D98–D105 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 4 November 2010*
doi:10.1093/nar/gkq1110

RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units)

Socorro Gama-Castro[1], Heladia Salgado[1], Martin Peralta-Gil[1], Alberto Santos-Zavaleta[1], Luis Muñiz-Rascado[1], Hilda Solano-Lira[1], Verónica Jimenez-Jacinto[2], Verena Weiss[1], Jair S. García-Sotelo[1], Alejandra López-Fuentes[1], Liliana Porrón-Sotelo[1], Shirley Alquicira-Hernández[1], Alejandra Medina-Rivera[1], Irma Martínez-Flores[1], Kevin Alquicira-Hernández[1], Ruth Martínez-Adame[1], César Bonavides-Martínez[1], Juan Miranda-Ríos[3], Araceli M. Huerta[1], Alfredo Mendoza-Vargas[4], Leonardo Collado-Torres[5], Blanca Taboada[6], Leticia Vega-Alvarado[6], Maricela Olvera[4], Leticia Olvera[4], Ricardo Grande[2], Enrique Morett[4] and Julio Collado-Vides[1,*]

[1]Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, A.P. 565-A, Cuernavaca, Morelos 62100, [2]Unidad Universitaria de Secuenciación Masiva de ADN, Instituto de Biotecnología, Universidad Nacional Autónoma de México, A.P. 510-3, Cuernavaca, Morelos 62100, [3]Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, D. F., [4]Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, A.P. 510-3, Cuernavaca, Morelos 62100, [5]Winter Genomics, D. F. CP 07300 and [6]Centro de Ciencias Aplicadas y Desarrollo Tecnológico, Universidad Nacional Autónoma de México, D. F., México

Received September 15, 2010; Revised October 15, 2010; Accepted October 18, 2010

ABSTRACT

RegulonDB (http://regulondb.ccg.unam.mx/) is the primary reference database of the best-known regulatory network of any free-living organism, that of *Escherichia coli* K-12. The major conceptual change since 3 years ago is an expanded biological context so that transcriptional regulation is now part of a unit that initiates with the signal and continues with the signal transduction to the core of regulation, modifying expression of the affected target genes responsible for the response. We call these genetic sensory response units, or Gensor Units. We have initiated their high-level curation, with graphic maps and superreactions with links to other databases. Additional connectivity uses expandable submaps. RegulonDB has summaries for every transcription factor (TF) and TF-binding sites with internal symmetry. Several DNA-binding motifs and their sizes have been redefined and relocated. In addition to data from the literature, we have incorporated our own information on transcription start sites (TSSs) and transcriptional units (TUs), obtained by using high-throughput whole-genome sequencing technologies. A new portable drawing tool for genomic features is also now available, as well as new ways to download the data, including web services, files for several relational database manager systems and text files including BioPAX format.

INTRODUCTION

The major challenge of genomic bioinformatics is to transform data into knowledge, by means of organizing and integrating the information for a more comprehensive understanding of biology. We have been aware of the limited knowledge on gene regulation that has been available in RegulonDB, and as a consequence, in this new version, a larger variety of mechanisms of gene regulation are included. Collections of these mechanisms are integrated within a biological context defined as Genetic Sensory Response Units, or Gensor Units (GUs).

*To whom correspondence should be addressed. Tel: +52 777 313 2063; Fax: +52 777 317 5581; Email: regulondb@ccg.unam.mx; collado@ccg.unam.mx

A GU represents the biological context that naturally provides a better understanding of gene regulation, integrating mechanisms and concatenated reactions in a physiological unit defined by the capacity of the genome to regulate gene expression in response to a signal or stimulus. RegulonDB version 7.0 contains now GUs associated with carbon utilization, metabolism of amino acids and regulation of sigma factors. We are committed to continue the curation of the corresponding GUs for all the remaining transcription factors (TFs).

The RegulonDB team has been expanded so that it now contains both, the Program of Computational Genomics, where the project originated, and the laboratory of Enrique Morett, which is devoted to experimental high-throughput (HT) strategies for genome-wide identification of regulatory components of the *Escherichia coli* network of gene regulation. This synergy has generated remarkable progress in the determination of more than 2000 transcriptional start sites (TSSs) and their corresponding promoters with high precision and several hundred transcriptional units (TUs) [(1) and Collado-Torres *et al.*, submitted for publication]. These data are now incorporated in RegulonDB version 7.0.

METHODS

Curation of Gensor Units

Curation of GUs was initiated by searching all PubMed articles containing information about signals and signal transduction of sigma factors and about TFs regulating carbon source utilization and metabolism of amino acids. Based on the abstracts, full papers were selected and each molecule interaction contained in the papers was saved for later representation in the diagrams. The signal transduction data included all the concatenated reactions from the signal to the modification of a TF to activate or repress transcription. Each reaction contains the reactants, the products, inhibitors, activators and growth conditions in which the reaction is active or inactive.

To obtain the data about the response, the function of all the products of genes regulated by a TF were revised in our database. Those TFs that have a role in signal transduction, metabolism and signal transport were assigned as part of the response. The transport and metabolism reactions were taken from and linked to other databases, such as EcoCyc and KEGG.

