SBML-Bridge: An XML Pathway Converting Tool

Kanaris Ioannis[1], Chatziioannou Aristotelis[2], Maglogiannis Ilias[1], Kolisis Fragiskos[2]

[1] University of the Aegean, Department of Information and Communication Systems Engineering, Karlovasi Samos
[2] National Hellenic Research Foundation, Institute of Biological Research and Biotechnology, Athens

Abstract

Many universities and research groups worldwide, working on life sciences, produce nowadays large amounts of biological data through the conduction of biological experiments. In initial research efforts found in literature, biological data have been coded in XML format, which is capable of expressing compound properties and relations between them, trying to simulate large metabolic pathways for “in silico” experiments. However, the existing coding schemas produced by the main biological organizations are different, due to the presence of syntactic, semantic and schematic heterogeneities. In this article we present an approach to “bridge” all these incompatibilities in XML based representation of biological data and thus to make the information easily accessible for further analysis and usage by other scientists. Two case studies of biological data conversion are presented. SBML - Systems Biology Markup Language - was selected as a target format and CellML - Cell Markup Language - and KGML - KEGG Markup Language - as source formats to be converted. The paper presents all the technical details of the proposed approach along with the corresponding findings and the appropriate discussion.

Keywords: Biomedical Databases**,** Bioinformatics, XML, Metabolic Pathways, SBML, CellML, KGML.

1. Introduction

Bioinformatics today is gradually evolving from a supporting action to the experimentalist to an autonomous scientific field with the potential for many applications, innovations and capabilities, thus revolutionizing biological research. Ontologies used in information science to categorize and characterize all kinds of data using semantics, gave the biologists a very powerful tool to aid them in representing all kinds of knowledge emerging from their laboratory research. Especially XML - eXtensible Markup Language (http://www.w3.org/XML/) - proved to be much more expressive than traditional structured database languages, giving the ability not only to semantically represent the knowledge with respect to different species but also the relations and linkages among them. This led to building complex models that describe various levels of biological processes inside organisms.

However, there seems to be a gap in the cooperation among researchers and as a consequence, a lack of standards with respect to the task of finding a unique semantic way of representing data. As a result, a multitude of biological databases offer information in so many different ways, following their own proprietary formats for representation, that cannot be easily used, combined or cross-checked. In order to overcome these obstacles, the task of integration amid various tools and platforms, using distributed databases and markup languages to make models easily comparable and even cooperative, is becoming imperative.

Trying to tackle this incompatibility problem, incurs reduction of the complexity of the problem by emphasizing in some factors. First of all, one of the existing formats is chosen to work upon, based on the needs of the researchers. Then, other existing databases are examined and with respect to the amount and quality of their content, they are transformed to the appropriate format. This proves to be necessary as no converter can be used, capable of handling all these different file formats. In this work, SBML - Systems Biology Markup Language - was selected as a target format and CellML - Cell Markup Language - and KGML - KEGG Markup Language - as source formats to be converted. From these three different markup languages, CellML and SBML describe models used for simulation purposes whilst KGML biological models are visually descriptive, acting as static representations of molecular pathways. SBML format was selected due to its ability to represent biochemical models with simulation capabilities and the continuing effort of many researchers to support it by implementing new tools and models based on it. On the other hand, CellML format, despite its limited number of models provides simulation-related information such as reaction kinetic laws, while KGML was selected mostly due to its large database of models and entities.

The rest of the paper is structured as follows: section 2 refers to previous approaches and related work and section 3 describes the basic layout of each format focusing on their heterogeneities followed by a catalog of semantically correlated features among them. In section 4, the proposed approach is described and finally, section 5 discusses the potentials of this research and concludes the paper.

