Genomes contain not only information for current biological functions, but also information for potential novel functions that may allow the host to adapt to new environments. The field of experimental evolution studies that potential by selecting for novel functions and deducing the means by which the function evolved, but until now it has not attempted to predict the outcomes of such experiments. Here I present a model system that is being developed specifically to examine the issue of what kind of information is most useful in predicting how novel functions will evolve. The system is the evolution of a Lac-PTS transport system and a phospho-β-galactosidase hydrolase system as a novel pathway for metabolism of lactose in Escherichia coli. Two kinds of information, sequence-based phylogenetic inference and biochemical activity, are considered as predictors of which E. coli genes will evolve the required new functions. Both biochemical data and phylogenetic inference predict that the cryptic celABC genes, which currently specify a PTS-β-glucoside transport system, are most likely to evolve into a PTS-lactose transport system. Phylogenetic inference predicts that the bglA gene, which currently specifies a phospho-β-glucosidase, is most likely to evolve into a phospho-β-galactosidase. In contrast, biochemical data predict that the cryptic bglB gene, which also currently specifies a phospho-β-glucosidase, is most likely to evolve into a phospho-β-galactosidase.

Introduction

For over 30 years, microorganisms have been used as model systems to study the evolution of new functions (Mortlock 1984). Those systems have the advantage that very specific selection pressures can be applied to very large populations under well-defined conditions. When the model is Escherichia coli, there are the additional advantages of a well-defined genetic system that has been thoroughly studied at the biochemical and physiological levels.

The usual strategy is to apply selection for utilization of a novel resource, typically a carbon source, that cannot be used by the wild-type organism. Such studies have demonstrated several means by which an organism can evolve the ability to use novel resources: (1) by constitutive expression of a gene whose product was already able to metabolize the novel resource but which was undetectable by that resource (reviewed in Mortlock 1984), (2) by activation of cryptic genes or operons that could not be expressed in the wild-type organism but could be expressed following a mutation or transposition of a mobile element (Hall, Yokoyama, and Calhoun 1983; Hall, Betts, and Kricker 1986; Hall and Betts 1987; Parker and Hall 1988; Hall and Xu 1992), and (3) by base substitutions that alter the catalytic properties of a protein so that it acquires a new or expanded substrate range to include the novel resource (Hall 1976, 1981; Hall et al. 1983; Clarke 1984; Hall, Betts, and Wootton 1989).

An alternative strategy is simply to select for better adaptation to a laboratory environment such as a chemostat or serial transfer in minimal medium. In those cases, the measure of evolution is not a specific new activity, but improved competitive fitness (Lenski and Travisano 1994).

None of those studies has attempted to predict which genes will evolve or what the nature of the evolved products will be. In effect, experimental evolutionists did exactly what evolutionary biologists have always done: explain what has happened by characterizing the outcomes.

Recently, attention has begun to shift in the direction of understanding the evolutionary potential that is inherent in current genomes (Hall 1995, 1999a, 1999b; Hall and Malik 1998). That shift is partially occasioned by the availability of complete sequences of microbial genomes, with the resulting sense that if we know the sequence of a genome we should surely be able to describe its potential for evolving new functions.

In at least one case, we can speak with some assurance about the evolutionary potential of a gene for a particular novel function. The ebgA gene of E. coli K12 can evolve to replace the lacZ gene as a source of β-galactosidase for the metabolism of lactose and other β-galactoside sugars (Hall 1986). That evolution is the result of amino acid substitutions that dramatically improve the catalytic efficiency of Ebg enzyme for the new substrates (Hall 1976, 1978, 1981; Hall, Betts, and Wootton 1989). The potential of ebgA for the improved efficiency is limited to exactly two base substitutions, one near the 5’ end and one near the 3’ end of the gene, that alter active-site amino acids (Hall 1995; Hall and Malik 1998). That potential, however, was not predicted; it was deduced after the fact.

Although the complete sequences of 45 microbial genomes are available in the public databases, and at least that many more are underway, it is not clear that sequence information is the most effective way to predict evolutionary potential. Biochemical information might be equally or more effective at predicting novel activities. Does a trace level of a desired novel activity
predict that the activity can be enhanced by one or several mutations? In at least one case it does: the double-mutant Class IV Ebg β-galactosidase does not permit growth on lactobionic acid, but it does have a low level of activity that is detectable in vitro with purified enzyme (Hall 1981). That activity is enhanced as the result of a third mutation, and the resulting triple mutant strains can grow on lactobionate as a sole carbon and energy source (Hall 1978). The problem is that neither the wild type nor either of the two single-mutation classes of Ebg enzyme has any detectable lactobionase activity in vitro, and biochemical measurement would not have predicted the evolution of that novel activity from the wild-type gene. Finally, might protein structural data assist us in predicting which proteins might evolve new functions and which residues might be most likely to evolve?

In this paper, I report the development of a model system for predicting potential of an organism based on knowledge of its current genome and its current biochemical activities. The intent is to compare the sequence information and limited protein structural information with biochemical information to determine (1) if they consistently make the same predictions, and (2) if they do not, which kind of information most accurately predicts evolutionary outcomes.

