HAL
open science

TIM/WIM: A Set of Tools to Interface Modelling in Biology.

Francois Vallee, M. Beurton-Aimar, N. Parisey, Florent Collot, Sophie Colombié

To cite this version:

Francois Vallee, M. Beurton-Aimar, N. Parisey, Florent Collot, Sophie Colombié. TIM/WIM: A Set of Tools to Interface Modelling in Biology.. European Conference on Object-Oriented Programming, Jul 2009, Genova, Italy. hal-00629019

HAL Id: hal-00629019

https://hal.science/hal-00629019v1

Submitted on 4 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TIM/WIM: A Set of Tools to Interface Modelling in Biology.

François Vallée
LaBRI - UMR 3800
Université Bordeaux 1
vallee@labri.fr

Marie Beurton-Aimar
LaBRI - UMR 3800
Université Bordeaux 1
beurton@labri.fr

Nicolas Parisey
LaBRI - UMR 3800
Université Bordeaux 1
parisey@labri.fr

Florent Collot
LaBRI - UMR 3800
Université Bordeaux 1

Sophie Colombié
INRA UMR 619
Biologie du fruit

ABSTRACT

Modelling in biology is a complex task. Many types of information are used by biologists but there is a lack of tools for integrating heterogeneous data in a core interface. TIM (Tools to Input Models) is a tool which allows to put in the same interface data which describe biological objects like enzymes, metabolites, DNA, ... and information about biological process modelling like those coming from NMR (Nuclear Magnetic Resonance) experiments and used to simulate flux through metabolic networks. TIM is able to manage the widely used database format, PGDB (BioCyc format), and uses a large part of a biological ontology, BioPAX format, to store information about biological processes. To manage all these data, WIM (Web Interface for Modelling) provides a set of web pages. It generates dynamic html sources using CL-WHO library and integrates an Hunchentoot web server. At present, the application is used by a biologist group to store data about the carbon metabolism of the tomato fruit. They have released the first version of the TomaCyc database and currently used the WIM interface to create automatically input files for softwares that simulate the activity of their metabolism network.

Keywords

System Biology, framework for modelling metabolic networks.

1. INTRODUCTION

In the last ten years, the quantity of biological data available increases exponentially. These data are stored in large databases and are easily accessible from international web sites. However, each of them usually represents a specific category of data: proteome, genome, metabolome and so on. One can note that modelling in biology often needs to take into account data from several sources like genomic databases and simulation models databases [13]. The problem is that in general, there are no direct links between these different kinds of data.

For a project, BioFiL, consisting of designing a generic framework to capitalise and to simulate models for metabolism, we have implemented two components: TIM and WIM to put together descriptive data about metabolism and mathematical models and to automatically generate configuration files for simulation programs. These tools are configured to be linked with PGDB (Pathway/Genome DataBase) data files created from the Pathway Tools application, an application distributed by the Stanford Research Institute (SRI). The first application is for the Tomaflux project. This project studies the biological network activities of the central carbon metabolism of the tomato fruit. It needs to store data describing the biological objects involved in the network and information from nuclear magnetic resonance (NMR) experiments, and to produce input files required by the simulation software.

We will now present the Pathway Tools and BioPAX formats that we have partially used to develop TIM and WIM object model. The third and the fourth sections explain how we have built these two components and we will finish with a brief description of their use to create and to manage the database TomaCyc concerning the tomato fruit metabolism.

