Published in final edited form as:

Curr Protoc Bioinformatics. 2018 March ; 61(1): 1.32.1–1.32.30. doi:10.1002/cpbi.43.

Using RegulonDB, the *Escherichia coli* K-12 gene regulatory transcriptional network database

Heladia Salgado¹, Irma Martínez-Flores¹, Víctor H. Bustamante², Kevin Alquicira-Hernández¹, Jair S. García-Sotelo³, Delfino García-Alonso¹, and Julio Collado-Vides^{*,1}

¹Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62100, México

²Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62100, México

³Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Santiago de Querétaro 76230, Querétaro, México

Abstract

In RegulonDB we have been gathering knowledge by manual curation from original scientific literature on the regulation of transcription initiation and the genome organization in transcription units of the *Escherichia coli* K-12 genome for over 25 years. This paper describes six basic protocols that can serve as a guiding introduction to the main content of the current version (v9.4) of this electronic resource. These include the general navigation as well as searching for specific objects such as genes, gene products, transcription units, promoters, transcription factors, coexpression, and Genetic sensory response units or GENSOR Units. Following the format of these protocols the user will find an initial introduction to the concepts pertinent to the protocol, what is the content when performing the given navigation, as well as the necessary resources. This easy to follow protocol shall help anyone interested to quickly see all what is currently offered in RegulonDB, including position weight matrices of transcription factors, coexpression values based on published microarrays, as well as the GENSOR Units unique to RegulonDB that offer regulatory mechanisms in the context of their signals and metabolic consequences.

Keywords

RegulonDB; gene regulation; *Escherichia coli* K-12; database; transcriptional regulatory network

INTRODUCTION

For 25 years, we have been manually curating the scientific literature containing knowledge of the regulation of transcription initiation and the genome organization in transcription units of *Escherichia coli* K-12 in RegulonDB (<http://regulondb.ccg.unam.mx>).

*Corresponding author. collado@ccg.unam.mx; regulondb@ccg.unam.mx.

This is one of the major electronic resources that facilitates access to information and analysis of gene regulation of a bacterial genome. As years have passed, we have kept up to date and have expanded the type of information of the major players involved in understanding the regulation of transcription initiation and its genomic organization of the best-characterized organism, the bacterium *E. coli* K-12. RegulonDB organizes large amounts of data, such as regulatory interactions, transcription units (TUs) and the overlapped TUs into operons, regulons, transcription factors (TFs) and their active and inactive conformation, as well as the elementary genetic-metabolic networks that we call GENSOR units (GUs), defined by all the components from the input signal to the output response of the products for genes controlled by a single regulator, among other concepts (Gama-Castro et al., 2008; Gama-Castro et al., 2011; Gama-Castro et al., 2016; Huerta, Salgado, Thieffry, & Collado-Vides, 1998; Salgado et al., 2006).

As a result of our manual curation, each participating genetic element is gathered, together with its corresponding evidence and references, as well as its interactions with other elements, when known. Evidence codes come from those in EcoCyc and Gene Ontology (GO) (Gene Ontology, 2015; Keseler et al., 2013). We defined what can be called an algebra of evidence, based on a classification of sources of knowledge, which means that multiple weak independent types of evidence for an object or interaction is enhanced and considered a strong evidence (Weiss et al., 2013). If users do not subscribe to our way of classifying evidence, in weak, strong or confirmed, they can always access to the direct evidence and take their own decisions. Currently, RegulonDB gathers 4653 genes, 5843 promoters, 210 TFs, 3261 TF interactions, and several other regulatory elements.

In the last release of RegulonDB version 9.4, we integrated a tool that quantifies the coexpression of all possible gene pairs in the genome assigning them a “coexpression distance” (Gama-Castro et al., 2016). This tool has a rank-based approach for a comparative selection, using data available in COLOMBOS version 2.0, which contains expression profiles for 2470 different contrasting conditions (Pannier, Merino, Marchal, & Collado-Vides, 2017). The rank-based approach normalizes the differences derived from the ranges of correlation coefficients of genes, which allows comparisons of coexpression strengths among genes despite the large variability of expression values. For a historical account see the Background section at the end of this paper.

The search engine was improved in the RegulonDB version 9.0, enabling logic operators and character wildcards. This was achieved through the integration of the ElasticSearch tool (<https://www.elastic.co/>); as well as the gathering of the properties of biological objects to be susceptible for searching.

The biological objects have a context or outlook; thus, RegulonDB contains pages that display the integration of the data related to the context elements. The integrative results pages are Gene, Operon, Regulon, Growth Condition, Sigmulon, GU and sRNA. For example, the Operon page includes information for the operon itself, its different possible TUs, its promoters, regulation, and terminators. Then, if a search is performed with the name of a promoter, the search will return the operon controlled by that promoter.

The main concepts of gene regulation are introduced in each protocol (but see also the background section at the end), followed by how to find, and what is found when searching any of the major players involved in regulation and their specific searches: Genes or Functional Class; Operons, TUs and Promoter; Regulons, TFs and their conformations; GUs, summaries and Coexpression. Except for the basic protocol 1, all the others initiate in the Search button of the main page, and in all of them the specific references link to the corresponding PubMed page. These protocols should be easy to follow by any student or user with no previous knowledge other than basic biology, interested in navigating within RegulonDB.

BASIC PROTOCOL 1. REGULONDB DATABASE: GETTING STARTED

RegulonDB can be accessed through the Web site <http://regulondb.ccg.unam.mx>. The site was designed to navigate through hyperlinked menus or hyperlinked text. The home page has a hyperlinked menu bar located at the top of the page. This menu bar allows getting additional information about RegulonDB; such as Datasets (available files: Experimental, Computational, and High-throughput data files), which are files in text format, frequently utilized by the users of RegulonDB because of their great interoperability, as a universal format. It is also possible to download the full database in several formats (txt, xml or dump files). Through this menu, you can access to several sections: News, Notes, Annotation Process, Glossary, and Evidence Classification.

Several database-associated tools are available for the prediction and visualization of data. For example, the Transcription factor browser, which allows the navigation through each TF. This browser displays a list of all TFs and their detailed statistics, such as total number of regulated genes, regulated operons, etc. Furthermore, with the TF PWMs (position weight matrix) Browser it is possible to navigate through the weight matrices of all TFs. Even more, the use of Web Services is available to obtain the regulatory networks for the analyzed object.

The first protocol shows the content of the RegulonDB site, indicating the options in the menus and the respective content, which can facilitate the users to find the information. Not all options are described, only those that we consider relevant and that are more frequently utilized by the users.

Necessary Resources

Hardware—Computer with Internet access

Software—Any modern Web browser (Firefox, Chrome, Opera, Safari, Internet Explorer, etc)

1. Start your Web browser and go to RegulonDB Web site at <http://regulondb.ccg.unam.mx>.

*The home page of RegulonDB (Figure 1) has a blue menu bar located below the logo RegulonDB, at the top of the page with 5 options: **Home, Features, Integrated Views & Tools, Downloads and Doc &***

Help. The menu and logo will always appear in the same place. The search box appears in the center only on the main page; once navigating through RegulonDB, it will be part of the header of all the pages, along with the logo and the menu, in order to facilitate access to all the features of RegulonDB at any time.

In the footer, there is information for RegulonDB contact and how to cite RegulonDB. This section also shows the version of RegulonDB data and the version of the genome used in the annotation.

2. Click on **Features** option in the menu, a web page with many sections and hyperlinked text is displayed to navigate.

This page contains information about What is RegulonDB, What is new, Summary History with a report integrating all objects through the years, Publications and Credits.

RegulonDB has 3 to 4 releases of data every year; each release includes a summary highlighting what is considered more significant in the most recent curation, as well as the totals of all encoded data.

3. Go to **Integrated Views & Tools**, where the user can find different sections of tools and browsers (Figure 2).
 - a. RegulonDB Overviews. It offers a collection of graphs that show the distributions of different biological objects encoded in RegulonDB. Also, in the transcriptional regulatory network links, different graphical representations of the regulatory network of *E. coli* are available.
 - b. Browse RegulonDB. Different browsers and tools are included in this section. Some of them are:
 - i. The Transcription factor browser allows getting the list of all TFs known in RegulonDB, and shows the total number of genes regulated by every TF, operons, TF Binding sites and regulatory interactions. Each line is a hyperlink to the corresponding regulon page for every TF.
 - ii. Other very useful browser is the *Gensor Unit Groups*. This browser groups the GUs either on the signal that initiates a flux of a regulated process (i.e. GUs related to carbon utilization or metabolism of amino acids), or based on their signal transduction mechanism (two component systems), (Figure 3), offering a direct link to each of them.
 - iii. The DrawingTracesTool is a friendly tool allowing the generation of images for elements related to DNA and involved in gene regulation (such as gene, operon, binding site, promoter, terminator, attenuator, riboswitch and small RNA). The user can type a pair of genome coordinates and select the biological object to be displayed; then the DTT get

all the objects annotated in the database that are in that interval. For example, click on in the DrawingTracesTool link, and for *Absolute genome left position* type 3852220 and 3853712 for *Absolute genome right position*, and deselect *operon* option from the check box. Scroll down and click the *Go* button, the result is a graph with all biological objects in that interval (Figure 4).

