D394–D397 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj156

RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions

Heladia Salgado, Socorro Gama-Castro, Martín Peralta-Gil, Edgar Díaz-Peredo, Fabiola Sánchez-Solano, Alberto Santos-Zavaleta, Irma Martínez-Flores, Verónica Jiménez-Jacinto, César Bonavides-Martínez, Juan Segura-Salazar, Agustino Martínez-Antonio and Julio Collado-Vides*

Program of Computational Genomics, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México. A.P. 565-A. Cuernavaca, Morelos 62100, Mexico

Received September 15, 2005; Revised and Accepted October 31, 2005

ABSTRACT

RegulonDB is the internationally recognized reference database of *Escherichia coli* K-12 offering curated knowledge of the regulatory network and operon organization. It is currently the largest electronically-encoded database of the regulatory network of any free-living organism. We present here the recently launched RegulonDB version 5.0 radically different in content, interface design and capabilities. Continuous curation of original scientific literature provides the evidence behind every single object and feature. This knowledge is complemented with comprehensive computational predictions across the complete genome. Literature-based and predicted data are clearly distinguished in the database. Starting with this version, RegulonDB public releases are synchronized with those of EcoCyc since our curation supports both databases. The complex biology of regulation is simplified in a navigation scheme based on three major streams: genes, operons and regulons. Regulatory knowledge is directly available in every navigation step. Displays combine graphic and textual information and are organized allowing different levels of detail and biological context. This knowledge is the backbone of an integrated system for the graphic display of the network, graphic and tabular microarray comparisons with curated and predicted objects, as well as predictions across bacterial genomes, and predicted networks of functionally related gene products. Access RegulonDB at http://regulondb.ccg.unam.mx.

INTRODUCTION

Escherichia coli K-12 is the, simplest, best known and natural biological model where novel ways of understanding and new modeling strategies can be tested. It is through *E.coli* that we will know how far systems biology will take us. With these aims in mind, a group of scientists have conformed the International *E.coli* Alliance, where the efforts of our group find their context (1). RegulonDB is the reference database of *E.coli* curated knowledge of genetic regulation and operon organization. The curation team in our laboratory feeds both RegulonDB and EcoCyc, providing two computational environments for the same biological content. This is important for users to know, and in fact, as detailed later, we have unified the content and releases of the two databases to avoid potential confusion by the users.

The amount of experimentally validated knowledge of the *E.coli* regulatory network is the largest currently available for any organism. To represent this knowledge in a computable and easy to use electronic database several levels are involved of modelling the complex biology of gene regulation and function. These levels include the database design which imposes a much more precise definition of biological concepts than what is frequently used by the biological community, graphic designs of both chromosomal and network images as well as navigation pathway design. RegulonDB represents a frontier in gene regulation design in the bioinformatics of bacterial genomics.

*To whom correspondence should be addressed. Tel +527 77 313 9877; Fax +527 77 317 5581; Email: ecoli-t1@ccg.unam.mx

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

The current version 5.0 has both new biological content and major navigation changes that makes this a much better version than the previous one. In this paper we describe the improvements and changes in regard to the previous version. First, we describe the computational aspects of improvements to the interfaces, then we show the new or expanded computational tools to visualize, analyse and compare the available knowledge, and finally, we finish with a summary of the major expansions of the biological content in RegulonDB. This includes a table with mechanisms of gene regulation beyond transcription initiation.

MATERIALS AND METHODS

On curation

We start by gathering abstracts from PubMed database using a set of pertinent keywords. These abstracts are read and selected to obtain the complete articles in order to read them. The data extracted are added to both RegulonDB and EcoCyc, through capture forms. The quality of the data is monitored automatically through inconsistency reports. Curation in our team is coordinated with colleagues in EcoCyc so that literature on gene regulation as well as observations from users of both databases pertintent to gene regulation are unified and curated by our team. The team of curators follows a unified set of criteria or guidelines that are expanded as experience accumulates.

Database and interfaces

RegulonDB is a relational database implemented in Oracle DBMS. Interfaces are written in Java, using the Java 2D API. The software was developed with Java 2 Platform, Enterprise Edition (J2EE) architecture. J2EE defines the standard for developing component-based multitier enterprise applications.