2. Related Work

Initially, publicly available life science databases were built based on traditional SQL systems. Stemming from the need for data exchange among different scientific organizations and research fields, many platforms begun to support flat file navigation approaches. Such systems are for example SRS [Etzold et. al. (1996)] and DBGET/LinkDB [Fujibuchi et. al. (1998)], working by accessing HTML flat files. Both the aforementioned approaches were not able to solve the problem mostly due to the complexity of biological data that demands a predefined and universally agreed data structure, based on a general schema. As a consequence, XML technologies begun to evolve, providing the descriptive power and machine readability that were missing. Many platforms based on this technology evolved such as the EMBL platform [Wang et. al. (2002)], the DDBJ [Miyazaki (2003)], Reactome [Joshi-Tope et. al. (2005)] and more including the aforementioned CellML. Although XML querying is more flexible and fast, incompatibilities among different formats still exist, demanding conversions that provide data exchange through all available databases.

Focusing on this particular compatibility problem, two tools have been developed until now: CellML2SBML [Schilstra (2004)] and KEGG2SBML [Funahashi et al. (2004)]. These applications apply conversions from these formats to SBML as their names indicate and can be found in the official site of SBML format (http://www.sbml.org). The CellML converter is based on XSL Transformations using four XSL files applied contiguously to the source document, while the KEGG converter uses string and XML parsing functions implemented in Perl. Their command line interfaces and the fact that are Linux implementations , make them non user-friendly especially to traditional “wet-lab” biologists who are not experienced in working on various sophisticated computer platforms. In this paper a unified converter for both formats is introduced, which to our best knowledge is the first application performing this conversion task for both significant from the biological point of view formats simultaneously. In addition the proposed tool is open and expandable, enabling its further enrichment with other converted models (CellDesignerSBML, BioPAX ontology web language, etc.), in contrast with the closed CellML2SBML and KEGG2SBML tools. Finally the described tool is implemented in a user friendly graphical environment based on windows platform improving significantly the usability of the aforementioned tools.

3. Semantic Heterogeneities in Biological XML Documents

To better understand the need of interconverting XML documents among various biological databases, it is imperative to identify the basis of their differences and in which ways it affects the description of biochemical models. As it may be commonly known, documents such as XML are based on a schema; a file that describes the basic guidelines of how the document is constructed. The schema defines namespaces, structure and semantics of document elements. The incompatibilities among the XML formats have their source mainly to the schema, and its descriptive power. The three selected markup languages have similarities in general structure but they have also many differences mostly in their namespaces and attribute values.

3.1 SBML Format

SBML is a ‘free, open, XML-based format for representing biochemical reaction networks’ [Hucka et. al. (2003)] [Finney et. al. (2003)]. It’s a standard which is software-independent that guarantees interoperability among multiple tools that are used in biological model simulations. This language is used to represent formally various models common in many areas of computational biology research, including cell signaling pathways, metabolic pathways, gene regulation, and others. Since its creation, developers have implemented richer versions of it, reaching level 2 as a stable version, with level 3 under development now.

An SBML document describes a single Model inside an SBML 'container'. The model describes a system that contains zero or more of the following core components (as described in the official site guide [Schilstra et. al. 2003]):

- **Species** – an entity that takes part in reactions
- **Compartment** – a spatially extended, well-stirred container for species
- **Reaction** – a transition in which one or more reactant species are transformed into one or more product species.

A model must contain at least one species and one compartment if it contains one or more reactions. A compartment has a size (which may be its volume, but in Level 2 may also be, for instance, a surface area), and may have an ‘outside’: another compartment that surrounds it. A reaction has a list of reactants, and a list of products. In Level 2, there may also be a list of modifiers. A kinetic law equation specifies the rate at which the reactants are transformed into products, as well as the effect of modifier species on this rate.

3.2 KGML Format

The KGML format is mostly an exchange format of graph objects made by the KEGG institute. Pathway models are hand crafted and updated and give a good view of many protein and chemical reactions. The pathway maps give a graphical representation of networks of interacting molecules responsible for specific cellular functions and are organized in two types:

- **Reference pathways** which are manually drawn and
- **Organism - specific pathways** which are computationally generated based on reference pathways.