- c. RegulonDB Web Services. Web services are used for machine-to-machine communication. RegulonDB offers several web services to extract information, especially about of the transcriptional regulatory network.
4. Go to the **Downloads** option from the main menu, where RegulonDB provides mechanisms to download datasets with different option of formats and contents: Experimental data; Computational predictions; and full RegulonDB database in different formats, such as: TXT, XMLS, DMP files for different database management systems, and BioPAX (Figure 5).
 - a. Go to the Experimental Dataset link. A list of downloadable datasets is shown (Figure 6), such as datasets of sequences of genes and the 5' and 3' UTR sequence of TUs. Datasets containing the complete list of genes, promoters, transcription unit, operons and regulatory interactions with evidence are also available. The most downloaded file in RegulonDB is the *TF - gene regulatory network interaction*, click on the link. A text file is shown, generally the data are tab separated and includes the release version (Figure 7).
5. The **Doc & Help** option includes information about the definitions of biological terms used in RegulonDB in a glossary; it also provides a link for the evidence classification supporting every piece of knowledge we have gathered in the database, essentially based on the methods used to generate them. The common criteria to distinguish our classification into weak, strong or confirmed evidence are explained.

BASIC PROTOCOL 2. SEARCH BY GENE, PRODUCT OR FUNCTIONAL CLASS

As mentioned before, RegulonDB currently has five types of pages that integrate the information in a context, all of them initiate in the Search button of the main page. One of the resulting pages is the Gene page, which contains the information of the gene, product, operon and transcriptional regulation, as well as the respective references.

Several properties of the biological objects integrated in Gene pages are utilized to initiate the navigation that bring users to the gene pages such as the gene name or synonym found in other databases, like GenBank and EcoCyc; the name or synonym of the product; and the

functional Multifun and GO classes. The authors of references are also part of the searching terms.

Additionally, some keywords commonly used for searching genes, for example, “gene”, “orf”, “product”, “all genes”, can also be added for searching a more specific result. For instance, searching for “araC gene” will be more specific than searching only for “araC”. In fact, if any of them is used without the name of the gene, the complete list of the genes registered in the database is obtained.

The logic operators (AND, OR, NOT) and wildcard characters (*, ? and parenthesis) can also be used to condition the searching or to broaden the selections of any query.

In this protocol, we show how to perform a search for genes, and how to revise the content of the gene page.

Necessary Resources

Hardware—Computer with Internet access

Software—Any modern Web browser (Firefox, Chrome, Opera, Safari, Internet Explorer, etc.)

1. Go to RegulonDB Web site at <http://regulondb.ccg.unam.mx>
2. Place your cursor into the Text Search Box and type *arabinose*. It will show a page of results with a list of all those hits where the word in any of the searching terms was found (Figure 8).

The hits found are organized by categories or context pages. Below the title of the page, all categories are shown indicating in parentheses the total of hits found for the query. The categories are also links to the following section where the results appear. The lines of one section represent objects having a match that is highlighted, with the searching term.

3. Go to gene category and select the *araC* gene; the gene page of the *araC* gene is displayed (Figure 9).
4. Use the scroll bar on the right side of the browser window to scroll down the Gene page. The main sections are:
 - a. Gene Tool Bar. Immediately below the main menu there is an icon section to navigate to different context pages (GUs, operon and coexpression). This tool bar will change depending on the context page.
 - b. Gene Context Graph. The graph displays a scale drawing of all objects located within the gene context region. The regulatory elements that have been reported to affect the gene are shown in colors. All other regulatory elements that are located in that region, for which there is no evidence of their effect on the selected gene, are shown in grey. The

display of both types of elements provides a complete picture of the context region.

Each element in the graph displays a tooltip giving details of the element. Also, the other genes in the image are clickable to navigate to their own Gene pages. The icons on top of the graph are used to zoom in and out the image; the expand 5' region icon is used to expand the regulatory region of the gene (Figure 10). The graphic code for objects and colors about the graph can be accessed in the help menu (Figure 11).

- c. Data sections. The five main sections highlighted by a blue bar are: gene, product, operon, transcriptional regulation and references. In Figure 9, all sections were collapsed to see the full page but by default they are expanded.
- d. Gene section. This section displays the information related to the genome region, such as genome positions; the strand; DNA sequence; the gene name and its synonyms. Identifiers to other databases are also shown as hyperlinks.
- e. Product section. The gene product section displays the name of the product and its synonymous; the amino acid sequence; the molecular weight and the isoelectric point, together with the gene ontology and other functional classifications of the products.

When the product of a gene is a transcriptional regulator, as the case of *araC*, in the “Product” section, its “Name” field is a hyperlink to the Regulon page, in this case the AraC regulon.

- f. Operon section. This section shows the operon arrangement to which the gene belongs, its transcriptional units and promoters. From this section, it is possible to connect to the corresponding operon main page.
- g. The Transcriptional Regulation section. It shows those TFs that activate or repress a gene; the *Display Regulation* link connects to the Operon page where the gene belongs to.
- h. Finally, the references section contains a list of all references used in the four sections, in the order in which they appeared in the text, and each of them is a hyperlink to the corresponding PubMed page.

*The genes and their products were initially extracted from GenBank, but RegulonDB in collaboration with Ecocyc, has complemented the information with the papers curated through the years. In fact, when there is any inconsistency in the data reported in a paper and those from RegulonDB, a note is added indicating the correction, for example see the *mak* gene (<http://regulondb.ccg.unam.mx/gene?>*

[term=ECK120001261&organism=ECK12&format=jsp&type=gene](http://regulondb.ccg.unam.mx/?term=ECK120001261&organism=ECK12&format=jsp&type=gene)).

When a new version of the E. coli sequence is released in GenBank, a process to re-annotate all coordinates of the objects having a position in the genome is executed.

BASIC PROTOCOL 3. SEARCH BY OPERON, TRANSCRIPTION UNIT OR PROMOTER

Operon is one of the most important concepts in RegulonDB; a common understanding is that an operon involves two or more genes and their associated regulatory elements, which are transcribed as a polycistronic unit (Jacob and Monod, 1961), but Jacob and Monod soon accepted operons with either one or more genes (Monod, Changeux & Jacob, 1963; Jacob, 1966). In order to capture all genes of the genome within operons, we include the possibility of operons with only one gene; therefore, an operon may contain one or more contiguous genes transcribed in the same direction. It should be noted that, according to this definition, an operon must contain at least one promoter and a terminator, upstream and downstream, respectively. Also, some operons can contain several promoters, and therefore several TUs, conforming complex operons, some of them internally located and thus transcribing only some genes of the full-length operon.

A Transcription Unit is formed by one or more genes transcribed from a single promoter. A TU may also include regulatory protein binding sites affecting this promoter and a terminator. A complex operon with several promoters will contain several TUs; however, according to the definition of operon, at least one TU must include all the genes of the operon (for instance see the *rpsU-dnaG-rpoD*, *glnALG*, *focA-pflB* operons).

The transcription factor binding sites (TFBSs) were, until a few years ago, the main regulatory elements annotated on RegulonDB. But in recent versions, other regulatory elements have been included in the operon context, such as those involved in the regulation by sRNAs and by the effector molecule ppGpp.

Given these definitions, in this protocol, we will focus in exemplifying the searches that bring us to the page of the operon, and its structural elements such as TUs, promoters, terminators, and their regulated genes.

Necessary Resources

Hardware—Computer with Internet access

Software—Any modern Web browser (Firefox, Chrome, Opera, Safari, Internet Explorer, etc.)

1. Go to RegulonDB Web site at <http://regulondb.ccg.unam.mx>
2. Different ways of accessing the operon context page will be tested.

Type `operon` or `operons` in the text search box. A list of many genes, operons and regulons are obtained that match with the query.

- a. The total matching records are shown below the title. Using the words *operon* or *operons* alone, all operons annotated in RegulonDB are retrieved. Click on *Operon (2630)* link to go the operon section list (currently 2630 operons, RegulonDB v9.4). The complete operon list is displayed ordered alphabetically. Every row, in any result section, begins with the name of the main object, in this case the operon name; the rest of the text is the field or term where the query matched. The results are clickable to go to a specific operon context page.
 - b. The same result is obtained using `transcription unit`, `all operons`, `all transcription units` or `all TUs` query words.
3. Go to the top in the main page and now type `all promoters`. The summary shows *Operons*, as a result for *Operon*, this result shows all the operons (1372 of them, in RegulonDB v9.4) with at least one promoter annotated in the database.

