# RegulonDB (version 3.0): transcriptional regulation and operon organization in *Escherichia coli* K-12

**Heladia Salgado, Alberto Santos-Zavaleta, Socorro Gama-Castro, Dulce Millán-Zárate, Frederick R. Blattner[1] and Julio Collado-Vides***

Centro de Investigación sobre Fijación de Nitrógeno, UNAM A.P. 565-A Cuernavaca, Morelos 62100, México and [1]Department of Genetics, University of Wisconsin, Madison, 445 Henry Mall Madison, WI 53706, USA

Received October 12, 1999; Accepted October 13, 1999

## ABSTRACT

**RegulonDB is a database on transcription regulation and operon organization in *Escherichia coli*. The current version describes regulatory signals of transcription initiation, promoters, regulatory binding sites of specific regulators, ribosome binding sites and terminators, as well as information on genes clustered in operons. These specific annotations have been gathered from a constant search in the literature, as well as based on computational sequence predictions. The genomic coordinates of all these objects in the *E.coli* K-12 chromosome are clearly indicated. Every known object has a link to at least one MEDLINE reference. We have also added direct links to recent expression data of *E.coli* K-12. The version presented here has important modifications both in the structure of the database, as well as in the amount and type of information encoded in the database. RegulonDB can be accessed on the web at URL: http://www.cifn.unam.mx/Computational_Biology/regulondb/**

## INTRODUCTION

RegulonDB is a relational database containing information on mechanisms of transcriptional regulation as well as operon organization in the *Escherichia coli* K12 chromosome. Our previous publications in this journal explain in detail the design of this relational database as well as subsequent modifications (1,2). The relational design has been modified in some aspects in order to add flexibility to the database. We have enriched the knowledge on transcriptional regulation by incorporating new regulatory elements, the Shine–Dalgarno ribosome binding sites (RBSs), as well as rho independent terminator signals (3). Furthermore, in addition to annotations gathered from the literature, basically from MEDLINE and PubMed (4), we have added predicted promoters, regulatory sites and operons, based on sequence analyses. These predictions are clearly marked, so that they can be easily distinguished from information gatherered from the literature. Methods for such predictions have been previously described (5,6). In this way, every gene in the chromosome is assigned to a known or predicted transcriptional unit. We have added individual links of genes in RegulonDB to genome-wide expression data gathered in the laboratory of F.R.B. (7).

## COMPUTATIONAL INFRASTRUCTURE AND RELATIONAL DESIGN

RegulonDB uses a relational database scheme. The design of the database has previously been explained in detail (1). This year we have migrated the database from Sybase to Oracle 8 Server. Forms to access the information from the web have been implemented using the software Developer 2000 and PL-SQL query language (Copyright © Oracle Corporation).

The relational design has been modified, motivated by the requirement to make it more flexible so that partial information can be continuously incorporated as it is gathered from the literature. Thus, operons do not necessarily require a promoter anymore and vice versa. Predicted objects based on computational analyses are internally distinguished by their absence of an ID, so that they can be more easily modified in future improved analyses. The user will easily identify predicted from known objects, although a given operon can be formed by a mixture of known and predicted regulatory elements. Predictions are kept so as to complement what has been gathered from the literature. New tables were added to deal with terminator signals, and Shine–Dalgarno RBSs. We have divided gene products into three classes: (i) regulatory polypeptides, (ii) polypeptides and (iii) RNAs, with their corresponding tables linked to the gene table. Regulatory polypeptides refer exclusively to gene products that define transcriptional DNA-binding proteins, whereas all other products are left in the general polypeptide table. RNAs contain the different types of stable RNAs: tRNAs, rRNAs and miscellaneous RNAs. A table with repeat elements has also been added.

We have modified the way of encoding operons by adding a table for transcriptional units. The information contained within the table of operons has been moved to a new table called 'transcription unit'. This enables the rich diversity of operon expression, that can involve multiple promoters transcribing different sets of genes under different conditions to be encoded in the database. At the same time, given the partial knowledge available, a promoter does not necessarily need to be associated to a known operon and reciprocally, an operon does not necessarily have a known promoter.

