Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Exploring molecular networks using MONET ontology

João Paulo Müller da Silva, Ney Lemke, José Carlos Mombach, José Guilherme Camargo de Souza, Marialva Sinigaglia and Renata Vieira

Laboratório de Bioinformática e Biologia Computacional;
Universidade do Vale do Rio dos Sinos - Unisinos,
Av. Unisinos, 950, 93022-000 São Leopoldo, RS, Brasil
Corresponding author: N. Lemke
E-mail: lemke@unisinos.br

Genet. Mol. Res. 5 (1): 182-192 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

ABSTRACT. The description of the complex molecular network responsible for cell behavior requires new tools to integrate large quantities of experimental data in the design of biological information systems. These tools could be used in the characterization of these networks and in the formulation of relevant biological hypotheses. The building of an ontology is a crucial step because it integrates in a coherent framework the concepts necessary to accomplish such a task. We present MONET (molecular network), an extensible ontology and an architecture designed to facilitate the integration of data originating from different public databases in a single- and well-documented relational database, that is compatible with MONET formal definition. We also present an example of an application that can easily be implemented using these tools.

Key words: MONET, Ontology, Molecular network

INTRODUCTION

One of the most important challenges for biology in the post-genomic era is to understand the structure and behavior of the complex web of molecular interactions that controls cell behavior (Barabási and Oltvai, 2004). The huge and complex amount of biological data collected during the last years includes information that requires an integrative approach (Uetz et al., 2002). It imposes on computer scientists and biologists to search for innovative methods to deal with these data in a way that will increase our understanding of the underlying biological processes that operate inside the cell (Ideker et al., 2001; Uetz et al., 2002; Barabási and Oltvai, 2004; Yeger-Lotem et al., 2004).

However, this integration task is difficult since biological data are disseminated in many different databases. These databases have different management systems, formats and manners of representing the stored data. Most of them are accessible by flat files or by web interfaces that allow some kind of query. The two main problems involved here are the difficulty in parsing the data when dealing with heterogeneous flat file formats and the inconsistencies due to the absence of a unified vocabulary.

In bioinformatics, ontologies are crucial for maintaining the coherence of a large collection of complex concepts and their relationships (Baker et al., 1999). An ontology is an explicit specification of a conceptualization (Gruber, 1993). While controlled vocabularies only restrict the words used to describe a domain, ontologies extend this simple control vocabulary feature and allow formal specification of terms and their relationships. They make it possible to share and reuse knowledge. They support the interoperability between systems, and they also allow inferences from the represented knowledge. Even when data integration proposals such as mediators are used, ontologies aid in the process of knowledge sharing, such as with Transparent Access to Multiple Bioinformatics Information Sources (Goble et al., 2001).

In this context, we present MONET (molecular network) ontology, an integrated model for the integration of the different molecular networks (Barabási and Oltvai, 2004) that exist inside of the cell. Such an integrated view aims to help understand the large-scale interactions responsible for the behavior of the cell, help predict aspects of cellular behavior that can be tested experimentally (Ideker et al., 2001), and help formulate new hypotheses, such as putative functions for open reading frames and candidates for essential genes.

Domain analysis

Ontologies can be used for communication between systems, people and organizations, to support generic knowledge-based system projects and development. However, only a few applications use ontologies. Gómez-Pérez (1999) indicated that this happens mainly because:

- ontologies are developed for a specific application, without regard for questions of reuse and sharing, making it difficult to reuse existing ontologies;
- ontologies are dispersed over several servers;
- formalization differs depending on the server on which the ontology is stored;
- ontologies on the same server are usually described with different levels of detail, and
- there is no common format for representing relevant information about ontologies so that users can decide which ontology best suits their propose.

