THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 2002 by The American Society for Biochemistry and Molecular Biology, Inc.

Vol. 277, No. 15, Issue of April 12, pp. 12507–12515, 2002
Printed in U.S.A.

Subdivision of the Helix-Turn-Helix GntR Family of Bacterial Regulators in the FadR, HutC, MocR, and YtrA Subfamilies*

Received for publication, November 15, 2001, and in revised form, December 20, 2001
Published, JBC Papers in Press, December 27, 2001, DOI 10.1074/jbc.M110968200

Sébastien Rigali‡, Adeline Derouaux, Fabrizio Giannotta, and Jean Dusart

From the Centre d'Ingénierie des Protéines, Université de Liège, Institut de Chimie B6, Sart-Tilman, B-4000, Liège, Belgium

Haydon and Guest (Haydon, D. J, and Guest, J. R. (1991) *FEMS Microbiol. Lett.* 63, 291–295) first described the helix-turn-helix GntR family of bacterial regulators. They presented them as transcription factors sharing a similar N-terminal DNA-binding (D-b) domain, but they observed near-maximal divergence in the C-terminal effector-binding and oligomerization (E-b/O) domain. To elucidate this C-terminal heterogeneity, structural, phylogenetic, and functional analyses were performed on a family that now comprises about 270 members. Our comparative study first focused on the C-terminal E-b/O domains and next on DNA-binding domains and palindromic operator sequences, has classified the GntR members into four subfamilies that we called FadR, HutC, MocR, and YtrA. Among these subfamilies a degree of similarity of about 55% was observed throughout the entire sequence. Structure/function associations were highlighted although they were not absolutely stringent. The consensus sequences deduced for the DNA-binding domain were slightly different for each subfamily, suggesting that fusion between the D-b and E-b/O domains have occurred separately, with each subfamily having its own D-b domain ancestor. Moreover, the compilation of the known or predicted palindromic *cis*-acting elements has highlighted different operator sequences according to our subfamily subdivision. The observed C-terminal E-b/O domain heterogeneity was therefore reflected on the DNA-binding domain and on the *cis*-acting elements, suggesting the existence of a tight link between the three regions involved in the regulating process.

Among transcription factors, several groups have been identified according to their conserved motifs and their modes of DNA binding such as helix-turn-helix, zinc-fingers, leucine-zipper, homeodomain, and β-sheet DNA-binding proteins (2, 3). The most studied and best characterized is the HTH[1] group (1, 4–8) in which the conserved DNA recognition motif consists of an α-helix, a turn, and a second α-helix, often called the "recognition" helix as it is the part of the HTH motif that fits into the DNA major groove. Generally, HTH proteins bind as dimers, 2-fold symmetric DNA sequences in which each monomer recognizes a half-site. This group is now considered as a reference for understanding the general rules that govern protein-DNA interactions (9, 10) and has also become a favorite target for evolutionary studies (8, 11).

Among HTH transcriptional regulators, families have been identified throughout sequence comparisons and phylogenetic, structural, and functional analyses focused on DNA-binding domains and almost exclusively on the HTH structure, which is the only active motif that shows strong similarities among all members of the group (1, 4, 6–8, 11). These comparative studies have led to the determination of a specific HTH consensus pattern or signature for each family, providing the basis for a simple method of classification and detection of new members (12).

The lack of significant similarity among regions involved in effector binding or oligomerization systematically excludes these domains during families signature establishment, although they have important roles in the regulating process. In fact, it is often the oligomerization between regulatory subunits and/or the conformational changes due to the binding or the removal of the inducing/repressing molecule that allows correct HTH motif disposition and the subsequent DNA binding ability of the whole regulatory protein. The link between the two regions is therefore more intimate than it first appears from a unique amino acids comparison and may also be reflected in the DNA operator sequences, the third structural element involved in gene regulation.

To argue for the existence of a link between regions involved in the regulating process, we analyzed the HTH GntR family of bacterial regulators. As determined thus far, the family comprises about 270 members distributed among the most diverse bacterial groups and regulating the most various biological processes. This family was first described by Haydon and Guest in 1991 (1) and was named after GntR, the repressor of the gluconate operon in *Bacillus subtilis* (13, 14). Our interest in the properties of these bacterial regulators arises from the identification by our laboratory of the *xlnR* gene (15) in which chromosomal disruption in *Streptomyces lividans* relieves various extracellular enzymatic systems from glucose repression.

The first purpose of this report is to present, 10 years after the first comparative study, an update of the GntR family description. Moreover, we decided to analyze the full-length sequence of the proteins through amino acid comparisons, secondary structure predictions, phylogenetic tree construction, and functional analysis in order to find hidden specific characteristics among the regions that are generally not considered. Analyses that extended to the regions outside of the DNA-binding domain could lead to a more precise family signature and should define the subfamilies.

* This work was supported by the "Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture" (FRIA, Brussels, Belgium). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed. Tel.: +32-4-366-33-77; Fax: +32-4-366-33-64; E-mail: srigali@student.ulg.ac.be.

[1] The abbreviations used are: HTH, helix-turn-helix; E-b/o domain, effector-binding and oligomerization domain; D-b domain, DNA-binding domain; PLP, pyridoxal 5′-phosphate; FadR, fatty acid-responsive regulator in *Escherichia coli*.

