318–321 Nucleic Acids Research, 2002, Vol. 30, No. 1

AraC-XylS database: a family of positive transcriptional regulators in bacteria

Raquel Tobes and Juan L. Ramos*

Department of Biochemistry and Molecular and Cellular Biology of Plants, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, 18008 Granada, Spain

Received July 27, 2001; Revised and Accepted October 2, 2001

ABSTRACT

The AraC-XylS database contains information about a family of positive transcriptional regulators broadly distributed in bacteria. This specific database focuses on protein sequences and on the biological and functional features of each of the proteins that belong to this family. Each entry provides information on the protein itself, the annotated protein sequence and, when the crystal is available, a comprehensive representation of its three-dimensional structure. The organization of the database is based on an exhaustive analysis of the scientific literature. The data are interconnected and linked with other databases. Multiple alignments of the members of the family, an extensive collection of references and a tutorial about the family provide additional information. The AraC-XylS database is accessible on the World Wide Web at http://www.AraC-XylS.org.

INTRODUCTION

During the last few years, notable developments in molecular and computational biology have yielded a huge amount of data related to protein and gene sequences. This raises the need to organize and represent the data in a manner that facilitates access to the information from different research fields. In addition, there is an ever increasing interest in obtaining comprehensive, structured knowledge about the function of biological molecules. Specific databases that gather information about a particular protein family can undoubtedly help increase knowledge in a research area where there is an obvious imbalance between the data and knowledge. Furthermore, an insight into a protein family broadly distributed among many organisms opens the possibility of designing comparative studies and further experiments that could lead to the unequivocal assignment of functions to different domains of these proteins.

New technologies, especially those related to DNA arrays, have emphasized the importance of transcriptional regulation in protein expression. Consequently, one of the next major objectives of post-genomics is to clarify gene regulatory networks. In bacteria, the adaptation to a changing environment is essentially mediated by systems that regulate gene expression; hence, it is especially important to trace the interconnections between these networks. Most of the available databases dealing with bacteria [RegulonDB (1) and PromEC (2) of *Escherichia coli*, DBTS (3) for *Bacillus subtilis* and TIGR CMR (4)] focus on a specific organism. In contrast, databases for a specific family of bacterial transcriptional regulators are not available.

We have developed a specific database for the AraC-XylS family of regulators. The profile that defines the AraC-XylS family of transcriptional regulators was generated by Gallegos *et al.* (5), based on previous studies by Ramos *et al.* (6) and Gallegos *et al.* (7), who had identified a segment of 99 conserved amino acids at the C-terminal end of the first 27 proteins recognized as members of the family. With the new profile the family was extended to almost 100 members. Analyses of a protein sequence with the help of a matrix defining the family profile assigned a value to the query sequence. The value given by the matrix to each of the family members was between 30.74 and 12.52, with small variations between two consecutive proteins identified as members. However, a difference of at least 4.7 points was observed between the last member of the family and the closest value of a protein not identified as a member of the family. Gallegos *et al.* (5) proposed that a protein belongs to the AraC-XylS family if the value after analysis with the PS01124 profile defined in PROSITE is above 12.52. Most members of the AraC-XylS proteins are positive transcriptional regulators involved in the control of many important processes related to carbon metabolism, stress responses and pathogenesis (reviewed in 5). The AraC-XylS database is built around a family of proteins that are broadly distributed in Gram-negative and Gram-positive bacteria. Members of the AraC-XylS family belong to an ancient lineage, as deduced from the great evolutionary distance between prokaryotes with AraC-XylS proteins, and from the existence of marked differences in their G+C content. The proteins belonging to this family have been found in 47 different genera and 84 different microbial species so far.

In broad terms, proteins of the family consist of two domains, a non-conserved one, which seems to be involved in effector/signal recognition and dimerization, and a conserved one, characterized by significant amino acid sequence homology extending over a 100 residue stretch containing the DNA-binding domain of the family members (5). The first high-resolution structures for the DNA-binding domain of AraC-XylS family proteins have recently become available. X-ray diffraction analyses of the co-crystals of MarA, a 109 amino acid protein that contains only the DNA-binding domain with its *mar* promoter binding site, shows that MarA binds as a

*To whom correspondence should be addressed. Tel: +34 958 121011; Fax: +34 958 129600; Email: jlramos@eez.csic.es

monomer with two helix–turn–helix (HTH) motifs inserted in two adjacent major groove segments (8). The recognition helices of the HTH motifs are held in place by a rigid 27 Å long α-helix, which is shorter than the 34 Å that separate the two major grooves that induce a 35° bend in the DNA. In addition to phosphate backbone contacts, the most important determinants of binding appear to be H-bonds made by both HTH motifs with several bases of the DNA.