RESULTS

Interfaces

RegulonDB version 4.0 had seven search forms corresponding to genes, promoters, transcription units (TUs), transcriptional regulators, regulons, effectors and growth conditions. This highly structured search generated interfaces with precise but limited content, thus involving substantial navigation efforts. In this new version 5.0, a radically different, much more simplified navigation structure has been implemented based on three main object streams: genes, operons and regulons, and a fourth to be re-implemented for conditions. These three navigation paths offer the user a rich and structured information always containing data on gene regulation. The site map shows the different available interfaces and links departing from them. Some common navigation features are the following. Pages frequently have graphic and textual information. Field descriptors of specific objects are described only when they have content. Text and tables can be exported in XML format, and images in jpg format. The main menu shows the first and second navigation options in each branch of alternative navigation, thus facilitating decisions to the user. This main menu is accessible from any page in RegulonDB. Pages have stable URLs enabling them to be referenced. Finally, graphic tools and interfaces are much faster than our previous version.

Gene navigation stream. This path enables visualization of detailed knowledge of genes, their products and their associated direct features such as the Shine Dalgarno sequence, strand and map position, gene sequence access; as well as the molecular weight, functional classification and access to the protein sequence. This page also displays information about the operon a given gene belongs to, and if known, the transcription factors that regulate the gene. The graphic display shows the genomic context of the queried gene (i.e. its operon and other neighboring genes), with the all objects located within the region, including promoters, binding sites and terminators. This may well include sites with no regulatory effect on the queried gene.

Operon navigation stream. An operon is defined as a group of two or more genes transcribed as a polycistronic unit. For database modeling reasons, we also accept monocistronic operons. Complex operons contain multiple promoters, some of which may transcribe a fraction of the genes, defining different transcription units. Operons are displayed on the top of the page, with all the regulatory elements affecting the different expression alternatives, more precisely, the different transcription units of the operon. The complete set of known transcription units are displayed below the operon, with detailed regulatory information of the corresponding promoters, terminators if known, and the regulatory sites and corresponding transcription binding factors.

Regulon navigation. To understand the organization of this form, we must first recall the biological definition of a regulon and the extended definition we use here. Regulons were defined for the first time for the arginine biosynthetic system regulated by ArgR, as a set of genes subject to the regulation of one and only one regulator (2). This is what we call a simple regulon. Following the same principle, complex regulons are defined as a group of genes subject to regulation by several and exactly the same regulators. For instance, global regulators participate in many different complex regulons, and some regulators, like AraC and NarL, have no known simple regulon, they always co-regulate genes. When searching with a regulator, the user will get the list of all simple and complex regulons where that regulator participates, in other words, its co-regulators. These co-regulators are links that bring the user to the whole set of regulons where they participate. The beauty of this regulon table is that it enables a higher level navigation of the network, since the user can go from, for instance, the regulons where FNR participates to those where GlpR participates. From there, a specific regulon can be selected bringing the user to tables of the regulators defining the complex regulon, containing the list of their co-regulators. Additionally, a larger table shows the set of sites, promoters and regulator function (activator, repressor or dual). Sites are displayed by groups of strictly co-regulated genes, for instance when looking at the (ArcA, NarL) complex regulon, sites are separated for ArcA being activator and ArcA being repressor, in both groups NarL is a repressor. Similarly, the (AraC, CRP) complex regulon can be decomposed in three strict groups: AraC dual, CRP activator; AraC activator, CRP activator and

AraC activator, CRP dual. The whole description of these simple and complex regulons can be accessed in RegulonDB.

Navigation of growth conditions. Since our previous version, RegulonDB contains what we called non-mechanistic information derived from gene expression experiments under certain growth conditions. These putative interactions are mostly based on experiments with knock-outs of regulatory genes. Interactions are defined although no precise site and mechanisms are yet described. At present this data can be accessed from the 'downloads' option in the main page. We are working to generate specific search and navigation of growth conditions and gene expression changes.

Tools and additional interfaces

RegulonDB is a knowledge database integrated in an environment of computational tools that facilitate the use and analysis of its content. Tools existing since our previous publication include GetTools (http://www.ccg.unam.mx/Computational_Genomics/GETools/) to facilitate the comparison with microarray experiments and links to the Regulatory Sequence Analysis (http://embnet.cifn.unam.mx/rsa-tools/). New or expanded versions include the genome browser, a new network graphic display or Network Tool, as well as links to Nebulon, a tool to predict groups of functionally related genes.

Genome browser. The previous version of our browser has been improved. When querying with a gene, the genome browser will point to the circular chromosome showing the position of the query gene. When the input is a group of genes—a TU or an operon—they all appear in their chromosomal position. Genes are colored based on their functional classes. A subsequent level of detail shows a linear display with genes and TUs, showing promoters and binding sites. The genome browser can be accessed from the gene form.