TABLE I
List of the HTH GntR-like regulators presented in our comparative study

Protein (Reference)	Function	Organism *(abbreviation)*	length in a.a.	Swiss-Prot number	Sub-family
AnsR (17)	L-asparagine permease operon repressor	*Rhizobium etli (Ret)*	239	Q9RFN7	FadR
AphS (18)	Regulator of genes involved in phenol utilization	*Comamonas testosteroni* TA441 *(Cte)*	239	Q9RHW8	
BphS (19)	Repressor of genes involved in biphenyl degradation	*Alcaligenes eutrophus (Aeu)*	243	Q9EV74	
DgoR	Putative galactonate operon repressor	*Escherichia coli (Eco)*	229	P31460	
EmoR	Unknown function	*EDTA-degrading bacterium* BNC1 *(Edb)*	207	Q9F9T1	
EsmR (20)	Epidemic strain marker regulator	*Burkholderia cepacia (Bce)*	277	P96570	
ExuR (21)	Hexuronate regulon repressor	*Escherichia coli*	258	P42608	
FadR (22-24)	Fatty acid metabolism regulator	*Escherichia coli*	238	P09371	
GlcC (25-27)	Glycolate oxidation operon activator	*Escherichia coli*	254	P52072	
GlcC	Putative glycolate oxidation operon regulatoror	*Pseudomonas aeruginosa (Pae)*	251	Q9HTK2	
GntR (13,14)	Gluconate operon repressor	*Bacillus subtilis (Bsu)*	243	P10585	
LldR (27,28)	Putative L-lactate dehydrogenase operon repressor	*Escherichia coli*	258	P33233	
LuxZ (29)	Lux operon enhancer	*Photobacterium leiognathi (Ple)*	221	Q9ZAP4	
MatR (30)	Activator for malonate metabolism	*Rhizobium leguminosarum (Rle)*	222	Q9JP74	
MdcY (31)	Malonate decarboxylase operon repressor	*Acinetobacter calcoaceticus (Aca)*	224	Q9F0Q8	
NtaR (32)	Putative nitrilotriacetate monooxygenase subunits regulator	*Chelatobacter heintzii (Che)*	210	P54988	
PdhR (33)	Pyruvate dehydrogenase complex repressor	*Escherichia coli*	254	P06957	
PipR (34)	Cytochrome P450 regulator	*Mycobacterium smegmatis* mc2155 *(Msm)*	245	Q9XDB1	
SCF55.06	Unknown function	*Streptomyces coelicolor*	253	Q9RJQ8	
SC6D7.29	Unknown function	*Streptomyces coelicolor*	231	Q9RKW9	
UxuR (35)	Glucuronic acid (GlcUA) gene cluster regulator	*Bacillus stearothermophilus* T-6 *(Bst)*	249	Q9ZFL9	
VanR (36)	Vanillate demethylase synthesis repressor	*Acinetobacter* sp. ADP1 *(Asp)*	251	O24839	
WhiH (37)	Sporulation transcription factor	*Streptomyces coelicolor (Sco)*	295	O50536	
FarR (38,39)	Fatty acyl responsive regulator	*Escherichia coli*	240	P13669	HutC
HutC (40)	Histidine utilization repressor	*Pseudomonas putida (Pp)*	248	P22773	
KorA (41)	pIJ701 kil-kor repressor	*Streptomyces lividans (Sli)*	241	P22405	
KorSA (42)	pSAM2 kil-kor repressor	*Streptomyces ambofaciens (Sam)*	259	Q07191	
PhnF (43)	Putative alkylphosphate uptake regulator	*Escherichia coli*	241	P16684	
PhnR	Unknown function	*Salmonella typhimurium (Sty)*	239	P96061	
SCD39.28	Unknown function	*Streptomyces coelicolor*	269	Q9F2T4	
SCE39.19	Unknown function	*Streptomyces coelicolor*	251	Q9X8E2	
SC7E4.28	Unknown function	*Streptomyces coelicolor*	254	Q9K492	
TraR (44)	pJV1 TraA operon repressor	*Streptomyces phaeochromogenes (Sph)*	245	Q54677	
TreR (45)	Trehalose operon repressor	*Bacillus subtilis*	238	P39796	
XlnR (15)	Regulator involved in catabolite repression	*Streptomyces lividans*	252	Q9ACN8	
YvoA	Unknown function	*Bacillus subtilis*	243	O34817	
MocR (46)	Probable rhizopine catabolism regulator	*Rhizobium meliloti (Rme)*	493	P49309	MocR
PdxR (47)	Pyridoxal phosphate synthesis regulator	*Streptomyces venezuelae (Sve)*	532	Q9FDB4	
PtsJ (48)	Putative phosphotransferase system regulator	*Salmonella typhimurium*	430	P40193	
YcxD	Putative surfactin operon regulator	*Bacillus subtilis*	444	Q08792	
YcnF	Unknown function	*Bacillus subtilis*	479	P94426	
YdfD	Unknown function	*Bacillus subtilis*	482	P96681	
YhdI	Unknown function	*Bacillus subtilis*	469	O07578	
YjiR	Unknown function	*Escherichia coli*	470	P39389	
YrdX	Unknown function	*Rhodobacter sphaeroides (Rsp)*	456	Q01856	
BH0651	Unknown function	*Bacillus halodurans (Bha)*	123	Q9KF35	YtrA
BH2647	Unknown function	*Bacillus halodurans*	123	Q9K9J9	
SA1748	Unknown function	*Staphylococcus aureus subsp. aureus* N315 *(Sau)*	126	Q99SV4	
SCF43A.13	Unknown function	*Streptomyces coelicolor*	146	Q9XAA2	
SCGD3.13	Unknown function	*Streptomyces coelicolor*	115	Q9XA65	
TA0736	Unknown function	*Thermoplasma acidophilum (Tac)*	103	Q9HK68	
TM0766	Unknown function	*Thermotoga maritima (Tma)*	121	Q9WZM5	
YhcF	Unknown function	*Bacillus subtilis*	121	P54590	
YtrA (49)	Acetoine utilization gene cluster repressor	*Bacillus subtilis*	130	O34712	
AraR (50)	Transcriptional repressor of the arabinose operon	*Bacillus subtilis*	384	P96711	LacI
FucR (51)	Repressor of L-fucose utilization gene cluster	*Bacteroides thetaiotaomicron (Bth)*	331	Q9RQ14	-

EXPERIMENTAL PROCEDURES

Selection of GntR-like Members—Members of the GntR family were identified from the SWISS-PROT/TrEMBL/GenBank^{TM} sequence data bases (last update, June 2001) by a keywords search on the ExPASy molecular Biology server and NCBI server.[2] All sequences proposed by the data bases as belonging to the GntR family were used as query sequences for a BLAST search to verify their N-terminal DNA-binding domain homology to other GntR-like regulators. Incorrectly GntR-like classified proteins by sequence data bases, *i.e.* the Irr protein from *Bradyrhizobium japonicum* (16), were rejected from our comparative study. Fragment of sequences were rejected too. We finally collected and analyzed about 270 members. For ease and usefulness of presentation, the best studied regulators (13–15, 17–51), most representative members, or proteins yielding data of specific interest were selected for publication. The 56 proteins discussed and presented in this paper are listed in Table I.

Secondary Structure Predictions—To identify homologous C-terminal sequences within the HTH GntR family, we started our comparative study from the level of the secondary structures, in which conservation is known to be less eroded during evolution. Secondary structure predictions result from the compilation of PSI-pred, Predict Protein, Sspro, and Jpred automated prediction programs on the PredictProtein server.[3] To improve the validity of our consensus prediction approach, we compared the theoretical model that we obtained for FadR (fatty acid-responsive regulator in *Escherichia coli*) to its experimentally resolved tertiary structure (52, 53). The method was revealed to have an accuracy of >90% for FadR with most of the inaccuracies occurring at the

[2] Found on the Web at www.expasy.ch and www.ncbi.nlm.nih.gov, respectively.

[3] Found on the Web at dodo.cpmc.columbia.edu.

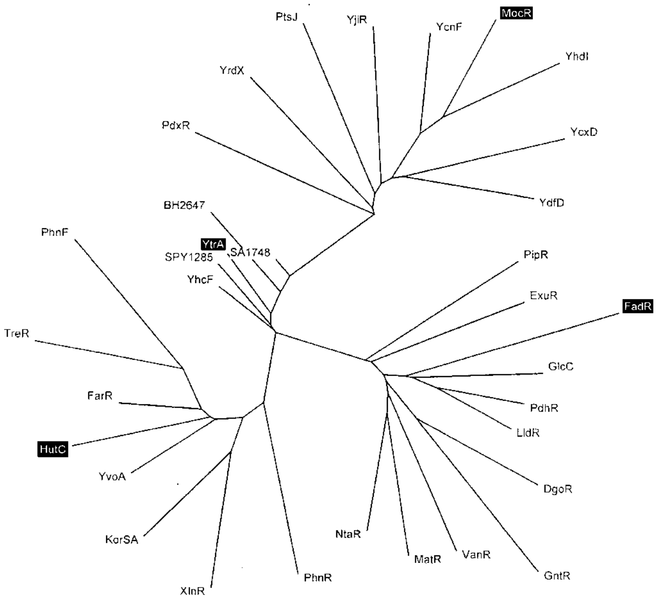

FIG. 1. **Unrooted tree of the proteins of the GntR family.** The abbreviations are as indicated in Table I. GntR-like regulators were classified in four subfamilies according to the four clusters of branches that emerged from the constructed tree and reflecting the observed C-terminal structural topology.

boundaries of the secondary structure elements.