The graphs were generated using the CellDesigner editor (2,3), the html maps were developed in java, and the resizing image map uses a javascript taken from http://home.comcast.net/~urbanjost/IMG/resizeimg.html.

TSSs detection and TU mapping

Modified 5′-RACE and pyrosequencing methods are detailed in (1). Using an Illumina Genome Analyzer IIx (GAIIx), we sequenced the 5′-ends (36 bp reads) of all transcripts from *E. coli* MG1655 using four experimental methods. Samples were treated with DNase I and ribosomal RNA was removed with the RiboMinus Transcriptome Isolation Kit (Invitrogen, Carlsbad, CA, USA). Then the RNA was either (i) enriched for 5′-monophosphate transcripts, (ii) enriched for 5′-triphosphate transcripts by degrating 5′-monophosphate transcripts with a specific exonuclease, (iii) enriched for 5′-triphosphate transcripts by ligating an adapter only to these transcripts or (iv) left untreated. Finally, an adapter was ligated to the 5′-end; all the above enables the identification of TSSs by filtering out products of degradation. Reads were mapped to the genome using Bowtie (4) –v mode allowing maximum three mismatches. If a 36 bp read did not align, we trimmed one base from the 3′-end and re-used the same mapping parameters. We ended this iterative alignment if the reads were too short (18 bp). We used the start position of the reads as the putative TSSs position and filtered out reads mapping to ribosomal operons. 5′-ends shared among the four methods with a frequency above a given cutoff enable us to generate the final set of reliable TSSs.

TUs were identified as cotranscribed genes in the same mRNA molecule detected by paired-end RNA-seq with different insert sizes, using Illumina GAIIx.

A detailed compendium of all these findings will be published elsewhere.

Building position-specific scoring matrices

In RegulonDB, each position-specific scoring matrix (PSSM) has been evaluated based on the methodology proposed by Medina-Rivera *et al.* (5). This evaluation combines theoretical and empirical score distributions to assess the predictive capability of PSSMs. The theoretical distribution provides an estimate of the false-prediction rate. Empirical distributions indicate the enrichment of binding sites in the upstream regions of the *E. coli* K-12 genome. Negative controls are performed to analyze the same sequence collections with column-permuted matrices.

PSSMs based on the published literature: Matrices were generated for 91 TFs with three or more annotated binding sites. For one TF, sites were extended 3 bp on each side. Extended binding sites were aligned using the program 'consensus' in order to obtain the PSSM based on the best alignment of the sites. The width of the PSSM was set to be equal to the annotated length of the binding sites. Variable PSSMs: Matrices were generated for TFs with four or more annotated binding sites. For one TF, sites were extended 10 bp on each side, and these sites were used to build several matrices by varying parameters, such as program used (meme or consensus), width ($\pm$ 5 bp from the annotated length of the binding sites), and background model (Markov zero, Markov one). A collection of 30 PSSMs per TF was created; from this collection the matrix with the best quality was selected. A more detailed explanation of this procedure can be found in Medina-Rivera *et al.* (5).

TFBS symmetry assignments

PSSM matrices are built using the annotated binding sites in RegulonDB for a TF, each matrix resumes the collection of transcription factor binding site (TFBS) information. Using PSSMs with width reported in the literature (available for 91 TFs), other reports from the literature,

and manual analysis by curators, we identified the internal structure within the TFBSs and assigned an internal symmetry to them. For this analysis we included also information from the RegPrecise database (6), which shows a comparative genomic approach for the TFBSs from a wide variety of prokaryotic genomes. For TFs with no available matrix, symmetry identification was based only on information in the published literature and the biological knowledge of the curators.

Repeats within the TFBSs were detected based on a matrix self-comparison approach, at different levels of shift. We measured the similarity distances between matrices with the program 'compare-matrices', which is included in the program suite RSA-tools (7), and we used as distance metric the 'Normalized Correlation'. Direct repeats were detected by measuring the distance of a set of TFBSs to themselves at shifted positions to show whether the first 'half' of the matrix was similar to the second 'half' Inverted repeats were detected by comparing a set of TFBSs with their reverse complement (RC) and using various shift values. When it was not possible to detect computationally a determined symmetry, we have annotated these TFBSs as asymmetrical.

The computational method for symmetry identification using a collection of TFBSs is available for use in the program 'matrix-symmetry', which is part of the *cvs* distribution of the RSA-tools suite (7).

RESULTS

Gensor Units

The ability of a cell to respond to changes inside the cell or in the environment initiates when a new signal or stimulus is sensed and transmitted through a series of molecular concatenated reactions, called signal transduction or transduction pathways. These events bring into action genetic switches that modify gene expression to produce a molecular response by the cell. This response is the sum of all changes in concentration of the target gene products induced and/or repressed. Together, these components define what we call a genetic sensory response unit, or Gensor Units (GUs). We have initiated their curation and integration within RegulonDB, bringing transcriptional regulation into the context of the corresponding reactions and interactions, including transport, synthesis, degradation, catalysis, formation of biomolecular complexes, induction and inhibition, among others.