2. MODELLING AND DATA MANAGEMENT IN BIOLOGY

Modelling in biology supposes to take into account, on one hand, biological descriptions of molecules and, on the other hand, how to model processes (through simulation procedures, set of equations, ...). In fact, these two kinds of information are rarely associated in the same database or format. Several international projects have the main goal to produce and to manage large databases to store biological data depending on their types: Reactome [1] [4], KEGG[2] [7], BioCyc[3] [15], BioModels[4] [3]. Each of these projects provides a standard format to structure data (resp. BioPAX [6], KEGGML, PGDB [11], SBML [1]), and usually they also provide tools to convert data from one format to another. Most of these formats are XML and are well defined by their respective xsd (XML Schema Description). But no format puts together (in the same entity) all the useful types of information required to model biological processes. For example, there is no way to specify equations or model pa-

[1] http://www.reactome.org
[2] http://www.genome.jp/kegg
[3] http://biocyc.org
[4] http://www.ebi.ac.uk/biomodels-main

rameters in the PGDB format and conversely, SBML cannot provide tags to store information such as DNA sequence or molecular weight.

After a consequent analysis of the different tools available to manage biological data, we have chosen the Pathway Tools framework [8] as a basis to develop our tools. Pathway Tools provides a way to store, to update and to distribute biological data through computer networks. It is written in Lisp, largely distributed in the biologist community and mature enough to be a robust basis for new developments. We will now present this framework before explaining how we have interleaved some parts of it with the BioPAX format to build our own set of modelling tools.

2.1 BioCyc and Pathway Tools

From the 90's on, SRI has developed a set of tools to collect and to re-distribute biological databases. It consists of two main parts: a web front-end, BioCyc, and the Pathway Tools application. BioCyc allows to access a collection of Pathway/Genome DataBases (PGDB).

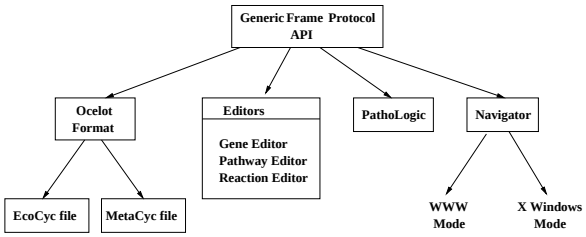


Figure 1: The Pathway Tools software.

The first of these PGDBs was the database for E. coli organism, EcoCyc [2]. Today, BioCyc provides access to 376 databases developed by a large amount of groups of biologists around the world. All these databases respect the PGDB format defined by P. Karp [10]. This format designed firstly for the EcoCyc database has been extended to comply with the complexity of biological data brought by others organisms.

The core of the Pathway Tools application has three components. **PathoLogic** allows users to create PGDB files, and to generate drawings of the biological networks. The **Editing Tools** contain a subset of operations to easily modify existing data. The **Navigator** prints information about PGDB like the BioCyc website. Figure 1 shows the architecture of the Pathway Tools framework. It can be used in a desktop mode or in a web mode. The desktop mode gives access to the three components. The web mode gives only access to the Navigator but has a comparative genomic tool, an advanced query tool and BLAST (Basic Local Alignment Searching Tool) which allows comparisons of DNA or proteic sequences.

Pathway Tools uses Common-Lisp to define classes for biological objects and the graphical user interface is written with CLIM-AllegroCL. The software also accepts external queries from Perl or Java programs thanks to the modules PerlCyc or JavaCyc. The PGDBs relies on the Ocelot object system database [11]. This database management system provides the multiuser access capabilities, scalability and robust operation of relational databases combined with a knowledge representation system based on frames. The PGDBs can be stored either by a classical relational database management system (as Oracle) or in flat files (Ocelot format).

Though the PGDB format has a detailed design, some information on metabolism is missing. This is the reason why we have decided to use also the BioPAX format for our application.

2.2 BioPAX format and biological modelling

BioPAX [6] is a cooperative project of the biologist community. The first release of the BioPAX format dates back to 2002. Large projects like Reactome are based on this format and a lot of softwares exists to manage BioPAX data files and to convert them into others formats (SBML and so on).