Multiple Alignments and Phylogenetic Tree Construction—Multiple alignments were developed with the MULTIALIN (54) and CLUSTALW (55)[4] programs, included in the ExPASy multiple alignment tool, followed by manual improvement by eye according to the predicted secondary structures. The advantage of these alignments resides in the integration of the structural reality of the proteins. Distances between aligned proteins were computed with the PRODIST program using maximum likelihood estimates on the Dayhoff PAM matrix (56). The FITCH program estimated phylogenies from distances in the matrix data using the Fitch-Margoliash algorithm (57), and phylogenetic trees were drawn using the TREEVIEW program (58). PRODIST and FITCH programs are included in the PHYLIP package developed by Feldenstein (59).

RESULTS

As mentioned by Haydon and Guest (1), members of the GntR family of bacterial regulators share similar N-terminal DNA-binding domains, but high heterogeneity has been observed among the various C-terminal effector-binding and oligomerization domains. In order to elucidate the C-terminal dissimilarity, the characterization of the N- and C-terminal domains was done separately.

The C-terminal Effector-binding and/or Oligomerization Domain—The construction of a phylogenetic tree deduced from the full-length multiple alignment of GntR-like members revealed that the C-terminal heterogeneity was limited to four E-b/O types. In fact, we can see in Fig. 1 four major and distinct clusters of branches. By the same way, two-dimensional structural predictions revealed four major types of E-b/O structural domain topologies (Fig. 2, *a*–*d*) with discrete variants in each subfamily and very few proteins (7%) escaping from this subdivision. The presence of four major types of C-terminal topologies suggests at least four different E-b/O domain donor-ancestors for the fusion to a common type of DNA-binding domain. Once the fusion occurred between the two domains, the high similarity level (55%) calculated suggests that proteins within a subfamily arose by duplication events.

The first GntR subfamily, which we called FadR, is the most represented one as it regroups 40% of GntR-like regulators. In this subfamily, the proteins consist of an all-helical C-terminal domain (Fig. 2*a*) with seven or six α-helices for the FadR and VanR subgroups, respectively. VanR-like regulators certainly derive from FadR-like proteins, as they only diverge by the loss of the first α-helix (α_4). The average C-terminal length of the FadR and VanR subgroups is, respectively, about 170 and 150 amino acids. The crystal structure of the C-terminal domain of FadR (Protein Data Bank code 1EX2) has been determined (52, 53) and, according to our comparative study, its relative three-dimensional data could be used as a scaffold to orient studies to the entire subfamily. Most of the FadR-like proteins are involved in the regulation of oxidized substrates related to amino acids metabolism or at the crossroads of various metabolic pathways such as aspartate (AnsR), pyruvate (PdhR), glycolate (GlcC), galactonate (DgoR), lactate (LldR), malonate (MatR), or gluconate (GntR).

In the second proposed subfamily, the C-terminal domain contains both α-helical and β-sheet structures arranged as shown in Fig. 2*b*. The subfamily is named HutC and comprises 31% of GntR-like regulators among which the cluster of proteins involved in conjugative plasmid transfer in various *Streptomyces* species (*i.e.* KorSA, KorA, and TraR proteins). The average length of the C-terminal domain is about 170 amino acids and, so far, no three-dimensional structural data on it are available. In this subfamily, the conservation of the structural elements has been altered at several positions (see for instance, $\beta_{3,}$ α_7, and β_6 in Fig. 2*b*). The observed altered E-b/O topology could be the result of structural accommodation in response to the most diverse biological processes regulated by HutC-like members.

In the third subfamily, called MocR, the E-b/O domain is immediately distinguishable from others because of its exceptional average length of about 350 amino acids and its homology to the class I of aminotransferase proteins (61) (see Fig. 2*d*). These proteins catalyze the reversible transfer of an amino group from the amino acid substrate to an acceptor α-keto acid. They require pyridoxal 5′-phosphate (PLP) as a cofactor to catalyze this reaction. Transamination reactions are of central importance in amino acid metabolism and in links to carbohydrate and fat metabolism. This class of aminotransferases acts as dimers in a head-to-tail configuration (62). Each subunit binds one molecule of PLP through an aldimide linkage with the ϵ-amino group of the conserved lysine residue in the PLP attachment site. The observed modular association to an aminotransferase-like C-terminal domain suggests that similar dimerization should occur in MocR-like proteins and that PLP is required as a cofactor for their regulating activity. The most relevant evidence comes from PdxR in *Streptomyces venezuelae*, which is involved directly in the regulation of pyridoxal phosphate synthesis (47).

The fourth subfamily possesses a reduced C-terminal domain with only two α-helices (Fig. 2*c*). The subfamily, that we called YtrA, is the less represented with only 6% of GntR-like regulators, most of these forming part of operons involved in ATP-binding cassette (ABC) transport systems. As it emerges from the alignment of YtrA-like proteins (Fig. 2*c*), the weaker identity observed between members suggest that the C-terminal domain has undergone some molecular recombinations or that the origins of the E-b/O domain could be multiple. The average length of the putative E-b/O domain is about 50 amino acids, and according to Yoshida *et al.* (49), this length should be too small to accommodate effector binding. Dimerization should remain possible, as numerous GntR-like palindromic operator sequences have been observed in the corresponding upstream regions (see "Operator Site Analysis" below). The presence of

[4] Found on the Web at protein.toulouse.inra.fr/multialin and npsapbil.ib.cp.fr, respectively.

