Parallel Evolution of Ligand Specificity Between LacI/GalR Family Repressors and Periplasmic Sugar-Binding Proteins

Kaoru Fukami-Kobayashi, Yoshio Tateno, and Ken Nishikawa

Center for Information Biology and DDBJ, National Institute of Genetics, Mishima, Japan

The bacterial LacI/GalR family repressors such as lactose operon repressor (LacI), purine nucleotide synthesis repressor (PurR), and trehalose operon repressor (TreR) consist of not only the N-terminal helix-turn-helix DNA-binding domain but also the C-terminal ligand-binding domain that is structurally homologous to periplasmic sugar-binding proteins. These structural features imply that the repressor family evolved by acquiring the DNA-binding domain in the N-terminal of an ancestral periplasmic binding protein (PBP). Phylogenetic analysis of the LacI/GalR family repressors and their PBP homologues revealed that the acquisition of the DNA-binding domain occurred first in the family, and ligand specificity then evolved. The phylogenetic tree also indicates that the acquisition occurred only once before the divergence of the major lineages of eubacteria, and that the LacI/GalR and the PBP families have since undergone extensive gene duplication/loss independently along the evolutionary lineages. Multiple alignments of the repressors and PBPs furthermore revealed that repressors and PBPs with the same ligand specificity have the same or similar residues in their binding sites. This result, together with the phylogenetic relationship, demonstrates that the repressors and the PBPs individually acquired the same ligand specificity by homoplasious replacement, even though their genes are encoded in the same operon.

Introduction

Many bacterial transcription regulatory proteins have been identified and classified into several families on the basis of sequence similarity. The LacI/GalR family is one of those families, consisting of repressors (Vartak et al. 1991; Weickert and Adhya 1992; Nguyen and Saier 1995). This family is also called LacI family in Pfam (Bateman et al. 2000), and the HTH lacI family in PROSITE (Hofmann et al. 1999).

The crystal structures of three members of the LacI/GalR family, PurR (purine nucleotide synthesis repressor), LacI (lactose operon repressor), and TreR (trehalose operon repressor), clearly show that the LacI/GalR family repressors have two structural domains (Schumacher et al. 1994*a*; Friedman, Fischmann, and Steitz 1995; Lewis et al. 1996; Hars et al. 1998). The N-terminal domain is a helix-turn-helix DNA-binding domain, and the C-terminal domain is a ligand-binding domain whose 3D structure is similar to those of periplasmic binding proteins (PBPs), as shown in figure 1. The C-terminal domain is especially similar to the PBPs that bind sugars (Fukami-Kobayashi, Tateno, and Nishikawa 1999). It also shows a weak sequence homology to the PBPs (Mauzy and Hermodson 1992). It has thus been suggested that the C-terminal domain of the repressors and the PBPs share a common ancestor, and that the progenitor repressor was formed when the common ancestor acquired the DNA-binding domain in its N-terminal.

The PBP is one of the components of the ABC transporter and is involved in the active transport of water-soluble ligands. When PBP binds its ligand in the periplasmic space (or out of the bacterial cell), it undergoes a conformational change to bind the permease in the plasma membrane. A similar conformational change is observed in the C-terminal domain of the LacI/GalR family repressor, when the domain binds its ligand, which affects the DNA-binding affinity of the repressor. This implies that the repressor and PBP share a common ancestor.

Each repressor has its own ligand specificity, and so does each PBP. Some repressors and PBPs share the same ligand specificity. For instance, GalS (repressor, R) and MglB (PBP, P) both bind D-galactose/D-glucose; XylR (R) and XylF (P) bind D-xylose; and RbsR (R) and RbsB (P) bind D-ribose (Bairoch and Apweiler 2000). Such pairs, made up of a repressor and a PBP, are often encoded in a single operon in the *Escherichia coli* and *Bacillus subtilis* genomes (Itoh et al. 1999). Similar operon structures have been suggested to be present in other bacterial species (Tomii and Kanehisa 1998).

The question that then arises is, which occurred to the common ancestor first, acquisition of the DNA-binding domain or divergence of ligand specificity? If the acquisition of the DNA-binding domain occurred first, then only one acquisition is enough to generate the repressor family. That possibility is consistent with the results of phylogenetic analysis of several protein families that show domain organization was established in an early stage of their evolution (Fukami-Kobayashi, Tomoda, and Go 1993; Fukami-Kobayashi et al. 1996; Koyanagi et al. 1998). On the one hand, if this is the case, then ligand specificity must have evolved in the LacI/GalR family and in the PBP family independently, even if a pair of repressor and PBP is coded in the same operon. On the other hand, if the divergence of ligand specificity occurred first, then the ligand specificity must have been unchanged while the ancestral PBP gene duplicated and one of the duplicates acquired the DNA-binding domain to evolve into a repressor. This process would explain why a pair of repressor and PBP with the same ligand specificity is often encoded in a single operon. In the latter case, however, we have to assume that the acquisition of the DNA-binding domain occurred independently in each operon.

It has been reported that sequence similarity of the N-terminal DNA-binding domain is higher than that of the C-terminal ligand–binding domain in the LacI/GalR family

Key words: parallel evolution, repressor, periplasmic binding protein, operon, functional genomics, ABC transporter.

E-mail: kfukami@genes.nig.ac.jp.

Mol. Biol. Evol. 20(2):267–277. 2003
DOI: 10.1093/molbev/msg038

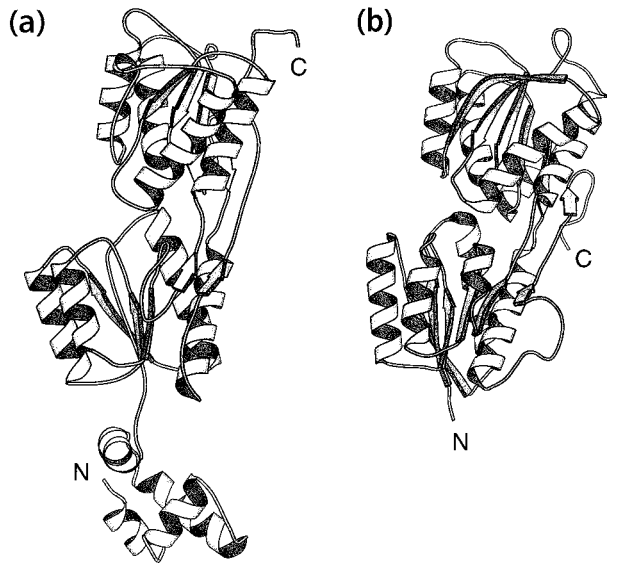

FIG. 1.—Crystal structure of (*a*) purine repressor and (*b*) D-galactose/D-glucose–binding protein. The C-terminal (upper) domain of purine repressor performs ligand binding, and its structure is quite similar to the D-galactose/D-glucose–binding protein. The N-terminal (lower) domain functions as DNA binding domain. The coordinates are from the PDB entries 1QPZ chain A (purR_Eco [Glasfeld et al. 1999]) and 2GBP (mglB_Eco [Vyas, Vyas and Quiocho 1988]).

