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 ITH2 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.

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.
EXPERIMENTAL PROCEDURES

Selection of GntR-like Members—Members of the GntR family were identified from the SWISS-PROT/TrEMBL/GenBank sequence data bases (last update, June 2001) by a keywords search on the ExPASy molecular Biology server and NCBI server. All sequences proposed by the databases 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 databases, i.e., the Irr protein from Bradyrhizobium japonicum, 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, most representative members, or proteins yielding data of specific interest were selected for presentation. 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. 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. The method was revealed to have an accuracy of >90% for FadR with most of the inaccuracies occurring at the

| Protein (reference) | Function | Organism (abbreviation) | Length in aa | Sequence number | Sub-family |
|---------------------|----------|-------------------------|--------------|-----------------|-----------|
| AnaR (17) | L-arginine permease operon repressor | Rhizobium etli (Rhe) | 239 | Q18370.1 |         |
| AphiR (14) | Regulator of genes involved in phenol utilization | Comamonas testosteroni TA414 (Cta) | 239 | Q18918 |         |
| BphS (15) | Repressor of genes involved in biphenyl degradation | Alcaligenes eutrophus (Aeu) | 243 | Q4EY74 |         |
| DgpR | Putative galactonate operon repressor | Escherichia coli (Eco) | 229 | P13160 |         |
| DnaF (16) | Unknown function | E. coli K12-magnesium sensitive bacterium (EKM) | 207 | Q8H71 |         |
| ExrR (12) | Epidemic strain marker regulator | Burkholderia cepacia (Bcp) | 277 | P09570 |         |
| ExrR (12) | Hexosamine regulon regulator | Escherichia coli | 258 | P29078 |         |
| FadR (42-244) | Fatty acid metabolism regulator | Escherichia coli | 258 | P29078 |         |
| GLG (210-221) | Glycolate oxidation operon activator | Escherichia coli | 254 | P25072 |         |
| GLG (210-221) | Putative glycolate oxidation operon regulator | Pseudomonas aeruginosa (Paer) | 251 | Q18943 |         |
| GntR (113-143) | Glutamine operon repressor | Bacillus subtilis (Bat) | 243 | P01385 |         |
| LiR (17-255) | Putative L-fucose dehydrogenase operon repressor | Escherichia coli | 258 | P33332 |         |
| LuxK (125) | Lux operon enhancer | Photorhabdus luminescens (Plum) | 221 | Q9124N |         |
| NahR (130) | Activator for malfonate metabolism | Rhizobium leguminosarum (Rle) | 222 | Q9P74 |         |
| MdbB (313) | Malonate decarboxylase operon repressor | Acinetobacter calcoaceticus (Aca) | 224 | Q95028 |         |
| PilK (131) | Putative nitrite/nitric oxide reductase subunits regulator | Chelatobacter heintzii (CHei) | 210 | P45918 |         |
| PipR (38) | Pivoyne dehydrogenase complex repressor | Escherichia coli | 254 | P69057 |         |
| PipR (38) | Cytadverse P450 regulator | Mycobacterium smegmatis mc2155 (msm) | 245 | P42362 |         |
| BGP5S06 | Unknown function | Streptomyces coelicolor | 253 | Q9R083 |         |
| BCGgly7 | Unknown function | Streptomyces coelicolor | 231 | Q9R8K9 |         |
| BGLU (53) | Glucaronic acid (GlcUA) gene cluster regulator | Bacillus steearothermophilus T-6 (Bat) | 249 | Q9FL9 |         |
| BmoR (149) | Fumarate dehydrogenase synthesis repressor | Acinetobacter sp. ADP1 (Aac) | 251 | Q24809 |         |
| BhuC (117) | Sporholatin transcription factor | Streptomyces coelicolor (Sco) | 251 | Q05536 |         |
| BhuC (117) | Fatty acyl responsive regulator | Escherichia coli | 240 | P13069 |         |
| HUC (40) | Histidine utilization regulator | Pseudomonas putida (Ppu) | 248 | P22273 |         |
| KorA (41) | p701 ki-kor repressor | Streptomyces lividans (Stl) | 241 | P22405 |         |
| KorR (42) | pNAM2 ki-kor regulator | Streptomyces ambifaciens (San) | 259 | Q07191 |         |
| PheE (43) | Pheylalnine tRNA synthetase | Escherichia coli | 241 | P16044 |         |
| PheR | Unknown function | Salmonella typhimurium (Sty) | 239 | P95001 |         |
| SCO93 2B | Unknown function | Streptomyces coelicolor | 249 | P9F274 |         |
| SCO93 19 | Unknown function | Streptomyces coelicolor | 251 | Q9N2K2 |         |
| SCO93 28 | Unknown function | Streptomyces coelicolor | 254 | Q9N4K2 |         |
| TRA (44) | pIV1 Traa operon regulator | Streptomyces phaocochromogenes (Sir) | 245 | Q9L677 |         |
| TRA (44) | Tetraspore operon regulator | Bacillus subtilis | 258 | P39790 |         |
| XILR (139) | Regulator involved in catalase repression | Streptomyces lividans | 292 | Q9AC8R |         |
| YQD | Unknown function | Bacillus subtilis | 243 | O34817 |         |
| YQD (44) | Probable diisopine catabolism regulator | Rhizobium meliloti (Rmel) | 405 | P49299 |         |
| TMAO (47) | Pyridoxal phosphate synthase regulator | Streptomyces venezuelensis (SVe) | 152 | Q94E54 |         |
| PtdJ (44) | Putative phosphotransferase system regulator | Salmonella enterica (Sene) | 450 | P41083 |         |
| XDP | Unknown function | Bacillus subtilis | 444 | P01972 |         |
| YDF | Unknown function | Bacillus subtilis | 479 | P06420 |         |
| YDF (28) | Unknown function | Bacillus subtilis | 482 | P06681 |         |
| YDF (28) | Unknown function | Bacillus subtilis | 609 | P05758 |         |
| YDF (28) | Unknown function | Rhodococcus rhodochrous (Rrh) | 470 | P01509 |         |
| BGD (52) | Unknown function | Bacillus halodurans (Bhd) | 125 | Q9W71P |         |
| BGD (52) | Unknown function | Bacillus halodurans | 123 | Q9W30O |         |
| SDF (67) | Unknown function | Staphylococcus aureus subsp. aureus | 126 | Q9S16V4 |         |
| SCO11-5 | Unknown function | Streptomyces coelicolor | 140 | Q9FAA2 |         |
| SCO11-5 | Unknown function | Streptomyces coelicolor | 115 | Q9FAA5 |         |
| TAD25 | Unknown function | Thermoplasma acidophilum (Taa) | 103 | Q9H0L8 |         |
| TAD25 | Unknown function | Thermoplasma acidophilum | 121 | Q9W855 |         |
| YPR (127) | Unknown function | Bacillus subtilis | 121 | P14000 |         |
| YPR (127) | Unknown function | Bacillus subtilis | 130 | O34712 |         |