Other useful query is searching by publication; RegulonDB accepts the author name, title or the PMID. Type `Muller V` in the search box; then, the genes and operons associated with the reference are displayed.

Other terms that can be used to search and get an operon result are the promoter; identifier; operon identifier; operon name; transcription unit name; promoter name; promoter sequence; the genes that belong to the operon; synonyms of operon; transcription unit; functional class or gene ontology (GOs) linked to the operon genes. For example, typing `fliFp1` in the text search box, the result obtained is a sigma70 promoter for the *fliFGHIJK* operon.

4. Finally, typing `fliF operon` in the Text Search Box. Since *fliF* is the first gene of the *fliFGHIJK* operon, then the search returns the operon where the *fliF* gene belongs to. Click in the result.
5. The Operon Result Page. The page is organized as follow: the toolbar (the gray bar at the top); the operon graph structure; the operon and its transcription unit(s); the list of evidence types and the list of references (Figure 12).
 - a. The operon tool bar displays in addition to the help, two options: The Genome browser and the coexpression tool. The last one is a tool to see the coexpression patterns in microarrays experiments using the genes from the operon (this tool is explained in protocol 6).
 - b. Operon Graph Structure. The graph has two parts: a) the first one is the DNA operon region where all DNA objects are drawn (the graph is not to scale, to see the gene context graph at scale, go to the gene page). The numbers at the left and right of the graph represent the genome position. The genes are clickable to navigate to their corresponding Gene pages; while a click in the promoters links them to their

corresponding TU in the data section below. b) the second one is the Transcription Unit graphs representing each TU with a line. The label at the beginning of each line corresponds to the name of the TU, and it is a hyperlink to the corresponding TU section.

- c. The operon and its TUs. Scroll down until you see the first TU section. There is a section for each TU in the operon; a TU section includes the list of transcribed genes; the promoter information that contains the absolute genome position and the transcriptional start initiation +1; the sigma factor; the distance from the promoter to the first gene of the TU; the promoter sequence showing 80 base pairs, containing the -35 and -10 boxes and the +1 position (Figure 13); notes from the curators about the promoter; the evidence and references that support the information. The TU also includes information about the site where transcription ends.

Different types of regulation are also part of a TU. In RegulonDB the registered regulation includes transcription factors, sRNA (see *ompR* operon) and RNA *cis*-regulatory elements (see *glnALG* operon). The regulation by TF is shown in the TF binding site section and it includes TF; the promoter; the effect of the TF on the promoter (activator, repressor or unknown); the absolute genome coordinates of the binding site; the distance between the promoter and the central position of the TFBS; the TFBS sequence in capital letters with additional 10 base pairs on each side; when the growth conditions are known this information is included; and finally, the evidence and references that support the data.
- d. The evidence code and description used in the Operon page is listed with links to our Evidence Classification.
- e. The last section is the references; the number used in the different sections of the Operon page, in the reference attribute, is listed here. As already mentioned, each reference is a link to its PubMed page.

BASIC PROTOCOL 4. SEARCH BY TRANSCRIPTION FACTOR, REGULON OR TF CONFORMATION

To exemplify the searches and the content of the website corresponding to a Regulon, we need first to know the definitions of the concepts as they are used in RegulonDB.

One of the main concepts is a TF or regulatory protein; it is a protein or a complex of proteins (since it can be a dimer or multimer) with the function of activating or repressing the transcription of a TU through binding to specific DNA sites. More precisely, we call them “DNA binding transcription factors”.

Another concept is the effector; in RegulonDB, the effector is a metabolite that binds to a TF altering its conformation. Usually one of the conformations is the active one capable of specific binding to its TFBSs, whereas the other one is inactive, unable to bind. Most effectors in *E. coli* are metabolites that bind non-covalently to an allosteric TF site. The covalent modifications are also included as effectors, i.e. phosphorylation by TFs belonging to the two-component regulators family.

Finally, we also group all the TFBS involved in the regulation of a promoter, in a so-called “regulatory phrase”. This usually consists of a combination of proximal and remote binding sites for one or more TFs, which control the expression of the same promoter. TFBSs are classified as *proximal* or *remote* based on their relative distance to their regulated promoter. (See the glossary in RegulonDB and the background section at the end of this paper).

TFs are usually allosteric proteins that bind specifically to their operator DNA sites, either in the *holo* (with the effector bound) or *apo* (without effector) conformation to regulate gene expression. RegulonDB refers to a functional *holo* conformation when the TF binds to DNA as a complex bound to an effector, which can be either a noncovalently bound small molecule, or a covalent modification, such as phosphorylation in a two-component system. A TF binds in an *apo* conformation when the TF does not require any effector to bind its sites; contrary, in this case, binding to an effector leads to unbinding of the TF. For instance, CRP binds to its specific binding sites once bound to cAMP, its allosteric small ligand; whereas the LacI repressor binds to DNA as a protein in *apo* conformation, and unbinds in the presence of allolactose, its allosteric modifier (Balderas-Martinez et al., 2013).

Another important concept used in RegulonDB is Regulon. The regulon was coined in the study the arginine biosynthetic system regulated by ArgR, as a set of genes subject to the regulation of one and only one regulator (Maas, 1964). 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 but 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 their genes with other regulators.

Finally, alignment and position scoring weight matrix (PSWM) are concepts related to TFBS. The PSWM represents a collection of aligned binding sequences for the same transcriptional regulator. It is derivative of a multiple alignment of such sites. Each row corresponds to one of the letters of the relevant alphabet -e.g., 4 rows in the case of DNA. Each column corresponds to one of the positions within the aligned sites. A frequency matrix contains the frequency of the four nucleotides at each position. Our previous work has provided a methodology to assess the quality of the binding sites. Users are invited to see the details in (Medina-Rivera et al., 2011).

In this protocol, we show how to perform searches to obtain the information of a Regulon and we will see what they contain; the regulon FlhDC is used as an example.

The FlhD and FlhC proteins are transcriptional factors that form heterohexamers (FlhD4C2) (Wang, Fleming, Westbrook, Matsumura, & McKay, 2006). The FlhDC TF is the main

regulator of genes for bacterial flagellum biogenesis and swarming migration (Claret & Hughes, 2002; Stafford, Ogi, & Hughes, 2005). This heteromeric regulator activates class II genes involved in the flagellar basal body (Claret & Hughes, 2002; Ikebe, Iyoda, & Kutsukake, 1999; Liu & Matsumura, 1994; Stafford et al., 2005); proteins of the export machinery; the flagellar σ subunit (FlhA); and its anti- σ factor, FlgM (Claret & Hughes, 2002; Stafford et al., 2005). A microarray analysis showed that the master regulator FlhD4C2 regulates several nonflagellar genes, but the direct effects have not been determined. The FlhDC Regulon was comprehensively mapped using ChIP-seq and RNA-seq (Fitzgerald, Bonocora, & Wade, 2014, 2015).

Necessary Resources

Hardware—Computer with Internet access

Software—Any modern Web browser (Firefox, Chrome, Opera, Safari, Internet Explorer, etc.)

1. Go to RegulonDB Web site at <http://regulondb.ccg.unam.mx>.
2. Type all regulons in the Search Text Box. The summary results below the title shows *Regulon (213)*, counting all the regulons available in RegulonDB to date.

Also, using all transcription factor and all tfs, the same regulon list is obtained.
3. Other terms that can be used for the regulon searching are the TF name; the active conformation; the effector name; the TFBS sequence; and their synonyms. Type or copy and paste the sequence
gaccatttttGCGTTTATTCCGCCGATAacgcgcgcg, which is a binding site of FlhDC TF. Two matches are found, one in the promoter sequence of the *flgBCDEFGHIJ* operon, in the Operon section, and the other in a binding sequence of FlhDC TF in the Regulon page.
4. Now type FlhDC regulon in the Search Text box. The Regulon Page is displayed with blue bars collapsed, to see a full visualization of this page (Figure 14).
5. Scroll down/up for full web page navigation. The different sections identified with a blue bar are expanded.
6. Regulon Toolbar is at the top of the page with only an icon for the Coexpression tool. All the genes of the regulon are sent as parameter to the coexpression tool to obtain their Coexpression values (see protocol 6).
7. A brief statement of current knowledge of the TF is shown in a summary section; almost all TFs have a summary, this is a rich integrated text generated by curators. To see the complete text, click on *Read more >* link.