(Weickert and Adhya 1992; Nguyen and Saier 1995). This conclusion implies that the functional divergence of ancestral PBPs took place prior to the acquisition of the N-terminal domain. In addition, the number of LacI/GalR repressors varies widely among bacterial species (Kawabata et al. 2002), indicating that the acquisition of the DNA-binding domain occurred independently to produce a repressor unique to each lineage.

When bacteria needed a repressor with novel ligand specificity, which of the above-mentioned two strategies did they employ? Did they make the repressor from an existing repressor by inventing the novel specificity, or from PBP by acquiring the DNA-binding domain? Because this problem is rooted in the acquisition of a new protein function, its solution will contribute to the prediction of unknown function of open reading frames (ORFs) identified in genome sequences. In this paper we report our approach to solving this problem.

Materials and Methods

Approach

Our approach to the solution is to reveal the evolutionary relationships of the repressors and the PBPs by constructing a phylogenetic tree. If the acquisition of the DNA-binding domain occurred first, the repressor and PBP can be expected to belong to separate clusters in the phylogenetic tree. If the divergence of ligand specificity took place first, however, pairs of repressors and PBPs with the same specificities will be considered to cluster together in the tree. Our approach is divided into five procedures as described in the paragraphs that follow.

The Database GTOP

To collect data on the repressors and their PBP homologues, we used the database GTOP (Genes TO Proteins, http://spock.genes.nig.ac.jp/~genome/gtop.html) (Kawabata et al. 2002) of December 2000. GTOP contains the results of various sequence analyses of all ORFs of the organisms for which the whole genome sequence has been reported. It particularly features an extensive utilization of protein 3D-structure information. The 3D structures of the ORFs have been predicted by PSI-BLAST (Altschul et al. 1997) search against PDB (Berman et al. 2000) in GTOP. The predicted 3D-structures are classified into categories according to the criterion of SCOP (Murzin et al. 1995). We also used GTOP for predicting the function of the ORF and the operon structure.

Sequence Data Collection

The C-terminal domains of the LacI/GalR family repressors are classified into "purine repressor (PurR) C-terminal domain," "lac-repressor (lacR) core C-terminal domain," and "trehalose repressor C-terminal domain" in SCOP. Our previous study (Fukami-Kobayashi, Tateno, and Nishikawa 1999) indicated that their closest PBP homologues are classified into "D-ribose–binding protein," "L-arabinose-binding protein," "D-allose–binding protein," and "galactose/glucose-binding protein." Those seven groups belong to "L-arabinose–binding protein-like" family (its hierarchy position is a/b class, periplasmic binding protein-like I fold, periplasmic binding protein-like I superfamily in SCOP). From GTOP, we selected the ORFs whose predicted 3D structures are classified into any of the seven groups.

Multiple Alignment

Some of the selected ORFs have a predicted all-alpha region in addition to the PBP-related sequence. When we made multiple alignments of the translated amino acid sequences from the ORFs, we removed the all-alpha regions for better alignment in the common region to all the ORFs.

We used ClustalW 1.81 (Thompson, Higgins, and Gibson 1994) and PRRN (Gotoh 1999) with the default parameters for multiple alignments of the amino acid sequences. To examine the performance of the two computer tools, we determined the spatially equivalent secondary structures by MATRAS (Kawabata and Nishikawa 2000) beforehand. The proteins and their PDB data subjected to the MATRAS determination were treR_Eco (1BYK chain A [Hars et al. 1998]), purR_Eco (1QPZ chain A [Glasfeld et al. 1999]), yjcX_Eco (1RPJ [Chaudhuri et al. 1999]), lacI_Eco (1TLF chain A [Friedman, Fischmann, and Steitz 1995]), rbsB_Eco (2DRI [Bjorkman et al. 1994]), mglB_Eco (2GBP [Vyas, Vyas, and Quiocho 1988]), and araF_Eco (8ABP [Vermersch et al. 1991]), which were selected from each of the seven SCOP groups by their resolution, R-factor, and source.

Construction of a Phylogenetic Tree

Using the result of the alignments, we constructed maximum likelihood (ML) trees by ProtML in MOLPHY (Adachi and Hasegawa 1996). A column of aligned sites that has one or more gaps was omitted from the tree construction. We adopted the JTT model (Jones, Taylor, and Thornton 1992) to compute the likelihood of a tree. First, 50 seed trees were generated by the quick add OTUs search mode of ProtML. Then, a tree with a larger likelihood value was searched by rearranging a part of the topology of a seed tree. After repeating this procedure for all seed trees, the tree with the largest likelihood value among the results was chosen as the ML tree. Reliability of each internal branch of the tree was then evaluated by the bootstrap probability (Felsenstein 1985), which was computed by the "resembling of estimated log-likelihood" (RELL) method (Hasegawa and Kishino 1994).

Identification of Ligand-Binding Sites in PBPs

We adopted the ligand-binding sites that have been identified by HBPLUS (McDonald and Thornton 1994) in the PDBsum database (Laskowski 2001). The PDB entries adopted were 1QPZ chain A (purine-binding sites), 1TLF chain A (lactose-binding sites), 2DRI (D-ribose–binding sites), 2GBP (D-galactose/D-glucose–binding sites), and 8ABP (L-arabinose–binding sites).

Results

LacI/GalR Family Is a Monophyletic Group

We selected ORFs of LacI/GalR family repressors and the PBP homologues from the published complete genome sequences. The number of the selected ORFs was 102. Table 1 shows the source organism, classification, number, name, predicted function (if any in the GTOP database) of the ORFs. We classified the ORFs into LacI/GalR family repressors (LGF), XylR type repressors (XR), PBPs, and others, according to the predicted domain organization and the length of the all-alpha domain. Each PBP precursor has a signal peptide to be transported to the periplasmic space in the N-terminal, and some of the signal peptides were predicted to consist of an all-alpha domain. The length of the domain is shorter than 40 aa in the ORFs that are well identified as PBP, whereas the DNA-binding domain of PurR and LacI is longer than 50 aa. We thus defined the ORF containing an all-alpha domain longer than 50 aa in the N-teminal as repressor. Because XylR and HI1106, unlike other repressors, have all-alpha domains in the C-terminal, they were classified as XylR type. We classified VCA0737 as "other," because it showed a strong sequence similarity to LuxP that has been proposed to function neither as repressor nor as a PBP, but as a periplasmic receptor in the Lux density-sensing system (Bassler, Wright, and Silverman 1994). Information about the ORFs—i.e., the source organism, classification, name, predicted function (if any in the GTOP database), predicted domain organization, and length of predicted all-alpha domain—are available in table 1S at http://spock.genes.nig.ac.jp/~kfukami/LGF/supplement/ or the masterfile of each ORF in the GTOP database.

