An Exchange Format for GenMAPP Biological Pathway Maps

by
Lynn M. Ferrante
San Francisco, California
August, 2004

A thesis submitted to the faculty of San Francisco State University
in partial fulfillment of the requirements for the degree
Master of Arts in Biology: Cell and Molecular

Thesis Committee:
Dr. Michael A. Goldman
Professor of Biology

Dr. Bruce Conklin
Associate Professor of Medicine and Cellular and Molecular Pharmacology
University of California, San Francisco

Dr. Maureen Whalen
Professor of Biology

Abstract:
A biological pathway map is a graphical representation of a known biological pathway. Each source of pathway map data has its own way of storing, analyzing, displaying, and exporting pathway data, and there is no standard for map exchange. Valuable opportunities for data interpretation are lost. The Gene Map Pathway Profiler (GenMAPP) is a freely available software package that helps visualize pathways and genome scale expression data. GenMAPP users must build their own pathway maps or use a small set of provided maps. I investigated a pathway map exchange format for GenMAPP as follows. First, I translated existing metabolic pathway maps to GenMAPP. Next I assessed the proposed BioPAX Pathways Exchange format for exchanging GenMAPP maps. Finally, I devised a pathway exchange format for GenMAPP.

ACKNOWLEDGEMENTS

This work was carried out at the Gladstone Institute of Cardiovascular Disease at the University of California, San Francisco, under the guidance of Dr. Bruce Conklin. I sincerely thank Dr. Conklin for his support of this work. I am also grateful to the many other members of the lab that assisted me in numerous ways, including Dr. Alex Pico who kindly reviewed this document.

This work would never have been possible without the support of many professors at San Francisco State University. I would like to thank my thesis committee, Dr. Michael Goldman and Dr. Maureen Whalen, for their unwavering support and confidence. I am also indebted to Drs. Maureen Whalen, R. L. Bernstein, and Leticia Marquez-Magana for their inspirational teaching of biology and bioinformatics. I am extremely grateful to Dr. John Stubbs who encouraged me to continue my education, supported me through the National Science Foundation Graduate K-12 Fellowship Program, and was instrumental in placing me at Gladstone.

I also wish to thank my family for their love and encouragement throughout my graduate education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

I. INTRODUCTION

A biological pathway map is a graphical representation of a known biochemical pathway that is often essential in understanding the basic biology of an organism. Biological pathway maps have been elucidated after many years of experimental research, yet it is often difficult for current researchers to obtain this information in a digitally accessible format. Recently developed high throughput experimental methods such as DNA microarrays and two hybrid screens have created huge amounts of genome-scale data that are difficult to interpret unless put in the context of known biological pathways (Wittig and de Beuckelaer, 2001). In attempts to meet this need several different sources of pathway information have been made available on the internet including KEGG (Kanehisa *et al*., 2004) (Kanehisa, 1997), BioCarta (Biocarta, 2004), WIT (Ergo-Light, 2004), BIND (Bader and Hogue, 2003), BioCyc (BioCyc, 2004), Genome KnowledgeBase (Joshi-Tope *et al.*, 2003) and GenMAPP (Dahlquist *et al.*, 2002). Unfortunately, the exchange of pathway data between these resources is hampered by the fact there are multiple pathway formats that are difficult to translate and collectively resemble the “Tower of Babel”. Pathway information is often available in graphical form and referred to as a pathway *map*. The data behind the graphical display of the pathway are sometimes stored in databases but often have only graphical representation. Additional information may be linked to pathway maps, such as relevant literature citations and detailed genomic and biochemical information, and other graphics.

Each pathway database or pathway map source has its own way of storing, analyzing, graphically displaying, and exporting pathway data (Schaefer, 2004). Frequently, the biological researcher must use multiple data analysis programs and sources of pathway data and maps to successfully process and analyze data. While there are now a variety of sources for pathway data, the challenge is to share and integrate pathway information among them. Currently, there is no standard for sharing these data. Biologists are often confused, and valuable opportunities for interpretation of data are lost.

The Gene Map Pathway Profiler (GenMAPP) is a freely available software package developed by the nonprofit, academically based organization GenMAPP.org (http://www.genmapp.org). GenMAPP helps visualize genomic scale expression data in the context of biological processes depicted graphically as maps. Viewing expression data in this context helps bring to light changes in gene expression in a particular pathway or relationships between genes that have changed their expression under specific conditions. GenMAPP currently has a base of approximately 8,000 registered users who primarily use the program for DNA microarray studies.

GenMAPP users must either build their own pathway maps with the GenMAPP drafting tool or use a small set of pathway maps provided by the GenMAPP staff and other contributors. Additional biological pathways are available on various academic and other internet sites and are sometimes displayed maps with map-viewer software. GenMAPP and other software packages cannot use these pathway maps since the format is incompatible with their software. The underlying problem is that there is no common data exchange format for pathway data in the scientific community. If there were a single standard format shared by the multiple sources of pathway data for import and export of data, it would make pathway information far more accessible and useful for research.

Pathway maps may contain much more than pathway data. Researchers can customize their maps to reflect the focus of their investigation, using different versions of gene product names, adding explanatory text, references, expression data, SNPs (single nucleotide polymorphisms), and graphics that assist a researcher in explaining their work. It is often beneficial for a researcher to share their pathway maps with others or to integrate a pathway map from an external source into their analysis. This can also be quite difficult with no standard pathway map exchange format.

My research involved investigating a data exchange format for pathway maps that would allow the exchange of maps between GenMAPP and other groups and individuals in the biologic research community. My goal was to provide a broad additional source of pathway information for the GenMAPP community at large that would allow users to export their completed maps in a format suitable for exchange with other software programs. The three major sections of this research are as follows:

Determine the feasibility of importing existing external metabolic pathway maps into GenMAPP. I selected the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic maps as my external pathway map source due to the recent availability in 2003 of hundreds of these maps in a version of XML[a] created specifically for this purpose by Genome Net[b]. XML files representing the KEGG catalog of metabolic maps were analyzed, parsed, and translated into a format suitable for GenMAPP display. This component was much more difficult than expected due to a lack of full documentation of the structure of the XML files and a lack of rigorous definition of different elements within those files. Close to 700 KEGG metabolic maps were converted to GenMAPP. There were difficulties in layout and not all objects and relationships present in the KEGG hand-drawn maps were present in the XML files. This information is therefore not present in the resulting GenMAPP maps. These converted maps are now available to GenMAPP users in GenMAPP version 2.

Assess the proposed BioPAX Pathways Exchange format as a future standard data exchange format for GenMAPP maps. BioPAX (BioPAX Workgroup, 2004), the Biopathways Data Exchange workgroup, is developing a common exchange format for biological pathway data to promote collaboration and accessibility**,** incorporating key elements from a range of pathway databases. The standard (version 1 level 1) exists as an ontology, which, unlike an XML Document Type Definition (DTD)[c] or schema[d], allows for rigorous definition of terms and nuances in semantics.

I found that the BioPAX standard specifically addresses pathway data, but not all data that might be on a pathway map. The components of a pathways exchange format such as BioPAX do not completely overlap with a map pathways exchange format as proposed by GenMAPP. The content of GenMAPP pathway maps is broader than the pathway data covered by BioPAX. One of the goals of this study was to investigate whether a subset of the proposed exchange format was appropriate for GenMAPP, either now or in the future. As expected, I found that while GenMAPP could use parts of the BioPAX ontology, there were many extra map objects that did not fit into BioPAX. Additionally, since GenMAPP maps implement interactions between objects as independent arrow and line objects, there was no overlap with BioPAX in this area.

