JOURNAL OF BACTERIOLOGY, Feb. 1989, p. 625–635
0021-9193/89/020625-11$02.00/0

Vol. 171, No. 2

Expression and Nucleotide Sequence of the *Lactobacillus bulgaricus* β-Galactosidase Gene Cloned in *Escherichia coli*

BRIAN F. SCHMIDT,* ROBIN M. ADAMS, CAROL REQUADT, SCOTT POWER, AND STANLEY E. MAINZER

Genencor, Inc., 180 Kimball Way, South San Francisco, California 94080

Received 4 August 1988/Accepted 24 October 1988

The *Lactobacillus bulgaricus* β-galactosidase gene was cloned on a ca. 7-kilobase-pair *Hin*dIII fragment in the vector pKK223-3 and expressed in *Escherichia coli* by using its own promoter. The nucleotide sequence of the gene and approximately 400 bases of 3′- and 5′-flanking sequences was determined. The amino acid sequence of the β-galactosidase, deduced from the nucleotide sequence of the gene, yielded a monomeric molecular mass of ca. 114 kilodaltons, slightly smaller than the *E. coli lacZ* and *Klebsiella pneumoniae lacZ* enzymes but larger than the *E. coli* evolved (*ebgA*) β-galactosidase. The cloned β-galactosidase was found to be indistinguishable from the native enzyme by several criteria. From amino acid sequence alignments, the *L. bulgaricus* β-galactosidase has a 30 to 34% similarity to the *E. coli lacZ*, *E. coli ebgA*, and *K. pneumoniae lacZ* enzymes. There are seven regions of high similarity common to all four of these β-galactosidases. Also, the putative active-site residues (Glu-461 and Tyr-503 in the *E. coli lacZ* β-galactosidase) are conserved in the *L. bulgaricus* enzyme as well as in the other two β-galactosidases mentioned above. The conservation of active-site amino acids and the large regions of similarity suggest that all four of these β-galactosidases evolved from a common ancestral gene. However, these enzymes are quite different from the thermophilic β-galactosidase encoded by the *Bacillus stearothermophilus bgaB* gene.

Little is known about the specific amino acids involved in the substrate binding and catalysis of the *Escherichia coli lacZ* β-galactosidase, even though this enzyme has been studied for many years (35). From iodination (8, 15), fluorotyrosine substitution (27), and active-site-directed inhibitor (11, 12, 23) experiments, Tyr-503 is thought to be the proton-donating species needed for catalysis. Primarily by the use of active-site-directed reagents (11, 13), Glu-461 is believed to be the catalytic residue that stabilizes the carboxonium ion intermediate. Recently, Cupples and Miller also confirmed that Glu-461 is essential for enzyme activity by substituting other amino acids at this position, with different amber suppressor strains being used (C. G. Cupples and J. H. Miller, J. Cell. Biochem., **11C**:242, 1987). Much earlier, Langridge used a genetic approach to map a series of mutations affecting the substrate binding (18), heat stability (19), and urea sensitivity (20) of *E. coli* β-galactosidase. Unfortunately, at the time his work was done, the β-galactosidase gene had not been cloned, and so the amino acids involved in these mutations could not easily be determined.

Presently, the complete amino acid sequences of four different β-galactosidases have been reported, two from *E. coli* and one each from *Klebsiella pneumoniae* and *Bacillus stearothermophilus*. The sequence of the *E. coli lacZ* (17) β-galactosidase is very similar (ca. 60%) to the *Klebsiella pneumoniae lacZ* gene product (5). The *E. coli ebgA* gene product is different from the other two (ca. 30% similarity) and has very low lactase activity unless two undefined point mutations are present (31). The other sequenced β-galactosidase, the *B. stearothermophilus bgaB* gene product, has no homology to the three other enzymes (14). Thus, the availability of an additional related sequence of a β-galactosidase from a different genus could be useful in determining which amino acids are important for enzyme function. To this end, the β-galactosidase from *Lactobacillus bulgaricus* has been cloned in *E. coli* and its amino acid sequence has been deduced and compared with the other known β-galactosidase sequences.

The β-galactosidase from *L. bulgaricus* has been purified previously (16) but not extensively studied. Also, the plasmid-encoded β-galactosidase from *Lactobacillus casei* has been cloned in *E. coli* (J. L. Flickinger, E. V. Porter, and B. M. Chassy, Abstr. Annu. Meet. Am. Soc. Microbiol. 1986, abstr. H-179, p. 156), but no detailed nucleotide sequence of this gene has yet appeared.

In addition, little is yet known about the genetics of the food grade lactobacilli (6). It is expected that the knowledge of the nucleotide sequence and codon usage of the relatively large β-galactosidase gene will be useful for future cloning and expression experiments.

MATERIALS AND METHODS

Bacterial strains and media. A lyophilized culture of *L. bulgaricus* B131 (Centre International de Recherche Daniel Carasso, BSN group, Le Plessis-Robinson, France) was grown overnight at 37°C in a capped tube containing 10 ml of litmus milk broth (Difco Laboratories) and then stored frozen at −70°C in 1-ml aliquots. When needed, a frozen aliquot was grown overnight in 20 ml of Lactobacilli MRS broth (Difco Laboratories) and was then used to inoculate a 500-ml culture of Lactobacilli MRS broth. The cells from this overnight culture were harvested by centrifugation and used for the purification of β-galactosidase or chromosomal DNA.

β-Galactosidase purification. Native or cloned β-galactosidase was purified from *L. bulgaricus* or *E. coli* JM105 (34), respectively, by essentially the same method (B. Chassy, personal communication). During the purification, the samples were kept at 4°C, except that the columns were run at room temperature. Enough HEPES buffer (25 mM *N*-2-hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid [pH 7.4], 1 mM $MgCl_2$) was added to 20 ml of dry glass beads (diameter, 0.45 to 0.50 mm; Braun) and 20 ml of cell paste to bring the total volume to 60 ml. The cells were lysed in a Braun

* Corresponding author.