We then performed multiple alignment of the translated amino acid sequences from the ORFs. Because identity was low among the sequences (<30% in many pairs), we expected that the results of alignment might vary with alignment algorithms. We thus used two different alignment programs, ClustalW and PRRN, and confirmed that the two programs performed in a similar fashion, so as to align the equivalent secondary structures. The result of the alignments is also available at http://spock.genes.nig.ac.jp/~kfukami/LGF/supplement/.

The ML tree was finally constructed using the ClustalW result (fig. 2). The LacI/GalR type repressors in magenta made one cluster containing the XylR type repressors (shown in green). Note that the bootstrap probability of the branch that divides the cluster from the others is 98%. We also constructed another ML tree using the PRRN result, which again shows that all the repressors are clustered together at the bootstrap probability of 100%. The tree is also available at http://spock.genes.nig.ac.jp/~kfukami/LGF/supplement/.

The two trees indicate that the domain organization of LacI/GalR type repressors emerged only once on the branch that divides the repressors and the PBPs, as indicated by the red arrow in figure 2. The establishment of the domain organization is thus predicted to predate the divergence of major lineage of eubacteria. Thereafter, the common ancestor diverged into repressors with a variety of ligand specificities. The XylR type repressor appeared during this process of divergence. The PBP also evolved to acquire a variety of ligand specificities after their divergence from the repressors.

Repressors and PBPs Acquired Their Ligand-Binding Sites Independently

There are pairs of PBP and repressor sharing the same ligand in the 102 ORFs selected (Bairoch and Apweiler 2000). The pairs are listed in table 2. In our analysis, we could deduce the ligands of two repressors VC2337 and TM0949 as D-galactose/D-glucose and D-ribose, respectively. The repressors and PBPs in table 2 are circled in color in figure 2 according to their ligand specificity as shown in the phylogenetic tree. Because the repressors and the PBPs are present in different clusters in the tree, the two must have acquired their ligand specificity independently, as illustrated in figure 3.

The genes of a pair of repressor and PBP sharing the same ligand specificity are often located close to each other on the same strand in the genome. Those pairs are shown as "vicinity" in table 2. Such a pair is highly likely to be located in the same operon. It is noted, however, that the repressor and the PBP acquired their ligand specificity independently, even in those cases.

In addition, the phylogenetic tree shows that the acquisition of specificity for ligands such as D-galactose/D-glucose and D-ribose might have occurred more than once within the repressor or PBP family. For example, because the D-galactose/D-glucose–binding protein from *Treponema pallidum* (TP0684) is not clustered with those of other species, *T. pallidum* probably acquired the ligand specificity independently. The phylogenetic tree also

Table 1
List of ORFs

Source Organism	Classification	Number	ORF name	Annotation
Escherichia coli (24)	PBP	9	araF_Eco	L-arabinose-binding periplasmic protein
			b1516_Eco	Putative lacI-type transcriptional regulator
			mglB_Eco	D-galactose-binding periplasmic protein
			rbsB_Eco	D-ribose-binding periplasmic protein precursor
			torT_Eco	Periplasmic protein TorT precursor
			yjcX_Eco	D-allose-binding periplasmic protein precursor
			yphF_Eco	ABC transporter periplasmic binding protein
			ytfQ_Eco	ABC transporter periplasmic binding protein
			xylF_Eco	D-xylose-binding periplasmic protein precursor
	LGF	14	ascG_Eco	Cryptic asc operon repressor
			cytR_Eco	Transcriptional repressor CytR
			ebgR_Eco	ebg operon repressor
			fruR_Eco	Fructose repressor(catabolite expressor/activator)
			galR_Eco	Galactose operon repressor
			galS_Eco	mgl repressor and galactose ultrainduction factor
			gntR_Eco	Gluconate utilization system Gnt-I transcriptional repressor
			idnR_Eco	L-iodinate regulatory protein
			lacI_Eco	Lactose operon repressor
			malI_Eco	Maltose regulon regulatory protein malI
			purR_Eco	Purine nucleotide synthesis repressor
			rbsR_Eco	Ribose operon repressor
			treR_Eco	Trehalose operon repressor
			ycjW_Eco	Hypothetical transcriptional regulator in ompG-tyrR intergenic region
	XR	1	xylR_Eco	Xylose operon regulatory protein
Vibrio cholerae (16)	PBP	4	VC1101	Conserved hypothetical protein
			VC1325	Galactoside ABC transporter, periplasmic D-galactose/D-glucose-binding protein
			VCA0130	Ribose ABC transporter, periplasmic D-ribose-binding protein
			VCA0710	Periplasmic protein TorT
	LGF	11	VC0289	Gluconate utilization system gnt-I transcriptional repressor
			VC0909	Trehalose operon repressor
			VC1286	Transcriptional regulator, LacI family
			VC1557	Transcriptional regulator, LacI family
			VC1721	Transcriptional regulator, LacI family
			VC2677	Transcriptional repressor, LacI family
			VC2337	Transcriptional regulator, LacI family
			VCA0654	Sucrose operon repressor ScrR, putative
			VCA0673	Transcriptional regulator, LacI family
			VCA0132	Ribose operon repressor
			VCA0519	Fructose repressor
	other	1	VCA0737	luxP protein
Haemophilus influenzae (7)	PBP	3	HI0504	D-ribose-binding periplasmic protein precursor
			HI0822	D-galactose-binding periplasmic protein
			HI1111	D-xylose-binding periplasmic protein precursor
	LGF	3	HI0506	Ribose operon repressor
			HI0821	Galactose operon repressor
			HI1635	Purine nucleotide synthesis repressor
	XR	1	HI1106	Xylose operon regulatory protein
Pseudomonas aeruginosa (5)	PBP	1	rbsB_Pae	Binding protein component precursor of ABC ribose transporter
		4	fruR_Pae	Fructose transport system repressor FruR
			gntR_Pae	Transcriptional regulator GntR
			ptxS_Pae	Transcriptional regulator PtxS
			rbsR_Pae	Ribose operon repressor RbsR
Bacillus subtilis (13)	PBP	1	rbsB_Bsub	D-ribose-binding protein precursor
	LGF	12	araR_Bsub	Transcriptional regulator (LacI family)
			ccpA_Bsub	Catabolite control protein A (glucose-resistance amylase regulator)
			degA_Bsub	Degradation activator
			lacR_Bsub	Transcriptional regulator (LacI family)
			kdgR_Bsub	kdg operon repressor
			msmR_Bsub	Transcriptional regulator (LacI family)
			rbsR_Bsub	Ribose operon repressor
			yhjM_Bsub	
			yjmH_Bsub	