Devise a pathway map exchange format for GenMAPP. The GenMAPP pathway map exchange format should include both pathway information and additional information on the map. Machine processing of this format should be able to distinguish between the two, since only the pathway information is necessary for data exchange with other pathway databases. A recommendation for GenMAPP is made in phases. For the initial phase, a version of XML entitled GenMAPP Markup Language (GMML) is used, and a DTD defined. GMML can represent both pathways and other map data. This implementation provides XML import and export of pathway maps and can be accomplished in a short time period. After GenMAPP implements an interaction type, the second phase of the recommendation will involve using a more rigorous definition of GenMAPP pathway map terms through the BioPAX ontology, and a closer association with the BioPAX standard. Most pathway information, including gene products and their relationship to other gene products, could be exported in a format compatible with BioPAX. All other map information would remain outside the realm of the BioPAX standard.

[a] XML (Extensible Markup Language) is a metalanguage designed to interchange structured data and improve Web functionality. The full XML specification is online at http://www.w3.org/TR/REC-xml

[b] GenomeNet is a Japanese network of database and computational services for genome research and related research areas in molecular and cellular biology. (http://www.genome.ad.jp/about_genomenet/)

[c] An XML Document Type Definition describes the markup and other components available in a specific type of document.

[d] An XML schema is similar in function to a DTD but allows for formal data typing and name spaces.

II MATERIALS AND METHODS

External Map Conversion. KGML files representing 120 KEGG version 0.2 reference maps and 692 organism specific maps were downloaded from ftp://ftp.genome.ad.jp/pub/kegg/xml/KGML_v0.2 /map. In addition, the KEGG Markup Language v0.2 Draft Specification and the DTD for KGML v0.2, available at http://www.genome.ad.jp/kegg/KGML/ KGML_v0.2/ and included in this document as appendices A and B, were used as reference documents. I developed two software programs based on these materials. These programs are freely available at www.GenMAPP.org. The first program, the KGML GenMAPP Reference Map Converter, converts KGML reference map files to GenMAPP pathway maps by parsing through the KGML for appropriate objects and accessing the KEGG ligand database (Goto *et al.,* 1998) for additional information.

The second program, the KGML GenMAPP Organism Map Converter uses the KGML specific to one of the organisms shown in Table 1 and the appropriate GenMAPP reference map (converted by the first program) and creates an organism specific map for each pathway. This program also accesses additional information from the KEGG ligand database. I programmatically converted 120 KEGG reference maps to GenMAPP and created 692 organism specific pathway maps for nine species. I used the Perl language (version 5.6.1) to create both programs. The programs use the process displayed in Figure 1.

ORGANISM	NUMBER OF METABOLIC MAPS CONVERTED	REASON FOR CONVERSION
Reference maps	120	Needed for all organisms
Homo sapiens	98	Model organism
Mus musculus	92	Model organism
Rattus norvegicus	88	Model organism
Caenorhabditis elegans	87	Model organism
Drosophila melanogaster	95	Model organism
Danio rerio	18	Model organism
Saccharomyces cerevisiae	83	Model organism
Pyrococcus furiosus	62	Requested by University of California, Santa Cruz, Dept. of Biomolecular Engineering
Plasmodium falciparum	69	Requested by University of California, San Francisco, Dept. of Biochemistry, Functional genomics of *Plasmodium falciparum* project
TOTAL excluding reference maps	692	

Table 1: KEGG metabolic maps converted to GenMAPP. KEGG contains approximately 200 hand-drawn reference maps, but only a portion of these were available in KGML in 2003. All reference maps available in KGML were converted to GenMAPP. GenMAPP Version 2 has model organism databases for the first seven species listed. Other species were added by request.

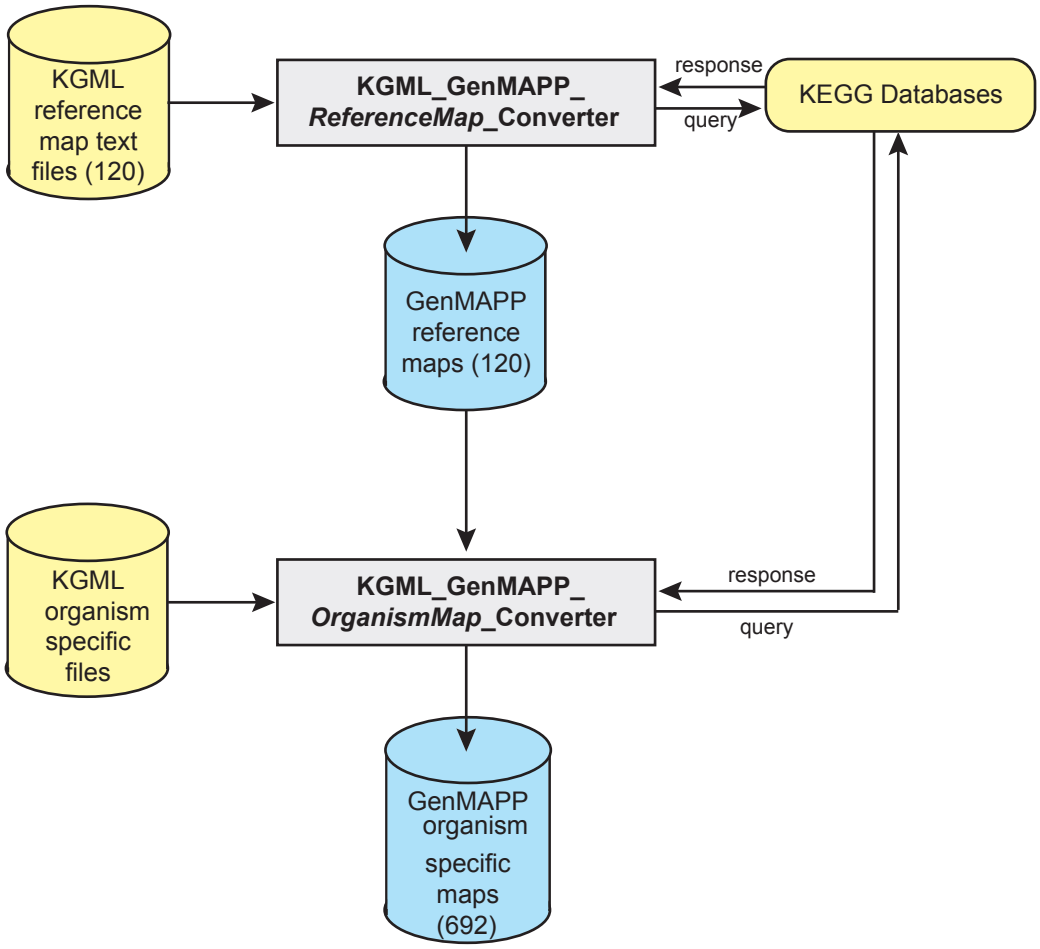
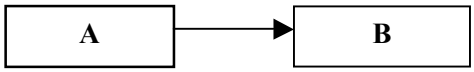


Figure 1: Overview of conversion process from KEGG to GenMAPP pathway maps. KGML reference map files were downloaded from the KEGG site and used as input into the reference map conversion program. This program parsed through the KGML to find the appropriate data and links to the KEGG databases, then queried the databases at KEGG for additional information. The final product was a GenMAPP reference map. These reference maps were used as a basis for the organism specific map conversion. KGML files for each species were downloaded from KEGG and used to update the reference maps for each species. Additional gene links were taken from the KGML files and the KEGG databases searched.

III. RESULTS AND DISCUSSION

GENMAPP PATHWAY MAPS

Each GenMAPP pathway map is currently stored as a relational database. The map format for pathway maps is vector based. A map consists of objects, each of which has a row in a relational table. Lines or arrows may graphically connect objects (or groups of objects):



Since GenMAPP was initially conceived as a drawing program (such as Adobe Illustrator®) it does not store relationships between objects. It does independently store the start and end coordinates of arrows and lines. These arrows or lines graphically represent relationships between objects on the displayed map. The arrows or lines may start and end outside of, within, or on the edge of the objects they are connecting. Objects on the map do not have intrinsic relationships, but to the human eye, they appear related since other objects (arrows and lines) connect them.