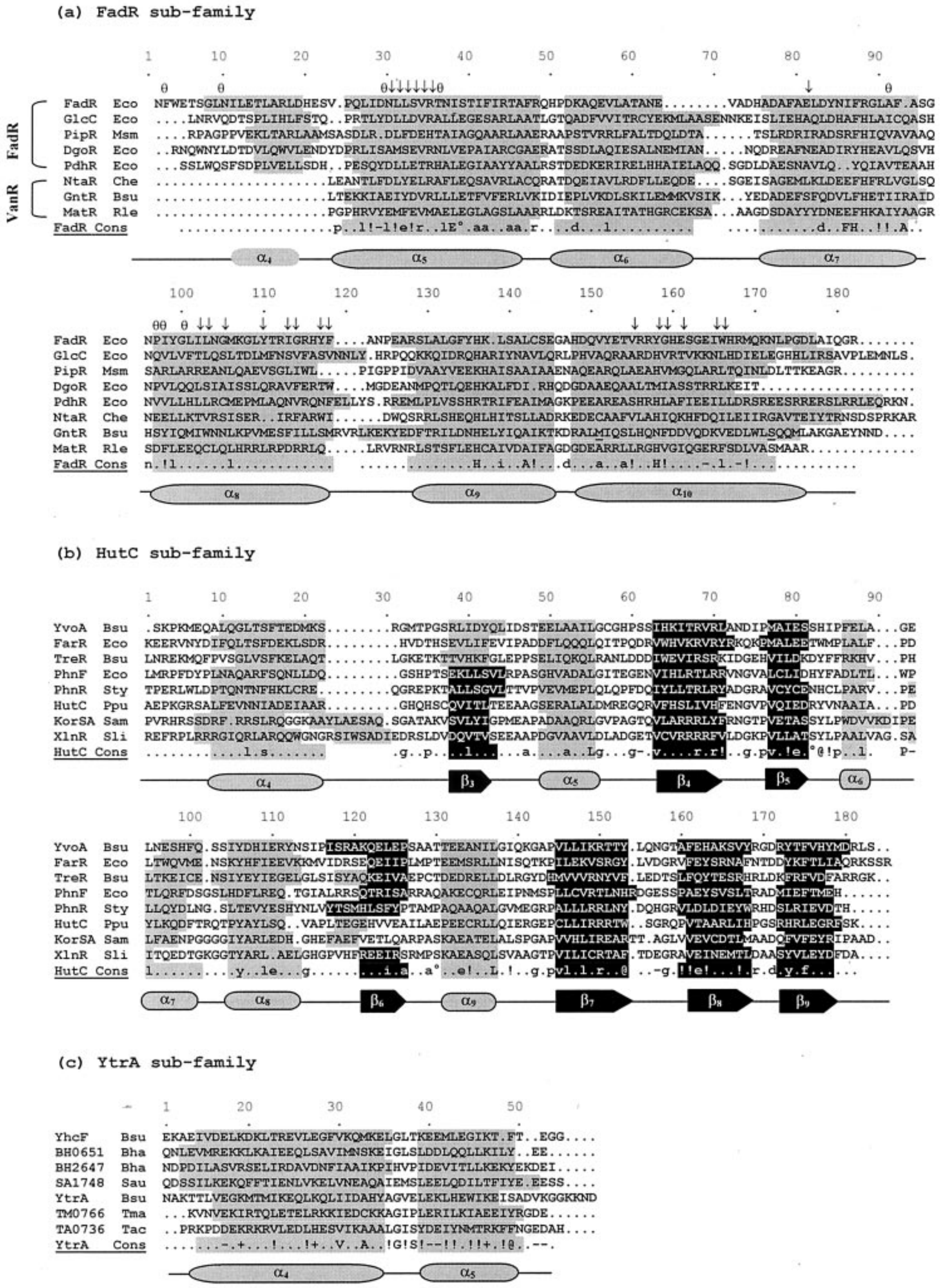

FIG. 2. **Structure-based sequence alignment of the C-terminal domains of proteins of the GntR family.** Abbreviations are as indicated in Table I. Consensus sequences result from the multiple alignment of all GntR-like members and not only those listed in Table I. The high and low consensus levels were fixed arbitrarily at 80 and 40% of identity and are represented, respectively, by *capital* and *lowercase letters*. The similarity level was fixed at 80%. Symbols for conserved amino acid properties are as follows: !, conserved hydrophobic residues (ILVAMFYW); @, aromatic residues (FYW); −, negatively charged residues (ED); +, positively charged residues (RKH); ○, small residues (GSATPN). ↓ and θ indicate, in *panel a*, residues implicated in effector binding and dimerization of the FadR protein (52, 53). Also in *panel a*, the *underlined* residue indicates mutations that affect gluconate binding ability in GntR (60). In *panel d*, the *underlined* residue in the consensus corresponds to the lysine that established the covalent link with pyridoxal phosphate in aminotransferases. *Spaces* in consensus sequences denote insertions within the alignment.

many positively or negatively charged as well as hydrophobic and aromatic residues at the end of the domain suggests that dimer formation should occur through classical salt bridges and side-chain-side-chain hydrophobic interactions.

The DNA-binding Domain—As shown in Fig. 3, structural predictions revealed that the DNA-binding (D-b) domain topology of the whole GntR family is rather well conserved and all of the secondary structure elements are in similar relative positions. It consists of three α-helices and two (sometimes three) β-sheets disposed as follow: $\alpha_1\alpha_2\alpha_3\beta_1\beta_2$. According to FadR structural data, we can consider that the N-terminal DNA-binding domain of all GntR-like members contains a small β-sheet core and three α-helices, the HTH motif being formed by helices α_2 and α_3.

The average amino acids identity obtained for the DNA-binding domain of the entire GntR-family is about 25%. The level obtained is relatively low compared, for instance, with the LacI/GalR HTH family (45%). Thus, evidences of a common DNA-binding domain ancestor for the whole GntR family are highlighted by the conserved structural topology rather than

