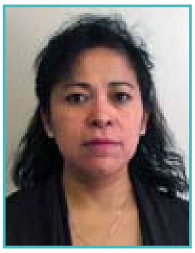
Expanding BioPAX format by integrating Gene Regulation



Irma Martínez-Flores, Verónica Jiménez-Jacinto, Alejandra C. López-Fuentes and Julio Collado-Vides

Programa de Genómica Computacional, Centro de Ciencias Genómicas-UNAM. Cuernavaca, Morelos, México

Biological information

Pathways

Knowing the complete sequencing of the genomes of organisms makes it possible to identify their genetic participants. However, the activity in an organism depends not only on the functionality of genes but also on relations and regulatory mechanisms established among them. The cell functions have a series of interactions, which often form pathways that can be grouped together for either organizational or functional reasons. These pathways assemble multiple biological data types, including metabolic pathways, signal transduction pathways, protein-protein interaction networks, gene regulatory pathways and genetic interactions.

Transcription regulation in *Escherichia coli* (genetic pathway)

Transcription is the first step leading to gene expression; therefore, transcriptional regulation is one of the most important regulatory mechanisms. This pathway describes how genes could be expressed, and transcription factors (TFs) as one of their main participants. The role of these factors is to coordinate and regulate the expression of the genes of an organism. The processes occur with the participation and interaction of different regulatory elements.

A chromosome contains several regions, which include the following:

- genes involved in producing a polypeptide chain or stable RNA;
- promoters that are comprised by a transcription start site (TSS) and short conserved sequences upstream of the TSS, which mediate the binding of RNA polymerase. Each promoter is specific to a particular sigma factor;
- binding sites (BS) for TFs; these sites are recognized by the transcription factors within a genome;
- terminators, which are required regions to finish the transcription process.

In addition to these DNA segments, there are other types of molecules acting in genetic pathways, including:

- proteins, which perform a large number of regulatory functions (RNA polymerase, sigma factors (σ), TFs, etc.);
- small molecules (such as phosphorus, cAMP, iron, etc.);
- RNAs, such as mRNA, tRNA, rRNA and sRNA.

The interactions among all the elements above mentioned participate in genetic regulation pathways.

The organization of genes in a genomic context allows making regulation process more efficient [1]; for instance, operons express coordinately a set of genes of a single promoter in bacteria [2]. However, some cases are complex because they may contain several promoters, out of which some can be internal. Each promoter related to an operon produces a single mRNA, this structure is known as transcriptional unit (TU) [3].

Regulon DB

RegulonDB contains information regarding regulatory elements and the events involving them (http://regulondb.ccg.unam.mx). This curated information is high quality, complete and updated [4]. The last version includes a much more precise definition of biological concepts; all the elements we have already mentioned are curated in this database.

The events involving these elements include the complexes formed between TFs and effectors generating specific conformations, which can be either active or inactive. The function of the TFs depends on the binding to DNA-BS under the control of a specific promoter. The action of integrating all these events is called regulatory interaction (RI).

Also, RegulonDB is linked directly to a number of bioinformatic tools that facilitate the analysis of data sets and tools for microarray, as well as direct access to full papers supporting all this knowledge. In summary, RegulonDB represents a "gold standard" in the bioinformatics of gene regulation design bacterial genomics [4].

Biological Databases and Bioinformatics tools

Needs and difficulties

The number of biological databases is growing fast. Nevertheless, the information contained in these databases has distinct data models, access methods, file formats and semantics. This diversity of implementation makes it extremely difficult to collect data from multiple sources, and therefore slows down the scientific research that involves pathways [5, 6]. This generates the need to develop a format for biological pathway data exchange.

Several standard exchange formats have been developed in order to make heterogeneous data sources easier to use and share. The Systems Biology Markup Language (SBML) [7] and CellML [8] represent mathematical models of pathways designed for quantitative simulation of concentration profiles of components over time. The Proteomics Standards Initiative's Molecular Interaction (PSI-MI) format enables to exchange molecular interaction data sets [5]. Finally, the Biological Pathway Exchange (BioPAX) format enables the exchange of biological pathways in general [9].