Each pathway map consists of two relational database tables, *Info* and *Object*. Info contains generalized information about the map, such as title, author, maintainer, and size. The Object table contains one row for each object on the map. Objects may be genes, labels, arrows, lines, or other objects. Each object has attributes, including x-y coordinates, width and height, color, and rotation. GenMAPP software displays the map data for these two tables on the GenMAPP drawing board.

The gene object is the backbone of GenMAPP. Additional information about each gene may be stored in the GenMAPP gene database and displayed in another window, called the backpage, when the gene is right clicked. DNA-microarray expression data, when available, is also linked to genes through other GenMAPP relational tables, and a user-defined color scheme can be used to differentiate changes in gene expression on the map.

Importing pathway maps from other sources to GenMAPP is possible but complicated. GenMAPP has its own internal representation of pathway data as described above. In GenMAPP versions 1.0 and 2.0, a pathway map is stored as a Microsoft Access database file, and each pathway is a separate Access database. To import a map into GenMAPP, the map data must be in a specific GenMAPP format readable by Access and must have the information necessary for GenMAPP to recognize the file as a GenMAPP pathway map. This involves placing specific information into the two Access tables, *Info* and *Objects*.

To import any maps, a software program must:

- Have access to an external source of pathway data that is program readable, e.g. XML
- Be able to interpret and relate the external pathway data to GenMAPP internal map structure
- Access additional information on genes or other map objects that may not be visible on the original pathway map but is available
- Write the translated data to a Microsoft Access database in GenMAPP specific format

GenMAPP 2.0 exports data in three graphical formats (bmp, PDF, and jpg) and html. These formats are suitable for graphical pathway display but not for exchanging the actual pathway data or maps so that the relationships can best be preserved in a database or a text file (such as XML).

KYOTO ENCYCLOPEDIA OF GENES AND GENOMES PATHWAY MAPS

KEGG was a good potential source of pathway information for GenMAPP due to the volume of maps available to the public, and the availability of these maps in XML. For this project, I converted all KEGG metabolic reference maps to the existing GenMAPP format, and then computationally produced organism specific maps for nine species.

KEGG maintains two types of pathways:

- Manually drawn *reference pathways* (236) (Kanehisa *et al.,* 2003) based on Enzyme Commission (EC) numbers[e] .
- Computationally generated, *organism specific pathways* based on the reference pathways and additional information for each species (currently representing 181 species[f]) resulting in over 13,000 maps.

Until 2003, these maps were unavailable to GenMAPP since they were maintained only in a graphical image format.

[e] Enzyme Commission (EC) numbers are a numbering scheme for enzymes based on the chemical reactions each catalyzes. http://www.chem.qmul.ac.uk/iubmb/enzyme/

[f] http://www.kegg.com/kegg/kegg1.html

KEGG gene, compound, and enzyme databases link to the KEGG graphical pathway maps by clicking on compounds or enzymes. Additional information about the object appears on a new browser page. KEGG (KGML v0.2) maintains the relationship between enzymes in each map as a pair of nodes (enzymes) and an edge that is a common compound between enzymes. There are close to 15,000 enzyme-compound-enzyme relationships in the KEGG dataset.

To investigate the issues involved in sharing pathway data, I attempted to bring the KEGG metabolic pathway maps into GenMAPP for use by the GenMAPP community in analyzing microarray and other data. GenMAPP version 1.0 had several maps manually copied from the KEGG catalog by GenMAPP staff. While the maps were useful, GenMAPP staff considered it too labor intensive to copy any more maps by hand, so an automated process was desired.

KEGG Markup Language and the KEGG DTD.

In late 2003, KEGG developed an XML representation of their graphical pathway maps to facilitate data exchange. XML is a meta-language used to create other specialized markup languages. The World Wide Web Consortiumg (W3C) developed XML as a data exchange language. It is a subset of the Standard Generalized Markup Languageh (SGML), the International Organization for Standardization'si meta-language for text markup systems (ISO 8879). W3C developed XML as a way to store data and the relationships between data elements in an electronic format. A markup language allows the encoding of the document's storage layout and logical structure and provides a mechanism to impose constraints on these. XML is used to store any kind of structured information and to store information in a way that makes it easier to pass it between different computing systems. The XML specification document (Bray et al., 2004) fully describes the language structure. XML is not a programming language but is a markup specification language used to describe information for storage, transmission, or processing by a program. It is important to note that XML is a meta-language and one must develop a specialized version of it before creating documents. KEGG calls their version of XML developed specifically for their pathway maps the KEGG Markup Language (KGML).

An XML document must conform to XML syntax, but it may also have another level of validation based on an XML schema or document type definition (DTD). The DTD is one tool typically used to create valid XML documents and describes the specific XML created language (in this case KGML). KEGG used a DTD to validate their KGML documents. It is a list of rules for representing your document type with XML (Goldfarb and Prescod, 1998). If two people exchanging documents follow the DTD, they can exchange data. In this case, the DTD functions to impose validation rules on the KGML files that KEGG created for each of its hand drawn maps.

The DTD defines the building blocks of a KGML document (elements), parent-child relationships between elements, and categorizes elements as required or optional. Any XML document may have a DTD referenced at the top of the document to establish the rules under which this document falls. For the KEGG KGML documents, a DTD reference exists at the top of each document:

<!DOCTYPE pathway SYSTEM "http://www.genome.ad.jp/kegg/xml/KGML_v0.2_.dtd">

This tells us that the name of the DTD is pathway and its location is on the KEGG website.

KEGG translated their reference and organism maps to KGML with the requirements of the DTD (Figures 2,3, and 4). If GenMAPP could generate similar maps based on the KGML, these would suitable as starter maps and would still save users a significant amount of time in pathway map development.

I analyzed these KGML files to see if a translation to GenMAPP format was possible. At the start of this project, KEGG exported their map catalog in KGML Version 0.2. Since that time, KEGG has released KGML versions 0.3 (October 2003, with minor changes) and 0.4 (April 2004, with major changes) (GenomeNet, 2004) which are not addressed in this document.

[g] The World Wide Web Consortium is an organization that develops common protocols to promote the interoperability of the web. http://www.w3.org/

[h] SGML is a generic markup language for representing documents and is an International Standard that describes the relationship between a document's content and its structure.

[i] (ISO) ISO, a voluntary, non-treaty organization founded in 1946, is responsible for creating international standards in many areas, including computers and communications. Its members are the national standards organizations of 89 countries.