Bioinformatics is an area in which a large variety of ontologies have been proposed in the last years. Here, we present some of the ontologies that are available in the molecular biology domain:

- The GO (Gene Ontology, http://www.geneontology.org) is the most ambitious project applied to biology. GO aims to provide an ontology that encloses diverse domains of cellular and molecular biology, being divided into three sub-ontologies: biological processes (formed by one or more sets of molecular functions), cellular functions (describes activities on a cellular level) and cellular components (enumerates the cell location considering sub-cellular structures). These sub-ontologies have a concept-hierarchy organization (Yeh et al., 2003). The sub-ontologies are used in gene annotation, products and sequences.
- SO (Sequence Ontology Project, http://song.sourceforge.net) is a project that joins efforts between genome annotation centers; its goal is to provide an ontology for application in annotation sequences and data exchange. It is under development, and its interim releases are made available as soon as considered to have a usable status.
- The PSI (Proteomics Standards Initiative, http://psidev.sourceforge.net) is a molecular interaction ontology that represents protein-protein interactions. The PSI-MI is a HUPO effort and was implemented through the specification of an ontology. The current level implements declarative representations of molecular interactions: feature detection (method used to determine features involved in the interaction), feature type (properties of subsequences that intervene in the binding of the proteins), interaction detection (method used to determine interactions between proteins), interaction type (method to determine the physical interaction between two proteins), and participant detection (used to detect the proteins involved in an interaction).
- The primary purpose of the Microarray Gene Expression Data (MGED, http://www.mged.org) ontology is to provide standard terms for the annotation of microarray experiments. The modeling of representation of microarray data requires complex structures; the inexistence of a universally accepted data format complicates various processes, such as data-interchange and data documentation (Spellman et al., 2002). Brazma et al. (2001) has shown that there are distinct representations for microarray data, which make the reproduction of experiments a problematic task. This ontology (still under development) enables unambiguous descriptions of how experiments are performed.
- The goal of BioPAX (Biological Pathways Exchange Language, http://www.biopax.org) is to facilitate the integration and data-exchange of biological databases concerning pathways. Despite all the efforts directed towards ontology development, exemplified by the work described above, data integration is still a challenge in Bioinformatics. A possible solution for this problem is to define a standard format to represent these data; however, there is no standard format applicable for pathways, despite the availability of about one hundred databases of pathways on the internet. The BioPAX project goal is to provide a format for data-exchange concerning pathways in the most popular databases, and to reach this goal BioPAX ontology was designed to support existing data models, such as BioCyc (http://www.biocyc.org), BIND (http://www.bind.ca) and KEGG (http://www.genome.jp/kegg).

The ontologies mentioned above give us an idea of the efforts of the community to cover the vast area of molecular biology. The ontology we are proposing is inspired by the construction of integrated topological models of molecular networks. None of the available ontologies deal with the integration of data for experiments involving the different molecular networks inside cells.

We acknowledge the usefulness of each of the efforts referred to above for this task; however, none of them is adequate for our general goal. To accomplish this task, we propose the MONET ontology.

MONET ontology

The definition of an ontology is time-consuming. Any aid can provide a significant productivity profit. To develop MONET ontology, we used Protégé 3.1 as the editing tool (http://protege.stanford.edu/), for two reasons: a) the need, not only for an ontology editor, but also for a Knowledge Base Management System, since we wanted to populate the database with information about various microorganisms, and b) its open-source Java-extensible architecture allows improvements in its functionalities through the aggregation of new plugins. This latter characteristic allows the ontology to be exported to the different formats required by different research groups. A variety of import/export plugins can be used to automatically read/write the ontology in different representations. We developed MONET as an OWL (Ontology Web Language) ontology. The technical vocabulary used to describe MONET, concerning the ontology (not the biological knowledge), is based on Protégé OWL. Its logic-based description defines an ontology as a formal explicit description of concepts in a domain of discourse (concepts or classes), properties of each concept describing various features and attributes of the concept (roles or properties), and restrictions on classes and roles.

MONET ontology is a model for the web of molecular interactions that determine cell behavior. It represents relevant biological knowledge and allows data integration regarding interactions involving the different biomolecules (Ideker et al., 2001; Uetz et al., 2002; Barabási and Oltvai, 2004; Yeger-Lotem et al., 2004). Since the necessary data are distributed among different databases over the internet, an ontology for this domain must satisfy the following requirements:

- it should minimize data redundancy and inconsistency;
- the data-interchange problem must be taken into consideration through the adoption of free and open standards, and
- it needs to be extensible, so that new knowledge can be easily implemented by the aggregation of new concepts.

MONET integrates information from metabolic pathways, protein-protein interaction networks for eukaryotes and prokaryotes, and transcription-regulatory interactions for prokaryotes, through a model able to minimize data redundancy and inconsistency. Figure 1 shows a schematic view of the proposed ontology.