More precisely, a GU is defined by four components: (i) the signal or stimulus, (ii) the signal transduction pathway, (iii) the mechanisms affecting the expression of genes and (iv) the response resulting from the modified gene expression of the affected set of target genes. A GU, in principle, always involves at least one feedback loop of information (signal, signal transduction, genetic switch, response, signal), as shown in Figure 1.

The name given to each GU consists of the name of the TF and the name of the effector. The GU describes a signal transduction circuit associated with a TF. This association is an essential conceptual difference from the notion of a stimulon, which, as proposed by Neidhardt, is mechanistically independent (8,9).

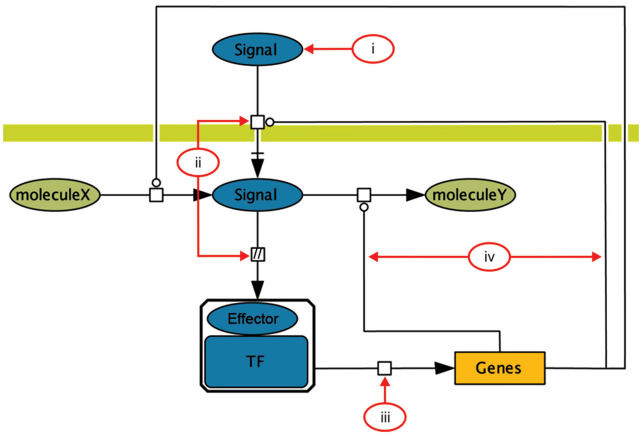

Figure 1. GU map. The elementary GU includes the signal (i), the concatenated set of reactions (ii) that elicit a genetic switch (iii) to induce or repress the expression of genes involved in the response (iv). The target genes in this diagram code for gene products involved in the synthesis of the signal (from molecule X to the signal), the utilization (from the signal to molecule Y), and induction of the gene(s) involved in the transport of the external signal.

GUs can be described as a set of reactions that together make an explainable response, as shown in Figures 2 and 3. By explainable we mean that the observed response affects genes whose products deal with the signal, in a way that seems reasonable to occur. However, it is not always clear what is the role of all regulated genes, either because of the high complexity of the biology, the fact that some genes have no known function, or the lack of their complete curation.

Graphic displays are called 'Gensor Unit maps.' In association with each GU map, RegulonDB contains both a summary text and a table of all genes within the GU. They may be connected to each other through molecules or reactions that belong to one particular unit. Such is the case of the units in which amino acids or carbon sources act as effectors of TFs.

The representation of GUs involves not only signal transduction and gene regulation but also metabolism and any reactions in which the gene products may be involved. The function of the regulated genes is the response, the fourth component of a GU. Sometimes metabolic reactions are represented inside the GU as one single compact transformation or superreaction. In that case, each of the proteins catalyzing individual reactions is linked to its corresponding superreaction.

The explainable part of a GU is captured by a single sentence or paragraph that we use to describe at the physiological level the function of a GU. For instance, the GU for AraC regulation summarizes the reactions involved in the following sentence: 'In the presence of arabinose, AraC activates transcription of genes that code for proteins necessary for the utilization and transport of arabinose' (Figure 2).

RegulonDB version 7.0 contains complete GUs of different TFs involved in carbon source utilization and