| Protein (reference) | Function | Organism (abbreviation) | Length in aa | Sequence number | Sub-family |
|---------------------|----------|-------------------------|--------------|-----------------|-----------|
| AnaR (17) | Transcriptional repressor of the arabinose operon | Bacillus subtilis | 384 | P09171 |         |
| FadR (51) | Repressor of L-arabinose utilization gene cluster | Bacillus subtilis | 331 | Q9R024 |         |

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.
Multiple Alignments and Phylogenetic Tree Construction—Multiple alignments were developed with the MULTIALIN (54) and CLUSTALW (55) 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 Felsenstein (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. 2a) 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 (α9). 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 (GlcE), galactonate (DgoR), lactate (LidR), 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. 2b. 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, β3, 57, and β6 in Fig. 2b). 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. 2d). 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 aldime 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. 2e). 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. 2e), 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
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: \( \alpha_1 \beta_1 \alpha_2 \beta_2 \). 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 \( \alpha_2 \) and \( \alpha_3 \).

The DNA-binding domain 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: \( \alpha_1 \beta_1 \alpha_2 \beta_2 \). 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 \( \alpha_2 \) and \( \alpha_3 \).

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
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 (Fig. 2). The most frequent anomalies were the lack of the first α-helix (α1) (NtaR from Chelatobacter heintzii and EmoR from the EDTA-degradating bacterium, BNC1) or the presence of an additional helix upstream of α1 (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 α1.

