THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 288, NO. 18, pp. 12993–13005, May 3, 2013

Structural Explanation for Allolactose (*lac* Operon Inducer) Synthesis by *lacZ* β-Galactosidase and the Evolutionary Relationship between Allolactose Synthesis and the *lac* Repressor[S]

Received for publication, January 23, 2013, and in revised form, March 6, 2013 Published, JBC Papers in Press, March 13, 2013, DOI 10.1074/jbc.M113.455436

Robert W. Wheatley, Summie Lo, Larisa J. Jancewicz, Megan L. Dugdale, and Reuben E. Huber[1]

From the Division of Biochemistry, Faculty of Science, University of Calgary, Calgary, Alberta T2N 1N4, Canada

Background: Synthesis of allolactose (*lac* operon inducer) from lactose requires a site for clasping glucose as an acceptor on β-galactosidase.
Results: The structure of the glucose site was defined, and its evolutionary conservation was determined.
Conclusion: The glucose binding site defines an "allolactose synthesis motif" that is co-selected with *lac* repressors.
Significance: Novel insights into evolutionary adaptations for regulation by allolactose are presented.

β-Galactosidase (*lacZ*) has bifunctional activity. It hydrolyzes lactose to galactose and glucose and catalyzes the intramolecular isomerization of lactose to allolactose, the *lac* operon inducer. β-Galactosidase promotes the isomerization by means of an acceptor site that binds glucose after its cleavage from lactose and thus delays its exit from the site. However, because of its relatively low affinity for glucose, details of this site have remained elusive. We present structural data mapping the glucose site based on a substituted enzyme (G794A-β-galactosidase) that traps allolactose. Various lines of evidence indicate that the glucose of the trapped allolactose is in the acceptor position. The evidence includes structures with Bis-Tris (2,2-bis(hydroxymethyl)-2,2′,2″-nitrilotriethanol) and L-ribose in the site and kinetic binding studies with substituted β-galactosidases. The site is composed of Asn-102, His-418, Lys-517, Ser-796, Glu-797, and Trp-999. Ser-796 and Glu-797 are part of a loop (residues 795–803) that closes over the active site. This loop appears essential for the bifunctional nature of the enzyme because it helps form the glucose binding site. In addition, because the loop is mobile, glucose binding is transient, allowing the release of some glucose. Bioinformatics studies showed that the residues important for interacting with glucose are only conserved in a subset of related enzymes. Thus, intramolecular isomerization is not a universal feature of β-galactosidases. Genomic analyses indicated that *lac* repressors were co-selected only within the conserved subset. This shows that the glucose binding site of β-galactosidase played an important role in *lac* operon evolution.

The classical model of the *lac* operon was developed based on the regulation (1) of β-galactosidase (EC 3.2.1.23) production in *Escherichia coli*. Two key genes in the *lac* operon are *lacZ*, which codes for β-galactosidase, and *lacI*, which codes for the *lac* repressor. The *lac* operon is not regulated directly by lactose (D-Gal-(β1–4)-D-Glc). Instead allolactose (D-Gal-(β1–6)-D-Glc) is an inducer (2, 3), binding to the *lac* repressor, stopping repression, and allowing the transcription of *lacZ* and related genes. Besides its function in hydrolyzing lactose, β-galactosidase synthesizes allolactose (3, 4) (Fig. 1*A*). The presence of allolactose is, however, transient because allolactose is also a substrate and is eventually hydrolyzed (Fig. 1*A*) (5).

E. coli β-galactosidase is a tetramer with four identical subunits, each of which is 1023 amino acids (6). The structure of the enzyme has been determined, and many molecular details of the reaction have been elucidated (7–9). The enzyme has broad specificity and in addition to lactose is capable of hydrolyzing a large variety of β-D-galactosides. Substrate binding occurs in two steps. Lactose initially binds in what has been described as the "shallow" binding mode (8), located parallel to and making interactions with Trp-999. Binding in this mode is nonproductive, and in order for the reaction to proceed the substrate must move ∼3 Å and rotate ∼90°. In this new position, described as the "deep" binding mode, the Gal[2] of lactose forms interactions with Trp-568 and is correctly positioned for reaction *vis à vis* two key catalytic residues, Glu-461 and Glu-537.

Specific conformational changes occur during catalysis. The side chain of Phe-601 rotates ∼60° while interacting with the C6 of the Gal, and a mobile loop composed of residues 794–803 can move up to ∼11 Å and close over the active site. Three residues, Arg-599 (10), Met-542 (11), and Glu-808 (12), are not part of the loop but are important in loop conformation. Structural studies (8) have shown that the loop and Phe-601 of the native enzyme are in the "closed" conformation when some transition state analogs are bound. Otherwise, when no ligand

[S] This article contains supplemental Figs. 1 and 2.

The atomic coordinates and structure factors (codes 4DUW, 4DUV, and 4DUX) have been deposited in the Protein Data Bank (http://wwpdb.org/).

[1] To whom correspondence should be addressed: Dept. of Biological Sciences, University of Calgary, 2500 University Dr. NW, Calgary, AB Canada T2N 1N4. Tel.: 403-220-7273; Fax: 403-289-9311; E-mail: huber@ucalgary.ca.

[2] The abbreviations used are: Gal, galactose; Bis-Tris, 2,2-bis(hydroxymethyl)-2,2′,2′′-nitrilotriethanol; LB, Luria-Bertini; *o*NP, *o*-nitrophenol; *o*NPG, *o*-nitrophenyl-β-D-galactopyranoside; TES, *N*-tris(hydroxymethyl) methyl-2-aminoethane-sulfonic acid; aLRT, approximate likelihood-ratio test.

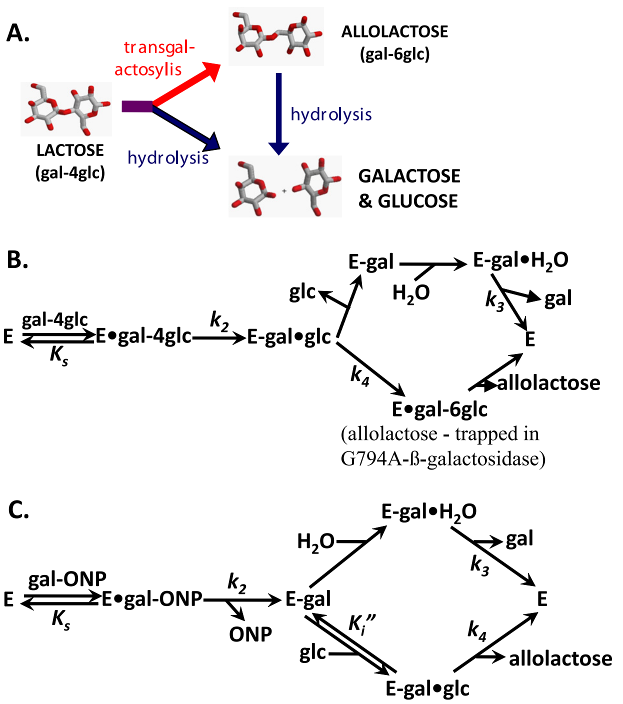

FIGURE 1. **The reactions of *β*-galactosidase.** *A*, shown is the general scheme presenting the structures of lactose and allolactose. The concurrent hydrolysis and intramolecular transgalactosylis reactions are diagrammed. Allolactose production is only transitory because it is also hydrolyzed to Gal and Glc by *β*-galactosidase. *B*, shown is the reaction mechanism when lactose is the substrate. *C*, shown is the reaction mechanism when *o*NPG and Glc are substrates. The addition of glucose is used to study the transgalactosylation reaction. For *B* and *C*, K_s is the dissociation constant of the enzyme·substrate complex, k_2 is the rate constant of the first displacement reaction, and k_3 is the rate constant for the second displacement reaction when water reacts. For the transgalactosylation reaction with Glc that forms allolactose, the rate constant is k_4. K_i'' is the dissociation constant for Glc dissociating from the covalent complex between the enzyme and Gal (*E-Gal·Glc*).

is bound or when substrate is bound, Phe-601 and the loop are in the "open" conformation. Specific amino acid substitutions (11–13) alter the loop open-closed equilibrium.

β-Galactosidase is a retaining glycosyl hydrolase with a double displacement mechanism. In the first displacement (k_2 in Fig. 1*B*), a nucleophile, Glu-537, reacts with the substrate anomeric carbon to break the glycosidic bond and form a covalent bond with Gal (14). Glu-461 acts as an acid catalyst (15), and the reaction probably proceeds via an oxocarbenium ion-like transition state. In the second displacement an acceptor hydroxyl group acts as a nucleophile to break the enzyme-Gal covalent bond. When water is the acceptor, hydrolysis occurs (k_3), and Gal is produced. When Glc is the acceptor, allolactose is formed (k_4).

Most, if not all, *β*-galactosidases (including *lacZ* *β*-galactosidase) catalyze intermolecular allolactose synthesis. Glc simply competes with water to form allolactose (16). The intermolecular route requires high concentrations of Glc (either added to the reaction or produced from the hydrolysis of lactose), and the proportion of the overall reaction that forms allolactose is dependent on the Glc concentration. In a cell with an uninduced *lac* operon, this reaction route would not be sufficient for operon induction. In contrast, when *lacZ* *β*-galactosidase reacts with lactose, ~50% of the initial product is allolactose (4). This high proportion is found even at low initial lactose concentrations, an observation incompatible with intermolecular transfer. Instead, an "intramolecular" transfer of the Gal moiety of lactose from the Glc O4 to O6 occurs without the Glc leaving the active site. Kinetic studies (4, 17) have shown that allolactose synthesis in *lacZ* *β*-galactosidase is facilitated by slow release of Glc after cleavage from lactose by the first displacement reaction. This putative Glc acceptor site has moderate specificity for Glc, with a dissociation constant of 17 mM (17). Affinity is a balance between the need for intramolecular allolactose synthesis for operon induction, which would be facilitated by high Glc affinity, and the need to produce Glc and Gal for glycolysis, which requires low Glc affinity.

The means by which *β*-galactosidase synthesizes allolactose is important to help understand the regulation of the *lac* operon. To date, the moderate affinity of Glc has prevented mapping of the putative Glc site. This report begins by presenting the first detailed crystallographic study of how *β*-galactosidase binds Glc at the acceptor site and identifies the residues involved. Kinetic studies in conjunction with site-directed mutagenesis confirmed that these residues are important for Glc affinity. These findings were then used for a phylogenetic analysis, which indicated that residues important for the intramolecular synthesis of allolactose are conserved in *β*-galactosidases of only a relatively small group of known organisms. Furthermore, *lac* repressors were only identified in organisms where this intramolecular allolactose production can occur. Taken together, these results suggest that allolactose production by *β*-galactosidase is not merely a fortuitous side reaction, as often portrayed in the literature and text books. Instead, allolactose production and hence the regulatory mechanism in the *lac* operon are features co-selected through evolution.

EXPERIMENTAL PROCEDURES

Chemicals

Complete Protease Inhibitor Tablets and ampicillin were from Roche Diagnostics. Na_2HPO_4 and NaH_2PO_4 were from BDH Inc., Toronto, ON. Sephacryl 300, *o*NPG, and L-arabinose were from MP Biomedicals, Solon, OH. $MgSO_4$ was from Mallinckrodt Baker. Imidazole, BSA, EDTA, Bis-Tris, PEG-8000 and TES were from Sigma. NaCl was from EMD Chemicals, Gibbstown, NJ. DTT was from Roche Applied Science.