The Transcription Factor section contains many characteristics related to the TF, such as the functional and non-functional conformations

*(shown when they are known); the sensing class based on the type of the effector, classified as external or internal, or both (dual), according to the origin of the effector or metabolites that usually bind allosterically to the TFs (Janga, Salgado, Collado-Vides, & Martinez-Antonio, 2007; Martinez-Antonio, Janga, Salgado, & Collado-Vides, 2006); the name of the genes producing the TF or a subunit of the TF (see the *fhID* and *fhIC* genes in the *FhlDC* example), and the operon or TU that encodes the TF genes, also appears in this section.*

8. Use the scroll bar and go to the Regulon section, here the full list of the regulated genes and regulated operons are shown, ordered alphabetically, with a link to their corresponding Gene and Operon pages. RegulonDB has incorporated some useful functional classification of the gene products, one of them is the multifunctional terms; they are listed with the total number of genes that belongs to the same category. The simple and complex regulons, previously defined above, appear with links to a web page that describes the TFBS of all TFs that bind in the promoter region.

- a. At the end of the Regulon section, you will find the Simple and Complex regulatory phrases. The *[FhlDC+] (10)* is other representation for a simple or strict regulon and represents the different promoters regulated by FhlDC as activator, where the binding position for this TF is known; the number in parentheses indicates the total of promoters.
- b. Click on the link *[FhlDC+] (10)* and the Regulatory Phrase is displayed (Figure 15). The web page contains a map of the TF behavior when it acts as an activator. The table indicates the FhlDC TFBS classified as proximal or remote with respect to the promoter.

Proximal sites are those within the interval from -93 to +30 with respect to the Transcription Start Site (+1); from where the TF can in principle directly interact with the RNA polymerase. All other sites are considered remote, either upstream or downstream (Collado-Vides, 1992; Collado-Vides et al., 2009).

9. Move down to the TFBS arrangements section to see all the binding sites (Figure 16).
 - a. The table has different functions to show the data, for example, there is a search text box, right at the top, to filter the results according to the written text. For instance, typing “-68” will only show the binding sites at -68.
 - b. Also, there are columns that can be sorted using the order symbol available at the end of the column name.
 - c. Each row is a regulatory interaction between the TF with its functional or active conformation (see for instance in the AraC regulon the different conformations); its regulatory function (activator, repressor,

dual or unknown); the regulated promoter with its sigma type; the distance from the central position of the TFBS to the TSS of the regulated promoter; the distance to the beginning of the first downstream regulated gene; the list of regulated genes; sequence of TFBS; evidence and references. Note that for a given TF-promoter interaction, each TFBS is described in separate rows.

10. Different views of the TFBS can be displayed. Go to the end of the blue bar of this table, and click on the $+$ symbol. Select the *DNA binding sites* option; the result is a table (Figure 17) showing only the unique or non-redundant binding sites, eliminating repetitions of those sites that regulate more than one promoter.
11. Now, in the same menu $+$, select *Graphs TF and coregulators* and select *relative distance to gene*, in the sub menu. The image displayed is a view of all binding sites for FlhDC and its coregulators on target genes, where each line is the first gene of a TU (Figure 18).
12. Scroll down and go to the Alignment and PSSM section. PSSMs are built using the annotated binding sites of TFs. A matrix is built for all TFs with four or more annotated sites. The logo is a clickable figure to go to the PSSM process where the evaluation of the quality of the matrix is done (Medina-Rivera et al., 2011).
13. Move down to the *Evolutionary conservation of regulatory elements* section.
 - a. Click on the TF-target gene evolutionary conservation link icon. This analysis shows how the regulatory interaction is conserved in gamma-proteobacteria (Figure 19).
 - b. To see an example, click on the icon in the footprint conservation of *flgA* gene. The result is the process to find conservation of the interactions (Figure 20). In general, an interaction is conserved if there is a significant overrepresentation of TFBSs in the set of orthologous promoters of the target gene. The set of orthologous promoters for *flgA* was scanned with the annotated matrix for FlhDC. The second link in this page with label *Ortholog predicted binding sites graph* is the graphical representation of the binding sites found in the orthologous regions.
14. The last two sections are Evidence and References, records made for all the codes and numbers used in the Regulon page; here the complete name of the evidence and the references are listed.

BASIC PROTOCOL 5. SEARCH BY GENSOR UNITS

The ability of a cell to respond to changes inside the cell or in the environment initiates when the new signal or stimulus is sensed and transmitted through a series of concatenated reactions, called signal transduction or transduction pathways. These events bring into action genetic switches (or regulatory interactions) that modify gene expression of regulated genes to produce the response given the signal that is encoded in the genome of the cell. We

defined elementary genetic Sensory Response Units, or Gensor Units to capture this larger context for every regulon. A GU is defined by four components: i) the signal or stimulus, ii) the signal transduction pathway, iii) the regulatory mechanisms governing the expression of genes, and iv) the response resulting from the modified gene expression of the affected set of target genes (Ledezma-Tejeida, Ishida, & Collado-Vides, 2017).

This protocol explains how to browse and identify the components of a GU, using the BetI GU as example. BetI is a small GU, facilitating the identification of the four components of a GU.

The BetI GU. Choline is the signal that is transported through the membrane; genetic switches cause the repression of two TUs, *betT* and *betIBA*, which are involved in the transformation of choline into glycine betaine. Choline binds to BetI preventing it to repress the genes coding for its transport and transformation into glycine betaine. The BetI regulon is an inducible system in response to the appearance of choline in the medium, which will induce the expression of the genes involved in its transport of choline and conversion. Note that since the appearance of the signals can provoke the effector to activate or inactivate the TF, repressible systems can be executed by repressors or activators, as well as inducible systems ((Balderas-Martinez et al., 2013).

Necessary Resources

Hardware—Computer with Internet access

Software—Any modern Web browser (Firefox, Chrome, Opera, Safari, Internet Explorer, etc.)

1. Go to RegulonDB Web site at <http://regulondb.ccg.unam.mx>
2. In the Text Search Box type BetI gensor unit (or BetI gu, BetI gensor or simply BetI).
3. Select the BetI GU and the Gensor Unit is displayed (Figure 21).

The Gensor Unit web page contains three main sections: the GU map, the GENSOR unit information, and its reactions.

- a. The Gensor map section. It is a graph representation of GENSOR units created using CellDesigner tool (Matsuoka, Funahashi, Ghosh, & Kitano, 2014). BetI GU shows all the components involved in response to the presence of choline. The graphic code and colors can be accessed in the blue bar, in the help icon.

All the GU maps manage four colors to fill the components. Those filled with blue color correspond to the components involved in the signal and signal transduction; the yellow ones represent the genetic switch; the green ones are the adequate response, and the gray ones are elements whose role in the GU is not clear. Some map components are clickable, those that

participate in the genetic switch are hyperlinks to the operon page, and the rest go to their gene page.

- b. The GENSOR unit information section. A description or summary of the GU is available in this section. Also, all the gene ontology terms associated to the molecules are displayed.
 - c. Reactions section. All the reactions are numbered on the map, and these numbers are used in the reaction section to access a specific reaction and to display all its reactants or products.
4. Go to the *GENSOR unit map* section and click on the plus (+) icon at the end of the blue bar. This is an option to export the GENSOR unit map to xml or sbml format for CellDesigner tool.

CellDesigner is a structured diagram editor for drawing gene-regulatory and biochemical networks (Matsuoka et al., 2014). SBML (Systems Biology Markup Language) is a machine-readable format (XML) for representing computational models in systems biology; its focus is to describe the systems of biochemical reactions (<http://sbml.org>).

5. Another way to search for a GENSOR unit is through the GENSOR Unit Groups option in the Integrated Views & Tools from the main menu.

To assemble a GU for each TFs in RegulonDB were used a data-driven approach. The program retrieved from the RegulonDB database the genes directly regulated by a TF (regulon); its known effectors; its active/inactive conformations; and the TF's regulatory effect over the regulated genes. From the EcoCyc database (Keseler et al., 2013) are automatically obtained the gene products of the regulated genes; the reactions catalyzed by the gene products; the substrates and products of the catalyzed reactions; and the protein complexes in which the gene products participate.

BASIC PROTOCOL 6. SUMMARY OF GENE REGULATION AND COEXPRESSION

The last version of RegulonDB incorporated coexpression analysis using the global gene expression data, from microarrays and RNA-seq assays, contained in the COLOMBOS database (Engelen et al., 2011; Meysman et al., 2014). High-throughput experiments, including microarrays and RNA-seq assays, enable massive expression measurements of thousands of genes simultaneously. With this information, it is possible to determine which genes show a similar expression pattern in a large set of growth conditions or genetic backgrounds. Coexpression of genes may strongly suggest a related cellular function and a link in transcriptional regulation.

In the previous protocols, we indicated that in distinct pages, like those for gene, operon, and regulon, there is an icon in the Toolbar connecting to the visualization tool of the results of

coexpression. Thus, although we will use a list of genes to exemplify the description of the tool, all steps will be valid regardless of where the tool is accessed.

In this protocol, it was searched if genes reported to be positively controlled only by FlhDC, the master transcriptional regulator of the flagellar genes, are indeed coexpressed. The first genes of different operons regulated by FlhDC are tested.

Necessary Resources

Hardware—Computer with Internet access

Software—Any modern Web browser (Firefox, Chrome, Opera, Safari, Internet Explorer, etc.)