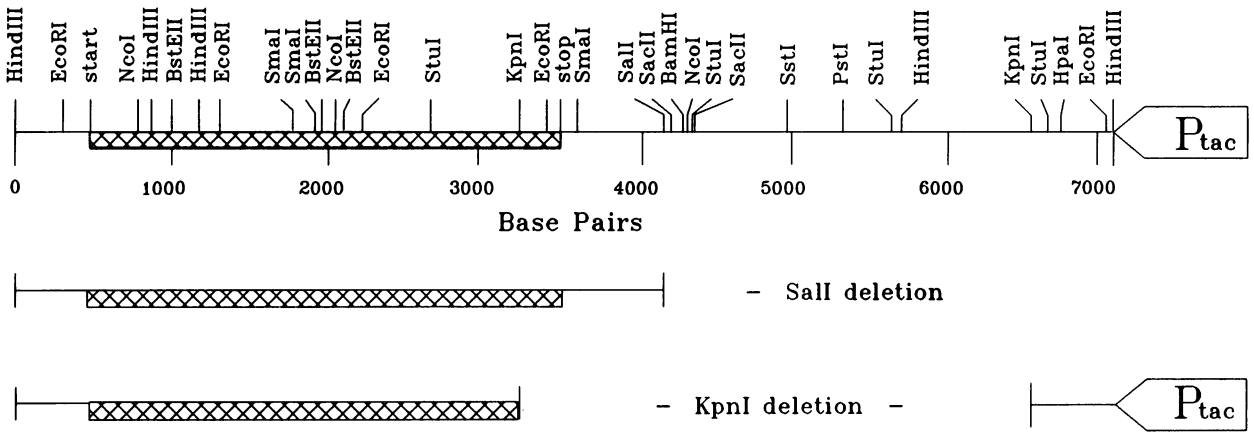

FIG. 1. Restriction map of the DNA insert carrying the *L. bulgaricus* gene (▬) in pKK223-3. The orientation of the insert with respect to the *tac* promoter (P_{tac}) on the vector is shown. The region remaining after deletion of the *Kpn*I or *Sal*I fragments is indicated below the map. The distances between restriction sites were determined by DNA sequencing or estimated by running cleaved DNA on 5% polyacrylamide gels with a 123-base-pair ladder (Bethesda Research Laboratories, Inc.) as a size marker. The ca. 7-kilobase-pair insert could not be cleaved by *Asu*II, *Bgl*II, *Cla*I, *Eco*RV, *Mlu*I, *Nsi*I, *Xba*I, or *Xho*I.

homogenizer by shaking three times for 30 s with 15-s intervals. After two low-speed centrifugations to remove the glass beads and cellular debris, the supernatant was centrifuged at 50,000 rpm for 30 minutes in a Beckman Ti65 rotor. The supernatant was fractionated by ammonium sulfate precipitation, and the 40 to 60% saturation cut was retained. The pellet was dissolved in HEPES buffer containing 1 M ammonium sulfate and applied to a phenyl-Sepharose CL-4B column (Pharmacia Fine Chemicals) that had been equilibrated in the same buffer. The β-galactosidase activity was eluted in a broad peak (0.7 to 0.4 M ammonium sulfate) from this column by using a 1 to 0 M ammonium sulfate gradient (0.5%/min) in HEPES buffer. Fractions with activity were pooled, concentrated (Centriprep-30; Amicon), and exchanged into HEPES buffer by using a PD10 desalting column (Pharmacia). The sample was then applied to a MonoQ HR5/5 column (Pharmacia) equilibrated with HEPES buffer and eluted with a 0 to 0.5 M KCl gradient (1%/min). The fractions with β-galactosidase activity (0.1 to 0.2 M KCl) were usually homogeneous at this stage as determined by sodium dodecyl sulfate-polyacrylamide gel electrophoresis. In some cases with the cloned enzyme, an additional gel filtration (HEPES buffer containing 0.2 M KCl) run with two Superose-12 HR10/30 columns in series (Pharmacia) was needed to obtain a homogeneous β-galactosidase preparation. N-terminal sequence analysis was determined for both the native and cloned enzyme by using the microsequencing technique as previously described (28).

Cloning of the β-galactosidase gene. The harvested *L. bulgaricus* cells were treated with 1 mg of lysozyme (chicken egg white; Sigma Chemical Co.) per ml in 100 mM Tris hydrochloride (pH 8.0)–20 mM disodium EDTA–20% sucrose for 30 min at 37°C. The cells were then subjected to several freeze-thaw cycles by freezing in a dry-ice–ethanol bath and thawing in a 40°C water bath. The cells were lysed by addition of 1/2 volume of a 1% sodium dodecyl sulfate solution. Chromosomal DNA was purified by multiple phenol extractions. The DNA was finally precipitated with ethanol, dried, and then dissolved in 10 mM Tris hydrochloride (pH 7.6)–1.0 mM disodium EDTA.

The DNA was either partially or completely digested with *Eco*RI, *Hin*dIII, or *Pst*I (Boehringer Mannheim Biochemicals). The cleaved DNA was fractionated by electrophoresis on a 5% polyacrylamide gel, and fragments of approximately 2 to 15 kilobase pairs were electroeluted from the gel. These fragments were ligated into the vector pKK223-3 (Pharmacia), which had been linearized with the appropriate restriction enzyme and dephosphorylated with alkaline phosphatase (Boehringer Mannheim). The resulting plasmids were used to transform *E. coli* JM105 (34), which was then grown on plates containing 50 μg of carbenicillin per ml and 40 μg of X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside; Sigma Chemical Co.) per ml. β-Galactosidase activity was detected by the appearance of blue colonies. Plasmid DNA from bacterial colonies was isolated by the alkaline lysis method of Birnboim and Doly (3).

DNA sequencing. With the restriction map of the cloned *L. bulgaricus* β-galactosidase gene as a guide, various DNA fragments were cloned into M13mp18 or M13mp19 (New England BioLabs, Inc.) by using the *Eco*RI, *Hin*dIII, *Sal*I, *Stu*I, and *Xma*I restriction sites (34). The nucleotide sequence of the fragments was determined in both directions by the dideoxy-chain termination method (29). Any ambiguities due to compression were resolved by replacing guanosine with inosine in the sequencing reactions (22).

RESULTS

Cloning of the β-galactosidase gene. Fractionated *L. bulgaricus* chromosomal DNA was shotgun cloned into the *E. coli* expression vector pKK223-3, and over 5,000 transformants were screened for β-galactosidase activity. The DNA fragments were inserted downstream of the *tac* promoter on pKK223-3 in case the *L. bulgaricus* β-galactosidase promoter was nonfunctional in *E. coli*. One blue colony from a partial *Hin*dIII digest of *L. bulgaricus* DNA was obtained. Plasmids isolated from this colony carried an insert of four *Hin*dIII fragments with a combined length of ca. 7 kilobase pairs. A restriction enzyme map of the *L. bulgaricus* insert was constructed and is shown in Fig. 1.

β-Galactosidase from *L. bulgaricus* was purified and the sequence of the first 29 amino acids was determined as described in the Materials and Methods. From this N-terminal sequence, a pool of DNA oligomers was synthesized (Genentech, Inc.) complementary to the β-galactosidase gene corresponding to amino acids 24 to 28. The

oligomers were end labeled with ^{32}P by using [γ-^{32}P]ATP (Amersham Corp.) and polynucleotide kinase (New England BioLabs) and then used to probe Southern blots (30) of the *L. bulgaricus* insert digested with *Hin*dIII. The labeled probes hybridized specifically to a ca. 880-base-pair *Hin*dIII fragment. This fragment mapped to the distal portion of the insert in relation to the *tac* promoter of the pKK223-3 vector (Fig. 1). Thus, the β-galactosidase gene was inserted in the opposite orientation of transcription from the *tac* promoter, indicating that the *E. coli* RNA polymerase recognized a promoter present on the ca. 880-base-pair *L. bulgaricus* fragment. It is also possible, but not very likely, that transcription occurs by a readthrough from a cryptic promoter on the vector.