In Figure 2, we show a schematic view of the different concepts present in the MONET ontology. Most of these concepts are organism-dependent, with the exception of general chemical reactions and some of its dependent concepts. Some of the concepts do not depend on the

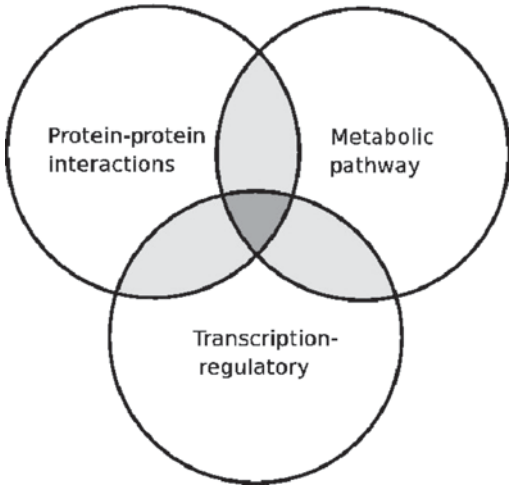

Figure 1. Current networks modeled in MONET intersections represent the existence of common concepts shared between these networks.

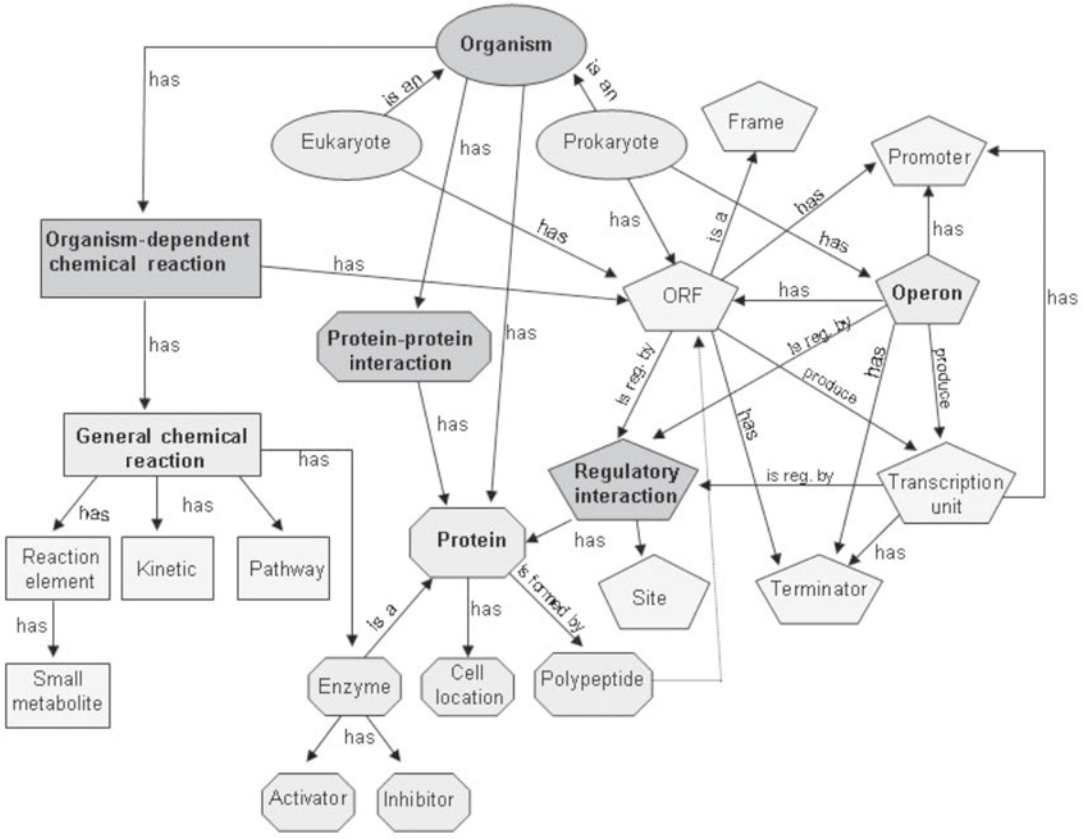

Figure 2. Concept relations in MONET ontology. Rectangles represent metabolic pathway networks; octagonal forms represent protein-protein interaction networks; the transcription-regulatory networks are shown as pentagons, and ellipses represent super kingdoms (eukaryotes and prokaryotes). ORF = open reading frame.

super kingdom, such as protein-protein interactions and open reading frames, while in the case of regulation there are important differences between the different super kingdoms. We plan to extend the ontology to include regulation in eukaryotes. The most important MONET concepts are described in Table 1.