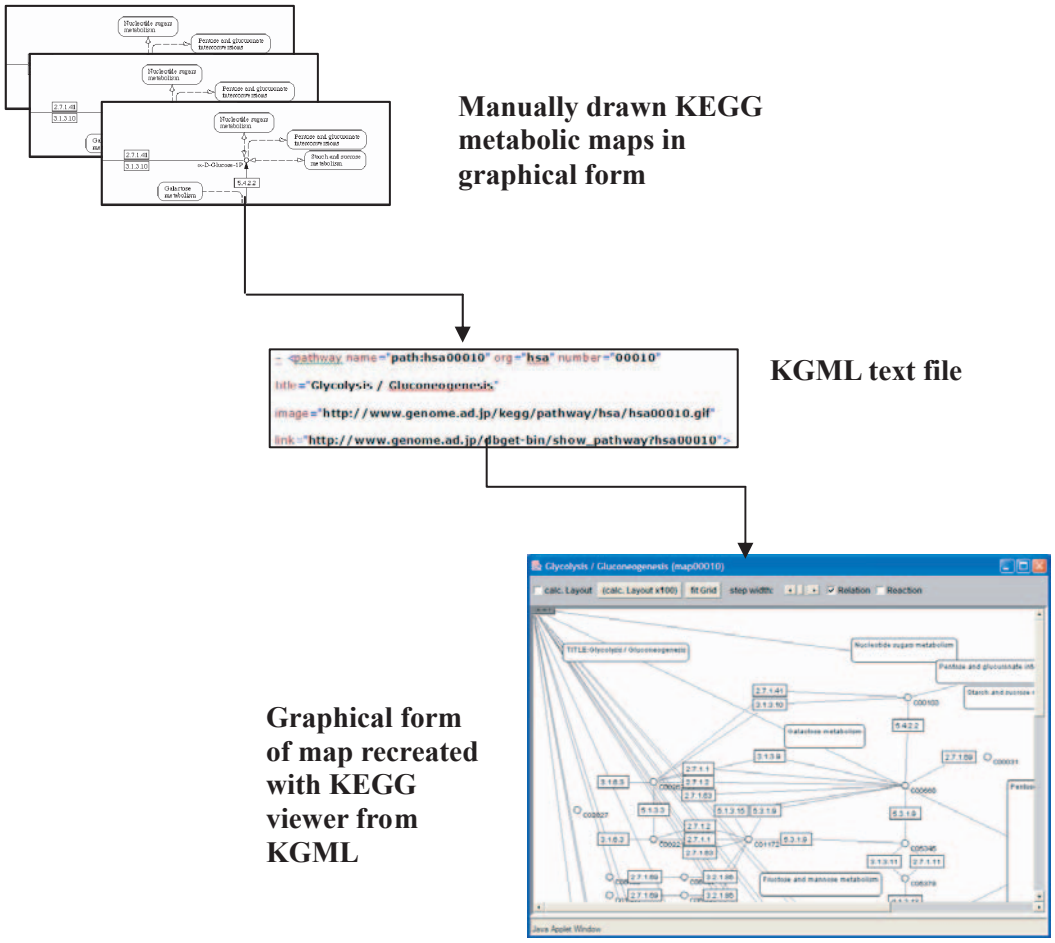


Figure 2: KEGG created KGML documents of their hand drawn pathway maps. KEGG also provides a KGML viewer that graphically recreates the pathway maps, though the layout is not the same as the hand drawn map.