To localize the 3′ end of the β-galactosidase gene on the insert, we digested the plasmid with either *Sal*I or *Kpn*I and then religated it. Plasmids with deletions of both *Sal*I fragments, one from the *L. bulgaricus* insert and the other from the pKK223-3 vector, were isolated from blue colonies on X-Gal plates, indicating that the β-galactosidase stop codon was still present in these clones. Deletion of the *Kpn*I fragment gave rise to colonies having no β-galactosidase activity; thus, the end of the gene mapped between the *Kpn*I and *Sal*I restriction sites (Fig. 1). This was later verified by nucleotide sequencing. Also, as expected, *Sal*I-deleted clones with the *tac* promoter removed still expressed β-galactosidase, confirming that the gene was not transcribed from the *tac* promoter.

DNA sequence. The DNA sequence of the entire *L. bulgaricus* β-galactosidase gene and some 700 base pairs of flanking sequences was determined (Fig. 2). The gene has a G+C content of 52.2%, compared with the reported G+C content of ca. 50% for genomic *L. bulgaricus* DNA (21). The sequence of the first 29 amino acids predicted from the *L. bulgaricus* gene sequence agreed exactly with that determined by sequencing the amino-terminal end of the purified protein. However, the N-terminal methionine did not appear in the mature protein, and amino acid 12, a glutamate, was not determined reliably by protein sequencing. In addition, the cloned β-galactosidase purified from *E. coli* had the same N-terminal sequence as the native enzyme. The cloned and native enzymes were also indistinguishable on the basis of their specific activities, cyanogen bromide digestion patterns, and mobilities on sodium dodecyl sulfate-polyacrylamide gels (data not shown).

From the DNA sequence, the β-galactosidase gene codes for a protein monomer of 1,006 amino acids with a calculated molecular weight of 113,915. This compared well with a monomeric molecular weight of ca. 110,000 observed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis of purified *L. bulgaricus* β-galactosidase.

Flanking nucleotide sequences. The nucleotide sequence upstream of the β-galactosidase gene has a G+C content of 53.8%, slightly higher than that of the gene. By comparison with the homologous region from *Streptococcus thermophilus*, the sequence upstream of the β-galactosidase gene appears to code for the carboxy-terminal end of a β-galactoside transport protein (26). Putative −10 and −35 promoter regions for *E. coli* RNA polymerase are shown in Fig. 2. Further experiments are required to show that this is indeed the promoter for *E. coli* RNA polymerase and to show whether a functional promoter exists in this region for the *L. bulgaricus* RNA polymerase.

The downstream nucleotide sequence was searched for possible terminator sequences by using a computer program (Genentech, Inc.). A region with dyad symmetry was found around the stop codon (Fig. 2), but the resulting hairpin does not resemble a typical rho-independent terminator (25). Putative hairpins have also been found at the end of the *lacZ* genes in *E. coli* (4) and *K. pneumoniae* (5), but as yet the function of these structures is unknown. It is possible that the terminator is further downstream, since preliminary sequencing data have revealed a putative open reading frame downstream of the β-galactosidase gene (B. F. Schmidt et al., unpublished results). Since there appears to be a β-galactoside transport gene upstream and another reading frame downstream, it is very likely that the *L. bulgaricus* β-galactosidase gene is part of an operon.

Comparison of β-galactosidases. The *L. bulgaricus* enzyme is similar to the two β-galactosidases from *E. coli* (encoded by *lacZ* and *ebgA*) (17, 31) and the *K. pneumoniae* enzyme (encoded by *lacZ*) (5) in terms of size and average amino acid weight (Table 1). However, the thermostable β-galactosidase from *B. stearothermophilus* (encoded by *bgaB*) (14) is quite different from the *L. bulgaricus* enzyme and the other sequenced β-galactosidases.

The percent similarity between two β-galactosidase enzymes was determined from a computer alignment of two protein sequences. We used a computer program developed at Genentech, Inc., which utilizes the Fitch and Smith (10) version of the Needleman and Wunsch (24) algorithm. The results are shown in Table 2 for the four β-galactosidases tested (only identical residues were considered in the alignments, and a gap penalty of eight plus four per residue was used). The *L. bulgaricus* β-galactosidase has a similarity of 30 to 34% to the other three β-galactosidases and no homology to the *B. stearothermophilus bgaB* β-galactosidase. In Fig. 3, the amino acid sequences of all four enzymes are aligned. From a comparison of the sequences, there are seven regions of high similarity (>50% identical residues in eight or more consecutive residues) in these four enzymes; this accounts for only about 10% of the total amino acids. The amino acids in the high-similarity regions are 201 to 219 (I), 296 to 304 (II), 373 to 413 (III), 451 to 468 (IV), 525 to 541 (V), 576 to 584 (VI), and 873 to 882 (VII) (*L. bulgaricus* β-galactosidase sequence) (Fig. 3). An additional region of high similarity, roughly centered around amino acid 975 in the *L. bulgaricus* sequence, is present in all but the *E. coli ebgA* β-galactosidase. There are also five regions of low similarity among the four different β-galactosidases, roughly corresponding to *L. bulgaricus* amino acids 1 to 90, 230 to 290, 620 to 860, and 890 to 960.

In addition to the β-galactosidases mentioned above, the sequences of the amino-terminal portions of three other β-galactosidases have been reported. Although N-terminal sequences are generally only moderately conserved, there is a rather high similarity (37%) between the first 119 amino acids of the yeast *Kluyveromces lactis* β-galactosidase (2) and the corresponding region of the *E. coli lacZ* enzyme (this compares with a 28% similarity between the *L. bulgaricus* and *E. coli lacZ* β-galactosidases in the same region). The first 86 amino acids of the β-galactosidase from *Streptomyces lividans* (7) are 26% and 31% similar to the first 100 amino acids of the *L. bulgaricus* and *E. coli lacZ* enzymes, respectively. Unfortunately, the rest of the *Kluyveromyces lactis* and *Streptomyces lividans* β-galactosidase sequences have not yet been reported. Also, the first 113 amino acids of the *L. bulgaricus* enzyme is ca. 50% similar to the homologous region of the β-galactosidase from *Streptococcus thermophilus* (26).

