



BacRegulators: a database of transcriptional regulators in bacteria and archaea

Manuel Martínez-Bueno^{1,†}, Antonio J. Molina-Henares^{1,†},
Eduardo Pareja², Juan L. Ramos¹ and Raquel Tobes^{1,*}

¹Department of Plant Biochemistry and Molecular and Cellular Biology, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, E-18008 Granada, Spain and ²Bioinformatics, Hospital Universitario Virgen de las Nieves, E-18014 Granada, Spain

Received on February 23, 2004; revised on April 27, 2004; accepted on May 10, 2004
Advance Access publication May 27, 2004

ABSTRACT

Motivation: The BacRegulators database is intended to collect and to integrate information on proteins belonging to defined families of transcriptional regulators in prokaryotes.

Results: The BacRegulators database currently contains data on two families of transcriptional regulators: AraC-XylS and TetR. The proteins included in the BacRegulators database have been identified by screening 123 genomes from archaea and bacteria and the SWISS-PROT and TrEMBL databases with profiles defining each family. As the result of an integration process, we have included 1326 different protein sequences from the AraC-XylS family and 1487 different protein sequences from the TetR family. The definition of an entry in BacRegulators is based on protein sequence, source organism, genome element and position in this genome element. The BacRegulators site allows the user to retrieve protein sequences, functional features and experimental evidence supporting the functions, references and the three-dimensional structure of the regulator when available. BacRegulators supplies an innovative tool that allows the researcher to obtain an integrated report that shows the data corresponding to other entries which are related by sequence similarity to the query entry. BacRegulators detects and classifies the regulators belonging to AraC-XylS and TetR families present in prokaryotic genomes, and thus contributes to a more accurate annotation of regulators in genomes. The information collected on each protein in the family can be useful to characterize a new regulator or compile information on the biological properties of a known regulator.

Availability: The BacRegulators is available at www.bactregulators.org

Contact: rtobes@bioinformatica.org

Supplementary information: www.bactregulators.org/supplementary

INTRODUCTION

The most important mechanisms that bacteria use to adapt their physiology to changing environmental conditions are based on regulation at the transcriptional level (Ramos *et al.*, 1997, 2001; Ishihama, 2000). Transcriptional regulators can fine-tune gene expression to face the specific environmental changes and stress conditions. This control of gene expression is achieved through the delicate interplay of sigma factors that confer promoter selectivity, and transcriptional regulators that modulate RNA polymerase activity (Ramos *et al.*, 1997; Gallegos *et al.*, 1997). Since 1990, we have been updating information on the AraC-XylS family of transcriptional regulators (Gallegos *et al.*, 1997), and we have compiled information on proteins of this family in the AraC-XylS database (Tobes and Ramos, 2002). In an effort to understand the regulatory networks of complex processes that involve regulators of different families, we decided to extend the database to incorporate new families of regulators. This new approach can be useful to define the regulatory networks that control microbial responses to different environmental challenges. In addition, the rapid appearance of newly sequenced prokaryotic genomes generates exponential growth in the number of protein sequences, but in contrast, the experimental acquisition of information about these proteins is relatively limited.

Another issue is that data from sequenced genomes and protein databases are often not integrated, so further efforts are needed in that direction. BacRegulators (www.bactregulators.org) has been created with the aim of integrating information about regulators from the NCBI microbial genome resource and from the SWISS-PROT and TrEMBL protein databases.

The BacRegulators database currently compiles data on two families of regulators: AraC-XylS and TetR. Information

*To whom correspondence should be addressed.

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

on AraC-XylS family members has been updated and at present covers 1326 sequences. These sequences have been obtained after screening the bacterial and archaeal genomes included in the NCBI Refseq database (June 11, 2003) and the SWISS-PROT and TrEMBL (SPTR) protein databases (June 16, 2003). The sequences have been integrated in a non-redundant set.

The new family we have analyzed and incorporated into this new database is the TetR family. We have developed a TetR profile, which has been validated (Tobes *et al.*, 2004) and used to detect TetR proteins both in genomes and in the SPTR database. In addition, we have compiled a non-redundant set of 1487 protein sequences belonging to the TetR family.