1. Go to <http://regulondb.ccg.unam.mx> web site
2. Place your cursor into the Text Search Box and type `ygbK recC fliL ycgR fliD gltL flgB flhE flgK` coexpression. The direct access to the coexpression page is using the word *coexpression* at the end of the query.

The result is the Coexpression page, as shown in Figure 22; the web page organizes differently the data compared to the pages previously seen; the data is organized in three tabs.

3. Click on *My query genes* tab section. This section gives an overview of different features about the query genes, such as their transcriptional regulation; their product; the operon to which they belong; and the description of associated gene ontologies.

In the example, all the genes are regulated only by FlhDC, which is a simple regulon. There are nine operons and only the first gene of each operon was used in the query list. Each column in the table has an icon next to the column name to sort the data. The gene name, the operon, and TFs are clickable to their corresponding Web page.

4. Select the *Coexpression by gene* tab section. This section allows the user to display the top 50 genes that are most highly coexpressed with each query gene (Figure 23).
 - a. The query genes are ordered alphabetically, and the top highly coexpressed genes are ordered by their rank. Each tab displayed represents one of the query genes.
 - b. Within each tab or gene two sections are shown, the first section is “GENE INFORMATION” that displays information about the transcriptional regulation of that gene; the second section is “THE TOP 50 COEXPRESSED GENES” that shows the top 50 genes highly coexpressed with the respective query gene and their information on transcriptional regulation.

Here, it is expected that the top coexpressed genes will be those located in the same operon of the respective query gene and, in this case, the other query genes, since all these belong to the FlhDC regulon. As shown in figure 23, the flgB, flgK, flhE, fliD, fliL, gltL and ycgR genes are evidently coexpressed with genes located in their respective operons and with some other genes. In contrast, the recC and ygbK query genes are not coexpressed with the other query genes tested, suggesting that the expression of recC and ygbK is also controlled by other regulators that could act in conditions where the regulation by FlhDC is not active.

5. The Matrix HeatMap tab. It is a graphical representation of the coexpression of the genes that belong to the query list, with their top 50 coexpressed genes (Figure 24).
 - a. The color palette shows the gradient associated with the ranks. The lowest ranks indicate high coexpression and are identified by green color; the highest ranks indicate no coexpression and are shown in red. The rank is a measure that shows how close is a gene from another, comparing its coexpression with all the other genes, being 1 the minimum difference between a pair of highly coexpressed genes.
 - b. By default, the query genes are ordered alphabetically and the top 50 coexpressed genes for the first query gene (*flgB* for this example) are displayed. For other query genes, it is shown the coexpression rank color of the top 50 genes coexpressed with the first query gene (*flgB*). There is a dropdown menu above the graph where is possible to select the query gene to be shown as the first query gene, and thus to get its respective top 50 coexpressed genes. Also, the rank value can be seen when clicking the *View coexpression value* check box.

The ranges go from 1 to 4167 (number of genes - 1). If two genes have SCR=1, it means that these two genes have the highest coexpression level with each other, compared to their coexpression with all the other genes. If two genes have SCR=4167, it means that these two genes have the lowest coexpression level with each other.

COMMENTARY

Background Information

The seed of RegulonDB initiated 25 years ago with the initial review of transcription initiation in sigma 70 and sigma 54 promoters in *E. coli*, as part of the postdoctoral stay of the Professor Julio Collado-Vides in the laboratory of the Professor Boris Magasanik, and with the collaboration and support of Jay Gralla, in the Department of Biology at MIT. An important observation with the 116 sigma70 regulated promoters analyzed, was that the

regulation of sigma 70 promoters requires a *proximal* binding site, in contrast to the regulation of sigma 54 promoters where the proximal site is not required (Collado-Vides, Magasanik, & Gralla, 1991). This tendency was confirmed years later, with a larger collection of promoters analyzed (Collado-Vides et al., 2009). The main initial motivation for this review was the interest in building a grammatical model to capture the combinatorial arrangements of multiple transcription binding sites, with proximal sites being obligatory while remote sites considered optional in the model (Collado-Vides, 1992). Years later, this model was the basis for a computational development that could diminished false positives by a factor of 8-fold (Thieffry, 1996). The first paper describing RegulonDB appeared in 1998 (Huerta et al., 1998), and since then, our laboratory has been curating continuously the literature on transcriptional regulation and operon organization in *E. coli* K-12. The collaboration with Prof. Fred Blattner with the sequencing of the full-length genome of *E. coli* K-12 (Blattner et al., 1997) made a strong impact, motivating us to new challenges, such as the whole genomic computational prediction of TFBS (Thieffry, Salgado, Huerta, & Collado-Vides, 1998), promoters and operons (Salgado, Moreno-Hagelsieb, Smith, & Collado-Vides, 2000). In those years, we initiated our collaboration with Prof. Peter Karp, as editors of transcription regulation, and since then our work feeds both EcoCyc and RegulonDB. We have expanded our work through the years, increasing the quality of evidence types, other types of objects, such as those related to the regulation mediated by small RNA, and their details; the evolutionary conservation of binding sites; computational predictions and the addition of high throughput datasets, among others. Furthermore, we have also enriched RegulonDB with concepts such as regulatory phrases and, more recently, with GUs. As mentioned, a GENSOR Unit is an integrative concept that links environmental signals, the signal processing, the associated regulatory interactions and the regulated response as metabolic and cellular capabilities (Ledezma-Tejeida, Ishida, & Collado-Vides, 2017). RegulonDB is an enabling resource for bioinformatics benchmarking in the development of novel methods, network analyses, and the design of systems and synthetic biology, among several other things. Ongoing projects include an ontology and controlled vocabulary for growth conditions as well as including knowledge from ChIP and related genomic methodologies. We hope to continue the enrichment and expansion of RegulonDB in the future.

Critical Parameters

Although not specified in any protocol, PWMS can be used to search for similar TFBSs in selected DNA sequences. In that case, users must be aware of the parameters used in their search. Each TF motif has different information content so that the quality of their results will vary from the excellent conserved motif of LexA to poor motifs like that of IHF. The best is to use the p value at the level where the curve of found sites separates from the random model. For details see (Medina-Rivera et al., 2011).

Troubleshooting

It is not precisely troubleshooting, but the reader must be aware that these specific protocols are those for RegulonDB version 9.4. Since we continue working on this resource, most likely there will be changes in the future; for instance, the Regulon page will contain regulatory interactions obtained from various ChIP experiment in the near future. It is

always important to double-check the genome sequence version, indicated at the bottom of the main page in small letters.

Acknowledgments

We want to thank to Muñiz Rascado Luis José, Del Moral Chávez Víctor and Bonavides Martínez César for their technical support. Also, we acknowledge funding from UNAM, from CONACyT Fronteras de la Ciencia (project 15), and the almost continuous funding, since the year 2000, from the National Institutes of Health, NIGMS (current grant number 5R01GM110597-03).

LITERATURE CITED

- Balderas-Martinez YI, Savageau M, Salgado H, Perez-Rueda E, Morett E, Collado-Vides J. Transcription factors in *Escherichia coli* prefer the holo conformation. PLoS One. 2013; 8(6):e65723.doi: 10.1371/journal.pone.0065723 [PubMed: 23776535]
- Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, ... Shao Y. The complete genome sequence of *Escherichia coli* K-12. Science. 1997; 277(5331):1453–1462. [PubMed: 9278503]
- Claret L, Hughes C. Interaction of the atypical prokaryotic transcription activator FlhD2C2 with early promoters of the flagellar gene hierarchy. J Mol Biol. 2002; 321(2):185–199. [PubMed: 12144778]
- Collado-Vides J. Grammatical model of the regulation of gene expression. Proc Natl Acad Sci U S A. 1992; 89(20):9405–9409. [PubMed: 1409648]
- Collado-Vides J, Magasanik B, Gralla JD. Control site location and transcriptional regulation in *Escherichia coli*. Microbiol Rev. 1991; 55(3):371–394. [PubMed: 1943993]
- Collado-Vides J, Salgado H, Morett E, Gama-Castro S, Jimenez-Jacinto V, Martinez-Flores I, ... Santos-Zavaleta A. Bioinformatics resources for the study of gene regulation in bacteria. J Bacteriol. 2009; 191(1):23–31. DOI: 10.1128/JB.01017-08 [PubMed: 18978060]
- Engelen K, Fu Q, Meysman P, Sanchez-Rodriguez A, De Smet R, Lemmens K, ... Marchal K. COLOMBOS: access port for cross-platform bacterial expression compendia. PLoS One. 2011; 6(7):e20938.doi: 10.1371/journal.pone.0020938 [PubMed: 21779320]
- Fitzgerald DM, Bonocora RP, Wade JT. Comprehensive mapping of the *Escherichia coli* flagellar regulatory network. PLoS Genet. 2014; 10(10):e1004649.doi: 10.1371/journal.pgen.1004649 [PubMed: 25275371]
- Fitzgerald DM, Bonocora RP, Wade JT. Correction: Comprehensive Mapping of the *Escherichia coli* Flagellar Regulatory Network. PLoS Genet. 2015; 11(9):e1005456.doi: 10.1371/journal.pgen.1005456 [PubMed: 26379184]
- Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, ... Collado-Vides J. RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res. 2008; 36(Database issue):D120–124. DOI: 10.1093/nar/gkm994 [PubMed: 18158297]
- Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, ... Collado-Vides J. RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Sensor Units). Nucleic Acids Res. 2011; 39(Database issue):D98–105. DOI: 10.1093/nar/gkq1110 [PubMed: 21051347]
- Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muniz-Rascado L, Garcia-Sotelo JS, ... Collado-Vides J. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 2016; 44(D1):D133–143. DOI: 10.1093/nar/gkv1156 [PubMed: 26527724]
- Gene Ontology C. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015; 43(Database issue):D1049–1056. DOI: 10.1093/nar/gku1179 [PubMed: 25428369]
- Huerta AM, Salgado H, Thieffry D, Collado-Vides J. RegulonDB: a database on transcriptional regulation in *Escherichia coli*. Nucleic Acids Res. 1998; 26(1):55–59. [PubMed: 9399800]