A computer alignment was also made for the nucleotide sequences of the *E. coli lacZ* and *L. bulgaricus* β-galactosi-

FIG. 2. DNA sequence of the *L. bulgaricus* B131 β-galactosidase gene and flanking regions determined by the dideoxy-chain termination method (29). The start and stop codons are underlined along with a putative *E. coli* transcriptional promoter (−10 and −35 regions) and a possible ribosomal-binding site (RBS). A region of dyad symmetry near the stop codon is indicated by arrows.

AGAACATGATGGCTCGGGACAAGAACCATGCTTCAATCCTAATCTGGTCTTTAGGCAATG

AGTCTTACGCCGGCACTGTCTTTGCCCAAATGGCTGATTACGTCCGGAAGGCTGATCCGA

CCCGGGTTCAGCACTATGAAGGGGTGACCCACAACCGGAAGTTTGACGACGCCACCCAGA

TTGAAAGCCGGATGTATGCTCCGGCCAAGGTAATTGAAGAATACTTGACCAATAAACCAG

CCAAGCCATTTATCTCAGTTGAATACGCTCACGCCATGGGCAACTCCGTCGGTGACCTGG

CCGCCTACACGGCCCTGGAAAAATACCCCCACTACCAGGGCGGCTTCATCTGGGACTGGA

TTGACCAAGGACTGGAAAAAGACGGGCACCTGCTTTATGGGGGCGACTTCGATGACCGGC

CAACCGACTATGAATTCTGCGGGAACGGCCTGGTCTTTGCTGACCGGACTGAATCGCCGA

AACTGGCTAATGTCAAGGCCCTTTACGCCAACCTTAAGTTAGAAGTAAAAGATGGGCAGC

TCTTCCTCAAAAACGACAATTTATTTACCAACAGCTCATCTTACTACTTCTTGACTAGTC

TTTTGGTCGATGGCAAGTTGACCTACCAGAGCCGGCCTCTGACCTTTGGCCTGGAGCCTG

GCGAATCCGGGACCTTTGCCCTGCCTTGGCCGGAAGTCGCTGATGAAAAAGGGGAGGTCG

TCTACCGGGTAACGGCCCACTTAAAAGAAGACTTGCCTTGGGCGGATGAGGGCTTCACTG

TGGCTGAAGCAGAAGAAGTAGCTCAAAAGCTGCCGGAATTTAAGCCGGAAGGGCGGCCAG

ATTTAGTTGATTCCGACTACAACCTAGGCCTGAAAGGAAATAACTTCCAAATTCTCTTCT

CCAAGGTCAAGGGCTGGCCGGTTTCCCTCAAGTATGCCGGTAGGGAATACTTGAAGCGGC

TGCCGGAATTTACCTTCTGGCGGGCCCTGACGGACAACGACCGGGGAGCTGGTTACGGCT

ATGATCTGGCCCGGTGGGAAAATGCCGGCAAGTATGCCCGCTTGAAAGACATCAGCTGCG

AGGTCAAGGAAGACTCCGTTTTGGTCAAGACTGCCTTTACGTTGCCTGTCGCCTTAAAGG

GTGATTTAACCGTGACCTATGAAGTCGATGGACGGGGCAAGATTGCTGTAACAGCTGACT

TCCCAGGCGCGGAAGAAGCTGGTCTCTTGCCAGCCTTTGGCTTGAACCTGGCCCTGCCAA

AAGAACTGACCGATTACCGCTACTATGGTCTGGGACCTAATGAGAGCTACCCAGACCGCT

TGGAAGGTAATTACCTGGGCATCTACCAGGGAGCGGTAAAAAAGAACTTTAGCCCATATC

GTCCGCAGGAAACGGGCAACCGGAGCAAGGTTCGCTGGTACCAGCTCTTTGATGAAAAGG

GCGGCTTGGAATTTACGGCCAATGGGGCAGACTTGAACTTGTCTGCTTTGCCATATTCTG

CCGCCCAAATTGAAGCAGCGGACCACGCTTTTGAACTGACTAACAATTACACTTGGGTTA

GAGCCTTAAGCGCCCAGATGGGGGTCGGCGGGGATGACTCCTGGGGGCAGAAGGTCCACC

CGGAATTCTGCCTGGATGCTCAAAAAGCCCGCCAGCTTCGCCTGGTGATTCAGCCCCTTT

STOP

TACTAAAATAAATGCTACAATTGACTTAACAGGATGAAATTTTAGTAAAAGCAAAGCGAG

TGAGGAAGATGGCAACGATCAGAGAAGTGCCAAGGCAGCCGGCGTGTCGCTAGCGACGGT

TTCCCGGGTCTTGAACTATGACCAGACCCTGTCAGTCAATGAGGCAACCCGGCAGATTTG

FIG. 2—*Continued*

TABLE 1. Comparison of some physical properties of the sequenced β-galactosidases

Source of enzyme	No. of amino acids	Mol wt	Avg residue mol wt	Empirical net charge[a]
L. bulgaricus	1,006	113,915	113.2	−31
E. coli lacZ	1,023	116,351	113.7	−23
E. coli ebgA	963	109,652	113.8	−8
K. pneumoniae lacZ	1,033	117,385	113.6	−14
B. stearothermophilus bgaB	672	78,052	116.2	−8

[a] The net charge is the sum of the potential positive residues (assuming that half of the histidines are charged) minus the sum of the potential negative amino acids.

dase genes (results not shown). Overall, there is a 47.9% similarity between these two genes. In general, gaps in the alignment of the nucleotide sequences correspond directly to gaps in the amino acid sequence alignments. As might be expected, gaps in the nucleotide sequence that cause a shift in the reading frame occur in regions of low similarity in the amino acid sequences (*L. bulgaricus* amino acids 701 to 755 are in such a frame-shifted region).

Conserved amino acids. The putative active-site glutamate in *E. coli lacZ* β-galactosidase (Glu-461), identified by active-site-directed inhibitors (13), is indeed conserved in the *L. bulgaricus* enzyme (also in the other two enzymes [Fig. 3]) and occurs in high-similarity region IV. Since this glutamate is only one of eight that are conserved in all four β-galactosidases, and since there are also nine conserved aspartate residues, it is possible that any of these could furnish the carboxylate group needed for catalysis.

Since no homology was found between the *L. bulgaricus* β-galactosidase and the *B. stearothermophilus ebgA* gene product, a search was made for regions of local similarity. From our computer search, amino acids 135 to 148 of the *B. stearothermophilus* β-galactosidase do have some similarity to the putative active-site regions of the other enzymes (residues 451 to 464 of the *L. bulgaricus* β-galactosidase). This suggests that Glu-148 may be the active-site carboxylate in the *B. stearothermophilus* enzyme, which is different from the one suggested by Hirata et al. (14). If these regions are indeed similar, a consensus sequence of KNHPXIXI WXLXNES may be the basis for a generic β-galactosidase catalytic site in which E is the active-site glutamate and X is an undefined amino acid (in most cases a small polar or nonpolar amino acid).