KGML documents are not standalone files because they depend on other database files for information about the model. These documents describe the pathway network in the form of a graph with **entry elements** as its nodes and **relations** and **reaction elements** as its edges. Every chemical compound, gene or protein that takes part in the model is defined as entry, while protein interactions and map links to other pathways are defined as relations. Finally reaction elements are referred on how compounds and proteins interact with each other in various ways to produce other elements. Although KGML models are very descriptive in terms of visual representation, the missing of kinetic laws that describe the chemical reactions makes it difficult to be used in simulations.

3.3 CellML format

The process of translating a biological process into CellML document capable of simulating this procedure in a computer system follows a specific template. This template describes the different sections required [Cuellar et. al. (2003)]. In detail they are:

- **Metadata** – Usually in RDF format, they contain information about the model's name and Author, creation date etc.
- **Units** – Here, all different units of time, volume and quantity are defined.
- **Components, Variables and Equations** – Always the model begins with a component, named environment and it contains all global variables such as time. Additionally components also include kinetic laws giving information on how every compound takes part in reactions in the unit of time.
- **Interfaces, Groups and Connections -** Each variable can have a public and/or private interface, with a value of either out, in or none. Interfaces describe the direction in which the value of a variable is passed between components, i.e., variable export, import, or no movement.

3.4 Heterogeneities summary

Following the description of the aforementioned formats it becomes obvious that each one of them represents data through a different perspective. Incompatibility mostly occurs due to schema conflicts, which are caused by different conceptualization of biological entities. Element and of course attribute conflicts are caused mainly by the use of different terms for the same concepts (synonyms) and the same terms for different ones (homonyms). For example, the words “protein” and “enzyme” are used to describe the same thing in different languages, but each format has different ID values to define the same compound. Finally, value conflicts occur when for example a specific compound or reaction has more than one names and it's difficult to verify the integrity of correlation between transformed objects. Generally, the differences in schema and element conflicts are represented in Table 1 in descending order from general (e.g. model) to more specific elements such as simulation-related parts (e.g. function definition). Semantically correlated features were placed in the same line indicating differences in term use.

Table 1. *Correlated features*

SBML	KGML	CellML
Sbml model	Pathway model/map	CellML model
Compartment		Group
Species	Entry	Component
Reaction	Relation	Component (reaction)
	Reaction	
Unit definition	-	Unit
Parameter		
Rule Event Function definition	-	Variable Equation

4. Proposed Approach

The main goal in the proposed approach is to create an easy to use graphical environment that can be easily improved and handled by users (Figure 4). Almost all needed conversion options together with an XML tree view structure of each document reside in the main format. Given the fact that the files to be converted are always written in XML, the best way to access and change their form is by XSL Transformations. A tool was developed using Microsoft's .NET Framework 2.0 and is mostly designed to run on a Windows platform. A high level view on the tool's architecture is shown in Figure 1. The conversion method followed in each case is described in the following sections.

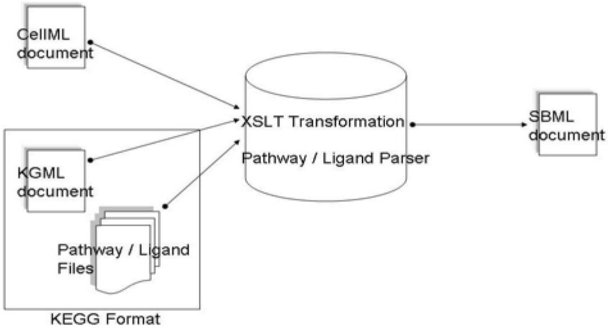


Figure 1. *Architecture of the proposed tool*

4.1 CellML Conversion

In the case of CellML files, all available information about the model, including compounds, reactions and even kinetic laws are included in a single XML file. So, with a proper XSL Transformation the desired features can be accessed and transformed giving a full model capable of simulating. For this task the Microsoft

MSXML 4.0 engine was used applying the four XSLT files provided by CellML2SBML tool (http://sbml.org/software/cellml2sbml/). These four style sheets applied contiguously on the source CellML file (Figure 2) can transform a vast variety of models and were left intact. In case of a new release, the end user can always replace the four files and ensure compatibility with the latest format.