Escherichia coli metabolizes lactose via the lacY-encoded permease that transports lactose into the cell against a concentration gradient, and the lacZ-encoded β-galactosidase that hydrolyzes lactose to its component simple sugars, glucose and galactose (fig. 1A). That is not only the lactose metabolism pathway in E. coli, it is by far the commonest pathway in microorganisms. Some Gram-positive organisms metabolize lactose by a different pathway, the Lac-PTS/phospho-β-galactosidase pathway, in which a phosphoenolpyruvate-dependent lactose transport system phosphorylates lactose, and the resulting lactose-6-phosphate is hydrolyzed by a phospho-β-galactosidase (fig. 1B).

Escherichia coli K12 has neither Lac-PTS transport nor phospho-β-galactosidase genes, but there is good reason to believe that E. coli has the potential to evolve those functions. The only Gram-negative bacterial species known to have a Lac-PTS/phospho-β-galactosidase pathway is Klebsiella (Hall 1979; Hall, Imai, and Romano 1982). A strain of Klebsiella lacking both the chromosomal and the plasmid-borne Lac operons is Lac−, but it mutates spontaneously to Lac+. The mutant expresses a phospho-β-galactosidase (Hall 1979) and a Lac-PTS transport system (Hall, Imai, and Romano 1982). The phospho-β-galactosidase is the phospho-β-glucosidase used by Klebsiella to hydrolyze phosphorylated cellobiose, and the transport activity is the result of activation of a cryptic Lac-PTS transport system (Hall, Imai, and Romano 1982). Since E. coli is a close relative of Klebsiella, and E. coli K12 possesses a cryptic operon for metabolism of cellobiose by a cellobiose-PTS transport system and a phospho-β-glucosidase, it seems likely that the E. coli has the potential to evolve a Lac-PTS/phospho-β-galactosidase pathway.

This paper predicts which genes in E. coli K12 have the potential to evolve Lac-PTS transport and phospho-β-galactosidase functions.

Materials and Methods
Phylogenetic Analysis
Alignment

Protein sequences were aligned using the multiple-alignment program ClustalX (Thompson et al. 1997). DNA sequences were aligned by introducing into coding sequences triplet gaps between codons corresponding to gaps in the aligned protein sequences via the program CodonAlign (unpublished program). CodonAlign for the Macintosh, and source code that can be compiled for other platforms, is available free at http://www.rochester.edu/College/BIO/labs/HallLab/CodonAlign.html.
Bayesian Inference of Phylogenies

Phylogenies were constructed by the Bayesian method (Rannala and Yang 1996; Mau and Newton 1997; Mau, Newton, and Larget 1999) as implemented by the program MrBayes (Huelsenbeck 2000). MrBayes is available from http://brahms.biology.rochester.edu/software.html. The Bayesian method seeks the most likely trees given the data (the alignment) and the evolutionary model. The posterior probabilities of the phylogeny, branch lengths, and substitution parameters cannot be calculated directly; however, they can be approximated by a Markov Chain Monte Carlo process by sampling trees from the posterior probability distribution. A variant of MCMC called “Metropolis-coupled Markov chain Monte Carlo” (MCMCMC) runs several chains, some of which are heated. A heated Markov chain has the posterior probability of a tree raised to some power i. Heated Markov chains can more easily cross deep likelihood valleys. The effect of heating is to fill in valleys and lower peaks; hence, a heated Markov chain can better explore the parameter space. Using the MCMCMC algorithm, a swap of the states between two chains is attempted at each step. If the swap is accepted, then the states for the two chains are exchanged. If a swap occurs between a heated chain and the cold chain, the cold chain might cross a large valley that it would normally cross with only a very small probability.

Trees are saved every N generations as specified by the user. At the end of the run, a consensus tree with branch lengths is calculated from the saved trees. As the chains run, the ln likelihoods of the trees converge on a stable value. Only trees saved well after ln likelihood has converged are used to calculate the consensus tree.

One of the advantages of Bayesian inference of phylogeny is that the results are easy to interpret. For example, the sum of the posterior probabilities of all trees will sum to 1. Moreover, the posterior probability of any single clade is simply the sum of the posterior probabilities of all trees that contain that clade. The consensus tree calculated by MrBayes does not include the posterior probabilities of the clades; thus, the entire set of trees was imported into PAUP* (Swofford 2000), and the same trees used by MrBayes to calculate a consensus were used to calculate a 50% majority rule consensus in PAUP* (Swofford 2000). The resulting tree shows the posterior probabilities of the clades.

The consensus trees calculated by MrBayes were imported into PAUP* for the purposes of displaying and printing the tree.

Structural Analysis

Structures of the AscB, BglA, and BglB proteins were modeled by the SWISS-MODEL program (Guex and Peitsch 1997; Peitsch et al. 1999; Schwede et al. 2000) as part of the 3Dc crunch project that modeled all of the proteins in the SwissProt database as of May 1998. For details of the project, see http://www.expasy.ch/swissmod/SM/3Dc crunch.html. For each of the proteins, the only crystal structures for which there was meaningful structural homology were those of the Lactococcus lactis phospho-β-galactosidase (Wiesmann et al. 1995).

The Swiss-Pdb Viewer program (Guex and Peitsch 1997) was used to view the aligned structures.