Modelling of metabolic activities consists of describing chains of enzymatic reactions. One chain corresponds to a biological function and is called a pathway. A set of pathways builds a metabolic network which can be common for several organisms. Characteristics like kinetic parameters or carbon markers are added to the description of the biological objects (enzymes, substrates, and so on) to describe the processes involved in the pathway. A part of these characteristics can be specified in BioPAX format or in PGDB files. Other specific parts of information, such as carbon markers, are not taken into account in these formats but they could be essential to use particular simulation softwares. For example, they are needed to reproduce NMR experiments in simulations like 13CFlux [16]. As far as we know there is no place to put this information in any standard format for biology modelling, including BioPAX.

We have chosen to design a new model to capture all the details of the biological object descriptions. The BioPAX format is used as basis of our own object model and new concepts are added to obtain an architecture more suitable for our own needs. We will describe our model in the next section.

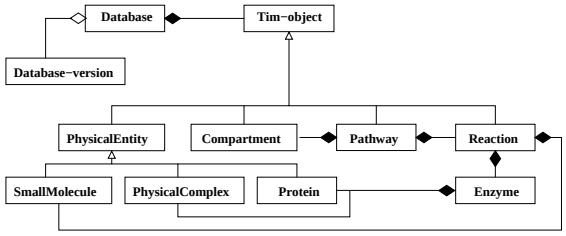


Figure 2: Object diagram from TIM.

3. TIM: TOOLS TO INPUT MODELS

The BioFiL framework is still a work in progress but at present it consists of several applications, such as a web interface to manage PGDB files and NMR data, tools to print biological graphs and a multi-agent simulator to build "in silico" experiments. We have chosen to implement BioFiL in Common Lisp because this language provides us all the nec-

essary tools to address these concerns. As the BioCyc framework is also built and managed by a lisp software, Pathway Tools, it is easy to adapt some functionalities provided by Pathway Tools to our own needs. Though a lot of languages exist to code web interface or simulation computing, using Lisp allows us to have only one way to make that, to give facilities to users to parametrize the tools as they want , for example to load a file with a new set of parameters without need to add a script language. Then, Common-Lisp has an object layer: the Common-Lisp Object System (CLOS) which implements the Meta-Object Protocol and so allows us to design our application respecting it. Finally, TIM and WIM depend on Pathway Tools for the main part of the data, having the same language in both applications is a good way to ensure coherence between functionalities.

The object-oriented architecture of TIM is shown in fig.2. All classes corresponding to biological objects in TIM inherit from the generic class `Tim-object`. This class defines the slots and the methods all these objects share. Then, subclasses of `Tim-object` are inspired by the BioPAX format: `PhysicalEntity`, `Pathway`, `SmallMolecule`, `PhysicalComplex` and `Protein` have a direct match in the BioPAX architecture whereas `Reaction` is the equivalent to the `biochemicalReaction` class and `Enzyme` stands for the `Catalysis` class.

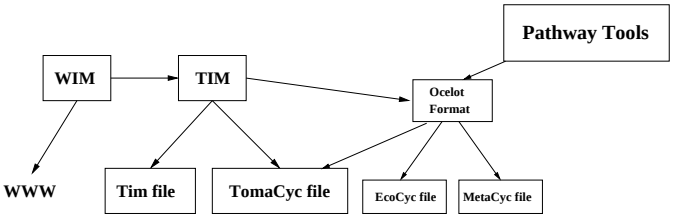


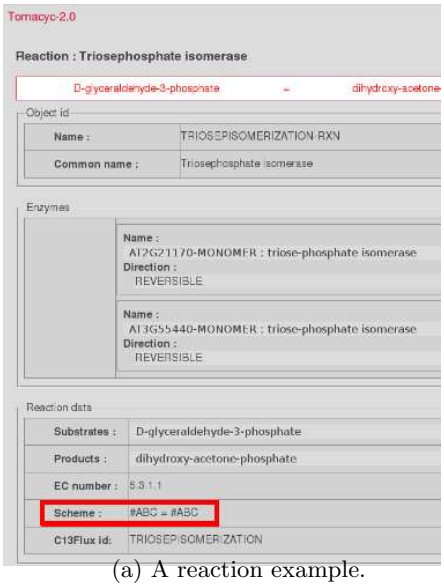
Figure 3: Pathway Tools and TIM-WIM.

