Using Protégé to build a Molecular Network Ontology

Eduardo Battistella, Renata Vieira, José G. C. de Souza,
Adriana N. dos Reis, João Paulo M. da Silva, Cláudia K. Barcellos,
Norma M. da Silva, Guilherme B. Bedin, José C. M. Mombach, Ney Lemke

UNISINOS, Brazil

One of the most important challenges for biology in the postgenomic era is to understand the structure and behavior of the complex intercellular web of molecular interactions that controls cell behavior [Barabási&Oltvai, 2004]. The huge and complex amount of data collected during the last years holds information that requires an integrative approach [Uetz *et al*., 2002]. It imposes to computer scientists and biologists the search for innovative methodologies to deal with the data in a way to increase our understanding of the underlying biological processes that operate inside the cell [Barabási&Oltvai, 2004,Yeger-Lotem *et al*., 2004,Uetz *et al*., 2002,Ideker *et al*., 2001]. However, the integration process is complicated because the data are spread geographically in the web. The databases, in turn, have diverse management systems, formats and different ways of representing the data. Most them are accessible by flat files or by web interfaces that allow some kind of query over it. 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 for definition of data.

In bioinformatics, ontologies are crucial for maintaining the coherence of a large collection of complex concepts and their relationships [Baker *et al*., 1999]. They allow the sharing and reusing of formally specified knowledge, and inferences can be made based on the represented knowledge. Examples of existing ontologies in the molecular biology domain are the Gene Ontology (GO) (www.geneontology.org), the Sequence Ontology (SO) (song.sourceforge.net), the Proteomics Standards Initiative Molecular Interaction (PSI MI) (psidev.sourceforge.net), the Microarray Gene Expression Data (MGED) (www.mged.org), and new efforts such as BioBabel (www.ebi.ac.uk/biobabel). Based on these ontologies we introduce MONET, the MOlecular NETwork ontology. An integrated model for the "*network of networks*" [Barabási&Oltvai, 2004] that exists inside the cell. Such integrated view aims the understanding of the large-scale interactions responsible for the behavior of the cell, to predict cellular behaviors that can be tested experimentally [Ideker *et al*., 2001], and to formulate new hypotheses.

MONET integrates information from metabolic pathway, transcription-regulatory, and protein-protein interaction networks with data from prokaryote and eukaryote organisms. Aiming to establish a model that minimizes data redundancies and data inconsistencies. The transcription-regulatory network contains concepts such as *Operon* (a set of genes transcribed under the control of an operator gene), *Transcription Unit* (part of DNA that will be transcribed into a RNA), *Terminator* (DNA region where the transcription supposedly stops), *ORF* (a portion of a gene sequence that could potentially encode a protein), *Site* (DNA sequence whose location and base sequence are known), *Promoter* (a segment of DNA which provides a site where the enzymes involved in the transcription process can bind to a DNA molecule, and initiate transcription), *Reg-*

ulatory Interaction (general information concerning the transcription-regulatory data being mapped) and *Protein* (a complex, high molecular weight organic compound that consists of aminoacids joined by peptide bonds).

Whereas the transcription-regulatory network is involved with interactions between DNA and proteins, and the consequent production of proteins, the metabolic network involves proteins characterized by its enzymatic function. In fact, proteins are the main common link between these networks. The protein-protein interaction network contemplates binary interactions among proteins. We adopted the concept of *interaction detection* from the PSI-MI ontology, the method to determine the interaction is then divided in the sub-methods *experimental* and *in silico*.

The small molecule metabolism (metabolic network) of MONET is a subset of the complete metabolism that excludes DNA replication and protein synthesis reaction. The concepts here are: *General Chemical Reaction* (referring to the chemical reactions that occur in different organisms), *Organism Dependent Chemical Reaction* (chemical reactions that occur in specific organisms), *Reaction Element* (in an enzymatic reaction, they are substrate (or reactant), product and enzyme), *Substrate* (a reactant (other than a catalyst) in a catalysed reaction), *Product* (a substance that is formed during a chemical reaction), *Pathway* (shows biochemical interactions - biochemical reactions), *Enzyme* (a type of protein that catalyses chemical reactions in the organisms), *EC* (the enzyme commission number) other concepts such as *Inhibitor* (a substance that diminishes the rate of a chemical reaction and the process is called inhibition), *Activator* (a substance, other than the catalyst or one of the substrates, that increases the rate of a catalysed reaction). 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 enzyme 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 *Compartment* to indicate the protein's subcellular location. The location of a protein and other chemicals is an important feature in the network modeling.

At its current stage, MONET includes concepts and data related to metabolic pathway networks, transcription-regulatory networks, and networks of protein-protein interactions. It is easily expandable, existing ontologies can be incorporated into the model to increase the coverage of molecular biology domain. In this way, MONET allows the construction of topological models of cells of microorganisms and the extension of these models as new biological knowledge becomes available.