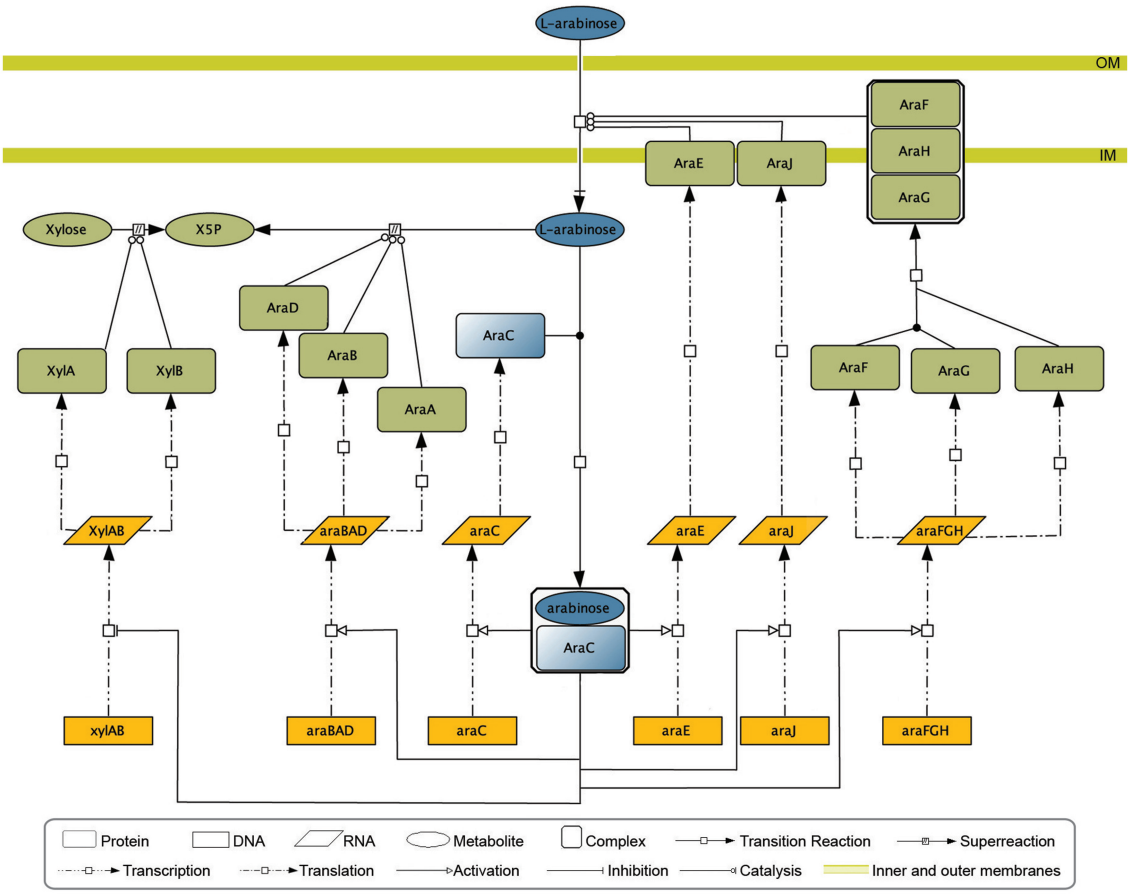

Figure 2. AraC-arabinose GU. In the presence of l-arabinose, AraC activates transcription of genes that code for proteins necessary for utilization and transport of l-arabinose. In the absence of glucose CRP coregulates with AraC. The four GU elements are represented by different colors in the image: blue, the signal (i) and signal transduction (ii); yellow, genetic switch (iii); green, the response (iv).

metabolism of amino acids. We also have partially curated other GUs, such as the signal transduction component for the sigma factors and for TFs of two-component systems. GUs can be accessed directly from the website typing the name of a TF, a sigma factor or a gene in the search menu, located in the main page, using the option 'Gensor Unit'. On the other hand, in the web pages of Gene, Operon, Sigmulon and Regulon there is a link to the corresponding GU page.

Signal transduction pathways

We have initiated the description of signal transduction pathways, the second component of GUs within RegulonDB, adding several types of new interactions, such as RNA–DNA, RNA–RNA, RNA–protein, protein–protein and metabolite–protein interactions; indirect interactions are also represented.

As case studies for curation of signal transduction, we took the two-component systems and the sigma factors (see Figure 3 for an illustration of σ^{24}). The signal transduction of sigma factors shows that their activity is mainly inhibited by anti-sigma proteins through the formation of complexes, in addition to proteolysis of sigma factors themselves. We include all the molecules involved in the activation cascade. Stress conditions activate sigma factors through inactivation of anti-sigma factors by degradation. All sigma factors are also regulated at the level of transcription by TFs. σ^{38} is also regulated at the level of translation by small RNAs.

In the case of two-component systems, we have represented autophosphorylation of sensory proteins and the transfer of the phosphate group to the response regulator. Our goal is to complete this information by adding other molecules involved in these systems, as we have done already for the two-component system CpxA/CpxR.

The graphic displays of signal transduction pathways include the inner and outer cell membranes and the corresponding cytoplasmic and periplasmic regions where each molecule is located. Different shapes represent the different molecule types (Figures 2 and 3). The details of each reaction are displayed interactively as tool tips. When two or more signal transduction pathways converge in a large diagram, some of these pathways are shown as submaps, with a specific shape representing the compacted pathway. The full pathway can be displayed by clicking on the submap icon (e.g. the gray square, CpxR, in Figure 3). A brief description of the signal transduction pathway is shown in each diagram.