A pathway is a set of `Reactions` and takes place in one or more `Compartments`. These reactions are composed of `Compounds` which are the substrates and products and of one or more `Enzyme`. Each `Enzyme` contains a `Protein` or a `PhysicalComplex` which represents the catalysis of a reaction. To be consistent with the Pathway Tools ontology [9], we have a direct link between TIM objects and Pathway Tools frames. `Pathway`, `Reaction`, `SmallMolecule`, `PhysicalComplex` and `Protein` in TIM architecture correspond respectively to the terms `Pathways`, `Reactions`, `Compounds`, `Complexes` and `Proteins` in the Pathway Tools ontology.

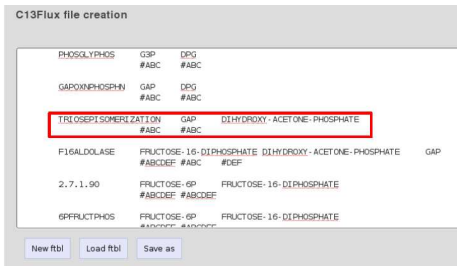
TIM allows users to create and to edit biological objects and to complete these objects with data that are not part of the PGDB format. For example, carbon markers are a slot of the `Reaction` class. To store these supplementary data, the TIM database is separated into two files: the Ocelot file which contains the PGDB from Pathway Tools and a specific file with TIM data (Fig. 3). In addition, these two files can be saved under different versions following PGDB conventions.

We assessed the correctness of the PGDB files generated by TIM by loading them with Pathway Tools. One interesting feature in Pathway Tools is its ability to generate automatically drawings close to what biologists are used to. Each pathways can be drawn from its chain of reaction, and, for each database there is a cellular overview which puts together the graphs of all the pathways in the PGDB. This overview is generated by PathoLogic according to the data and allows a fast way to navigate through the PGDB (see figure 5). We verified that these drawings can be obtained with our new file.

TIM is not designed to be used by itself, but rather with an interface such as the WIM application.



(a) A reaction example.



(b) 13CFlux output.

Figure 4: WIM interface. (a) Red rectangle: the slot for NMR data. (b) Red rectangle: the reaction of the figure a.

4. WIM: THE USER INTERFACE

As we explained before, Pathway Tools does not provide any access to modify the databases from the web server. We have preferred to permit that from a new web interface WIM. With WIM, it is possible to download a PGDB file, to modify it and to add new information that will be stored in TIM files. To build this interface, the web server is implemented with Hunchentoot[5] and the dynamics web pages are built thanks to the CL-WHO[6] library. It is possible to use Hunchentoot with an Apache web server (with mod-lisp) but