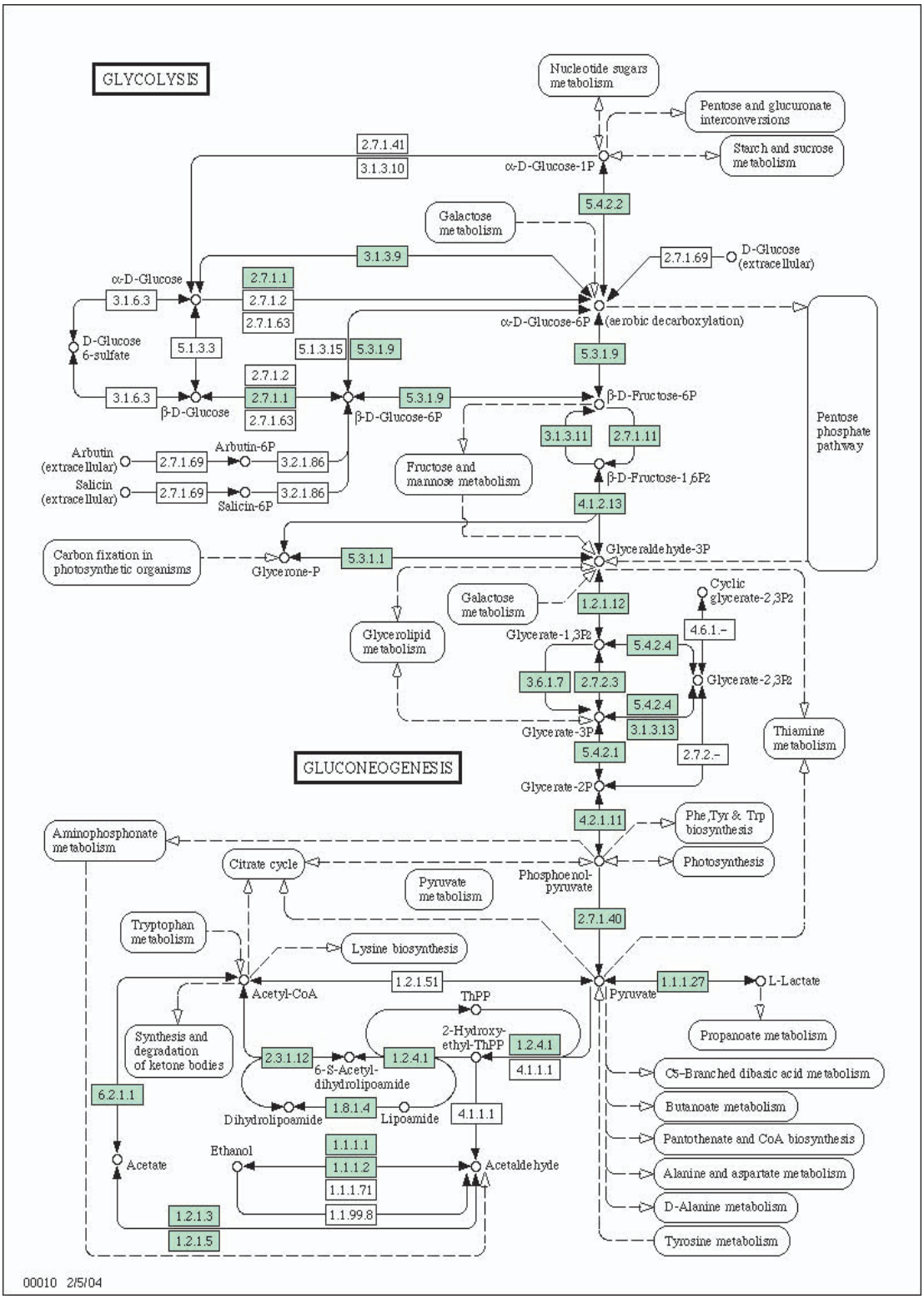


Figure 3: KEGG manually drawn map of Glycolysis/Gluconeogenesis pathway

http://www.genome.ad.jp/dbget-bin/get_pathway?org_name=hsa&mapno=00010

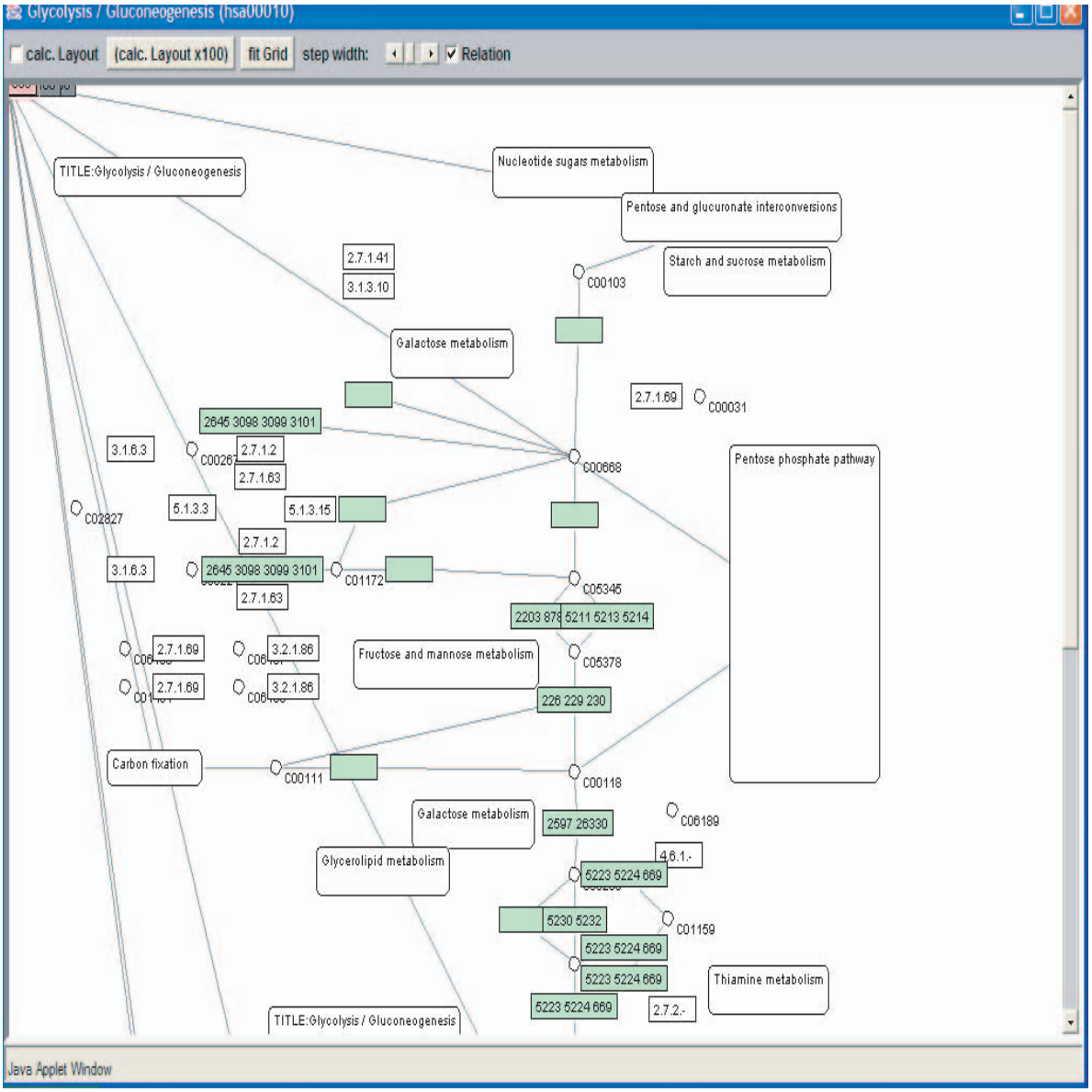


Figure 4: KEGG viewer converts KGML to a viewable map. This is a screen shot of a portion of the KEGG Glycolysis/Gluconeogenesis map reconstructed from KGML by the KEGG viewer. This map, and other KEGG viewer reconstructed maps, had many of the same problems as the GenMAPP converted maps, such as overwriting and orphan objects.
http://www.genome.ad.jp/kegg/KGML/KGML_v0.2/hsa/

Appendices A and B contain the KEGG Markup Language v0.2 Draft specification 0.2 DTD and the DTD for KGML v0.2. These documents specify the necessary and allowed contents of the KGML text file for each pathway map. The DTD is essential to deciphering the content of the KGML files. Figure 5 shows the DTD specification for the top-level pathway element. A pathway element in the DTD represents a single KEGG biological pathway. The first line shows the element pathway may contain any number of additional elements of the type **entry, reaction**, and **relation** (described below). The DTD uses occurrence indicators[j] that specify how many times each element may occur. The following lines list the attributes (ATTLIST) of the pathway element (name, number, org, title, image, and link). These attributes have types (such as keggid.type) defined in another part of the DTD (not shown). The attribute values are either REQUIRED or IMPLIED, where the latter means optional. For example, a pathway must have a name, number, and org, but a title, image, and link are optional. Figure 6 shows the KGML for the pathway element.

```
<!ELEMENT pathway (entry*, reaction*, relation*)>
<!ATTLIST pathway name          %keggid.type;      #REQUIRED>
<!ATTLIST pathway number        %mapnumber.type;   #REQUIRED>
<!ATTLIST pathway org           %maporg.type;      #REQUIRED>
<!ATTLIST pathway title         %string.type;      #IMPLIED>
<!ATTLIST pathway image         %url.type;         #IMPLIED>
<!ATTLIST pathway link          %url.type;          #IMPLIED>
```

Figure 5: DTD for pathway element of KEGG map. KGML KEGG generated for the glycolysis pathway element and shows that this pathway has each attribute described in the DTD (name, number, org title, image, and link) with values specific to the glycolysis pathway. The KGML for each element on the map follows.

```
<pathway name="path:hsa00020" org="hsa" number="00010"

title="Glycolysis / Gluconeogenesis"

image="http://www.genome.ad.jp/kegg/pathway/hsa/hsa00020.gif"

link="http://www.genome.ad.jp/dbget-bin/show_pathway?hsa00020">
```

Figure 6: KGML for a typical pathway element. This KGML represents the glycolysis and gluconeogenesis *Homo sapiens* map for the pathway element. The image and link elements point to the graphical image files at the KEGG site.

[j] The plus sign means that there may be one or more occurrences of the element; the question mark means that there may be at most one and possibly no occurrence; the asterisk means that the element may either be absent or appear one or more times

Figure 7 shows a graphical representation of the DTD. At the upper level of a hierarchy is the **pathway** element that represents one pathway map. Below pathway are the **entry** elements, which are the map nodes, and the **reaction** and **relation** elements, which are the graph edges that link nodes together. An **entry** element may contain a **graphics** element, which specifies layout and rendering information for the node. A **reaction** element specifies a reaction (reversible or irreversible) between two **entry** elements categorized as **products** and **substrates.** A **relation** element indicates how two **entry** elements are related through a **compound**. Each KGML file represents a pathway map by using this hierarchy of nodes (with associated layout and rendering information), and edges.