The aim of the BioPAX project is to develop a standard data exchange of biological pathways based on XML. The project is structured in several levels. The first level provides an exchange language of metabolic pathways; the second level allows the access to molecular interaction databases. The development of the third level has been initiated and will include signal transduction and regulatory networks. We are currently working in collaboration with the BioPAX team, specifically in the expansion of the ontology related to transcriptional regulation of bacteria.

The aim of this work is:

- to collaborate with the BioPAX team to expand the scope of BioPAX;
- to extend the format to support gene regulation information;
- to establish a relationship so that data contained in RegulonDB can be translated into the BioPAX format;
- to implement a process to generate a BioPAX file containing the full data of RegulonDB.

Methodology

The process of translating data from RegulonDB into the BioPAX format can be divided in four general activities:

- understanding the BioPAX format. To make a correct translation of the data is an important step to understand the BioPAX format in terms of structure, concepts and definitions; and by means of this, the representation of knowledge;
- indentifying entities in RegulonDB. As we mentioned earlier, RegulonDB contains the complete detailed information of the elements of the transcriptional regulatory network of *E. coli*, thus, the RegulonDB physical entities involved in this process need to be identified;
- expanding the format with transcriptional regulation and adaptation of RegulonDB entities into BioPAX. This activity consists on mapping RegulonDB into BioPAX. During the mapping process, it was found that there was not a right place or class to map some RegulonDB data; it was therefore necessary to create or expand a number of BioPAX classes;
- automation of the translation process. A computer program was developed and implemented in order to translate the RegulonDB data into the BioPAX format.

Implementation

a) Understanding the BioPAX ontology

Considering that BioPAX is an ontology based format, it is necessary to understand the generic components of ontologies and how they represent biological knowledge.

Ontology components

Individuals: also known as instances of a class, they represent objects in the domain we are interested in [10]. For example: the araC gene.

Properties: the characteristics of individuals that can also express relations with others individuals. For example: sequence, binds to.

Classes: they can be interpreted as sets that contain individuals. They are described in formal ways that state precisely the requirements for membership to the class. For example, the Protein class would contain all the individuals that are proteins within our domain of interest [10].

In the BioPAX format, the root class for all the interacting components in the ontology is called "Entity" and represents a discrete biological unit. The entities include pathways, interactions and physical entities. The class that serves to represent entities with a physical structure is called "Physical Entity" and contains several subclasses.

An example of a class, its properties and an individual belonging to that class is presented below.

A class:

Class	Protein
Properties	Name
	Cellular Location
	Comment
	References

Individual of this class:

Name	AraC
Cellular Location	Cytoplasm
Comment	Molecular Weigth: 33.384 Isoelectric point: 6.95
References	Hendrickson W., et a., 1990.

b) Identifying entities in RegulonDB

We identified RegulonDB physical entities involved in the transcriptional regulation process to establish a relation (mapping) between BioPAX and RegulonDB (Fig.1).