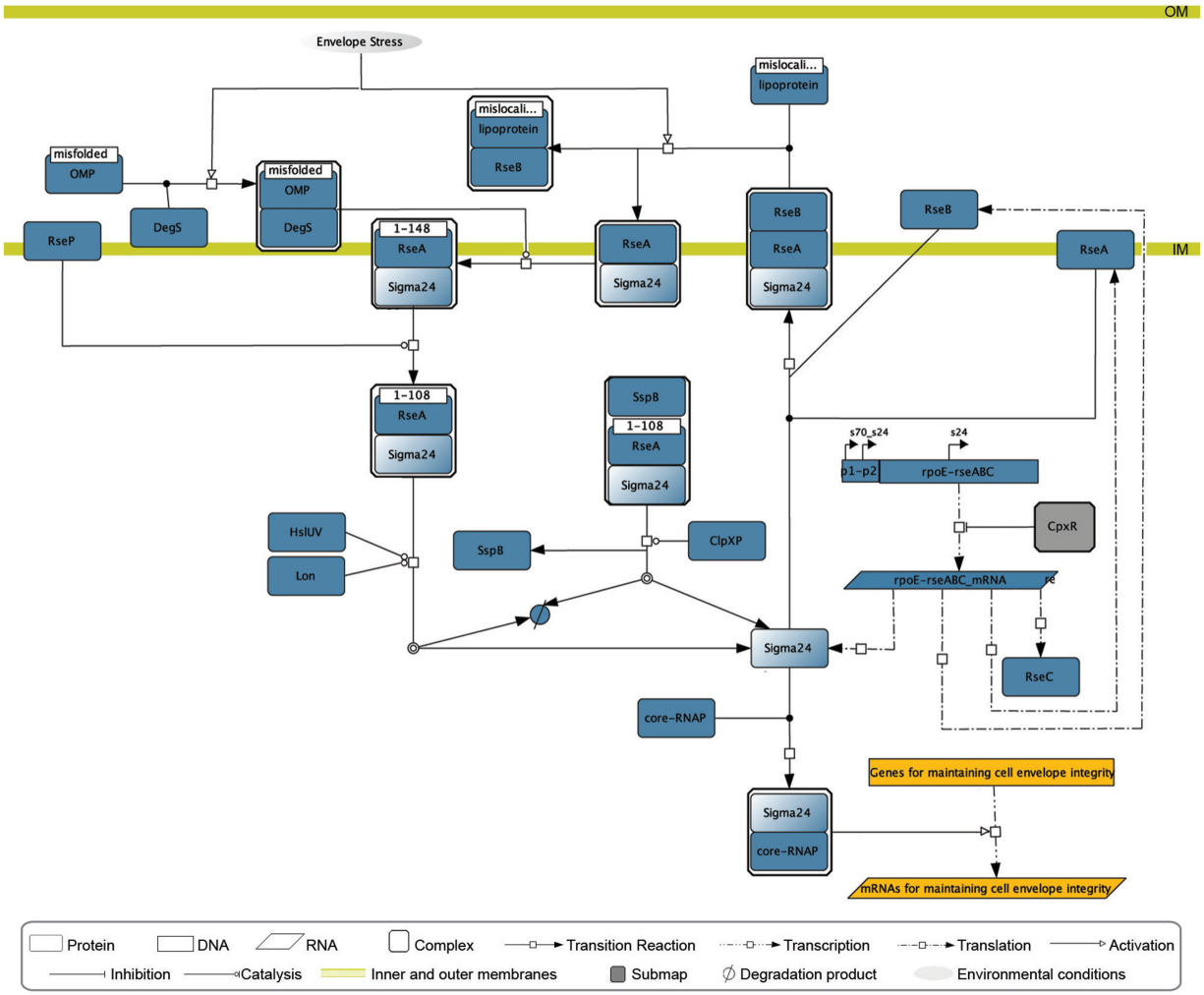

Figure 3. Signal transduction of the σ^{24} (σ^{E}) GU. In the absence of stress, the sigma factor σ^{24} is complexed with RssA and RssB in the inner membrane, but under envelope stress, misfolded proteins and lipoproteins appear in the periplasm. The lipoproteins release RssB from the complex RssB-rseA-σ^{24}, and then a proteolytic cascade starts when DegS, bound to unfolded proteins, partially degrades the RseA protein in the RseA-σ^{24} complex, producing the RseA-1-108 polypeptide that is cleaved by RseP, and the complex is released from the inner membrane. In the cytoplasm, HslUV or Lon can degrade RseA, releasing σ^{24}, which can bind to core RNA polymerase to activate transcription of genes for maintaining cell envelope integrity. Another protease that also degrades RseA in the cytoplasm is ClpXP, but the protein SspB has to be bound to the RseA-1-108-RpoE complex. On the other hand, the *rpoE* gene, which forms part of the rpoE-rseAB operon, is induced by the TF CpxR, which is shown in the image as a submap in a gray box. The submap is expandable by clicking on it. Details of each reaction with supporting evidence and references are provided in the database.

Whole-genome high-throughput mapping of transcription start sites and transcriptional units

We have developed HT whole-genome TSS and TU mapping strategies, initially by using pyrosequencing by 454 and more recently with Illumina sequencing technology [(1) and Collado-Torres *et al.*, submitted for publication]. These methodologies are highly sensitive and detect any 5′-end, as well as cotranscribed genes from the same mRNA molecule by paired-end sequencing. *Bona fide* transcription initiation events have a triphosphate at the 5′-RNA end, while 5′-monophosphate ends derive mainly from mRNA degradation. Two methods to enrich for 5′-triphosphate ends were implemented that rely on selective elimination of RNA molecules with 5′-monophosphate ends; one method uses 5′-phosphate-dependent exonuclease, and the other is based on the ligation of a biotinylated oligonucleotide to the 5′-monophosphate molecules and their elimination by magnetic streptavidin binding. Together, these efforts allowed us to map more than 2500 putative TSSs. We identified promoter sequences and sigma factor types that control the expression of ~80% of these genes, with σ^{70} being the most common, followed by σ^{38}. The majority of the putative TSSs were located between 20 and 40 nt from the start ATG. Interestingly, there is a high occurrence of putative TSSs in unexpected positions in the genome, such as within genes and operons, and in an antisense orientation, although generally at lower expression levels than putative TSSs located at the beginning of genes or operons. In addition, the RNA-seq paired-end strategy allowed us to detect several hundred TUs with high confidence (Collado-Torres *et al.*, submitted for publication).