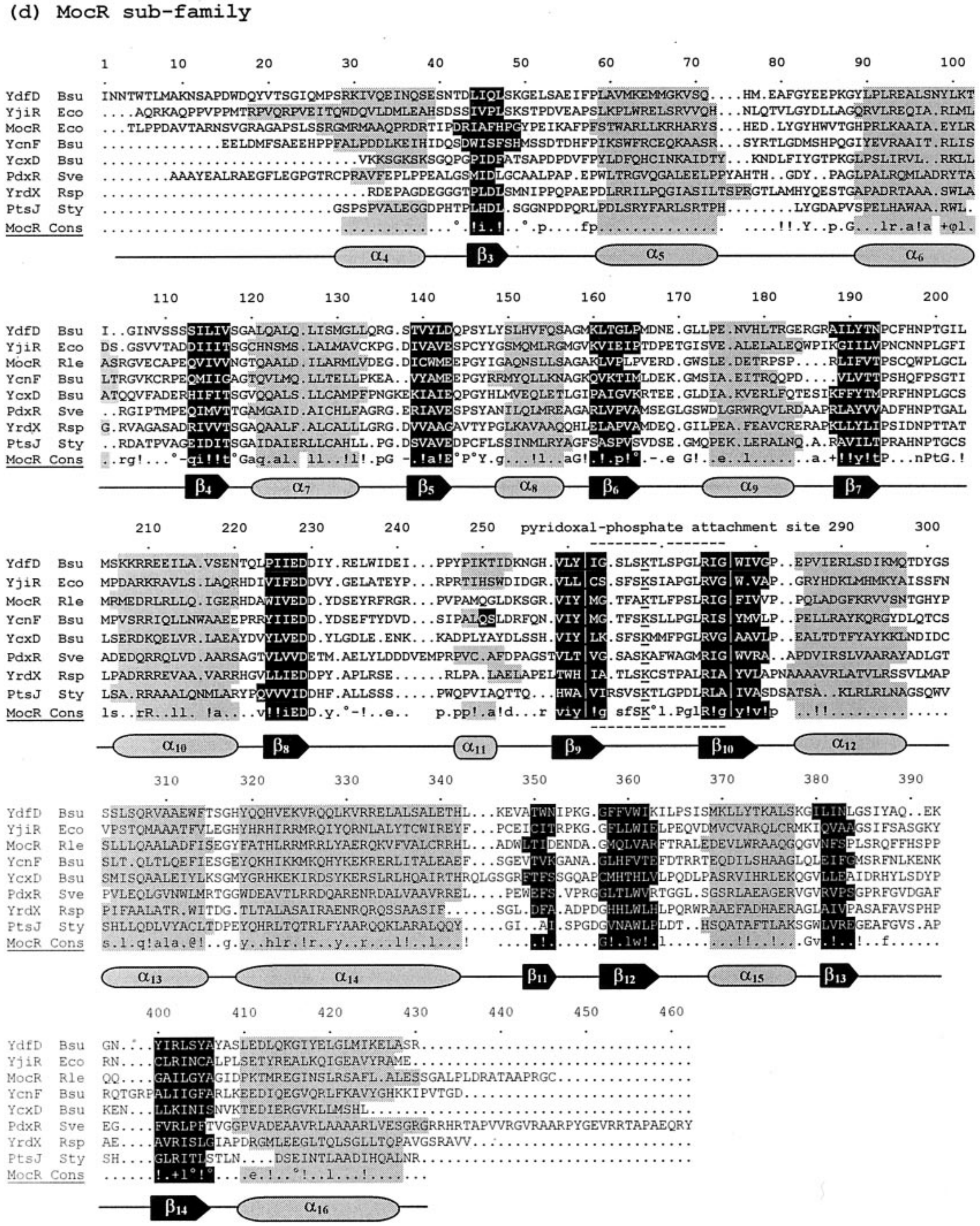

FIG. 2—*continued*

by amino acids conservation. When subfamilies are analyzed separately, the levels of identity and similarity rise to 40 and 60%, respectively. Therefore, the C-terminal structural subdivision is reflected on the DNA-binding domain and on the HTH motif itself. In fact, significantly different HTH consensus sequences have been obtained for each subfamily (Fig. 3) except between MocR and YtrA, where the differences are very weak. The fusion between the D-b domain and the E-b/O domain should have occurred separately for the FadR, HutC, and MocR/YtrA subfamilies, and none of the four subfamilies has emerged from one of the three others by internal molecular rearrangements. The high level of similarity observed between the D-b domains of the MocR and YtrA subfamilies also appears in the phylogenetic tree obtained from full-length multiple alignment (Fig. 1). In fact, the two clusters arise from a common branch, highlighting a conserved amino acids composition in their N-terminal region. One of these two subfamilies could have emerged from the other through C-terminal domain replacement.

Only a few "anomalies" have been found in the two-dimensional N-terminal structural consensus ($\alpha_1\alpha_2\alpha_3\beta_1\beta_2$). The most frequent anomalies were the lack of the first α-helix (α_1) (NtaR from *Chelatobacter heintzii* and EmoR from the EDTA-degrading bacterium, BNC1) or the presence of an additional helix upstream of α_1 (*i.e.* WhiH from *Streptomyces aureofaciens* or PdxR from *S. venezuelae*). We have also noticed that among YtrA regulators, a third, additional β-sheet is frequently predicted before α_1.

Operator Sites Analysis—Although there is no precise "recognition code" involving a one-to-one correspondence between

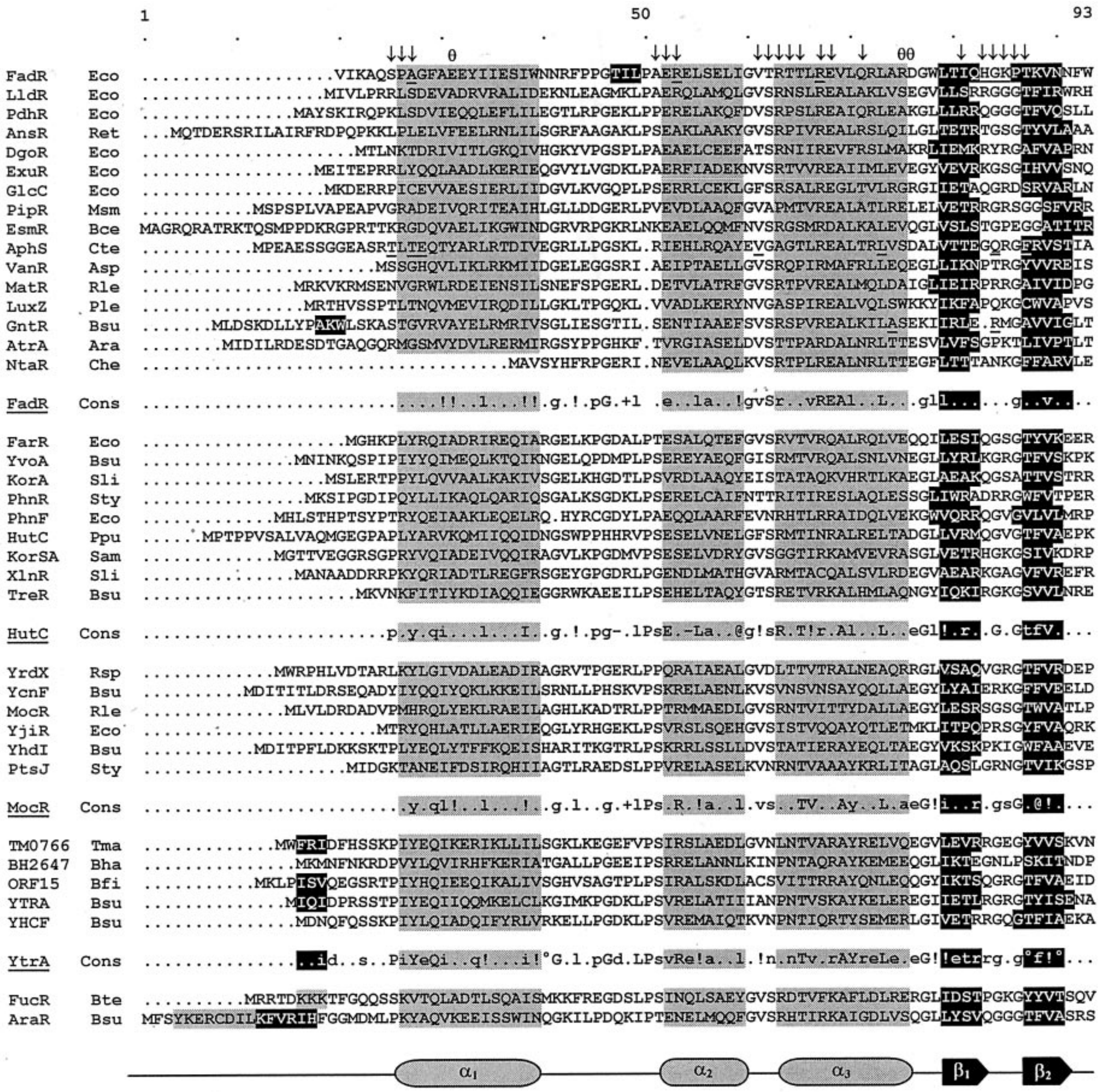

FIG. 3. **Structure-based sequence alignment of the N-terminal DNA-binding domain of proteins of the GntR family.** Abbreviations are as indicated in Table I. Consensus sequences result from the multiple alignment of all GntR-like members and not only those listed in Table I. The high and low consensus levels were fixed arbitrarily at 80 and 40% of identity and are represented, respectively, by *capital* and *lowercase letters*. The similarity level was fixed at 80%. Symbols for conserved amino acid properties are as follows: !, conserved hydrophobic residues (ILVAMFYW); @, aromatic residues (FYW); −, negatively charged residues (ED); +, positively charged residues (RKH); ○, small residues (GSATPN). ↓ and θ indicate, in FadR, residues implicated in DNA binding and dimerization (52, 53). The mutation of the *underlined* residues affects the DNA binding ability of AphS (17), FadR (63), and GntR (64). *Spaces* in the consensus sequences denote insertions within the alignment.

amino acid side chains and the base pairs in the DNA (9), it is logical to suppose that highly conserved DNA-binding motifs may bind similar operator sequences. The known or putative inverted repeat operator sites recognized by some GntR-like proteins are compiled in Table II according to our previous C-terminal classification. Looking at the entire family, we observed that almost all bound sites are organized around a constant palindromic 5′-$(N)_y$GT$(N)_x$AC$(N)_y$-3′ sequence. The most important divergence among the various operator sites resides in the number (y) and the nature (N) of the nucleotides that surround the above consensus sequence. Therefore, as observed by Weickert and Adhya (6) for the LacI/GalR family, the center of the palindrome seems to be highly conserved, whereas the peripheral regions diverge. The similar structural environment that resides at the center of the operator is generally considered the molecule-attracting region for these regulators, whereas the peripheral zones perform the operator discrimination role.