*To whom correspondence should be addressed. Tel: +527 313 2063; Fax: +527 317 5581; Email: ecoli-t1@cifn.unam.mx

**Table 1.** Summary of the contents in RegulonDB version 3.0

| Objects | | Known | Predicted |
|---|---|---|---|
| Transcription units | | 374 | 2271 |
| Genes | | 4405 | |
| Promoters | | 432 | 4602 |
| Sites | | 406 | 270 |
| Regulatory interactions | | 433 | |
| Terminators | | 40 | |
| RBSs | | 59 | |
| Gene products (4405): | | | |
| | Regulatory polypeptides | 83 | |
| | RNAs | 115 | |
| | Other polypeptides | 4207 | |
| Protein complexes | | 83 | |
| Signals | | 36 | |
| Regulatory phrases | | 126 | |
| Alignments | | 33 | |
| Matrices | | 34 | |
| Synonyms | | 6090 | |
| External references | | 4394 | |

Number of objects as of October 1, 1999 in RegulonDB.

Each transcriptional unit encodes information about one promoter, the set of contiguously located genes that are being transcribed and a terminator, all of these associated to a given condition of expression. One single operon may involve several transcriptional units. The structure of an operon can always be recuperated as the longest transcriptional unit that may share genes with smaller ones. The expression levels for each gene are accessed via a link to an external file, so the relational design is not affected.

Within the site table we have added the sequence of the binding site for a specific regulatory protein. This sequence may differ from the sequence in the alignment associated to the regulatory protein. This will eliminate the error of the user considering the aligned sequence to be exactly the experimental site. Protein complexes, limited to regulatory proteins in RegulonDB, describe the type of symmetry of the binding site for that protein.

## OVERVIEW OF THE CURRENT DATA

Table 1 shows the number of objects as of October 1, 1999 in RegulonDB. The updated version of this overview table can be accessed on the web. The web address for the overview table, together with additional links to related databases and tools for upstream analysis are described in Table 2.

As can be observed, the current version of RegulonDB has an increased amount of information, as well as new regulatory elements describing operons and transcriptional regulation in a more comprehensive way. Current information on transcription units correspond in fact to operons, except for eight cases which are multiple transcripts of some few operons. The complete set of genes for the *E.coli* K12 chromosome have been incorporated. Their names and synonyms were obtained from the annotation of the *E.coli* genome and kept strictly conserved so that the name in RegulonDB corresponds to the name of F.R.B.'s database. All known promoters, regulatory interactions and operons have at least one link to an external database (usually MEDLINE, but also GenBank). We have tried to reference the original publications wherever possible.

## OVERVIEWS AND EXAMPLES

Complex operons can transcribe subsets of genes under different physiological conditions by means of different promoters and internal terminators, as illustrated, among many others, by the nlpD-rpoS, rpsU-dnaG-rpoD, focA-pflB and rpoH transcription units (8–12). The design of RegulonDB can describe such complex operons, as well as their rich complex regulatory interactions. Some examples of complex operons with several promoters, as well as operons with a rich upstream transcriptional regulation are illustrated with figures and more detailed explanations at the following URL: http://www.cifn.unam.mx/Computational_Biology/regulondb/docs/complex_operons.html

Furthermore, in order to provide a better overview of the biology contained in RegulonDB we have added the distribution of several objects and their relative position in the chromosome. These overviews can be obtained at URL: http://www.cifn.unam.mx/Computational_Biology/regulondb/docs/overviews.html