Table 1
Continued

Source Organism	Classification	Number	ORF name	Annotation
			ykvZ_Bsub	
			yvdE_Bsub	Hypothetical transcriptional regulator in clpP-crh intergenic region
			yyaG_Bsub	Catabolite control protein B
Bacillus halodurans (22)	PBP	7	BH2323	Sugar ABC transporter (sugar-binding protein)
			BH3442	Multiple sugar transport system (multiple sugar-binding protein)
			BH3448	Sugar ABC transporter (sugar-binding protein)
			BH3840	Sugar ABC transporter (sugar-binding protein)
			BH3901	Rhizopine ABC transporter (rhizopine-binding protein)
			med_Bhal	Transcriptional activator of comK gene
			rbsB_Bhal	Ribose ABC transporter (ribose-binding protein)
	LGF	15	BH0901	Transcriptional regulator
			BH1250	Transcriptional regulator (LacI family)
			BH1516	Glucose-resistance amylase regulator
			BH1855	Transcriptional regulator
			BH1928	Transcriptional regulator (LacI family)
			BH2313	Transcriptional regulator
			BH2923	Transcriptional regulator involved in carbon catabolite control
			BH3230	Transcriptional regulator (LacI family)
			BH3692	Transcriptional regulator
			araR_Bhal	Transcriptional repressor of the arabinose operon
			ccpA_Bhal	Transcriptional regulator involved in carbon catabolite control
			degA_Bhal	Transcriptional regulator involved in degradation
			lacR_Bhal	Transcriptional repressor (beta-galactosidase gene)
			msmR_Bhal	Transcriptional regulator (AraC/XylS family)
			rbsR_Bhal	Transcriptional repressor of the ribose operon
Mycobacterium tuberculosis (1)	LGF	1	Rv3575c	Hypothetical protein Rv3575c
Borrelia burgdorferi (4)	PBP	4	BB0382	Basic membrane protein B precursor
			BB0383	Basic membrane protein A precursor (immunodominant antigen P39)
			BB0384	Basic membrane protein C precursor
			BB0385	Basic membrane protein D precursor
Treponema pallidum (2)	PBP	2	TP0545	Methylgalactoside abc transporter, periplasmic galactose-binding protein (mglB-1)
			TP0684	Glucose/galactose-binding lipoprotein precursor
Thermotoga maritima (8)	PBP	2	TM0114	Sugar ABC transporter, periplasmic sugar-binding protein
			TM0958	Ribose ABC transporter, periplasmic ribose-binding protein
	LGF	6	TM0275	Transcriptional regulator, GntR family
			TM0299	Transcriptional regulator, LacI family
			TM0949	Transcriptional regulator, LacI family
			TM1200	Transcriptional regulator, LacI family
			TM1218	Transcriptional regulator, LacI family
			TM1856	Transcriptional regulator, LacI family

indicates that the ribose operon repressors have evolved in four lineages; the first to *Escherichia coli*, *Vibrio cholerae*, and *Haemophilus influenzae*, the second to *Pseudomonas aeruginosa*, the third to *B. subtilis* and *B. halodurans*, and the fourth to *T. pallidum*. As for the PBPs for D-ribose, it is not clear whether the acquisition occurred once or more, because of the low bootstrap probability (48%) of the branch that divides the ribose-binding protein of *P. aeruginosa* (rbsB_Pae) from the others.

Figure 4 shows multiple alignments in three regions including ligand-binding sites of three sets of the amino acid sequences. The upper set refers to ribose-binding proteins and ribose operon repressors, the middle set to repressors of known 3D structure and the lower set to repressors and binding proteins for D-galactose/D-glucose. The ligand-binding sites are known in rbsB_Eco, purR_Eco, lacI_Eco, treR_Eco, and mglB_Eco from their crystal structure. They are enclosed in a red box in the alignments in figure 4. Although overall amino acid similarity is higher within each of the PBPs and the repressors than between them as shown in the tree, the similarity of the ligand-binding sites alone is higher between the PBP and the repressor sharing the same ligand than within each PBP and repressor pair. For example, in figure 4 all ribose-binding proteins and ribose operon repressors have asparagine (N) at site 54, whereas the proteins for D-galactose/D-glucose, except TP0684, have aspartic acid (D) at that site. Most proteins for D-ribose have DR/DW at sites 134–135 or