The iodination of tyrosine residues (8, 15) and the substitution of tyrosines with fluorotyrosine (27) have implicated a tyrosine residue as the general acid and general base catalyst required for lactose hydrolysis. Although Tyr-253 and Tyr-285 (*E. coli lacZ*) are rapidly iodinated (8, 15), they are probably not the catalytic residues, since they are not conserved in the other β-galactosidases. There are 12 tyrosines (out of a total of 27 in the *K. pneumoniae* enzyme, which has the fewest tyrosines) that are conserved in all four enzymes; 6 of these occur in high-similarity regions (II, III, V, and VII). It has been suggested that Tyr-503 (in *E. coli lacZ*) is an active-site group, since it is adjacent to a methionine which can be labeled with active-site-directed reagents (12, 23). This methionine itself does not seem to be catalytically important, since it can be replaced with a norleucine without a significant loss in enzyme activity (23). This Met-Tyr pair is conserved in all four β-galactosidases, but it is in a region of rather low similarity. In all four enzymes there is an arginine residue one amino acid away from this tyrosine; however, in the *L. bulgaricus* β-galactosidase it is toward the N-terminal side, whereas in the others in is on the C-terminal side.

TABLE 2. Percent similarities among β-galactosidases

Enzyme source	% Similarity with enzyme from:		
	E. coli lacZ	*E. coli ebgA*	*K. pneumoniae lacZ*
E. coli ebgA	32.6		
K. pneumoniae lacZ	58.6	31.4	
L. bulgaricus	33.7	30.1	32.6

It has been suggested that His-464 might be the positively charged group near the active-site Glu-461 in the *E. coli lacZ* β-galactosidase (13), although it is also possible that a different amino acid (e.g., lysine) is involved. From our similarity data, and assuming that a positively charged histidine is indeed necessary for catalysis, His-464 seems a poor choice, since there is no corresponding histidine in any of the other β-galactosidases. There are a total of 5 conserved histidines (out of a total of 18 in the *L. bulgaricus* enzyme, which has the fewest histidines); 4 of these occur in high-similarity regions (III, IV, and V), making them the more likely candidates for involvement in lactose hydrolysis.

The conservation of one disulfide bond is possible in these β-galactosidases, since only two cysteines are conserved. There are only five cysteines in the *L. bulgaricus* β-galactosidase, and so there can be no more than two disulfide bonds in this protein.

Codon usage. The codon usage for the β-galactosidase gene (Table 3) has some notable differences from that found in *E. coli* (33). Specifically, several minor codons in *E. coli* (CGG, GCC, and GUC) are used frequently in the *L. bulgaricus* β-galactosidase gene, whereas some *E. coli* major codons (CGU, GGU, and GUA) are used very little. There is a preference for G and C residues in codon 3 in the *L. bulgaricus* gene, such that the G+C content in position 3, 62.8%, is even higher than that of the entire gene (52.5%). No significant trends in codon usage for the *Lactobacillus* genus can be seen by comparing the codon usage of the *L. bulgaricus* β-galactosidase gene with that of the histidine decarboxylase gene from *Lactobacillus* strain 30a (32) or the dihydrofolate reductase gene from *L. casei* (1).

DISCUSSION

From the high degree of similarity among the *L. bulgaricus*, *E. coli lacZ*, *E. coli ebgA*, and *K. pneumoniae* β-galactosidases, it is likely that all these enzymes evolved from a common ancestral gene. The *bgaB* β-galactosidase from *B. stearothermophilus*, however, is not homologous to these enzymes (14). From amino acid sequence alignments, we have been able to define seven regions of high similarity among the divergent β-galactosidases. In particular, it is striking that the putative active-site residues in the *E. coli lacZ* β-galactosidase (Glu-461 and Tyr-503) are conserved in all the enzymes studied so far.

Recently, Edwards et al. (9) reported that cleavage by chymotrypsin between Trp-585 and Ser-586 completely inactivated the *E. coli lacZ* β-galactosidase, whereas cleavage by elastase between Ala-732 and Ala-733 had no effect on enzyme activity. Chymotryptic cleavage is near the border of high-similarity region VI, whereas the elastase site is in a region of low similarity (Fig. 3). This suggests that amino