Network display tool. When placed in a query gene coding for a regulatory protein, the network tool displays the sub-network formed by the query gene and its immediate neighbors in the regulatory network. Genes are colored based on whether they are regulators (green) or regulated genes (yellow) and furthermore can be distinguished as global regulators (blue). The definition of global regulators is obtained from (3). The directed arcs of the graph indicate regulatory interactions with green for activation, red for repression, blue for dual effect and black for unknown. This display can be reached starting from any of the three major navigation objects: genes, operons and regulons. We plan to enrich this tool by adding options that allow second-order neighbors.

Nebulon-tool. Nebulon is a recently published computational strategy that generates sets of genes predicted to be functionally related (4). The computational strategy relies on the co-occurrence of genes within operons in the whole bacterial kingdom, based on an initial method of operon prediction in *E.coli* (5). This tool is accessed from the gene interface.

Curation of the biology of gene regulation

Our group has been curating regulation of transcription initiation and operon organization for several years (6–10). This effort feeds two databases, RegulonDB and EcoCyc (11).

Because of the lack of synchronization in public releases, the multiple public versions had differences that could confuse different users, as shown in (12). We have identified all these differences, and have as well curated additional interactions published in (12,13) generating a unified description of the regulatory network. Files to directly re-construct the network as protein interactions (where heterodimer regulators like IHF are also a single node) can be directly accessed from the main menu option of 'downloads'.

Duplicated sites in divergent regulatory regions were eliminated. RegulonDB version 5.0 is equivalent to EcoCyc version 9.1, and from now on, public releases will be strictly synchronized so that a single version of the network will be available. The annotations of this version are based on the latest update of the *E.coli* K-12 genome sequence entry U00096.2 from the Blattner laboratory. The coordinates of genes, transcription start sites, promoters, binding sites and terminators used here derive from work performed by NCBI and by the EcoCyc team since version 8.6. The names of transcriptional regulators have been modified describing whether they act as repressors, activators or have a dual effect. The current total number of annotated and predicted objects can be accessed from the main menu at 'About Regulon'.

Regulation beyond transcription initiation. This year we have expanded our curation efforts to gather mechanisms of gene regulation beyond transcription initiation. Encoding them within the database will require an important expansion in the design of the database. Therefore, we are currently making them available in a table within the Downloads option of the main menu. The variety of mechanisms are grouped into types depending on the interacting molecules as protein–protein, protein–RNA, RNA–RNA or protein–metabolite interactions.

Computational predicted objects. Computational predictions of promoters, regulatory sites and operons are derived from methods implemented in our laboratory and are updated as new versions of such programs are generated (5,14,15). They are, as we have done since years, clearly distinguished as computational predictions different from other types of evidences.

Uses of RegulonDB. As mentioned in the abstract, the regulatory network of *E.coli* K-12 is currently the largest and most detailed regulatory network of any living organism electronically-encoded in a database. Of course many other organisms have much larger and complex regulatory networks, but the amount of interactions with experimental support is still limited. Therefore, it is no surprise to see RegulonDB knowledge playing an essential role in the generation of novel concepts, perspectives and methods in bioinformatics, genomics and systems biology. Some examples among a large set of cases include the notion of topological network motifs (13); novel or expanded computational methods of promoters; operons (16–18); microarray analyses (19,20); and chIP–chip experiments (21); formal models of transcriptional processes (22); metabolic and regulatory network reconstruction (23); and experimental studies (24,25).

DISCUSSION

Databases in molecular biology have become an essential tool like Rosetta stones to decipher different aspects of the biology from genomic sequences. They are also increasingly useful for younger students to learn from electronic publications. The changes implemented in RegulonDB version 5.0 have strongly modified the interfaces with graphic and text information and simplified the major navigation alternatives enriching the amount of knowledge and information each page contains. Particularly attractive is a new navigation capability within the topology of the network across regulators working together and their associated complex regulons.

In terms of content, we are aware that much more remains to be added, with other mechanisms beyond transcriptional regulation and up to date literature. Initial steps in that direction have been taken by curating other mechanisms of genetic regulation beyond transcription initiation. Their integration in the database will require a good investment in design expansion in the future. Currently, these additional mechanisms are available in a table.

Research on the biology of regulatory network may well generate new entities with novel definitions (like that of complex regulons), or novel classifications which will require expanding or increasing the database content and/or graphic display capabilities. For instance, we have classified transcriptional regulators based on their allosteric (generalized) interactions and the origin of such signal-metabolite as either of external origin if internalized, internal origin if synthetized or both. In a similar way, as our understanding of the network expands, the database will acquire new knowledge and reflect it in the modeling of gene regulation.

ACKNOWLEDGEMENTS

We acknowledge Bruno Contreras-Moreira and Mónica Peñaloza-Spínola for their contribution to the curators team; Delfino García-Alonso, Christian A. Avila-Sánchez and Víctor del Moral for their computer support; Rosemary Martínez for her contribution to the design of web pages. We acknowledge suggestions by anonymous referees. This work was supported by NIH grants GM62205-02 and 1-R01-RR07861, and UNAM grant 214905. Funding to pay the Open Access publication charges for this article was provided by the same grants.