Media and Buffers

Media—LB broth consisted of 5 g/liter yeast extract, 10 g/liter Tryptone, 10 g/liter NaCl, 0.05 g/liter ampicillin, pH 7.5 at 37 °C. LB plates contained LB broth supplemented with 15 g/liter agar.

Assays—TES buffer consisted of 30 mM TES, 145 mM NaCl, 10 mM $MgSO_4$, 0.1 mM EDTA pH 7.0 at 25 °C.

Chromatography—Native binding buffer was 200 mM sodium phosphate, 5 mM NaCl, pH 7.8; native wash buffer was 200 mM sodium phosphate, 5 mM NaCl, pH 6.0.

Crystallization—Crystallization buffer was 100 mM Bis-Tris, pH 6.5 (14 °C), 100 mM NaCl, 200 mM $MgCl_2$, 10 mM DTT. Mother liquor was 100 mM Bis-Tris, pH 6.5 (14 °C), 100 mM NaCl, 200 mM $MgCl_2$, 10 mM DTT, and 10% (w/w) PEG 8000.

Strain and Plasmid

β-Galactosidase was expressed in *E. coli* strain LMG194 (F-Δ*lac*X74 Gal E *thi rps*L ΔphoA (*Pvu*) Δ*ara*714 *leu*::Tn10 (Invitrogen^{TM}) containing the pBAD/His/*lac*Z plasmid (Invitrogen^{TM}). The following substituted β-galactosidases were created by Bio S&T Inc., Montreal, QC: N102A, H418N, N460S, K517A, G794A, S796A, and E797A.

Cultures

Single colonies of *E. coli* cells (from LB agar plates grown overnight at 37 °C) containing the desired plasmid were inoculated into 20 ml of LB broth. The starter culture was grown for 16 h (37 °C, 250 rpm) and then used to inoculate a Fernbach flask (LB broth, 1 liter). This culture was incubated for 2 h (37 °C, 250 rpm) and then induced with 10 ml of sterile L-arabinose (20% w/v). Growth was continued overnight. Cultures were harvested by centrifugation at 2000 × *g* for 10 min (4 °C).

Purification of β-Galactosidases

Pellets of the *E. coli* cells were resuspended in 3 ml of native binding buffer supplemented with one-half a Complete Mini protease tablet. The cells were disrupted with two passes through a French press (800 p.s.i.). The homogenate was clarified by centrifugation (11,200 × *g*, 55 min) and then applied to a nickel Probond column (3.4 ml, Invitrogen) pre-equilibrated with native binding buffer. The column was washed sequentially with several column volumes of native wash buffer and then several column volumes of native wash buffer supplemented with 50 mM imidazole. Next, β-galactosidase was eluted with native wash buffer supplemented with 500 mM imidazole. Fractions containing the highest activity were pooled and dialyzed exhaustively against TES buffer (4 °C). At this point the enzyme could be used for kinetic studies. However, crystallization required further purification. After being concentrated to ~2 ml, the enzyme was applied to a Sephacryl S-300 column (300 ml) equilibrated with TES buffer and eluted at 2 ml/min. Fractions with the highest activity were pooled and concentrated with ammonium sulfate (50% saturation, 25 °C). The recovered protein was dialyzed exhaustively against crystallization buffer (4 °C). Aliquots were frozen in liquid nitrogen and stored at −86 °C until needed.

Crystallography

Diffraction quality crystals were obtained by microseeding methods similar to those of Juers *et al.* (9). Seed crystals were obtained from starting crystals suspended in mother liquor and disrupted with a Seed Bead kit (Hampton Research). Different dilutions of seed crystals were utilized; the dilution required was determined empirically. Crystals were grown at 14 °C by hanging drop vapor diffusion over 1 ml of mother liquor in 24-well plates (Hampton Research). The hanging drops (4 μl) consisted of equal volumes of protein (10 mg/ml in crystallization buffer) and seed solution in mother liquor. Pyramid-shaped orthorhombic crystals appeared after a few hours and reached their maximum size about 36 h after seeding. Complexes were obtained by soaking (30 min) crystals of the desired enzyme in mother liquor supplemented with 50 mM lactose, 50 mM galactal, or 50 mM L-ribose. For cryoprotection, each crystal was placed into solutions of mother liquor with increasing DMSO content (6, 12, 19, 25% v/v) for ~5 s at each concentration before being flash-cooled in liquid N_2 at 100 K.

Diffraction intensities were collected from single crystals at the Canadian Light Source beamline 08B1-1 (1.11589 Å, 100 K). Data were processed and scaled with Mosflm and Scala (18–20). Previously determined structures were used as starting models for refinement as all space group $P2_12_12_1$ crystals had isomorphous unit cell dimensions. The R_{free} sets were preserved from the starting model data. Final models were obtained by iterative cycles of model building with COOT (21) and refinement with Refmac (22). Stereochemical restraints were defined manually as this strategy produced lower R_{free} values than the Refmac automatic algorithm. Omit electron density maps were produced by simulated annealing and re-refinement with Phenix (23) using models with the acceptor ligand removed.

Enzyme Assays

Enzyme aliquots were thawed slowly in tepid water and diluted into TES buffer supplemented with BSA (10 mg/ml). The BSA was important for minimizing enzyme denaturation (24). Small aliquots of the diluted enzyme were added to assay mixtures consisting of TES buffer and the substrate, *o*NPG. Activities were measured at 25 °C using a Shimadzu UV 2101PC spectrophotometer (420 nm). The extinction coefficient of the *o*NP (product) at 420 nm is 2.67 $\text{mM}^{-1}\text{cm}^{-1}$, pH 7.0, 25 °C.

Kinetic Studies

Equation 1 describes the effect of the Glc concentration ([glucose]) on appk_{cat}. K_i'' is the equilibrium constant for dissociation of Glc from E-Gal·Glc, and k_4 refers to the reaction forming allolactose starting with E-Gal·Glc (Fig. 1*C*). The equation was derived using the kinetic scheme in Fig. 1*C* in a manner analogous to that described for similar equations by Deschavanne *et al.* (25). Estimates of K_i'' were obtained using Equation 1 below. For each substituted enzyme, appk_{cat} values (k_{cat} values in the presence of Glc) were determined using non-linear regression (PRISM 4^{TM}) of Michaelis-Menten plots obtained at a series of Glc concentrations. Six concentrations of *o*NPG were used for each analysis; three above the K_m and three below. Assays at each substrate concentration were performed in duplicate. Protein concentration was determined by the absorbance at 280 nm (extinction coefficient: 2.09 mg ml^{-1} cm^{-1}).

$$\text{app}k_{\text{cat}} = \frac{\dfrac{k_2k_3}{k_2 + k_3} + \left(\dfrac{k_2k_4}{k_2 + k_3}\right)\dfrac{[\text{glucose}]}{K_i''}}{1 + \left(\dfrac{k_2 + k_4}{k_2 + k_3}\right)\dfrac{[\text{glucose}]}{K_i''}} \qquad \text{(Eq. 1)}$$

The first term in the numerator of Equation 1 is the k_{cat} of *o*NPG expressed as a ratio of rate constants. This value as well as the values of the second term in the numerator and the second term in the denominator were determined by non-linear regression of plots of appk_{cat} *versus* [glucose]. The k_2 and k_3

values were determined separately using the k_{cat} values of *p*- and *o*-NPG (which had different rate determining steps) as well as with added acceptors (mainly methanol). The second term in the numerator was divided by the second term of the denominator, and the values of k_2 were substituted. This allowed estimation of k_4. The k_2, k_3, and k_4 values obtained in this way were substituted into the estimates of the second terms of the numerator and of the denominator, and one K_i'' estimate was obtained from each term. These two estimates were similar for each point mutant and were averaged.

Bioinformatics

General—Homologous proteins were identified by searching NCBI databases with Blast 2.2.25+ (26). Because of the over-representation of *E. coli* sequences in databases, *E. coli* sequences were excluded from all Blast searches, and the *E. coli* sequences were manually appended to the results. Sequences were aligned with ClustalX 2.1 using default parameters (27). Maximum likelihood phylogenies were estimated with PhyML 3.0 (28). Tree searching operations used the "best of NNI and SPR" algorithm. Branch support was estimated with the "aLRT" method. Sequence data were extracted and analyzed algorithmically using BioPerl (29) scripts.

Identification of β-Galactosidases with an "Allolactose Synthesis Motif"—Fourteen amino acids were singled out as comprising the allolactose synthesis motif (Lys-517, loop residues 795–803, as well as Asn-804, Ala-805, Arg-599, and Glu-808). Details of the selection process are presented under "Results." An initial test analysis of the motif used a dataset of homologous β-galactosidase sequences. First, a Blast search of the NCBI non-redundant protein database was performed using the *E. coli* β-galactosidase sequence as the query sequence. The 251 top matches were retained. Second, truncated sequences were removed. The *E. coli* enzyme has 1023 amino acids, and 245 of the 251 sequences were longer than 950 amino acids. The 6 truncated sequences (all having less than 850 residues) were discarded to produce a working set of 245 sequences. Third, the sequences were aligned with Clustal, and the number of conserved motif residues in each sequence was determined. Fourth, maximum likelihood phylogenies were estimated with PhyML. The conservation of motif residues (as determined in the previous step) in various branches of the phylogenetic tree was calculated and compared, as indicated under "Results."

lac Repressor Identification—The purpose of the repressor sequence analysis was to separate *lac* repressor homologues from other related repressors. Seven residues (Ile-79, Asn-125, Asp-149, Phe-161, Ser-193, Phe-293, and Leu-296) were selected to distinguish *lac* repressors from other repressors, as described under "Results." An initial test analysis was performed in a similar manner to that involving β-galactosidase sequences, as described above. In this case, a Blast search of the NCBI non-redundant protein database was performed using the *E. coli lac* repressor amino acid sequence as the query sequence. The 252 top matches were retained. The *E. coli lac* repressor is 360 amino acids long, and 249 of the 252 sequences were greater than 310 amino acids in length. The 3 truncated sequences (all less than 270 amino acids in length) were discarded. Analysis of the conservation of the motif residues in the retained 249 sequences then proceeded as described above.

Genomic Analysis—Complete microbial genome sequences were analyzed to determine if an organism contained a β-galactosidase catalyzing allolactose formation (defined by the allolactose synthesis motif and/or a *lac* repressor (defined by the residues identified above). 1087 genomes were searched, representing all completed genomes available from the NCBI that could be searched against protein sequences.

Datasets were obtained with Blast searches using either the *E. coli* β-galactosidase or *lac* repressor as the query sequence. Searches were performed separately against each of the 1087 genomes. For each genome, only the top scoring β-galactosidase and/or repressor sequences were retained. Additionally, for the β-galactosidase sequence set, only sequences greater than 900 amino acids in length and greater than 30% sequence identity with *E. coli* β-galactosidase were retained. For the repressor sequence set, sequences from Blast search results with a length greater than 200 amino acids and a greater than 30% sequence identity with the *E. coli lac* repressor were retained. The *E. coli* ribose repressor sequence was also added to the repressor sequence set. For each dataset (the β-galactosidase sequences and the repressors sequences), the sequences were aligned with Clustal, and the number of conserved motif residues in each sequence was determined. Maximum likelihood phylogenies were estimated with PhyML. Using the criteria established in the previous analyses, *lac* repressors and β-galactosidases containing the allolactose synthesis motif were identified. Finally, the presence of each of the identified proteins in each of the genomes was compared. The number of genomes that contained only one or both of these proteins was determined.