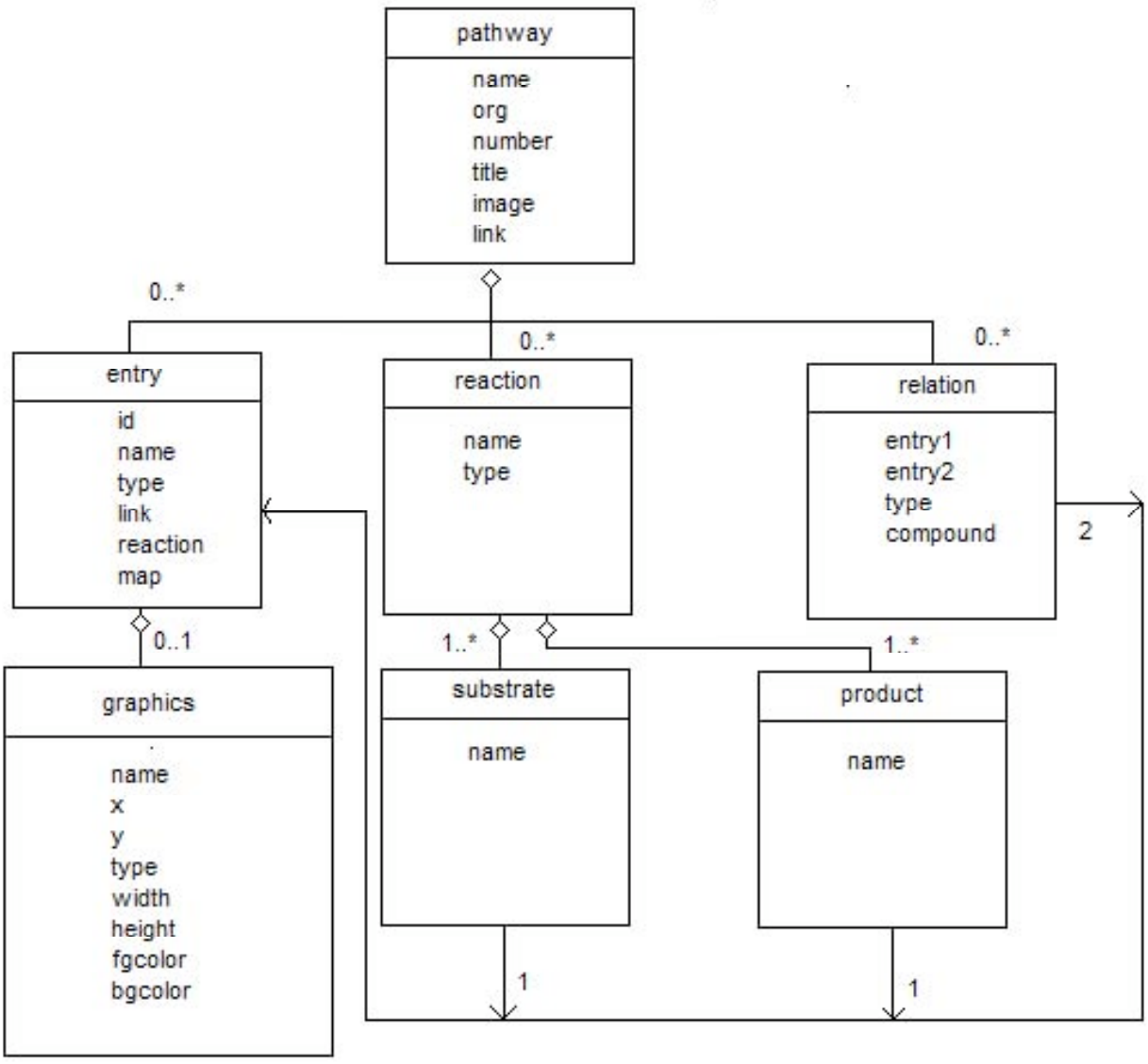


Figure 7: Graphical depiction of KGML DTD hierarchy. A pathway consists of zero or more entry, reaction, and relation elements. An entry optionally contains a graphic element. A reaction element must have one or more substrate and product elements. The entry1 and entry2 attributes in the relation element actually point to two entry ids.

Entry elements may be one of several types: enzyme, product, compound, ortholog, or another pathway map. Enzymes are gene products represented by Enzyme Commission numbers. Products are generally references to a gene or genes for a specific species and exist primarily on organism specific maps and not on reference maps. Compounds are chemical compounds. Orthologsk are not included in the scope of this project. Figure 5 shows a small piece of the KEGG Glycolysis map showing several **entries**. Solid or dotted lines connect the nodes and represent relationships between these objects. An **entry** object may also have a **graphics** element that describes its physical location, color, and dimensions on the map.

The edges of the graph are **relations**, which link two gene products together through a compound; and **reactions**, which link two or more compounds together. **Relations** and **reactions** are element types that have their own DTD section specifying necessary and optional attributes.

KEGG Map Conversion to GenMAPP

All KEGG reference and organism maps were converted from KGML to GenMAPP format. I found the converted KEGG maps could be a good starting point for some GenMAPP users but would require some user editing once displayed. Figure 8 shows the GenMAPP converted KEGG Glycolysis/Gluconeogenesis map for *Homo sapiens*. This map is typical in appearance to most of the converted maps. Since the KGML does not contain all relationships shown on the manually drawn maps, there are some orphan objects such as “Tyr metabolism” on the bottom right of the map. Since there are often multiple genes per EC number, the genes must be tiled or stacked on the GenMAPP map, sometime causing placement challenges (maps with over 20 genes per EC number are difficult to read). The KGML converted pathway maps are currently downloadable from the genmapp.org site and through the GenMAPP application.

[k] Orthologs are genes in different species that evolved from a common ancestral gene by speciation.

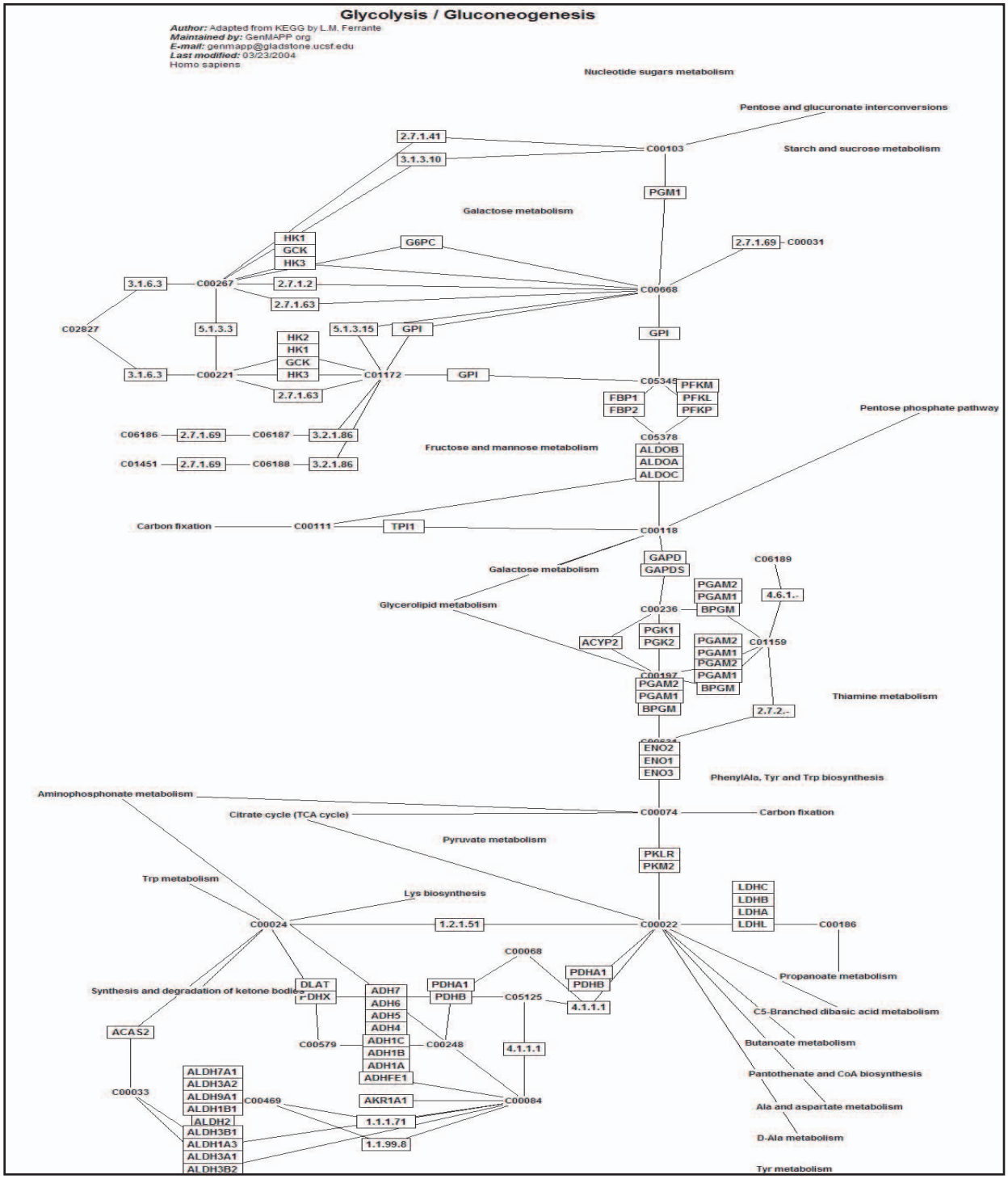


Figure 8: KEGG Glycolysis / Gluconerogenesis map converted to GenMAPP. EC numbers have been converted to genes for *Homo sapiens*.

BIOPAX INITIATIVE

In 2002, the Biopathways Data Exchange[l] (BioPAX) workgroup formed to address the significant time commitment the bioinformatics community was placing in pathways data exchange and integration. BioPAX plans to develop a common exchange format for biological pathway data to promote collaboration and accessibility**,** incorporating key elements from a range of pathway databases. BioPAX is not a formal standard, nor does it plan on going through a formal certification process, such as W3C at this time (Bader 2004). This initiative will live or die on community support and utilization. The intended users of BioPAX are pathways exchange databases and software tools that access these databases.

