Arac/XylS family of transcriptional regulators

GALLEGOS, Maria Trinidad, et al.

Abstract

The ArC/XylS family of prokaryotic positive transcriptional regulators includes more than 100 proteins and polypeptides derived from open reading frames translated from DNA sequences. Members of this family are widely distributed and have been found in the gamma subgroup of the proteobacteria, low- and high-G + C-content gram-positive bacteria, and cyanobacteria. These proteins are defined by a profile that can be accessed from PROSITE PS01124. Members of the family are about 300 amino acids long and have three main regulatory functions in common: carbon metabolism, stress response, and pathogenesis. Multiple alignments of the proteins of the family define a conserved stretch of 99 amino acids usually located at the C-terminal region of the regulator and connected to a nonconserved region via a linker. The conserved stretch contains all the elements required to bind DNA target sequences and to activate transcription from cognate promoters. Secondary analysis of the conserved region suggests that it contains two potential alpha-helix-turn-alpha-helix DNA binding motifs. The first, and better-fitting motif is supported by [...]

Reference


PMID : 9409145
INTRODUCTION

Searches for homology among protein sequences can identify well-conserved motifs such as cofactor binding domains, transient peptides, helix-turn-helix and zinc finger DNA-binding motifs, and others. This approach can also identify families of related proteins in which homology extends over one or several domains of proteins that possess similar functions (5, 7, 21, 102, 103, 167, 195–197, 257).

Within the current sequence databases, the AraC/XylS family of regulators is one of the most common positive regulators (84, 209, 243). Other common regulator families are ArsR (175), AsnC (140), Crr (233), DeoR (17), GntR (29, 104, 213), IdR (214), LacI (253), LuxR/UpA (112), LysR (111), MarR (56), MerR (107), NutR (189), TetR (198), YedF/YeeD/YhhP (14), and YhdG/YgbN/YohI (15).

The AraC/XylS family is characterized by significant amino acid sequence homology extending over a 100-residue stretch constituting the DNA binding domain of the family members. The domain is most often found in oligomeric proteins, but in a few natural cases (4, 18, 47, 86, 151, 260) and in artificial cases (37, 143, 170) the single conserved domain itself can bind to DNA and activate transcription from cognate promoters.

The domain does not appear to bind effector molecules, this function being provided either by additional domains in the family members or by other proteins that regulate the synthesis of AraC/XylS family members.

AraC, the regulator of the L-arabinose operon in Escherichia coli, was the first member to be identified, purified, and characterized biochemically (95, 223–226). Tobin and Schleif (243) envisaged that AraC, RhaS, and RhaR defined a group of transcriptional regulators. Later, Ramos et al. (209) and Henikoff et al. (112) suggested that eight proteins (AraC, RhaR, RhaS, MelR, and Rns from E. coli; XylS from Pseudomonas putida; AraC from Erwinia carotovora; and VirF from Yersinia enterocolitica) formed an incipient family. In 1993, Gallegos et al. (84) extended the family to include 27 proteins with the addition of AdaA from Bacillus subtilis; AraC from Citrobacter freundii; AppY (also called M5); CelD; CfaD; EmV; FapR; SoxS; TetD from E. coli; EcrA; and MmsR from Pseudomonas aeruginosa; VirF from Shigella flexneri; AraC and RhaS from Salmonella typhimurium; TcpN (also called ToxT) from Vibrio cholerae; LcrF from Yersinia pestis; and several natural XylS proteins from different TOL plasmids. These proteins were aligned with the PILEUP program, which made it possible to define a 99-amino-acid stretch of homology at the C terminus of these proteins.

In this review, we have extended the family to include more than 100 proteins and polypeptides derived from open reading frames (ORFs) translated from DNA sequences. Here we summarize and discuss the general distinguishing characteris-
tics of the family, the structure-function organization of the AraC/XylS family of polypeptides, and the biochemical and molecular aspects of their mode(s) of action.

CURRENT MEMBERS OF THE AraC/XylS FAMILY

Successive Search for Members of the Family

To identify new members of the AraC/XylS family of transcriptional regulators, the 99 amino acids of the C-terminal end of the 27 proteins identified as members of the family (84) were aligned and analyzed by the algorithm of Lüthy et al. (160). This allowed us to define a matrix for the profile of the aligned sequences. This profile was then used to search for new members of the family within protein databases (SWISSPROT and PIR). The newly identified proteins were retrieved and aligned with the 99 amino acids of the family, and a new profile was again defined. This new profile was then used to search for putative polypeptides as members of the family by searching in nucleic acid databases (EMBL, GenBank, Genpept, and TREMBL), where we identified ORFs whose translated sequences gave a polypeptide that was a probable regulatory protein and a member of this family. Finally, a new alignment of all found sequences was carried out. This again defined the 99-amino-acid stretch as the most highly conserved region of this family of proteins, and a new profile was defined by analyzing the segment with the algorithm of Lüthy et al. (160). This new profile was used to search for members of the family in all available protein and nucleic acid databases (March 1997), but no new members were identified. This profile therefore now defines the AraC/XylS family of transcriptional regulators. It can be accessed from the PROSITE database as entry PS01124.

Table 1 lists the current proteins identified from protein and DNA databases (March 1997) as members of the AraC/XylS family. All characterized proteins of the family are positive transcriptional regulators except CelD, which seems to be a repressor (199). As shown in Table 1, members of the AraC/XylS family regulate very diverse genes and functions. Some members of the family control single operons or genes; others control multiple, unlinked target genes (regulons); while others are themselves regulated by other genes, forming complex regulatory networks (184).

Analyses of a protein sequence with the matrix assigned a value to the query sequence. The value assigned by the matrix to each of the family members ranged from 30.74 to 12.52, with small variations between two consecutive proteins identified as members of the family (the complete set of values is available from K. Hofmann). However, a difference of 4.7 points was observed between the last member of the family assigned by the profile, namely, Hrp from Xanthomonas oryzae, and the closest value of a protein not identified as a member of the family, MutS from Thermus thermophilus. Therefore, we propose that a protein belongs to the AraC/XylS family if the value after analysis with the profile defined in PROSITE database is above 12.52.

Nonetheless, comparison of a query sequence with the conserved domain of any of the family members can identify the query sequence as a member of the family. The sequence can then be rapidly aligned to any of the homologs with the FASTA program.

Functions Regulated by Members of the Family

As mentioned above, all proteins in the AraC/XylS family are positive transcriptional activators except CelD, which seems to act as a repressor (199). Two members of the family, the AraC protein from E. coli and the YbtA protein from Y. pestis, can function both as an activator and as a positive regulator (73, 224, 225) in different promoters or in the same promoter depending on the presence or the absence of appropriate effectors.

Two types of proteins are distinguished in the family: in one group, the signal receptor resides in the same polypeptide as the regulatory function (i.e., AraC, XylS, RhaR, and UreR) (59, 63, 172, 185, 210, 232, 243); in the other group, transcription of the regulatory protein is controlled by another regulator. This regulator can be an activator or a repressor, so that stimulation or derepression of transcription leads to the overexpression of the member of the AraC/XylS family, which in turn regulates transcription from cognate promoters (i.e., MarA, SoxS, and TepN) (4, 113, 129, 190, 260). The proteins belonging to the family have three main regulatory functions in common: carbon metabolism, stress response, and pathogenesis.