E. coli Strains

Details of the strain constructions are available from the author upon request.

PEP14: cellR1::IS2 Δ(bgl-pho) rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 dTn10kan::bglA. Constitutive for celABCDF operon expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP24: cellR1::IS2 ΔcelBCDF bglR::IS1 bglG33 rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 dTn10kan::bglA1000. All β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP25: cellR1::IS2 ΔcelBCDF bglR::IS1 bglG33 rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 dTn10kan::bglA1000. Constitutive for bglGFB operon expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP33: cellR1::IS2 ΔcelBCDF (bgl-pho) rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 ArbT+. Constitutive for bglA and ArbT expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP54: ascG1::IS186 + 2 unknown mutations cellR1::IS2 ΔcelBCDF (bgl-pho) rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 dTn10kan::bglA1000. Constitutive for ascGFB operon expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP25: cellR1::IS2 ΔcelBCDF bglR::IS1 bglG33 rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 dTn10kan::bglA. Constitutive for celABCDF operon expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP54: ascG1::IS186 + 2 unknown mutations cellR1::IS2 ΔcelBCDF Δ(bgl-pho) rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 dTn10kan::bglA. Constitutive for ascGFB operon expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP25: cellR1::IS2 ΔcelBCDF bglR::IS1 bglG33 rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 dTn10kan::bglA. Constitutive for celABCDF operon expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP36: cellR1::IS2 F Δ(bgl-pho) rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 ArbT+. Constitutive for celABCDF, bglA, and ArbT expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP37: ascG1::IS186 + 2 unknown mutations cellR1::IS2 ΔcelBCDF Δ(bgl-pho) rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 ArbT+. Constitutive for celABCDF, bglA, and ArbT expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP38: cellR1::IS2 ΔcelBCDF bglR::IS1 bglG33 rpsL ara-14 leuB6 ΔlacZ4680 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR*5322 ArbT+. Con-
stutive for bglGFB, bglA, and ArbT expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP56: ascG1::IS186 + 2 unknown mutations celR1::IS2 Δ(bgl-pho) rpsL ara-14 leuB6 ΔlacZΔ820 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR+:L5322 dTn10kan::bglA1000. Constitutive for ascGFB operon and celABCD operon expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

PEP57: ascG1::IS186 + 2 unknown mutations celR1::IS2 ΔcelBCDF bglR::IS1 bglG33 rpsL ara-14 leuB6 ΔlacZΔ820 lacY his-208 metA160 galK2 dTn10cam::ebgA5100 ebgR+:L5322 dTn10kan::bglA1000. Constitutive for ascGFB operon and bglGFB operon expression; other β-glucoside transport and hydrolysis genes are either silent, disrupted, or deleted.

DHF5-ae: phi80dlacZΔM15 Δ(laczY-argF)U169 endA1 recA1 hsdR17m+) deoR thi-1 phoA supE44 λ- gyrA96 relA1 gal-.

**Growth Media**

Minimal media consisted of 423 mg sodium citrate, 100 mg MgSO₄·7H₂O, 1 g (NH₄)₂SO₄, 540 μg FeCl₃, 1 mg thiamine, 3 g KH₂PO₄, 7 g K₂HPO₄, 100 mg of each required amino acid, and a carbon source. Carbon sources were as indicated in the text. Solid media included 15 g/liter purified agar (Sigma). LB broth consisted of LB broth (Miller 1972) plus 1 g/liter glucose.

**PTS Transport Assays**

Phosphoenolpyruvate-dependent transport of lactose and β-glucoside sugars was measured by a modification of the Imai and Hall (1981) method. Briefly, cells growing exponentially in minimal medium were concentrated to an A₆₀₀ of 10 in TP buffer (50 mM Tris-HCl [pH 7.6], 10 mM MgCl₂) and were permabilized by adding 10 μl of 40% (v/v) tolune in ethanol to 1.0 ml of concentrated cells and vortexing thrice with 3 s with a 3-s pause between vortexing cycles. Permabilized cells were stored on ice and were assayed without 10 min. Freshly prepared assay cocktail was maintained at 30°C until used. One hundred microliters of permabilized cells was added to 4.4 ml of assay cocktail, and 900-μl aliquots were distributed to four cuvettes maintained at 30°C in the temperature-controlled chamber of a Pharmacia Biotech Utlrospec 2000 UV-Vis spectrophotometer. Assays were initiated by adding 100 μl of 40 mM sugar to three of the cuvettes and mixing. The fourth cuvette, which received 100 μl of buffer, served as a control. Each reaction thus contained 20 μl cells, 100 μM NADH, 0.4 U lactate dehydrogenase, 1.0 mM phosphoenolpyruvate, and 4.0 mM of either sugar or buffer. The sugar-dependent release of pyruvate was monitored as the decrease in Δ₆₄₅ resulting from the oxidation of NADH to NAD⁺. The reaction rate was determined by subtracting the ΔA₆₄₅ of the control cuvette from the mean ΔA₆₄₅ of the experimental cuvettes. Units are expressed as nanomoles of sugar transported per minute.