To build MONET we use Protégé-2000 (http://protege.stanford.edu). The main reasons for choosing this tool are:(a) the need, not only for an ontology editor, but for a Knowledge Base Management System (KBMS) since we want to populate the database with instances from various microorganisms; (b) its open source Java extensible architecture allows improvements in its functionalities through the aggregation of new plugins. A variety of import/export plugins can be used to automatically read/write the ontology in different representation data standards like Web Ontology Language (OWL), Resource Description Framework (RDF), Extensible Markup Language (XML) and others. We populated our knowledge base with instances from some microorganisms

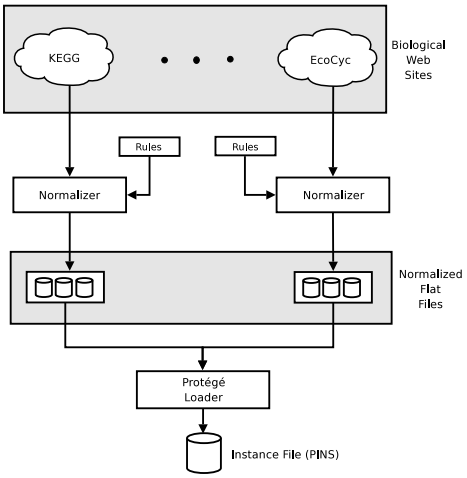

Fig. 1. Unifying biological data available over the Internet into MONET Ontology.

(*Escherichia coli, Saccharomyces cerevisiae and Helycobacter pylori*). We have incorporated KEGG's Ligand database[1], as instances to the metabolic pathway networks. And we have an automated procedure to include organism dependent metabolic information. To achieve this, we developed python scripts to normalize the data available in the flat files, execute a series of consistency checks to correct the existing inconsistencies, and automate the generation of the instance flat file of Protégé (Figure 1). The result of this process in number of instances for each concept are presented in Table 1.

Concept	Instances	Concept	Instances	Concept	Instances
General Chemical Reaction	4496	Enzyme	3407	Operon	785
Organism Dependent C. R.	3238	ORF	4410	Organism	3
Small Metabolite	3361	Product	8990	Promoter	973
Protein-Protein Interaction	12248	Reaction Element	17757	Protein	10201
Regulatory Interaction	1376	Site	1216	Pathway	126
Transcription Unit	833	Substrate	8767	Terminator	137

Table 1. Number of instances for each concept of MONET Ontology.

Considering other ontologies such as GO, PSI MI, MGED, and SO, MONET has a different point of view on knowledge modeling. GO attacks the annotation problem, MONET is not in this stage yet. PSI MI deals with molecular interactions, MONET also deals with this problem and incorporates most of the concepts available at PSI

[1] http://www.genome.ad.jp/kegg

MI. MGED covers microarray experiments, MONET does not. SO offers a way for sequence annotation and for data-interchange of this annotation, MONET also does it by incorporating most of SO concepts. While these ontologies are specific to a particular aspect in the molecular biology domain, MONET extends and integrates them into a holistic perspective of the cell. In our view, our proposal is a way to achieve the understanding of the cell internal organization. It is not a static or a complete model, but we consider it is an important step in a direction that can lead us to a comprehensive modeling of various networks that control the cellular behavior.

Besides ontologies, there are other alternatives for integration of biological data, such as BioWarehouse. This is an open source environment that integrates data originating from different public databases in a single database designed to help data management, mining and exploration. This system is restricted to metabolism, taxonomy, and genomic data. The main differences between our approach and the latter is that we manually create the data model and we include different databases using an strategy that minimizes data inaccuracies.

It remains a challenge to integrate data from the myriad of interactions of the cellular constituents. Our model is one of the multiple possible variations concerning the complex, constantly changing, and not yet completely understood area of molecular biology. This approach follows the idea of a "*functional bioinformatics*" [Karp, 2000], a bioinformatics that makes possible the development of new algorithms, graphical visualization interface, and many other tools that help the investigation of principles that govern cellular function. The next steps in our work are to refine MONET to include concepts such as cellular signalling and to use this ontology to build a knowledge base for *Mycoplasma pneumonia* microorganisms. As a result of this topological integrated model of an organism we expect to simplify and speed up the formulation of new models. We plan to use Jena (http://jena.sourceforge.net/) to make inferences about data and to use Racer (http://www.cs.concordia.ca/ haarslev/racer/) to execute consistency checks. For more details about MONET see http://www.inf.unisinos.br/∼lbbc/monet.html.

Acknowledgement This work was developed in collaboration with HP Brazil R&D. This work was partially supported by CNPq, process numbers 401999/2003-3 and 550215/2003-4.

References

[Baker *et al*., 1999] Baker,P. *et al*. (1999) An ontology for bioinformatics applications, *Bioinformatics*, **15**, 510-520.

[Barabási&Oltvai, 2004] Barabási,A., Oltvai,Z. (2004) Network biology: understanding the cell's functional organization, *Nature Reviews Genetics*, **5**, 101-113.

[Ideker *et al*., 2001] Ideker,T. *et al*. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network, *Science*, **292**, 929-934.

[Karp, 2000] Karp,P. (2000) An ontology for biological function on molecular interactions, *Bioinformatics*, **16**, 269-285.

[Uetz *et al*., 2002] Uetz,P., Ideker,T., Schwikowski,B. (2002) Visualization and integration of protein-protein interactions. In Golemis,E.(ed), Protein-Protein Interactions - A Molecular Cloning Manual. Cold Spring Harbor Laboratory Press, 623-646.

[Yeger-Lotem *et al*., 2004] Yeger-Lotem,E. *et al*. (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction, *PNAS*, **101**, 5934-5939.