JOURNAL OF BACTERIOLOGY, June 1987, p. 2579–2590
0021-9193/87/062579-12$02.00/0
Copyright © 1987, American Society for Microbiology

Vol. 169, No. 6

β-Glucoside (*bgl*) Operon of *Escherichia coli* K-12: Nucleotide Sequence, Genetic Organization, and Possible Evolutionary Relationship to Regulatory Components of Two *Bacillus subtilis* Genes

KARIN SCHNETZ, CHRISTIAN TOLOCZYKI, AND BODO RAK*

Institut für Biologie III, University of Freiburg, D-7800 Freiburg, Federal Republic of Germany

Received 23 December 1986/Accepted 5 March 1987

Wild-type *Escherichia coli* cells are unable to grow on β-glucosides. Spontaneous mutants arise, however, which are able to utilize certain aromatic β-glucosides such as salicin or arbutin as carbon sources, revealing the presence of a cryptic operon called *bgl*. Mutations activating the operon map within (or close to) the promoter region of the operon and are due to the transposition of an IS*1* or IS*5* insertion element into this region. This operon was reported to consist of three genes coding for a phospho-β-glucosidase, a specific transport protein (enzyme II^{Bgl}), and a positively regulating protein. We have defined the extent and location of three structural genes, *bglC*, *bglS*, and *bglB*, and have determined their DNA sequence. The amino acid sequences deduced from the open reading frames together with deletion and subcloning analyses suggest that the first gene, *bglC*, codes for the regulatory protein, the second, *bglS*, codes for the transport protein, and the third, *bglB*, for phospho-β-glucosidase. A fourth gene may exist which codes for a product of unknown function. We discuss structural features of the DNA sequence which may bear on the regulation of the operon. Homologies to sequences preceding the gene for an excreted levansucrase of *Bacillus subtilis*, which are known to be involved in the regulation of this gene, and to sequences preceding the gene for an excreted β-endoglucanase of *B. subtilis*, for which data pertaining to regulation are not yet available, suggest a close evolutionary relationship among the regulatory components of all three systems.

Members of the family *Enterobacteriaceae* differ in their capacity to ferment the various β-glucosides. Wild-type strains of *Escherichia coli* are β-glucoside negative but mutate spontaneously to Bgl$^+$, enabling them to grow on aryl-β-glucosides such as salicin or arbutin (40). The spontaneously occurring Bgl$^+$ mutations uncover a cryptic operon residing at 83.5 min on the genetic map of *E. coli* K-12 (4). The operon contains a regulatory site, *bglR*, where the Bgl$^+$ mutations map, and codes for at least three proteins (31): a phospho-β-glucosidase with high specificity for aryl-β-glucosides, a transport protein (enzyme II^{Bgl}) that is a member of the phosphoenolpyruvate-dependent carbohydrate-phosphotransferase system (11, 30, 39) mediating the intracellular accumulation of aryl-β-glucoside-phosphates, and a positive regulatory protein specific for the operon.

Substrates for the phospho-β-glucosidase, which is encoded by gene *bglB* of the operon, are phosphorylated salicin and arbutin. A second gene for a phospho-β-glucosidase, *bglA*, is not linked to the *bgl* operon and is expressed constitutively. This enzyme accepts arbutin as a substrate but not salicin (40).

Genetic (20) and molecular evidence (K. Schnetz and B. Rak, manuscript in preparation) demonstrates positive regulation of the operon and suggests that the positive regulation is exerted via specific antitermination of transcription.

Analysis of spontaneous mutations leading to the activation of the operon showed that they were due to integration of either IS*1* or IS*5* into a small region proximal to the *bgl* operon (33)—upstream of the *bgl* promoter and cyclic AMP binding protein binding site—causing increased activity of the *bgl* promoter (34, 35). Indeed, out of about 1,000 spontaneous Bgl$^+$ mutations isolated in *E. coli* K-12 carrying the wild-type operon on a plasmid, only 17 were not due to transposition of one of these elements (H. Ronecker, K. Schnetz, and B. Rak, unpublished data). Activation could, at least in the case of IS*5*, be caused by specific sequences internal to the element, exerting their effect in an orientation-independent manner from positions upstream as well as downstream of the promoter, analogous to the eucaryotic enhancing sequences (Schnetz and Rak, in preparation).

In this communication we present the nucleotide sequence of a 5,270-base-pair (bp) segment of the *E. coli* chromosome which includes all functions necessary for regulated uptake and degradation of aryl-β-glucosides but possibly not the 3′ end of the *bgl* operon. These data extend the known sequence information in the region of *oriC* to a total of >25 kbp, the longest contiguous sequence of the *E. coli* genome reported to date.

Our sequence data together with subcloning experiments confirm the existence of three genes, which are sufficient for regulated uptake and degradation of aryl-β-glucosides. According to recent mapping studies, the genes and regulatory sites of the *bgl* operon are arranged in the following order: *bglR*, *bgl* promoter, *bglC*, *bglS*, and *bglB*. It has been claimed that *bglC* codes for the transport protein, *bglS* codes for the positive regulator protein, and *bglB* codes for phospho-β-glucosidase (4, 33). In the interest of a uniform nomenclature we use the gene symbols in the order defined previously (4, 33). It is now clear, however, that *bglC* codes for the positive regulator protein and *bglS* codes for the

* Corresponding author.

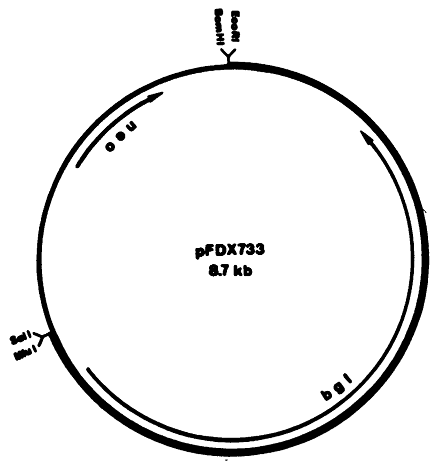

FIG. 1. Construction of plasmid pFDX733. Plasmid pAR6 was digested with *Eco*RI, and the protruding ends were filled in with DNA polymerase I (large Klenow fragment). A *Bam*HI linker (decamer) was ligated to the ends followed by digestion of the DNA with *Mlu*I. After polymerase treatment as above, the DNA was digested with *Bam*HI and a 6-kbp DNA fragment was isolated from an agarose gel. Vector plasmid pACYC177 was linearized with *Dra*I, ligated with a *Sal*I linker (octamer), and digested with *Sal*I. The ends were filled in with polymerase as above, and the DNA was digested with *Bam*HI, treated with alkaline phosphatase (calf intestinal), and run on an agarose gel. A 2.7-kbp fragment containing the origin of replication and the *neo* gene was isolated and ligated to the DNA fragment from pAR6. The ligation mixture was used to transform strain CSH26, and kanamycin-resistant colonies were selected.

transport protein. Thus the functional assignments for these two genes must be exchanged.

Gene *bglC* is preceded by a leader of 130 bases containing a potential stem-loop structure reminiscent of rho-independent terminators (37) and is followed by a long intercistronic sequence of 136 bases, which again contains a potential rho-independent transcriptional stop signal. These two regions contain elements sharing extensive sequence homology. Furthermore, similar sequences are also found 5′ to two *Bacillus subtilis* genes, one coding for an excreted levansucrase (45) and the other coding for an excreted β-endoglucanase (27). In the latter case, this homology extends—on the level of amino acid sequence—into the C terminus of *bglC*, suggesting a conserved evolutionary relationship between the respective regulatory components. Expression of levansucrase has recently been shown to be regulated by a mechanism of antitermination, acting at a site close to the region of homology (43).

MATERIALS AND METHODS

Bacterial strains and plasmids. The following strains are derivatives of *E. coli* K-12: R1068 *galK2 rpsL recA52* (from our collection); CSH26 *ara* Δ(*lac-proAB*) *thi* (26); JF201 Δ*lac*(*X74*) Δ(*pho-bgl*) *ara* B1$^-$ (35). Strain SL5235 is *Salmonella typhimurium* LT2 *metA metD trpD leu rpsL hsdL* (r$^-$m$^+$) *hsdSA* (r$^-$m$^+$) *hsdSB* (r$^-$m$^+$) (from B. Stocker). This strain should be identical to strain LB5000 (8) but contains additional unidentified auxotrophies. Plasmid pAR6 is a derivative of pBR322 (49) that contains a chromosomal *Eco*RI fragment with part of the *bgl* operon (34); plasmid pACYC177 (10) is a multicopy vector compatible with pBR322; plasmid pUC12 (25) is a high-copy-number derivative of pBR322 with a polylinker sequence. Plasmid pFDX53-Sal is a derivative of plasmid pFD53 (32) in which a *Sal*I linker (octamer) was ligated into the singular *Xmn*I site of the resident insertion element IS*5*, 25 bp away from its left end (H. Eibel and B. Rak, unpublished data). Plasmid pFDX99 carries gene *lacI*^{q} (9) substituted into vector plasmid pACYC177 providing overproduction of *lac* repressor (T. Khosaka and B. Rak, unpublished data).

Media for bacterial growth. LB (26) was used as standard liquid medium. If used for plates, 15 g of agar per liter was added. Bgl indicator plates were prepared as follows: 22.5 g of antibiotic medium no. 2 (Difco Laboratories) was dissolved in 1 liter of H_2O and autoclaved. Filter-sterilized salicin (BS plates) or arbutin (BA plates) was added to a final concentration of 0.5% followed by 10 ml of a solution containing 2% bromothymol blue, 50% ethanol, and 0.2 N NaOH. All media were supplemented with ampicillin or kanamycin (final concentration, 50 μg/ml) or both when necessary. MacConkey indicator plates contained 0.5% salicin or arbutin.

Isolation of Bgl$^+$ mutants. Bacteria were streaked onto BS plates and incubated at 37°C. Bgl$^+$ mutants grew to orange papillae on the light-green background. They were picked and restreaked several times for purification.

DNA manipulations. DNA manipulations were done essentially as described previously (21, 32). DNA restriction fragments for cloning and sequencing were eluted from agarose gels by using DEAE-membrane NA45 (Schleicher & Schuell Co.) as suggested by the manufacturer, except that the buffer for elution (high NaCl-EDTA-Tris hydrochloride) was 2 M for NaCl, which increased the yield. A 1:2 dilution with H_2O was made before ethanol precipitation.

DNA sequencing. DNA was sequenced by the chemical degradation method of Maxam and Gilbert (23). Fragments to be sequenced were prepared from plasmid pFDX733 (Fig. 1) and either processed directly or first subcloned into the *Sma*I site of plasmid vector pUC12. For subcloning the ends of the fragments were made blunt with DNA polymerase (Klenow large fragment) where necessary. The relevant structures of the various subclones are schematically given in Fig. 2. For preparation of fragments, plasmid DNA was cut with the restriction enzyme specific for the site to be labeled. The 5′ ends were dephosphorylated by using calf intestinal phosphatase followed by agarose gel electrophoreses and isolation of the appropriate fragment or by phenol-chloroform extraction and ethanol precipitation. The DNA was recut with an appropriate restriction enzyme, fragments were separated on an agarose gel, and the fragment to be sequenced was isolated. Fragments of the subclones were isolated by using restriction sites within the polylinker of the vector. Fragments prepared in this fashion accept label at only one end when treated with polynucleotide kinase, circumventing the use of preparative DNA gels with radioactively labeled fragments. For sequencing five degradation reactions with the following specificity were used: G, A>C, G+A, C+T, C. The resulting degradation products were separated on thermoregulated (50°C), 1-m-long, 0.26-mm-thick 4, 6, and 16% polyacrylamide gels containing 7 M urea (2). The 6% gel contained a buffer step gradient (5). On the average, about 500 bases could be read from a single end-labeled fragment under the conditions used. The sequencing strategy is outlined in Fig. 2.