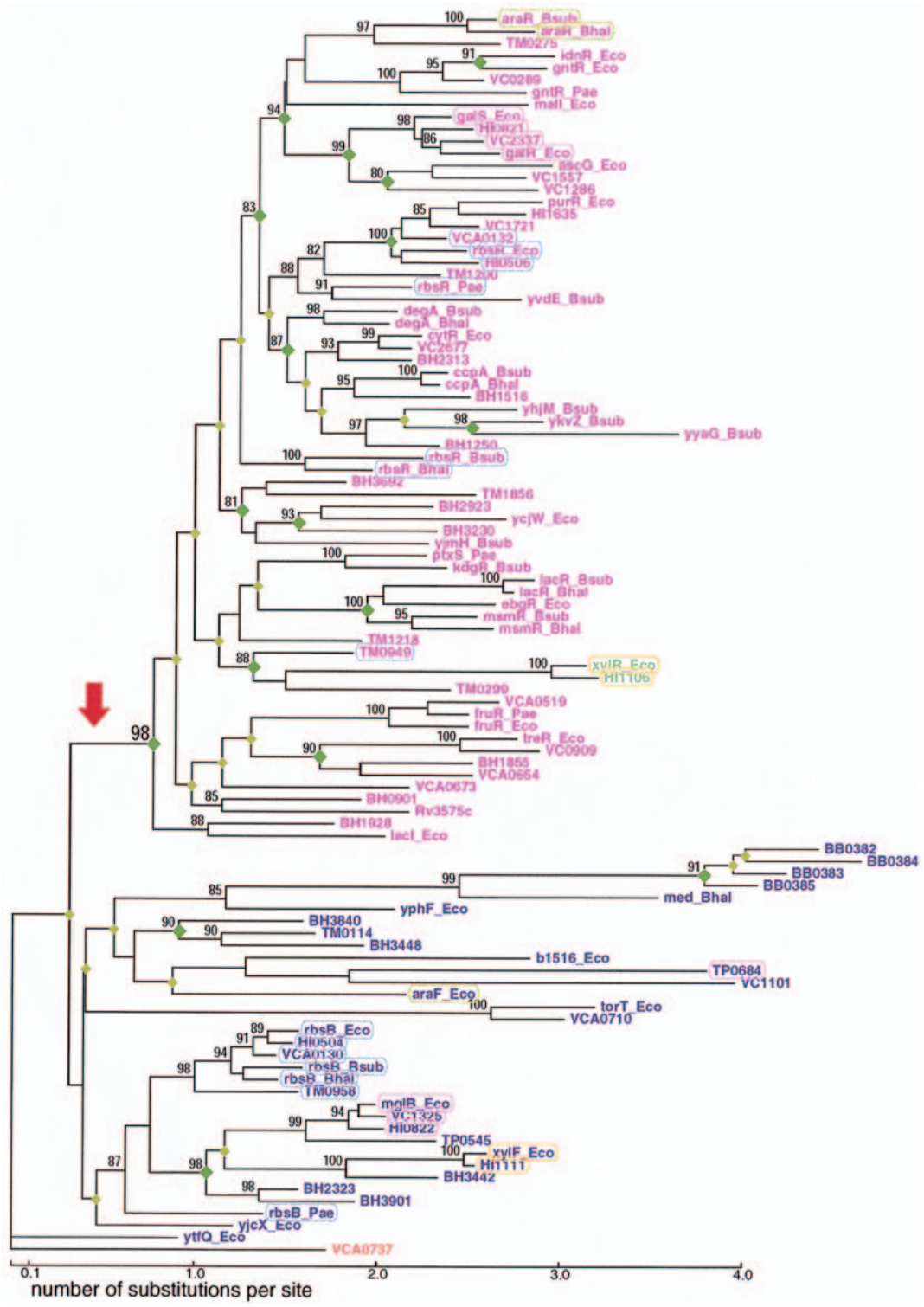

FIG. 2.—Phylogenetic tree of LacI/GalR type repressors, PBPs, and their homologues. This tree is constructed by the maximum likelihood method from ClustalW multiple alignment. Numbers at the branches show bootstrap probabilities. Green diamonds at the nodes indicate gene duplications that give rise to paralogous genes. The names of LacI/GalR type repressors appear in magenta; XylR type repressors, in green; PBPs, in blue; others, in orange. Repressors and PBPs with the same ligand specificity are circled in the matching colors: the specificity for L-arabinose is shown in green; D-galactose/D-glucose, in magenta; D-ribose, in blue; and D-xylose, in orange.

nearby, whereas the proteins for D-galactose/D-glucose have NR/NK. All proteins for D-galactose/D-glucose, except TP0684, have aspartic acid (D) at site 198. The aligned residues indicate that the PBP and the repressor independently acquired the same or a similar amino acid at their binding sites by homoplasious replacement to bind the same ligand, even if the pair is located in the same operon.

Table 2
List of Repressors and PBPs with Specificity for the Same Ligand

Ligand	Repressor	PBP	Vicinity in genome sequence
D-ribose	rbsR_Eco	rbsB_Eco	y
	VCA0132	VCA0130	y
	HI0506	HI0504	y
	rbsR_Pae	rbsB_Pae	y
	rbsR_Bsub	rbsB_Bsub	y
	rbsR_Bhal	rbsB_Bhal	y
	TM0949[a]	TM0958	y
D-galactose/D-glucose	galS_Eco	mglB_Eco	y
	galR_Eco		
	VC2337[b]	VC1325	
	HI0821	HI0822	y
		TP0684	
D-xylose	xylR_Eco	xylF_Eco	y
	HI1106	HI1111	y
L-arabinose		araF_Eco	
	araR_Bsub		
	araR_Bhal		

NOTE.—The following repressors were listed, despite lack of annotation about ligand specificity, for respective reasons:

[a] Seems to be on the same transcriptional unit as TM0958.

[b] Closely related to other D-galactose/D-glucose repressors.

Extensive Gene Duplication/Loss in the LacI/GalR and the PBP Families

The number of LacI/GalR repressors and PBPs varied from species to species, in the range of zero to 15 for the repressors and zero to 9 for the PBPs. This implies that gene duplication/loss took place in the evolution of these protein families. The green diamonds at the nodes of the phylogenetic tree in figure 2 indicate that gene duplication occurred there, which indicates that gene duplication/loss occurred more frequently than expected from the number of genes at present. ORFs of each species are distributed almost evenly throughout the tree, implying that the common ancestor of the major bacterial lineages acquired many genes for LacI/GalR type repressors and PBPs. At the same time, because some clusters contain ORFs from only one species, it is probable that gene duplication is still occurring.

Discussion

Monophyly of the LacI/GalR Family

The structural features of the LacI/GalR family are consistent with our finding that this family is a monophyly, which is to say that all of the all-alpha domains of LacI/GalR family repressors were classified into a single family "bacterial repressors," which is described as an all-alpha class, lambda repressor–like DNA-binding domains fold, lambda repressor–like DNA-binding domains superfamily in the SCOP hierarchy. Conversely, all members of the "bacterial repressors" family in SCOP are the N-terminal domains of the LacI/GalR family repressors. The tight linkage of the family and the N-terminal domains of the repressor probably reflects the single origin of the domain organization of the LacI/GalR family. Incidentally, the XylR-type repressor has a different origin from that family, and its all-alpha domains are classified into a different family "araC type transcriptional activator" in SCOP.

The regulatory mechanism of DNA binding is also consistent with our finding. The mechanisms of LacI and PurR are very similar to each other: they use a hinge helix for DNA binding, as well as the helix-turn-helix structure (Schumacher et al. 1994*a*, 1994*b*; Nagadoi et al. 1995; Bell and Lewis 2000). The hinge helix is formed in the presence of DNA in a region between the helix-turn-helix structure of the DNA-binding domain and the ligand-binding domain. Although the helix-turn-helix structure and recognition of the DNA major groove by the structure are found in other DNA-binding protein families, recognition of the DNA minor groove by the hinge helix is unique to the LacI/GalR family. The residues in the hinge helix play a crucial role in DNA binding of the repressors (Choi and Zalkin 1994; Pace et al. 1997). Even insertion of one glycine between the hinge region and the ligand domain causes a ~100-fold decrease in the affinity of the lactose repressor for its target DNA sequence (Falcon and Matthews 1999). In addition, the repressor functions as a dimer/tetramer, where the DNA binding is regulated by interactions not only between the two domains in one subunit but also between subunits. These findings together suggest that the regulatory mechanism is so elaborated that it is hardly possible that such a system was repeatedly formed in evolution. This suggestion is in agreement with our result that the LacI/GalR family is a monophyly.