The co-crystal structure of Rob, another member of the family, and its target *micF* promoter have been resolved. Because the two HTH motifs of Rob are superimposable on the MarA structure (9), this way of recognizing DNA may be common to all members of the AraC-XylS family of regulators. Furthermore, most AraC-XylS members may share a common mechanism to enhance transcription, i.e. they may facilitate recruitment of RNA polymerase and isomerization to open complexes, as demonstrated for AraC and SoxS proteins (10–12).

The α-CTD end of RNA polymerase is required at several promoters that are activated by AraC-XylS family members, such as AraC, MelR, MarA, RhaS, Rob, XylS and SoxS (13,14). However, direct interaction with RNA polymerase, as opposed to interaction with DNA UP elements or factors such as CRP, has not been demonstrated conclusively.

Taking into account the functions that the AraC-XylS family of transcriptional proteins regulates, it is reasonable to consider these proteins as possible targets for designing a broad spectrum of applications in the fields of biotechnology and medicine. Thus, their participation in the regulation of the use of C-sources and in the metabolism of recalcitrant pollutants establishes a relationship between this family and the catabolism of biogenic and xenobiotic compounds. The involvement of AraC-XylS proteins in antibiotic resistance and in pathogenesis of common diseases such as enterocolitis, respiratory and urinary infections, emphasizes the importance of these regulators as targets for new drugs. In parallel, their role in abiotic stress responses establishes a link with the ecological adaptation of microbes to changing environmental conditions (15).

DATABASE OVERVIEW

The AraC-XylS database includes all the proteins that fulfil two requirements: (i) the value obtained after matching with the PROSITE profile PS01124 must be above the threshold (12.52) set to define the family; and (ii) the protein must be included in SWISS-PROT or TrEMBL databases. One entry is considered different to another if its sequence is different or if it is present in a different organism. With these criteria, the database includes 280 entries at the time of writing.

The database is structured to cover general knowledge and sequence information.

Knowledge database

The information on each protein is organized into 25 fields that contain data related to the protein, its corresponding gene, the genes that it regulates, its function and pathogenicity, the three-dimensional structure, published mutation studies and bibliographic references.

ID Identification number in the database.
NA Name.
AN Accession number in SWISS-PROT or TrEMBL.
OR Organism.
DA Annotation update.
GE Name of the gene encoding the family member.
OP Orientation and position in chromosome or plasmid.
PP Name, location and characteristics of the promoter of the gene encoding the family member.
RE Regulation of the expression of the regulatory protein.
TR Type of regulation: activation or repression.
GR Genes regulated by the family member and their functions.
PR Name, location and characteristics of the promoters of the regulated genes.
RG Other proteins or regulating mechanisms that also participate in the regulation of the genes regulated by the family member.
IN Interaction with RNA polymerase subunits.
EF Effector: molecules and conditions (temperature, pH, etc.) that trigger the regulatory protein.
FU Function: overview of the bacterial function in which the regulator is involved.
PA Involvement of the family member in pathogenicity.
MU Published data about mutations.
ST Three-dimensional structure.
DO Functions associated with the different protein domains.
OL Oligomerization.
VA Value obtained with the PROSITE PS01124 profile.
SI Similarity with other members of the family: BLAST restricted to this family.
CO Comments.
RF References.

The knowledge data were gathered in a database management system to help sustain the database. The basic sources of information that constitute this knowledge database were bibliographic references that were specifically cited for each protein and information available in databases (SWISS-PROT, TrEMBL, PROSITE, INTERPRO, bacterial databases, etc.). A direct link to MEDLINE makes it possible to immediately retrieve published articles that are related to a given member of the family. In addition, direct links to SWISS-PROT and TrEMBL make it possible to retrieve information that has been specifically stored in these databases.

Sequence database

Each entry has four sequences associated with it.