**Phospho-β-Glycosidase Assays**

Phospho-β-glycosidase activity and phospho-β-galactosidase were assayed, respectively, on p-nitrophenyl-β-glucoside-6-PO₄ (PNPGluP) or p-nitrophenyl-β-galactoside-6-PO₄ (Sigma) (OPNGalP), in 50 mM Tris-HCl, 10 mM MgCl₂ (pH 7.5) buffer at 30°C by continuously monitoring the ΔA₄₂₀ in a Pharmacia Biotech Ultraspec 2000 UV-Vis spectrophotometer. Protein concentrations were determined by the Bradford (1976) assay.

**Molecular Methods**

The bglA and ascB genes were cloned into the expression vector pPROEX-HTa (Gibco). Expression of the cloned genes results in a protein with a 6x histidine tag separated from the mature peptide by a seven-amino-acid spacer. Sequences of bglA and ascB were taken from the complete sequence of the genome of E. coli K12 (GenBank accession number U00096) (Blattner et al. 1997). The host for the expression of the cloned genes was PEP24, a strain in which all phospho-β-glycosidase genes are either deleted or silent and in which no phospho-β-glucosidase activity is detectable (Thompson et al. 1999).

The bglA and ascB genes were amplified with the FailSafe (Epigen) polymerase system, which has a lower error rate than Taq polymerase.

The bglA gene was amplified using primer 1 (ATCTGGACCCCATGTTAAGGAAGAAA), which has an NcoI site at the ATG start of the bglA coding region, and primer 2 (GGAGCTCGATTTGCATCCGGTACTCTCATCGAC), which has a SacI site (boldface) at the 5’ end of the primer. The resulting amplicon and the vector were each digested with restriction endonucleases NcoI and SacI, the digests were purified with Qiagen PCR purification columns according to the manufacturer’s instructions, and the purified fragments were ligated and transformed into strain PEP24 by electroporation. The inserts in several plasmids were sequenced, and one whose sequence was identical to the published sequence of bglA was retained and the plasmid was designated pBglA.

The ascB gene was amplified using primer 3 (AACTAGTCCGACCGTCCCGAAAAGGAAGATA), which has a SpeI site at the 5’ end, and primer 4 (GGAGCTCGATTTGCATCCGGTACCATGAC), which has a SacI site at the 5’ end. The amplicon was digested with SpeI and SacI and ligated into similarly digested pBluscriptKS+ (Stratagene) to construct plasmid pAscBlu. The ATG start codon of ascB in pAscBlu was converted to an NcoI site by site-directed metagenesis using the Transformer kit (Clonetech) according to the manufacturer’s instructions to create pAscBluM. The NcoI/Sacl fragment was excised from pAscBluM and ligated into NcoI/Sacl-digested pPROEX-HTa to create plasmid pAscBNco. Because creation of the NcoI site had changed Ser₁ to an Ala, site-directed mutagenesis was used to restore Ser₁. Sequencing of the ascB coding region revealed that the various manipulations had resulted in a Ser₁→Gly substitution, so site-directed mutagenesis was used to restore the correct sequence at that
point. The sequence of ascB coding region of the final product, pAscB, was confirmed by DNA sequencing. pAscB was transformed into strain PEP24.

Purification of BglA and AscB Proteins

Strain PEP24/pBglA was grown at 37°C to mid-log phase (A\textsubscript{600} = 0.5) in 500 ml L broth containing 100 µl/ml ampicillin, at which point isopropyl-β-D-thioglactopyranoside (IPTG) was added to a final concentration of 0.6 mM to induce expression of bglA. After 3 h, the cells were harvested and resuspended in 20 ml B-PEP24.

Monocultures of the two strains were grown overnight at 30°C in minimal medium containing 0.1 g/liter glucose, a concentration that limits growth to about 10\textsuperscript{8} cells/ml. Equal volumes of the two cultures were mixed, and the mixed cultures were used to fill the growth chambers of 30-ml chemostats that were maintained at 30°C. One chemostat was connected to a reservoir containing minimal medium with 0.1 g/liter glucose, and the other to a reservoir containing minimal medium with 0.1 g/liter glucose and 0.5 g/liter lactose. Media from the reservoirs was pumped into the growth chambers at 10 ml/h and was mixed well by bubbling sterile air into the growth chamber. Overflow from the growth chamber was collected sterilely. Samples were collected from the overflow at the beginning and end of each day, and the total overflow volumes between samples were recorded. Samples were suitably diluted and were plated onto 1 g/liter glucose minimal medium to estimate the total number of cells per milliliter and onto 1 g/liter arbutin minimal medium to estimate the number of β-glucoside-positive cells per milliliter. The number of arbutin-negative (PEP24) cells per milliliter was estimated by subtraction. Each experiment was run for 5 days, about 170–180 generations. The selection coefficient for the β-glucoside-positive strain, s, was estimated from a least-squares fit of the natural log of the ratio of the β-glucoside-positive to the β-glucoside-null strain versus generation, where a generation is defined as one growth chamber volume of medium collected from the overflow. Fitness of the β-glucoside-positive strain is defined as 1 + s.