RESULTS

Subcloning. An approximately 6-kbp *Mlu*I-*Eco*RI fragment thought to contain the genes and sites of the *bgl* operon

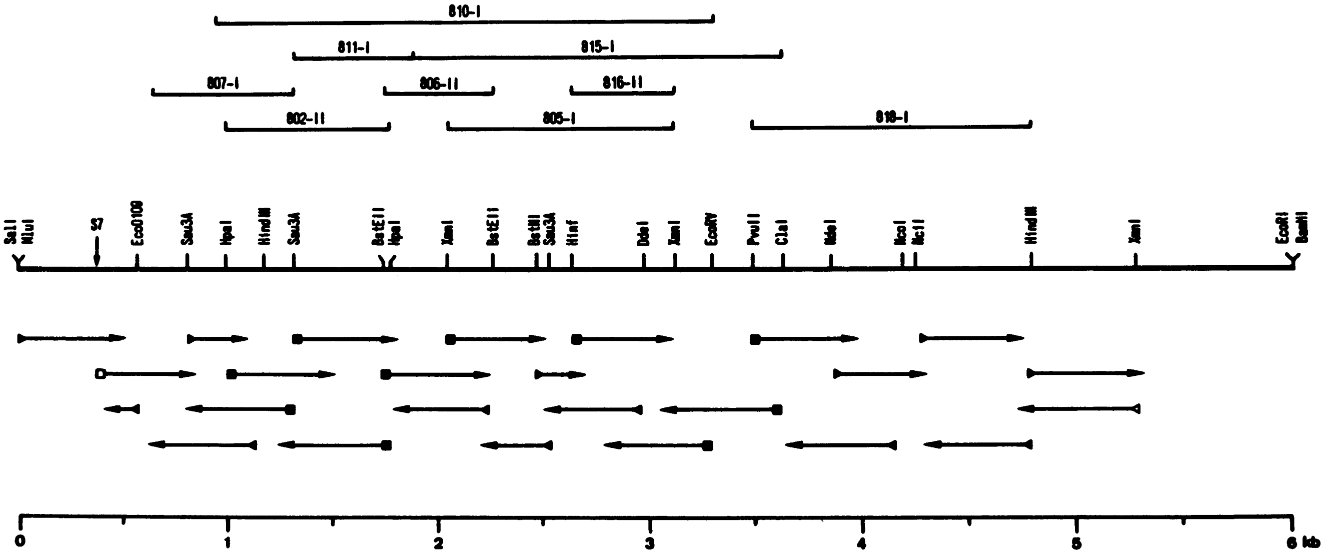

FIG. 2. Relevant restriction map, sequencing strategy, and subclones constructed for sequencing. Only restriction sites used for sequencing or subcloning are given. The integration site of IS5 in mutant *bglR-S7*::IS5-Sal is marked. The lines at the top give the extent of the individual subclones; the numbers refer to the subclones constructed in vector pUC12, and the roman numerals behind the hyphen give the relative orientation of the fragment. The arrows mark the direction and extent of the sequence analysis: ■, pUC12 polylinker site was used for labeling; ▲, restriction site of fragment was used for labeling; △, *Bam*HI site of deletion derivative pFDX750-S7::IS5-Sal (Fig. 5) was used for labeling; □, *Sal*I site within the mutant insertion element in plasmid pFDX733-S7::IS5-Sal (Fig. 5) was used. Overlapping sequences were determined for all parts.

involved in regulated uptake and degradation of aryl-β-glucosides was subcloned in vector plasmid pACYC177, resulting in plasmid pFDX733 (Fig. 1).

Selection of a Bgl$^+$ derivative of plasmid pFDX733 carrying a mutant IS5 element. Strains of *S. typhimurium* do not mutate to Bgl$^+$ (40; data not shown), nor do they contain IS*1* (29) or IS*5* (41; H. Eibel and B. Rak, unpublished observation). *S. typhimurium* SL5235 was transformed with plasmids pFDX733 and pFDX53-Sal as a donor for IS5 (a mutant element carrying a *Sal*I linker inserted 25 bp away from the left end). Double transformants were selected on BS-kanamycin-ampicillin plates. Bgl$^+$ mutants were picked, and the plasmid structures were analyzed. A large proportion of these carried the tagged IS5 element integrated in a small region between about 320 and 380 bp away from the *Mlu* site, previously identified as *bglR* (34, 35). Plasmids isolated from these mutants conferred an (inducible) Bgl$^+$ phenotype upon retransformation into various Bgl$^-$ strains. One such isolate, carrying the *Sal*I site of the mutant IS5 element close to the *bgl* operon genes, was designated pFDX733-S7; the mutation is referred to as *bglR-S7*::IS5-Sal.

Restriction map. Relevant restriction sites within the *Mlu*I-*Eco*RI fragment are given in Fig. 2. The integration site of IS5 in mutant *bglR-S7*::IS5-Sal is indicated by an arrow.

DNA sequence. The nucleotide sequence given in Fig. 3 (5,270 bp) starts at the *Mlu*I site and ends at an *Xmn*I site approximately 740 bp proximal to the *Eco*RI site (Fig. 2). The first 376 bp (48) and the first 586 bp (1) are identical to the 3′ end of the sequence published for gene *phoU*. Thus, our sequence includes the C-terminal part of the *phoU*-gene, which is followed by an inverted repeat structure (Fig. 3). No deviation from the published sequences was found. The sequence at bp 205 to bp 552 is identical to the published sequence of the *bgl* control region (34, 35). At position 458, however, we determined a G instead of an A residue. An A residue at this position would result in a *Mae*I site, and a G residue would result in a *Bst*NI site. Enzyme *Mae*I did not cut at this position, whereas we detected cleavage by *Bst*NI (data not shown).

The cyclic AMP binding protein binding site and the −35 and −10 sequences of the *bgl* promoter (Fig. 3) as well as the transcription start site of the *bgl* mRNA (+1) have been determined (34, 35).

Open reading frames. Aside from the C-terminal sequence of gene *phoU*, three large open reading frames were found, all reading from left to right. These are designated *bglC*, *bglS*, and *bglB* in Fig. 3. The first reading frame has an ATG in its 5′ region (at bp 582), preceded by a translational start signal (Shine-Dalgarno sequence [44]), AGGT, at a distance of 7 bp, suggesting that the first gene of the operon, *bglC*, starts with this ATG. Consequently, *bglC* is preceded by an untranslated leader RNA of 130 bases. The reading frame stops at bp 1415 and thus contains 278 codons corresponding to a protein of M_r 32,067. The second open reading frame, 625 codons long, could code for a protein of M_r 66,400. It starts with ATG at bp 1552 and stops at bp 3426. Here again, the initiator codon is preceded by a translational start signal, GAGG, at a distance of seven bases, suggesting that it is the beginning of gene *bglS*. The intercistronic region between genes *bglC* and *bglS* would then be 136 bases long. A third open reading frame of 471 codons (protein M_r, 53,120), a candidate for gene *bglB*, extends from the ATG at bp 3448 to 4860. This ATG is preceded at a distance of six bases by the translational start signal AGGAG. In this case, the apparent intercistronic region is only 21 bases in length. The first possible start codon (ATG) of a fourth open reading frame, which does not terminate within the sequenced fragment, is found at bp 4932, 72 bp distal to the preceding reading frame. The ATG of this open reading frame is preceded by the sequence AGGGA, which could qualify as a translational start signal.