1. The complete protein sequence.
2. The conserved region. The length of this region is approximately 100 amino acid residues and the limits are defined by alignment with the profile. It contains the DNA binding domain and critical features essential for transcription activation.
3. The C-terminal region with respect to the conserved region. This fragment is often short because the conserved region is usually located at the C-terminal portion of the protein.
4. The N-terminal region with respect to the conserved region. This region usually contains the dimerization, the effector binding pocket and the signal transmission domains.

We provide the sequences as independent files and as global files available at www.AraC-XylS.org. Each global file contains one type of associated sequence for all the entries. This format facilitates the use of the sequences for theoretical studies.

In AraC-XylS proteins, the different domains can be totally independent from a functional and evolutionary point of view. Consequently, we detected some members of the family that bear, in addition to the AraC-XylS conserved domain, a region corresponding to a receiver module of a response regulator of bacterial sensory transduction systems. We have adapted the sequence database to the modular structure of the proteins to facilitate specific theoretical studies to achieve insights on the modular evolution of the different regulators. A multiple alignment (CLUSTAL) is given for the conserved domain of all the database proteins (www.AraC-XylS.org). In the alignment each sequence can be 'clicked on' to display information about this protein entry.

GRAPHICAL USER INTERFACE

We have designed a web interface that provides an open, interconnected representation of the data included in the AraC-XylS database. Access to the proteins in the database is facilitated by three different indexes: the alphabetical order index, the organism index and the numerical index. The multiple alignment of the conserved domain sequences is also a useful index to access protein information. Finally, a search engine to find a particular protein by its name has also been included.

As an introduction to the database, we provide general tutorial information about the features of this protein family for researchers from different scientific fields who may consult the AraC-Xyls family web page.

The data are interconnected and each protein entry has both external and intra-database links. Within the latter are links that join each protein entry with its associated sequences, with the most similar protein (BLAST) of the family, and with the complete results of a BLAST comparison with all family members. Each entry also includes a link with a representation of the crystal structure if the molecule has been crystallized. We provide a didactic description of the crystal to provide a rapid, comprehensive view of the three-dimensional structure of the molecule. In the three-dimensional representation of the molecule, we have added labels with explanations to indicate the different domains, the most critical residues for interactions, the residues related to mutations and other information of interest. This didactic presentation enables scientists and even non-expert scientists to rapidly grasp the information available on the molecule.

Many fields of each protein entry are hyperlinked to related information provided in other databases. Thus, the AN field (accession number) is linked to the corresponding entry in SWISS-PROT or TrEMBL, the field OR (organism) is linked to the database of the genome of the microorganism if available, the ST field (three-dimensional structure) is linked to the PDB entry and the VA field (profile value) is linked to the corresponding profile in PROSITE. Each reference is linked to its PubMed abstract.

UTILITY AND FUTURE PERSPECTIVES

The AraC-XylS database supplies information about a specific family of transcriptional regulators and helps maintain the right balance between data and knowledge. We are confident that this database will be a valuable tool for experimental researchers and for the design of theoretical studies by bioinformaticians.

The increasing number and diversity of protein sequences require methods to predict details regarding functions. The broad distribution of proteins of the AraC-XylS family makes it possible to hypothesize the function of new proteins on the basis of sequence similarity studies.

We expect the number of new entries in this database to increase in parallel with the analysis of available genomes. *Escherichia coli* contains 38 proteins of the AraC-XylS family. In the unfinished (~6 Mb) genome of *Pseudomonas putida* KT2440 (16), we have found 41 chromosomal open reading frames corresponding to proteins that could be included in the AraC-XylS protein family. Note that only two out of these 41 proteins have been deposited previously in the SWISS-PROT database, and therefore only these two proteins are recognized at present as members of the AraC-XylS family. If we assume the number of regulators of the AraC-XylS family to be approximately 40, and if we take into account the number of microorganisms whose genomes are currently being sequenced, together with the wide distribution of members of this family in different genera of Gram-positive and Gram-negative bacteria, a considerable increase in the number of AraC-XylS proteins can be expected in the very near future. We will update the database by frequent additions of new family members. We anticipate an increase in the number of orthologs of each protein family, and a new organizational structure will then be incorporated to the database.