Results Phylogenetic Analyses

Phospho-β-Galactosidases

Phospho-β-galactosidases have been studied experimentally in Lactococcus lactis, Staphylococcus aureus, Streptococcus mutans, Lactobacillus casei (Witt, Frank, and Hengstenberg 1993), and Lactobacillus gasseri (Saito et al. 1998). The protein sequences of each of those genes were used to search the K12 genome by the BLASTP program (Altschul et al. 1990, 1997). Each of those searches identified the same three genes, bglA, bglB, and ascB, whose products were significantly similar (P < 10\textsuperscript{-15}) to the phospho-β-galactosidase probe sequence. Each of those genes encodes a phospho-β-glucosidase.

The bglA gene encodes a constitutively expressed enzyme, phospho-β-glucosidase A (Prasad, Young, and Schaeffer 1973) that is specific for hydrolysis of phosphorylated arbutin.

The bglB gene is part of the cryptic bglGFB operon and encodes phospho-β-glucosidase B, which hydrolyzes phosphorylated arbutin and salicin (Prasad and Schaeffer 1974; Mahadevan, Reynolds, and Wright 1987; Schnetz, Toloczky, and Rak 1987; Schnetz and Rak 1988). The bgl operon is maintained in a silent state by silencer elements that are located both upstream and downstream of the promoter and the CAP binding site (Schnetz 1995). The operon can be activated by muta-
tions that disrupt those flanking silencer element sequences. The most common disruptions involve insertions of IS1 or IS5 into that 223-bp region (Schnetz and Rak 1988), but rare IS2, IS10, or TN1000 insertional activation and activation by deletion of portions of the sequence upstream of the CAP binding site have also been reported (Schnetz 1995; Schnetz and Rak 1992). The bgl operon can also be activated by either of two base substitutions in the CAP binding site (Schnetz and Rak 1992).

The ascB gene is part of the cryptic ascGFB operon and encodes a phospho-β-glucosidase that hydrolyzes phosphorylated arbutin, salicin, and cellobiose (Hall and Xu 1992). The asc operon is activated by insertion of IS186 into ascG at a single site near the 3’ end of that gene (Hall and Xu 1992).

Sequences of each of the three E. coli phospho-β-glucosidase proteins were used to search the nonredundant protein database at NCBI using the BLASTP program. The 22 most similar sequences (table 1) were aligned with BglA, BglB, and AscB as described in Materials and Methods. The resulting DNA alignment was used to construct a phylogenetic tree (fig. 2) by Bayesian analysis using the program MrBayes as described in Materials and Methods. The evolutionary model was the General Time Reversible model, and among-site variation in evolutionary rate was estimated separately for first, second, and third positions of sites within codons. Four chains, with a “temperature” of 0.2 for the heated chains, were run for 496,500 generations, sampling trees every 100 generations. The ln likelihood of the trees had converged on a constant value by generation 10,000, i.e., after saving 100 trees. The consensus tree was calculated from the final 4,200 trees visited, well after convergence had occurred.

Among the E. coli phospho-β-glucosidases, the closest relative of a known phospho-β-galactosidase, phospho-β-galactosidase 1 of Lactobacillus gasseri, is BglA (fig. 2).

**PTS Enzyme II-lac**

Each of the known PTS Enzyme II-lactose protein sequences was used to search the E. coli K12 genome
by the BLASTP program. Only one protein, EnzymeII-
cellulobiose, encoded by the cryptic cellB gene, showed
significant homology (P < 10^{-3}) to any of the Enzyme
II-lactose proteins. The E. coli CelB protein sequence
was used to search the NCBI nonredundant protein se-
quenence data base. Eleven proteins, including all of
the known Enzyme II-lactose proteins, that are significantly
similar to CelB (P < 10^{-4}) were identified. The DNA
sequences encoding those proteins and the E. coli CelB
protein were aligned as described in Materials and
Methods. The aligned sequences were used to construct
a Bayesian phylogenetic tree (fig. 3) exactly as for the
phospho-β-galactosidases above.

Structural Analysis

Each of the proteins AscB, BglA, and BglB were
submitted to the 3Dcrunch project (Peitsch et al. 1997)
to identify structural homologs, and each identified the
same four structures: 1PBG, 2PBG, 3PBG, and 4PBG,
the four crystal structures of the L. lactis phospho-β-
galactosidase (Wiesmann et al. 1995). One of those
structures, 4PBG, includes galactose-6-phosphate at the
reaction center.

The majorities of the three E. coli proteins (residues
58–454 out of 479 amino acids of BglA, 57–437 out of
470 for BglB, and 56–471 out of 474 for AscB) were
structurally alignable with the phospho-β-galactosidase
structures and were used to construct the model struc-
tures. Swiss-Pdb Viewer (Guex and Peitsch 1997) was
used to visualize the superimposed models of the four
proteins.

Figure 4 shows the active-site residues of the L.
lactis phospho-β-galactosidase including hydrogen

---

Fig. 2.—Bayesian phylogeny of glycosidases related to known phospho-β-galactosidases.

---

Fig. 3.—Bayesian phylogeny of PTS transport genes related to known PTS Enzyme II-lac’s.