RESULTS

Statistics for diffraction data collection and refinement are presented in Table 1. Coordinates and structure factors were deposited in the Protein Data Bank. Accession codes are also given in Table 1. Structures showing ligand binding are presented in Fig. 2. The substitutions caused no significant differences in overall protein structure. Fig. 2 also includes omit maps showing the electron density of the ligand in the glucose acceptor site (*stereoview insets*). For each model, the *first inset* shows the Glc site ligand in the same orientation as in the main panel, whereas the *second inset* is rotated to more clearly show the ligand fit to the electron density.

Structure of G794A-β-galactosidase after Reaction with Lactose—Crystals of G794A-β-galactosidase incubated with 50 mM lactose for 30 min had an allolactose in the deep binding mode (Fig. 2*A*). The active site loop was in the closed conformation as expected with G794A-β-galactosidase (13). The presence of allolactose in the active site is consistent with allolactose being formed from the lactose and trapped *in situ*. In contrast, when either lactose or allolactose is added to catalytically inert E537Q-β-galactosidase crystals, they bind in the shallow mode with the active site loop open (8).

The electron density of the Glc of allolactose is well defined and allows for the identification of interactions of the Glc with the enzyme. Typical of sugar binding, the foundation of the Glc interaction is an aromatic residue, Trp-999. Some apparent

TABLE 1
Crystallographic data collection and refinement statistics
Data in parentheses refer to the highest resolution shell.

	G794A allolactose	G794A Bis-Tris	N460S L-ribose
Data collection statistics			
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit cell: *a, b, c* (Å)	151.8, 162.6, 203.7	150.6, 167.7, 201.6	150.6, 168.1, 201.9
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Resolution range (Å)	24.81-2.20 (2.32-2.20)	200-2.10 (2.21-2.10)	169-2.30 (2.42-2.30)
Unique reflections	235,634 (24540)	295,878 (42829)	224,130 (31819)
Completeness (%)	92.6 (66.6)	100.0 (100.0)	99.0 (97.1)
Redundancy	3.9 (1.8)	7.6 (7.6)	4.0 (3.3)
R_{merge}[a] (%)	7.7 (28.3)	11.0 (48.2)	8.9 (43.7)
Average ($I/\sigma(I)$)	10.0 (2.3)	12.7 (3.8)	8.9 (2.6)
Refinement statistics			
Resolution range (Å)	24.82-2.20 (2.257-2.200)	86.41-2.10 (2.155-2.100)	86.45-2.30 (2.360-2.300)
Number of reflections	232,012 (11,070)	291,478 (21,319)	220,754 (15,865)
Number of atoms			
Total	36,668	37,229	35,621
Protein	32,632	32,732	32,620
Solvent[b]	3,918	4,378	2,889
Ions	26	25	22
Ligand	92	96	90
R_{free} set size	3,378 (155)	4,267 (322)	3,223 (213)
R (%)	16.0 (23.3)	15.9 (20.6)	17.0 (22.2)
R_{free} (%)	21.3 (28.2)	20.6 (26.6)	22.1 (28.9)
Average B-factor (Å^2)	57	32	37
Root mean square deviation from ideal bonds (Å)	0.007	0.007	0.008
Root mean square deviation angles (°)	1.08	1.06	1.12
PDB ID	4DUW	4DUV	4DUX

[a] R_{merge} was calculated by Scala (18–20).
[b] Solvent includes atoms from water and DMSO.

C-H···π bonds (30) are shown between the partially positive hydrogens of the C3, C4, and C5 groups of Glc and the π electron cloud of Trp-999 (Fig. 2*A*, *red dashed lines*), but there are more such interactions than shown. Asn-102, Ser-796, and Glu-797 interact with the C1 hydroxyl of Glc in the β-anomeric configuration. There was no electron density suggesting the presence of an α-anomer. The C1 of Glc and the β carbon of Ser-796 are close enough for a hydrophobic interaction (Fig. 2*A*, *orange dashed line*). Asn-102 is also in position to interact with the O5 of the Glc. The two bonds with Asn-102 are longer than ideal and individually probably somewhat weak. However, the two interactions would be strengthened through cooperativity. His-418 and Lys-517 form indirect, water-bridged interactions with the Glc C3 and C4 hydroxyls, respectively. The bonds to the bridging waters are short (~2.7 Å) and at ideal H-bond angles. Glu-461 is within H-bonding distance of the allolactose glycosidic oxygen (Fig. 2*A*, *light blue dashed line*). During allolactose formation, this geometry would be correct for orienting the O6 hydroxyl of Glc for nucleophilic attack on the Gal anomeric carbon, and Glu-461 would be in position to act as a general base.

Structure of G794A-β-galactosidase with Bis-Tris—When D-galactal is incubated with native β-galactosidase, 2-deoxy-Gal becomes covalently bound to Glu-537 (8). The same reaction takes place when G794A-β-galactosidase is incubated with D-galactal. The active site loop is closed (Fig. 2*B*). Interestingly, a well defined Bis-Tris buffer molecule is located adjacent to 2-deoxy-Gal, and the residues that interact with the Bis-Tris are identical to those that bind the Glc of allolactose.

Trp-999 again forms C-H···π interactions (Fig. 2*B*, *red dashed lines*). Ser-796, Glu-797, and Asn-102 form direct interactions with the two hydroxyls at the nitrogen end of the Bis-Tris. On the opposite end of Bis-Tris, Lys-517 forms an indirect H-bond (via a water) with one hydroxyl, whereas His-418 forms a bifurcated H-bond to the other two hydroxyls. None of the five hydroxyl groups of Bis-Tris are near the anomeric carbon of the 2-deoxy-Gal, explaining why a Bis-Tris adduct with 2-deoxy-galactose is not formed.

Weak electron density for a Glc in the acceptor position was found by Juers *et al.* (8) when 2-deoxy-Gal was covalently attached to wild type β-galactosidase but only when crystallized at low (1 mM) Bis-Tris concentrations. Similar attempts with 2-deoxy-Gal and G794A-β-galactosidase were not successful (13). Even when G794A-β-galactosidase was crystallized in the presence of 500 mM Glc and 1 mM Bis-Tris, only Bis-Tris was found in the acceptor site. Bis-Tris concentrations could not be reduced further without interfering with crystallization. Bis-Tris at 1 mM must bind more tightly to the acceptor site of G794A-β-galactosidase than Glc at 500 mM.

Structure of N460S-β-galactosidase with L-Ribose—When N460S-β-galactosidase (31) is incubated with L-ribose, the crystals show that the active site loop is in the closed conformation. One L-ribose is positioned somewhat similarly to the Gal moiety of the allolactose (Fig. 2*C*, *line structure in the background*). In some ways this L-ribose behaves like a transition state analog (32). Of more importance for this present study, a second L-ribose is seen in the active site, in a similar position to the Glc moiety of allolactose and the Bis-Tris in the structures described above. Compared with those ligands, the second L-ribose is not as clearly defined by electron density. Fig. 2*C* (*insets C1* and *C2*) shows the electron density is shaped like a bowl with a high lip on one side. The three hydroxyls (O_2, O3, and O4) of L-ribose fit into this lip, and this orientation appears to be the only way that the L-ribose can be positioned. This model is also supported by the fact that the L-ribose orientation was consistent through the four active sites of the enzyme. Structures with

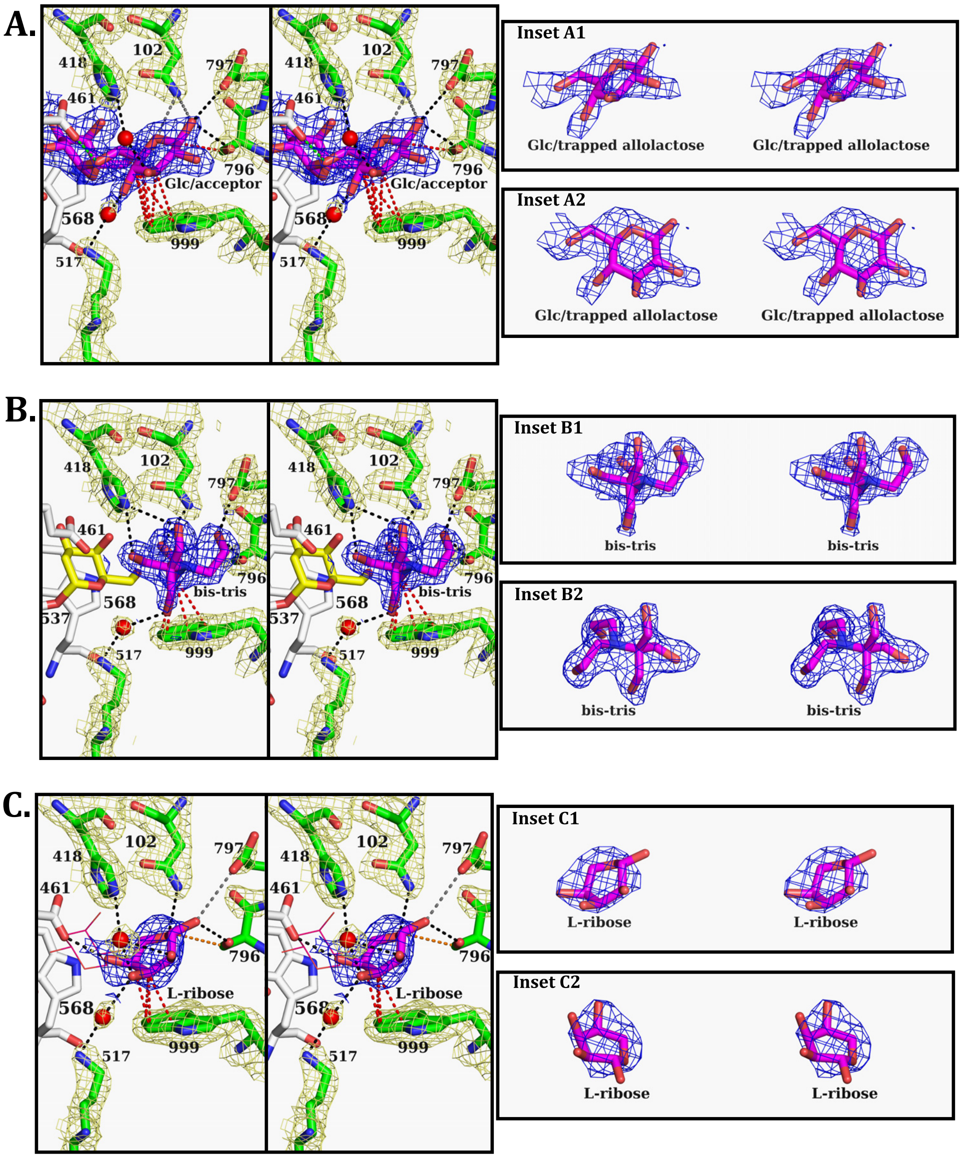

FIGURE 2. **Stereoviews of β-galactosidases with allolactose, Bis-Tris, and L-ribose ligands.** Oxygen atoms are *red*, nitrogen atoms are *blue*, and water molecules are represented by *red spheres*. Carbon atoms from residues involved in Glc binding are *green* and the electron density ($2F_o - F_c$) is *pale yellow* and contoured at 1.5σ, those from Glu-461 and Trp-568 are *white* and the electron density is left out, and the carbon atoms of allolactose, Bis-Tris, and L-ribose are *magenta*. The *dashed lines* show the interactions between the enzyme and the compounds in the Glc site: *black*, H-bonds of 3.2 Å length or less; *gray*, H-bonds between 3.2 and 3.4 Å; *red*, C-H···π bonds between hydrogens on sugars and the π electrons of Trp-999 (note that there are probably more of these, only bonds with short distances are shown); *orange*, hydrophobic interactions; *light blue*, bonds with Glu-461. *A*, shown is G794A-β-galactosidase active site structure with a trapped allolactose. The electron density of the allolactose ($2F_o - F_c$) is contoured at 1σ. *B*, shown is the G794A-β-galactosidase active site structure with a covalently bound 2-deoxy-galactose (*yellow carbons* with the electron density left out for clarity) and Bis-Tris in the Glc binding site. The electron density of Bis-Tris ($2F_o - F_c$) is contoured at 1σ. *C*, N460S-β-galactosidase with L-ribose in the position normally occupied by Gal in the deep mode and a second L-ribose in the Glc binding site is shown. The L-ribose in the Gal position is shown as *magenta lines* and without electron density so as not to obscure the rest of the structure. The electron density of the L-ribose in the acceptor site is contoured at 0.8σ. To the *right of each panel* (*A*, *B*, and *C*) are two *insets* showing simulated annealing omit maps ($F_o - F_c$) contoured at 3σ. The orientation of the omit densities in the *top inset* is identical to that in the *main panel*. The *lower inset* is re-oriented to more clearly show the ligand fit to the electron density.