The full set of these TSSs and TUs with their corresponding HT new evidences can be downloaded from the data sets menu.

Summaries for TFs

We have completed extensive summaries for 175 TFs that have at least one annotated gene regulatory interaction. These summaries contain information about the evolutionary family to which the TF belongs to, the domain composition, relevant characteristics of the TF and the cellular processes in which the regulated genes are involved. An indication of the active conformation of a complex (dimer, tetramer, etc.) is provided. These summaries also have descriptive information about binding site features: size, consensus sequence, relative position to the transcription start and spatial arrangement of the site sequences. The summary of each TF is found in its Regulon web page Summary label.

These TFs can be involved in different functional classes, for instance, flagellar and chemotaxis systems, metabolism of nucleosides, transport and synthesis of fatty acids, DNA replication, quorum sensing, toxin-antitoxin systems, adaptation and resistance under different conditions of stress, aerobic and anaerobic phases and carbon utilization, among others.

Matrix and symmetry for collection of TF binding sites

For years, RegulonDB has made available a collection of PSSMs for all TFs with three or more annotated binding sites; these PSSMs represent the set of TFBSs of each TF.

During the last 3 years, for the sake of producing predictions for new binding sites, a complete evaluation of the available matrices in RegulonDB has been performed based on the methodology proposed by Medina-Rivera *et al.* (5).

As a result of the evaluation of RegulonDB matrices some corrections on binding site annotations were performed (see below); also, the parameters used for building the matrices were modified. In addition, this evaluation will provide the necessary knowledge for the utilization of matrices for discovering putative binding sites in the *E. coli* K-12 genome.

PSSMs can be accessed in either of two ways: (i) from the website when accessing simple or complex regulons through links located in the section 'Transcription Factors', and (ii) the set can be downloaded as a data set (see the 'tutorial' web link in RegulonDB).

Some TFs bind to particular sequences that tend to be repeated two or more times conforming the TFBS. On the other hand, since a PSSM represents the set of TFBSs for one TF, by using a PSSM it is possible to predict automatically an internal structure contained in the collection of TFBSs. In order to assign the internal symmetry of the TFBSs, we combined information from the computational prediction of internal symmetry on the available PSSMs ('Methods' section), the published biological knowledge, the alignments constructed and manual editions made by the curators.

In RegulonDB there is information on regulatory interactions for 175 TFs, and we searched for a defined symmetry for 120 of them, resulting in: inverted repeat (46.3%), direct repeat (5.7%), asymmetric (16.6%) and undetected (31.4%). The asymmetrical binding sites generally are located in tandem with a definite orientation with respect to the neighboring binding sites but with no internal symmetry. An important portion of sites with undetected symmetry correspond to sites with long sequences, degenerated binding sites or sites with no specific position.

At the sequence level every TFBS is represented in the database with its defined sequence, length, complementary strand, central position, symmetry and orientation. Arrows showing orientations, in agreement with the consensus, represent the symmetries of sites.

Redefining TFBSs

Given the findings of the exhaustive analysis done in (5), and the fact that the number of sites for some TFs has grown considerably, we decided to revise systematically the TFBSs with lengths ranging from 40 to 60 bp whose orientations and consensus sequences were not clear.

We have relocated, reassigned and corrected binding sites for the CytR and OxyR regulons, which are involved in 52 regulatory interactions. This was done based on manual alignments of the respective upstream regions and the corresponding evidence obtained in the bibliography of every operon, such as similarity to the consensus sequence, and on the data from footprinting assays, site-directed mutagenesis, and profiling experiments.

As a result of this detailed analysis, the optimal CytR-binding site is now modeled as two octamer repeats, GTTGCATT, in a direct or inverted orientation and preferably separated by 2 bp (10,11). For this reason, the specific length of the CytR DNA-binding site assigned in the database is variable.

In the case of OxyR we have identified two inverted repeat motifs of 17 bp separated by 5 bp, covering ~40 bp (GATAGGTTnAACCTATCnnnnnGATAGGT TnAACCTATC). Therefore, we propose this motif (GA TAGGTTnAACCTATC) as a new consensus sequence. This is in agreement with the report of Toledano *et al.* (12), who proposed that this regulator binds as a dimer of dimers to long sequences in the DNA.

We also corrected and relocated the binding sites for FhlA, Ada, CaiF and YiaJ (see the full description at 'news' web link in RegulonDB).

New features in RegulonDB

New type of search. Two new forms of search are available: by sigmulon and by small RNA (sRNA). The first allows searching by name of a sigma factor or by the name of the gene that encodes the sigma factor. The sigmulon web page displays the properties of the sigma factor, including the transcription factors involved in the transcription of genes, and the promoters recognized by the sigma factor.