The body of information on the promoters that are recognized by the AraC-XylS proteins is restricted to about 20 members of the family. We expect to enlarge the present database with the DNA sequences recognized by each protein, and are currently searching for the characteristic pattern that defines sequences recognized by these regulators. We expect to find the distinguishing features for the promoters that make it possible to predict the location of the regulator binding site with respect to that of the RNA polymerase. This information will make it possible to predict interactions with the different subunits of RNA polymerase.

We also envisage a project in which thorough reviews of the most important proteins of the family are added to the database. These reviews will include abundant graphical information, schematic drawings and structured knowledge.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Comisión Interministerial de Ciencia y Tecnología BIO 2000-0964 and a grant from the European Commission QLTR-2000-00170. We thank Carmen Lorente and Karen Shashok for checking the English of the manuscript.

REFERENCES

1. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millán-Zarate,D., Díaz-Peredo,E., Sánchez-Solano,F., Pérez-Rueda,E., Bonavides-Martínez,C. and Collado-Vides,J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.
2. Hershberg,R., Bejerano,G., Santos-Zavaleta,A. and Magalit,H. (2001) PromEC: an updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Res.*, **29**, 277–280.

3. Takahiro,I., Yosida,K.-I., Terai,G., Fujita,Y. and Nakai,K. (2001) DBTBS: a database of *Bacillus subtilis* promoters and transcription factors. *Nucleic Acids Res.*, **29**, 278–280.
4. Peterson,J.D., Umayam,L.A., Dickison,T., Hickey,E.K. and White,O. (2001) The comprehensive microbial resource. *Nucleic Acids Res.*, **29**, 123–125.
5. Gallegos,M.T., Schleif,R., Bairoch,A., Hofmann,K. and Ramos,J.L. (1997) AraC/XylS family of transcriptional regulators. *Microb. Mol. Biol. Rev.*, **61**, 393–410.
6. Ramos,J.L., Rojo,F., Zhou,L. and Timmis,K.M. (1990) A family of positive regulators related to the *Pseudomonas putida* plasmid XylS and the *Escherichia coli* AraC activators. *Nucleic Acids Res.*, **18**, 2149–2152.
7. Gallegos,M.T., Michán,C. and Ramos,J.L. (1993) The XylS/AraC family of regulators. *Nucleic Acids Res.*, **21**, 807–810.
8. Rhee,S., Martin,R.G., Rosner,J.L. and Davies,D.R. (1998) A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. *Proc. Natl Acad. Sci. USA*, **95**, 100413–100418.
9. Known,H.J., Bennik,M.H., Demple,B. and Ellenberger,T. (2000) Crystal structure of the *Escherichia coli* Rob transcription factor in complex with DNA. *Nature Struct. Biol.*, **7**, 424–430.
10. Zhang,X., Reeder,T. and Schleif,R. (1996) Transcription activation parameters at *ara* pBAD. *J. Mol. Biol.*, **258**, 14–24.
11. Johnson,C.M. and Schleif,R.F. (2000) Cooperative action of the catabolite activator protein and AraC *in vitro* at the *araFGH* promoter. *J. Bacteriol.*, **182**, 1995–2000.
12. Li,Z. and Demple,B. (1996) Sequence specificity for DNA binding by *Escherichia coli* SoxS and Rob proteins. *Mol. Microbiol.*, **20**, 937–945.
13. Holcroft,C.C. and Egan,S.M. (2000) Roles of cyclic AMP receptor protein and the carboxyl-terminal domain of the subunit in transcription activation of the *Escherichia coli rhaBAD* operon. *J. Bacteriol.*, **182**, 3529–3535.
14. Ruíz,R., Ramos,J.L. and Egan S.M. (2001) Interactions of the XylS regulators with the C-terminal domain of the RNA polymerase α subunit influence the expression level from the cognate Pm promoter. *FEBS Lett.*, **491**, 207–211.
15. Ramos,J.L., Gallegos,M.T., Marqués,S., Ramos-González,M.I., Espinosa-Urgel,M. and Segura,A. (2001) Responses of gram-negative bacteria to certain environmental stresses. *Curr. Opin. Microbiol.*, **4**, 166–171.
16. Ramos-Díaz,M.A. and Ramos,J.L. (1998) Combined physical and genetic map of the *Pseudomonas putida* KT2440 chromosome. *J. Bacteriol.*, **180**, 6352–6363.