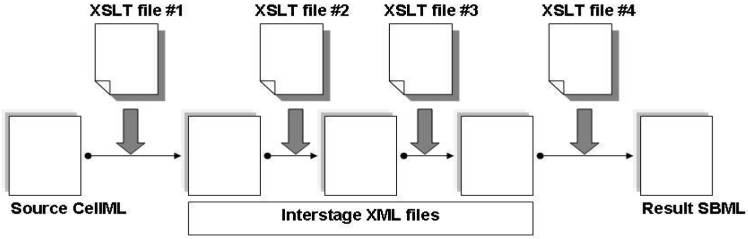


Figure 2. *CellML Phases of transformation*

4.2 KGML Conversion

The KEGG pathway models are not the same. The KEGG Markup Language or KGML, is based on the need of a graphical representation of each pathway. As a consequence, the mere use of XSLT is inadequate. An approach similar to CellML's can only solve schema and some element conflicts. The rest of element and all the value conflicts occur because of the internal ID naming of every compound and reaction inside the KEGG organization. Since KEGG database is much older than KGML format, the information regarding names and element values and their connection to IDs, are scattered in many flat files (plain text or HTML format). These files are available from the site of KEGG database and they form the so called Pathway and Ligand databases. Every pathway model, after its transformation to its SBML analogue must be processed from the DOM engine in cooperation with a Pathway/Ligand parser to apply proper names to every element of the file so it can be easily read by scientists who want to examine the model (Figure 3). In detail the steps followed during the conversion of KGML files are:

- **XSL Transformation** – Using a single style sheet document the elements of the KGML format are transformed to their equivalent ones in SBML. More specifically, each 'entry element' becomes a 'species element' and reaction together with relation element forms the SBML reaction element. The reaction element in KGML defines the reactants and products while modifiers are defined in relation elements that link to specific enzymes in the entry group.

- **Pathway/Ligand Parsing** – After the first step the model is ready for importing to any SBML-compatible tool for further editing. But the names of all compounds and reactions are still in KEGG ID format. To make the model easily understood by humans it's necessary to fill in the proper biological names. This is done by accessing the pathway files that are correlated with the file in question depending on the pathway ID and the organism it refers to. Three files, with the same name as the KGML but with extensions .cpd, .rn and .enz/.gene are responsible for assigning every ID value with the corresponding name. If some of the elements still remain unidentified after pathway parsing the ligand files are parsed too. Five files named compound, enzyme, glycan, reaction and map_title contain almost all the information available from KEGG. Ligand files are parsed only if needed, because of their large sizes which make name parsing relatively slow.