SYSTEMS AND METHODS

Selection of a non-redundant set of protein sequences of each family by analyzing the SPTR database and NCBI proteomes

The first step in the development of the BacRegulators database was to select the sets of transcriptional regulators with profiles defining each family (Gallegos *et al.*, 1997; Tobes *et al.*, 2004). The screening included searches in SPTR and in all the proteins from whole genome sequencing projects (available at the NCBI microbial genomes resource). The HAMAP project (High-quality Automated and Manual Annotation of microbial Proteomes) (Gattiker *et al.*, 2003) is oriented to automatically annotate in SWISS-PROT a significant percentage of proteins originating from bacterial and archaeal genome sequencing projects. This project has initiated a process of incorporation of transcriptional regulators from genome sequencing projects to SPTR. At present, the overlap between the sets of regulatory protein sequences detected in SPTR and in genome sequencing projects is only partial. We have developed a tool that integrates protein sequences from SPTR with protein sequences from genomes available at NCBI in a non-redundant set for each family (Supplementary Figure 1). This information is available at the BacRegulators website (www.bactregulators.org).

BacRegulators entry

The extraction of information about regulators is a complex and time-consuming task. When we extract data about a regulator it is always possible to assign these data to a protein sequence and to a microorganism. Furthermore, if the microorganism has been sequenced we can assign the information to a protein encoded by a specific gene located in a precise position of a genome element. The expanding availability of sequenced genomes emphasizes the importance of genomic data in the definition of a protein. On the other hand, to ensure flexibility and precision in a database it is necessary to maintain the information at the maximum level of granularity, considering that it is easy to combine information by common features but it is impossible to separate merged information.

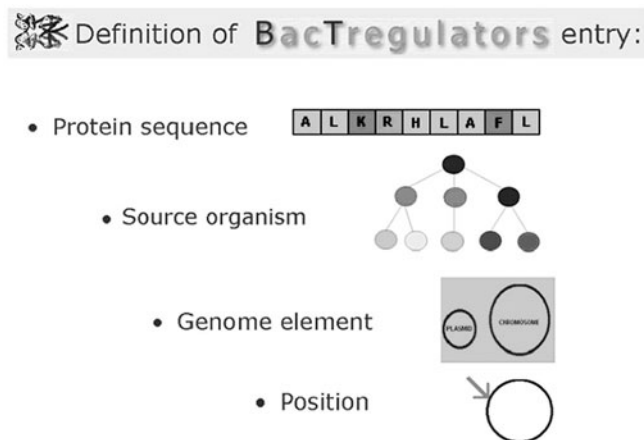


Fig. 1. Each BTR entry is defined considering its protein sequence, the taxonomic classification of the source organism, the genome element and the position in which is located its corresponding gene.

Thus, the definition of an entry in BacRegulators is based on the protein sequence, the source microorganism, the genome element (chromosome, plasmid, phage, etc.) and the specific location of the gene in the corresponding genome element (Fig. 1). The source microorganism is registered according to the taxonomic criteria of the NCBI taxonomy database (Wheeler *et al.*, 2000) (<ftp://ftp.ncbi.nih.gov/pub/taxonomy/>). In addition, the BacRegulators database stores the taxonomic tree structure, thus making it possible to access the lineage of each source microorganism. Proteins with differences at any of the four levels are recorded as different entries (Fig. 2). Thus, the number of entries in the BacRegulators database (5119) is higher than the number of sequences (2813) because several entries can share the same sequence. However, all entries with identical sequences are interconnected.

Most of the entries included in BacRegulators have been automatically generated after screening the SWISS-PROT, TrEMBL and TrEMBL-new databases (Boeckmann *et al.*, 2003) and the complete proteomes available at the Refseq NCBI microbial genomes resource (Pruitt and Maglott, 2001) for the TetR and XylS/AraC family profiles. The automatic generation of entries is different depending on the source database. When the protein comes from the NCBI Microbial Genome database the process is simple: each regulator detected in a proteome corresponding to a complete genome generates an entry with information for all four levels of definition (sequence, source microorganism, genome element and position). However, if the source database is SPTR the process is more complex. The first difference is that the definition of an entry in the SPTR database is based solely on the protein sequence (Boeckmann *et al.*, 2003), consequently the same SPTR entry can correspond to more than one organism and to more than one genome element. Another difference to

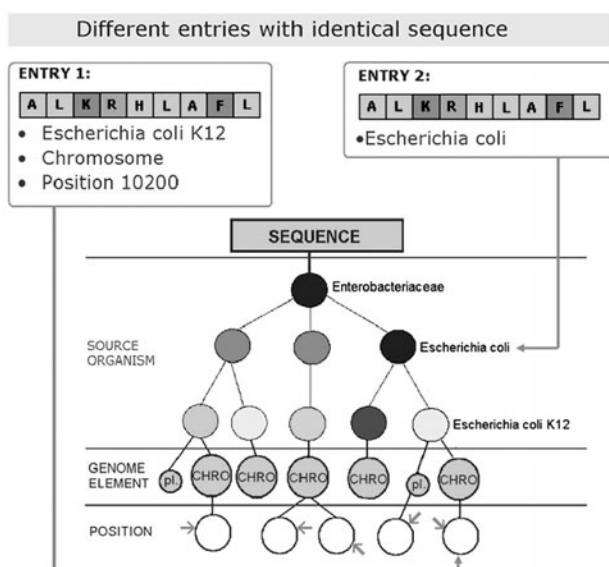


Fig. 2. Two identical sequences can have different entries depending on differences at any of the four levels that define a BTR entry: protein sequence, source organism, genome element and position. In this figure, for example, entry 1 and entry 2 share the same sequence but differ in the source organism, in the genome element and in the position in which the gene encoding the protein is located.