Searching on the website by sRNA can be made through the name of the gene, the GI or the bnumber. This page shows the sRNA sequence as well as the

regulatory interactions of the sRNA with the genes, each with their own supporting evidence and references.

DrawingTracesTool. We have recently developed the user-friendly DrawingTracesTool that allows generation of images from any region in the chromosome and can display all the elements contained in RegulonDB. This tool allows users to make graphs with their own data, defining all details, colors and shapes of objects. Users can save their figures and obtain a table listing all the elements that have been displayed. The DrawingTracesTool executable program is available at the 'download' web link in RegulonDB and runs on any computer platform. All documentation, the user manual and examples are available in the RegulonDB website.

Access to RegulonDB and links to external databases. All data stored in RegulonDB can be accessed through the web at http://regulondb.ccg.unam.mx/. Additionally, RegulonDB has diverse forms by which to access data: precomputed data sets from the 'Downloads' section of the RegulonDB website, web services, and direct remote connection to MySQL databases. Complementing the previous data sets, new data set files have been incorporated in order to have a complete repertoire of the genetic network; specific dump files for the Oracle, MySQL, PostgreSQL and Derby database manager systems (DBMS) are now available. RegulonDB is also accessible in BioPAX format, as we have contributed also in the expansion of this pathway exchange format (13).

RegulonDB now has direct connections to the Many Microbe Microarrays Database (M3D) at http://m3d.bu.edu/ (14) and to COLections Of Microarrays for Bacterial OrganismS (Colombos) database at http://ibiza.biw.kuleuven.be/colombos/, both of which are related to the analyzed genetic expression of bacterial organisms. Expression profiles of different microarray conditions are linked to Regulon web pages instead of individual Gene pages, because simple and complex regulons are coregulated.

Tutorials on frequent questions, which were published in a minireview on the study of gene regulation in bacteria (15), are available in RegulonDB. We keep curation of transcriptional regulation up to date with a delay of one or two months from the release date. We make three to four releases per year. A summary table of the total curated objects is available at the 'database summary' web link in RegulonDB.

DISCUSSION AND CONCLUSIONS

The accumulated work along 3 years, since publication of the previous RegulonDB paper (16), show considerable progress. The major conceptual change is that of embedding transcriptional regulation and operon organization as elements of the genetic sensory response units or organs, the so-called GUs. We are aware that the current version of the database is the beginning of a higher-level curation of gene regulation, which eventually shall pervade within our database for all regulatory mechanisms and their regulated genes. The addition of this biological context is an important change as we move forward with RegulonDB.

Three important aspects to consider and that will be addressed as we expand curation of more GUs are the following: First, the logics of regulation vis-à-vis the complexity of biology will be clearly visible. In some cases a GU will be easily understood and explainable in a simple sentence, i.e., 'in the presence of methionine, MetJ represses the genes that code for proteins necessary for its synthesis and transport.' However, not necessarily every GU will be understood directly or easily at this qualitative level of description. On the other hand, there may well be regulated genes that belong to a GU but whose function could not be well understood in terms of the physiology, signaling or whose explanation requires more detailed quantitative knowledge and/or even modeling simulations.

Second, in most cases we may not be certain when a GU has been completely described. Given its anatomical definition, any gene affected by the TF defining a GU is part of it. We do not necessarily have all this information at hand. Thus, as with any other type of object describing gene regulation (e.g. transcription units, promoters, DNA-binding sites), GUs are and will be subject to systematic curation.

Third, as mentioned before, we believe this high-level curation shall contribute to link mechanisms and physiology, with the long-term aim of addressing how the cell integrates its local circuits of regulation into global and interdependent organized components. As mentioned before, increased understanding of biological integration is a major challenge in genomics.

It is true that there are several databases that describe parts of the reactions and components of GUs. For instance, MiST (17) deals specifically with signal transduction reactions and KEGG (18) and EcoCyc (19) contain information on metabolism and the functions of regulated genes. EcoCyc also gathers information on transport, signal transduction and gene regulation—with the component on transcriptional regulation contributed by our team. The specific content and scope of some of these databases have changed, sometimes dramatically, over time, as is the case for EcoCyc, which was initially a database on metabolism and has been transformed into a Model Organism Database (MOD) (20). Given the availability of all these pieces of information, we have designed our curation and representation strategy in order to enrich in an important way the knowledge available in RegulonDB, without duplicating efforts performed by other teams. As mentioned before, GUs frequently involve a high level description by means of superreactions, providing the global view necessary to have a physiological understanding of gene regulation. Within GUs, superreactions are linked to other databases that describe them in more detail.

Furthermore, the major data enhancement to report now is the addition of thousands of TSSs and hundreds of TUs by global high-throughput mapping strategies performed by our joint team. These data will improve the accuracy of promoter prediction, operon structure and regulatory networks and support a new understanding

from a genome perspective of the complex regulatory network that governs transcription and regulation in *E. coli*.