L-ribose bound to this position of both native and N460S-β-galactosidase are available. The N460S structure is presented here because this structure gave the best resolution. The substitution for Asn-460 is distant from the putative Glc site.

The L-ribose in the acceptor site again forms C-H···π interactions (30) with Trp-999. Both Ser-796 and Glu-797 form H-bonds with the L-ribose C1 hydroxyl, although the Glu-797 bond to the C1 hydroxyl is longer than the bond with Ser-796. The Ser-796 β carbon forms a hydrophobic interaction with the L-ribose anomeric carbon (Fig. 3*C*, *orange dashed line*) similar to that with the Glc of the trapped allolactose. Asn-102 forms an H-bond with the L-ribose ring O5, again comparable to the interaction with Glc in the allolactose structure. His-418 interacts with a water that in turn forms H-bonds with the C2 and C3 hydroxyls. Lys-517 also forms an indirect interaction with the L-ribose C3 hydroxyl through a water ideally positioned to form H-bonds. Glu-461 forms an H-bond with the L-ribose C4 hydroxyl.

Dissociation Constants of Glc as an Acceptor—Table 2 shows that substituting for each of Asn-102, Lys-517, His-418, Ser-

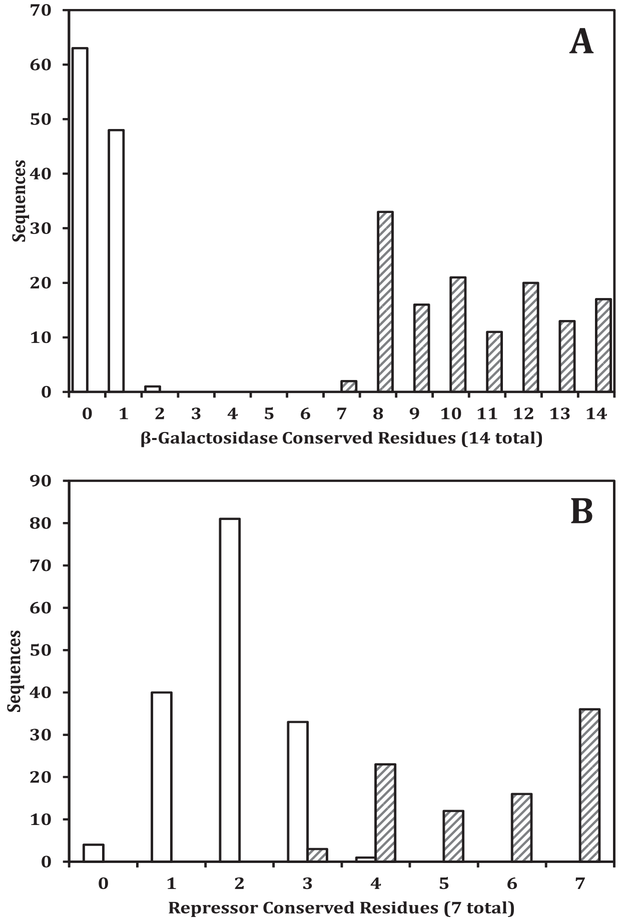

FIGURE 3. *A*, shown is the prevalence of allolactose synthesis motif residues in *β*-galactosidase sequences. The *bars with cross hatches* represent sequences identified as having an allolactose synthesis motif, and *open bars* represent the remaining sequences. *B*, shown is the prevalence of conserved *lac* repressor residues in *LacI-GalR* family repressor sequences. The *bars with cross hatches* represent sequences identified as *lac* repressors, and *open bars* represent the remaining sequences. The datasets of protein sequences were obtained as described in the text ("see Experimental Procedures").

TABLE 2

The K_i'' values are shown of ß-galactosidases with substitutions for the residues thought to be important for binding Glc for the intramolecular acceptor action of the enzyme

K_i'' is the dissociation constant of Glc from the acceptor site. k_4 values for glucose are also shown. Values are also given for N460A-ß-galactosidase, a point mutant that affects transition state binding but is not expected to affect Glc binding. A dash indicates that the value was not meaningful.

	K_i''	Fold poorer binding	k_4
	mM		s^{-1}
Native	17	—	280
Substituted enzymes			
N102A	~670	~40	185
K517A	~175	~10	135
H418N	>335	>20	50
S796A	~390	~23	20
E797A	~139	~8	285
N460A	~23	~1.3	20

796, and Glu-797 caused the K_i'' values to increase 8-fold or more. Previous studies have shown that binding of Glc is very poor upon substituting for Trp-999 (33). Also reported is a K_i'' value of a substitution (N460A-*β*-galactosidase, one of several that could have been chosen) that is unlikely to influence Glc binding. Indeed, this K_i'' does not change significantly. In addition, the table shows the k_4 values. The differences in the k_4 values of the substituted enzymes from the k_4 of the native enzyme do not correlate with the binding differences. Some of the residues have other roles that could affect the k_4 value (*e.g.* His-418, Asn-102, Ser-796) or slight differences in positioning of the Glc could alter k_4.

Bioinformatics

Identification of β-Galactosidases with an Allolactose Synthesis Motif—To identify *β*-galactosidases expected to produce allolactose using the intramolecular mechanism, 14 amino acids important for Glc binding were selected as comprising an allolactose synthesis motif. Based on the structural and kinetic findings in this study, Lys-517, Ser-796, and Glu-797 were selected because these residues appear to function primarily in Glc binding. In addition, two of these residues, Ser-796 and Glu-797, are part of the active site loop and are only in contact with Glc when the loop is closed. Thus the rest of the loop (amino acids 795 and 798–803) and residues known to function in the loop conformational change (Arg-599, Asn-804, Ala-805, and Glu-808) were also selected. The structures described above indicated that His-418, Asn-102, and Trp-999 are also important for binding Glc. However, these three residues were not useful for uniquely identifying *β*-galactosidases that synthesize allolactose as they have other important mechanistic roles (8, 33–35) and are highly conserved among all *β*-galactosidases examined.

An initial "test analysis" of the motif was done using a set of 245 homologous *β*-galactosidase sequences[3] obtained from the NCBI non-redundant protein database. Results are shown in the phylogenetic tree in supplemental Fig. 1*A*. Conservation of the motif residues was confined to one branch of the phylogenetic tree (branch A, aLRT branch support 1.00). Within the 133 sequences in branch A, most of the motif residues were highly conserved. However, little conservation of the 14 residues was found among the 112 sequences not part of branch A. For example, Ser-796, Glu-797, Asp-802, and Pro-803 were >95% conserved in branch A but averaged only 1.6% conservation in the remainder of the tree.

Two other observations obtained from this analysis are significant. First, both Lys-517, whose only known function is to bind Glc, and Arg-599, whose only known function involves loop closure (10), were highly conserved (>95%) in branch A but not conserved at all (0%) in the remaining sequences. The co-conservation of these two specific amino acids (Lys-517 and Arg-599) in addition to the co-conservation of the other 12 amino acids would be unlikely unless the residues directly involved in binding Glc (Lys-517, Ser-796, and Glu-797), the rest of the loop residues, and residues important for loop mobility share a common function. Second, all enzymes in branch A of the tree had at least 7 of the 14 motif residues strictly conserved (Fig. 3*A*), with an average of about 11. In addition, in

[3] All sequences in this dataset were predicted to have *β*-galactosidase activity based on the sequence homology with the *E. coli β*-galactosidase and particularly the conservation of Asp-201, Phe-601, and Asn-604, residues known to be conserved in this family of *β*-galactosidases but not in related glycohydrolases.

each case many of the 14 amino acids that were not strictly conserved were replaced by homologous amino acids. Branch A did not contain any organisms with enzymes having less than 7 of the 14 residues strictly conserved, and the remaining 112 enzymes outside of branch A had 2 or less of the 14 residues conserved. No enzymes had 3–6 residues conserved. Based on these analyses, the strict conservation of at least 7 of the 14 motif residues were used as the criterion to identify β-galactosidases capable of intramolecular allolactose synthesis (the allolactose synthesis motif).

Identification of lac Repressors—Repressor sequences were analyzed to determine criteria for identifying *lac* repressor homologs from among the larger family of *LacI-GalR* repressor proteins. Residues in the effector binding site were used as distinguishing features. Although a structure of allolactose complexed with *lac* repressor has not been reported, effector binding has been studied with the *lac* repressors having isopropyl 1-thio-β-D-galactopyranoside and *o*NPG bound (PDB IDs 2P9H and 2PE5, respectively) (36). The Gal moieties of isopropyl 1-thio-β-D-galactopyranoside and *o*NPG would define the binding of the Gal moiety of allolactose, whereas the Glc pocket is defined in repressor structures by residues near the thioisopropyl group of isopropyl 1-thio-β-D-galactopyranoside and the *o*NP portion of *o*NPG. Initial analyses, however, showed that the residues that comprise the Gal binding site were also highly conserved in ribose and other repressors. For example, structural alignments show that Asp-274 and Gln-291 are conserved in both the *E. coli lac* repressor and the *Lactobacillus acidophilus* ribose (PDB ID 3HS3) repressors. Residues that would comprise the binding pocket for the Glc portion of allolactose were more unique, presumably because allolactose is a disaccharide, whereas most of the related effectors and ribose, in particular, are monosaccharides. Inspection of the two *lac* repressor structures identified seven residues, Ile-79, Asn-125, Asp-149, Phe-161, Ser-193, Phe-293, and Leu-296, likely to be involved in Glc binding. Mutagenesis studies have also shown that these seven residues are important for effector binding to the *lac* repressor (37). Thus, these residues were selected to distinguish *lac* repressors from other repressors.