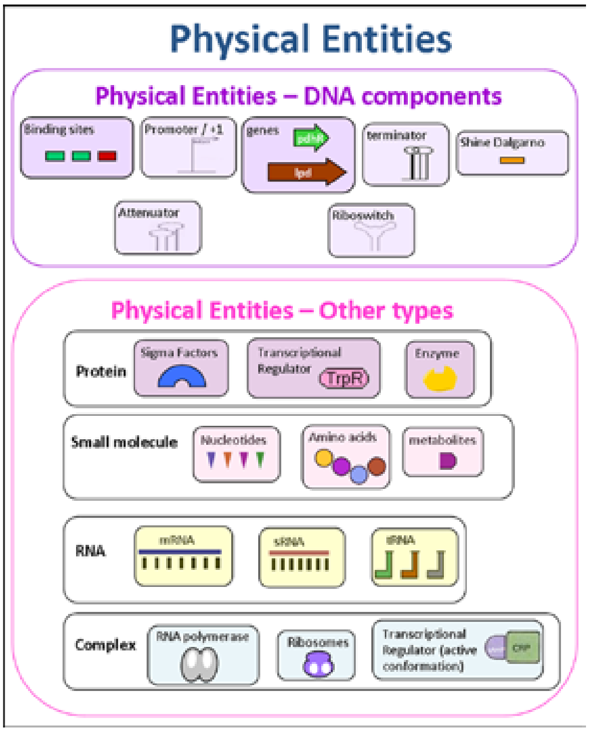


Figure 1. RegulonDB physical entities to mapping into the BioPAX format

c) Expanding the format that supports transcriptional regulation and adaptation of RegulonDB entities into BioPAX

The mapping process consisted on matching the relationships of each element of the regulatory network with the overall framework of the BioPAX ontology.

This was achieved by using the description of BioPAX classes to relate each RegulonDB entity with one or more classes. However, as BioPAX did not have enough support to represent genetic regulation, there was no place to map some data from RegulonDB. We collaborated with the BioPAX team to include details of the genetic regulation specifically; as a result, the format was improved, obtaining a better representation of it. This also allowed us to create rules to validate errors or warnings that will be useful in a future in terms of validating gene regulation behavior. Thus, classes, such as Template Reaction Regulation, Template Reaction, Dna Region, Dna Region Reference, Rna Region and Rna Region Reference were created and/or modified to have a suitable representation of the genetic information (Fig. 2).
The final relationship (mapping) between the RegulonDB and BioPAX is shown on Table 1.

This table contains an overview of the actual mapping; although, we also created a detailed document describing the specific properties of the involved elements. This document was

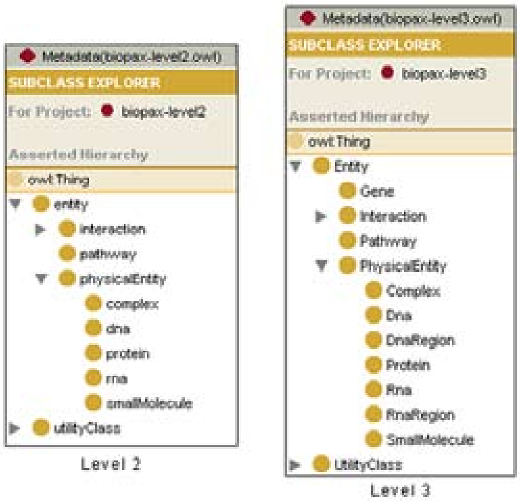


Figure.2. Example of changes on the BioPAX ontology.

used as a guide or template to translate all the RegulonDB data to BioPAX [11].

d) Automation of the translation process

RegulonDB contains a huge amount of data; therefore, translation must be an automated process. A computer program was implemented in Java; the algorithm used to translate the data is, in general terms, the following: load and open the BioPAX ontology; connect to RegulonDB; create individuals from each table. The class these individuals belong to in BioPAX is determined by a mapping that uses the document described above. For instance: there are 4000 genes stored in the RegulonDB table Gene, along with their characteristics. The program creates 4000 individuals and places them with all their properties in the BioPAX DnaRegion class, following the relation established in the mapping document.

Finally, when the program completes the translation, a file containing all the RegulonDB data represented in BioPAX format is generated (http://regulondb.ccg.unam.mx/download/RegulonDBSignIn.jsp); this file can be visualized using an ontology editor such as Protégé or a similar one.

Figure 3 shows the process described above.