Regarding carbon metabolism, members of the family control the degradation of sugars such as arabinose (AraC), cellobiose (CelD), melibiose (MelR and MsmR), raffinose (RaFR and MsmR), rhamnose (RhaR and RhaS), and xylose (XylR); certain amino acids such as valine (MmsR), arginine (AdiY), and ornithine (OruR); alcohols such as 1,2-propanediol (PocR); alkylbenzoates (XylS); p-hydroxyphenylacetic acid (HpaA); and herbicides such as S-ethyl dipropylthiocarbamate (TchR). These transcriptional regulators are characterized by the fact that they stimulate transcription from cognate promoters in response to the presence of the effector. All of them are about 300 amino acids long (30, 32, 74, 105, 169, 178, 199, 204, 218, 238, 248, 251). The three best-characterized proteins in this subgroup of the family are AraC, RhaR, and XylS (84, 224, 244; see below for further details). Certain regulatory proteins are involved in the production of virulence factors in infections of plants (HrpB from Burkholderia solanacearum) or mammals. Among the latter, these regulatory factors have been found in microbes that colonize mainly the gastrointestinal tract but also the respiratory tract or the urinary system. These factors include AfR, AaR, CsaD, CsaR, TepR, perA, and Rss from E. coli; CaR and LerF from Y. pestis; EssA and PchR from P. aeruginosa; InvF from S. typhimurium; MxiE from Shigella flexneri; TcpN from V. cholerae; and VirF from Shigella and Y. pestis (36, 55, 77, 85, 91, 106, 113, 127, 129, 133, 183, 192, 222, 258, 263). These proteins are involved in stimulation of the synthesis of proteins that play a role in adhesion to epithelial tissues, such as fimbiae (AfrR, AggR, CsaD, CsaR, TepR, perA, Rss, and TcpN), components of the cell capsule (CaR), and invasins (EssA, HrpB, InvF, MxiE, and VirF). Some members of the family control the production of other virulence factors such as siderophores (PchR) and urease (UreR). These regulators can be plasmid or chromosomally encoded.

Except for UreR, which binds urea to become active (58), it has not been demonstrated that regulators of this group bind specific effectors, although all of them respond to environmental factors such as temperature, osmolality of the medium, and concentration of Ca2+ (20, 53, 55, 76, 115, 124, 141, 182, 202, 230, 242, 261). Certain regulators are involved in the response to stressors, e.g., response to alkylating agents (Ada from E. coli, S. typhimurium, and Mycobacterium tuberculosis and AdaA from Bacillus subtilis) (57, 100, 176, 177); response to oxidative stress (SoxS from E. coli and S. typhimurium) (4, 260); tolerance to antibiotics, organic solvents, and heavy metals (AarP from Providencia stuartii, MarA and Rob from E. coli, PgrA from Proteus vulgaris, and RamA from Klebsiella pneumoniae) (90, 121, 161, 229, 240); and transition from exponential growth to

at Bibliotheque Faculte Medecine Genève on August 27, 2009
<table>
<thead>
<tr>
<th>Protein</th>
<th>Microorganism</th>
<th>Accession no.</th>
<th>Function</th>
<th>No. of residues</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AarP</td>
<td>Providencia stuartii</td>
<td>SP:P43463</td>
<td>Transcriptional activator of acc (2') la gen for 2'-N-acetyltransferase</td>
<td>135</td>
<td>161</td>
</tr>
<tr>
<td>Ada</td>
<td>Escherichia coli</td>
<td>SP:P06134</td>
<td>Repair of alkylated guanine in DNA by stoichiometrically transferring the alkyl group at the O-6 position to a cysteine residue in the enzyme in a suicide reaction, because the enzyme is irreversibly inactivated; can also repair O-4-methylthymine. The methylated Ada protein is a positive regulator of its own synthesis (<em>ada</em>) and that of other genes, such as <em>alkA</em>, <em>alkB</em>, and <em>aidB</em></td>
<td>354</td>
<td>57, 138, 150, 180</td>
</tr>
<tr>
<td>Ada</td>
<td>Mycobacterium tuberculosis</td>
<td>SP:Q10630</td>
<td>Similar to <em>E. coli</em> Ada</td>
<td>496</td>
<td>177</td>
</tr>
<tr>
<td>Ada</td>
<td>Salmonella typhimurium</td>
<td>SP:P26189</td>
<td>Similar to <em>E. coli</em> Ada</td>
<td>352</td>
<td>100</td>
</tr>
<tr>
<td>AdaA</td>
<td>Bacillus subtilis</td>
<td>SP:P19219</td>
<td>One of the two proteins required for the adaptive response to alkylating agents. It accepts a methyl group from methylphosphotriesters and then acts as a transcriptional activator of the <em>ada</em> operon</td>
<td>211</td>
<td>176</td>
</tr>
<tr>
<td>AdiY</td>
<td>Escherichia coli</td>
<td>SP:P33234</td>
<td>Transcriptional activator of the <em>adiA</em> gene for biodegradative acid-induced arginine decarboxylase</td>
<td>253</td>
<td>32, 237</td>
</tr>
<tr>
<td>AfrR</td>
<td>Escherichia coli</td>
<td>TE:Q07681</td>
<td>Probable transcriptional activator of the <em>afrABRS</em> operon for expression of AF/R1 fimbria in <em>E. coli</em> RDEC-1, a rabbit pathogen</td>
<td>272</td>
<td>258</td>
</tr>
<tr>
<td>AggR</td>
<td>Escherichia coli</td>
<td>SP:P43464</td>
<td>Transcriptional activator of the <em>aggA</em> gene for aggregative adherence fimbria I (AAF/I) expression in enteraggregative <em>E. coli</em> strains</td>
<td>265</td>
<td>183</td>
</tr>
<tr>
<td>AppY</td>
<td>Escherichia coli</td>
<td>SP:P05052</td>
<td>Transcriptional activator of the <em>cycAB</em>, <em>hyaAB</em>, <em>cdef</em>, and <em>appA</em> operons during the deceleration phase of growth</td>
<td>243</td>
<td>12, 131</td>
</tr>
<tr>
<td>AraC</td>
<td>Citrobacter freundii</td>
<td>SP:P11765</td>
<td>Regulator of several operons involved in the transport and catabolism of L-arabinose (similar to <em>E. coli</em> AraC)</td>
<td>281</td>
<td>30</td>
</tr>
<tr>
<td>AraC</td>
<td>Escherichia coli</td>
<td>SP:P03021</td>
<td>Activator of the expression of the <em>araBAD</em>, <em>araFGH</em> and <em>araE</em> operons, which are involved in the transport and catabolism of L-arabinose. Repressor of its own synthesis</td>
<td>292</td>
<td>174, 238, 248, 266</td>
</tr>
<tr>
<td>AraC</td>
<td>Erwinia chrysanthemi</td>
<td>SP:P07642</td>
<td>Similar to <em>E. coli</em> AraC</td>
<td>310</td>
<td>149</td>
</tr>
<tr>
<td>AraC</td>
<td>Salmonella typhimurium</td>
<td>SP:P03022</td>
<td>Similar to <em>E. coli</em> AraC</td>
<td>281</td>
<td>46</td>
</tr>
<tr>
<td>AraL</td>
<td>Streptomyces antibioticus</td>
<td>SP:Q03320</td>
<td>Unknown</td>
<td>303</td>
<td>265</td>
</tr>
<tr>
<td>AraL</td>
<td>Streptomyces lividans</td>
<td>SP:P35319</td>
<td>Unknown</td>
<td>304</td>
<td>43</td>
</tr>
<tr>
<td>CalR</td>
<td>Yersinia pestis</td>
<td>SP:P26950</td>
<td>Positive regulator of <em>cafIMA</em> and <em>cafl</em> operons for the production and transport of the capsule antigen F1</td>
<td>301</td>
<td>85, 128</td>
</tr>
<tr>
<td>CelID</td>
<td>Escherichia coli</td>
<td>SP:P17410</td>
<td>Repressor of the <em>celABCF</em> operon involved in the degradation of cellobiose, arbutin, and salicin</td>
<td>280</td>
<td>199</td>
</tr>
<tr>
<td>CfaD</td>
<td>Escherichia coli</td>
<td>SP:P25393</td>
<td>Transcriptional activator of the <em>cfaABCE</em> operon for the production of CFA/I fimbriae in enterotoxigenic <em>E. coli</em> strains</td>
<td>265</td>
<td>222</td>
</tr>
<tr>
<td>CsvR</td>
<td>Escherichia coli</td>
<td>SP:P43460</td>
<td>Transcriptional activator of the operon involved in the production of CS5 fimbriae in enterotoxigenic <em>E. coli</em> strains</td>
<td>301</td>
<td>55</td>
</tr>
<tr>
<td>EnvY</td>
<td>Escherichia coli</td>
<td>SP:P10805</td>
<td>Transcriptional temperature-dependent activator of several <em>E. coli</em> envelope proteins, most notably the porins OmpF and OmpC and the λ receptor, LamB</td>
<td>253</td>
<td>159</td>
</tr>
<tr>
<td>ExsA</td>
<td>Pseudomonas aeruginosa</td>
<td>SP:P26993</td>
<td>Transcriptional activator of the <em>extCBA</em> operon and <em>exsD</em>, <em>exsS</em>, and <em>ORF1</em> genes required for the synthesis and secretion of exoenzyme S</td>
<td>298</td>
<td>77</td>
</tr>
<tr>
<td>FapR</td>
<td>Escherichia coli</td>
<td>SP:P23774</td>
<td>Transcriptional activator of the 987P operon for fimbrial proteins in enterotoxigenic <em>E. coli</em> strains</td>
<td>260</td>
<td>133</td>
</tr>
<tr>
<td>HpaA</td>
<td>Escherichia coli</td>
<td>TE:Q46985</td>
<td>Transcriptional activator of <em>hpaBC</em> operon for catabolism of p-hydroxyphenylacetic acid</td>
<td>295</td>
<td>204</td>
</tr>
</tbody>
</table>