As with the β-galactosidase analysis, an initial "test" of the selected residues utilized a dataset of homologous repressors (249) obtained by a Blast search with the *E. coli lac* repressor as the query sequence. Results are shown in the phylogenetic tree in supplemental Fig. 1*B*. The conservation of the seven selected residues had less of a bimodal distribution (Fig. 3*B*) than did the β-galactosidase analysis (Fig. 3*A*). This result was not unexpected as the seven amino acids selected are common in the interior of most proteins. However, when the degree of conservation was examined in the context of the phylogenetic tree branching, there was a clear dichotomy. One branch of the tree containing 90 sequences had 87 sequences with 4–7 of the pocket residues conserved (branch B, supplemental Fig. 1*B*, aLRT branch support 0.93), and many of the non-matching amino acids of these 87 sequences were homologous to the *E. coli* residues. In the remainder of the tree (159 sequences), all but one repressor, that of *Deinococcus maricopensis*, had 3 or less residues conserved, with no clear pattern of sequence conservation or homologous residues discernible. The *E. coli* ribose repressor was located in this part of the tree and had only two of the seven residues conserved. This pattern of conservation was consistent with the structural and mutagenesis data and indicated that the seven amino acids were suited for identifying putative *lac* repressors.

Genome Analysis—Of 1087 genomes analyzed, 197 contained β-galactosidase sequences of which 53 were identified as possessing an allolactose synthesis motif (supplemental Fig. 2*A*). These 53 were in one branch of the phylogenetic tree (branch C, supplemental Fig. 2*A*, aLRT branch support 1.00), and a bimodal distribution was again found in the number of conserved motif residues. The 53 sequences had 7 or more of the 14 residues conserved; the other 144 proteins had 3 or less conserved.

From the same set of 1087 genomes, 306 were identified[4] that contained repressor sequences (supplemental Fig. 2*B*). A clear phylogenetic clustering of proteins containing the *lac* repressor residues was observed (branch G, supplemental Fig. 2*B*, aLRT branch support 1.00). This branch consisted of 33 repressors of which 31 had 4 or more of the 7 selected residues conserved. The remaining 2 sequences had only 3 residues strictly conserved, but homologous substitutions suggested these two proteins are probably also *lac* repressors. Two additional sequences with four conserved pocket residues were dispersed in the remainder of the tree (supplemental Fig. 2*A*) and were considered false positives. The *lac* repressor was only found among Gammaproteobacteria and, with the exception of two bacteria from the order Vibrionales, was confined to members of the order Enterobacteriales. Interestingly, six other Vibrionales species included in this study did not possess a *lac* repressor.

Next, the correlation between allolactose synthesis and the *lac* repressor in the microbial genomes was examined. Fig. 4 shows the branch of the β-galactosidase phylogenetic tree with the 53 species with the allolactose synthesis motif (branch C of supplemental Fig. 2*A*). Of these 53 species, 32 possessed a *lac* repressor and are shown in *blue*, whereas those without *lac* repressor are colored *black*. Fig. 5 shows the branch of the repressor phylogenetic tree with the 33 species identified as containing *lac* repressors (branch *G* of supplemental Fig. 2*B*). The 32 species that also contain β-galactosidases with an allolactose synthesis motif are shown in *purple*. The one species that does not contain a β-galactosidase is colored *black* (not shown in Fig. 4). This exception is *Sodalis glossinidius*, where a *lac* repressor was found but no β-galactosidase either with or without the allolactose synthesis motif could be identified. This bacteria is an endosymbiont whose genome has undergone massive erosion (38), so the missing gene was not significant. In summary, except for the endosymbiont, all species that have a *lac* repressor also have a β-galactosidase with an allolactose synthesis motif. However, some species have β-galactosidases with an allolactose synthesis motif but do not possess a *lac* repressor.

To further examine the relationship between the allolactose synthesis motif and *lac* repressors, the genomic arrangement of the corresponding genes was examined. In *E. coli* these genes are adjacent on the chromosome. However, they are regulated separately, and thus this arrangement is not required. Still, if a

[4] 307 sequences were analyzed, as two sequences, the *lac* and ribose repressors, were included from the *E. coli* genome.

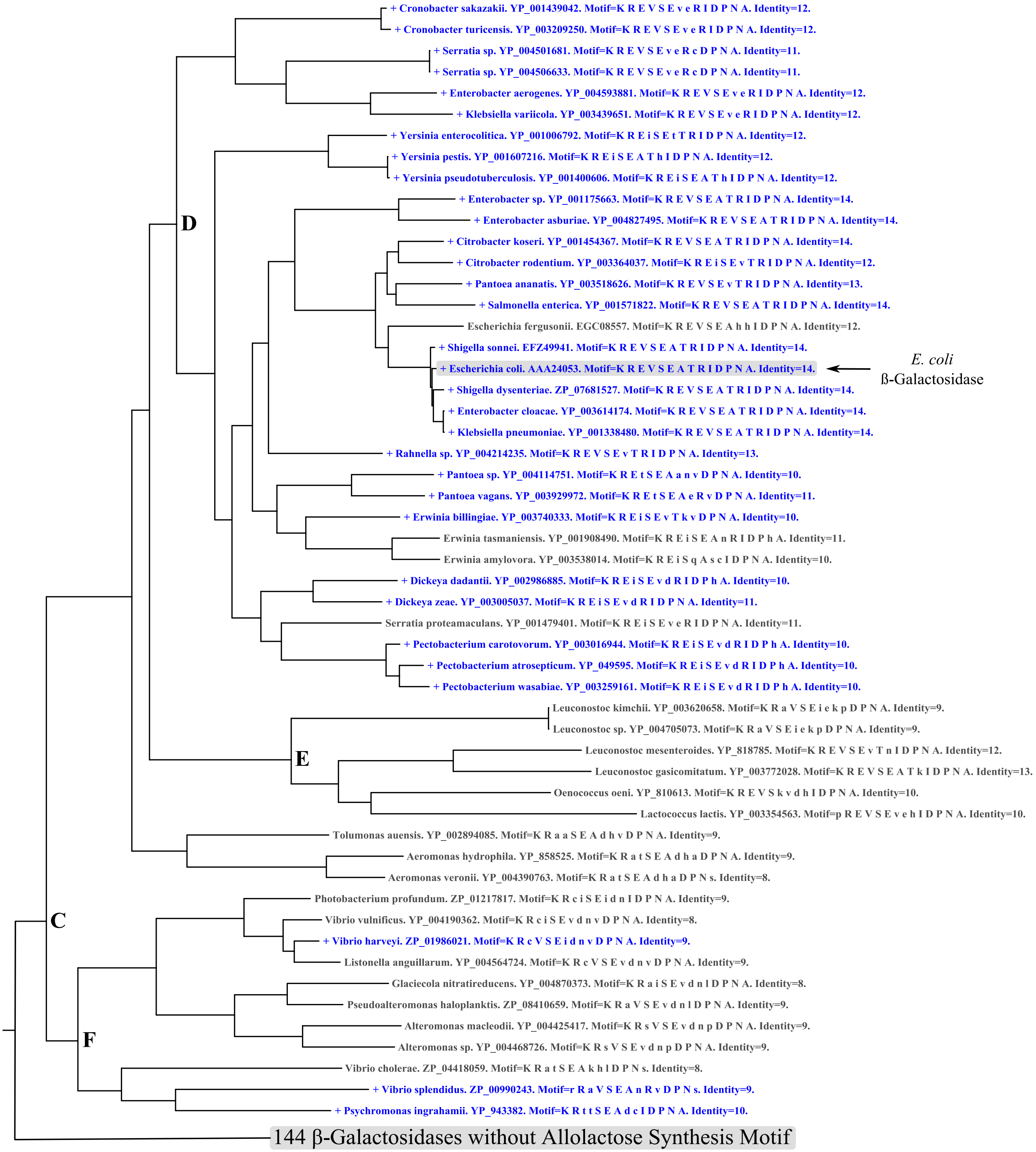
+ Cronobacter sakazakii. YP_001439042. Motif=K R E V S E v e R I D P N A. Identity=12.
+ Cronobacter turicensis. YP_003209250. Motif=K R E V S E v e R I D P N A. Identity=12.
+ Serratia sp. YP_004501681. Motif=K R E V S E v e R c D P N A. Identity=11.
+ Serratia sp. YP_004506633. Motif=K R E V S E v e R c D P N A. Identity=11.
+ Enterobacter aerogenes. YP_004593881. Motif=K R E V S E v e R I D P N A. Identity=12.
+ Klebsiella variicola. YP_003439651. Motif=K R E V S E v e R I D P N A. Identity=12.
+ Yersinia enterocolitica. YP_001006792. Motif=K R E i S E t T R I D P N A. Identity=12.
+ Yersinia pestis. YP_001607216. Motif=K R E i S E A T h I D P N A. Identity=12.
+ Yersinia pseudotuberculosis. YP_001400606. Motif=K R E i S E A T h I D P N A. Identity=12.
D
+ Enterobacter sp. YP_001175663. Motif=K R E V S E A T R I D P N A. Identity=14.
+ Enterobacter asburiae. YP_004827495. Motif=K R E V S E A T R I D P N A. Identity=14.
+ Citrobacter koseri. YP_001454367. Motif=K R E V S E A T R I D P N A. Identity=14.
+ Citrobacter rodentium. YP_003364037. Motif=K R E i S E v T R I D P N A. Identity=12.
+ Pantoea ananatis. YP_003518626. Motif=K R E V S E v T R I D P N A. Identity=13.
+ Salmonella enterica. YP_001571822. Motif=K R E V S E A T R I D P N A. Identity=14.
Escherichia fergusonii. EGC08557. Motif=K R E V S E A h h I D P N A. Identity=12.
+ Shigella sonnei. EFZ49941. Motif=K R E V S E A T R I D P N A. Identity=14.
+ Escherichia coli. AAA24053. Motif=K R E V S E A T R I D P N A. Identity=14.
E. coli
ß-Galactosidase
+ Shigella dysenteriae. ZP_07681527. Motif=K R E V S E A T R I D P N A. Identity=14.
+ Enterobacter cloacae. YP_003614174. Motif=K R E V S E A T R I D P N A. Identity=14.
+ Klebsiella pneumoniae. YP_001338480. Motif=K R E V S E A T R I D P N A. Identity=14.
+ Rahnella sp. YP_004214235. Motif=K R E V S E v T R I D P N A. Identity=13.
+ Pantoea sp. YP_004114751. Motif=K R E t S E A a n v D P N A. Identity=10.
+ Pantoea vagans. YP_003929972. Motif=K R E t S E A e R v D P N A. Identity=11.
+ Erwinia billingiae. YP_003740333. Motif=K R E i S E v T k v D P N A. Identity=10.
Erwinia tasmaniensis. YP_001908490. Motif=K R E i S E A n R I D P h A. Identity=11.
Erwinia amylovora. YP_003538014. Motif=K R E i S q A s c I D P N A. Identity=10.
+ Dickeya dadantii. YP_002986885. Motif=K R E i S E v d R I D P h A. Identity=10.
+ Dickeya zeae. YP_003005037. Motif=K R E i S E v d R I D P N A. Identity=11.
Serratia proteamaculans. YP_001479401. Motif=K R E i S E v e R I D P N A. Identity=11.
+ Pectobacterium carotovorum. YP_003016944. Motif=K R E i S E v d R I D P h A. Identity=10.
+ Pectobacterium atrosepticum. YP_049595. Motif=K R E i S E v d R I D P h A. Identity=10.
+ Pectobacterium wasabiae. YP_003259161. Motif=K R E i S E v d R I D P h A. Identity=10.
Leuconostoc kimchii. YP_003620658. Motif=K R a V S E i e k p D P N A. Identity=9.
Leuconostoc sp. YP_004705073. Motif=K R a V S E i e k p D P N A. Identity=9.
E
Leuconostoc mesenteroides. YP_818785. Motif=K R E V S E v T n I D P N A. Identity=12.
Leuconostoc gasicomitatum. YP_003772028. Motif=K R E V S E A T k I D P N A. Identity=13.
Oenococcus oeni. YP_810613. Motif=K R E V S k v d h I D P N A. Identity=10.
Lactococcus lactis. YP_003354563. Motif=p R E V S E v e h I D P N A. Identity=10.
Tolumonas auensis. YP_002894085. Motif=K R a a S E A d h v D P N A. Identity=9.
Aeromonas hydrophila. YP_858525. Motif=K R a t S E A d h a D P N A. Identity=9.
Aeromonas veronii. YP_004390763. Motif=K R a t S E A d h a D P N s. Identity=8.
Photobacterium profundum. ZP_01217817. Motif=K R c i S E i d n I D P N A. Identity=9.
C
Vibrio vulnificus. YP_004190362. Motif=K R c i S E v d n v D P N A. Identity=8.
+ Vibrio harveyi. ZP_01986021. Motif=K R c V S E i d n v D P N A. Identity=9.
Listonella anguillarum. YP_004564724. Motif=K R c V S E v d n v D P N A. Identity=9.
Glaciecola nitratireducens. YP_004870373. Motif=K R a i S E v d n l D P N A. Identity=8.
Pseudoalteromonas haloplanktis. ZP_08410659. Motif=K R a V S E v d n l D P N A. Identity=9.
Alteromonas macleodii. YP_004425417. Motif=K R s V S E v d n p D P N A. Identity=9.
F
Alteromonas sp. YP_004468726. Motif=K R s V S E v d n p D P N A. Identity=9.
Vibrio cholerae. ZP_04418059. Motif=K R a t S E A k h l D P N s. Identity=8.
+ Vibrio splendidus. ZP_00990243. Motif=r R a V S E A n R v D P N s. Identity=9.
+ Psychromonas ingrahamii. YP_943382. Motif=K R t t S E A d c I D P N A. Identity=10.
144 β-Galactosidases without Allolactose Synthesis Motif