The other relevant divergence between operators resides between the 5′-GT and 3′-AC conserved base pairs. In fact, although there are almost exclusively A and T residues, their number (x) and disposition seems to differ from a subfamily to another. In the FadR and HutC subfamilies we deduced as the consensus 5′-t.GTa.tAC.a-3′ and 5′-GT.ta.AC-3′, respectively. Moreover, the distance between the half-sites is known to be of maximal importance for a correct operator site presentation on the DNA surface according to the flexibility of the linker between the DNA-binding and the E-b/O domains (72–75). This distance varies weakly among the FadR and HutC subfamilies, although it fluctuates widely among the YtrA-like regulators. In this last subfamily, the conserved 5′-GT and 3′-AC residues are found sometimes far from the center of the palindrome. This larger variation among YtrA operators could be attributed to the low complexity of their C-terminal domains, which, added to weaker amino acid conservation, results in a mode of dimer formation specific for each member of the subfamily.

TABLE II

Comparison of known and predicted palindromic operator sites of GntR-like bacterial

For function, bacterial strain, and accession numbers related to the protein abbreviations, see Table I. $^{p, k, \text{and } c}$Putative, known, and consensus sequences, respectively. [1]GlcC from *Pseudomonas aeruginosa*; [2]Half-site of a directed repeat. Mismatched bases are not highlighted and are shown in lowercase letters. TreR$_{01}$ means operator number one of the TreR protein.

Sub-family	Regulator	Known or putative operator site sequence	Reference
FadR	AnsR	gTtGCTG....aCAGCtAt	(17)^{p}
	DgoR	TTGTAcTACAA	this work^{p}
	ExuR	AAATTGGTA.TACCAATTT	(65)$^{k;c}$
	FadR	ATCTGGTacgACCAGAT	(23)^{k}
	GlcC[1]	ATCTGGTagcACCAGAT	this work$^{p;c}$
	GntR	ATACTTGTA.TACAAGTAT	(1)^{k}
	LldR	AATTGGccctACCAATT	(66)^{p}
	LuxZ	CTGTAaAGTt.cACTgTACAG	this work^{p}
	MatR	TCtTGTA.TACAcGA	(30)^{k}
	MdcY	ATTGTA.TACAAT	(67)^{k}
	NtaR	CGGTGGT...ACCACCG	this work^{p}
	PdhR	AATTGGTaagACCAATT	(67)^{k}
	SCF55.6	AAATTGGT.c.ACCAATTT	this work^{p}
	SC6D7.29	GGGATCGTTgAACGATCCC	this work^{p}
	UxuR	CTAGTA.TACTAG	(35)^{p}
	Consensus	...t.GT...AC.a...	
HutC	FarR[2]	TGTATTGTA.T	(39)^{k}
	HutC	AtgCTTGTa.T.A.gACAAGtAT	(68)^{k}
	KorSA	TCACTCATGT.......ACATGAGTGA	(69)$^{k;c}$
	PhnR	TtTGGTc.T.A.tACCAgA	this work^{p}
	TreR o$_1$	CCTGTA.T.A.TACAGG	(70,71)^{k}
	TreR o$_2$	AAgTTGTA.T.A.TACAAgTT	
	SCD39.28	ACAGTCCT.AGGACTGT	this work^{p}
	SCE39.19	AAgCTAGT..T.A..ACTAGgTT	this work^{p}
	SC7E4.28	AcTGGTc.T.A.cACCAtT	this work^{p}
	SC4G1.22	TGGTc.T.A.aACCA	this work^{p}
	Consensus	.....GT. T A .AC.....	
YtrA	BH0651	TATATAtaGT...ATA.......TAT...ACatTATATA	this work^{p}
	SA1748	TcTGT...ATA.......TAT...ACAaA	this work^{p}
	SCF43A.13	CACGTCCAGT....c....ACTGGACGTG	this work^{p}
	SCGD3.13	CATGGTG..aTAgTttcAtTAg..CACCATG	this work^{p}
	TA0736	TGTTCTATA..aga..TATAGAACA	this work^{p}
	TM0766	TGT..AATAtTA.TAcTATT..ACA	this work^{p}
	YtrA	TtaAGTGTA..cTAaTTgAAgTAa..TACACTatA	(49)^{k}
	Consensus	...GT. .TA ... TA. .AC...	

So far, no *cis*-acting elements have been determined experimentally for the actual studied regulators of the MocR subfamily (PtsJ, PdxR, and MocR), preventing us from determining homologous putative sequences in their promoter regions. This subfamily presents another problem; most of these proteins are of unknown function, and therefore most of the regions upstream of the regulated genes are not available. A comparative study of the upstream regions of MocR-like genes did not revealed any palindromic sequence common to the whole subfamily, and very few MocR-like proteins presented weakly similar putative GntR-like operator. These results suggest either that there is another type of *cis*-acting element specific to the MocR-like regulators or that autoregulation is not widespread among them. To have an idea of the topology of *cis*-acting elements typical of the MocR subfamily, interesting data should come from crystallographic studies of the class I aminotransferases. In fact, as highlighted for the tyrosine aminotransferase (TyrB; Swiss-Prot accession no. P04693, Protein Data Bank code 3TAT) from *E. coli* (61), these proteins present a head-to-tail type of dimerization. As shown in Fig. 4, the head-to-tail configuration is not adapted to inverted repeats but is more appropriate to binding directed repeats that are sufficiently spaced to form DNA looping. Therefore, the lack of typical GntR-like operator sequences in the promoter regions of MocR-like regulators could be attributed to how these proteins should form dimers.

The deduced consensus operator sequences presented in Table II can be used as rapid operator site predicting tools. We tried to detect some of these on *Streptomyces coelicolor* genome to highlight genes in which expression could be regulated by a member of the HTH GntR-family. We chose the *S. coelicolor* genome for our investigation because of the exceptional large quantity of GntR-like members sequenced in this strain. A rapid and non-exhaustive search using the DNA motif program[5] revealed about 20 promoter regions that possess a putative GntR-like palindromic sequence. According to the observed reflected C-terminal heterogeneity on operator sequences, the number of putative candidates in binding a specific GntR-like operator site is now reduced, as an investigation of the members of a subfamily would be preferred.

However, we must also mention that few GntR-like regulators recognize operator sites that do not fit into the consensus sequences presented in Table II. It is the case for TraR (44, 76),

[5] Found on the Web at sanger.ac.uk/Projects/Scoelicolor/.