*Continued on following page*
<table>
<thead>
<tr>
<th>Protein</th>
<th>Microorganism</th>
<th>Accession no.</th>
<th>Function</th>
<th>No. of residues</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HrpB</td>
<td><em>Burkholderia solanacearum</em></td>
<td>SP:P31778</td>
<td>Transcriptional activator of the hypersensitive response genes (<em>hrp</em>) involved in plant pathogenicity</td>
<td>477</td>
<td>89</td>
</tr>
<tr>
<td>HrpXc</td>
<td><em>Xanthomonas campestris</em></td>
<td>TE:Q56801</td>
<td>Similar to <em>B. solanacearum</em> HrpB</td>
<td>503</td>
<td>193</td>
</tr>
<tr>
<td>HrpXv</td>
<td><em>Xanthomonas campestris</em></td>
<td>TE:Q56790</td>
<td>Similar to <em>B. solanacearum</em> HrpB</td>
<td>476</td>
<td>254</td>
</tr>
<tr>
<td>HrpXo</td>
<td><em>Xanthomonas oryzae</em></td>
<td>TE:Q56831</td>
<td>Similar to <em>B. solanacearum</em> HrpB</td>
<td>502</td>
<td>193</td>
</tr>
<tr>
<td>InvF</td>
<td><em>Salmonella typhimurium</em></td>
<td>SP:Q39437</td>
<td>Transcriptional activator of the <em>inv</em> operon required for epithelial tissue invasion</td>
<td>216</td>
<td>127</td>
</tr>
<tr>
<td>LcrF</td>
<td><em>Yersinia pestis</em></td>
<td>SP:P28808</td>
<td>Transcriptional activator of the virulence regulon (similar to <em>Y. enterocolitica</em> VirF)</td>
<td>271</td>
<td>116</td>
</tr>
<tr>
<td>LumQ</td>
<td><em>Photobacterium leiognathi</em></td>
<td>SP:Q51872</td>
<td>Unknown</td>
<td>248</td>
<td>153</td>
</tr>
<tr>
<td>LumQ</td>
<td><em>Synechocystis</em> sp.</td>
<td>TE:P73364</td>
<td>Unknown</td>
<td>241</td>
<td>126</td>
</tr>
<tr>
<td>MaoB</td>
<td><em>Escherichia coli</em></td>
<td>SP:Q47129</td>
<td>Transcriptional activator of the <em>maoA</em> gene coding a monoamine oxidase</td>
<td>301</td>
<td>263</td>
</tr>
<tr>
<td>MarA</td>
<td><em>Escherichia coli</em></td>
<td>SP:P27246</td>
<td>Transcriptional activator of the sodA, zwf, micF, slp, fpr, fumC, and nfo genes, which are involved in the multiple antibiotic resistance (mar) phenotype</td>
<td>129</td>
<td>47, 86</td>
</tr>
<tr>
<td>MelR</td>
<td><em>Escherichia coli</em></td>
<td>SP:Q56070</td>
<td>Similar to <em>E. coli</em> MarA</td>
<td>129</td>
<td>240</td>
</tr>
<tr>
<td>MmsR</td>
<td><em>Pseudomonas aeruginosa</em></td>
<td>SP:P28809</td>
<td>Transcriptional activator of the <em>mmsAB</em> operon for valine catabolism</td>
<td>307</td>
<td>236</td>
</tr>
<tr>
<td>MsrR</td>
<td><em>Streptococcus mutans</em></td>
<td>SP:Q00753</td>
<td>Transcriptional activator of the <em>msm</em> operon (<em>msmEFGK, aga, dexB, gftA</em>) required for the transport of melibiose, raffinose, and isomaltotriose and for melibiose, saccharose, and isomaltosaccharide catabolism</td>
<td>278</td>
<td>219</td>
</tr>
<tr>
<td>MxiE</td>
<td><em>Shigella flexneri</em></td>
<td>SP:Q04642</td>
<td>Transcriptional activator of the <em>mxi</em> and <em>spa</em> operons involved in the synthesis and secretion of the <em>Ipa</em> proteins required for the epithelial tissue invasion</td>
<td>210</td>
<td>3</td>
</tr>
<tr>
<td>MxiE</td>
<td><em>Shigella sonnei</em></td>
<td>SP:Q55292</td>
<td>Similar to <em>S. flexneri</em> MxiE</td>
<td>210</td>
<td>6</td>
</tr>
<tr>
<td>NitR</td>
<td><em>Rhodococcus rhodochrous</em></td>
<td>TE:P72312</td>
<td>Transcriptional activator of <em>nitA</em>, which codes for a nitrilase</td>
<td>319</td>
<td>137</td>
</tr>
<tr>
<td>OroR</td>
<td><em>Pseudomonas aeruginosa</em></td>
<td>TE:P72171</td>
<td>Probable transcriptional activator of the ornithine utilization operon</td>
<td>339</td>
<td>105</td>
</tr>
<tr>
<td>PcrR</td>
<td><em>Synechocystis</em> sp.</td>
<td>TE:P72600</td>
<td>Unknown</td>
<td>346</td>
<td>126</td>
</tr>
<tr>
<td>PchR</td>
<td><em>Synechocystis</em> sp.</td>
<td>TE:P72595</td>
<td>Unknown</td>
<td>326</td>
<td>126</td>
</tr>
<tr>
<td>PchR</td>
<td><em>Synechocystis</em> sp.</td>
<td>TE:P72608</td>
<td>Unknown</td>
<td>330</td>
<td>126</td>
</tr>
<tr>
<td>PerA</td>
<td><em>Escherichia coli</em></td>
<td>SP:Q43459</td>
<td>Transcriptional activator of the pyochelin and ferrisypochelin receptor</td>
<td>205</td>
<td>91</td>
</tr>
<tr>
<td>PobR</td>
<td><em>Pseudomonas aeruginosa</em></td>
<td>TE:Q51543</td>
<td>Probable transcriptional activator of <em>pobA</em>, which codes the p-hydroxybenzoate hydroxylase</td>
<td>288</td>
<td>68</td>
</tr>
<tr>
<td>PocR</td>
<td><em>Salmonella typhimurium</em></td>
<td>SP:Q05587</td>
<td>Transcriptional activator of <em>chiABCDEF-GHKLMNQOP</em> and <em>cobUST</em> operons, required for the adenosylcobalamine (vitamin B&lt;sub&gt;12&lt;/sub&gt;) synthesis, and <em>pduABC</em> and <em>pduF</em>, required for 1,2-propanediol catabolism. Also regulates its own synthesis</td>
<td>303</td>
<td>42, 218</td>
</tr>
<tr>
<td>PqRA</td>
<td><em>Proteus vulgaris</em></td>
<td>SP:Q52620</td>
<td>Probable transcriptional activator of genes and/or operons responsible of multidrug resistance</td>
<td>122</td>
<td>121</td>
</tr>
<tr>
<td>RfrR</td>
<td><em>Pediococcus pentosaceus</em></td>
<td>SP:Q43465</td>
<td>Transcriptional activator of the operon for rhamnose catabolism</td>
<td>277</td>
<td>147</td>
</tr>
<tr>
<td>RamA</td>
<td><em>Klebsiella pneumoniae</em></td>
<td>SP:Q48413</td>
<td>Probable transcriptional activator that confers multidrug resistance phenotype</td>
<td>113</td>
<td>90</td>
</tr>
<tr>
<td>RhaR</td>
<td><em>Escherichia coli</em></td>
<td>SP:P09378</td>
<td>Transcriptional activator of the operon <em>rhaSR</em> involved in the regulation of rhamnose catabolism</td>
<td>312</td>
<td>201, 243</td>
</tr>
<tr>
<td>RhaR</td>
<td><em>Salmonella typhimurium</em></td>
<td>SP:P40865</td>
<td>Similar to <em>E. coli</em> RhaR</td>
<td>106 (partial)</td>
<td>241</td>
</tr>
</tbody>
</table>