FIGURE 4. **The correlation between the β-galactosidase allolactose synthesis motif and *lac* repressors in microbial genomes.** The figure shows part of a larger tree of β-galactosidase sequences, restricted to the one branch of the tree where the allolactose synthesis motif was identified. A node label colored *blue* and preceded by a + symbol indicates a *lac* repressor was identified in the same genome. The node labels indicate genus, species, and β-galactosidase sequence accession number. *Motif* lists the residues that align with *E. coli* β-galactosidase allolactose synthesis motif in the order 517, 599, 808, 795, 796, 797, 798, 799, 800, 801, 802, 803, 804, and 805. Those residues that are conserved relative to the *E. coli* protein are capitalized. *Identity* indicates the total number of the 14 motif residues that are strictly conserved with *E. coli* β-galactosidase. The complete version of this tree is shown in supplemental Fig. 2*A*, but note in this figure the node labels are colored by different criteria.

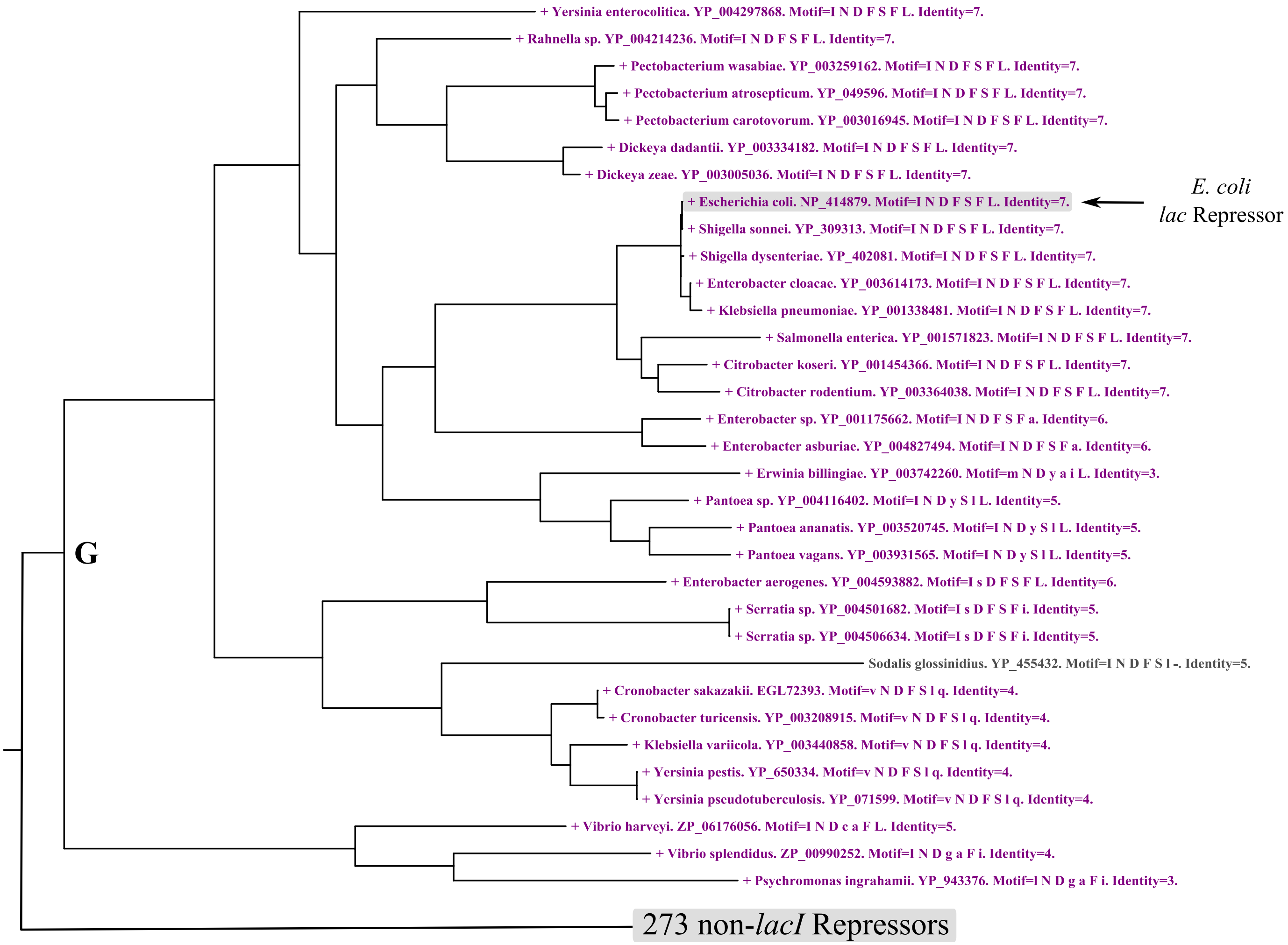
+ Yersinia enterocolitica. YP_004297868. Motif=I N D F S F L. Identity=7.
+ Rahnella sp. YP_004214236. Motif=I N D F S F L. Identity=7.
+ Pectobacterium wasabiae. YP_003259162. Motif=I N D F S F L. Identity=7.
+ Pectobacterium atrosepticum. YP_049596. Motif=I N D F S F L. Identity=7.
+ Pectobacterium carotovorum. YP_003016945. Motif=I N D F S F L. Identity=7.
+ Dickeya dadantii. YP_003334182. Motif=I N D F S F L. Identity=7.
+ Dickeya zeae. YP_003005036. Motif=I N D F S F L. Identity=7.
+ Escherichia coli. NP_414879. Motif=I N D F S F L. Identity=7.
E. coli
lac Repressor
+ Shigella sonnei. YP_309313. Motif=I N D F S F L. Identity=7.
+ Shigella dysenteriae. YP_402081. Motif=I N D F S F L. Identity=7.
+ Enterobacter cloacae. YP_003614173. Motif=I N D F S F L. Identity=7.
+ Klebsiella pneumoniae. YP_001338481. Motif=I N D F S F L. Identity=7.
+ Salmonella enterica. YP_001571823. Motif=I N D F S F L. Identity=7.
+ Citrobacter koseri. YP_001454366. Motif=I N D F S F L. Identity=7.
+ Citrobacter rodentium. YP_003364038. Motif=I N D F S F L. Identity=7.
+ Enterobacter sp. YP_001175662. Motif=I N D F S F a. Identity=6.
+ Enterobacter asburiae. YP_004827494. Motif=I N D F S F a. Identity=6.
+ Erwinia billingiae. YP_003742260. Motif=m N D y a i L. Identity=3.
+ Pantoea sp. YP_004116402. Motif=I N D y S l L. Identity=5.
+ Pantoea ananatis. YP_003520745. Motif=I N D y S l L. Identity=5.
+ Pantoea vagans. YP_003931565. Motif=I N D y S l L. Identity=5.
G
+ Enterobacter aerogenes. YP_004593882. Motif=I s D F S F L. Identity=6.
+ Serratia sp. YP_004501682. Motif=I s D F S F i. Identity=5.
+ Serratia sp. YP_004506634. Motif=I s D F S F i. Identity=5.
Sodalis glossinidius. YP_455432. Motif=I N D F S l -. Identity=5.
+ Cronobacter sakazakii. EGL72393. Motif=v N D F S l q. Identity=4.
+ Cronobacter turicensis. YP_003208915. Motif=v N D F S l q. Identity=4.
+ Klebsiella variicola. YP_003440858. Motif=v N D F S l q. Identity=4.
+ Yersinia pestis. YP_650334. Motif=v N D F S l q. Identity=4.
+ Yersinia pseudotuberculosis. YP_071599. Motif=v N D F S l q. Identity=4.
+ Vibrio harveyi. ZP_06176056. Motif=I N D c a F L. Identity=5.
+ Vibrio splendidus. ZP_00990252. Motif=I N D g a F i. Identity=4.
+ Psychromonas ingrahamii. YP_943376. Motif=I N D g a F i. Identity=3.
273 non-*lacI* Repressors

FIGURE 5. **The correlation between the *lac* repressor and the β-galactosidase allolactose synthesis motif in microbial genomes.** The figure shows part of a larger tree of repressor sequences, restricted to the one branch of the tree where *lac* repressors were identified. A node label colored *purple* and preceded by a + symbol indicates that a β-galactosidase containing an allolactose synthesis motif was identified in the same genome. The node labels indicate genus, species, and repressor sequence accession number. *Motif* lists the residues that align with *E. coli lac* repressor residues in the order 79, 125, 149, 161, 193, 293, and 296. Those residues that are conserved with the *E. coli* protein are *capitalized*. *Identity* indicates the total number of the seven residues that are conserved with the *E. coli lac* repressor. The complete version of this tree shown in supplemental Fig. 2*B*, but note in this figure the node labels were colored by different criteria.

putative *lac* repressor gene is located close to the β-galactosidase gene as in an operon structure, its identity as a *lac* repressor is more likely. Among the 32 identified *lac* repressors, the majority (23) were in proximity with the matched β-galactosidase gene, either adjacent or within the same operon (as identified by the sequence annotation). Of the remaining 134 β-galactosidases without an allolactose synthesis motif, the genomic organization of a 10% random sampling was examined. None of these repressors was adjacent to the β-galactosidase gene, within the same operon, or within ±10 genes on the chromosome, consistent with those repressors not functioning in regulating β-galactosidase expression.