Conflict of interest statement. None declared.

REFERENCES

1. Holden,C. (2002) Cell biology. Alliance launched to model *E.coli*.. *Science*, **297**, 1459–1460.
2. Maas,W.K. (1964) Studies on the mechanism of represión of arginine biosíntesis in *Escherichia coli*. II. Dominance of repressibility in diploids. *J. Mol. Biol*., **78**, 365–370.
3. Martinez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol*., **6**, 482–489.
4. Janga,S.C., Collado-Vides,J. and Moreno-Hagelsieb,G. (2005) Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. *Nucleic Acids Res*., **33**, 2521–2530.
5. Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
6. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al*. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli K-12*. *Nucleic Acids Res*., **32**, D303–D306.
7. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Diaz-Peredo,E., Sanchez-Solano,F., Perez-Rueda,E., Bonavides-Martinez,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.
8. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Blattner,F.R. and Collado-Vides,J. (2000) RegulonDB (version 3.0): transcriptional regulation and operon organization in *Escherichia coli K-12*. *Nucleic Acids Res*., **28**, 65–67.
9. Salgado,H., Santos,A., Garza-Ramos,U., van Helden,J., Diaz,E. and Collado-Vides,J. (1999) RegulonDB (version 2.0): a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res*., **27**, 59–60.
10. Huerta,A.M., Salgado,H., Thieffry,D. and Collado-Vides,J. (1998) RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res*., **26**, 55–59.
11. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res*., **33**, D334–D337.
12. Ma,H.W., Kumar,B., Ditges,U., Gunzer,F., Buer,J. and Zeng,A.P. (2004) An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res*., **32**, 6643–6649.
13. Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet*., **31**, 64–68.
14. Huerta,A.M. and Collado-Vides,J. (2003) Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol*., **333**, 261–278.
15. Gonzalez, A.D., Espinosa,V., Vasconcelos,A.T., Perez-Rueda,E. and Collado-Vides,J. (2005) TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes. *Nucleic Acids Res*., **33**, D98–D102.
16. Zwir,I., Shin,D., Kato,A., Nishino,K., Latifi,T., Solomon,F., Hare,J.M., Huang,H. and Groisman,E.A. (2005) Dissecting the PhoP regulatory network of *Escherichia coli* and *Salmonella enterica*. *Proc. Natl Acad. Sci. USA*, **102**, 2862–2867.
17. Li,H., Rhodius,V., Gross,C. and Siggia,E.D. (2002) Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl Acad. Sci. USA*, **99**, 11772–11777.
18. Gordon,L., Chervonenkis,A.Y., Gammerman,A.J., Shahmuradov,I.A. and Solovyev,V.V. (2003) Sequence alignment kernel for recognition of promoter regions. *Bioinformatics*, **19**, 1964–1971.
19. Sabatti,C., Rohlin,L., Oh,M.K. and Liao,J.C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res*., **30**, 2886–2893.
20. Cooper,T.F., Rozen,D.E. and Lenski,R.E. (2003) Parallel changes in gene expression after 20 000 generations of evolution in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **100**, 1072–1077.
21. Herring,C.D., Raffaelle,M., Allen,T.E., Kanin,E.I., Landick,R., Ansari,A.Z. and Palsson,B.O. (2005) Immobilization of *Escherichia coli* RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. *J. Bacteriol*., **187**, 6166–6174.
22. Sneppen,K., Dodd,I.B., Shearwin,K.E., Palmer,A.C., Schubert,R.A., Callen,B.P. and Egan,J.B. (2005) A mathematical model for transcriptional interference by RNA polymerase traffic in *Escherichia coli*. *J. Mol. Biol*., **346**, 399–409.
23. Herrgard,M.J., Covert,M.W. and Palsson,B.O. (2004) Reconstruction of microbial transcriptional regulatory networks. *Curr. Opin. Biotechnol*., **15**, 70–77.
24. Nickels,B.E., Mukhopadhyay,J., Garrity,S.J., Ebright,R.H. and Hochschild,A. (2004) The sigma 70 subunit of RNA polymerase mediates a promoter-proximal pause at the *lac* promoter. *Nature Struct. Mol. Biol*., **11**, 544–550.
25. Adams,M.A. and Jia,Z. (2005) Structural and biochemical evidence for an enzymatic quinone redox cycle in *Escherichia coli*: identification of a novel quinol monooxygenase. *J. Biol. Chem*., **280**, 8358–8363.