Conclusions

We have collaborated with the BioPAX team to include details of genetic regulation to generate a complete and robust format. This expanded schema proved to be adequate to contain transcriptional regulation information, since we could successfully translate our data into the standard

RegulonDB	BioPAX
Object External DB Link	Unification Xref
Publication	Publication Xref
Evidence	Evidence and Evidence Code
Gene	Dna Region and Dna Region Reference
Product	Rna Region and Rna Region Reference
	Protein and Protein Reference
	Cellular Vocabulary
Gene Product Link	Template Reaction
Site	Dna Region and Dna Region Reference
Promoter	Dna Region and Dna Region Reference
Promoter Feature	Dna Region and Dna Region Reference
Terminator	Dna Region and Dna Region Reference
Effector	Small Molecule
Conformation	Complex
Transcription Unit	Dna Region and Dna Region Reference
Regulatory Interaction	Template Reaction Regulation
Attenuator	Dna Region and Dna Region Reference
	Dna Region and Dna Region Reference
Attenuator Terminator	Dna Region and Dna Region Reference
	Dna Region and Dna Region Reference
Shine Dalgarno	Dna Region and Dna Region Reference
Rfam	Dna Region and Dna Region Reference
Motif	Dna Region and Dna Region Reference
Operon	Pathway

Table 1. Relation between RegulonDB and BioPAX

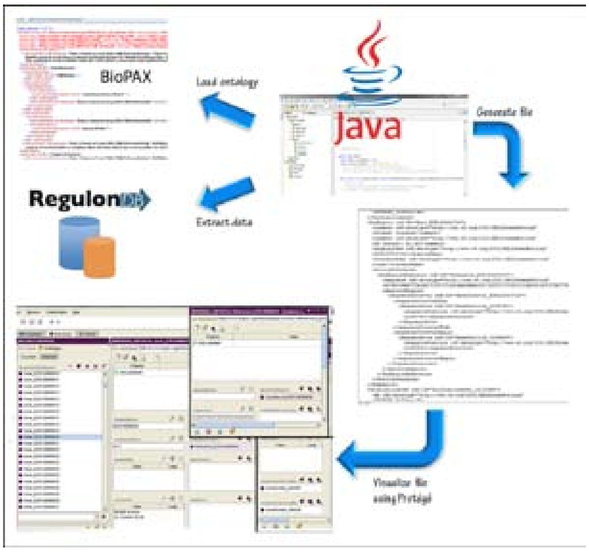


Figure.3. Diagram showing the process of mapping

format; the BioPAX file containing the full version of RegulonDB can be downloaded from the RegulonDB Web page at the main menu in Downloads/Full Version option.

RegulonDB data into the BioPAX format gives a consistent format to the database, which also strengthens this bioinformatics platform. Therefore, it will facilitate the process of sharing information with other databases, as well as making RegulonDB compatible with other standards and allow the addition of new types of data.

Acknowledgements

We would like to thank Gary D. Bader for his support in biological discussions; to Emek Demir and Luis José Muñiz for their invaluable computational assistance, and to Victor Del Moral and Romualdo Zayas for their technical support.

This work was supported by the National Institutes of Health grant GM071962.

References

1. Jacob F, Monod J: Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 1961, 3:318-356.
2. Patrick R. Murray MAP: Microbiología Médica Genética bacteriana, 5 edn. España: Elsevier; 2006.
3. Pierce BA: Genetics. . In: A conceptual approach. Edited by Company WHFa, 2nd edition edn; 2005.
4. 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, Santos-Zavaleta A: Bioinformatics resources for the study of gene regulation in bacteria. J Bacteriol 2009, 191(1):23-31.
5. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C et al: The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol 2004, 22(2):177-183.
6. Buetow KH: Cyberinfrastructure: empowering a "third way" in biomedical research. Science 2005, 308(5723):821-824.
7. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A et al: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 2003, 19(4):524-531.
8. Lloyd CM, Halstead MD, Nielsen PF: CellML: its future, present and past. Prog Biophys Mol Biol 2004, 85(2-3):433-450.
9. BioPAX: Biological pathway exchange. [http://www.biopax.org]
10. The Protégé Ontology Editor and Knowledge Acquisition System [http://protege.stanford.edu/]
11. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H et al: 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.