DISCUSSION

LacZ β-galactosidase is a bifunctional enzyme catalyzing the hydrolysis of lactose to Gal and Glc as well as intramolecular isomerization of lactose to allolactose. Allolactose induces the *lac* operon (2, 3), and its synthesis is important for understanding operon regulation. The rate of intramolecular allolactose synthesis can only be significant if the Glc cleaved from the Gal of lactose (Fig. 1*B*, *E-Gal·Glc*) remains bound long enough for the isomerization reaction to occur (4).

This work begins by mapping the Glc acceptor site of β-galactosidase. We took a study by Juers *et al.* (8) as the starting point. In that investigation weak electron density suggestive of a Glc was seen when concentrated Glc (500 mM) was added to crystals of β-galactosidase that had 2-deoxy-Gal covalently bound to Glu-537. Although the signal for the Glc was too poor to define binding interactions, the structure showed that the active site loop was in the closed conformation and that part of the weak Glc electron density was near a loop residue, Ser-796. We reasoned that Ser-796 may be involved in binding the Glc and that loop closure was important to correctly bind this residue. Thus we studied G794A-β-galactosidase, a variant that favors the closed loop conformation (13). Note that this substitution affects the hydrolytic reaction by reducing the k_3 (13), a change also expected to favor trapping intermediates.

When a crystal of G794A-β-galactosidase was incubated with lactose, an allolactose was in the active site. Presumably the lactose was converted to allolactose, and the G794A substitution prevented its release by hindering loop opening. A previous kinetic study had implied that allolactose was formed much more slowly with G794A-β-galactosidase than with native enzyme, and it was suggested that the rate of transgalactosylation (k_4; see Fig. 1, *B* and *C*) was significantly reduced (13). However, the results here indicate that allolactose production is slowed due to strong binding interactions that stabilize the enzyme-product complex and thereby impede catalysis.

The density defining the Glc moiety of the trapped allolactose was in a similar location as the weak signal for Glc found by Juers *et al.* (8). The signal for the Glc of allolactose is, however, much stronger and allows for identification of the enzyme-Glc interactions. We argue here that the location of the Glc moiety of the trapped allolactose defines the position of Glc when it binds as an acceptor. Several lines of evidence support this. First, as already stated, the density is in a similar position as the weak density for Glc with the native enzyme (8). Second, structures with Bis-Tris and with L-ribose in the putative Glc site identify precisely the same binding residues. Third, kinetic data show that replacement of the residues in the postulated binding site increased the value of the Glc dissociation constant (K_i'') by an order of magnitude or more (Table 2). Fourth, the geometry of the postulated Glc binding, *vis à vis* the C6 hydroxyl and Glu-461, correctly positions Glu-461 to act as a base catalyst for allolactose formation. Finally, authentication is provided by the co-conservation of residues important for binding Glc and the *lac* repressor in microbial genomes.

Glc Binding Interactions—Six amino acids interact with Glc (Fig. 2*A*). Five (Ser-796, Glu-797, Asn-102, His-418, and Lys-517) are in position to form H-bonds. The H-bonds with His-418 and Lys-517 are indirect (via waters), but the distances are short (~2.7 Å), and the geometry is ideal for strong bonds. The sixth residue, Trp-999, is an important platform of the acceptor site because its π cloud forms C-H···π (30) interactions with the hydrogens of the C3, C4, and C5 atoms of Glc.

Previous studies with various acceptor monosaccharides showed that sugars with hydroxyls in a different orientation at the C3 and/or C4 positions from those of Glc bound poorly (17). His-418 and Lys-517 interact with the O3 and O4 hydroxyls of Glc (via waters), respectively, and appear to be responsible for this specificity.

It is interesting that there are strong interactions with the β-anomeric hydroxyl of allolactose (with Asn-102, Ser-796, and Glu-797) and no electron density to suggest occupancy of the hydroxyl in the α-configuration. In addition, the location of Ser-796 would sterically hinder binding of the α-isomer. It is also of interest that only the β-anomer of allolactose binds to the *lac* repressor (39). The reason that β-allolactose is favored by both proteins has no conclusive explanation. When β-allolactose is released from the enzyme, it would undoubtedly mutarotate and not necessarily be in the β-form when it reaches the repressor. Possibly there are transitive interactions between β-galactosidase and the *lac* repressor so that direct transfer of β-allolactose from β-galactosidase without migration to bulk water takes place. A protein-protein interaction study of *E. coli* K12 proteins did show that if β-galactosidase is the "bait," *lac* repressor is captured as "prey" (40).

Conservation of the Allolactose Synthesis Motif—The structural findings led us to propose that 14 residues of β-galactosidase make up an allolactose synthesis motif. Of 1087 microbial genomes, 197 were identified (supplemental Fig. 2*A*) as containing β-galactosidases homologous to the *E. coli* enzyme. However, the allolactose synthesis motif was only conserved in 53 of these and is confined to one branch of the phylogenic tree (branch C, supplemental Fig. 2*A* and Fig. 4). Thus, intramolecular allolactose production does not seem to be a universal feature of β-galactosidases. In addition to the bioinformatics analysis, this non-universality was also suggested by crystal structures of β-galactosidases from *Arthrobacter* sp. (41) and *Kluyveromyces lactis* (42). Structural alignments reveal that the allolactose synthesis motif residues are not conserved in the two enzymes and neither enzyme has a structure homologous to the active site loop of the *E. coli* enzyme. In *E. coli* β-galactosidase, the loop is located between two α-helices. Both the *Arthrobacter* sp. and *K . lactis* enzymes have homologous α-helices. However, in the *Arthrobacter* sp. β-galactosidase, the connection between these α-helices is much longer than the *E. coli* loop and extends away from the active site, making its involvement in catalysis improbable. With the β-galactosidase from *K. lactis*, the α-helices are connected by only a single amino acid (Pro-798), not a loop. Functional studies showed that the *K. lactis* enzyme only synthesizes allolactose at high (*i.e.* 1 M) lactose (43) and that this activity depends on the Glc concentration (44). This reaction is thus intermolecular, not intramolecular, consistent with the absence of a Glc binding site. Two other β-galactosidases (*ebg* β-galactosidase (45) and *Thermoanaerobacterium thermosulfurigenes* EM1 β-galactosidase (46, 47)) are known not to synthesize allolactose by intramolecular isomerization. Structures of these enzymes have not been reported, but sequence analysis shows that the allolactose synthesis motif is missing in both these two β-galactosidases.

Phylogenetic Analysis of the lac Repressor—Repressors homologous to the *lac* repressor (*E. coli*) are more widely distributed than β-galactosidases homologous to *lacZ* β-galactosidase (306 rather than 197 of the 1087 genomes). However, only 33 of these 306 were identifiable as *lac* repressors (supplemental Fig. 2*B* and Fig. 5) by the criteria established in this study. Significantly, with the exception of one endosymbiont that has undergone massive genome erosion, *lac* repressors were only identified in microbes with the allolactose synthesis motif. Also, *lac* repressors were not found in the *K. lactis* or *Arthrobacter* sp.[5] genomes. The finding that *lac* repressors could not be identified in genomes unless residues important for allolactose synthesis are also present strongly argues that the allolactose synthesizing residues have been correctly identified and suggests co-selection of allolactose synthesis and *lac* repressors.

Intriguingly, there are β-galactosidases that synthesize allolactose in organisms without a *lac* repressor. There are several possible explanations. First, gene loss may have occurred.

[5] As only the genus and not the species was known for the *Arthrobacter* sp. enzyme, the entire *Arthrobacter* genus was searched for *lac* repressors.

For example, *Escherichia fergusonii* does not have a *lac* repressor even though this organism is located in the midst of a group of Enterobacteriales that have *lac* repressors. Allolactose synthesis may have been retained in this β-galactosidase as a vestigial feature. Second, there may have been horizontal transfer of the *lacZ* gene. Six Firmicutes (Fig. 4, *branch E*) do not have a *lac* repressor or indeed any *LacI-GalR* family repressor identifiable by the criteria of this study. In the phylogenetic tree, these Firmicutes form a sub-branch among a larger group of Gammaproteobacteria, indicative of horizontal gene transfer. Additionally, in a branch of mainly marine Gammaproteobacteria (Vibrionales and Alteromonadales, *branch F*, Fig. 4) 11 species contain β-galactosidases that have the machinery to synthesize allolactose, whereas a *lac* repressor was only in 3 of these. Horizontal gene transfer is again suggested as the β-galactosidase phylogenetic trees did not cluster bacteria from the *Vibrio* genus together but segregated them into two separate sub-branches. This analysis is also consistent with a previous study showing that the *lac* operon has a complicated evolutionary history even among closely related Gammaproteobacteria (48). The data also suggest that the metabolic activity of the *lac* operon (*i.e.* the β-galactosidase enzyme) is more conserved than the regulatory mechanism (*i.e.* the *lac* repressor).

Role of the Active Site Loop—The interactions of the two loop residues (Ser-796 and Glu-797) with Glc can only occur when the loop is closed. This indicates that the loop as well as the residues involved in its opening and closing are required for intramolecular allolactose synthesis. The studies showing that the loop and residues important for its opening and closing are co-conserved with the Glc binding residues strongly supports this interpretation as does the finding that the *lac* repressor is only present in organisms where loop and related residues are conserved.

It was previously suggested (10–13, 39, 49) that the role of the active site loop is to stabilize the transition state. The reason for that supposition was that transition state analog binding is significantly improved with substituted enzymes that favor the closed loop conformation. However, in retrospect, even though transition state analogs bind better, the binding of substrates decreases by similar factors. So, although an enzyme that favors the closed loop conformation stabilizes the transition state, the same loop conformation destabilizes substrate binding (13). The overall catalytic efficiencies (k_{cat}/K_m) of these substituted enzymes thus do not change (10, 12, 13), and no catalytic advantage is gained. A better explanation for the role of the loop is that loop closure facilitates Glc binding and that transition state formation is the trigger that induces loop closure.

Thus, it appears that loop closure, which helps form the Glc binding site, is a key component of transgalactosylation. Vice versa, it follows that opening of the loop promotes the release of Glc from the active site and in turn promotes the hydrolytic reaction. This argument is consistent with crystal structures of β-galactosidase complexed with transition state analogs. Those structures show some electron density for the loop in both the open and closed conformation (8), a finding suggestive of transient loop closure. Thus, the bifunctional nature of β-galactosidase depends on the presence of a Glc binding site when the loop is closed and the loss of the site when the loop opens.

Allolactose Regulation—It is not readily apparent why the *lac* operon is regulated indirectly by allolactose and not directly by lactose. Indeed, it has been suggested that glycerolgalactoside, a β-galactoside found in plants, is the true *lac* operon substrate (50, 51). Besides being hydrolyzed by β-galactosidase, glycerolgalactoside is a good inducer and binds to the *lac* repressor without processing. This theory has been supported by the misconception that allolactose production is a fortuitous or occasional side reaction of β-galactosidase and not a specific bifunctional reaction. By identifying the Glc binding site and the role of the loop in allolactose formation as well as the co-conservation with *lac* repressor, this work demonstrates that allolactose production is not due to chance. The complex adaptations needed to produce allolactose indicate that the bifunctionality of β-galactosidase results from selection though evolution and suggests that it confers a biological advantage. It also follows that lactose is indeed the natural substrate of the *lac* operon.