Homoplasy of Binding Specificity

The phylogenetic tree, in contrast to the above monophyly, indicates that the ligand specificity did evolve more than once in both the PBP and the LacI/GalR families. Such a homoplasy is probable. The crystal structures of these families show that the number of the ligand-binding sites is no more than 20 (Laskowski 2001), which corresponds to only a small percentage of the total number of sites. In addition, ligand specificity is determined only by a few of them: the same residue is conserved at several ligand-binding sites regardless of the specificities. As shown in the multiple alignments in figure 4, ligand specificity can be acquired by replacement with the corresponding residue at a small number of sites. It is thus probable that parallel replacements at such limited sites occurred in evolution. In fact, convergent evolution by amino acid replacements at specific sites of a protein has already been observed in lysozyme (Stewart, Schilling, and Wilson 1987; Kornegay, Schilling, and Wilson 1994), color vision pigment (Yokoyama and Yokoyama 1990; Briscoe 2000), and blood group antigens (O'hUigin, Sato, and Klein 1997; Kitano et al. 2000; Sumiyama et al. 2000). It is expected that site-specific mutagenesis experiments in the ligand-binding sites will verify our prediction.

We consider that the homoplasy at the ligand-binding sites was evolutionarily fixed by selection, not just by chance. It is observed that genes in an operon have related functions. For example, PBP, a repressor, and other genes in an operon have related function for the same ligand

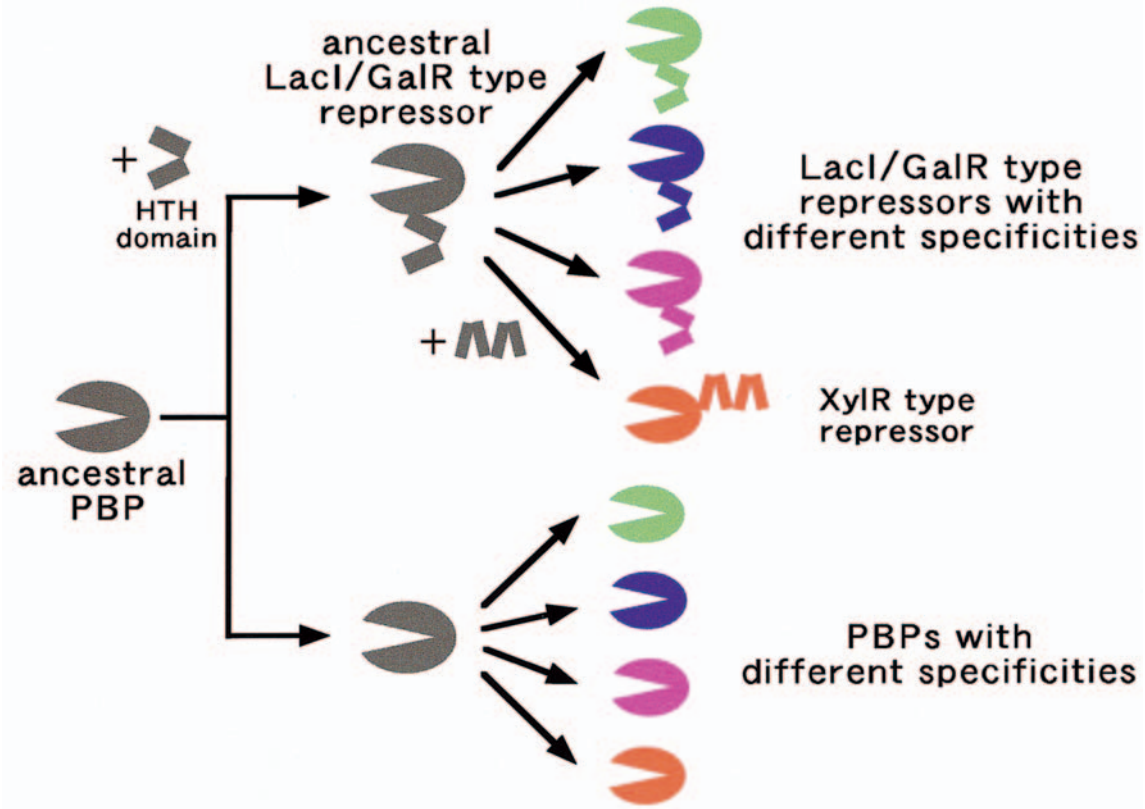

FIG. 3.—Evolutionary path of LacI/GalR repressor and PBP families deduced from the phylogenetic tree in figure 2. A progenitor of LacI/GalR type repressors emerged by acquring a DNA-binding domain in the N-terminal in one of the ancestral PBPs. The establishment of the domain organization occurred only once before the divergence of the major lineage of eubacteria. Then gene amplification and acquisition of various ligand specificities occurred independently in both the LacI/GalR repressor family and the PBP family. The XylR type repressor, which has HTH domains in the C-terminal, appeared during the divergence of the repressors.

(Tomii and Kanehisa 1998; Itoh et al. 1999). Most pairs of PBP and repressor genes sharing the same ligand are located in the "vicinity, as shown in table 2. In addition, the vicinity cannot be explained by a historical event that the ancestral operon possessed the same operon structure. Gene members and their order are not always the same in orthologous operons, which implies gene rearrangement in the operons in evolution. Nevertheless those operons have kept the genes with related functions together. These observations suggest that it is advantageous for those genes to be in the same operon, and that PBP and a repressor in the same operon are under constraint to have the same ligand specificity.

While ligand specificity of the PBP and the LacI/GalR families evolved from different origins, ligand-binding domains evolved from their common ancestor. Convergent evolution of protein functions is common, whereas that of protein structure is rare (Doolittle 1994). It is thus unlikely that the ligand-binding domains in the two families evolved independently into the same 3D structure. In addition, there is little functional necessity for those domains to assume the same structure. Functional homoplasy is often brought about by different mechanisms. For example, different molecules, heme and a pair of copper ions, are involved in oxygen binding of hemoglobin and hemocyanin, respectively. Even the same catalytic mechanism is derived from different origins. The catalytic triad of serine protease has evolved at least three times, as evidenced by subtilisin, trypsin, and alpha/beta hydrolase fold enzyme (Ollis et al. 1992). It is certain that these enzymes have different origins, because they have different protein folds and different sequence arrangements in the catalytic triad. In particular, the catalytic triad of the alpha/beta hydrolase fold enzyme is a "mirror image" of that of serine protease. Thus, there is often more than one solution to a biochemical problem. In the case of the PBP and repressor, for example, the same ligand specificity would have evolved from different protein folds and/or different binding modes even if those proteins had not shared a homologous domain. In fact, the ligand-binding domains of the two proteins do have the same protein fold, and they probably have the same binding mode as well, because the same residues at the structurally equivalent sites seem to be involved with the binding for the same ligand between the two proteins.