**Table 2.** Related links on the web

| | |
|---|---|
| Main page of RegulonDB (v.3.0) | http://www.cifn.unam.mx/Computational_Biology/regulondb/ |
| Summary table | http://www.cifn.unam.mx/Computational_Biology/regulondb/docs/summary.html |
| Overview tables | http://www.cifn.unam.mx/Computational_Biology/regulondb/docs/overview.html |
| Examples of complex operons | http://www.cifn.unam.mx/Computational_Biology/regulondb/docs/complex_operons.html |
| Current paper in pdf format | http://www.cifn.unam.mx/Computational_Biology/regulondb/docs/regulondb3.0.pdf |
| Web page of the Laboratory of Computational Biology | http://www.cifn.unam.mx/Computational_Biology/ |
| **Related databases** | |
| *E.coli* K-12 expression data | http://www.genetics.wisc.edu/html/expression2.html |
| EcoCyc | http://ecocyc.PangeaSystems.com/ |
| GenProtEC | http://dbase.mbl.edu/genprotec/start |
| Colibri | http://bioweb.pasteur.fr/GenoList/Colibri/ |
| *E.coli* genetic stock center | http://cgsc.biology.yale.edu/ |
| ECO2DB | http://janis.proteome.med.umich.edu/Eco2DBase/ |
| **Analysis tools for gene regulation** | |
| Yeast-tools | http://copan.cifn.unam.mx/~yeast/ |
| Dyad-detector | http://copan.cifn.unam.mx/~yeast/dyad-detector.html |
| Gibbs sampler | http://bayesweb.wadsworth.org/gibbs/gibbs.html |
| AlignACE | http://arep.med.harvard.edu/mrnadata/ |

The information contained in RegulonDB is pertinent to compare and analyze transcriptional global studies of gene regulation in *E.coli*, as well as, potentially, in other related bacteria.

## AVAILABILITY

RegulonDB 3.0 can be accessed though the URL: http://www.cifn.unam.mx/Computational_Biology/regulondb/ . We kindly ask users of RegulonDB to cite this article.

## ACKNOWLEDGEMENTS

We particularly thank fruitful discussions with Araceli Huerta on the changes to the design, as well as for sharing the promoter predictions. We wish to thank Guy Plunkett III for the availability of the repeat elements, gene names and synonyms; Monica Riley for her updated functional annotation of gene products and Claude Thermes for sharing his collection of terminators. We also appreciate useful discussions with Craig Richmond and Jeremy Glasner. This work was supported by grants from DGAPA and Conacyt to J.C.-V. Part of this work was done while J.C.-V. was on sabbatical leave in the laboratory of F.R.B. at the University of Wisconsin, Madison.

## REFERENCES

1. Huerta,A.M., Salgado,H., Thieffry,D. and Collado-Vides,J. (1998) *Nucleic Acids Res.*, **26**, 55–59.
2. Salgado,H., Santos,A., Garza-Ramos,U., van Helden,J., Díaz,E. and Collado-Vides,J. (1999) *Nucleic Acids Res.*, **27**, 59–60.
3. Benson,D.A., Bogusky,M., Lipman,D.J. and Ostell,J. (1998) *Nucleic Acids Res.*, **26**, 1–7. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 15–18.
4. Carafa,Y., Brody,E. and Thermes,C. (1990) *J. Mol. Biol.*, **216**, 835–858.
5. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) *Science*, **277**, 1453–1462.
6. Thieffry,D., Salgado,H., Huerta,A.M. and Collado-Vides,J. (1998) *Bioinformatics*, **14**, 391–400.
7. Richmond,C.S., Glasner,J.D., Mau,R., Jin,H. and Blattner,F.R. (1999) *Nucleic Acids Res.*, **27**, 3821–3835.
8. Lange,R. and Hengge-Aronis,R. (1994) *Mol. Microbiol.*, **13**, 733–743.
9. Lupski,J.R. and Godson,G.N. (1984) *Cell*, **39**, 251–252.
10. Sawers,G. (1993) *Mol. Microbiol.*, **10**, 737–747.
11. Sirko,A., Zehelein,E., Freundlich,M. and Sawers,G. (1993) *J. Bacteriol.*, **175**, 5769–5777.
12. Kallipolitis,B.H. and Valentin-Hansen,P. (1998) *Mol. Microbiol.*, **29**, 1091–1099.