BioPAX includes pathways relating to cellular and molecular biology. The initial release (BioPAX v1.0, July 2004) encompasses metabolic pathways only. A timeline for other pathway types, including signal transduction pathways and molecular interaction networks, is available at the BioPAX website (www.biopax.org/Docs/BioPAX_Roadmap.html). BioPAX acknowledges that different pathway groups will use different internal representations for their data, based on their individual needs for optimization and that switching to a common internal representation is not feasible. However, a common exchange format is feasible and desirable and must be flexible enough to support the different internal representations in use in the pathway community. Several pathway database groups, such as BioCyc, BIND, Reactome, and WIT, have been involved in the development effort, and collaborating organizations include the Proteomics Standard Initiative[m], CellML[n], SBML[o], and the Chemical Markup Language[p].

It is important to note that the exchange format proposed by BioPAX is for **biological pathways**, and not **biological pathway *maps***. In BioPAX terminology, a biological pathway is a series of molecular interactions and reactions, often forming a network, which biologists have found useful to group together (BioPAX Workgroup, 2004). This pathway has no graphical components but may have direction. It contains steps that lead to the function the pathway represents. There are numerous ways to represent a pathway graphically; there is no one standard way to visually represent pathways other than to maintain the integrity of the relationships between the steps or interactions. Therefore, graphic layout information is not part of a biological pathway.

I investigated the BioPAX standard as a possible biological pathway and biological pathway map exchange format for GenMAPP. However, it was immediately clear that the BioPAX pathway exchange format is only a partial solution for GenMAPP. The scope of the BioPAX initiative is necessarily much broader than GenMAPP maps require. The content of GenMAPP pathway maps is broader than the pathway data covered by BioPAX. One of the goals of this study was to investigate whether a subset of the proposed exchange format was appropriate for GenMAPP, either now or in the future.

BioPAX Ontology.

Ontology is a tool used to accurately model the information content of a complex domain. Ontologies consist of definitions, concepts, and relationships used to represent a domain. Ontology provides the means to have rigorous definitions of terms and to capture semantic nuances.

Ontologies are useful to both people and software that need to share information. BioPAX is implementing biological pathway ontology to represent pathway concepts and their relationships. They will use their ontology to create a biological pathways exchange format. The implementation of this pathways exchange format ontology will be in OWL, a W3C recommendation[q] for the Semantic Web (Hendler *et al*., 2002). The Semantic Web is a

[l] The BioPAX core group includes representation from the Memorial Sloan-Kettering Cancer Center, SRI, Argonne National Labs, Harvard Center for Genomics Research, NIST, Columbia University, and the University of Colorado. http://www.biopax.org/index.html

[m] The Proteomics Standards Initiative (PSI) at the European Bioinformatics Institute aims to define community standards for data representation in proteomics to facilitate data exchange and verification. http://www.ebi.ac.uk/Information/meetings/psi.html

[n] The CellML language is an open standard based on the XML markup language. The purpose of CellML is to store and exchange computer-based mathematical models. http://www.cellml.org/public/about/what_is_cellml.html

[o] The Systems Biology Markup Language (SBML) is a computer-readable format for representing models of biochemical reaction networks. http://sbml.org/index.psp

[p] CML (Chemical Markup Language) is a new approach to managing molecular information using recently developed Internet tools such as XML and Java. http://www.xml-cml.org/

[q] W3C Recommendation (REC). A W3C recommendation is a specification or set of guidelines that, after extensive consensus-building, has received the endorsement of W3C members. W3C recommends the wide deployment of its recommendations. W3C recommendations are similar to the standards published by other organizations. http://www.w3.org/2003/06/Process-20030618/tr.html#RecsW3C

W3C vision incorporating standards, ontologies, and infrastructure to give web information a well defined meaning and make it easier to process by machines and humans (Daconta *et al.,* 2003). OWL is a language designed to describe formally the meaning of terminology used in Web documents by assisting machine interpretability of documents. It provides the means to represent semantic relationships of pathway data, cardinality, characteristics of properties (e.g., symmetry), relations between classes (e.g., disjointness), and enumerated classes.

The overall concept behind the Level 1 BioPAX metabolic pathways ontology (BioPAX Workgroup, 2004) is as follows: biologic ***pathways*** consist of ***interactions*** between ***physical entities***, such as proteins or other molecules. A pathway (e.g., glycolysis) is a set of interactions, an interaction is a set or sets of physical entities with some relationship between them, and "physicalEntitys" are building blocks (e.g., proteins, rna) of interactions. The ontology describes the abstract root class as an entity[r] which contains second level classes pathway, physicalEntity, and interaction.

Figure 9 is an overview of the BioPAX ontology. **Pathway** is a class defined as a set of molecular interactions and reactions, often forming a network, grouped together for organizational, biophysical or other reasons.

The **interaction** class defines a single biochemical interaction between entities and cannot be defined without the entities it interrelates. This class includes only biochemical reactions at this time but may later be expanded to include temporal, logical, genetic, and other types of relationships.

The **physicalEntity** class describes an entity that has physical structure and is limited to molecules encompassed by subclasses **protein, small molecule**, **RNA**, and **complex. Proteins** are a sequence of amino acids. **RNA** includes all sequences of ribonucleic monophosphates, such as mRNA, microRNA, and rRNA. A **small molecule** is a bioactive molecule that is not a peptide, protein, or RNA .Complex carbohydrates and DNA are forced into the small molecule subclass since they do not have a class of their own in this version of the ontology. A **complex** consists of other physicalEntities bound non-covalently, at least one of which is a protein or RNA, and must be stable enough to function as a biological unit.

The **interaction** class defines two types of interaction subclasses. The **control** interaction subclass describes an interaction where one entity regulates, modifies, or influences another. Two types of control interaction subclasses are **catalysis** and **modulation.** In a **catalysis** control interaction, a catalyst increases the interaction rate by lowering the activation energy. **A modulation** control interaction involves a physical entity (such as a small molecule) and alters the ability of a catalyst to catalyze a reaction. **Conversion** interactions describe a single step conversion process, such as a biochemical reaction where one or more entities are physically transformed, and include the following subclasses: **biochemical reaction, complex assembly,** and **transport**. In a **biochemical reaction,** one or more entities change covalently to become other entities. In a **complex assembly**, a set of physical entities aggregate non-covalently. **Transport** involves the change of location of an entity. Since some interactions can be classified as both biochemical and transport, an additional subclass, **transportWithBiochemicalReaction**, is also provided.

Classes and subclasses have attributes (called slots in OWL). Attributes may be inherited or specifically defined in a class or subclass. For example, the classes pathway, interaction, and physicalEntity all inherit a number of attributes from the top level entity object, including name, short-name, availability, data-source, synonyms, and xref, but the pathway class has, in addition, organism and pathway-components attributes. The BioPAX OWL implementation describes each class, subclass, and attribute in detail.

The BioPAX ontology also contains a top level Utility class that provides custom data types when a simple type, such as a string, is not sufficient. The BioPAX Level 1 documentation describes the Utility class and its subclasses.

The BioPAX ontology and its implementation as a standard exchange format are still in a developmental phase and no doubt will change as the pathways community uses it and provides feedback to BioPAX.

[r] Entity is the root class and contains any concept that is referred to as a discrete unit when describing pathways, such as pathway, interaction or physical entity.