---

Fig. 4.—Structure of the active site of the Lactococcus lactis 6-
phospho-β-galactosidase. Active-site residues and the substrate galac-
tose-6-phosphate are shown. Dashed lines indicate hydrogen bonds to
the substrate. Homologous residues of the BglA, BglB, and AscB pro-
tein are shown where those residues are not identical to those of the
6-phospho-β-galactosidase.
bonding with galactose-6-phosphate (Wiesmann et al. 1995). All but two of those residues are identical in the three E. coli phospho-β-glucosidases, the exceptions being Asn297, which is Ser in all three phospho-β-glucosidases, and Trp429, which is Ala in BglB and AscB, but Phe in BglA. Weisman et al. (1995) point out that these exchanges suggest that the binding patterns of the phospho-β-glucosidases for glucoside substrates is likely to be quite different from that of the phospho-β-galactosidase for the galactoside substrate because, “Presumably, the 6-phospho-glucosides with their equatorial O4-atom show nothing equivalent to the intrinsic hydrogen bond between the axial O4-atom and the phosphate in Gal-6-P.” The similarity between Trp429 of the phospho-β-galactosidase and the Phe433 of the BglA phospho-β-glucosidase suggests that BglA might be able to evolve phospho-β-galactosidase by substitution of Asn for Ser315, which is homologous to Asn297 of the L. lactis phospho-β-galactosidase, while BglB and AscB would require two substitutions to fully superimpose the active-site residues.

Biochemical Analyses

**PTS Transport Activities**

In addition to the cryptic cel operon, which includes the celABC genes for PTS-mediated transport of cellobiose (Kricker and Hall 1987; Parker and Hall 1990), E. coli K12 has three other genes for PTS-mediated transport of β-glucoside sugars. The bglF and ascF genes belong to the same cryptic operons, respectively, as the bglB and ascB genes. The location of the cryptic arbT gene, which encodes a system for the PTS-mediated transport of arbutin (Kricker and Hall 1987), is not known. Spontaneous arbT- mutants are easily isolated in ΔbglFG and ΔcelBCDF strains (unpublished results) and are able to utilize arbutin because of the constitutively expressed bglA gene. Inactivation of bglA by insertion of a transposon renders arbT+ strains phenotypically arbutin-negative.

The strains listed in table 2 express the indicated genes constitutively, and all other PTS β-glucoside genes are either deleted or silent (see Materials and Methods), and all strains are lactose-negative. Cultures grown in glucose minimal medium were assayed for sugar transport as described in Materials and Methods. Because the glucose transport system is expressed constitutively, glucose transport served as an internal control for day-to-day variation in PTS activity. Activities relative to glucose are probably better measures of effective transport rates than are the absolute measurements.

The bglF-encoded protein was not previously known to transport cellobiose, and indeed strain PEP25, which grows well on both arbutin and salicin, is cellobiose negative. The observation that strain PEP25 transports cellobiose at about the same rate that it does salicin is consistent with the notion that the bglF-encoded phospho-β-glucosidase, part of the same operon as bglF, is inactive toward cellobiose-6-phosphate. The observation that strain PEP25 transports lactose at 7% of the rate at which it transports glucose is surprising and is, to the best of my knowledge, the first evidence for transport of a β-galactoside sugar by the bglF-encoded protein.

The relative rates at which the celABC-encoded PTS system transports glucose, cellobiose, arbutin, and salicin in strain PEP14 is consistent with earlier observations (Kricker and Hall 1987), but this is the first evidence for lactose transport by that system. The celABC system transports lactose fairly efficiently, at 24% of the glucose transport rate. The lactose transport rate by the celABC system is as high as the salicin transport rate by the bglF system, which allows good growth on salicin. The lactose-negative phenotype of strain PEP14 cannot, therefore, be attributed to a low lactose transport rate and must result from the failure of the celF-encoded phospho-β-glucosidase to hydrolyze lactose-phosphate at a rate sufficient for growth. Neither ArbT nor AscF transports lactose at a detectable rate.

**Phospho-β-Glucosidase Activities**

The BglA and AscB proteins were expressed and purified as described in Materials and Methods. BglA was estimated to be 98% pure and AscB to be 85% pure by SDS acrylamide gel electrophoresis. The purified proteins were assayed for phospho-β-glucosidase and phospho-β-galactosidase activities as described in Materials and Methods (table 3). The Michaelis-Menten kinetic parameters $K_m$ and $K_{cat}$ were determined in duplicate trials at $2\times$ increasing concentrations from 0.015625 mM to 1.0 mM PNPGluP and were estimated as previously described (Hall 1976) (table 3). Activities toward the galactoside substrate ONPGalP were deter-