REFERENCES

1. Jacob, F., and Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3,** 318–356
2. Jobe, A., and Bourgeois, S. (1972) lac Repressor-operator interaction. VI. The natural inducer of the lac operon. *J. Mol. Biol.* **69,** 397–408
3. Burstein, C., Cohn, M., Kepes, A., and Monod, J. (1965) Role of lactose and its metabolic products in the induction of the lactose operon in *Escherichia coli*. *Biochim. Biophys. Acta* **95,** 634–639
4. Huber, R. E., Kurz, G., and Wallenfels, K. (1976) A quantitation of the factors which affect the hydrolase and transgalactosylase activities of β-galactosidase (*E. coli*) on lactose. *Biochemistry* **15,** 1994–2001
5. Huber, R. E., Wallenfels, K., and Kurz, G. (1975) The action of β-galactosidase (*Escherichia coli*) on allolactose. *Can. J. Biochem.* **53,** 1035–1038
6. Fowler, A. V., and Zabin, I. (1978) Amino acid sequence of β-galactosidase. XI. Peptide ordering procedures and the complete sequence. *J. Biol. Chem.* **253,** 5521–5525
7. Jacobson, R. H., Zhang, X. J., DuBose, R. F., and Matthews, B. W. (1994) Three-dimensional structure of β-galactosidase from *E. coli*. *Nature* **369,** 761–766
8. Juers, D. H., Heightman, T. D., Vasella, A., McCarter, J. D., Mackenzie, L., Withers, S. G., and Matthews, B. W. (2001) A structural view of the action of *Escherichia coli* (lacZ) β-galactosidase. *Biochemistry* **40,** 14781–14794
9. Juers, D. H., Jacobson, R. H., Wigley, D., Zhang, X. J., Huber, R. E., Tronrud, D. E., and Matthews, B. W. (2000) High resolution refinement of β-galactosidase in a new crystal form reveals multiple metal-binding sites and provides a structural basis for α-complementation. *Protein Sci.* **9,** 1685–1699
10. Dugdale, M. L., Vance, M. L., Wheatley, R. W., Driedger, M. R., Nibber, A., Tran, A., and Huber, R. E. (2010) Importance of Arg-599 of β-galactosidase (*Escherichia coli*) as an anchor for the open conformations of Phe-601 and the active-site loop. *Biochem. Cell Biol.* **88,** 969–979
11. Dugdale, M. L., Dymianiw, D. L., Minhas, B. K., D'Angelo, I., and Huber, R. E. (2010) Role of Met-542 as a guide for the conformational changes of Phe-601 that occur during the reaction of β-galactosidase (*Escherichia coli*). *Biochem. Cell Biol.* **88,** 861–869
12. Jancewicz, L. J., Wheatley, R. W., Sutendra, G., Lee, M., Fraser, M. E., and Huber, R. E. (2012) Ser-796 of β-galactosidase (*Escherichia coli*) plays a key role in maintaining a balance between the opened and closed conformations of the catalytically important active site loop. *Arch. Biochem. Biophys.* **517,** 111–122
13. Juers, D. H., Hakda, S., Matthews, B. W., and Huber, R. E. (2003) Structural basis for the altered activity of Gly-794 variants of *Escherichia coli* β-galactosidase. *Biochemistry* **42,** 13505–13511
14. Gebler, J. C., Aebersold, R., and Withers, S. G. (1992) Glu-537, not Glu-461, is the nucleophile in the active site of (lac Z) β-galactosidase from *Escherichia coli*. *J. Biol. Chem.* **267,** 11126–11130
15. Richard, J. P., Huber, R. E., Lin, S., Heo, C., and Amyes, T. L. (1996) Struc-

ture-reactivity relationships for β-galactosidase (*Escherichia coli*, lac Z). 3. Evidence that Glu-461 participates in Brønsted acid-base catalysis of β-D-galactopyranosyl group transfer. *Biochemistry* **35,** 12377–12386

16. Mahoney, R. R. (1998) Galactosyl-oligoscchride formation during lactose hydrolysis. A review. *Food Chemistry* **63,** 147–154
17. Huber, R. E., Gaunt, M. T., and Hurlburt, K. L. (1984) Binding and reactivity at the "glucose" site of galactosyl-β-galactosidase (*Escherichia coli*). *Arch. Biochem. Biophys.* **234,** 151–160
18. Evans, P. R. (1993) *Proceedings of the CCP4 Study Weekend on Data Collection and Processing.* 29–30 January 1993 (Sawyer, L., Isaacs, N., Bailey, S., eds.) pp. 114–122, Warrington, Daresbury Laboratory, Great Britain
19. Kabsch, W. (1988) Evaluation of single-crystal X-ray diffraction data from a position-sensitive detector. *J. Appl. Crystallogr.* **21,** 916–924
20. Leslie, A. G. W. (1991) *Crystallographic Computing 5, from Chemistry to Biology*, pp. 50–61, Oxford University Press, Oxford
21. Emsley, P., and Cowtan, K. (2004) Coot. Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2126–2132
22. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **50,** 240–255
23. Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX. A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 213–221
24. Edwards, R. A., and Huber, R. E. (1992) Surface denaturation of proteins. The thermal inactivation of β-galactosidase (*Escherichia coli*) on wall-liquid surfaces. *Biochem. Cell Biol.* **70,** 63–69
25. Deschavanne, P. J., Viratelle, O. M., and Yon, J. M. (1978) Conformational adaptability of the active site of β-galactosidase. Interaction of the enzyme with some substrate analogous effectors. *J. Biol. Chem.* **253,** 833–837
26. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST. A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402
27. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948
28. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies. Assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321
29. Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002) The Bioperl toolkit. Perl modules for the life sciences. *Genome Res.* **12,** 1611–1618
30. Kumari, M., Balaji, P. V., and Sunoj, R. B. (2011) Quantification of binding affinities of essential sugars with a tryptophan analogue and the ubiquitous role of C-H···π interactions. *Phys. Chem. Chem. Phys.* **13,** 6517–6530
31. Wheatley, R. W., Kappelhoff, J. C., Hahn, J. N., Dugdale, M. L., Dutkoski, M. J., Tamman, S. D., Fraser, M. E., and Huber, R. E. (2012) Substitution for Asn-460 cripples β-galactosidase (*Escherichia coli*) by increasing substrate affinity and decreasing transition state stability. *Arch. Biochem. Biophys.* **521,** 51–61
32. Juers, D. H., Matthews, B. W., and Huber, R. E. (2012) LacZ β-galactosidase. Structure and function of an enzyme of historical and molecular biological importance. *Protein Sci.* **21,** 1792–1807
33. Huber, R. E., Hakda, S., Cheng, C., Cupples, C. G., and Edwards, R. A. (2003) Trp-999 of β-galactosidase (*Escherichia coli*) is a key residue for binding, catalysis, and synthesis of allolactose, the natural lac operon inducer. *Biochemistry* **42,** 1796–1803
34. Juers, D. H., Rob, B., Dugdale, M. L., Rahimzadeh, N., Giang, C., Lee, M., Matthews, B. W., and Huber, R. E. (2009) Direct and indirect roles of His-418 in metal binding and in the activity of β-galactosidase (*E. coli*). *Protein Sci.* **18,** 1281–1292
35. Lo, S., Dugdale, M. L., Jeerh, N., Ku, T., Roth, N. J., and Huber, R. E. (2010) Studies of Glu-416 variants of β-galactosidase (*E. coli*) show that the active site Mg^{2+} is not important for structure and indicate that the main role of Mg^{2+} is to mediate optimization of active site chemistry. *Protein J.* **29,** 26–31
36. Daber, R., Stayrook, S., Rosenberg, A., and Lewis, M. (2007) Structural analysis of lac repressor bound to allosteric effectors. *J. Mol. Biol.* **370,** 609–619
37. Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S., and Miller, J. H. (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues as well as "spacers," which do not require a specific sequence. *J. Mol. Biol.* **240,** 421–433
38. Toh, H. (2006) Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res.* **16,** 149–156
39. Turner, C. L., and Huber, R. E. (1977) Differential binding of allolactose anomers to the lactose repressor of *Escherichia coli*. *J. Mol. Biol.* **115,** 195–199
40. Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., Ara, T., Nakahigashi, K., Huang, H.-C., Hirai, A., Tsuzuki, K., Nakamura, S., Altaf-Ul-Amin, M., Oshima, T., Baba, T., Yamamoto, N., Kawamura, T., Ioka-Nakamichi, T., Kitagawa, M., Tomita, M., Kanaya, S., Wada, C., and Mori, H. (2006) Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. *Genome Res.* **16,** 686–691
41. Skálová, T., Dohnálek, J., Spiwok, V., Lipovová, P., Vondrácková, E., Petroková, H., Dusková, J., Strnad, H., Králová, B., and Hasek, J. (2005) Cold-active β-galactosidase from Arthrobacter sp. C2–2 forms compact 660-kDa hexamers. Crystal structure at 1.9 Å resolution. *J. Mol. Biol.* **353,** 282–294
42. Pereira-Rodríguez, A., Fernández-Leiro, R., González-Siso, M. I., Cerdán, M. E., Becerra, M., and Sanz-Aparicio, J. (2012) Structural basis of specificity in tetrameric *Kluyveromyces lactis* β-galactosidase. *J. Struct. Biol.* **177,** 392–401
43. Martínez-Villaluenga, C., Cardelle-Cobas, A., Corzo, N., A., O., and Villamiel, M. (2008) Optimization of conditions for galactooligosaccharide synthesis during lactose hydrolysis by β-galactosidase from *Kluyveromyces lactis* (Lactozym 3000 L HP G). *Food Chemistry* **107,** 258–264
44. Kim, C. S., Ji, E.-S., and Oh, D.-K. (2004) A new kinetic model of recombinant β-galactosidase from *Kluyveromyces lactis* for both hydrolysis and transgalactosylation reactions. *Biochem. Biophys. Res. Commun.* **316,** 738–743
45. Hall, B. G., and Hartl, D. L. (1974) Regulation of newly evolved enzymes. I. Selection of a novel lactase regulated by lactose in *Escherichia coli*. *Genetics* **76,** 391–400
46. Burchhardt, G., and Bahl, H. (1991) Cloning and analysis of the β-galactosidase-encoding gene from *Clostridium thermosulfurogenes* EM1. *Gene* **106,** 13–19
47. Huber, R. E., Roth, N. J., and Bahl, H. (1996) Quaternary structure, Mg^{2+} interactions, and some kinetic properties of the β-galactosidase from *Thermoanaerobacterium thermosulfurigenes* EM1. *J. Protein Chem.* **15,** 621–629
48. Stoebel, D. M. (2005) Lack of evidence for horizontal transfer of the lac operon into *Escherichia coli*. *Mol. Biol. Evol.* **22,** 683–690
49. Jancewicz, L. (2008) *Biological Sciences*, Ph.D. Thesis. The role of Ser-796 in the loop (residues 794–803) of β-galactosidase in *Escherichia coli*. University of Calgary, Calgary, Canada
50. Boos, W. (1982) in *Methods in Enzymology* Vol. 89, pp. 59–64, (Willis, A. W., ed.) Academic Press, New York
51. Egel, R. (1979) The lac-operon for lactose degradation or, rather, for the utilization of galactosylglycerols from galactolipids? *J. Theor. Biol.* **79,** 117–119