Continued on following page
<table>
<thead>
<tr>
<th>Protein</th>
<th>Microorganism</th>
<th>Accession no.</th>
<th>Function</th>
<th>No. of residues</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RhaS</td>
<td>Escherichia coli</td>
<td>SP:P09377</td>
<td>Transcriptional activator of genes required for the L-rhamnose catabolism (<em>rhaBAD</em>) and the genes which codify the rhamnose transporter (<em>rhaT</em>)</td>
<td>278</td>
<td>201, 243</td>
</tr>
<tr>
<td>RhaS</td>
<td>Salmonella typhimurium</td>
<td>SP:P27029</td>
<td>Similar to <em>E. coli</em> RhaS</td>
<td>277</td>
<td>187</td>
</tr>
<tr>
<td>Rns</td>
<td>Escherichia coli</td>
<td>SP:P16114</td>
<td>Transcriptional activator of the cosBACE operon, which codes the protein for CS1 or CS2 fimbriae in enterotoxigenic <em>E. coli</em> strains</td>
<td>265</td>
<td>30, 229</td>
</tr>
<tr>
<td>Rob</td>
<td>Escherichia coli</td>
<td>SP:P27292</td>
<td>Binds to the right arm of the replication origin oriC of the <em>E. coli</em> chromosome; also involved in resistance to antibiotics, heavy metals, and superoxide stress and in tolerance to organic solvents</td>
<td>289</td>
<td>106, 18, 151, 260</td>
</tr>
<tr>
<td>SoxS</td>
<td>Escherichia coli</td>
<td>SP:P22539</td>
<td>Transcriptional activator of the superoxide response regulon which includes at least 10 genes such as <em>acnA</em> (aconitase), <em>fpr</em> (NADPH-ferredoxin oxidoreductase), <em>fumC</em> (fumarase C), <em>inaA</em> (unknown), <em>micF</em> (an antisense inhibitor of ompF), <em>nfo</em> (endonuclease IV), <em>qpi-5</em> (unknown), <em>ribA</em> (GTP cyclohydrolase), <em>sodA</em> (Mn-superoxide dismutase), and <em>zwf</em> (glucose-6-phosphate dehydrogenase)</td>
<td>106</td>
<td>4, 18, 151, 260</td>
</tr>
<tr>
<td>SoxS</td>
<td>Salmonella typhimurium</td>
<td>SP:Q56143</td>
<td>Similar to <em>E. coli</em> SoxS</td>
<td>106</td>
<td>4, 18, 151, 260</td>
</tr>
<tr>
<td>TcpN</td>
<td>Vibrio cholerae</td>
<td>SP:P29492</td>
<td>Transcriptional activator of tcpABYCDZEF-MONlafBC, tcpI, tcpH, acfA, acfD, ctxAB operons required for epithelial tissue colonization</td>
<td>276</td>
<td>113, 130, 192</td>
</tr>
<tr>
<td>TetD</td>
<td>Tn10</td>
<td>SP:P28816</td>
<td>Unknown</td>
<td>138</td>
<td>22, 227</td>
</tr>
<tr>
<td>TheR</td>
<td>Rhodococcus sp.</td>
<td>SP:P43462</td>
<td>Transcriptional activator of the thc operon for the degradation of the thiocarbamate herbicide EPTC</td>
<td>332</td>
<td>178</td>
</tr>
<tr>
<td>UreR</td>
<td>Enterobacteriaceae</td>
<td>SP:P32326</td>
<td>Transcriptional activator of the ureDABCEFG operon for urease production; <em>P. stuartii</em> and <em>Salmonella</em> proteins are 98% identical to the <em>E. coli</em> protein</td>
<td>296</td>
<td>58, 185</td>
</tr>
<tr>
<td>V38K</td>
<td>Mycobacterium tuberculosis</td>
<td>SP:Q06861</td>
<td>Probable role in the regulation of proteins necessary for virulence</td>
<td>339</td>
<td>97</td>
</tr>
<tr>
<td>VirF</td>
<td>Shigella dysenteriae</td>
<td>SP:P04248</td>
<td>Transcriptional activator of the <em>virB</em> and <em>virG</em> genes. VirB is itself an activator of the <em>ipaABCD</em> virulence regulon; <em>S. flexneri</em> and <em>S. sonnei</em> proteins are identical to the <em>S. dysenteriae</em> protein</td>
<td>262</td>
<td>129, 220, 264</td>
</tr>
<tr>
<td>VirF</td>
<td>Shigella flexneri</td>
<td>SP:Q13225</td>
<td>Transcriptional activator of the <em>Yersinia</em> virulence regulon comprising <em>yop</em>, <em>ysc</em>, <em>yadA</em> and <em>yfaA</em> genes; the <em>Y. pseudotuberculosis</em> protein is 99% identical to the <em>Y. enterocolitica</em> protein</td>
<td>271</td>
<td>53</td>
</tr>
<tr>
<td>VirF</td>
<td>Shigella sonnei</td>
<td>SP:P33390</td>
<td>Transcriptional activator of the <em>Yersinia</em> virulence regulon comprising <em>yop</em>, <em>ysc</em>, <em>yadA</em> and <em>yfaA</em> genes; the <em>Y. pseudotuberculosis</em> protein is 99% identical to the <em>Y. enterocolitica</em> protein</td>
<td>392</td>
<td>231</td>
</tr>
<tr>
<td>XylR</td>
<td>Haemophilus influenzae</td>
<td>SP:P45043</td>
<td>Similar to <em>E. coli</em> XylR</td>
<td>387</td>
<td>74</td>
</tr>
<tr>
<td>XylR</td>
<td>Pseudomonas putida</td>
<td>SP:P07859</td>
<td>Transcriptional activator of the pWW0 plasmid <em>meta</em> operon (<em>xylXYZLTEGFJQKIH</em>), required for the degradation of benzoate and substituted derivatives</td>
<td>321</td>
<td>119, 169, 235</td>
</tr>
<tr>
<td>XylR</td>
<td>Pseudomonas putida</td>
<td>SP:P04710</td>
<td>Transcriptional activator of the pWW53 plasmid <em>meta</em> 1 and 2 operons for benzoate catabolism and substituted derivatives</td>
<td>321</td>
<td>10</td>
</tr>
</tbody>
</table>

*Continued on following page*
the stationary phase (AppY from *E. coli*) (12, 131). Some members of this group of proteins are highly homologous to each other, and some of them—SoxS, MarA, and Rob—cross-regulate certain genes (8, 48). These proteins apparently need each other, and some of them—SoxS, MarA, and Rob—cross.`

No specific regulatory function has yet been assigned to any of these regulators, although some have been shown to be overproduced to exert their regulatory role (9, 86, 181). A few have been found in low G+C gram-positive bacteria and in cyanobacteria, but none have been found in archaeabacteria or eukaryotes (194). However, because many prokaryotic genera have not been subjected to extensive genetic characterization, the observed distribution of AraC/XylS proteins may be nonrepresentative.

The large genetic distances between prokaryotes with AraC/XylS regulators and the vast differences in G+C content suggest that a progenitor arose early in prokaryotic evolution. The distribution of AraC/XylS proteins may be nonrepresentative. Because the conserved sequences within the members of the AraC/XylS are a series of well-established domains involved in DNA binding and stimulation of transcription, this family probably evolved through the recruitment of new domains of key importance in determining which function the regulator carries out. A phylogenetic tree in which no relationship between the branches and the function regulated by each subgroup is evident can be obtained upon request from M. T. Gallegos.

### Distribution and Evolution

Members of the AraC/XylS family are widely distributed in diverse prokaryote genera (Table 1). The G+C content of genes encoding AraC/XylS family members vary from 28% for *E. coli* rns (36) to at least 67% for Streptomyces *araL* (43, 265). Most of the genes encoding members of this family are in the genomes of the gamma subdivision of the proteobacteria (purple bacteria) (194). A few have been found in low G+C and high G+C gram-positive bacteria and in cyanobacteria, but none have been found in archaeabacteria or eukaryotes (194). However, because many prokaryotic genera have not been subjected to extensive genetic characterization, the observed distribution of AraC/XylS proteins may be nonrepresentative.