- Ikebe T, Iyoda S, Kutsukake K. Promoter analysis of the class 2 flagellar operons of *Salmonella*. *Genes Genet Syst.* 1999; 74(4):179–183. [PubMed: 10650844]
- Jacob F. Genetics of the Bacterial Cell. *Science.* 1966; 152(3728):1470–1478. [PubMed: 5327200]
- Jacob F, Monod J. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J Mol Biol.* 1961; 3:318–356. [PubMed: 13718526]
- Janga SC, Salgado H, Collado-Vides J, Martinez-Antonio A. Internal versus external effector and transcription factor gene pairs differ in their relative chromosomal position in *Escherichia coli*. *J Mol Biol.* 2007; 368(1):263–272. DOI: 10.1016/j.jmb.2007.01.019 [PubMed: 17321548]
- Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, ... Karp PD. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* 2013; 41(Database issue):D605–612. DOI: 10.1093/nar/gks1027 [PubMed: 23143106]
- Ledezma-Tejeda D, Ishida C, Collado-Vides J. Genome-Wide Mapping of Transcriptional Regulation and Metabolism Describes Information-Processing Units in *Escherichia coli*. *Front Microbiol.* 2017; 8:1466.doi: 10.3389/fmicb.2017.01466 [PubMed: 28824593]
- Liu X, Matsumura P. The FlhD/FlhC complex, a transcriptional activator of the *Escherichia coli* flagellar class II operons. *J Bacteriol.* 1994; 176(23):7345–7351. [PubMed: 7961507]
- Maas WK. Studies on the Mechanism of Repression of Arginine Biosynthesis in *Escherichia coli*. II. Dominance of Repressibility in Diploids. *J Mol Biol.* 1964; 8:365–370. [PubMed: 14168690]
- Martinez-Antonio A, Janga SC, Salgado H, Collado-Vides J. Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. *Trends Microbiol.* 2006; 14(1):22–27. DOI: 10.1016/j.tim.2005.11.002 [PubMed: 16311037]
- Matsuoka Y, Funahashi A, Ghosh S, Kitano H. Modeling and simulation using CellDesigner. *Methods Mol Biol.* 2014; 1164:121–145. DOI: 10.1007/978-1-4939-0805-9_11 [PubMed: 24927840]
- Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M, Salgado H, Collado-Vides J, van Helden J. Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res.* 2011; 39(3):808–824. DOI: 10.1093/nar/gkq710 [PubMed: 20923783]
- Meysman P, Sonogo P, Bianco L, Fu Q, Ledezma-Tejeda D, Gama-Castro S, ... Engelen K. COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia. *Nucleic Acids Res.* 2014; 42(Database issue):D649–653. DOI: 10.1093/nar/gkt1086 [PubMed: 24214998]
- Monod J, Changeux JP, Jacob F. Allosteric Proteins and Cellular Control Systems. *J Mol Biol.* 1963; 6:306–329. [PubMed: 13936070]
- Pannier L, Merino E, Marchal K, Collado-Vides J. Effect of genomic distance on coexpression of coregulated genes in *E. coli*. *PLoS One.* 2017; 12(4):e0174887.doi: 10.1371/journal.pone.0174887 [PubMed: 28419102]
- Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, ... Collado-Vides J. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.* 2006; 34(Database issue):D394–397. DOI: 10.1093/nar/gkj156 [PubMed: 16381895]
- Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci U S A.* 2000; 97(12):6652–6657. DOI: 10.1073/pnas.110147297 [PubMed: 10823905]
- Stafford GP, Ogi T, Hughes C. Binding and transcriptional activation of non-flagellar genes by the *Escherichia coli* flagellar master regulator FlhD2C2. *Microbiology.* 2005; 151(Pt 6):1779–1788. DOI: 10.1099/mic.0.27879-0 [PubMed: 15941987]
- Thieffry D. *Escherichia coli* as a model system with which to study cell differentiation. *Hist Philos Life Sci.* 1996; 18(2):163–193. [PubMed: 9122371]
- Thieffry D, Salgado H, Huerta AM, Collado-Vides J. Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12. *Bioinformatics.* 1998; 14(5):391–400. [PubMed: 9682052]
- Wang S, Fleming RT, Westbrook EM, Matsumura P, McKay DB. Structure of the *Escherichia coli* FlhDC complex, a prokaryotic heteromeric regulator of transcription. *J Mol Biol.* 2006; 355(4):798–808. DOI: 10.1016/j.jmb.2005.11.020 [PubMed: 16337229]

Weiss V, Medina-Rivera A, Huerta AM, Santos-Zavaleta A, Salgado H, Morett E, Collado-Vides J. Evidence classification of high-throughput protocols and confidence integration in RegulonDB. Database (Oxford). 2013; 2013:bas059.doi: 10.1093/database/bas059 [PubMed: 23327937]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Significance Statement

Escherichia coli K-12 is most likely the best studied free-living organism. The basic concepts we use to understand regulation of gene expression were proposed after research done in this organism by Francois Jacob and Jacques Monod. RegulonDB is the database that gathers knowledge of the regulation of transcription initiation and the genome organization in transcription units in the genome. This organized knowledge is the product of manual continuous curation of the original scientific literature during the last 25 years. This paper explains the access to the current (version 9.4) contents of this evolving resource.

Guiding access in RegulonDB to the organized knowledge of transcriptional regulation in
Escherichia coli K-12 Unit Title

Figure 1.

The RegulonDB home page contains the main menu at the top of the page to help users navigate through the site; the *Search Text Box* allows users to search the site.

Figure 2.
The *Integrated Views & Tools* menu option provides the user with a set of tools, browsers, and data views to perform different activities.

Figure 3. The GENSOR Unit Browser provides access to each of the GUs, curated to date, according to their mechanism of transduction or with the signal that initiates the flux of regulated processes.

DrawingTracesTool Result

Image full size view and dataset file

Figure 4.

DrawingTracesTool is a user-friendly tool that allows the user to generate images of elements related to DNA, involved in gene regulation. The image shows an example of the results page that contains the graph of a DNA region of the genome and the regulatory elements contained in that region, annotated in RegulonDB.

Figure 5.
Downloads menu provides options to download datasets and even the full database in different formats.

Downloadable Experimental Datasets		
RegulonDB 9.4, 05-08-2017.		
We kindly ask you to cite RegulonDB properly. By downloading any of the files below, the user agrees to be bound to the terms and conditions of the electronic End User License Agreement		
Datasets supported by literature with experimental evidence:		
Description	File	
Sequences	Gene Sequence	Download
	5' and 3' UTR sequence of TUs	Download
Gene - Product	All gene products	Download
	sRNA genes	Download
Transcriptional Factors - Functional conformation	Download	
TF binding sites	Download	
Regulatory Network Interactions	TF - gene interactions	Download
	TF - operon interactions	Download
	TF - TU interactions	Download
	TF - TF interactions	Download
	Sigma - gene interactions	Download
	Sigma - TU interactions	Download
	Alon and MA interactions	Download
	sRNA - gene interactions	Download
Promoters	All Promoters	Download
	Sigma 70	Download
	Sigma 54	Download
	Sigma 38	Download
	Sigma 32	Download
	Sigma 28	Download