Table 1. Definition of MONET ontology concepts.

Concept	Definition
Activator	An activator is a substance, other than the catalyst or one of the substrates, which increases the rate of a catalyzed reaction. An activator of an enzyme-catalyzed reaction may be called an enzyme activator if it acts by binding to the enzyme.
Cell location	Indicates the protein's sub-cellular location.
Enzyme	A type of protein that catalyzes chemical reactions in the organisms.
Frame	Refers to different phases of possible readings of a DNA sequence; the determination of the correct phase has specific criteria, thus, when this phase is determined, the DNA sequence can be translated to its corresponding amino acids.
General chemical reaction	Chemical reaction is a process that results in the interconversion of chemical species. General chemical reaction is not related to an organism. It refers to the all possible chemical reactions.
Inhibitor	An inhibitor is a substance that diminishes the rate of a chemical reaction, and the process is called inhibition. In enzyme-catalyzed reactions an inhibitor frequently acts by binding to the enzyme, in which case it may be called an enzyme inhibitor.
Kinetic	Kinetic equations are commonly expressed in terms of concentrations of the chemical species involved. Concentration is the amount of -the-substance (for which the SI unit is the mole, symbol mol) divided by the volume.
Operon	An operon consists of an operator, promoter, regulator, and structural genes. The genes in the operon are located contiguously on a stretch of DNA and are under the control of one promoter (a short segment of DNA to which the RNA polymerase binds to initiate transcription). A single unit of messenger RNA (mRNA) is transcribed from the operon and is subsequently translated into separate proteins.
ORF	Open reading frame occurs when successive nucleotide triplets can be read as codons specifying amino acids and where the sequence of these triplets is not prematurely interrupted by stop codons; the sequence is (potentially) translatable into a protein.
Organism-dependent chemical reaction	Chemical reactions related specific to an organism.
Pathway	Specifies biochemical interactions (biochemical reactions) in a metabolic map.
Polypeptides	A single linear chain of amino acids; proteins are composed of one or more polypeptide chains.
Promoter	Regulatory region of DNA involved in binding of RNA polymerase to initiate transcription.
Protein	A complex molecule consisting of one or more polypeptides; proteins vary in structure according to their function.

Continued on next page

Table 1. Continued.

Concept	Definition
Reaction element	The substrate (or reactant), product and enzyme in an enzymatic reaction. A substrate is a reactant (other than a catalyst) in a catalyzed reaction. Product is a substance that is formed during a chemical reaction. Enzyme is a type of protein that catalyzes chemical reactions in the organism. In the reaction element, the number of molecules in each species involved in the chemical reaction is called stoichiometric number.
Regulatory interaction	Occurs through the interaction between a specific protein and a DNA region (a sequence regulatory function).
Site	Refers to a sequence of DNA with a regulatory region.
Small metabolite	A small metabolite is any intermediate or product resulting from metabolism.
Terminator	Sequence of DNA, found at the end of the transcript, which causes RNA polymerase to terminate transcription.
The small molecule metabolism (metabolic network) of MONET	Is a subset of the complete metabolism that excludes DNA replication and protein synthesis reaction.
Transcription unit	Entire DNA sequence that is read from the start of transcription to the termination site; it may include more than one gene.

Whereas the transcription-regulatory network is involved with interactions between DNA and proteins, and with the consequent production of proteins, the metabolic network involves proteins characterized by their enzymatic function. In fact, proteins are the main common link between these networks. The protein-protein interaction network contemplates binary interactions among proteins.

The term metabolism comprises the entire groups of physical and chemical processes involved in the maintenance and reproduction of life, in which nutrients are broken down to generate energy and to give simpler molecules (catabolism), which may be used to form more complex molecules (anabolism).

Although the structures of metabolic networks and protein interaction networks are similar, there are a number of significant differences. While metabolic networks focus on the conversion of small molecules and the enzymes responsible for these conversions, protein interaction networks concentrate mainly on physical contacts without obvious chemical conversions (Uetz et al., 2002).