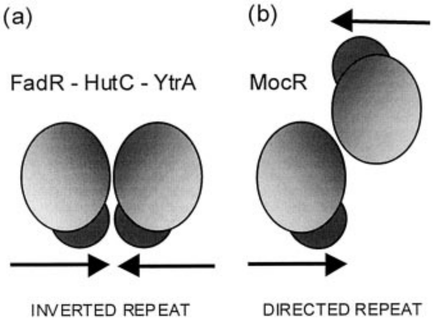

FIG. 4. **Hypothetical modes of dimerization for the FadR, HutC, YtrA, and MocR subfamilies.** Head-to-tail and anti-parallel dimer configurations are predicted, respectively, for the MocR subfamily and the FadR, HutC, and YtrA subfamilies. Directed repeat operator sequences at wide intervals are more appropriate for a head-to-tail configuration.

AphS and BphS (18, 19), and FucR (51), which bind boxes with no clearly defined symmetrical properties. Thus, although the consensus sequences presented in Table II should be regarded as interesting tools, for instance, in making sequencing projects maximally useful, they certainly should not be considered as unerring references, and some GntR regulators should not fit with the general properties highlighted in this study.

DISCUSSION

The structural, phylogenetic, and functional analysis of about 270 members of the bacterial HTH GntR-family led us to limit the C-terminal E-b/O domain heterogeneity to four major subfamilies that we called FadR, HutC, MocR, and YtrA. The presence of a few proteins escaping from this subdivision suggests that other subfamilies may be identified soon. Among members presenting a C-terminal domain that diverges from the four subfamilies defined above, the most interesting case comes from AraR in *B. subtilis*. The protein presents a GntR-like DNA-binding domain and a C-terminal domain that is GntR-like and a C-terminal domain typical of the HTH LacI/GalR family. AraR is a hybrid protein that is able to bind operator sites (AaACTTGT/A/T/ACAAGTaT) (50) that presents the typical GntR signature, and its C-terminal domain binds to a carbohydrate effector molecule (L-arabinose) as do most of the members of the LacI/GalR family. Recently, some proteins presenting this mosaic modular association have been sequenced (*i.e.* RliB from *Lactococcus lactis*, ssp. lactis, Swiss-Prot accession no. Q9CFH6; SPY1602 from *Streptococcus pyogenes*, Swiss-Prot accession no. Q99YP7; CAC1340 from *Clostridium acetobutylicum,* Swiss-Prot accession no. Q97JE6), confirming in a short time the emergence of new subfamilies.

The fact that C-terminal E-b/O heterogeneity seems to be reflected in the DNA-binding domain and in operator sequences suggests the existence of a tight link between the three regions involved in the regulatory process. This is not really surprising as *in vivo*, in the evolutionary process, once a gene and its upstream region present a successful functional combination between the three regions involved in gene regulation, it seems legitimate that descendants emerging through gene duplication would present a relative conservation throughout the duplicated sequence. Conservation between the three regions could also be explained from a structural and functional point of view. Dimerization certainly imposes steric constraints on the D-b domain, reducing its mobility with respect to the rest of the protein. According to the studies realized on AraC (72, 74, 75) (XylS/AraC HTH family) and LexA (73), both from *E. coli*, such a restricted mobility is thought to be due to interactions between the D-b and E-b/O domains and/or to interactions of part of the linker region with one of the two structural domains. These interactions might explain why a regulatory protein is limited, for instance, in its ability to accommodate a wide variation in distances between half-sites of palindromic operator sequences or to form DNA looping when *cis*-acting elements are separated by a nonintegral number of helix turn. Works on LexA show that the DNA binding ability of a specific domain can be enhanced or diminished by fusing the D-b domain with some alternative dimerization domains (73). These results obtained *in vitro* could explain why *in vivo*, among a family that presents a conserved DNA-binding domain, we observed different operator consensus sequences according to the E-b/O heterogeneity.

Finally, we have also delimited how far the information relative to a unique protein can constitute the theoretical and experimental framework of the other members of the family. According to our comparative study, the structural data relative to the FadR protein (52, 53) should be regarded as a reference for the whole GntR-family concerning the DNA-binding domain but must be limited to the FadR subfamily concerning the E-b/O domain. Moreover, because of the daily increasing amount of genome sequences listed, it seems essential to update and extend the early comparative studies realized on other families to make sequencing projects maximally useful.

Acknowledgments—We thank Dr. Josette Lamotte-Brasseur for technical help in using the programs from the PHYLIP package and Maria Colombo for valuable assistance and kind support in the preparation of this manuscript.

REFERENCES

1. Haydon, D. J, and Guest, J. R. (1991) *FEMS Microbiol. Lett.* **63,** 291–295
2. Harrison, S. C. (1991) *Nature* **353,** 715–719
3. Pabo, C. O., and Sauer, R. T. (1992) *Annu. Rev. Biochem.* **61,** 1053–1095
4. Henikoff, S., Haughn, G. W., Calvo, J. M., and Wallace, J. C. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85,** 6602–6606
5. Brennan, R. G., and Matthews, B. W. (1989) *J. Biol. Chem.* **264,** 1903–1906
6. Weickert, M. J., and Adhya, S. (1992) *J. Biol. Chem.* **267,** 15869–15874
7. Gallegos, M.-T., Michán, C., and Ramos, J. L. (1993) *Nucleic Acids Res.* **21,** 807–810
8. Nguyen, C. C., and Saier, M. H., Jr. (1995) *FEBS Lett.* **377,** 98–102
9. Pabo, C. O., and Sauer, R. T. (1984) *Annu. Rev. Biochem.* **53,** 293–321
10. Wintjens, R., and Rooman, M. (1996) *J. Mol. Biol.* **262,** 294–313
11. Rosinski, J. A., and Atchey, W. R. (1999) *J. Mol. Evol.* **49,** 301–309
12. Karmirantzou, M., and Hamodrakas, S. J. (2001) *Protein Eng.* **14,** 465–472
13. Fujita, Y., Fujita, T, Miwa, Y., Nihashi, J.-I., and Aratani, Y. (1986) *J. Biol. Chem.* **261,** 13744–13753
14. Reizer, A., Deutscher, J., Saier, M. H., Jr., and Reizer, J. (1991) *Mol. Microbiol.* **5,** 1081–1089
15. Giannotta, F. (1998) *Eléments cis et Trans dans la Régulation de la Xylanase C de Streptomyces sp. EC3.* Ph.D. thesis, Université de Liège, Belgium
16. Hamza, I., Chauhan, S., Hassett, R., and O'Brian, M. (1998) *J. Biol. Chem.* **273,** 21669–21674
17. Ortuño-Olea, L., and Durán-Vargas, S. (2000) *FEMS Microbiol. Lett.* **189,** 177–182
18. Arai, H., Akahira, S., Ohishi, T., and Kudo, T. (1999) *Mol. Microbiol.* **33,** 1132–1140
19. Watanabe, T., Inoue, R., Kimura, N., and Furukawa, K. (2000) *J. Biol. Chem.* **275,** 31016–31023
20. Mahenthiralingam, E., Simpson, A. A., and Speert, D. P. (1997) *J. Clin. Microbiol.* **35,** 808–816
21. Robert-Baudouy, J., Portalier, R., and Stoeber, F. (1981) *J. Bacteriol.* **145,** 211–220
22. DiRusso, C. C. (1988) *Nucleic Acids Res.* **16,** 7995–8009
23. DiRusso, C. C., Heimert, T. L., and Metzger, A. K.(1992) *J. Biol. Chem.* **267,** 8685–8691
24. DiRusso, C. C., Metzger, A. K., and Heimert, T. L. (1993) *Mol. Microbiol.* **7,** 311–322
25. Pellicer, M.-T., Badía, J., Aguillar, J., and Baldomà, L. (1996) *J. Bacteriol.* **178,** 2051–2059
26. Pellicer, M.-T., Fernandez, C., Badía, J., Aguilar, J., Lin, E. C. C., and Baldomà, L. (1999) *J. Biol. Chem.* **274,** 1745–1752
27. Núñez, M. F., Pellicer, M.-T., Badía, J., Aguilar, J., and Baldomà, L. (2001) *Microbiology* **147,** 1069–1077
28. Dong, J. M., Taylor, J. S., Latour, D. J., Iuchi, S., and Lin, E. C. C. (1993) *J. Bacteriol.* **175,** 6671–6678
29. Lin, J.-W., Lu, H. C., Chen, H.-Y., and Weng, S.-F. (1997) *Biochem. Biophys. Res. Commun.* **239,** 228–234
30. Lee, H. Y. L., An, J. H., and Kim, Y. S. (2000) *Eur. J. Biochem.* **267,** 7224–7229
31. Koo, J. H., and Kim, Y. S. (1999) *Eur. J. Biochem.* **266,** 683–690
32. Knobel, H.-R., Egli, T., and van der Meer, J. R. *(1996) J. Bacteriol.* **178,** 6123–6132
33. Stephens, P. E., Darlison, M. G., Lewis, H. M., and Guest, J. R. (1983) *Eur. J. Biochem.* **133,** 155–162