Codon usage. Many attempts have been made to relate the relative usage of synonymous codons (i.e., the frequency with which the different codons encoding identical amino

```
   1 ACGCGTTCGCGCGGATGGACATTGACGAAGCGGTACGTATTTATCGTGAAGATAAAAAAGTCGATCAGGAATACGAAGGTATTGTTCGTCAACTGATGAC
       AlaPheAlaArgMetAspIleAspGluAlaValArgIleTyrArgGluAspLysLysValAspGlnGluTyrGluGlyIleValArgGlnLeuMetTh
     phoU ———>

 101 CTACATGATGGAAGATTCGCGTACCATTCCGAGCGTACTTACTGCGCTGTTCTGCGCGCGTTCTATCGAACGTATTGGCGACCGCTGCCAGAATATTTGT
     rTyrMetMetGluAspSerArgThrIleProSerValLeuThrAlaLeuPheCysAlaArgSerIleGluArgIleGlyAspArgCysGlnAsnIleCys

 201 GAGTTTATCTTCTACTACGTGAAGGGGCAGGATTTCCGTCACGTCGGTGGCGATGAGCTGGATAAACTGCTGGCGGGGAAAGATAGCGACAAATAATTCA
     GluPheIlePheTyrTyrValLysGlyGlnAspPheArgHisValGlyGlyAspGluLeuAspLysLeuLeuAlaGlyLysAspSerAspLys-*-
                       IR-phoU                                            S7::IS5          CAP binding site
       ——————————>          <——————————
 301 CCAGACAAATCCCAATAACTTAATTATTGGGATTTGTTATATATAACTTTATAAATTCCTAAAATTACACAAAGTTAATAACTGCGAGCATGGTCATATT

                    -35    Pbgl             -10         +1                                              boxA^prox
 401 TTTATCAATAGCGCATTGCTATTTTCTCTGCACGCAATTAAATTAATTTCCGAACCTGGATGTTCGTTATAAAAACCATTAATAAATGACTGGATTGTTA
                                                                                                    MetThrGlyLeuLeu
                 boxB^prox  -35        IR-bglC^prox  -10                         SD           start orf19
               ——————————————————>           <———————————————
 501 CTGCATTCGCAGGCAAAACCTGACATAACCAGAGAATACTGGTGAAGTCGGGTTTTTTTGTTTATAAAAAAGGTCCTTGCTATGAACATGCAAATCACCA
     LeuHisSerGlnAlaLysProAspIleThrArgGluTyrTrp-*-                                    MetAsnMetGlnIleThrL
                                                                               start bglC

 601 AAATTCTCAACAATAATGTTGTGGTGGTTATTGATGATCAACAGCGGGAAAAAGTCGTCATGGGGCGCGGAATTGGCTTTCAAAAACGCGCTGGCGAAAG
     ysIleLeuAsnAsnAsnValValValValIleAspAspGlnGlnArgGluLysValValMetGlyArgGlyIleGlyPheGlnLysArgAlaGlyGluAr

 701 AATTAACTCAAGTGGAATAGAAAAAGAGTATGCCTTGAGCAGTCATGAACTGAACGGGCGATTAAGCGAACTCTTAAGTCATATTCCTCTTGAGGTGATG
     gIleAsnSerSerGlyIleGluLysGluTyrAlaLeuSerSerHisGluLeuAsnGlyArgLeuSerGluLeuLeuSerHisIleProLeuGluValMet

 801 GCAACCTGTGATCGTATTATCTCTTTAGCGCAGGAGCGCTTGGGAAAATTACAGGACAGTATTTATATCTCGCTAACTGACCATTGCCAGTTTGCGATTA
     AlaThrCysAspArgIleIleSerLeuAlaGlnGluArgLeuGlyLysLeuGlnAspSerIleTyrIleSerLeuThrAspHisCysGlnPheAlaIleL

 901 AACGCTTTCAGCAAAACGTGTTGCTGCCCAACCCGTTGCTGTGGGATATCCAGCGGCTTTACCCGAAAGAGTTCCAGCTAGGGGAAGAAGCGTTAACCAT
     ysArgPheGlnGlnAsnValLeuLeuProAsnProLeuLeuTrpAspIleGlnArgLeuTyrProLysGluPheGlnLeuGlyGluGluAlaLeuThrIl

1001 TATTGATAAACGGTTGGGCGTGCAGTTACCGAAAGATGAAGTGGGCTTTATTGCCATGCATCTGGTCAGTGCCCAAATGAGCGGAAATATGGAGGATGTT
     eIleAspLysArgLeuGlyValGlnLeuProLysAspGluValGlyPheIleAlaMetHisLeuValSerAlaGlnMetSerGlyAsnMetGluAspVal

1101 GCAGGTGTCACGCAGTTAATGCGCGAAATGCTGCAATTAATAAAATTTCAGTTCAGCCTTAATTACCAGGAAGAAAGCTTGAGTTATCAGCGACTGGTTA
     AlaGlyValThrGlnLeuMetArgGluMetLeuGlnLeuIleLysPheGlnPheSerLeuAsnTyrGlnGluGluSerLeuSerTyrGlnArgLeuValT

1201 CACATCTGAAGTTTTTATCCTGGCGTATTCTTGAACATGCTTCAATTAACGATAGTGATGAATCATTACAACAAGCAGTAAAACAAAATTACCCGCAAGC
     hrHisLeuLysPheLeuSerTrpArgIleLeuGluHisAlaSerIleAsnAspSerAspGluSerLeuGlnGlnAlaValLysGlnAsnTyrProGlnAl

1301 ATGGCAATGTGCGGAACGGATCGCCATTTTTATTGGTTTGCAGTATCAACGTAAAATTTCACCCGCAGAGATTATGTTTTTAGCCATAAATATAGAGCGC
     aTrpGlnCysAlaGluArgIleAlaIlePheIleGlyLeuGlnTyrGlnArgLysIleSerProAlaGluIleMetPheLeuAlaIleAsnIleGluArg
                                              boxA^dist              boxB^dist              IR-bglC^dist
                                                                   ———————————————————>       <——————
1401 GTGCGCAAAGAACACTGAAATATTATTACTGAGTAAAGGATTGTTACCGCACTAAGCGGGCAAAACCTGAAAAAAATTGCTTGATTCACGTCAGGCCGTT
     ValArgLysGluHis-*-
                                            SD
     ————————
1501 TTTTTCAGGTTTTTTTTTGGAGTTTTGCCGCAAAGCGGTAGAGGGCAAGTTATGACGGAGTTAGCCAGAAAAATAGTCGCAGGAGTCGGGGGCGCAGATA
                                                       MetThrGluLeuAlaArgLysIleValAlaGlyValGlyGlyAlaAspA
                                                  start bglS

1601 ACATTGTGAGTCTGATGCATTGCGCAACGCGATTACGTTTTAAATTAAAGGATGAAAGCAAAGCGCAAGCAGAGGTACTGAAAAAGACCCCCGGTATTAT
     snIleValSerLeuMetHisCysAlaThrArgLeuArgPheLysLeuLysAspGluSerLysAlaGlnAlaGluValLeuLysLysThrProGlyIleIl

1701 TATGGTGGTGGAAAGCGGTGGCCAGTTTCAGGTGGTCATAGGTAACCATGTGGCCGATGTCTTCCTGGCGGTTAACAGTGTGGCAGGCCTTGACGAAAAA
     eMetValValGluSerGlyGlyGlnPheGlnValValIleGlyAsnHisValAlaAspValPheLeuAlaValAsnSerValAlaGlyLeuAspGluLys

1801 GCGCAACAGGCACCGGAAAATGATGATAAAGGTAATCTGCTAAACCGCTTTGTTTATGTTATTTCAGGTATTTTTACGCCTCTGATCGGTTTGATGGCGG
     AlaGlnGlnAlaProGluAsnAspAspLysGlyAsnLeuLeuAsnArgPheValTyrValIleSerGlyIlePheThrProLeuIleGlyLeuMetAlaA

1901 CAACCGGGATCTTGAAAGGTATGCTGGCTCTGGCGCTCACTTTTCAGTGGACGACCGAACAAAGTGGTACTTATTTAATTTTATTCAGCGCCAGTGATGC
     laThrGlyIleLeuLysGlyMetLeuAlaLeuAlaLeuThrPheGlnTrpThrThrGluGlnSerGlyThrTyrLeuIleLeuPheSerAlaSerAspAl

2001 CTTGTTTTGGTTCTTCCCGATAATCCTGGGATACACCGCGGGGAAACGCTTCGGCGGTAATCCATTTACTGCCATGGTGATTGGTGGAGCGTTAGTGCAT
     aLeuPheTrpPhePheProIleIleLeuGlyTyrThrAlaGlyLysArgPheGlyGlyAsnProPheThrAlaMetValIleGlyGlyAlaLeuValHis

2101 CCATTAATTCTGACTGCTTTCGAGAACGGGCAAAAAGCGGATGCGCTGGGGCTGGATTTCCTGGGTATTCCGGTCACATTGTTGAATTACTCGTCATCGG
     ProLeuIleLeuThrAlaPheGluAsnGlyGlnLysAlaAspAlaLeuGlyLeuAspPheLeuGlyIleProValThrLeuLeuAsnTyrSerSerSerV
```

FIG. 3. Nucleotide sequence of the *Mlu*I-*Xmn*I fragment containing *bgl*. See the text for details.

```
2201 TTATTCCCATTATTTTTTCTGCCTGGTTGTGCAGCATTCTGGAACGCCGACTTAATGCGTGGTTACCGTCGGCAATCAAAAATTTCTTCACACCATTGCT
     alIleProIleIlePheSerAlaTrpLeuCysSerIleLeuGluArgArgLeuAsnAlaTrpLeuProSerAlaIleLysAsnPhePheThrProLeuLe

2301 ATGTCTGATGGTTATCACACCCGTCACCTTTCTGCTGGTGGGGCCGCTATCAACCTGGATAAGCGAACTGATTGCCGCCGGTTATCTCTGGCTTTATCAG
     uCysLeuMetValIleThrProValThrPheLeuLeuValGlyProLeuSerThrTrpIleSerGluLeuIleAlaAlaGlyTyrLeuTrpLeuTyrGln

2401 GCGGTTCCTGCATTTGCGGGCGCGGTAATGGGCGGCTTCTGGCAAATCTTCGTCATGTTCGGACTGCACTGGGGCCTGGTGCCGCTGTGTATCAATAACT
     AlaValProAlaPheAlaGlyAlaValMetGlyGlyPheTrpGlnIlePheValMetPheGlyLeuHisTrpGlyLeuValProLeuCysIleAsnAsnP

2501 TCACCGTGCTGGGCTACGACACCATGATCCCGCTGTTAATGCCCGCCATTATGGCGCAGGTCGGGGCGGCGCTCGGCGTCTTCCTCTGCGAACGCGATGC
     heThrValLeuGlyTyrAspThrMetIleProLeuLeuMetProAlaIleMetAlaGlnValGlyAlaAlaLeuGlyValPheLeuCysGluArgAspAl

2601 GCAGAAAAAAGTGGTGGCGGGATCAGCGGCGTTGACGAGTCTGTTTGGTATCACCGAACCAGCGGTATATGGCGTCAACCTGCCGCGTAAGTACCCCTTT
     aGlnLysLysValValAlaGlySerAlaAlaLeuThrSerLeuPheGlyIleThrGluProAlaValTyrGlyValAsnLeuProArgLysTyrProPhe

2701 GTTATCGCCTGTATCAGTGGGGCTTTGGGGGCCACCATTATTGGCTACGCGCAAACGAAAGTCTACTCCTTTGGTTTGCCAAGTATTTTCACCTTCATGC
     ValIleAlaCysIleSerGlyAlaLeuGlyAlaThrIleIleGlyTyrAlaGlnThrLysValTyrSerPheGlyLeuProSerIlePheThrPheMetG

2801 AAACCATCCCGTCAACGGGAATTGATTTCACCGTCTGGGCCAGCGTTATTGGCGGTGTCATTGCCATCGGTTGCGCATTTGTCGGTACGGTGATGCTTCA
     lnThrIleProSerThrGlyIleAspPheThrValTrpAlaSerValIleGlyGlyValIleAlaIleGlyCysAlaPheValGlyThrValMetLeuHi

2901 TTTCATCACCGCTAAACGTCAGCCAGCGCAGGGTGCCCCGCAAGAGAAAACACCAGAGGTTATTACACCACCTGAGCAGGGCGGTATCTGTTCACCGATG
     sPheIleThrAlaLysArgGlnProAlaGlnGlyAlaProGlnGluLysThrProGluValIleThrProProGluGlnGlyGlyIleCysSerProMet

3001 ACGGGAGAGATTGTGCCGCTCATTCACGTCGCTGATACCACGTTTGCCAGTGGCCTGTTGGGTAAAGGTATTGCCATTCTGCCCTCGGTTGGTGAAGTGC
     ThrGlyGluIleValProLeuIleHisValAlaAspThrThrPheAlaSerGlyLeuLeuGlyLysGlyIleAlaIleLeuProSerValGlyGluValA

3101 GTTCTCCGGTTGCGGGTCGAATTGCTTCGTTGTTCGCCACATTACACGCCATTGGCATTGAGTCAGATGATGGTGTGGAGATCCTGATTCATGTCGGTAT
     rgSerProValAlaGlyArgIleAlaSerLeuPheAlaThrLeuHisAlaIleGlyIleGluSerAspAspGlyValGluIleLeuIleHisValGlyIl

3201 CGACACCGTAAAACTGGACGGCAAATTCTTTTCCGCTCACGTCAACGTGGGTGACAAGGTCAATACAGGCGATCGGCTGATTTCTTTTGATATCCCTGCT
     eAspThrValLysLeuAspGlyLysPhePheSerAlaHisValAsnValGlyAspLysValAsnThrGlyAspArgLeuIleSerPheAspIleProAla

3301 ATTCGCGAGGCCGGATTTGATCTGACGACGCCGGTATTAATCAGTAATAGCGATGATTTTACGGACGTATTACCCCACGGCACGGCGCAGATAAGCGCAG
     IleArgGluAlaGlyPheAspLeuThrThrProValLeuIleSerAsnSerAspAspPheThrAspValLeuProHisGlyThrAlaGlnIleSerAlaG
                                            SD
                                         ─────
3401 GTGAACCGCTGTTATCCATCATTCGCTAACGATAAAAGGAGTTAATTATGAAAGCATTTCCAGAAACATTTCTTTGGGGTGGCGCAACAGCTGCCAATCA
     lyGluProLeuLeuSerIleIleArg-*-                  MetLysAlaPheProGluThrPheLeuTrpGlyGlyAlaThrAlaAlaAsnGl
                                              start bglB
3501 GGTGGAAGGTGCCTGGCAGGAAGATGGCAAAGGGATCTCGACCTCAGATTTACAGCCTCATGGCGTAATGGGAAAAATGGAACCGCGCATCCTGGGGAAA
     nValGluGlyAlaTrpGlnGluAspGlyLysGlyIleSerThrSerAspLeuGlnProHisGlyValMetGlyLysMetGluProArgIleLeuGlyLys

3601 GAGAATATCAAAGATGTCGCCATCGATTTTTATCACCGTTACCCGGAAGATATCGCGTTATTTGCCGAGATGGGCTTCACCTGTCTGCGTATTTCCATTG
     GluAsnIleLysAspValAlaIleAspPheTyrHisArgTyrProGluAspIleAlaLeuPheAlaGluMetGlyPheThrCysLeuArgIleSerIleA

3701 CCTGGGCGCGAATTTTCCCTCAGGGCGACGAAGTCGAACCGAATGAAGCGGGGTTAGCGTTTTACGATCGGCTGTTTGATGAAATGGCGCAGGCGGGGAT
     laTrpAlaArgIlePheProGlnGlyAspGluValGluProAsnGluAlaGlyLeuAlaPheTyrAspArgLeuPheAspGluMetAlaGlnAlaGlyIl

3801 CAAGCCGCTGGTAACGTTATCCCATTACGAAATGCCATATGGGCTGGTGAAAAACTACGGCGGTTGGGCTAATCGAGCGGTCATCGGTCACTTCGAGCAT
     eLysProLeuValThrLeuSerHisTyrGluMetProTyrGlyLeuValLysAsnTyrGlyGlyTrpAlaAsnArgAlaValIleGlyHisPheGluHis

3901 TACGCCCGCACGGTCTTTACTCGCTACCAACATAAAGTGGCGTTATGGCTGACGTTTAATGAAATCAACATGTCGTTACACGCGCCATTCACGGGCGTGG
     TyrAlaArgThrValPheThrArgTyrGlnHisLysValAlaLeuTrpLeuThrPheAsnGluIleAsnMetSerLeuHisAlaProPheThrGlyValG

4001 GGCTGGCAGAAGAGAGTGGCGAGGCGGAAGTTTATCAGGCTATCCACCATCAACTGGTTGCCAGTGCGCGGGCAGTTAAAGCCTGTCATAGCCTGCTCCC
     lyLeuAlaGluGluSerGlyGluAlaGluValTyrGlnAlaIleHisHisGlnLeuValAlaSerAlaArgAlaValLysAlaCysHisSerLeuLeuPr

4101 CGAAGCGAAAATCGGCAATATGCTTCTCGGTGGGCTCGTTTACCCCCTCACCTGCCAGCCACAGGATATGTTGCAGGCCATGGAAGAGAACCGGCGCTGG
     oGluAlaLysIleGlyAsnMetLeuLeuGlyGlyLeuValTyrProLeuThrCysGlnProGlnAspMetLeuGlnAlaMetGluGluAsnArgArgTrp

4201 ATGTTCTTTGGTGATGTTCAGGCGCGTGGCCAGTATCCCGGCTATATGCAGCGTTTCTTCCGCGACCACAATATCACCATTGAGATGACTGAAAGTGACG
     MetPhePheGlyAspValGlnAlaArgGlyGlnTyrProGlyTyrMetGlnArgPhePheArgAspHisAsnIleThrIleGluMetThrGluSerAspA

4301 CAGAAGATTTAAAACATACCGTCGATTTCATCTCTTTTAGTTATTACATGACTGGTTGTGTTTCCCACGACGAAAGCATTAATAAAAATGCGCAGGGCAA
     laGluAspLeuLysHisThrValAspPheIleSerPheSerTyrTyrMetThrGlyCysValSerHisAspGluSerIleAsnLysAsnAlaGlnGlyAs
```