#### TABLE 1—Continued

<table>
<thead>
<tr>
<th>Protein</th>
<th>Microorganism</th>
<th>Accession no.</th>
<th>Function</th>
<th>No. of residues</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XylS2</td>
<td><em>Pseudomonas putida</em></td>
<td>SP-Q05092</td>
<td>Pseudogen present in pDK1 and pWW53 plasmids</td>
<td>157</td>
<td>10</td>
</tr>
<tr>
<td>XylS3</td>
<td><em>Pseudomonas putida</em></td>
<td>SP-Q05335</td>
<td>Transcriptional activator of the pWW53 plasmid meta 1 and 2 operons for benzoate catabolism and substituted derivatives</td>
<td>331</td>
<td>10</td>
</tr>
<tr>
<td>XylS4</td>
<td><em>Pseudomonas putida</em></td>
<td>SP-Q04713</td>
<td>Transcriptional activator of the pDK1 plasmid meta 1 and 2 operons for benzoate catabolism and substituted derivatives</td>
<td>331</td>
<td>10</td>
</tr>
</tbody>
</table>

* SP, SWISSPROT; TE, TREMBL; GP, Genpept.

**Domain Organization of AraC/XylS Polypeptides**

### Size and Location of the Conserved Domain in AraC/XylS Members

Most members of the AraC/XylS family of regulators are 250 to 300 residues long, although a few exceptions are found:
HprB from *Burkholderia solanacearum*, Ada from *M. tuberculosis*, Ybb from *B. subtilis*, and Hrp from *X. campestris* and *X. oryzae* are about 500 amino acids long (Table 1). A few proteins and hypothetical polypeptides were found to be particularly short (106 to 166 residues), e.g., AarP from *Providencia stuartii*, MarA and SoxS from *E. coli*, PqrA from *Proteus vulgaris*, RamA from *Klebsiella pneumoniae*, TetD from Tn10, and YgK from *Alteromonas carragennovora* (Table 1).

The region of greatest amino acid sequence homology identified in Yxls/AraC members is a nearly 100 residues with an overall similarity greater than 25% and that the region of similarity extends for a region of 27 proteins. Analyses of the structures and sequences of proteins present sequence is more accurate, because it includes 109 residues. Analyses of the structures and sequences of proteins have established that sequence homology greater than 25% is conserved in at least 60% of the aligned proteins. The sequence is similar but not identical to that previously proposed by Gallegos et al. (84) based on the alignment of 27 proteins. From a statistical point of view, the consensus for the family (A-----S---L---F----G----------R---A---

**Conserved Domain**

The alignment of the 99 amino acids that are highly conserved in the proteins of the AraC/Yxls family of regulators is shown in Fig. 1. By using Matrix Blocks45 (112), a histogram showing the degree of similarity at each position was obtained (Fig. 2). With a cutoff point of 0.5 for similarity, 17 residues showed a high degree of conservation and represent the consensus for the family (A-----S---L---F----G----------R---A---

This sequence was conserved in at least 60% of the aligned proteins. The sequence is similar but not identical to that previously proposed by Gallegos et al. (84) based on the alignment of 27 proteins. From a statistical point of view, the present sequence is more accurate, because it includes 109 proteins and extends for 75 amino acids within the stretch of 99 residues. Analyses of the structures and sequences of proteins have established that sequence homology greater than 25% between two proteins extending for 50 amino acids is sufficient to ensure their identical tertiary structure (221). Given that members of the AraC/Yxls family are transcriptional regulators and that the region of similarity extends for a region of nearly 100 amino acids with an overall similarity greater than 20%, these proteins can be assumed to possess identical tertiary structures in the conserved region. However, no tertiary structure for this domain is available, mainly because of the low solubility of the proteins of this family (64, 226).

Secondary-structure predictions were made with the entire alignment of the 99-amino-acid homologous segment by using the algorithm of Rost and Sander (217). This analysis suggested the existence of two potential α-helix-turn-α-helix (HTH) DNA binding motifs (23, 195–197, 257). In the XylS regulator, the first HTH motif is located at positions 228 to 251 and the second HTH motif is located at positions 281 to 305; these correspond to positions 195 to 218 and 245 to 270, respectively, in the AraC regulator (26–28, 169).

Evidence that the first HTH motif constitutes the DNA binding motif in AraC is based on the following findings. (i) Interference binding assays suggested that residues in the second α-helix of the motif made specific contacts with target DNA sequences at the *PargoBAD* promoter (27, 28). (ii) Mutations within residues in this region in AraC (Cys204→Tyr, Ser208→Ala, Arg210→Cys, and His212→Tyr or Ala) reduced binding to and decreased transcriptional activation from the *PargoBAD* promoter (26–28, 39, 78). The presence of two mutations in the XylS protein supports a role for these helices in promoter recognition: Ser229→Ile (the first amino acid of the first α-helix) and substitution of Cys for Phe248 (in the second α-helix) resulted in mutant regulators with increased affinity for target sequences and the ability to mediate transcription from the cognate Pm promoter constitutively (81, 83, 162, 267).

The second HTH motif has been proposed for all proteins in the family. This motif contains an extra amino acid in the turn with respect to canonical HTH DNA binding motifs. Its biochemical role is unknown. Mutations within these helices are available for some members of the family: the substitution of Ala and Asn for Ser271 and Arg272, respectively, has been achieved in MelR, and Val has been substituted for Asp288 in XylS. These mutants behaved similarly to the wild-type regulator (41, 208). In the case of AraC, the picture arising from the analysis of mutants with mutations in these helices (Gly249→Asp, Arg250→His, Gly253→Ser, Asp256→Ala, Gln257→Ala, Ser261→Ala, and Val264→Ile) is more complex, since certain mutants lost contact with multiple bases or bound to DNA in a pattern not fully consistent with a canonical HTH DNA binding motif (28, 39, 78).

It was recently suggested that AraC might contact target DNA sequences through the two HTH motifs (186). Although this might be the case for AraC (see below), it may not be a general rule for members of the family. This statement is based on comparisons of each of the HTH motifs of each member in the family with the corresponding aligned HTH motif in the rest of the family. Our results showed that sequence conservation at the HTH comprising the first HTH motif was low and that with certain pairs of sequences it was highly divergent. In contrast, no such variation was found when the second HTH motifs were compared (Fig. 2). We suggest that the variation in the first HTH motif represents the recognition of different target sequences at the cognate promoters by different regulators; conservation at the second HTH motif may thus represent a common function for all members of the family, e.g., contact with the transcriptional machinery. However, this hypothesis needs to be tested in vitro.

A small region of high sequence conservation was found outside the second HTH motif and toward the C-terminal end. Its most characteristic feature was the presence of a proline in more than 90% of the proteins in the family.

Given that AarP, MarA, PqrA, RamA, SoxS, and TetD, the shortest members of the Yxls family (106 to 166 amino acids long), consist mainly of the homologous segment, the stretch of conserved residues most probably contains all the domains necessary for these regulators to interact with target DNA sequences and RNA polymerase and thus activate transcription from target promoters. Furthermore, for regulators whose recognition site has been defined, the target sequences in the cognate promoters have been located adjacent to or overlapping the 35 region of the promoter, as is the case in other positively regulated promoters (33, 49, 122). This suggests that the mechanism of transcription activation by AraC/Yxls family members may involve direct interactions with RNA polymerase.

**Nonconserved Domain**

Data available for the nonconserved domain are scarce and basically limited to the AraC protein of the family *Enterobacteriaceae*; much less is known about the XylS and the other proteins. The nonhomologous N-terminal and central regions of the regulators recognizing chemical signals are presumed to contain binding sites for activator molecules that confer specificity (41, 172, 208, 232). Whether this information also holds for other members of the family is unknown.