The spatial aspect was also considered; MONET implements a concept entitled cell location to indicate the protein's sub-cellular location. The location of a protein and other chemicals is an important feature in network modeling.

Data integration

In the current state, the MONET database has data on prokaryotic (*Escherichia coli*

and *Helicobacter pylori*) and eukaryotic (*Saccharomyces cerevisiae*) organisms. It is easy to include data for new organisms, since we developed and implemented an architecture to facilitate data integration. This architecture is based on the data transformation approach. This method merges data from different databases in a unique storage repository that allows for data manipulation and study. We used the database PostgreSQL, as it is open source, fast and robust, and because it has more programming functions than other databases with the same purpose. The data integration process is divided into three stages, summarized below:

- Data acquisition (the data is acquired from downloads and is loaded in our database).
- Normalization and integration (in this stage it is necessary to create parsers to manipulate the data that were acquired in the previous stage and to translate the data to a standard format, because each database has its own format; later these data are loaded into the database).
- Data cleaning (the purpose of this process is to correct inexact or incorrect data; this process is executed together with a molecular biology specialist, by comparing data from different databases and technical literature).

Figure 3 shows a schematic representation of this process.

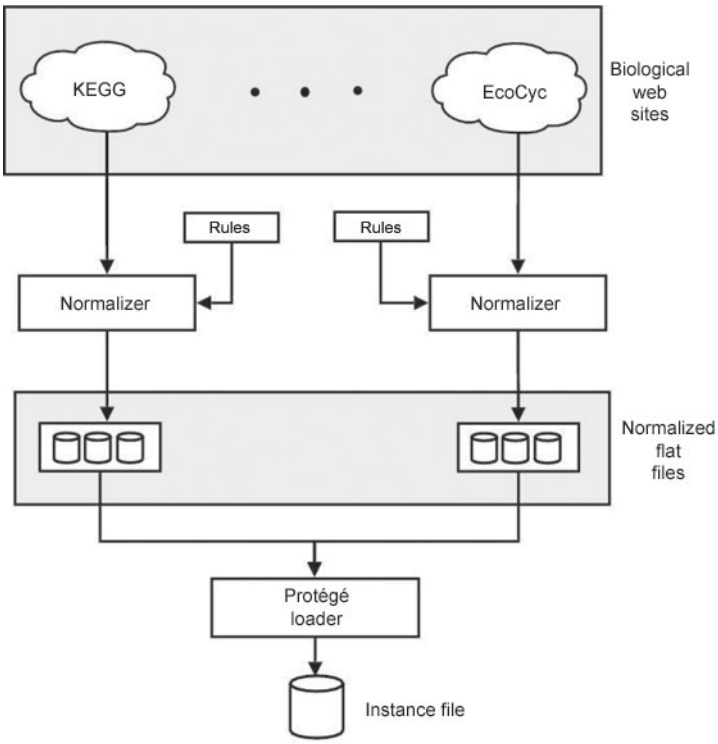

Figure 3. A schematic representation of the data integration architecture that is used to populate MONET ontology.

The next task is to populate the MONET ontology from the data storage. At this level, the necessary relationships between the data are created to generate the tables in accordance with the relationships between concepts (Table 1) defined by the ontology. After creating the relationships of the tables, the data are first exported to text files, and then an OWL file is generated with instances of the ontology. Table 2 presents the data sources used to populate the ontology, while Table 3 presents the number of instances for each concept.

Table 2. Data sources used for the construction of MONET ontology.

Network	Database
Metabolism	KEGG [http://www.genome.jp/kegg/]
Regulation	RegulonDB [http://www.cifn.unam.mx/Computational_Genomics/regulondb/]
Protein-protein interaction	Butland G et al. (2005)

Table 3. MONET concepts and the number of instances of each concept.

Concept	Instances	Concept	Instances
Enzyme	4650	Protein	9970
General chemical reaction	6469	PPI	12248
Operon	784	Reaction element	18194
ORF	10614	Regulatory interaction	1376
Organism	3	Site	1216
Organism-dependent chemical reaction	5838	Small metabolite	23954
Pathway	238	Terminator	137
Promoter	973	Transcription unit	833

ORF = open reading frame; PPI = protein-protein interaction.