Continued on next page

```
4401 CATACTGAATATGATCCCCAATCCGCATCTGAAAAGTTCAGAGTGGGGGTGGCAAATTGATCCGGTTGGATTACGGGTTCTGTTAAATACGCTTTGGGAT
     nIleLeuAsnMetIleProAsnProHisLeuLysSerSerGluTrpGlyTrpGlnIleAspProValGlyLeuArgValLeuLeuAsnThrLeuTrpAsp

4501 CGTTATCAAAAACCGTTATTTATTGTCGAGAACGGATTAGGCGCAAAAGACAGCGTTGAAGCGGATGGTTCGATACAGGACGATTATCGAATTGCCTATT
     ArgTyrGlnLysProLeuPheIleValGluAsnGlyLeuGlyAlaLysAspSerValGluAlaAspGlySerIleGlnAspAspTyrArgIleAlaTyrL

4601 TAAACGATCACCTGGTACAGGTAAATGAAGCGATTGCCGATGGTGTGGATATTATGGGGTACACCAGTTGGGGGCCAATTGATTTAGTCAGTGCATCTCA
     euAsnAspHisLeuValGlnValAsnGluAlaIleAlaAspGlyValAspIleMetGlyTyrThrSerTrpGlyProIleAspLeuValSerAlaSerHi

4701 TTCACAAATGTCTAAGCGCTACGGCTTTATTTATGTGGATCGTGATGATAATGGCGAAGGAAGCCTCACAAGAACACGCAAGAAAAGCTTTCGGATGGTA
     sSerGlnMetSerLysArgTyrGlyPheIleTyrValAspArgAspAspAsnGlyGluGlySerLeuThrArgThrArgLysLysSerPheArgMetVal
                                                                                    IR-bglB
4801 TGCGCAGAGGTGATCAAGACGCGGGGGCTGTCATTAAAAAAAATAACCATTAAAGCACCTTAATTATCGTCGCATTCAGAACAGTCTGGATGCGATGCGT
     CysAlaGluValIleLysThrArgGlyLeuSerLeuLysLysIleThrIleLysAlaPro-c-
                           SD
4901 TAATTCTTTCTTTGCACCATAAAGGGATATTATGTTTAGACGAAATCTTATTACCTCTGCCATCTTATTAATGGCACCGTTAGCCTTTTCTGCACAATCA
                                    MetPheArgArgAsnLeuIleThrSerAlaIleLeuLeuMetAlaProLeuAlaPheSerAlaGlnSer
                               start orf
5001 TTGGCTGAATCATTAACGGTGGAACAACGCCTTGAGTTATTAGAAAAGGCGTTAAGAGAAACGCAAAGCGAACTCAAAAAGTATAAAGATGAAGAGAAGA
     LeuAlaGluSerLeuThrValGluGlnArgLeuGluLeuLeuGluLysAlaLeuArgGluThrGlnSerGluLeuLysLysTyrLysAspGluGluLysL

5101 AAAAGTATACGCCAGCGACGGTGAATCGTAGCGTAAGTACGAATGATCAAGGGTATGCCGCCAATCCGTTCCCGACCAGTAGTGCCGCAAAACCTGATGC
     ysLysTyrThrProAlaThrValAsnArgSerValSerThrAsnAspGlnGlyTyrAlaAlaAsnProPheProThrSerSerAlaAlaLysProAspAl

5201 TGTACTGGTCAAAAATGAAGAGAAAAATGCCAGTGAGACAGGCTCGATTTATTCTTCCATGACTCTGAAA
     aValLeuValLysAsnGluGluLysAsnAlaSerGluThrGlySerIleTyrSerSerMetThrLeuLys
```

acids are used) to the rate of expression of individual genes. Codon usage has been discussed as a factor influencing translational fidelity and thus as a means to determine whether a gene belongs to the highly or weakly expressed class (13, 15). For comparisons of relative synonymous codon usage of the *bgl* operon we chose the most recent investigation of this issue (42), based on the most extensive compilation of genes to date. We compared the codon usage of the open reading frames found in the *bgl* operon with that of the two extreme groups, containing genes expressed at a high level on the one hand and moderately and weakly expressed genes on the other hand. This latter class theoretically contains genes not subject to selection for efficient translation (42). Table 1 gives the relative synonymous codon usage values (42), which are the observed frequency of a codon divided by the expected frequency, assuming that all codons for any particular amino acid are used equally.

It is apparent that the relative synonymous codon usage values for the three reading frames (*bglC*, *bglS*, and *bglB* in Table 1) are closely related to the low-bias group, indicating that translation of the genes belonging to the *bgl* operon may be quite low.

Hydropathy. One of the genes of the *bgl* operon (*bglC* in references 31 and 33) codes for a specific transport protein. The corresponding polypeptide thus should contain typical hydrophobic transmembrane domains. We therefore analyzed the primary amino acid sequences deduced from the nucleotide sequence by using the data of Kyte and Doolittle (16). Whereas the hypothetical proteins encoded by *bglC* and *bglB* in Fig. 3 are soluble and hydrophilic with no major hydrophobic domains (data not shown), the one encoded by *bglS* shows a hydropathy pattern characteristic of a transmembrane protein (potentially spanning the membrane several times) and contains an intermediate to hydrophilic N-terminal and C-terminal part (Fig. 4A). The *bgl* transport protein belongs to the group of phosphotransferase system-coupled transport systems, whose members phosphorylate the substrate concomitantly with the transport process (11, 30, 31, 36, 39). We therefore compared the hydropathy patterns of the product of *bglS* with the mannitol-specific transport protein, enzyme II^{Mtl}, the only other enzyme II (i.e., phosphotransferase system-coupled transport protein) of known protein sequence (17). Interestingly, the size of this protein (637 amino acids) is almost identical to that of the protein encoded by *bglS* (625 amino acids). In Fig. 4 the pattern of the mannitol-specific protein has been aligned to give maximal match with the pattern of the *bglS*-encoded protein. The comparison revealed striking similarities between them, the main difference being in their termini. Whereas enzyme II^{Mtl} has a relatively hydrophilic tail of about 280 amino acids, the *bglS*-encoded protein has an N-terminal stretch of about 100 amino acids and a C-terminal tail of about 180 amino acids showing a considerably less hydrophobic pattern than the core.