The AraC protein, which regulates the l-arabinose operons in *E. coli*, consists of two domains that function in chimeric proteins. One provides the ability to form dimers (residues 1 to 170) and binds the ligand arabinose, and the other provides...
FIG. 1. Multiple alignment of proteins belonging to the AraC/XylS family. We excluded from the alignment those sequences found in closely related microorganisms and exhibiting a high degree of sequence conservation. Multiple alignments were found with the algorithm of Lüthy et al. (160). If the residue is identical to the defined consensus (see the text), it appears printed on a black background. If the residue is similar but not identical to the consensus, it appears on a gray background.
site-specific DNA binding capability and activates transcription (residues 178 to 286) (35, 143, 168). These domains are connected by a flexible linker (69). In vivo and in vitro experiments showed that a chimeric protein consisting of the N-terminal half of the AraC protein and the DNA binding domain of the LexA repressor dimerizes, binds to a LexA operator, and represses the expression of a LexA operator in an arabinose-responsive manner (35). This suggests that at least in the case of AraC, the ligand domain and the DNA binding domain are independent (35).

Conclusive evidence for effector binding and dimerization of AraC in this nonconserved domain is provided by the crystal structure of this domain in the presence of arabinose (232). This domain contains an eight-stranded antiparallel β-barrel with “jelly-roll” topology, followed by two turns of 3_10 helix, followed by a ninth β-strand that form part of one sheet of the barrel. Each monomer of AraC binds one molecule of α-L-arabinose. The sugar stacks against the indole ring of Trp-95 and is stabilized by direct hydrogen bonds with the side chains of Pro-8, Thr-24, Arg-38, Tyr-82, and His-92, as well as hydrogen bonds with water molecules in the binding pocket. The sugar binding site is completed by the N-terminal arm of the protein (residues 7 to 18), which loops around to close off the end of the β-barrel in which arabinose is bound.

AraC is a dimer in both the presence and the absence of arabinose (255). Crystallographic data for the N-terminal domain of AraC showed that the two monomers are associated by an antiparallel coiled coil formed between the terminal α-helix of each monomer, with each end of the coiled coil anchored by a triad of leucine residues that pack together in a knobs-into-holes manner.

Schleif’s group investigated whether any of the amino acids in the linker region between the nonconserved and the conserved domains play active, specific, and crucial structural roles or whether these amino acids merely serve as passive spacers between the functional domains. They found that all but one of the linker amino acids could be substituted by other amino acids individually and in small groups with no substantial effect on the ability of AraC protein to activate transcription when arabinose is present. However, when the entire linker region is replaced with linker sequences from other proteins, the functioning of AraC is impaired (69, 70).

MECHANISMS OF ACTION OF INDIVIDUAL FAMILY MEMBERS

The XylS Regulator Controls Expression from the Pm Promoter

The growth of P. putida (pWW0) on alkylbenzoates requires expression of the meta pathway operon, mediated by the XylS protein (79, 83, 118, 208). The xylS gene is expressed at low constitutive levels from a σ^70-dependent promoter called P2; on the addition of a meta-cleavage pathway substrate, expression from the Pm promoter occurs immediately, suggesting that the regulator becomes active after effector binding (82). The XylS protein is 321 amino acid residues long (119, 169, 235). The first two-thirds of the protein sequence, i.e., the
amino-terminal and central regions of the protein, seem to be involved in interactions with effectors (172, 208). Interactions between effector molecules and the regulator have been studied by analyzing XylS-dependent transcriptional activation from the Pm promoter in the presence of different benzoate analogs. These studies revealed that substituted benzoates are XylS effectors, although not all positions in the planar benzoate molecule are equivalent. For example, position 3 is highly permissible (-CH₃, -C₂H₅, and -OCH₃ groups and F, Cl, Br, and I atoms are permissible substituents), whereas positions 2 and 4 pose some restrictions to substituents (-CH₃, -F, and -Cl groups are allowed, whereas -C₂H₅ and -I are not) (210, 211). Although disubstitutions involving positions 2 and 3 and positions 3 and 4 are permissible, other combinations are usually nonpermissible, which suggests that interactions between the effector and the regulator are nonsymmetrical. Ramos et al. (208) and Michán et al. (172) isolated and sequenced a series of mutant regulators able to recognize substituted benzoate effectors that are not recognized by the wild-type regulator. Critical mutations were found to be clustered at positions 37 to 45. Arg41 seems to be a critical residue for interaction(s) with effectors, since changes at this position result in multiple different phenotypes. For example, XylSArrg41Gly is a mutant regulator that has lost the ability to recognize o- and p-methylbenzoate, although it remained activatable by m-methylbenzoate. Substitution of Arg41 with Leu resulted in a mutant unable to respond to benzoate effectors (172).

XylS mutants such as XylSArrg41Cys, XylISPro37Gly, XylSSer229Ile, XylISasp274Val, and XylISasp274Glu mediated transcription from Pm in the absence of effectors (172, 267). These results support the hypothesis that XylS exists in vivo in a dynamic equilibrium between an inactive and an active form with respect to transcriptional stimulation. Therefore, transition from the inactive to the active form may be mediated by effector binding. How the interaction between the inactive and active forms, and XylS leads to an active regulator is not yet understood, but regardless of the mechanism, the effector binding pocket and the DNA binding motif are not independent domains, as shown by intramolecular dominance of C-terminal mutations over N-terminal ones and by the reversal of this dominance in double mutants constructed in vitro (171).

Overproduction of XylS via a natural cascade regulatory system—involving expression from tandem Ps1 and Ps2 promoters (82; see reference 206 for a review)—or after expression from strong promoters (120, 169, 207, 234) leads to stimulation of transcription from Pm in the absence of effectors (172, 267). These results support the hypothesis that XylS exists in vivo in a dynamic equilibrium between an inactive and an active form with respect to transcriptional stimulation. Therefore, transition from the inactive to the active form may be mediated by effector binding. How the interaction between the XylS molecules and XylS leads to an active regulator is not yet understood, but regardless of the mechanism, the effector binding pocket and the DNA binding motif are not independent domains, as shown by intramolecular dominance of C-terminal mutations over N-terminal ones and by the reversal of this dominance in double mutants constructed in vitro (171).

Stimulation of transcription from the Pm promoter requires a DNA sequence extending to about 80 bp upstream of the transcription initiation point (83, 132, 207). In the architecture of the Pm promoter, two regions can be distinguished on the basis of genetic data: the XylS interaction region, which extends from about 10 bp to about 55 bp, and the region between about 60 and +1 for RNA polymerase recognition, which exhibits atypical –35 and –10 DNA sequences. XylS-dependent transcription from Pm can be mediated by RNA polymerase with either σ⁷⁰ or σ⁷⁷ (163).

Gallegos et al. (83) and González (92) have studied in detail the organization of XylS binding sites in the Pm promoter. They generated a series of 5′ sequential deletions and a large series of point mutations in the promoter and analyzed transcription from the resulting mutant promoters mediated by the wild-type XylS protein and by mutant XylS regulators that were constitutive. It was found that Pm promoter variants deleted up to –60 could be activated by constitutive XylS mutants (but not by the wild-type regulator) and that extension of the deletion to –51 prevented transcription. On the basis of sequence analyses, it was proposed that the XylS binding site was probably represented by the motif T(C/A)CAN,TGCA, which appears twice, such that the exact location of the RNA polymerase binding site proximal motif was between –46 and –47 and the distal motif was between –67 and –78 (82). The –46 to –57 proximal site constitutes the minimum sequence required for transcription stimulation. Point mutations suggest that the TGCA submotif may be the primary recognition site, with the remaining sequences contributing to overall affinity (92, 132).

Kaldalu et al. (125) reported the immunopurification of a functionally active XylS protein bearing a hemagglutinin epitope fused at its N terminus (N-XylS). This N-XylS variant was able to specifically bind and retain a DNA fragment bearing the proposed XylS binding region in Pm. A set of footprinting experiments indicated that N-XylS binds along one side of the DNA, covering four helix turns (from –28 to –72) and making base-specific contacts in four adjacent major groove regions on the same helix face. This footprinting extended beyond the site defined by genetic means; as in other members of the family (28, 37, 59, 71, 72, 110, 152, 250), this may reflect oligomerization of N-XylS after recognition of a primary binding site. Further in vitro studies with purified RNA polymerase and XylS are needed to determine whether the binding sites for each protein overlap. The observation that overproduction of the regulator is sufficient to activate Pm in vivo in the absence of effector (120, 169, 207, 234) supports the hypothesis that effectors increase the cellular concentration of XylS in its active conformation (XylS may exhibit higher affinity for its target DNA sequence) at the DNA target site.