34. Poupin, P., Ducrocq, V., Hallier-Soulier, S., and Truffaut, N. (1999) *J. Bacteriol.* **181,** 3419–3426
35. Shulami, S., Gat, O., Sonenshein, A. L., and Shoham, Y. (1999) *J. Bacteriol.* **181,** 3695–3704
36. Morawski, B., Segura, A., and Ornston, L. N. (2000) *FEMS Microbiol. Lett.* **187,** 65–68
37. Ryding, N. J., Kelemen, G. H., Whatling, C. A., Flärdh, K., Buttner, J., and Chater, K. F. (1998) *Mol. Microbiol.* **29,** 343–357
38. Buck, D., and Guest, J. R. (1989) *Biochem. J.* **260,** 737–747
39. Quail, M. A., Dempsey, C. E., and Guest, J. R. (1994) *FEBS Lett.* **356,** 183–187
40. Allison, S. L., and Phillips, A. T. (1990) *J. Bacteriol.* **172,** 5470–5476
41. Kendall, K. J., and Cohen, S. N. (1988) *J. Bacteriol.* **170,** 4634–4651
42. Hagège, J., Pernodet, J.-L., Sezonov, G., Gerbaud, C., Friedmann, A., and Guérineau, M. (1993) *J. Bacteriol.* **175,** 5529–5538
43. Makino, K., Kim, S.-K., Shinagawa, H., Amemura, M., and Nakata, A. (1991) *J. Bacteriol.* **173,** 2665–2672
44. Servín-González, L., Sampieri, A., III, Cabello, J., Galván, L., Juárez, V., and Castro, C. (1995) *Microbiol.* **141,** 2499–2510
45. Schoeck, F., and Dahl, M. K. (1996) *Gene* **175,** 59–63
46. Rossbach, S., Kulpa, D. A., Rossbach, U., and de Bruijn, F. I. (1994) *Mol. Gen. Genet.* **245,** 11–24
47. Magarvey, N., He, J., Aidoo, K. A., and Vining, L. C. (2001) *Microbiology* **147,** 2103–2112
48. Titgemeyer, F., Reizer, J., Reizer, A., Tang, J., Parr, T. R., Jr., and Saier, M. H., Jr. (1995) *DNA Seq.* **5,** 145–152
49. Yoshida, K.-I., Fujita, Y., and Ehrlich, S. D. (2000) *J. Bacteriol.* **182,** 5454–5461
50. Mota, L. J., Tavares, P., and Sá-Nogueira, I. (1999) *Mol. Microbiol.* **33,** 476–489
51. Hooper, L. V., Xu, J., Falk, P. G., Midtvedt, T., and Gordon, J. I. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 9833–9838
52. van Aalten, D. M. F., DiRusso, C. C., Knudsen, J., and Wierenga, R. K. (2000) *EMBO J.* **19,** 5167–5177
53. van Aalten, D. M. F., DiRusso, C. C., and Knudsen, J. (2001) *EMBO J.* **80,** 2041–2050
54. Corpet, F. (1988) *Nucleic Acids Res.* **16,** 10881–10890
55. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680
56. Young, C. L., Barker, W. C., Tomaselli, C. M., and Dayhoff, M. O. (1979) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed) Vol. 5, Suppl. 3, pp. 73–93, National Biochemical Foundation, Silver Spring, MD
57. Fitch, W. M., and Margoliash, E. (1967) *Science* **155,** 279–284
58. Page, R. D. M. (1996) *Comput. Appl. Biosci.* **12,** 357–358
59. Felsenstein, J. (1989) *Cladistics* **5,** 164–166
60. Yoshida, K.-I., Ohmori, H., Miwa, Y., and Fujita, Y. (1995) *J. Bacteriol.* **177,** 4813–4816
61. Sung, M. H., Tanizawa, K., Tanaka, H., Kuramitsu, S., Kagamiyama, H., Hirotsu, K., Okamoto, A., Higuchi, T., and Soda, K. (1991) *J. Biol. Chem.* **266,** 2567–2572
62. Ko, T.-P., Wu, S.-P., Yang, W.-Z., Tsai, H., and Yuan, H. S. (1999) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55,** 1474–1477
63. Raman, N., Black, P. N., and DiRusso, C. C. (1997) *J. Biol. Chem.* **272,** 30645–30650
64. Yoshida, K.-I., Fujita, Y., and Sarai, A. (1993) *J. Mol. Biol.* **231,** 167–174
65. Rodionov, D. A., Mironov, A. A., Rakhmaninova A. R., and Gelfand, M. S. (2000) *Mol. Microbiol.* **38,** 673–683
66. Quail, M. A., and Guest, J. R. (1995) *Mol. Microbiol.* **15,** 519–529
67. Koo, J. H., Cho, I. H., and Kim, Y. S. (2000) *J. Bacteriol.* **182,** 6382–6390
68. Hu, L., Allison, S. L., and Phillips, A. T. (1989) *J. Bacteriol.* **171,** 4189–4195
69. Sezonov, G., Possoz, Ch., Friedmann, A., pernodet, J.-L., and Guérineau, M. (2000) *J. Bacteriol.* **182,** 1243–1250
70. Schöck, F., and Dahl, M. K. (1996) *J. Bacteriol.* **178,** 4576–4581
71. Bürken, L., Schöck, F., and Dahl, M. K. (1998) *Mol. Gen. Genet.* **260,** 48–55
72. Dunn, T. M., Hahn, S., Odgen, S., and Schleif, R. F. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81,** 5017–5020
73. Oertel-Buchheit, P., Schmidt-Dörr, T., Granger-Scharr, M., and Schnarr, M. (1992) *J. Mol. Biol.* **229,** 1–7
74. Carra, J. H., and Schleif, R. F. (1993) *EMBO J.* **12,** 35–44
75. Harmer, T., Wu, M., and Schleif, R. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 427–431
76. Kataoka, M., Kosono, S., Seki, T., and Yoshida, T. (1994) *J. Bacteriol.* **176,** 7291–7298