Functional assignment of the *bgl* genes. A second gene coding for a phospho-β-glucosidase (*bglA*) is present on the chromosome of *E. coli*. This gene is expressed constitutively (31, 40). The enzyme encoded by *bglA* is specific for arbutin, whereas the enzyme encoded by *bglB* hydrolizes salicin as well as arbutin. There is, however, only one transport system supplying both enzymes, and this is encoded by the *bgl* operon. To map the genetic functions of *bgl* we constructed several deletions of Bgl$^+$ mutant S7::IS*5*-Sal affecting the distal part of the operon and tested them for the phenotypes they conferred on *E. coli* JF201 (which is Δ *bgl*, but contains *bglA*) and on *S. typhimurium* (lacking *bgl* genes [40]). Bacterial strains containing the various deletion plasmids were streaked out on BS and BA plates containing kanamycin, and the phenotypes were scored (Fig. 5). A deletion removing the distal part (ca. 740 bp) of the insert (plasmid pFDX750-S7::IS*5*-Sal) is phenotypically indistinguishable from pFDX733-S7::IS*5*-Sal. A *bglB* deletion (pFDX751-S7::IS*5*-Sal) is salicin and arbutin negative in *Salmonella* sp. and salicin negative and arbutin positive in *E. coli* JF201. These phenotypes indicate that *bglB* codes for phospho-β-glucosidase B. Deletions extending into *bglS*

TABLE 1. Relative synonymous codon usage

Amino acid	Codon	Relative synonymous codon usage[a]				
		Standard[b]		Genes of this paper[c]		
		High	Low	*BglC*	*BglS*	*BglB*
Ala	GCA	1.10	0.74	1.25	0.70	0.92
	GCC	0.23	1.24	1.25	1.21	1.13
	GCG	0.80	1.49	1.00	1.59	1.64
	GCU	1.88	0.53	0.50	0.51	0.31
		(104)	(106)	(57)	(101)	(82)
Arg	AGA	0.02	0.13	0.35	0.40	0.26
	AGG	0.00	0.09	0.00	0.00	0.00
	CGA	0.02	0.29	0.71	1.20	0.78
	CGC	1.56	2.76	2.47	2.40	1.83
	CGG	0.02	0.57	1.41	0.40	1.56
	CGU	4.39	2.17	1.06	1.60	1.56
		(57)	(56)	(61)	(25)	(49)
Asn	AAC	1.90	1.09	1.08	0.94	0.60
	AAU	0.10	0.91	0.92	1.06	1.40
		(43)	(42)	(47)	(27)	(43)
Asp	GAC	1.39	0.72	0.36	0.50	0.43
	GAU	0.61	1.28	1.64	1.50	1.57
		(63)	(51)	(39)	(39)	(60)
Cys	UGC	1.33	1.21	0.67	1.00	0.80
	UGU	0.67	0.79	1.33	1.00	1.20
		(5)	(11)	(11)	(12)	(10)
Gln	CAA	0.22	0.66	0.96	0.84	0.50
	CAG	1.78	1.34	1.04	1.16	1.50
		(36)	(48)	(90)	(31)	(43)
Glu	GAA	1.59	1.37	1.36	1.05	1.35
	GAG	0.41	0.63	0.64	0.95	0.65
		(74)	(59)	(79)	(34)	(66)
Gly	GGA	0.02	0.33	1.23	0.51	0.41
	GGC	1.65	1.74	1.23	1.21	1.54
	GGG	0.04	0.59	0.92	0.57	1.13
	GGU	2.28	1.34	0.62	1.71	0.92
		(88)	(76)	(46)	(100)	(82)
His	CAC	1.55	0.86	0.29	1.00	0.88
	CAU	0.45	1.14	1.71	1.00	1.12
		(16)	(23)	(26)	(16)	(34)
Ile	AUA	0.01	0.12	0.46	0.26	0.30
	AUC	2.53	1.24	0.58	1.03	1.40
	AUU	0.47	1.64	1.96	1.71	1.30
		(78)	(56)	(93)	(93)	(64)
Leu	CUA	0.04	0.18	0.36	0.26	0.00
	CUC	0.20	0.64	0.36	0.44	0.65
	CUG	5.33	3.12	1.27	2.65	2.43
	CUU	0.22	0.54	0.73	0.35	0.49
	UUA	0.11	0.74	2.00	1.15	2.27
	UUG	0.11	0.79	1.27	1.15	0.16
		(73)	(104)	(118)	(109)	(78)
Lys	AAA	1.60	1.51	1.86	1.64	1.64
	AAG	0.40	0.49	0.14	0.36	0.36
		(75)	(39)	(51)	(35)	(46)
Met	AUG	(21)	(26)	(36)	(24)	(38)
Phe	UUC	1.54	0.89	0.40	1.03	0.76
	UUU	0.46	1.11	1.60	0.97	1.24
		(33)	(38)	(36)	(62)	(45)

Continued

TABLE 1—*Continued*

Amino acid	Codon	Relative synonymous codon usage[a]				
		Standard[b]		Genes of this paper[c]		
		High	Low	*BglC*	*BglS*	*BglB*
Pro	CCA	0.44	0.75	0.00	0.94	1.05
	CCC	0.04	0.52	1.14	0.82	0.84
	CCG	3.29	2.19	2.29	1.76	1.47
	CCU	0.23	0.55	0.57	0.47	0.63
		(33)	(42)	(25)	(54)	(40)
Ser	AGC	1.05	1.93	1.58	1.37	1.20
	AGU	0.22	0.87	2.21	1.54	1.68
	UCA	0.20	0.59	1.26	1.20	0.96
	UCC	1.91	0.83	0.32	0.51	0.72
	UCG	0.04	0.95	0.32	0.86	0.72
	UCU	2.57	0.83	0.32	0.51	0.72
		(44)	(61)	(69)	(56)	(52)
Thr	ACA	0.14	0.48	0.67	0.68	0.80
	ACC	1.87	1.78	2.00	1.56	1.40
	ACG	0.18	1.13	0.67	1.37	1.20
	ACU	1.80	0.62	0.67	0.39	0.60
		(59)	(50)	(23)	(65)	(42)
Trp	UGG	(5)	(12)	(11)	(14)	(21)
Tyr	UAC	1.61	0.82	0.86	1.09	1.00
	UAU	0.39	1.18	1.14	0.91	1.00
		(27)	(23)	(25)	(18)	(42)
Val	GUA	1.11	0.60	0.25	0.45	0.71
	GUC	0.15	0.89	1.00	1.36	1.00
	GUG	0.50	1.53	1.75	1.36	1.00
	GUU	2.24	0.98	1.00	0.83	1.29
		(87)	(71)	(57)	(86)	(60)

[a] Values given are observed frequency of a codon divided by the expected frequency if all codons for any particular amino acid are used equally (42). Total occurrence of each amino acid (per thousand) is also given within parentheses.

[b] Values for standard are taken from reference (42); "high" is the group defined as highly expressed genes (15 genes with a total of 9,223 codons), and "low" is the low codon bias group (58 genes, 22612 codons) (42).

[c] Genes *bglC* (278 codons), *bglS* (625 codons), and *bglB* (471 codons).

(plasmids pFDX752-S7::IS*5*-Sal and pFDX753-S7::IS*5*-Sal) are negative for salicin and arbutin in *E. coli* as well as in *Salmonella* sp., indicating that *bglS*, alone or in conjunction with *bglC*, is required for the uptake of arbutin. To delimit the minimum requirement for transport of substrate we inserted gene *bglS* downstream of the *lac* operator-promoter, transformed this plasmid into strain JF201 harboring plasmid pFDX99 (directing overproduction of the *lac* repressor), and screened the phenotype. JF201 was arbutin positive only when the *lac* promoter was induced (Fig. 5, plasmid pFDX841) supporting the conclusion, drawn from hydropathy analysis, that gene *bglS* codes for the transport protein mediating accumulation of the different aryl-β-glucosides. On the basis of these results we assume that *bglC* codes for the regulatory protein.

Potential signal sequences. The evidence presented above suggests that gene *bglC* codes for the regulatory protein, *bglS* codes for the transport protein, and *bglB* codes for phospho-β-glucosidase B. The functioning of the *bglC* gene product in the regulation of the *bgl* operon was further substantiated and analyzed (Schnetz and Rak, in preparation). The arrangement with the regulatory gene at the head of the operon is unusual. Are there any structures in the primary sequence that shed light on the mechanism of

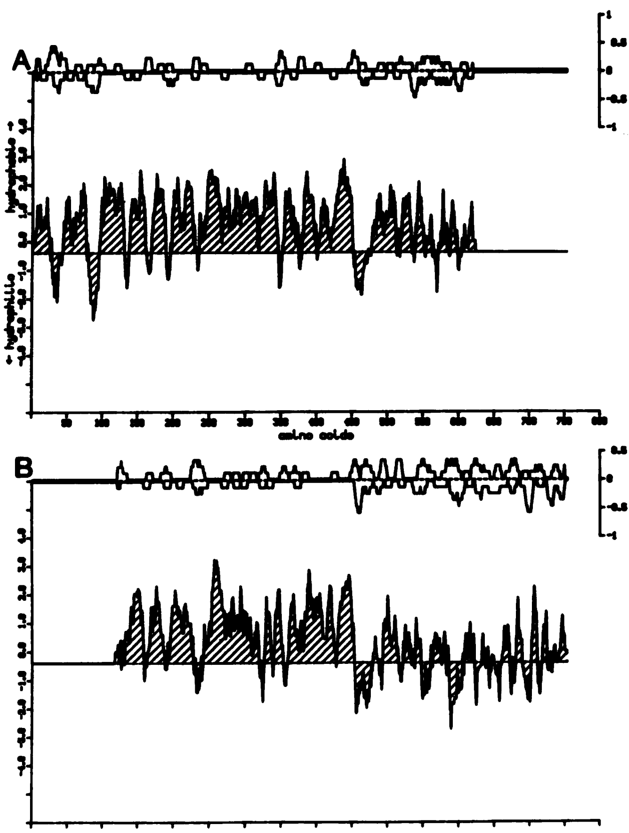

FIG. 4. Hydropathy plot using the standard parameters of Kyte and Doolittle (16). (A) Product of gene *bglS* as deduced from the DNA sequence. (B) enzyme II^{Mtl} (17). The window size was 9 amino acids. Relative charge distributions are given at the top of each plot with the same window.

regulation? The regulatory gene is preceded by a rather long leader and followed by an intercistronic region, which again is unusually long. When we investigated both of these untranslated regions we found a potential stem-loop structure within each of them (Fig. 3). Both are followed by oligo(dT) sequences typical of rho-independent terminators (37). The free energy of formation of stem-loop structures (ΔG) can be calculated as -21 and -26 kcal/mol (ca. -87.9 and -108.8 kJ/mol, respectively) for the first and second inverted repeats, respectively, when formed as RNA (50). Within the leader, an open reading frame of 19 amino acids which terminates within the potential terminator, and thus could interfere with its functioning if expressed, can be found. Expression of this possible leader peptide, however, seems unlikely, because no decent translation start signal is present. When we aligned the leader sequence with that of the intercistronic region, we found two sequence boxes of extensive homology (Fig. 3 and 6). Box A is located at the foot of the potential terminator stems, and box B reaches well into the stem structures. This suggests that these structures play a role in the regulation of the operon. When we screened the sequence for possible promoters in addition to the one mapped previously (34, 35), we recognized a sequence within the first stem-loop structure, which could qualify for a promoter (-35 and -10 in Fig. 3). An additional potential stem-loop forming structure (ΔG of -16 kcal/mol [ca. -66.9 kJ/mol]), which is followed by a stretch of T residues, is located between the C terminus of gene *bglB* and the start of the open reading frame (Fig. 3). This sequence is suggestive of a signal terminating the *bgl* operon. However, our own preliminary evidence (derived from in-frame fusions in *lacZ*) suggests that the open reading frame is, at least in part, expressed coordinately with the *bgl* genes (data not shown).