[5] `http://weitz.de/hunchentoot/`
[6] `http://weitz.de/cl-who/`

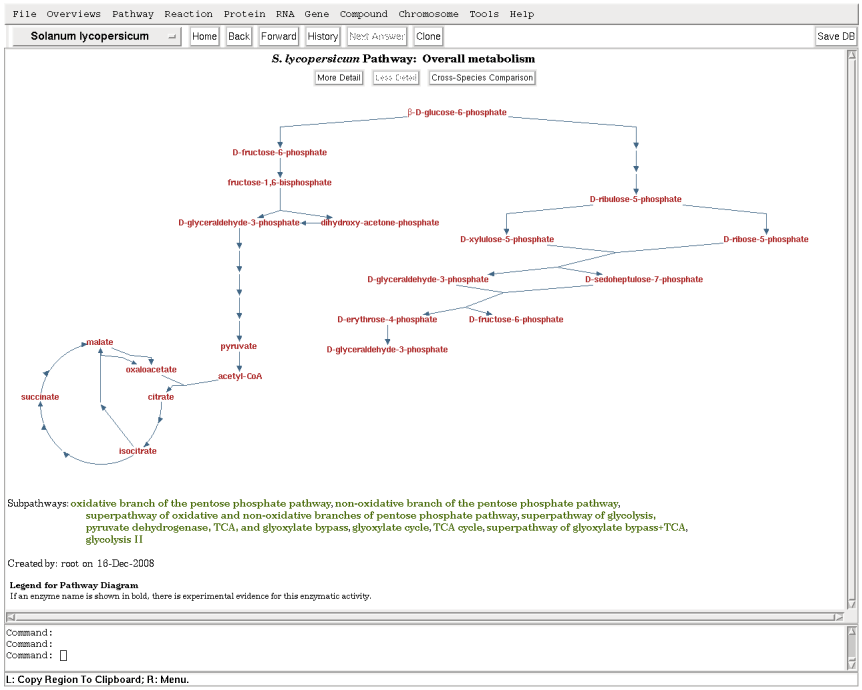


Figure 5: The cellular overview of TomaCyc.

we have chosen to use Hunchentoot directly as web server for developing conveniences. CL-WHO allows generation of html output. The piece of Lisp code in figure 6 gives a rapid example of how to create an html form for the pathway web page.

The current version of WIM provides a form to input NMR data results from experiments. Figure 4.a shows an example of reaction in the WIM interface. In addition to general information, such as a name or enzyme, the user can specify how the carbon markers are given from the substrates to the products during the reaction with the scheme slot (the red rectangle). This slot is used by WIM to write output file for simulation softwares. We have implemented a form in this version to produce 13CFlux files for pathways. This kind of file describes the distribution of carbon atoms through the reactions of pathways, and, writing it manually for large pathways is not easy. Figure 4.b shows an example of 13CFlux file automatically generated by WIM where the reaction of the figure 4.a is written (the red rectangle). The output can be edited by the user in order to change the name of the reaction or anything needed.

The next version will also provide the capability to analyze Metatool[7] format files we frequently use to study the topology of the metabolic networks.

The whole application has been installed and used by a group of biologists to create the first version of the database TomaCyc. Firstly, a part of useful biological data was imported from the AraCyc PGDB [5]. Secondly, these data were modified to match with the biologists' knowledge about the tomato fruit. As a result, there are 14 pathways and more than one hundred enzymatic reactions. These pathways are about central carbon metabolism from several classes: glycolysis, pentose phosphate pathways, biosynthesis and degradation of carbohydrates, TCA cycle, a respiratory chain which were imported from AraCyc and a pathway of transporters for the vacuole compartment.

This database can be accessed with Pathway Tools as a common PGDB. Several visualizations were generated with PathoLogic and no difference appears between TomaCyc and the others PGDB. Figure 5 shows the overview schema obtained when we have analyzed TomaCyc with PathoLogic.

5. CONCLUSION AND FUTURE WORKS

Modelling metabolism networks leads us to define a new application in order to fill the missing information in existing databases. As a result, the development of TIM and WIM gives an appropriate response for the needs of metabolic studies, and especially NMR experiments. With these new applications, we have tools to build new databases, to modify them and to export data to simulation softwares. The TomaCyc PGDB is currently used by biologists to create input files for flux analysis with 13CFLux. We are developing a new extension to allow output to Metatool file for another database called MitoCyc which deals with mitochondria metabolism.

TIM and WIM are important components of the BioFiL framework which contains also a Multi-Agents System BASiL [12] to simulate enzymatic reactions and oxydoreduction phenomena [14]. We hope that we will eventually be able to propose a complete set of tools to work on metabolism networks.

