changes so that active alleles are favored, those alleles that have accumulated
enough mutations to make them inactive would be purged from the population quickly.
The result of that purging would be a population bottleneck. In effect, purifying se-
lection, instead of being applied gradually over the long period when the gene was silent, is applied all at once when the environment changes; that is, selection is episodic, rather than fairly constant. Over long periods of time that include many cycles of silencing and reactivation, the intensity of selection on a cryptic gene is thus expected to be about the same as it would be if the gene had been expressed continuously. This view is consistent with the recent finding that the DNA sequence diversity of 12 naturally occurring alleles of the cryptic celC gene (encoding PTS enzyme III<sub>cellobiose</sub>) is about the same as that for alleles of the functional gutB gene (encoding PTS enzyme III<sub>glucitol</sub>) that are from the same 12 strains (Hall and Sharp 1992). The consistent finding that cryptic-gene activation involves insertion sequences suggests that these mobile elements may not be entirely selfish DNA after all—but that, instead, they may play an important (if only occasional) role in controlling expression of normally cryptic genes.

**Sequence Availability**

This sequence has been submitted to GenBank and has been assigned accession number M73326.

**Acknowledgments**

We thank Yansu Ju and Shariff Bayoumy for their contributions to some of the experiments in this study. We are grateful to Milton Saier, Ken Rudd, Jonathan Reizer, Bernard Henrisat, Frédéric Barras, and Michael O'Neill for their interest in this project, their encouragement and advice, and their gracious help with computer analyses of this sequence. We particularly thank two anonymous reviewers for their comments. This study was supported by USPHS grant GM-37110.

**LITERATURE CITED**


Cryptic operon of E. coli


MARTIN KREITMAN, reviewing editor

Received July 22, 1991; Revision received December 18, 1991

Accepted December 27, 1991