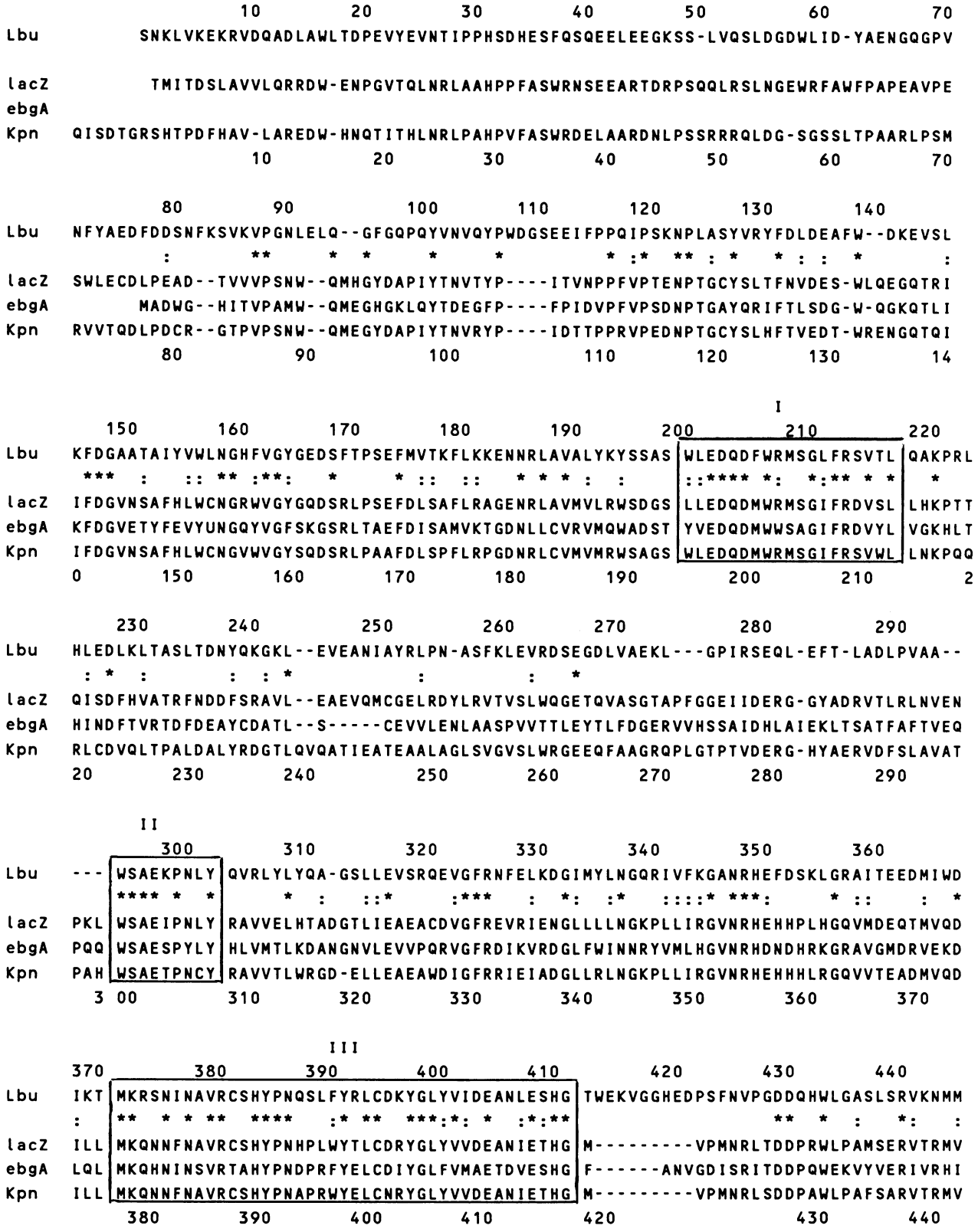

FIG. 3.