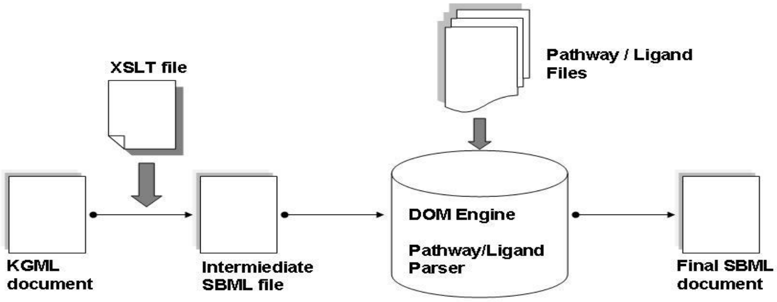
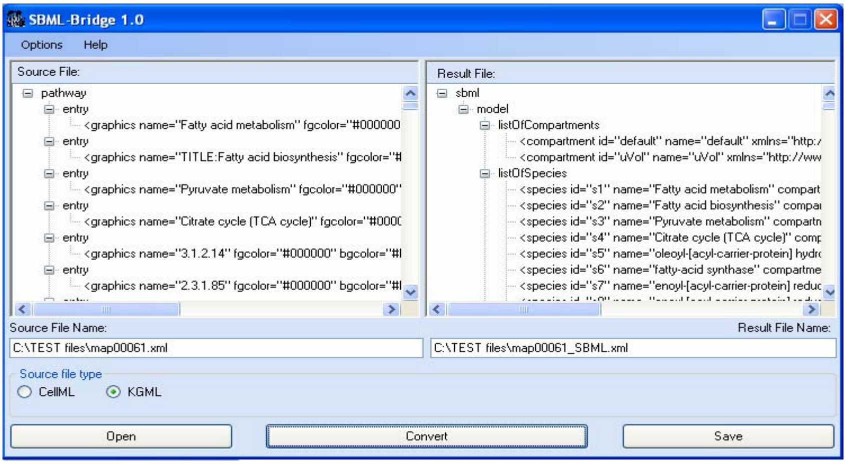


Figure 3. *KGML Phases of transformation*

There are two kinds of pathways: Reference and organism-specific models. Files that belong to the first category are richer; in most cases they have all the information regarding element names, linkages and enzymes catalyzing the respective reactions. On the other hand, when a pathway relates to an organism is always smaller, includes only the parts of the pathway that are active (or known) in each organism and cites the genes encoding the enzymes of the depicted reactions. The main problem that biologists face in transforming organism specific pathways is that the resulting models define reaction networks among metabolites, which according to the specie depicted may present a radically different structure, according to the dispensable set of genes, each specie has adopted. As a result, the derived network of a specie, can be much sparser in terms of edges (reactions) than the reference pathway, especially when the specie depicted is relatively low in the developmental climax. The presented tool manages to resolve this problem using the Ligand library. In order to keep the models well interconnected, the initial transformation is applied to the reference pathway. After creating the target SBML model and completing the appropriate names, the end user can provide the organism he is interested in and the tool can find all the known related genes that correlate to the enzymes of the reactions and connect them appropriately, giving a more accurate view of the biological procedure which is very important in biological research. For the same reason, any available information about pathways linking somehow to the model in process, is preserved giving biologists a better view of the interaction among pathways. This is a novel feature, supported for the first time, by a conversion tool for KEGG pathway models. At present, the tool supports only local data files parsing. An upgrade task for the application will also be the support of information parsing, directly from the KEGG web site, through the use of the dbget platform.

The tool was designed to export files in pure SBML, providing maximum compatibility with any SBML supporting tool. Since one of the most commonly used tools for biochemical model representation and simulation is CellDesigner (http://celldesigner.org/), in the case of KGML transformation a second XSLT template was implemented making the tool capable of supporting extra namespaces and tags for this tool. The proprietary format of CellDesigner exploits standard SBML by expanding it through fields that contain useful information about graphical representation, enable support of representation of more biochemical entities such as types of species (genes, proteins, receptors, etc.) and reactions, increasing the expressive power of models.



Source File:

```
⊟ pathway
    ⊟ entry
        <graphics name="Fatty acid metabolism" fgcolor="#000000
    ⊟ entry
        <graphics name="TITLE:Fatty acid biosynthesis" fgcolor="#
    ⊟ entry
        <graphics name="Pyruvate metabolism" fgcolor="#000000"
    ⊟ entry
        <graphics name="Citrate cycle (TCA cycle)" fgcolor="#0000
    ⊟ entry
        <graphics name="3.1.2.14" fgcolor="#000000" bgcolor="#I
    ⊟ entry
        <graphics name="2.3.1.85" fgcolor="#000000" bgcolor="#I
```