---

**Table 2: Sugar Transport in Strains Expressing β-Glucoside Utilization Operons**

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genes</th>
<th>Glucose ($\mu$M)</th>
<th>Arbutin ($\mu$M)</th>
<th>Salicin ($\mu$M)</th>
<th>Cellobiose ($\mu$M)</th>
<th>Lactose ($\mu$M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEP24</td>
<td>null</td>
<td>472 ± 6.3</td>
<td>2.0 ± 1.8</td>
<td>17.9 ± 20.6</td>
<td>−11 ± 5.6</td>
<td>−3.4 ± 1.6</td>
</tr>
<tr>
<td>PEP25</td>
<td>bglF</td>
<td>352 ± 10.6</td>
<td>163 ± 13.6</td>
<td>42.7 ± 7.4</td>
<td>37.2 ± 10.5</td>
<td>26.2 ± 6.4</td>
</tr>
<tr>
<td>PEP14</td>
<td>celABC</td>
<td>246 ± 2.1</td>
<td>174 ± 12.4</td>
<td>297 ± 11.7</td>
<td>387 ± 5.3</td>
<td>58 ± 10.3</td>
</tr>
<tr>
<td>PEP33</td>
<td>arbT</td>
<td>529 ± 46</td>
<td>541 ± 7.9</td>
<td>282 ± 10.9</td>
<td>74.5 ± 5.7</td>
<td>−7.8 ± 5.4</td>
</tr>
<tr>
<td>PEP54</td>
<td>ascF</td>
<td>201 ± 3.7</td>
<td>30.7 ± 3.5</td>
<td>126 ± 9.8</td>
<td>7.6 ± 1.3</td>
<td>4.3 ± 7.5</td>
</tr>
</tbody>
</table>

**Note.**—Values are nanomoles of sugar transported per minute per milliliter of cells at an $A_{600}$ of 1.0 ± SE. All assays were done at 30°C at an external concentration of 4 mM sugar. Numbers in parentheses are transport rates relative to glucose transport in that strain.
Phospho-β-glucosidase Activities on PNPGluP and OPNGalP

<table>
<thead>
<tr>
<th>Property</th>
<th>BglA</th>
<th>AscB</th>
<th>BglB</th>
<th>CelF</th>
</tr>
</thead>
<tbody>
<tr>
<td>V_max PNPGluP</td>
<td>133,333</td>
<td>b</td>
<td>(1.48 ± 0.06) × 10^3</td>
<td>6.5 × 10^2</td>
</tr>
<tr>
<td>K_m (mM) PNPGluP</td>
<td>6.0 ± 0.01</td>
<td>0.08</td>
<td>0.18</td>
<td>6.0 ± 0.01</td>
</tr>
<tr>
<td>K_cat /PNPGluP</td>
<td>6.136</td>
<td>b</td>
<td>12.6</td>
<td>6.0 ± 0.5</td>
</tr>
<tr>
<td>K_cat /K_m /s/mole</td>
<td>(1.32 ± 0.04) × 10^6</td>
<td>(7.12 ± 2.5) × 10^4</td>
<td>ND c</td>
<td>(3 × 10^5)</td>
</tr>
</tbody>
</table>

Units are nanomoles substrate hydrolyzed at 2 mM substrate at 30°C. Velocity at 2 mM PNP-glucoside-6-P was calculated from K_m and K_cat. ND: not determined because instability of protein precludes quantifying the fraction of protein that is active.

Table 3: Phospho-β-glucosidase Activities on PNPGluP and OPNGalP

Preferences for the phospho-β-active and the BglA protein is only 2.1-fold that activity relative to phospho-β-arbT pose, the constructed (see Materials and Methods) operons would permit lactose utilization, strains constitutively expressing the six possible pairs of operons were constructed (see Materials and Methods). For this purpose, the arbtT-encoded transport system and the bglA-encoded phospho-β-glucosidase were considered an “operon,” although there is no evidence that they are so organized. All six combinations were lactose negative; thus, none of the operons under consideration, singly or in pairs, allow utilization of lactose as a sole carbon and energy source.

Competition Experiments

Because the natural substrates arbutin, salicin, cellobiose, and lactose are not commercially available in the phosphorylated form, it is necessary to assay phospho-β-glucosidase and phospho-β-galactosidase activities using the analogs p-nitrophenyl-β-D-glucopyranoside and o-nitrophenyl-β-D-galactopyranoside. Nitrophenyl substrates are often hydrolyzed more efficiently than corresponding natural sugars; thus, inferences of lactose-phosphate hydrolysis rates from ONPGalP hydrolysis rates are not as strong as one might like. Thus, the observation that BglB enzyme hydrolyzes ONPGalP about 20% as well as it does PNPGluP and that BglF transports lactose at more than half of the rate at which it transports salicin does not necessarily mean that constitutive expression of the bglGFB operon is likely to permit utilization of lactose by the PTS pathway. Indeed, all of the strains in table 2 are completely lactose-negative; i.e., they are unable to form colonies on lactose minimal plates.

On the other hand, lactose utilization might require a PTS transport system encoded by one operon in conjunction with a phospho-β-galactosidase encoded by another operon. To determine whether any combination of β-glucoside transport genes and phospho-β-galactosidase genes would permit lactose utilization, strains constitutively expressing the six possible pairs of operons were constructed (see Materials and Methods). For this purpose, the arbtT-encoded transport system and the bglA-encoded phospho-β-glucosidase were considered an “operon,” although there is no evidence that they are so organized. All six combinations were lactose negative; thus, none of the operons under consideration, singly or in pairs, allow utilization of lactose as a sole carbon and energy source.
Although none of the combinations permit utilization of lactose as a sole carbon and energy source, one or more combinations might allow sufficient cometabolism of lactose to provide a selective advantage when other carbon sources are limiting. To test that possibility, strains expressing pairs of β-glucoside utilization operons were competed with strain PEP24, which is null for all four operons, in glucose-limited chemostats and in glucose-limited chemostats with excess lactose (Materials and Methods). The relative fitnesses of the β-glucoside-positive strains were estimated from the change in the ratio of β-glucoside-positive to β-glucoside-null cells per generation.