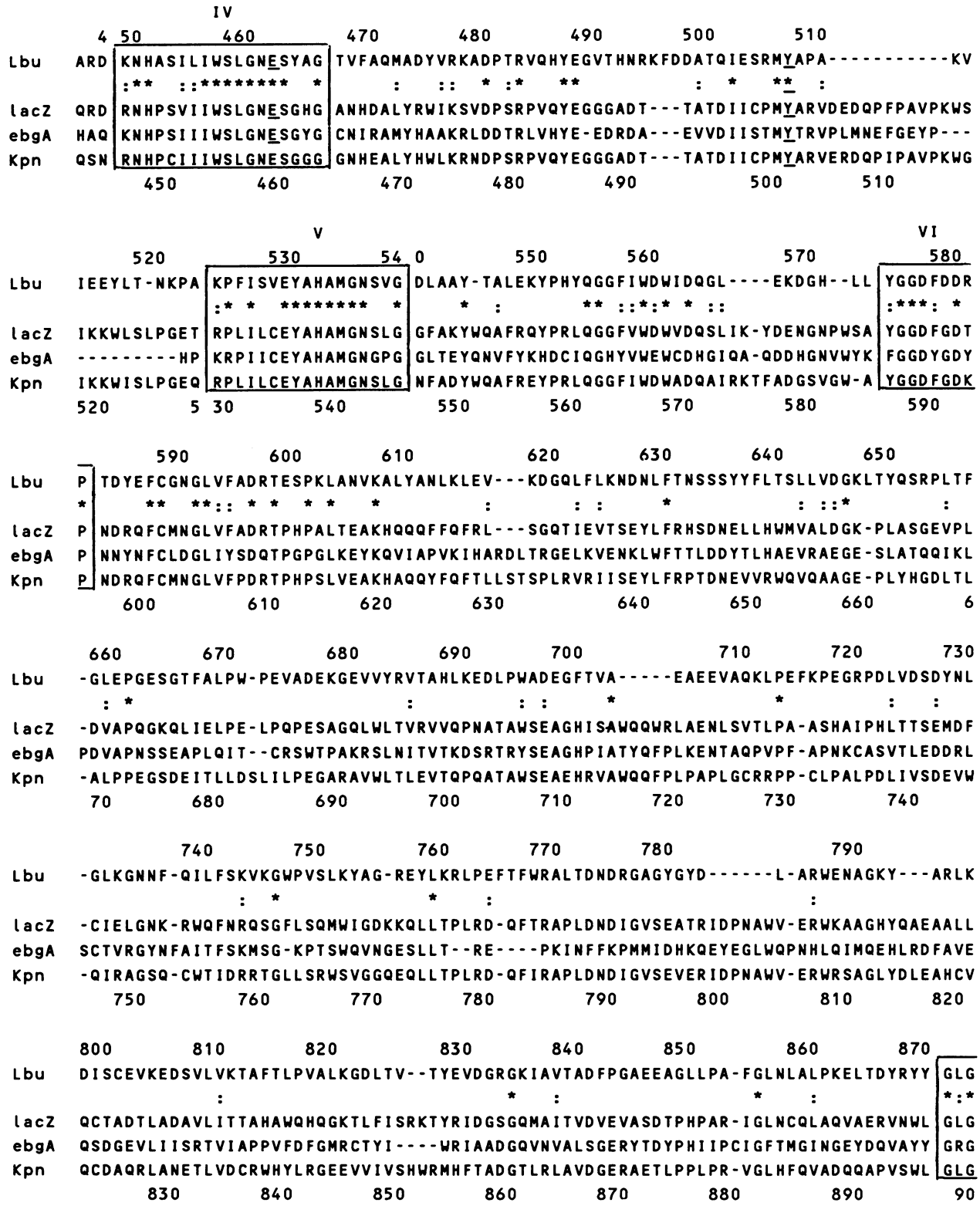

FIG. 3—*Continued*

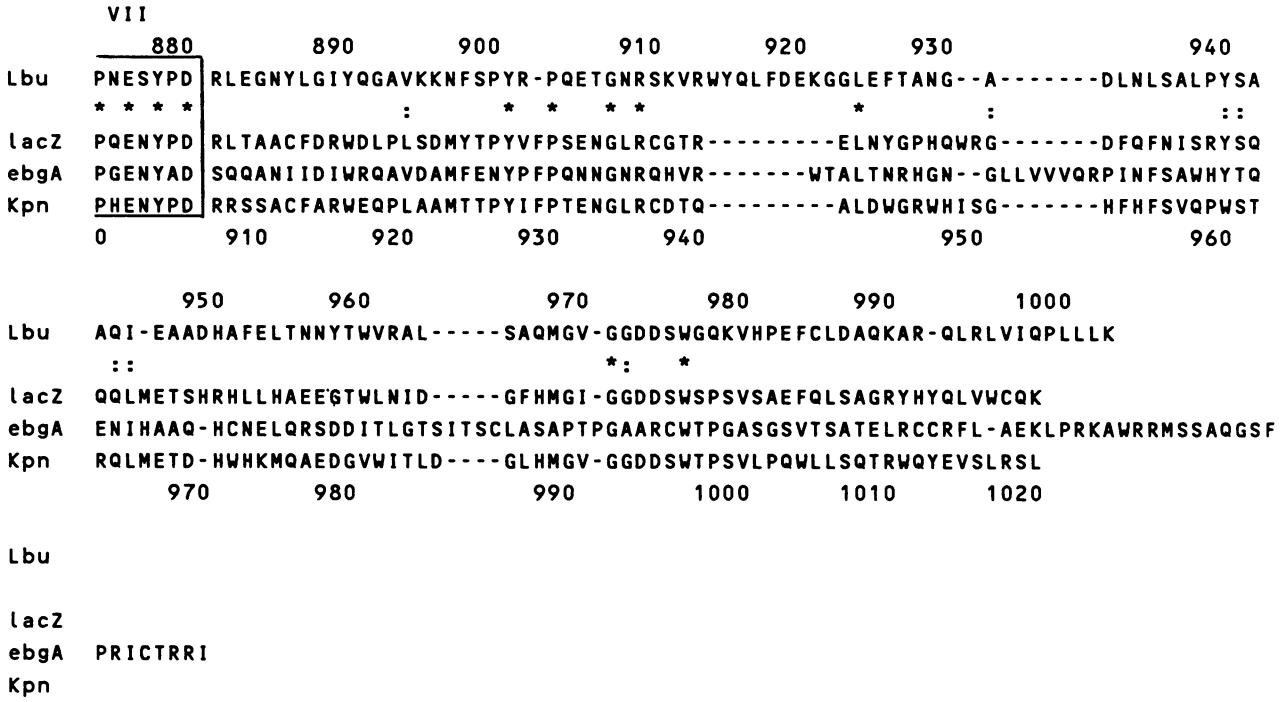

FIG. 3. Amino acid sequence alignment. From the computer alignments used to generate the data in Table 2, the amino acid sequences (from top to bottom) of the *L. bulgaricus* (Lbu), *E. coli lacZ* (lacZ), *E. coli ebgA* (ebgA), and *K. pneumoniae lacZ* (Kpn) β-galactosidases were aligned manually. The amino acids for the *L. bulgaricus* and *E. coli lacZ* enzymes are numbered above and below the alignments, respectively. Identical (*) or conserved (:) amino acids in all four sequences are indicated. Seven regions of high similarity (>50% identical residues in eight or more consecutive amino acids) are indicated by boxes and Roman numerals. The putative active-site residues (Glu-464 and Tyr-509 in the *L. bulgaricus* sequence) are underlined.

acid sequences in high-similarity regions are indeed likely to be important for enzyme function.

By producing *E. coli* mutants that were unable to grow on lactose, Langridge (18) characterized β-galactosidase mutants having a lower affinity for substrate. The mutations in these mutants were genetically mapped to five separate regions on the β-galactosidase gene. Four of the five regions mapped by Langridge coincide well with areas of high similarity noted in our computer alignments (regions I, III, VII, and the amino acids around position 975 in Fig. 3). Somewhat surprisingly, high-similarity region IV, which includes the putative active-site residue, Glu-461, was not mapped by Langridge. The remaining region identified by Langridge maps to an area of relatively low similarity among

TABLE 3. Codon usage of the *L. bulgaricus* β-galactosidase gene

No.	Codon	Amino acid	No.	Codon	Amino acid	No.	Codon	Amino acid	No.	Codon	Amino acid
26	UUU	Phe	9	UCU	Ser	18	UAU	Tyr	1	UGU	Cys
23	UUC	Phe	16	UCC	Ser	36	UAC	Tyr	4	UGC	Cys
14	UUA	Leu	5	UCA	Ser	1	UAA	OC[a]	0	UGA	OP[a]
24	UUG	Leu	1	UCG	Ser	0	UAG	AM[a]	2	UGG	Trp
12	CUU	Leu	8	CCU	Pro	3	CAU	His	2	CGU	Arg
12	CUC	Leu	6	CCC	Pro	15	CAC	His	12	CGC	Arg
4	CUA	Leu	18	CCA	Pro	11	CAA	Gln	1	CGA	Arg
42	CUG	Leu	16	CCG	Pro	27	CAG	Gln	23	CGG	Arg
14	AUU	Ile	10	ACU	Thr	19	AAU	Asn	5	AGU	Ser
13	AUC	Ile	18	ACC	Thr	30	AAC	Asn	24	AGC	Ser
0	AUA	Ile	3	ACA	Thr	22	AAA	Lys	5	AGA	Arg
12	AUG	Met	8	ACG	Thr	43	AAG	Lys	1	AGG	Arg
13	GUU	Val	28	GCU	Ala	25	GAU	Asp	11	GGU	Gly
30	GUC	Val	39	GCC	Ala	46	GAC	Asp	37	GGC	Gly
9	GUA	Val	8	GCA	Ala	62	GAA	Glu	9	GGA	Gly
11	GUG	Val	6	GCG	Ala	16	GAG	Glu	19	GGG	Gly

[a] OC, Ocher; AM, amber; OP, opal.

the enzymes (amino acids 670 to 760). Langridge also characterized *E. coli* β-galactosidase mutants with altered heat or urea sensitivities as a probe of tertiary or secondary structure (19, 20). In general, the mutations in these mutants map to regions essentially identical to those in the substrate-binding mutants. However, mutations in some heat and urea mutants map to an area encompassed by high-similarity regions IV and V and the intervening sequence, which includes Tyr-503, the putative catalytic residue. An additional area in heat-sensitive mutants corresponds approximately to amino acids 100 to 120, a moderately conserved region. This suggests that most but not all of the highly conserved sequences among this family of related enzymes are involved in functional enzyme properties (e.g., substrate binding or subunit interaction) and may serve as useful targets for mutagenesis directed at altering enzyme properties. Furthermore, regions conserved between only some members of related enzymes may indicate important differences between the proteins. For example, it has been suggested that the lack of lysine residues in the amino- and carboxy-terminal portions of the *E. coli lacZ* β-galactosidase may mean that these areas are buried and might be important for quaternary interactions (35). Indeed, the *E. coli ebgA* and *K. pneumoniae* β-galactosidases also have very few lysines in these regions. However, this is not the case for the *L. bulgaricus* enzyme. Since the oligomeric structure of the *L. bulgaricus* enzyme is unknown, but preliminary data suggest that it is not tetrameric like the *E. coli lacZ* β-galactosidase (R. M. Adams et al., unpublished results), differences in these enzymes may be expected for the amino acids involved in subunit interaction. The heat-sensitive region mentioned above (amino acids 100 to 120) may in fact be an area where amino acid substitutions can affect quaternary structure. We hope that in the absence of a three-dimensional X-ray structure for β-galactosidase, chemical modification, mutagenesis, and sequence comparisons will help define the specific amino acids that are involved in substrate binding, subunit interaction, and catalysis.

ACKNOWLEDGMENTS

We thank Bruce Chassy and Jeannette Flickinger, National Institutes of Health, Bethesda, Md., and Akram Fazel (BSN) for helpful discussions, advice, and materials supplied during the course of this work.

This work was supported in part by the Centre International de Recherche Daniel Carasso (BSN group).