Origin and Evolution of Operon Structures of PBPs

It is reported that PBP is often encoded in an operon not only with repressor but also with permease and ABC protein that cooperate with their partner PBP in transportation of ligand (Tomii and Kanehisa 1998; Itoh et al. 1999). This suggests the following model on the origin and evolution of an operon containing those protein genes:

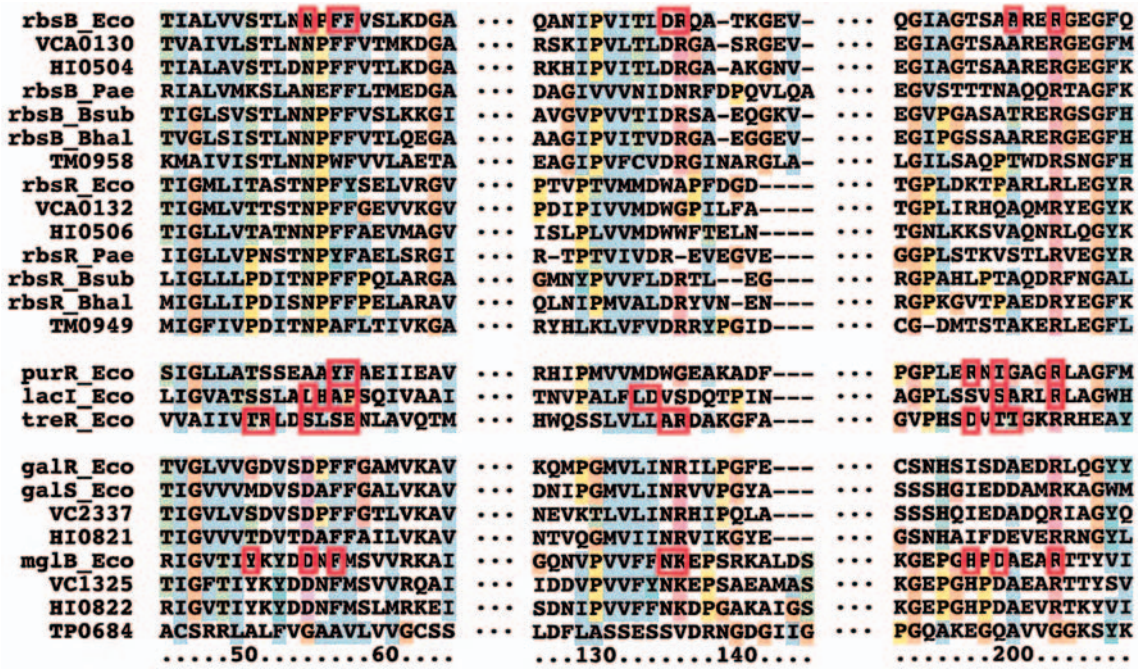

FIG. 4.—Multiple alignments in three regions that include ligand-binding sites. Each region consists of amino acid sequences of D-ribose–binding proteins and repressors (upper part), repressors with known 3D-structure (middle part), D-galactose/D-glucose repressors and binding proteins (lower part). Red boxes indicate ligand-binding sites obtained from the crystal structure.

First there were ancient operons encoding three genes of PBP, permease, and ABC protein that functioned for the same ligand. Then a PBP gene in one of the ancient operons duplicated, and one of those duplicates acquired a DNA-binding domain in the 5′ end. This operon amplified next in genome, and diverged to those with a variety of ligand specificities.

In this model, the order of the four genes in the operon is expected to be the same among the descendants. However, as mentioned in the previous section, the order has been conserved only among the orthologous operons of closely related species, and the gene members are not always the same among the descendants. It is thus considered likely that the operon occasionally rearranged the gene order in itself, translocated a gene outside, or acquired a gene with a related function into itself, while it amplified in the genome and acquired new ligand specificities. Such gene context conservation has also been found in glutamate ABC transporter genes, translation-associated genes, and flagellum-related genes (Lathe, Snel, and Bork 2000). If this is the case, we need to impose the evolutionary constraint that would have kept the functionally related genes in an operon. This imposition should be reasonable, because it is expected to be advantageous for the functionally related genes to be encoded together in a cotranscribed and so coregulated unit.

The results of this study demonstrate that the LacI/GalR and the PBP families can be distinguished by analyzing their overall structure, whereas their ligand specificities are determined mainly by the ligand-binding sites. The sites are a limited number of residues and compose local structure of a protein. Our finding suggests that it is more effective to consider spatial arrangement of functionally important residues than to compare overall similarity when we attempt the empirical prediction of unknown protein function in functional genomics.

Acknowledgments

We thank Dr. Takeshi Kawabata for his help and advice on the use of GTOP. This work was supported in part by a Grant-in-Aid for Science Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Literature Cited

Adachi, J., and M. Hasegawa. 1996. MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood. The Institute of Statistical Mathematics, Tokyo.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Bairoch, A., and R. Apweiler. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28**:45–48.

Bassler, B. L., M. Wright, and M. R. Silverman. 1994. Multiple signalling systems controlling expression of luminescence in *Vibrio harveyi*: sequence and function of genes encoding a second sensory pathway. Mol. Microbiol. **13**:273–286.

Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. Sonnhammer. 2000. The Pfam protein families database. Nucleic Acids Res. **28**:263–266.

Bell, C. E., and M. Lewis. 2000. A closer view of the conformation of the Lac repressor bound to operator. Nat. Struct. Biol. **7**:209–214.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Res. **28**:235–242.

Bjorkman, A. J., R. A. Binnie, H. Zhang, L. B. Cole, M. A. Hermodson, and S. L. Mowbray. 1994. Probing protein-protein interactions. The ribose-binding protein in bacterial transport and chemotaxis. J. Biol. Chem. **269**:30206–30211.

Briscoe, A. D. 2000. Six opsins from the butterfly *Papilio glaucus*: molecular phylogenetic evidence for paralogous origins of red-sensitive visual pigments in insects. J. Mol. Evol. **51**:110–121.

Chaudhuri, B. N., J. Ko, C. Park, T. A. Jones, and S. L. Mowbray. 1999. Structure of D-allose binding protein from *Escherichia coli* bound to D-allose at 1.8 A resolution. J. Mol. Biol. **286**:1519–1531.

Choi, K. Y., and H. Zalkin. 1994. Role of the purine repressor hinge sequence in repressor function. J. Bacteriol. **176**:1767–1772.

Doolittle, R. F. 1994. Convergent evolution: the need to be explicit. Trends Biochem. Sci. **19**:15–18.

Falcon, C. M., and K. S. Matthews. 1999. Glycine insertion in the hinge region of lactose repressor protein alters DNA binding. J. Biol. Chem. **274**:30849–30857.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

Friedman, A. M., T. O. Fischmann, and T. A. Steitz. 1995. Crystal structure of lac repressor core tetramer and its implications for DNA looping. Science **268**:1721–1727.