Homology to other systems. A search of the EMBO Sequence Data Bank yielded some interesting homologies. Highly significant is the occurrence of homologous sequences proximal to a gene of *B. subtilis* coding for β-endoglucanase, an excreted endohydrolase degrading mixed-linked polymers of the type 1,3-1,4-β-D-glucan (6, 27) (Fig. 6 and 7). Coincidentally, this *B. subtilis* gene is also called *bgl*. The gene for β-endoglucanase is preceded by a stem-loop structure. This structure overlaps a block of obvious homology to box B at the same position as the *bgl* hairpin. Lying 5′ is a sequence highly homologous to box A. Thus, a motif found twice within the *bgl* operon is present proximal to the β-glucanase gene of *B. subtilis*. The stretch of sequence preceding the β-glucanase gene and the box A-box B motif contains 85 codons of the 3′ end of an unidentified open reading frame. Alignment of the amino acid sequence of this open reading frame with the C-terminal 85 amino acids of gene *bglC* showed a homology of 38% (54%, allowing exchanges for functionally similar amino acids; Fig. 7). Significantly, no homology on the level of DNA sequence could be detected between these two coding regions.

Particularly noteworthy is the occurrence of homology to the box A-box B motif in the control region of another *B. subtilis* gene, *sacB*, which codes for an excreted levansucrase (45). Again box A-box B can be found at the same relative position to a stem-loop structure (Fig. 6). In this case the site has been shown to function as a transcriptional terminator involved in regulation of the *sacB* gene (43). Homology between the *B. subtilis* genes *bgl* and *sacB* in this region has been noted previously (3).

DISCUSSION

The 5,270 bp of sequence presented extends the longest uninterrupted block of nucleotide sequence data presently available for the *E. coli* chromosome to a total length of >25 kbp. The sequenced region (from 83.4 to 84.1 min on the *E. coli* genetic map [4]) includes *asnA*, *oriC*, *gidA*, *gidB*, the nine genes of the *unc* operon, *glmS*, *phoS*, *phoW*, *pstA*, *pstB*, *phoU*, and three genes of the *bgl* operon (*bglC*, *bglS*, and *bglB*) (1, 7, 18, 19, 24, 28, 46–48, 51). Interestingly, all of the genes up to the origin of replication (*oriC*) are transcribed counterclockwise.

The *bgl* operon is known to contain three structural genes, designated *bglB*, *bglS*, and *bglC*, which code for a phospho-β-glucosidase, a phosphotransferase system-coupled transport protein, and a positive regulator protein, respectively (31). The nucleotide sequence of the sequenced segment spans a region sufficient for regulated utilization of aryl-β-glucosides. It shows three tandemly arranged large open reading frames, each with an ATG codon at the 5′ end. Translational start sequences precede each of the start codons at a distance of six to seven bases, suggesting that the above three genes of the operon do indeed start at the positions indicated in Fig. 3. Assuming that this is the case, proteins of 278 amino acids (M_r 32,067), 625 amino acids (M_r 66,400), and 471 amino acids (M_r 53,120) can be deduced for the gene products of *bglC*, *bglS*, and *bglB*, respectively.

Previous studies have shown that gene *bglB* of the *bgl* operon encodes phospho-β-glucosidase B, an enzyme which

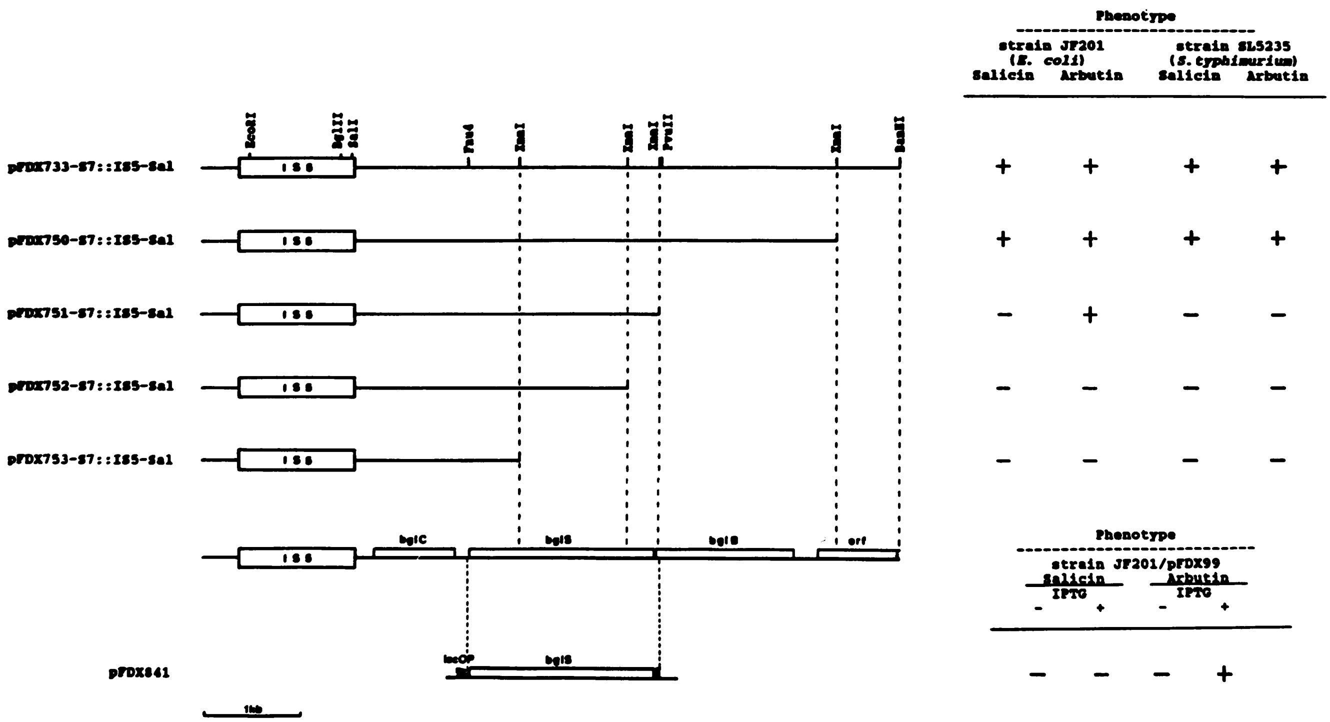

FIG. 5. Functional mapping of the *bgl* operon genes. Phenotypes were scored with BA and BS plates and MacConkey salicin and arbutin plates. Strain JF201 carries a deletion removing the *bgl* operon. Construction of the different deletion derivatives was as follows: plasmid pFDX733-S7::IS5-Sal was partially digested with *Xmn*I, and the linearized form was isolated from a gel and digested with *Bam*HI. After polymerase treatment (Klenow large fragment) the DNA was loaded onto a gel, and the DNA fragments were separately eluted and ligated. Strain R1068 was transformed, and transformants were selected on LB-kanamycin plates. Plasmid pFDX841 contains gene *bglS* cloned as an *Fnu*4HI-*Pvu*II fragment (positions 1529 to 3491 in Fig. 3) into the *Eco*RI site of plasmid vector pUC12. Details of the construction will be published elsewhere. Expression of gene *bglS* in this plasmid is controlled by the *lac* operator-promoter (lacOP). The compatible plasmid pFDX99 codes for the *lac* repressor. Isopropyl-β-D-thiogalactoside (IPTG; 2 mM) was used for induction.

catalyzes the hydrolysis of phosphorylated salicin as well as arbutin. Another gene of *E. coli*, *bglA*, codes for an enzyme that hydrolyzes arbutin but not salicin. Expression of *bglA* is constitutive (31, 40). The import of both substrates is mediated by the same transport protein encoded by the *bgl* operon (31, 40). Strains of *S. typhimurium* encode neither an aryl-β-glucosidase nor a corresponding transport protein (40). Transformation of *S. typhimurium* with a plasmid-borne *bgl* operon deleted for the distal region including *bglB* (plasmid pFDX751-S7::IS5-Sal in Fig. 5) resulted in Bgl⁻ cells, i.e., salicin and arbutin negative. When the same deleted plasmid was introduced into *E. coli* deleted for the chromosomal *bgl* operon, the cells were salicin negative but arbutin positive (Fig. 5), indicating that *bglB* does in fact code for phospho-β-glucosidase B. This result is in accordance with the mapping data reported previously (31) as well as with a more recent map of the *bgl* genes (4, 33).

As to the genes coding for the regulator protein and the transport protein, the functional assignment must be exchanged. Hydropathy analysis of the three hypothetical proteins revealed that the product of gene *bglS* alone has the characteristics of a membrane-spanning transport protein (Fig. 4A). Moreover, comparison of the hydropathy plots of the *bglS* gene product and the mannitol-specific enzyme II transport protein reveals extensive similarity (Fig. 4B), suggesting that the second gene of the operon codes for the aryl-β-glucoside-specific transport protein (enzyme II^{Bgl}). Evidence supporting this assignment was obtained from plasmid deletion mutants and by a gene fusion. Transformation of an *E. coli* Δ*bgl* strain with plasmids deleted for all of *bglB* and part of *bglS* (plasmids pFDX752-S7::IS5-Sal and pFDX753-S7::IS5-Sal in Fig. 5) resulted in a salicin-negative, arbutin-negative phenotype, whereas a plasmid carrying the *bglS* structural gene under the control of the *lac* promoter-operator (plasmid pFDX841 in Fig. 5) gave an arbutin-positive phenotype with the *lac* promoter in the

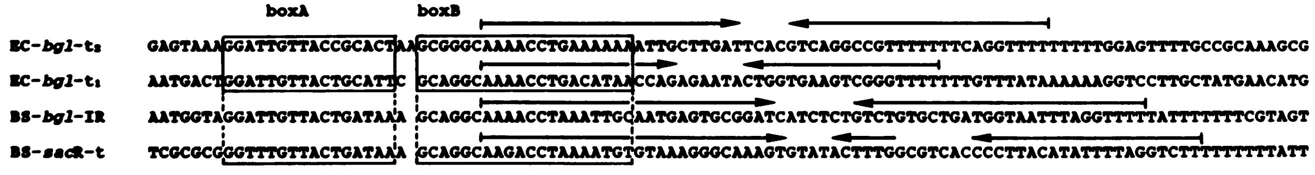
boxA
boxB
EC-bgl-t₂ GAGTAAAGGATTGTTACCGCACTAAGCGGGCAAAACCTGAAAAAAATTGCTTGATTCACGTCAGGCCGTTTTTTTCAGGTTTTTTTTTGGAGTTTTGCCGCAAAGCG
EC-bgl-t₁ AATGACTGGATTGTTACTGCATTC GCAGGCAAAACCTGACATAACCAGAGAATACTGGTGAAGTCGGGTTTTTTTGTTTATAAAAAAGGTCCTTGCTATGAACATG
BS-bgl-IR AATGGTAGGATTGTTACTGATAAA GCAGGCAAAACCTAAATTGCAATGAGTGCGGATCATCTCTGTCTGTGCTGATGGTAATTTAGGTTTTTATTTTTTTCGTAGT
BS-sacR-t TCGCGCGGGTTTGTTACTGATAAA GCAGGCAAGACCTAAAATGTGTAAAGGGCAAAGTGTATACTTTGGCGTCACCCCTTACATATTTTAGGTCTTTTTTTTTATT

FIG. 6. DNA sequence comparison of the box A-box B-terminator motifs. Ec-bglC^{dist}, Motif seen distal to *bglC* (Δ*G*, −26 kcal [ca. −108.8 kJ]); EC-bglC^{prox}; motif seen proximal to gene *bglC* (Δ*G*, −21 kcal [ca. −87.9 kJ]); BS-bgl, corresponding sequence from the *B. subtilis* gene *bgl*(Δ*G*, −17 kcal [ca. −71.1 kJ]); BS-sacR, sequence from the leader of *B. subtilis* gene *sacB* (Δ*G*, −39 kcal [ca. −158.2 kJ]). Arrows indicate inverted repeats. The Δ*G*s of inverted repeats (if formed as mRNA) were calculated as described previously (50).