RESULTS

Monet ontology was designed to facilitate the construction of integrated molecular networks of organisms. We discuss here results for *E. coli*, which was chosen since it has very comprehensive datasets, especially for regulation and metabolism. The nodes of this network are genes; genes g1 and g2 that code for proteins p1 and p2 are linked if:

- p1 and p2 interact physically,
- g1 regulates the transcription of gene g2, or
- a product generated by a reaction catalyzed by p1 is consumed in a reaction catalyzed by p2 (we excluded from this analysis the most frequently used compounds, such as ATP, NAD, H_2O, etc.).

Following this procedure we obtained a network with 1508 genes and 37,636 interactions. For each gene we calculated the number of links it has, this quantity is called connectivity.

We expect that the most important genes have high connectivities. The vast majority of the genes are connected to less than five genes, while some of the genes are highly connected with almost 300 connections. Table 4 presents the most connected genes.

Table 4. The most connected genes in the *Escherichia coli* network.

Gene	Connectivity
crp	272
metK	229
aceE	182
aceF	181
trpD	176
lpdA	175
Ihf	170
ygdP	162
Fis	156
sixA	151

It is not surprising that cyclic AMP receptor protein (CRP) is the most connected gene in *E. coli*, since it codes an important transcription factor that regulates transcription initiation for more than 100 genes that are mainly involved in the catabolism of carbon sources other than glucose. *E. coli* preferentially uses glucose over other sugars, and it only catabolizes other sugars when the supply of glucose has become depleted. The presence of glucose prevents *E. coli* from catabolizing alternative sugars through several mechanisms; one of which is that glucose lowers the level of cAMP, the inducer for CRP. CRP also regulates the transcription of genes required for energy production, amino acid metabolism, nucleotide metabolism, and ion transport systems. In addition, CRP can regulate the transcription of other transcription factors, such as MelR, RpoH, BlgG, Fis, and PdhR (this analisys was performed using the Kegg database http://www.genome.jp/kegg/).

CONCLUSIONS

We presented MONET ontology as a tool to facilitate the construction of integrated models for organisms, by introducing an architecture that reduces data inconsistencies and redundancy. This ontology was applied to the construction of an integrated model for the *E. coli* molecular network; in this network, the most connected component is the *crp* gene, an important molecule in the regulation of *E. coli* metabolism.

ACKNOWLEDGMENTS

Research developed in collaboration with HP Brazil R&D and partially supported by CNPq, process number 401999/2003-3.

REFERENCES

Baker PG, Goble CA, Bechhofer S, Paton NW, et al. (1999). An ontology for bioinformatics applications. *Bioinformatics* 15: 510-520.

Barabasi AL and Oltvai ZN (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet*. 5: 101-113.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, et al. (2001). Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat. Genet*. 29: 365-371.

Butland G, Peregrin-Alvarez JM, Li J, Yang W, et al. (2005). Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* 433: 531-537.

Goble C, Stevens R, Ng G, Bechhofer S, et al. (2001). TAMBS: Transparent Access to Multiple Bioinformatics Information Sources. *IBM Syst. J.* 40: 532-552.

Gómez-Pérez A (1999). Ontological engineering: a state of the art. *Expert Update* 2: 33-43.

Gruber TR (1993). Toward principles for the design of ontologies used for knowledge sharing. In: Formal ontology in conceptual analysis and knowledge representation (Guarino N and Poli R, eds.). Kluwer Academic, Dordrecht, Netherlands.

Ideker T, Thorsson V, Ranish JA, Christmas R, et al. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292: 929-934.

Ogata H, Goto S, Sato K, Fujibuchi W, et al. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 27: 29-34.

Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, et al. (2004). RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res*. 32: D303-D306.

Spellman PT, Miller M, Stewart J, Troup C, et al. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*. 3: 461-469.

Uetz P, Ideker T and Schwikowski B (2002). Visualization and integration of protein-protein interactions. In: Protein-protein interactions - a molecular cloning manual (Golemis E, ed.). Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.

Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, et al. (2004). Network motifs in integrated cellular networks of transcription-regulation and protein-protein interactions. *Proc. Natl. Acad. Sci. USA* 101: 5934-5939.

Yeh I, Karp PD, Noy NF and Altman RB (2003). Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics* 19: 241-248.