Fukami-Kobayashi, K., M. Nosaka, A. Nakazawa, and M. Go. 1996. Ancient divergence of long and short isoforms of adenylate kinase: molecular evolution of the nucleoside monophosphate kinase family. FEBS Lett. **385**:214–220.

Fukami-Kobayashi, K., Y. Tateno, and K. Nishikawa. 1999. Domain dislocation: a change of core structure in periplasmic binding proteins in their evolutionary history. J. Mol. Biol. **286**:279–290.

Fukami-Kobayashi, K., S. Tomoda, and M. Go. 1993. Evolutionary clustering and functional similarity of RNA-binding proteins. FEBS Lett. **335**:289–293.

Glasfeld, A., A. N. Koehler, M. A. Schumacher, and R. G. Brennan. 1999. The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions. J. Mol. Biol. **291**:347–361.

Gotoh, O. 1999. Multiple sequence alignment: algorithms and applications. Adv. Biophys. **36**:159–206.

Hars, U., R. Horlacher, W. Boos, W. Welte, and K. Diederichs. 1998. Crystal structure of the effector-binding domain of the trehalose- repressor of *Escherichia coli*, a member of the LacI family, in its complexes with inducer trehalose-6-phosphate and noninducer trehalose. Protein Sci. **7**:2511–2521.

Hasegawa, M., and H. Kishino. 1994. Accuracies of the simple methods for estimating the bootstrap probability of a maximum likelihood tree. Mol. Biol. Evol. **11**:142–145.

Hofmann, K., P. Bucher, L. Falquet, and A. Bairoch. 1999. The PROSITE database, its status in 1999. Nucleic Acids Res. **27**:215–219.

Itoh, T., K. Takemoto, H. Mori, and T. Gojobori. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol. Biol. Evol. **16**:332–346.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.

Kawabata, T., S. Fukuchi, K. Homma, M. Ota, J. Araki, T. Ito, N. Ichiyoshi, and K. Nishikawa. 2002. GTOP: a database of protein structures predicted from genome sequences. Nucleic Acids Res. **30**:294–298.

Kawabata, T., and K. Nishikawa. 2000. Protein structure comparison using the Markov transition model of evolution. Proteins **41**:108–122.

Kitano, T., R. Noda, K. Sumiyama, R. E. Ferrell, and N. Saitou. 2000. Gene diversity of chimpanzee ABO blood group genes elucidated from intron 6 sequences. J. Hered. **91**:211–214.

Kornegay, J. R., J. W. Schilling, and A. C. Wilson. 1994. Molecular adaptation of a leaf-eating bird: stomach lysozyme of the hoatzin. Mol. Biol. Evol. **11**:921–928.

Koyanagi, M., K. Ono, H. Suga, N. Iwabe, and T. Miyata. 1998. Phospholipase C cDNAs from sponge and hydra: antiquity of genes involved in the inositol phospholipid signaling pathway. FEBS Lett. **439**:66–70.

Laskowski, R. A. 2001. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res. **29**:221–222.

Lathe, W. C. III, B. Snel, and P. Bork. 2000. Gene context conservation of a higher order than operons. Trends Biochem. Sci. **25**:474–479.

Lewis, M., G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Lu. 1996. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science **271**:1247–1254.

Mauzy, C. A., and M. A. Hermodson. 1992. Structural homology between rbs repressor and ribose binding protein implies functional similarity. Protein Sci. **1**:843–849.

McDonald, I. K., and J. M. Thornton. 1994. Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. **238**:777–793.

Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**:536–540.

Nagadoi, A., K. Nakazawa, S. Morikawa, H. Nakamura, G. Sampei, K. Mizobuchi, H. Yamamoto, M. A. Schumacher, R. G. Brennan, and Y. Nishimura. 1995. Conformational changes of purine repressor DNA-binding domain upon complexation with DNA. Nucleic Acids Symp. Ser. **34**:63–64.

Nguyen, C. C., and M. H. Saier, Jr. 1995. Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. FEBS Lett. **377**:98–102.

O'hUigin, C., A. Sato, and J. Klein. 1997. Evidence for convergent evolution of A and B blood group antigens in primates. Hum. Genet. **101**:141–148.

Ollis, D. L., E. Cheah, M. Cygler et al. (10 co-authors). 1992. The alpha/beta hydrolase fold. Protein Eng. **5**:197–211.

Pace, H. C., M. A. Kercher, P. Lu, P. Markiewicz, J. H. Miller, G. Chang, and M. Lewis. 1997. Lac repressor genetic map in real space. Trends Biochem. Sci. **22**:334–339.

Schumacher, M. A., K. Y. Choi, H. Zalkin, and R. G. Brennan. 1994*a*. Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. Science **266**:763–770.

———. 1994*b*. Crystallization and preliminary X-ray analysis of an Escherichia coli purine repressor-hypoxanthine-DNA complex. J. Mol. Biol. **242**:302–305.

Stewart, C. B., J. W. Schilling, and A. C. Wilson. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature **330**:401–404.

Sumiyama, K., T. Kitano, R. Noda, R. E. Ferrell, and N. Saitou. 2000. Gene diversity of chimpanzee ABO blood group genes elucidated from exon 7 sequences. Gene **259**:75–79.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Tomii, K., and M. Kanehisa. 1998. A comparative analysis of ABC transporters in complete microbial genomes. Genome Res. **8**:1048–1059.

Vartak, N. B., J. Reizer, A. Reizer, J. T. Gripp, E. A. Groisman, L. F. Wu, J. M. Tomich, and M. H. Saier, Jr. 1991. Sequence and evolution of the FruR protein of *Salmonella typhimurium*: a pleiotropic transcriptional regulatory protein possessing both activator and repressor functions which is homologous to the periplasmic ribose-binding protein. Res. Microbiol. **142**:951–963.

Vermersch, P. S., D. D. Lemon, J. J. Tesmer, and F. A. Quiocho. 1991. Sugar-binding and crystallographic studies of an arabinose-binding protein mutant (Met108Leu) that exhibits enhanced affinity and altered specificity. Biochemistry **30**:6861–6866.

Vyas, N. K., M. N. Vyas, and F. A. Quiocho. 1988. Sugar and signal-transducer binding sites of the *Escherichia coli* galactose chemoreceptor protein. Science **242**:1290–1295.

Weickert, M. J., and S. Adhya. 1992. A family of bacterial regulators homologous to Gal and Lac repressors. J. Biol. Chem. **267**:15869–15874.

Yokoyama, R., and S. Yokoyama. 1990. Convergent evolution of the red- and green-like visual pigment genes in fish, *Astyanax fasciatus*, and human. Proc. Natl. Acad. Sci. USA **87**:9315–9318.

Naruya Saitou, Associate Editor

Accepted October 11, 2002