consider is the possibility of redundancy, especially between TrEMBL and TrEMBL-new entries (Gattiker *et al.*, 2003). Thus, for the automatic generation of BacTregulators entries from SPTR database we have used the following strategy: (1) If the SPTR entry corresponds to only one organism and to only one genome element, only one BacTregulators entry is generated. (2) If the SPTR entry corresponds to more than one organism or to more than one genome element, several BacTregulators entries are generated. To avoid entry redundancy, the presence of each entry generated in the database is checked before its inclusion as a new BacTregulators entry.

During the phase of data extraction from published papers, the curators assign information to the appropriate entry, assisted by tools specifically designed for this task. If occasionally the appropriate entry is not included in the set of the automatically generated entries, the curators can generate additional entries to assign the information in the most specific way. Curators are continuously updating BacTregulators database but new data about knowledge and sequences will be available by releases to ensure high quality and integrity of the data.

RESULTS

Information content and structure of the BacTregulators database

The data structure of BacTregulators database is especially designed to include data from genomes. The information is

stored in a highly structured way, organized in 158 fields grouped into 33 tables in a relational database managed by the MySQL database management system and the PHP scripting language.

BacTregulators contains data of three types: sequences, knowledge and references.

- (1) Information about sequences has been specially treated by integrating, filtering and checking each protein sequence. We supply three sets of non-redundant complete protein sequences for each family of regulators:
 - (i) The whole set of protein sequences obtained as the result of the integration of sequences from SPTR and NCBI prokaryotic genomes.
 - (ii) The set of protein sequences that result from the integration of sequences from the SWISS-PROT, TrEMBL and TrEMBL-new databases.
 - (iii) The set of protein sequences from NCBI prokaryotic genomes.

As in the AraC-XylS database described previously (Tobes and Ramos, 2002), we also supply these three sets of sequences by domains (DNA-binding domain, N-terminal domain and C-terminal domain) in 9 additional sets of sequences for each family of regulators.

- (2) In BacTregulators, the knowledge extraction process combines automatic data extraction with manual extraction. Some data assigned to each entry have been obtained automatically from NCBI Microbial resources and from the SPTR knowledgebase (see below). The rest of the knowledge associated with each entry has been extracted manually from bibliographic references. We have adopted a special structure for this type of data, using the text paragraph as the information unit. All manually obtained data are organized in text paragraphs that are always individually referenced. This structure identifies the source of each piece of data. During the manual extraction of knowledge, we have dissected experimental data that support biological features. These data are termed 'experimental evidence' and receive a special treatment in BacTregulators. When there is experimental evidence to support the knowledge expressed in a paragraph, a link to the experimental data is displayed. These data are useful to evaluate the reliability of the knowledge data.
- (3) Each text paragraph is referenced and linked to the Medline abstract. Currently the reference database contains 457 references corresponding to AraC-XylS and TetR families.

Graphical information about the three-dimensional (3D)-structures of crystallized proteins and tutorials are also available at the BacTregulators website.

Access to BacTregulators information

The information contained in the BacTregulators database can be accessed in various ways. Data can be searched by BacTregulators identifier or by SWISS-PROT-TrEMBL or NCBI identifiers. A search by family, microorganism, name and COG allows the user to retrieve specific sets of proteins. Finally, a text search in all the fields of the database allows more flexible browsing. The reference database can also be independently browsed by text searching.

Information about a specific entry is displayed in four sections (Supplementary Figure 2). The first section (blue vertical bar) shows the data that define the BacTregulators entry: sequence identifier, source organism, genome element and position in the genome element. The second section (blue-green vertical bar) contains data automatically obtained from databases: SPTR or NCBI accession number, source database, the protein's full name and short name, gene name, gene orientation, COG code, NCBI functional code and last update of this set of data. The third section (green vertical bar) displays knowledge manually extracted from published research articles. This knowledge is structured in the following fields: function, genes regulated, regulatory networks, effectors, genomic allocation, promoters of regulated genes, promoters of the regulator, sigma dependence, pathogenicity, applications, mutational data, 3D-structure, oligomerization, similarities, comments and last update of this group of manually extracted data (Tobes and Ramos, 2002). For each of these knowledge fields, we also provide downloadable text files containing all the information corresponding to each field. The last section (orange vertical bar) provides access to information related to protein sequences. The subsection on protein sequences provides access to the complete protein sequence and to the sequence of the different domains. Finally, the subsection of BLAST similarities allows the user to access the file with the BLAST results as compared to the rest of the members of the family, and a list of entries with a similar sequence ranked by significance (BLAST *E*-value).