ACKNOWLEDGEMENTS

We thank Victor del Moral Chávez and Romualdo Zayas-Lagunas for technical support, Gerardo Salgado Osorio for supporting project management and Ingrid Keseler for periodically sending us selected literature references for curation.

FUNDING

National Institutes of Health (GM071962, GM077678); Consejo Nacional de Ciencia y Tecnologia (CONACyT), Mexico (83686 G.I., CB2008-103686-Q, PROINNOVA-134817 and INNOVAPYME 137117). Funding for open access charge: Consejo Nacional de Ciencia y Tecnologia (CONACyT), Mexico.

Conflict of interest statement. None declared.

REFERENCES

1. Mendoza-Vargas,A., Olvera,L., Olvera,M., Grande,R., Vega-Alvarado,L., Taboada,B., Jimenez-Jacinto,V., Salgado,H., Juarez,K., Contreras-Moreira,B. *et al*. (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. *PLoS One*, **4**, e7526.
2. Funahashi,A., Tanimura,N., Morohashi,M. and Kitano,H. (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, **1**, 159–162.
3. Funahashi,A., Matsuoka,Y., Jouraku,A., Morohashi,M., Kikuchi,N. and Kitano,H. (2008) CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proc. IEEE*, **96**, 1254–1265.
4. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R2.
5. Medina-Rivera,A., Abreu-Goodger,C., Thomas-Chollier,M., Salgado,H., Collado-Vides,J. and van Helden,J. (2010) Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res.*, doi:1093/nar/GKQ710.
6. Novichkov,P.S., Laikova,O.N., Novichkova,E.S., Gelfand,M.S., Arkin,A.P., Dubchak,I. and Rodionov,D.A. (2010) RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res.*, **38**, D111–118.
7. Thomas-Chollier,M., Sand,O., Turatsinze,J.V., Janky,R., Defrance,M., Vervisch,E., Brohee,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–127.
8. Smith,M.W. and Neidhardt,F.C. (1983) Proteins induced by aerobiosis in *Escherichia coli*. *J. Bacteriol.*, **154**, 344–350.
9. Neidhardt,F.C. and Savageau,M.A. (1996) Regulation beyond the operon. In Neidhardt,F.C., Curtiss,R. III, Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik,B. *et al*. (eds), *Escherichia coli and Salmonella: Cellular and Molecular Biology*, Vol. 2. ASM Press, Washington, D.C., pp. 1310–1324.
10. Pedersen,H. and Valentin-Hansen,P. (1997) Protein-induced fit: the CRP activator protein changes sequence-specific DNA recognition by the CytR repressor, a highly flexible LacI member. *EMBO J.*, **16**, 2108–2118.
11. Jorgensen,C.I., Kallipolitis,B.H. and Valentin-Hansen,P. (1998) DNA-binding characteristics of the *Escherichia coli* CytR regulator: a relaxed spacing requirement between operator half-sites is provided by a flexible, unstructured interdomain linker. *Mol. Microbiol.*, **27**, 41–50.
12. Toledano,M.B., Kullik,I., Trinh,F., Baird,P.T., Schneider,T.D. and Storz,G. (1994) Redox-dependent shift of OxyR-DNA contacts along an extended DNA-binding site: a mechanism for differential promoter selection. *Cell*, **78**, 897–909.
13. Demir,E., Cary,M.P., Paley,S., Fukuda,K., Lemer,C., Vastrik,I., Wu,G., D'Eustachio,P., Schaefer,C., Luciano,J. *et al*. (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 935–942.
14. Faith,J.J., Driscoll,M.E., Fusaro,V.A., Cosgrove,E.J., Hayete,B., Juhn,F.S., Schneider,S.J. and Gardner,T.S. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.
15. Collado-Vides,J., Salgado,H., Morett,E., Gama-Castro,S., Jimenez-Jacinto,V., Martinez-Flores,I., Medina-Rivera,A., Muniz-Rascado,L., Peralta-Gil,M. and Santos-Zavaleta,A. (2009) Bioinformatics resources for the study of gene regulation in bacteria. *J. Bacteriol.*, **191**, 23–31.
16. Gama-Castro,S., Jiménez-Jacinto,V., Peralta-Gil,M., Santos-Zavaleta,A., Peñaloza-Spinola,M., Contreras-Moreira,B., Segura-Salazar,J., Muñiz-Rascado,L., Martínez-Flores,I., Salgado,H. *et al*. (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12, beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36(Database issue)**, D120–D124.
17. Ulrich,L.E. and Zhulin,I.B. (2007) MiST: a microbial signal transduction database. *Nucleic Acids Res.*, **35**, D386–D390.
18. Kanehisa,M. (2002) The KEGG database. *Novartis Found. Symp.*, **247**, 91–101, discussion 101–3, 119–28, 244–252.
19. Keseler,I.M., Bonavides-Martinez,C., Collado-Vides,J., Gama-Castro,S., Gunsalus,R.P., Johnson,D.A., Krummenacker,M., Nolan,L.M., Paley,S., Paulsen,I.T. *et al*. (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.*, **37**, D464–D470.
20. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18(Suppl. 1)**, S225–S232.