Figure 6.
The Experimental Datasets section contains the RegulonDB Datasets supported by literature, with experimental evidence, release 9.4. The most downloadable dataset corresponds to the Regulatory Network Interactions.


```

# Contact
#
#   Person: RegulonDB Team
#   Email: regulondb@ccg.unam.mx
#   Contact Web Page: http://regulondb.ccg.unam.mx/menu/about_regulondb/contact_us
#
#
# Release: 9.4 Date: 05-08-2017
#
# Columns:
# (1) Transcription Factor (TF) name
# (2) Gene regulated by the TF (regulated gene)
# (3) Regulatory effect of the TF on the regulated gene (+ activator, - repressor, +/- dual, ? unknown)
# (4) Evidence that supports the existence of the regulatory interaction
# (5) Evidence Type [Weak, Strong, Confirmed] For details see: http://regulondb.ccg.unam.mx/evidenceclassification
#
AcrR   acrA   -   [BCE, BPP, GEA, HIBSCS] Weak
AcrR   acrB   -   [BCE, BPP, GEA, HIBSCS] Weak
AcrR   acrR   -   [AIBSCS, BCE, BPP, GEA, HIBSCS] Weak
AcrR   flhC   -   [GEA, HIBSCS] Weak
AcrR   flhD   -   [GEA, HIBSCS] Weak
AcrR   marA   -   [BPP, GEA, HIBSCS] Strong
AcrR   marB   -   [BPP, GEA, HIBSCS] Strong
AcrR   marR   -   [BPP, GEA, HIBSCS] Strong
AcrR   micF   -   [AIBSCS] Weak
AcrR   soxR   -   [BPP, GEA, HIBSCS] Strong
AcrR   soxS   -   [BPP, GEA, HIBSCS] Strong
Ada    ada    +   [BPP, GEA, HIBSCS, SM] Strong
Ada    ada    -   [BPP, GEA, HIBSCS, SM] Strong
Ada    aldB   +   [BPP, GEA, HIBSCS] Strong
Ada    alka    +   [BPP, GEA, HIBSCS, SM] Strong
Ada    alkb    +   [BPP, GEA, HIBSCS, SM] Strong
Ada    alkb    -   [BPP, GEA, HIBSCS, SM] Strong
AdiY   adiA    +   [BPP, GEA] Strong
AdiY   gadA    +   [BPP] Strong
AdiY   gadB    +   [BPP] Strong
AdiY   gadC    +   [BPP] Strong
AdiY   gadX    +   [BPP] Strong
AdiY   gltB    +   [BPP] Strong
AdiY   gltD    +   [BPP] Strong
AdiY   gltF    +   [BPP] Strong

```

Figure 7.

The Regulatory network interactions dataset is the most popular data file. The file contains objects and interactions supported by experimental evidence from RegulonDB; its contents include contact information, columns description, and the data.

Figure 8.

The results web page for arabinose query displays the list of all those hits where the word in any of the searching terms was found in RegulonDB.

Figure 9. The *araC* Gene web page displays the query gene and their regulatory elements affecting the gene in a colored graph; and their gene context in gray, followed by a section containing all its related data.

Figure 10.

The expanded 5' region for *araC* gene showing its regulatory elements is a view from the icon upper to the graph in the gene page. Objects with strong evidence are drawn with a continuous line, while those of weak evidence are with a dotted line. The genes are painted according to their functional class. Three colors are also used for the effect of TF on the promoter, red for repressor, green for activator, blue for dual and gray for unknown effect.

Biological Name (see RegulonDB glossary)	Properties	Graphic feature, strong evidence	Graphic feature, weak evidence
Promoter	Not applicable
TF binding site	Activator
	Dual
	Repressor
	Unknown
Terminator	Not applicable
Riboswitch	Not applicable
Translational attenuator	Not applicable
Transcriptional attenuator	Not applicable
Gene	Complete gene (Color Code)
	Incomplete gene (Color Code)
Small RNA binding site	Activator
	Dual
	Repressor
	Unknown
Allosteric regulation of RNAP	Activator
	Dual
	Repressor

Figure 11.
The graphic code for objects and colors used for DrawingTracesTool.

Figure 12. Partial view of Operon web page shows, below the title, a graphic displaying the query operon containing all the genes of its different TUs, as well as all the regulatory elements involved in the transcription and regulation of those TUs. This is followed by a section containing all its related data.

```
acaaaaaacgggtttatttggcggatagaaataacgaaagcacaaataatgggagcgtcAattttcgagtttgctgacc
-35-10+1
```

Figure 13.

The *fliFp1* promoter sequence region showing 80 bp, containing the -35 and -10 boxes, and the $+1$ site (in red upper case).

FlhDC DNA-binding transcriptional dual regulator	
Synonyms: FlhDC Summary: The FlhD and FlhC proteins are transcriptional factors that form heterohexamers (FlhD ₄ C ₂) [10]. The complex transcription factor is the principal regulator of bacterial flagellum biogenesis and swarming migration [1, 11]. This heteromeric regulator activates class II genes involved in the flagellar basal body [1, 3, 4, 11], proteins of the export machinery, the flagellar α subunit (FlaA), and its anti- α factor, FlgM [1, 11, 12]. Currently, no inducer for this regulator has been reported in the literature. A microarray analysis showed that the master regulator, FlhD ₄ C ₂ , regulates several nonflagellar genes, but the direct effect has not been determined. FlhD and FlhC are part of the unusual <i>flhDC</i> operon that encodes two transcriptional regulators [3, 13]. Read more >	
Transcription factor	?
Regulon	+ ▲
Regulated gene(s)	ccmA, ccmB, ccmC, ccmD, ccmE, ccmF, ccmG, ccmH, flgA, flgB, flgC, flgD, flgE, flgF, flgG, flgH, flgI, flgJ, flgK, flgL, flgM, flgN, flhA, flhB, flhE, flhA, flhD, flhE, flhF, flhG, flhH, flhI, flhJ, flhK, flhL, flhM, flhN, flhO, flhP, flhQ, flhR, flhS, flhT, flhZ, glpA, glpB, glpC, gltI, gltJ, gltK, gltL, hydN, hypF, mdh, mglA, mglB, mglC, napA, napB, napC, napD, napF, napG, napH, nrfA, Read more >
Multifun term(s) of regulated gene(s)	MultiFun Term (List of genes associated to the multifun term) flagella (34) motility, chemotaxis, energytaxis (aerotaxis, redotaxis etc) (32) flagellum (29) anaerobic respiration (19) membrane (17) Read more >
Regulated operon(s)	flgAMN, flgBCDEFGHIJ, flgKL, flhBAE, flhAZ-tyj, flhDST, flhE, flhFGHIJK, flhLMNOPQR, glpABC, gltI-sroC-gltJLK, hydN-hypF, mdh, mglBAC, napFDAGHBC-ccmABCDEF, nrfABCDEF, pdeH, ppdAB-ygdB-ppdC-recC, ycgR, yecR, ygbJK
First gene in the operon(s)	flgA, flgB, flgK, flhB, flhA, flhD, flhE, flhF, flhI, glpA, gltI, hydN, mdh, mglB, napF, napF, nrfA, pdeH, ppdA, ycgR, yecR, ygbK
Simple and complex regulons	ArcA,CRP,DpiA,FlhDC ArcA,CRP,FNR,Fis,FlhDC,GlpR CRP,FlhDC,GalR,GalS CsgD,FlhDC CsgD,FlhDC,H-NS,Mata,NsrR,SutR Read more >
Simple and complex regulatory phrases	Regulatory phrase (List of promoters regulated by the phrase) [FlhDC,+](10)
Transcription factor binding sites (TFBSs) arrangements	+
Alignment and PSSM for FlhDC TFBSs	
Evolutionary conservation of regulatory elements	
Evidence	
Reference(s)	

Figure 14.
The FlhDC web Regulon page contains different sections with information about their related objects. Only the Regulon section is displayed and the rest are collapsed.

Figure 15. The FlhDC+ Simple Regulon web page shows all the promoters regulated exclusively (known) by FlhDC as activator. Each line in the graph represents a promoter and the boxes a binding interaction of the TF. The yellow boxes are proximal sites.