**Arabinose Metabolism in E. coli**

Four transcriptional units in *E. coli* are involved in the utilization of L-arabinose: *araBAD*, which encodes three enzymes responsible for L-arabinose catabolism (65); *araE* and *araFGH*, which encode proteins responsible for low-affinity and high-affinity transport of L-arabinose (25, 239); and the regulatory gene *araC*, which encodes a protein that controls the expression of these genes as well as autoregulating its own synthesis (38, 66, 67, 95, 256).

The AraC protein is predominantly a dimer in solution (35, 168, 255). In the absence of arabinose, AraC protein represses expression of the *araBAD* and *araC* promoters (called P*araBAD* and P*araC*, respectively) (62, 99, 114, 117, 144–146, 156, 157, 165). With arabinose, AraC activates transcription from the promoters of the catabolic operons (Fig. 3). The response of the wild-type *ara* operons to arabinose was found to occur within 3 s of inducer addition (114), and mRNA was detected within 15 to 30 s (114, 123). Expression from all four of these promoters is also regulated by the cyclic AMP-catabolite activator protein (99). AraC protein interactions with the *ara* promoters were determined by chemical interference assays and by mutagenesis of the protein and the promoters (26–28, 37, 40, 101, 109, 144, 156, 168). A consensus sequence for AraC binding was obtained by comparing the sites from *E. coli* and *S. typhimurium ara* promoters (28, 108–110, 158). An AGCN,TCCATA sequence is conserved in all sites and appears as a tandem repeat (Fig. 4). The araO₁ site, which is needed for inhibition of transcription at P*araBAD* (see below), is apparently only half of a site (Fig. 4).
Regulation of the araBAD (P_{araBAD}) and araC (P_{araC}) promoters by the AraC protein has been extensively characterized (37, 156). In the absence of arabinose, one monomer of the AraC dimer occupies the araI site while the other occupies a half-site approximately 200 bp away, known as araO2 (Fig. 3). The dimer bound to target sequences in this way generates a DNA loop, which prevents transcription from P_{araBAD} and P_{araC} (37, 156, 157). When arabinose is added, the AraC protein undergoes a conformational change and shifts to occupy the adjacent half-sites, araI and araI2 (156, 157). As a result, P_{araBAD} is induced. Therefore, the main consequence of arabinose binding on AraC protein is to change the affinity of AraC for different spatial arrangements of half-sites. In the absence of arabinose, AraC favors binding to half-sites separated by more than one helical turn of DNA (Fig. 3), whereas in the presence of arabinose, AraC favors binding to half-sites separated by less than one helical turn of DNA (37). Therefore, arabinose destabilizes AraC protein binding to the I1-O2 looped complex but stabilizes binding to the I1-I2 site. Furthermore, because the loop is disorganized, free access of RNA polymerase to the P_{araC} promoter is transitorily facilitated and transcription increases. Subsequently, P_{araC} shuts down as a result of AraC protein binding to the araO1 site, which blocks the access of RNA polymerase to the P_{araC} promoter (225).

It was shown that to activate transcription in P_{araBAD}, the AraC protein binding site must overlap the −35 region of the promoter by 4 bp (212). AraC protein was located on one side of four adjacent helix-turn regions of the DNA, and there is evidence that each AraC monomer requires two direct repeats in successive turns of the DNA helix for binding (37, 110, 158). In light of the strict spacing and orientation requirements for AraC activation, interactions between AraC and RNA polymerase are likely to be specific and inflexible. Providing further support for this theory is the almost identical arrangement of the protein binding sites for araBAD and araE. Surprisingly, the araFGH promoter (P_{araFGH}) possesses a radically different structure. In P_{araFGH}, the catabolite activator protein site, rather than the AraC site, overlaps the −35 recognition sequence of RNA polymerase. In addition, the AraC sites in araFGH are arranged in the opposite direct-repeat orientation (108).

Niland et al. (186) systematically substituted every base pair in a synthetic 17-bp araI target (5-TAGCATTTTTTAACC TA-3' [the underlined bases correspond to those conserved in the consensus]) with each of the three possible alternatives and then used qualitative gel shift analysis to test the binding of AraC to these 51 DNA targets in the presence of L-arabinose. They found that every substitution of the underlined bases reduced AraC binding to 1/10 or less whereas substitutions at other bases had little or no effect. In the absence of L-arabinose, the binding of AraC to araI was reduced to one-sixth or less.

Two possible HTH motifs were proposed in the C-terminal domain of AraC (78), but contact to DNA was demonstrated only for the first (27, 28). This first motif binds the first major groove of the DNA. These results were confirmed by Niland et al. (186). The finding of Niland et al. (186) with AraC mutant Asp256 → Ala (in the second helix of the second HTH motif) provided evidence that the second HTH contacts the second major groove.

**SoxS Regulator and Sox-Box**

Redox cycling compounds such as paraquat and menadione are a continuous source of superoxide in the cell as a consequence of repeated cycles of oxidation and reduction. Exposure of *E. coli* cells to these compounds induces the synthesis of about 40 proteins (93). A subset of these proteins are produced by a regulon controlled by two genes, soxR and soxS, which constitute the so-called soxRS regulon (4, 190, 260). The following genes are known to be members of this regulon: acnA (aconitate), fpr (NADPH:ferredoxin oxidoreductase), fumC (fumarase C), imaA (function unknown), mifC (antisense regulator of ompF), nfo (DNA repair enzyme endonuclease IV), pgi-5 (function unknown), rhaA (GTP cyclohydrolase II), sodA (manganese superoxide dismutase), soi-17 (function unknown), soi-28 (function unknown), and zwf (glucose-6-phosphate dehydrogenase) (44, 94–96, 134–136, 154, 155, 173, 216, 246). Both SoxR (17 kDa) and SoxS (13 kDa) are DNA binding proteins. Induction of the soxRS regulon occurs in two steps. An intracellular signal of oxidative stress (reduction in the cellular NADPH/NADP⁺ ratio or exposure to superoxide) converts preexisting SoxR protein into a transcriptional activator of the soxS gene. The overproduced SoxS protein in turn activates the transcription of target genes of the regulon (4, 87, 190, 191, 260).

In vitro studies have demonstrated that purified SoxS and...
MalE-SoxS fusion proteins activate transcription from the promoter of target genes and can specifically bind and form multiple DNA-protein complexes thanks to the presence of multiple binding sites at cognate promoters (71, 72, 151, 152). DNase I footprinting assays have shown that promoters whose transcription is activated by SoxS seem to fall into two classes with respect to the location of the proximal site relative to the −35 hexamers of the promoters (Fig. 5). In one class, the primary protected region completely covers the −35 hexamers of the micF, nfo, P1-psiA, and sodA promoters, whereas it is adjacent to or only partially overlaps the −35 hexamers of the funC and zwf promoters (Fig. 5). ribA seems to be an exception, since the putative SoxS binding site is located from −146 to −118, far upstream from the −35 element (134). The SoxS distal sites at the micF and zwf promoters (Fig. 5) have been characterized by a combination of DNase I footprinting and methylation interference assays (71, 72, 151, 152). The alignment of the protected regions (Fig. 5) revealed a “Sox-box” consensus whose sequence is ANNGCAPyNAPuANNNPuNN, where N is any base, Py is a pyrimidine, and Pu is a purine (72).

A potentially important feature of the 19-bp consensus sequence is the GCAPy motif that lies near the 5′ end. This short sequence is conserved among the proximal and distal sites of funC, micF, nfo, P1-psiA, sodA, and zwf. Therefore, the GCAPy motif may be a primary recognition element for SoxS, with the remaining positions of the SoxS box sequence contributing to the overall affinity. The dissociation constant for chimeric MalE-SoxS binding to DNA sequences that contain this element is about 10−8 to 10−9 M. This relatively weak interaction suggests that additional free energy for binding might come from cooperative interactions with either RNA polymerase or a second SoxS molecule.