[7] pinguin.biologie.uni-jena.de/bioinformatik/networks

The CL-WHO code:

```
(defmethod html-object-view ((object pathway))
  `(:div
    (:h3 ,(concatenate 'string
       "Pathway : "
       (common-name object)))
    ,(html-field-view object 'object-id)
    ,(html-field-view object 'pthwy-data)
    ,(html-field-view object 'general-data)))

(defmethod html-field-view (object field)
  `(:div
    (:fieldset
     :id "object-field"
     (:legend (str ,(field-namestring field)))
     (:table :id "object-table"
     ,@(loop for slot in
     (field-slots field)
     collect
     (html-slot-view object slot))))
    (:br)))
```

The html output obtained by this code is:

```
<div>
 <h3>Pathway : glycolysis I</h3>
  <div>
    <fieldset id='object-field'>
      <legend>Object id</legend>
        <table id='object-table'>
          <tr id='tr-slot'>
            <td id='td-slot-name'>Name :</td>
            <td id='td-slot-value'>PWY-1042</td>
          </tr>

          <tr id='tr-slot'>
            <td id='td-slot-name'>Common name :</td>
            <td id='td-slot-value'>glycolysis I</td>
          </tr>
        </table>
      </fieldset>
   <br />
  </div>
</div>
```

Figure 6: Example of html code generated with CL-WHO.

6. ACKNOWLEDGEMENTS

This work was supported by the Epigenomic project (Programme d'Epigénomique - Génopole Evry) and a Tomatoflux ANR grant. We would especially thanks Inge Bylemans for her work on the first version of this project.

7. REFERENCES

[1] Hucka et al. The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network ModelsPaper. *Bioinformatics*, 9(4):524–531, 2003.

[2] Karp et al. The EcoCyc Database. *Nucleic Acids Research*, 30:56–58, 2002.

[3] Le Novère et al. Biomodels Database: A Free, Centralized Database of Curated, Published, Quantitative Kinetic Models of Biochemical and Cellular Systems. *Nucleic Acids Res*, 34:D689–D691.

[4] Matthews L. et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, 37, 2009.

[5] Zhang P et al. MetaCyc and AraCyc. metabolic pathway databases for plant research. *Plant Physiology*, 138(1):27–37, 2005.

[6] Luciano JS. Pax of mind for pathway researchers. *Drug Discovery Today*, 10(13):937–42, 2005.

[7] M. Kanehisa and S. Goto. Kegg: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28:27–30, 2000.

[8] P. Karp, S. Paley, and P. Romero. The Pathway Tools Software. *Bioinformatics*, 18:225–32, 2002.

[9] P.D. Karp. An ontology for biological function based on molecular interactions. *Bioinformatics*, 16:269–85, 2000.

[10] P.D. Karp. Pathway databases: a case study in computational symbolic theories. *Science*, 293:2040–2044, 2001.

[11] P.D. Karp, V.K. Chaudhri, and S.M. Paley. A collaborative environment for authoring large knowledge bases. *J Intelligent Information Systems*, 13:155–94, 1999.

[12] Beurton-Aimar M. and Parisey N. An agent-based framework to simulate metabolic processes. *ELW at European Conference on Object-Oriented Programming, 20th edition July 3-7, Nantes (France)*, 2006.

[13] Beurton-Aimar M., Ballet P., Zemirline A., and Mazat J.P. Hybrid system to model biological objects: application to mitochondria organel. In *Proceeding of the spring school on Modelling and simulation of biological processes in the context of genomics. Dieppe (France).*, pages 97–100, May 2003.

[14] N. Parisey, J.P. Mazat, and M. Beurton-Aimar. Mitochondrial Oxydoreduction Simulation using Multi-Agent System. In *European Simulation and Modelling Conference,ESM'2007*, ISBN 978-90-77381-36-6, pages 385–390, Malta, October 22-24 2007.

[15] C. Moore-Kochlacs L. Goldovsky P. Kaipa D. Ahren S. Tsoka N. Darzentas V. Kunin P.D. Karp, C.A. Ouzounis and N. Lopez-Bigas. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 19:6083–89, 2005.

[16] Wiechert W. 13C metabolic flux analysis. *Metabolic Engineering*, 3(3):195–206, 2001.