Transcription factor binding sites (TFBSs) arrangements

Show 10 entries

Search:

Functional conformation	Function	Promoter	Sigma factor	Central Rel-Pos	Distance to first Gene	Genes	Sequence	LeftPos	RightPos	Evidence (Confirmed, Strong, Weak)	References
FlhDC	activator	flgAp	Sigma70	-68.0	-90.0	flgA, flgM, flgN	taacgcgcgcgTATCGCGCAATAAACgcaaaatggg	1130945	1130961	[GEA], [HIBSCS]	[1]
FlhDC	activator	flgAp	Sigma70	-40.0	-62.0	flgA, flgM, flgN	caaaatgggtCCTATTATTCCTTCatggtcattg	1130917	1130933	[GEA], [HIBSCS]	[1]
FlhDC	activator	flgBp	Sigma70	-67.5	-93.0	flgB, flgC, flgD, flgE, flgF, flgG, flgH, flgI, flgJ	caatgacctCAACGCGATTAATAGGaccacatttg	1130917	1130933	[BPP1], [GEA], [HIBSCS]	[1], [2]
FlhDC	activator	flgBp	Sigma70	-41.5	-67.0	flgB, flgC, flgD, flgE, flgF, flgG, flgH, flgI, flgJ	gaccacatttgCCTTTATTCCGCAaagcgcgcg	1130943	1130959	[BPP1], [GEA], [HIBSCS]	[1]
FlhDC	activator	flgKp	Sigma28	nd	nd	flgK, flgL	nd	nd	nd	[BPP1], [GEA]	[2]
FlhDC	activator	flhBp	Sigma70	-67.5	-92.5	flhB, flhA, flhE	tggtggcgagCAAAACCGCTAAATCCGccgtgttggc	1966275	1966292	[BPP1], [GEA], [HIBSCS]	[1], [3]
FlhDC	activator	flhBp	Sigma70	-39.0	-64.0	flhB, flhA, flhE	ctgttttgccCCTTACTCAAAACCATTCgaagcgtttgc	1966247	1966263	[BPP1], [GEA], [HIBSCS]	[1], [3]
FlhDC	activator	flhAp1	Sigma70	-67.0	-97.0	flhA, flhZ, tcyJ	gaatttacctGTAAACCCCAAAATAGCccctcatttca	2001878	2001894	[BPP1], [GEA], [HIBSCS], [SM]	[1], [2], [3], [4], [5]
FlhDC	activator	flhAp1	Sigma70	-38.0	-68.0	flhA, flhZ, tcyJ	tcatttcaccCACTAAATCTCCATTAaaaacccctgc	2001849	2001865	[BPP1], [GEA], [HIBSCS], [SM]	[1], [2], [3], [4], [5], [6]
FlhDC	activator	flhDp	Sigma70	-69.0	-124.0	flhD, flhS, flhT	cattttttgtTATCGCGCAATAGTCcttttcctgc	2003740	2003756	[BPP1], [GEA], [HIBSCS]	[1], [2]

Showing 1 to 10 of 32 entries

First Previous 1 2 3 4 Next Last

Alignment and PSSM for FlhDC TFBSs

Figure 16. The Transcription Factor binding sites arrangements section displays the information for the first 10 of 32 interactions for FlhDC regulon.

Transcription factor binding sites (TFBSs) arrangements	
DNA binding site length:	17
DNA binding site symmetry:	asymmetric
Non-redundant binding site(s)	
Show 10 entries	
Search:	
Sequence	
gtcaataagaCAAAGGCCTAAATAGCAacaaaaaac	[-40.5, flEP1, reverse] [-69.5, flIP1, forward]
cattttttgtTACTCGCCGAAATACTCttttctctgc	[-69.0, flIDp, forward]
caaaaaaacgGGTTTATGGCGGATAGaaaaaacga	[-69.0, flEP1, reverse] [-41.0, flIP1, forward]
tacgcgcgcgTATCGCGCAATAAACgcaaatggg	[-68.0, flgAp, reverse]
gacacctacaGATGGCAAGAAATAGCGcctgccaggc	[-68.0, yecRp, forward]
caatgaccatCAACGGCATAAATAGCGacccattttg	[-40.0, flgAp, reverse] [-67.5, flgBp, forward]
tgtgtggcagCAAAAGCCCATAATCCCgcctgttttgc	[-67.5, flhBp, reverse]
gaatttacctGTAAACCCCAAAATAACccctatttea	[-67.0, flIP1, reverse]
tattttcgccTAACGTCAGAGGTAGCAccgtaatecg	[-65.0, flIP1, forward]
aatggcgaagAACGGTCTAAATAGGCTgattcaagg	[-62.0, gltIP1, reverse]
Showing 1 to 10 of 17 entries	
First Previous 1 2 Next Last	

Figure 17. The Non-redundant binding sites view displays 10 of the 17 binding sites of the FlhDC regulon. The binding sites are in upper case, and the DNA letters are colored only to help to see patterns if they exist. The list of regulated promoters comes in the second column, the number is the distance calculated from the central position of the binding site to the promoter.

Figure 18. This view shows all binding sites for FlhDC TF, as well as the binding sites for other TFs in the same set of genes. In this case, each line represents a gene; the 0 number in the scale is the start of the gene; the color is the effect and the intensity of the color the evidence of the interaction. It helps to have a complete sight of the complexity of regulation of the set of genes.

Summary of evolutionary conservation of FlhDC gene regulated (first gene of a TU) interactions

Evolutionary conservation in RegulonDB is limited to gammaproteobacteria. Conservation of a regulatory interaction, i.e. a TF regulating a target gene, is inferred from the overrepresentation of TFBSs in a set of orthologous regulatory regions for the specified target gene. Following the regulon definition, for an ortholog gene to be taken into account for the collection of orthologous regulatory regions, it is necessary the existence of the TF gene ortholog in the same organism. Orthologs are selected based on a bidirectional best-hit procedure using RSAT tools package. Overrepresentation is calculated based on the binomial distribution; the binomial significance is shown in following table as significance of conservation.

First Gene	Significance of conservation (nd)	Footprint conservation
figA	20.52
figB	14.60
flhB	8.46
fliA	9.13
fliD	6.83
fliE	11.00
fliF	13.69
fliL	22.53
glpA	1.09 (Below threshold)
gltI	0.20 (Below threshold)
hydN	0 (Below threshold)
mdh	0 (Below threshold)
mglB	0 (Below threshold)
napF	0 (Below threshold)
nrfA	0 (Below threshold)
ppdA	-2.54 (Below threshold)
ycgR	0 (Below threshold)
yecR	5.17
ygbK	nd
yhjH	0.05 (Below threshold)

Figure 19.

The web page shows the evolutionary conservation of the interaction between FlhDC TF and the regulated genes. The process gets all first genes from TUs regulated by FlhDC, then for all genes, the regions of the promoter sequences of orthologous promoters are obtained in the gammaproteobacteria. The FlhDC PSSM is used to scan the regions and predict the binding sites in these orthologous upstream regions.

Evolutionary conservation of the FlhDC regulatory interaction in flgA orthologous genes

In general an interaction is conserved if there is a significant overrepresentation of TFBSs in the set of orthologous promoters of the target gene. The set of orthologous promoter for flgA was scanned with the annotated matrix for FlhDC. In order to avoid bias due to an excessive representation of close genomes (i.e. several *E. coli* genomes) we masked the regulatory sequences that where to similar. Overrepresentation is calculated based on the binomial distribution, comparing the observed scores vs. the expected ones for the masked sequence set.

Ortholog predicted binding sites	flgA_Escherichia_coli_K_12_substr_MG1655_uid57779_Gammaproteobacteria_FlhD_sites.tab
Ortholog predicted binding sites graph	flgA_Escherichia_coli_K_12_substr_MG1655_uid57779_Gammaproteobacteria_FlhD_sites.jpg
genes	flgA_Escherichia_coli_K_12_substr_MG1655_uid57779_Gammaproteobacteria_query_genes.tab
Query sequence	flgA_Escherichia_coli_K_12_substr_MG1655_uid57779_Gammaproteobacteria_query_seq.fasta
Orthologs	flgA_Escherichia_coli_K_12_substr_MG1655_uid57779_Gammaproteobacteria_ortho_bbh.tab
ortho sequences	flgA_Escherichia_coli_K_12_substr_MG1655_uid57779_Gammaproteobacteria_ortho_seq.fasta
Purged sequences	flgA_Escherichia_coli_K_12_substr_MG1655_uid57779_Gammaproteobacteria_ortho_seq_purged.fasta
Occ sig	flgA_Escherichia_coli_K_12_substr_MG1655_uid57779_Gammaproteobacteria_FlhD_occ_sig.tab
Occ sig graph	flgA_Escherichia_coli_K_12_substr_MG1655_uid57779_Gammaproteobacteria_FlhD_occ_sig.jpg

(Click on image below)

Figure 20. The evolutionary conservation process for each gene creates a results page with graphs, showing the distribution of the significance of hits and the input/output files used in the pipeline.

BetI Genetic Sensory Response Unit in *Escherichia coli* K-12 genome

Figure 21.
The BetI genetic sensory response unit web page. The page has three sections: the map, the GENSOR unit and the reactions.

Figure 22.
The RegulonDB Gene Coexpression web page showing results of searching nine genes regulated by FlhDC as activator.

Figure 23. The Coexpression by Gene Tab shows 25 of 50 top coexpressed genes for *flgB* gene. All of the top coexpressed genes for *flgB* are regulated by the FliHDC TF (except the genes that code for the regulator *flhC* and *flhD*), according to the genetic regulation annotated in RegulonDB.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 24. The Matrix HeatMap tab displays a graphical representation of the coexpression for *flgB* gene and the rank color between its top 50 coexpressed genes with the rest of the query list. The color palette shows the gradient associated with the ranks, lowest ranks are high coexpression in green, and the highest ranks are not coexpressed, shown in red color.