DISCUSSION

Integration of knowledge customizable at the retrieval step

When a user is accessing a specific entry, the knowledge associated with related entries can be useful to infer functional features. This was our rationale for designing an innovative tool to display an integrated report of a set of entries related to a single initial entry. Although the information is maintained at the maximum level of granularity associated with each entry, the information corresponding to a given set of related entries can be combined at the retrieval step. We have considered similarity between sequences as the key criterion to define a group of related entries. This first selection of related entries can be restricted by taxonomic requirements. In this way the

level of integration of knowledge can be tuned by the user, who sets similarity parameters and taxonomic level.

To select the set of related entries, a list is displayed with all BacTregulators entries that have significant BLAST similarity with the initial sequence. All entries that share a sequence identical to the initial entry are displayed against a yellow background. In a first step, the user selects an *E*-value as the threshold to define the set of related sequences. In a second step, this set can be limited by taxonomic criteria: the user chooses the taxonomic level for including an entry in the set of related entries. The lineage corresponding to the source microorganism of the initial entry is displayed to allow the user to fix taxonomic limits (Supplementary Figure 3). Once the similarity threshold and taxonomic limits are set, an integrated report of the selected set of entries is displayed. The data are displayed by field in different sections. Each section includes all the data for a field from all entries in the set. The corresponding BacTregulators entry identifier is indicated at the beginning of each paragraph. A list of integrated entries is also accessible at the integrated report page. (Supplementary Figure 4).

It is unquestionable that orthologous proteins frequently share many functional features. Hence, inference of function from sequence similarity (Gattiker *et al.*, 2003) makes particular sense for orthologous proteins. However, in practise, it is extremely difficult to define orthologous proteins. We have avoided this problem by allowing the user to choose the set of entries from which function may be inferred. This tool can help orient the user to the possible function of regulators with no defined function. Thus, a selected group of entries with a similar sequence can include entries with functional features and entries with no known function. If the function is similar in all the characterized proteins in the group, it is likely that the uncharacterized proteins of the group have a similar function. The tool described here thus facilitates the selection of functional clusters of entries for which knowledge can be propagated to entries with no published function.

New families of transcriptional regulators will be added to BacTregulators database to extend the understanding of the complex circuitry of gene regulatory networks in bacteria and archaea.

BacTregulators is especially intended to detect the regulators in prokaryotic genomes and classify them into families. As designed, BacTregulators facilitates the accurate annotation of regulators in genomes. In addition, the information collected on each protein of the family can be useful for researchers who wish to characterize a new regulator or investigate the biological properties of a known regulator.

ACKNOWLEDGEMENTS

We thank Carmen Lorente and Karen Shashok for checking the language of the manuscript and Wilson Teran for the information about TtgR. Work in J.L.R.' laboratory was

financed by European Commission grants (BIO-CT3-00435 and BIO-CT2000-00170) and by grants from the CICYT (GEN2001-4698-C05-3 and BIO2003-00515).

REFERENCES

- Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
- Gallegos,M.T., Schleif,R., Bairoch,A., Hofmann,K. and Ramos,J.L. (1997) AraC/XylS family of transcriptional regulators. *Microbiol. Mol. Biol. Rev.*, **61**, 393–410.
- Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.
- Ishihama,A. (2000) Functional modulation of *Escherichia coli* RNA polymerase. *Annu. Rev. Microbiol.*, **54**, 499–518.
- Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
- Ramos,J.L., Marques,S. and Timmis,K.N. (1997) Transcriptional control of the *Pseudomonas* TOL plasmid catabolic operons is achieved through an interplay of host factors and plasmid-encoded regulators. *Annu. Rev. Microbiol.*, **51**, 341–373.
- 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 stressors. *Curr. Opin. Microbiol.*, **4**, 166–171.
- Tobes,R. and Ramos,J.L. (2002) AraC-XylS database: a family of positive transcriptional regulators in bacteria. *Nucleic Acids Res.*, **30**, 318–321.
- Tobes,R., Martínez-Bueno,M., Molina-Henares,A.J., Pareja,E. and Ramos,J.L. (2004) The TetR family of transcriptional repressors, in preparation.
- Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L. and Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.