Operator Sites Analysis—Although there is no precise "recognition code" involving a one-to-one correspondence between
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 \(5'-\text{GT(N)}\_y\-3'\) 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 \(5'-\text{GT}\) and \(3'-\text{AC}\) 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 \(5'-\text{GT}\_a\-3'\) and \(5'-\text{GT}\_a\-3'\), 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 \(5'-\text{GT}\) and \(3'-\text{AC}\) 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 more dimer formation specific for each member of the subfamily.
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 program5 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/.
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 (AaACTGTTA/ATACAAGTTaT) (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. Q5CFH6; 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, in fact, 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 noninteger 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. B. (1991) *FEBS Microbiol. Lett.* 63, 291–295
2. Harrison, S. C. (1991) *Nature* 353, 715–719
3. Pabo, C. O., and Sauer, R. T. (1991) *Annu. Rev. Biochem.* 60, 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. Bieneman, 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., Michan, C., and Ramos, J. L. (1993) *Nucleic Acids Res.* 21, 807–810
8. Nguyen, C. C., and Saier, M. H., Jr. (1995) *FEMS 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. Biol.* 49, 301–309
12. Karmirantzou, M., and Hamodrakas, S. J. (2001) *Protein Eng.* 14, 465–472
13. Fujita, Y., Fujita, T., Miwa, Y., Nishashi, 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) *J. Mol. Biol.* 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. Ortego-Olea, L., and Durán-Vargas, S. (2000) *FEBS Microbiol. Lett.* 189, 177–182
18. Araı, H., Akahira, S., Ohishi, T., and Kudo, T. (1999) *Mol. Microbiol.* 35, 1132–1140
19. Watanabe, T., Inoue, R., Kimura, N., and Furukawa, K. (2000) *J. Biol. Chem.* 275, 31016–31023
20. Menonhiralingam, E., Simpson, A. A., and Speert, D. P. (1997) *J. Clin. Microbiol.* 35, 808–816
21. Robert-Baudouy, J., Portalier, R., and Stoever, F. (1981) *J. Bacteriol.* 145, 211–220
22. DiRusso, C. C. (1988) *Nucleic Acids Res.* 16, 7905–8009
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. Pellier, M.-T., Badia, J., Aguillar, J., and Baldomà, L. (1996) *J. Bacteriol.* 178, 2051–2059
26. Pellier, M.-T., Fernandez, C., Badia, J., Aguilar, J., Lin, E. C. C., and Baldomà, L. (1999) *J. Biol. Chem.* 274, 1745–1752
27. Núñez, M. F., Pellier, M.-T., Badia, 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, 653–660
32. Knobel, H.-R., Eglı, T., and van den Meir, 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

---

**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.
34. Pospisil, P., Durocq, V., Hallier-Soulier, S., and Truffaut, N. (1999) J. Bacteriol. 181, 3419–3426
35. Shulami, S., Gat, O., Sosznehsin, 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. Ryting, N. J., Klemm, G. H., Whaling, 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. Hazgüe, J., Perneaud, J.-L., Sexonov, 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. Servin-Gonzalez, L., Sampieri, A., III, Cabello, J., Galván, L., Juárez, V., and Castro, C. (1995) Microbiol. 141, 2499–2519
45. Schwock, F., and Dahl, M. K. (1996) Gene 175, 59–63
46. Rossbach, S., Kulpa, D. A., Rossbach, U., and de Bruijn, P. I. (1994) Mol. Gen. Genet. 245, 11–24
47. Magarvey, N., He, J., Aido, K. A., and Vining, L. C. (2001) Microbiology 147, 2103–2112
48. Tögezmeier, 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, 476–486
51. Biesemeyer, 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. 20, 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 Margolash, 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., Taniawaza, 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, 174–177
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. Sexonov, G., Possee, Ch., Friedmann, A., perneaud, J.-L., and Guérineau, M. (2000) J. Bacteriol. 182, 1243–1250
70. Schock, F., and Dahl, M. K. (1996) J. Bacteriol. 178, 4576–4581
71. Burken, L., Schick, P., and Dahl, M. K. (1998) Mol. Gen. Genet. 260, 48–55
72. Dunn, T. M., Hahn, S., Odgen, S., and Schlief, R. F. (1984) Proc. Natl. Acad. Sci. U. S. A. 81, 5017–5020
73. Oertel-Buchheit, P., Schmidt-Doerning, K., Granger-Scharr, M., and Schnarr, M. (1992) J. Mol. Biol. 229, 1–7
74. Carra, J. H., and Schlief, R. F. (1993) EMBO J. 12, 35–44
75. Harner, T., Wu, M., and Schlief, R. (2001) Proc. Natl. Acad. Sci. U. S. A. 98, 427–431
76. Katoaoka, M., Kosono, S., Seki, T., and Yoshida, T. (1994) J. Bacteriol. 176, 7291–7298