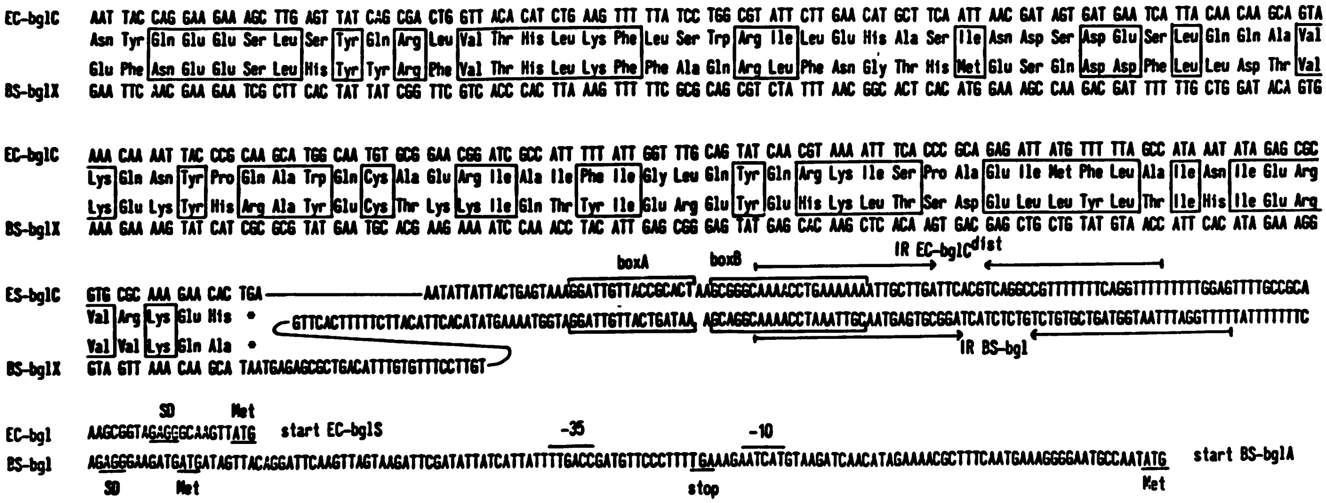

FIG. 7. Alignment of the C-terminal part of *bglC* and the intercistronic region with the *B. subtilis* sequence proximal to the β-endoglucanase gene. The −35 and −10 motif of a hypothetical promoter (27) is indicated. Also marked is an ATG in the *B. subtilis* sequence located 3′ of the box A-box B motif, which can be found at an almost identical relative position as the probable start codon of *bglS*. As in the case of *bglS* it is preceded at a distance of seven nucleotides by the identical Shine-Dalgarno sequence GAGG. The corresponding reading frame, however, terminates after 22 codons. EC-bglC, Distal portion of gene *bglC* and the corresponding amino acid sequence; BS-bglX, sequence proximal to *B. subtilis* β-glucanase gene and its hypothetical translation product; BS-bglA, *B. subtilis* β-glucanase gene. Identical and homologous amino acids are boxed. Homologous amino acids are as follows (12, 22): (i) Lys, Arg, and His; (ii) Asp and Glu; (iii) Asn and Gln; (iv) Ile, Leu, Val, and Met; (v) Ser and Thr; (vi) Phe, Trp, and Tyr; (vii) Ala and Gly. For other details, see the text.

induced state and was arbutin negative with the *lac* promoter uninduced.

It has been speculated that the different enzyme II proteins may have evolved by duplication and subsequent mutational diversification from an ancestral fusion of genes for a porin-like protein and a phosphoenolpyruvate-accepting molecule (14, 30, 38). We have compared gene *mtlA* with gene *bglS* on the DNA as well as protein level and were unable to detect any clear homology. This indicates that, if the above hypothesis is true, either there must be more than one ancestor or the common root of genes *mtlA* and *bglS* lies too far back to be easily recognized.

Comparison of codon usages of the *bgl* genes and of highly expressed genes on the one hand and, at the other extreme, of the low-bias group as defined previously (42) revealed that selection among codons encoding identical amino acids is more related to the latter group for all three genes (Table 1). This seems to indicate that translation of all three genes is relatively poor. Preliminary expression studies with the minicell system support this interpretation (K. Fuchs and B. Rak, unpublished data).

Where is the 3′ boundary of the operon? The functions of the *bgl* operon sufficient for regulated uptake and degradation of aryl-β-glucosides occupy 4,493 bp extending from the cyclic AMP binding protein binding site of *bglR* to and including the translation stop codon of *bglB*. Distal to gene *bglB* and separated by only 68 bp from its translational stop signal is an ATG preceded by a Shine-Dalgarno sequence. This open reading frame does not terminate within the segment of DNA sequenced (113 codons). On the other hand, a sequence can be found between gene *bglB* and this open reading frame, which could qualify as a rho-independent terminator. We are hesitant, however, to interpret this structure as the 3′ limit of the *bgl* operon, because our preliminary evidence indicates that the open reading frame is expressed and controlled—at least partially—coordinatively with the *bgl* genes (data not shown). Noteworthy in this context is the location of the potential terminator: the left inverted repeat overlaps the last codon of *bglB*. It is conceivable that translation of *bglB* could attenuate or eliminate activity of the terminator, providing for a coordinated expression of *bglB* and the open reading frame.

The *bgl* operon is cryptic in wild-type *E. coli* K-12 cells. Spontaneous mutations to Bgl$^+$ arise, the majority of which are due to integration of insertion element IS*1* or IS*5* into a region proximal to the cyclic AMP binding protein binding site of the operon (34, 35; K. Schnetz, H. J. Ronecker, and B. Rak, unpublished data), leading to an enhancement of the activity of the *bgl* promoter (34, 35; Schnetz and Rak, in preparation). These events saturate the DNA sequence from the cyclic AMP binding protein binding site to a region within the potential *phoU* terminator (34, 35; data not shown). All of the mutants of this class which have been tested are inducible by the substrate (31, 33; Schnetz and Rak, in preparation). Thus it is unlikely that signal sequences involved in substrate-dependent positive regulation map upstream of the *bgl* promoter (Fig. 3). What kind of model for the regulation of the *bgl* operon can we make? The structural gene coding for the positive regulation protein should be *bglC*. If this is correct, then its position as the first gene of the operon is unusual. Other remarkable features are the 130-nucleotide-long leader and the extensive intercistronic region between *bglC* and *bglS*. A search for signal sequences that could be involved in the regulation of the operon revealed that *bglC* is bracketed by potential stem-loop structures similar in motif to rho-independent terminators. Preceding these structures and partially overlapping them we found highly homologous sequences (box A and box B in Fig. 3 and 6). If the stem-loop structures represent terminators, then in the induced state transcription initiating at the *bgl* promoter must overcome both of these terminators, i.e., antitermination must take place. This model would involve the *bglC* gene product as a specific antitermination factor which, when charged with effector, may recognize the box A-box B-terminator motif. How then could the system provide for the synthesis of enough *bglC* gene product (and transport protein) in the uninduced state to ensure inducibility? Either the terminators are leaky enough to allow for sufficient synthesis or a second promoter is present directing low-level expression of *bglC* (and *bglS*) in the uninduced state. A sequence which could qualify for such a second promoter has been found (Fig. 3). On the other hand, the stem-loop structure that could be formed by the proximal inverted repeat is less stable than that one predicted for the distal inverted repeat (ΔG, −21 versus −26 kcal/mol; Fig. 6). Thus a hierarchy of termination could exist.

Our own experimental evidence supports the above scenario of regulation by antitermination. In a separate communication (Schnetz and Rak, in preparation) we show that the stem-loop structures bracketing gene *bglC* both function as efficient transcriptional terminators and that the *bglC* gene product acts as a specific antiterminator at these signal structures, but only in the presence of an inducer. A mechanism of regulation involving antitermination of transcription is also postulated in the accompanying paper (20).

An excreted β-endoglucanase (1,3-1,4-β-D-glucan-4-glucanohydrolase) is synthesized by *B. subtilis* 168 from a gene which coincidentally is called *bgl* (6). The nucleotide sequence of this gene and several hundred base pairs of the upstream region has been determined (27). Not surprisingly, no homology between the *B. subtilis* gene and *bglB* (or the other two genes of the *bgl* operon) could be detected on the nucleotide or protein level (data not shown). The region proximal to the *B. subtilis* gene, however, showed homology to the *bgl* operon on two levels (Fig. 7). The DNA sequence preceding the β-endoglucanase gene contains the 3′ part of an open reading frame 85 codons in length. Alignment of the corresponding amino acid sequences of this open reading frame and the C-terminal 85 amino acids of gene *bglC* showed matches at 32 positions (38%). If exchanges of chemically similar amino acids are allowed, homology is 54%. The presence of a single cysteine at identical positions may also be of significance. These observations strongly suggest evolutionary conservation of the C-terminal part of the *bglC* gene product and the hypothetical product encoded by the unidentified reading frame. No relationship is apparent on the level of the nucleotide sequence, indicating that selective pressure acted to conserve the functional structure of the proteins. This leads to the question of whether the expression of the β-glucanase gene of the gram-positive bacterium *B. subtilis* and the *bgl* operon of the gram-negative bacterium *E. coli* is regulated in a similar fashion. Nothing is known about the regulation of the *B. subtilis* gene. Again striking, however, is the high degree of homology—in this case on the level of DNA—between the region downstream of the unidentified reading frame of *B. subtilis* and the box A-box B-terminator motif flanking the *bglC* gene of *E. coli* (Fig. 3, 6, and 7) which, according to our model, plays an essential role in the regulation of the *E. coli bgl* operon.

Inspection of the sequence of another *B. subtilis* gene, *sacB*, which encodes an excreted levansucrase (3, 45), also revealed significant homologies to the box A-box B motif. The homologous sequences are located in the leader between the promoter and the translation start signal of the gene (*sacR*). No other homologies on the nucleotide or amino acid levels were found in this case. In Fig. 6 the relevant sequences of the *E. coli bgl* operon are shown aligned with the respective sequences of *B. subtilis* genes *bgl* and *sacB*. In all four cases box A-box B overlaps with an inverted repeat beginning at identical positions within box B.

In contrast to *B. subtilis bgl*, regulation of *sacB* has been studied in some detail (3, 43). A gene necessary for induction but not linked to *sacB* (*sacS*) has been defined. Expression can be induced by sucrose, and it has recently been shown that in the uninduced state transcription is constitutive but terminates at the inverted repeat. Termination is overcome upon induction (43), resulting in the expression of the distal coding region. Thus, a mechanism of inducer-mediated transcriptional antitermination probably involving a common sequence motif is found both in a gram-negative system and in a gram-positive system.