LITERATURE CITED

1. **Andrews, J., G. M. Clore, R. W. Davies, A. M. Gronenborn, B. Gronenborn, D. Kalderon, P. C. Papadopoulos, S. Schäfer, P. F. G. Sims, and R. Stancombe.** 1985. Nucleotide sequence of the dihydrofolate reductase gene of methotrexate-resistant *Lactobacillus casei*. Gene **35**:217–222.
2. **Breunig, K. D., U. Dahlems, S. Das, and C. P. Hollenberg.** 1984. Analysis of a eukaryotic β-galactosidase gene: the N-terminal end of the yeast *Kluyveromyces lactis* protein shows homology to the *Escherichia coli lacZ* gene product. Nucleic Acids Res. **12**:2327–2341.
3. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7**:1513–1523.
4. **Büchel, D. E., B. Gronenborn, and B. Müller-Hill.** 1980. Sequence of the lactose permease gene. Nature (London) **283**: 541–545.
5. **Buvinger, W. E., and M. Riley.** 1985. Nucleotide sequence of *Klebsiella pneumoniae lac* genes. J. Bacteriol. **163**:850–857.
6. **Chassy, B. M.** 1987. Prospects for the genetic manipulation of lactobacilli. FEMS Rev. **46**:297–312.
7. **Eckhardt, T., J. Strickler, L. Gorniak, W. V. Burnett, and L. R. Fare.** 1987. Characterization of the promoter, signal sequence, and amino terminus of a secreted β-galactosidase from "*Streptomyces lividans*." J. Bacteriol. **169**:4249–4256.
8. **Edwards, L. A., and R. E. Huber.** 1986. A detailed examination of the iodination of β-galactosidase stoichiometric inactivation by nonspecific iodination. Biochem. Cell Biol. **64**:523–527.
9. **Edwards, L. A., M. R. Tian, R. E. Huber, and A. V. Fowler.** 1988. The use of limited proteolysis to probe interdomain and active site regions of β-galactosidase (*Escherichia coli*). J. Biol. Chem. **263**:1848–1854.
10. **Fitch, W. M., and T. F. Smith.** 1983. Optimal sequence alignments. Proc. Natl. Acad. Sci. USA **80**:1382–1386.
11. **Fowler, A. V., and P. J. Smith.** 1983. The active site regions of *lacZ* and *ebg* β-galactosidases are homologous. J. Biol. Chem. **258**:10204–10207.
12. **Fowler, A. V., I. Zabin, M. L. Sinnott, and P. J. Smith.** 1978. Methionine 500, the site of covalent attachment of an active site-directed reagent of β-galactosidase. J. Biol. Chem. **253**: 5283–5285.
13. **Herrchen, M., and G. Legler.** 1984. Identification of an essential carboxylate group at the active site of *lacZ* β-galactosidase from *Escherichia coli*. Eur. J. Biochem. **138**:527–531.
14. **Hirata, H., T. Fukazawa, S. Negoro, and H. Okada.** 1986. Structure of a β-galactosidase gene of *Bacillus stearothermophilus*. J. Bacteriol. **166**:722–727.
15. **Huber, R. E., A. V. Fowler, and I. Zabin.** 1982. Inactivation of β-galactosidase by iodination of tyrosine-253. Biochemistry **21**:5052–5055.
16. **Itoh, T., M. Ohashi, T. Toba, and S. Adachi.** 1980. Purification and properties of β-galactosidase from *Lactobacillus bulgaricus*. Milchwissenschaft **35**:593–597.
17. Kalnins, A., K. Otto, U. Rüther, and B. Müller-Hill. 1983. Sequence of the *lacZ* gene of *Escherichia coli*. EMBO J. **2**:593–597.
18. **Langridge, J.** 1968. Genetic evidence for the disposition of the substrate binding site of β-galactosidase. Proc. Natl. Acad. Sci. USA **60**:1260–1267.
19. **Langridge, J.** 1968. Genetic and enzymatic experiments relating to the tertiary structure of β-galactosidase. J. Bacteriol. **96**: 1711–1717.
20. **Langridge, J.** 1974. Genetic and enzymatic experiments relating to the quaternary structure of β-galactosidase. Aust. J. Biol. Sci. **27**:321–330.
21. **London, J.** 1976. The ecology and taxonomic status of the lactobacilli. Annu. Rev. Microbiol. **30**:279–301.
22. **Mills, D. R., and F. R. Kramer.** 1979. Structure-independent nucleotide sequence analysis. Proc. Natl. Acad. Sci. USA **76**:2232–2235.
23. **Naider, F., Z. Bohak, and J. Yariv.** 1972. Reversible alkylation of a methionyl residue near the active site of β-galactosidase. Biochemistry **11**:3202–3207.
24. **Needleman, S. J., and C. D. Wunsch.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**:443–453.
25. **Platt, T.** 1986. Transcription termination and the regulation of gene expression. Annu. Rev. Biochem. **55**:339–372.
26. **Poolman, B., T. J. Royer, S. E. Mainzer, and B. F. Schmidt.** 1989. Lactose transport system of *Streptococcus thermophilus*: a hybrid protein with homology to the melibiose carrier and enzyme III of phosphoenolpyruvate-dependent phosphotransferase systems. J. Bacteriol. **171**:244–253.
27. **Ring, M., I. M. Armitage, and R. E. Huber.** 1985. m-Fluorotyrosine substitution in β-galactosidase; evidence for the existence of a catalytically active tyrosine. Biochem. Biophys. Res. Commun. **131**:675–680.
28. **Rodriguez, H., W. J. Kohr, and R. N. Harkins.** 1984. Design and operation of a completely automated Beckman micro sequencer. Anal. Biochem. **134**:538–547.
29. **Sanger, F., S. Nicklen, and A. R. Coulsen.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
30. **Southern, E. M.** 1975. Detection of specific sequences among

DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98**:503–517.

31. **Stokes, H. W., P. W. Betts, and B. G. Hall.** 1985. Sequence of the *ebgA* gene of *Escherichia coli*: comparison with the *lacZ* gene. Mol. Biol. Evol. **2**:469–477.
32. **Vanderslice, P., W. C. Copeland, and J. K. Robertus.** 1986. Cloning and nucleotide sequence of wild type and a mutant histidine decarboxylase from *Lactobacillus* 30a. J. Biol. Chem. **261**:15168–15191.
33. **Varenne, S., J. Buc, R. Lloubes, and C. Lazdunski.** 1984. Translation is a non-uniform process. J. Mol. Biol. **180**:549–576.
34. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33**:103–119.
35. **Zabin, I., and A. V. Fowler.** 1980. β-Galactosidase, the lactose permease protein, and thiogalactoside transacetylase, p. 89–121. *In* J. Miller and W. Reznikoff (ed.), The operon. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.