Table 4 shows the relative fitnesses of each β-glucoside-positive strain in the presence and absence of lactose based on two independent paired chemostat experiments. Paired t-tests showed that in no case did lactose confer a significant fitness advantage upon a β-glucoside-positive strain. Thus, despite the fact that the celABC system permits PTS-mediated transport of lactose at 24% of the glucose transport rate (table 2), none of the phospho-β-glucosidases hydrolyze the lactose phosphate at a rate that provides any growth advantage. While these experiments do not help predict which genes might have more potential for evolving a lactose PTS/phospho-β-galactosidase system, they do confirm that any potential is strictly that, a future potential, and not a present capability.

Discussion

On the basis of phylogenetic distance from a known phospho-β-galactosidase, bglA is the E. coli gene most likely to evolve phospho-β-galactosidase activity, followed in descending order by ascB and bglB; i.e., bglA > ascB > bglB.

On the basis of structural homology with the L. lactis phospho-β-galactosidase active-site residues, bglA is the E. coli gene most likely to evolve phospho-β-galactosidase activity, and that activity is predicted to arise via substitution of an asparagine for the serine at position 315 of the BglA protein.

On the basis of biochemical activity, the E. coli gene most likely to evolve phospho-β-galactosidase activity is bglB. The order bglB > ascB > bglA > celF is the opposite of the order predicted by phylogenetic considerations. Both phylogenetic and biochemical considerations predict that the celABC genes are the most likely to evolve into a lactose-PTS transport system.

With the predictions in hand, we can now begin to test those predictions by experimental evolution of lactose utilization using each of the four strains that expresses a single β-glucoside utilization operon, each of the six strains that express a pair of operons, and the null strain, which expresses none of the operons. Both in vivo selection of spontaneous Lac+ mutants and in vitro evolution of the individual transport and phospho-β-glucosidase genes will be used to test the predictions.

There are three possible outcomes of the experiments. It might prove impossible to isolate Lac+ mutants, Lac+ mutants might arise from mutations in genes other than the β-glucoside utilization genes, or they might arise from mutations in some of the predicted β-glucoside utilization genes. The first possibility has been ruled out. Spontaneous Lac+ mutants have been isolated from strain PEP60, a strain that expresses all four of the β-glucoside utilization genes (unpublished results). A parallel attempt to isolate Lac+ mutants from the β-glucoside null strain PEP24 failed. While not definitive, that result makes the second possibility unlikely.

Acknowledgments

I am grateful to Ross LaRosa for expert technical assistance. I am deeply grateful to Jack Thompson for the gift of p-nitrophenyl-β-glucopyranoside-6-phosphate, which is not commercially available, and to Tadao Saito for the gift of half of the world’s remaining supply of commercially available o-nitrophenyl-β-galactopyranoside-6-phosphate. I am grateful to Tony Dean for insisting that I consider structural homology as a predictor of enzyme evolution.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genes Expressed</th>
<th>Fitness in Glucose ± SE</th>
<th>Fitness in Glucose + Lactose ± SE</th>
<th>Ratio of Glucose + Lactose to Glucose Fitness ± SE</th>
<th>P*</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEP 27</td>
<td>celABCD and bglGFB</td>
<td>0.933 ± 0.011</td>
<td>0.943 ± 0.016</td>
<td>1.010 ± 0.029</td>
<td>0.69</td>
</tr>
<tr>
<td>PEP 36</td>
<td>bglA, arbT, and celABCD</td>
<td>0.898 ± 0.047</td>
<td>0.896 ± 0.006</td>
<td>0.999 ± 0.060</td>
<td>0.97</td>
</tr>
<tr>
<td>PEP 37</td>
<td>bglA, arbT, and ascGFB</td>
<td>0.882 ± 0.023</td>
<td>0.907 ± 0.005</td>
<td>1.028 ± 0.031</td>
<td>0.43</td>
</tr>
<tr>
<td>PEP 38</td>
<td>bglA, arbT, and bglGFB</td>
<td>0.875 ± 0.019</td>
<td>0.885 ± 0.008</td>
<td>1.012 ± 0.031</td>
<td>0.39</td>
</tr>
<tr>
<td>PEP 56</td>
<td>ascGFB and celABCD</td>
<td>0.952 ± 0.007</td>
<td>0.959 ± 0.008</td>
<td>1.007 ± 0.001</td>
<td>0.09</td>
</tr>
<tr>
<td>PEP 57</td>
<td>ascGFB and bglGFB</td>
<td>0.995 ± 0.002</td>
<td>0.990 ± 0.010</td>
<td>0.995 ± 0.009</td>
<td>0.56</td>
</tr>
</tbody>
</table>

* P is the probability that the ratio is not different from 1.0

**LITERATURE CITED**


HUELSMANN, J. P. 2000. MrBayes: Bayesian inferences of phylogeny. Rochester, N.Y.


Antony Dean, reviewing editor

Accepted March 23, 2001