The importance of the GCAPy motif is also substantiated by the properties of several sodA mutants. Naik and Hassan (179) and Compan and Touati (50) described sodA mutants that do not respond to superoxide stress. In one mutant, the 5′-GCAT-3′ sequence, which lies within the proximal protected site of sodA, was changed to 5′-TACG-3′; in another mutant, the sequence was deleted. Presumably the inducible phenotype of these mutants was derived from the destruction of this GCAT sequence. Furthermore, single base pair substitutions at any position in the GCAY motif greatly reduced SoxS binding to synthetic oligonucleotides bearing the micF-proximal site (152).

MalE-SoxS fusion proteins activate transcription from the promoter of target genes and can specifically bind and form multiple DNA-protein complexes thanks to the presence of multiple binding sites at cognate promoters (71, 72, 151, 152). DNase I footprinting assays have shown that promoters whose transcription is activated by SoxS seem to fall into two classes with respect to the location of the proximal site relative to the −35 hexamers of the promoters (Fig. 5). In one class, the primary protected region completely covers the −35 hexamers of the micF, nfo, P1-psiA, and sodA promoters, whereas it is adjacent to or only partially overlaps the −35 hexamers of the funC and zwf promoters (Fig. 5). ribA seems to be an exception, since the putative SoxS binding site is located from −146 to −118, far upstream from the −35 element (134). The SoxS distal sites at the micF and zwf promoters (Fig. 5) have been characterized by a combination of DNase I footprinting and methylation interference assays (71, 72, 151, 152). The alignment of the protected regions (Fig. 5) revealed a “Sox-box” consensus whose sequence is ANNGCAPyNAPuANNNPuNN, where N is any base, Py is a pyrimidine, and Pu is a purine (72).

A potentially important feature of the 19-bp consensus sequence is the GCAPy motif that lies near the 5′ end. This short sequence is conserved among the proximal and distal sites of funC, micF, nfo, P1-psiA, sodA, and zwf. Therefore, the GCAPy motif may be a primary recognition element for SoxS, with the remaining positions of the SoxS box sequence contributing to the overall affinity. The dissociation constant for chimeric MalE-SoxS binding to DNA sequences that contain this element is about 10−8 to 10−9 M. This relatively weak interaction suggests that additional free energy for binding might come from cooperative interactions with either RNA polymerase or a second SoxS molecule.

The importance of the GCAPy motif is also substantiated by the properties of several sodA mutants. Naik and Hassan (179) and Compan and Touati (50) described sodA mutants that do not respond to superoxide stress. In one mutant, the 5′-GCAT-3′ sequence, which lies within the proximal protected site of sodA, was changed to 5′-TACG-3′; in another mutant, the sequence was deleted. Presumably the inducible phenotype of these mutants was derived from the destruction of this GCAT sequence. Furthermore, single base pair substitutions at any position in the GCAY motif greatly reduced SoxS binding to synthetic oligonucleotides bearing the micF-proximal site (152).
revealed a consensus sequence, aACCCgWWcWcGG (where W is A or T), which appears twice in each promoter. No clear symmetry was found in this sequence, although the nature of the highly conserved residues suggested that the binding sites are recognized as two direct repeats (73).

RhaS is one of the regulators of rhamnose metabolism in E. coli. RhaS activates transcription from rhaBAD, and transcription from the rhaS gene is controlled by RhaR (63, 244). Both RhaR and RhaS bind rhamnose to stimulate transcription from the corresponding cognate promoters. Full transcription from the rhaBAD promoter requires CRP (catabolite repression protein). Deletion analysis at the promoter of the rhaBAD operon revealed the requirement for a stretch of about 80 bp upstream from the main transcription initiation point. The CRP binding site was located adjacent to this sequence and was centered at bp −92.5 (63, 243, 244). By DNase footprinting and mutational analysis, it has been shown that RhaS binds in P_rhaBAD to an inverted repeat of two 17-bp half-sites separated by 16 bp. These findings were made possible by the discovery that the normally insoluble RhaS protein could be renatured in active form by the slow removal of urea while in the presence of DNA. This technique will probably prove useful in the study of the AraC/XylS family members (243).

The MeiR regulator controls transcription from the melAB operon promoter. The melAB operon encodes proteins essential for melibiose metabolism in E. coli. Transcription initiation from P_melAB is stimulated by the MeiR regulator with melibiose. Scratyly of the nucleotide sequence at this promoter revealed a relatively well-conserved σ^70 – 10 box and an unconserved −35 region. Upstream in the P_melAB promoter, two identical 18-bp elements are organized as an inverted repeat from positions −109 to −92 and from positions −54 to −71; these are the MeiR binding sites (41, 252).

**CONCLUDING REMARKS AND FUTURE PERSPECTIVES**

Profile PROSITE PS01124 defines for AraC/XylS family members a matrix that has been established on the basis of successive searches in protein and nucleic acid databases. More than 100 proteins and polypeptides deduced from ORFs have been included in the family. The matrix assigns to these proteins and polypeptides a normalized score equal to or higher than 12.52. Once a more exhaustive analysis of proteins assigned to the AraC/XylS family on the basis of PROSITE profile PS01124 will increase significantly.

Multiple alignments of the proteins in this family revealed a stretch of 99 conserved amino acids. This conserved domain comprises all functions required for DNA binding and RNA polymerase contact and stimulation of transcription. Secondary-structure analysis predicts the presence of two potential HTH structures. The set formed by the first HTH motif seems to be the actual DNA binding domain of the members of the family; however, the possibility that the other HTH domain also functions as a DNA binding domain cannot be ruled out. Exhaustive analyses of mutations in different members of the family are needed to further define the role of these HTH. Furthermore, efforts to crystallize this stretch are needed to determine the actual tertiary structure of the members of this family.

One of the striking features of the AraC/XylS family is the paucity of biochemical data. This reflects the difficulty of handling these proteins. Most of them are highly insoluble and are thus difficult to purify. Because several members of the family possess this property and because the dimerization domain of AraC is soluble (232), it seems that it is the DNA binding domain which makes these proteins poorly soluble (64). Efforts to improve the solubility of this domain are essential to facilitate purification and crystallization.

The conserved domain is usually connected to a nonconserved domain via a linker. The nonconserved domain is critical for signal recognition in members of the family activated by effector binding. However, it is not known how the linker transfers a signal from the signal reception site to the DNA binding site or how the active regulator interacts with RNA polymerase to drive transcription from cognate promoters. The role of the nonconserved domain in proteins involved in pathogenesis is an area that deserves particular attention, since practically no data are available.

No general conclusions can be drawn regarding the promoters regulated by members of the family. However, it has been found that these promoters usually contain more than one binding site for the regulator. Many sites for which the regulator has high and low affinity have been identified. The site proximal to the RNA polymerase binding site has been found in most cases to overlap or abut the −35 region of the promoter, but cases exist in which sites are located at about 100 bp from the main transcription initiation point. Whether this reflects the possibility that different members of the family contact RNA polymerase in different ways is unknown (11, 24, 142, 163, 188, 247). Another feature is the organization of the binding sites. It has been suggested that for some promoters the binding sites are organized as inverted repeats whereas for others they are organized as direct repeats. However, few symmetry studies are available, and this deserves attention.

It should be noted that in spite of the high homology among AraC/XylS members, transcription stimulation mediated by these proteins from the corresponding promoters shows interesting diversity. In addition to the specific regulator, transcription from certain promoters regulated by members of this family requires other proteins for maximal activity (e.g., the CRP in the P_rhaBAD promoter), or histone-like proteins that act as negative regulators (e.g., YomA in the VirF-regulated P_ysop promoters) (52, 124, 242). This is clear evidence that the expression of genes controlled by members of this family is integrated in overall cellular control.

**ACKNOWLEDGMENTS**

Work in Granada was supported by a grant from the CICYT (AMB 1038-C02-1) and a grant from the Biotechnology Program of the EC (BIO4-CT97-2270). Work in Baltimore was supported by a grant from NIH (GM18277).

We thank Karen Shashok for linguistic improvement of the manuscript, Maxi Manzanera for assistance with artwork, and Javier González for help with the computers. We are indebted to M. M. Fandila for preparing the manuscript.

**REFERENCES**


Demple, B., S. E. F. V. Thomas, and C. M. Collins. 1996. Activation of transcription at divergent uaa-depressed promoters by the uaae gene

Downloaded from mbr.asm.org at Bibliothèque Faculte Medecine Genève on August 27, 2009


protein at 2.9 Å resolution suggests binding to left-handed DNA-Nature 290:744–749.