Result File:

```
⊟ sbml
    ⊟ model
        ⊟ listOfCompartments
            <compartment id="default" name="default" xmlns="http:/
            <compartment id="uVol" name="uVol" xmlns="http://ww
        ⊟ listOfSpecies
            <species id="s1" name="Fatty acid metabolism" compart
            <species id="s2" name="Fatty acid biosynthesis" compai
            <species id="s3" name="Pyruvate metabolism" compartn
            <species id="s4" name="Citrate cycle (TCA cycle)" comp
            <species id="s5" name="oleoyl-[acyl-carrier-protein] hydrc
            <species id="s6" name="fatty-acid synthase" compartme
            <species id="s7" name="enoyl-[acyl-carrier-protein] reduc
```

Source File Name: C:\TEST files\map00061.xml

Result File Name: C:\TEST files\map00061_SBML.xml

Source file type
○ CellML ● KGML

Figure 4. *A snapshot of the tool's graphical environment*

5. Discussion and Future Work

As the number of various biological databases and records grows exponentially, data-integration platforms and tools will be developed to facilitate interconnection among all those different format types. The SBML format seems to be the most promising, mostly due to large amount of tools for modeling and simulation, already implemented in the specific format. The SBML-Bridge tool supplies a user-friendly graphical environment for converting models from two large databases. Its ability to produce files in both pure SBML format as well as Celldesigner annotated files, giving semantically well formed models ready to be processed, makes it a valuable tool in the effort to finely depict and simulate biochemical or molecular pathway models. Especially in KGML format it manages to link related pathways as compounds to each model giving a better conceptualization. Also the proposed tool has implemented gene – enzyme correlation support, a novel element that enables reverse engineering methodologies offering significant insight with respect to the study of how the molecular pathways response is orchestrated by genes. Future work concerns efforts to further enrich the pathway models by adding kinetic laws to KGML model files. Another issue that is under development is the ability to add whole networks of metabolic pathways combining multiple models until full scale cell simulation is possible. The specific feature will enable scientists to examine consequences of environmental changes or even medical treatments to the cell's life cycle.

References

Akira Funahashi, et. al. Converting the KEGG Pathway Database to SBML. Presented at ICSB 2004, Heidelberg, Germany, October 9-13, 2004.

Catherine Lloyd. Best Practices when writing a CellML Model. Available from: http://www.cellml.org/tutorial/best_practice/html2pdf

Cuellar, et. al. CM. CellML Specification 1.1. Draft—30 September 2003. 2003; Available from: http://www.cellml.org/public/specification/index.html.

Finney A, et. al. Systems Biology Markup Language (SBML) Level 2: Structures and Facilities for Model Definitions. 2003; Available from: http://sbml.org/specifications/sbml-level-2/version-1/html/sbml-level-2.html.

Goto S., et. al. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Research, 30:402-404 (2002). [PMID:11752349]

Joshi-Tope G, et. al. 2005. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005 Jan 1;33 Database Issue:D428-32. PMID: 15608231

L. Wang et. al. XEMBL: Distributing EMBL data in XML format. Bioinformatics, vol. 18, no. 8, pp. 1139–1140, 2002

M. Hucka, et. al. The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models. Bioinformatics 2003, 19(4): 524-531.

Maria J. Schilstra, et. al. A Practical Introduction to SBML. Available from: http://sbml.org/documents/presentations/ebi-2003/A%20PRACTICAL%20INTRODUCTION%20TO%20SBML.htm

S. Miyazaki, et. al. DNA data bank of Japan (DDBJ) in XML. Nucleic Acids Research, vol. 31, no. 1, pp. 13–16, 2003.

Schilstra, M.J., (2004).Conversion of CellML 1.1 into SBML L2v1 [online] http://sbml.org/software/cellml2sbml/conversion_of_CellML2SBML.pdf

T. Etzold, et. al. SRS: Information retrieval system for molecular biology data banks. Methods Enzymol., vol. 226, pp. 114–128, 1996

W. Fujibuchi, et. al. DBGET/LinkDB: An integrated database retrieval system. In Proc. Pacific Symp. Biocomputing, 1998, pp. 681–692.