We believe that the protein and DNA homologies taken together point to a conserved regulatory principle common to all three systems. The regulation of the β-endoglucanase gene of *B. subtilis* and its relationship to the regulatory mechanisms of *sacB* and the *bgl* operon certainly merits further investigation. A comparative examination of the proteins encoded by *bglC* and *sacS* would be most interesting in this context, and it will become tempting to speculate about the evolutionary significance of the relationships found.

ACKNOWLEDGMENTS

Plasmid pAR6 and strain JF201 were received from G. Hobom. B. Stocker sent us strain SL5235. M. Scheufens, M. Trippel, and A. Uhlmann provided us with computer programs. R. Hertel, H. Kössel, and E. Schwartz critically read the manuscript. We gratefully acknowledge their contributions.

This work was supported by Deutsche Forschungsgemeinschaft through grants Ra276/3-6 and SFB31.

ADDENDUM IN PROOF

Gene *sacS* of *B. subtilis* has now been cloned, and its nucleotide sequence has been determined. It encodes a protien of M_r 32,000 which has an evenly distributed homology of about 30% (identical amino acids) to the *bglC* gene product. The same degree of homology was detected between the SacS protein and the protein potentially encoded by the open reading frame preceding the β-endoglucanase gene of *B. subtilis* (M. Steinmetz, personal communication).

LITERATURE CITED

1. **Amemura, M., K. Makino, H. Shinagawa, A. Kobayashi, and A. Nakata.** 1985. Nucleotide sequence of the genes involved in phosphate transport and regulation of the phosphate regulon in *Escherichia coli*. J. Mol. Biol. **184**:241–250.
2. **Ansorge, W., and R. Barker.** 1984. System for DNA sequencing with resolution of up to 600 bp. J. Biochem. Biophys. Methods **9**:33–47.
3. **Aymerich, S., G. Gonzy-Tréboul, and M. Steinmetz.** 1986. 5′-Noncoding region *sacR* is the target of all identified regulation affecting the levansucrase gene in *Bacillus subtilis*. J. Bacteriol. **166**:993–998.
4. **Bachmann, B. J.** 1983. Linkage map of *Escherichia coli* K-12, edition 7. Microbiol. Rev. **47**:180–230.
5. **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.
6. **Borriss, R., K. H. Suess, M. Suess, R. Manteuffel, and J. Hofmeister.** 1986. Mapping and properties of *bgl* (β-glucanase) mutants of *Bacillus subtilis*. J. Gen. Microbiol. **132**:431–442.
7. **Buhk, H.-J., and W. Messer.** 1983. The replication origin region of *Escherichia coli*: nucleotide sequence and functional units. Gene **24**:265–279.
8. **Bullas, L. R., and J. I. Ryu.** 1983. *Salmonella typhimurium* LT2 strains which are r^- m^+ for all three chromosomally located systems of DNA restriction and modification. J. Bacteriol. **156**:471–474.
9. **Calos, M. P.** 1978. DNA sequence for a low-level promoter of

the *lac* repressor gene and an 'up' promoter mutation. Nature (London) **274**:762–765.

10. **Chang, A. C. Y., and S. N. Cohen.** 1978. Construction and characterization of the amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. J. Bacteriol. **134**:1141–1156.
11. **Fox, C. F., and G. Wilson.** 1968. The role of phosphoenolpyruvate-dependent kinase system in β-glucoside catabolism in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **59**:988–995.
12. **Grantham, R.** 1974. Amino acid difference formula to help explain protein evolution. Science **185**:862–864.
13. **Grantham, R., C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier.** 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9**:r43–r74.
14. **Jacobson, G. R., C. A. Lee, J. E. Leonard, and M. H. Saier, Jr.** 1983. Mannitol-specific enzyme II of the bacterial phosphotransferase system. I. Properties of the purified permease. J. Biol. Chem. **258**:10748–10756.
15. **Konigsberg, W., and G. N. Godson.** 1983. Evidence for use of rare codons in the *dna*G gene and other regulatory genes of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **80**:687–691.
16. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.
17. **Lee, C. A., and M. H. Saier, Jr.** 1983. Mannitol-specific enzyme II of the bacterial phosphotransferase system. III. The nucleotide sequence of the permease gene. J. Biol. Chem. **258**:10761–10767.
18. **Lichtenstein, C., and S. Brenner.** 1982. Unique insertion site of Tn*7* in the *E. coli* chromosome. Nature (London) **297**:601–603.
19. **Magota, K., N. Otsuji, T. Miki, T. Horiuchi, S. Tsunasawa, J. Kondo, F. Sakiyama, M. Amemura, T. Morita, H. Shinagawa, and A. Nakata.** 1984. Nucleotide sequence of the *phoS* gene, the structural gene for the phosphate-binding protein of *Escherichia coli*. J. Bacteriol. **157**:909–917.
20. **Mahadevan, S., A. E. Reynolds, and A. Wright.** 1987. Positive and negative regulation of the *bgl* operon in *Escherichia coli*. J. Bacteriol. **169**:2570–2578.
21. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
22. **Margolin, W., and M. M. Howe.** 1986. Localization and DNA sequence analysis of the *C* gene of bacteriophage *mu*, the positive regulator of *mu* late transcription. Nucleic Acids Res. **14**:4881–4897.
23. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavage. Methods Enzymol. **65**:499–560.
24. **Meijer, M., E. Beck, F. G. Hansen, H. Z. N. Bergmans, W. Messer, K. von Meyenburg, and H. Schaller.** 1979. Nucleotide sequence of the origin of replication of the *Escherichia coli* K-12 chromosome. Proc. Natl. Acad. Sci. USA **76**:580–584.
25. **Messing, J.** 1983. New M13 vectors for cloning. Methods Enzymol. **101**:20–79.
26. **Miller, J. H.** 1972. Experiments in molecular genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
27. **Murphy, N., D. J. McConnell, and B. A. Cantwell.** 1984. The DNA sequence of the gene and genetic control sites for the excreted *B. subtilis* enzyme β-glucanase. Nucleic Acids Res. **12**:5355–5367.
28. **Nakamura, M., M. Yamada, Y. Hirota, K., Sugimoto, A. Oka, and M. Takanami.** 1981. Nucleotide sequence of the *asn*A gene coding for asparagine synthetase of *E. coli* K-12. Nucleic Acids Res. **9**:4669–4676.
29. **Nyman, K., K. Nakamura, H. Ohtsubo, and E. Ohtsubo.** 1981. Distribution of insertion sequence IS1 in Gram-negative bacteria. Nature (London) **289**:609–612.
30. **Postma, P. W., and J. W. Lengeler.** 1985. Phosphoenolpyruvate:carbohydrate phosphotransferase system of bacteria. Microbiol. Rev. **49**:232–269.
31. **Prasad, I., and S. Schaefler.** 1974. Regulation of the β-glucoside system in *Escherichia coli* K-12. J. Bacteriol. **120**:638–650.
32. **Rak, B., and M. von Reutern.** 1984. Insertion element IS5 contains a third gene. EMBO J. **3**:807–811.
33. **Reynolds, A. E., J. Felton, and A. Wright.** 1981. Insertion of DNA activates the cryptic *bgl* operon in *E. coli*. Nature (London) **293**:625–629.
34. **Reynolds, A. E., S. Mahadevan, J. Felton, and A. Wright.** 1985. Activation of the cryptic *bgl* operon: insertion sequences, point mutations, and changes in superhelicity affect promoter strength. UCLA Symp. Mol. Cell. Biol. New Series **20**: 265–277.
35. **Reynolds, A. E., S. Mahadevan, St. F. J. LeGrice, and A. Wright.** 1986. Enhancement of bacterial gene expression by insertion elements or by mutation in a CAP-cAMP binding site. J. Mol. Biol. **191**:85–95.
36. **Rose, S. P., and C. F. Fox.** 1971. The β-glucoside system of *Escherichia coli*. II. Kinetic evidence for a phosphoryl-enzyme II intermediate. Biochem. Biophys. Res. Commun. **45**:376–380.
37. **Rosenberg, M., and D. Court.** 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. **13**:319–353.
38. **Saier, M. H., Jr.** 1977. Bacterial phosphoenolpyruvate:sugar phosphotransferase systems: structural, functional, and evolutionary interrelationships. Bacteriol. Rev. **41**:856–871.
39. **Saier, M. H., Jr.** 1985. Mechanism and regulation of carbohydrate transport in bacteria. Academic Press, Inc., London.
40. **Schaefler, S., and A. Malamy.** 1969. Taxonomic investigations on expressed and cryptic phospho-β-glucosidases in *Enterobacteriaceae*. J. Bacteriol. **99**:422–433.
41. **Schoner, B., and R. G. Schoner.** 1981. Distribution of IS5 in bacteria. Gene **16**:347–352.
42. **Sharp, P. M., and W. H. Li.** 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. Nucleic Acids Res. **14**:7737–7749.
43. **Shimotsu, H., and D. J. Henner.** 1986. Modulation of *Bacillus subtilis* levansucrase gene expression by sucrose and regulation of the steady-state mRNA level by *sacU* and *sacQ* genes. J. Bacteriol. **168**:380–388.
44. **Shine, J., and L. Dalgarno.** 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementary to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. USA **71**:1342–1346.
45. **Steinmetz, M., D. Le Coq, St. Aymerich, G. Gonzy-Tréboul, and P. Gay.** 1985. The DNA sequence of the gene for the secreted *Bacillus subtilis* enzyme levansucrase and its genetic control sites. Mol. Gen. Genet. **200**:220–228.
46. **Sugimoto, K., A. Oka, H. Sugisaki, M. Takanami, A. Nishimura, Y. Yasuda, and Y. Hirota.** 1979. Nucleotide sequence of *Escherichia coli* K-12 replication origin. Proc. Natl. Acad. Sci. USA **76**:575–579.
47. **Surin, B. P., D. A. Jans, A. L. Fimmel, D. C. Shaw, G. B. Cox, and H. Rosenberg.** 1984. Structural gene for the phosphate-repressible phosphate-binding protein of *Escherichia coli* has its own promoter: complete nucleotide sequence of the *phoS* gene. J. Bacteriol. **157**:772–778.
48. **Surin, B. P., H. Rosenberg, and G. B. Cox.** 1985. Phosphate-specific transport system of *Escherichia coli*: nucleotide sequence and gene-polypeptide relationships. J. Bacteriol. **161**: 189–198.
49. **Sutcliffe, J. G.** 1978. Complete nucleotide sequence of the *Escherichia coli* plasmid pBR322. Cold Spring Harbor Symp. Quant. Biol. **43**:77–90.
50. **Tinoco, I., P. N. Borer, B. Dengler, M. D. Levine, O. C. Uhlenbeck, D. M. Crothers, and J. Gralla.** 1973. Improved estimation of secondary structure in ribonucleic acids. Nature (London) New Biol. **246**:40–41.
51. **Walker, J. E., N. J. Gay, M. Saraste, and A. N. Eberle.** 1984. DNA sequence around the *E. coli unc* operon. Biochem. J. **224**:799–815.