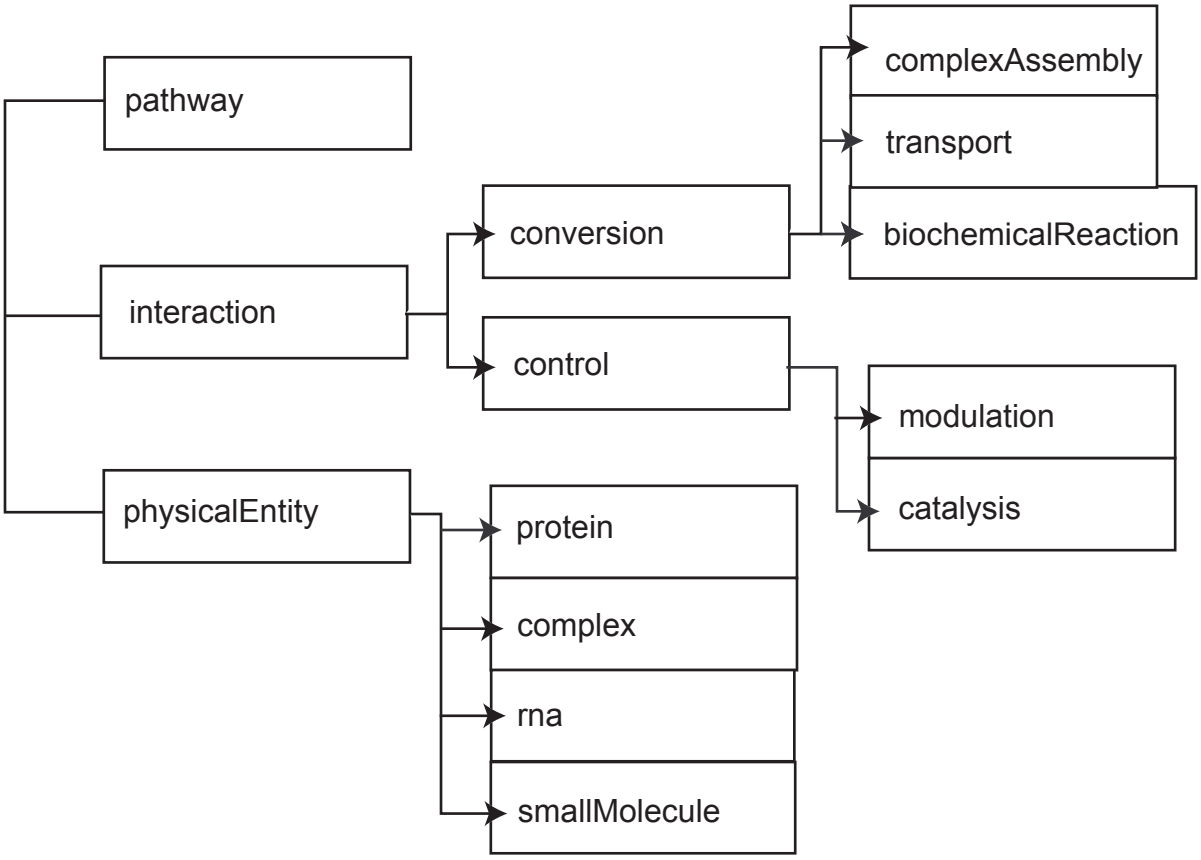


Figure 9: Overview of the BioPAX ontology top level classes and their subclasses

Mapping GenMAPP Map Objects to BioPAX

To relate a biological pathway map to a pathway exchange standard, we must acknowledge the difference between a biological pathway and a biological pathway map. A biological pathway map is one visual representation of the steps in a pathway and may have additional information on it that helps the viewer understand the pathway, customizes it for a particular purpose, or has significance to the author of the map. This type of information is not included in the BioPAX exchange format.

To assist the translation of GenMAPP biological pathway maps to the BioPAX biological pathway exchange format, I classified the data on GenMAPP maps into two categories based on the above discussion, pathway information and map information. Pathway information consists of general information about the pathway, such as name, core information about GenMAPP "gene" objects (excluding layout), and their inter-relationships with other "gene" objects. These objects describe primarily the pathway itself and not an interpretation of or customization of the pathway. This is the core information that BioPAX is targeting in its proposed exchange format for pathways. There should be no layout information in this category. General information about the GenMAPP map corresponds to the BioPAX pathway class. GenMAPP "gene" objects correspond to BioPAX physicalEntitys (protein, RNA, or complex) with the exception of their graphical layout information. Inter-relationships between gene objects, when defined as edges between nodes on a graph, would normally fall into this category. However, GenMAPP "gene" objects relate to other "gene" objects visually through a vector representation of lines and arrows. These lines and arrows have x-y start and end coordinates that point to the "genes" they connect and can be moved independently of these objects using the GenMAPP Drawing Board. There is no true internal representation of interaction (as defined by BioPAX) in GenMAPP maps (note, GenMAPP.org expects to develop an interaction type in a future release to meet this need). GenMAPP lines and arrows that represent relationships between objects cannot be considered pathway information in the context of BioPAX.

Map information consists of everything else on the map, including the GenMAPP object types line, arrow, label, shape (including cell, cell component, ligand, receptor, and protein complex shapes) and all graphical information about each object on the map including genes (e.g., as x-y coordinates, width and height). This information helps us build a map from digital data but is out of the realm of a pathways exchange format like BioPAX.

Table 2 shows the mapping of pathway attributes from the GenMAPP map database *info* table to BioPAX classes and subclasses. Most attributes, with the exception of pathway-components, email, and map maintainer, can be mapped between GenMAPP and BioPAX. Certain GenMAPP pathway attributes are specific to the GenMAPP environment and show as "N/A" in the table

Table 3 shows the mapping of all other GenMAPP objects to BioPAX classes. As noted previously, only the GenMAPP gene object maps to BioPAX physicalEntity protein, RNA, or complex subclasses (without layout or rendering information). All other GenMAPP objects, such as the cell, ribosome, organelle, line, arrow, and label, are not mapped to the BioPAX pathways exchange format.

There are several ways to handle other map objects and layout data that falls outside the scope of BioPAX. The first possibility is to have separate map export and pathways data exchange formats, introducing BioPAX as the latter. Another possibility is to have one format but discard all objects that do not fit into BioPAX, such as GenMAPP shapes. The GenMAPP staff is currently discussing the necessity of the shape objects available in the drafting board and will decide on how to phase out any extraneous shapes. We could still maintain layout data (and perhaps labels) within BioPAX by encoding this information into the BioPAX comment attribute of each node object. GenMAPP could design a format for encoding this layout information or follow an existing format, such as the Systems Biology Markup Language Layout Extension (Hucka *et al.*, 2003) (Gauges *et al.*, 2004). The SBML Layout Extension includes layout but not rendering information. SBML will support this extension and since BioPAX wants to remain compatible with SBML, it may also be supported by BioPAX in the future (Bader, 2004). Other possibilities include the Scalable Vector Graphics (SVG) specification (Andersson *et al.*, 2003) and Graph Manipulation Language (GML) (Brainsys, unknown). This report does not assess these layout specifications.

Table 2: GenMAPP 2.0 pathway object mapped to the BioPAX exchange format V1.0.

GenMAPP pathway attribute	GenMAPP V2 attribute definition	BioPAX	BioPAX Level 1.0 attribute definition
title	The title that appears at the top of the graphic	name or short-name	The preferred full name for this entity An abbreviated name for this entity. Preferably a name that is short enough to be used in a visualization application to label a graphical element that represents this entity.
version	The version of the GenMAPP program used to create this MAPP	N/A, GenMAPP specific	
author	Author of the map	data-source	A free text description of the source of this data (e.g. a database or person name)
maint	**Person who maintains the MAPP**	**No match, though possibly xref**	**When exporting GenMAPP maps, place in the BioPAX pathway comments attribute**
email	**Email address for MAPP information**	**No match**	**When exporting GenMAPP maps, place in the BioPAX pathway comments attribute**
copyright	Copyright date and entity	availability	Describes the availability of this data (e.g., a copyright statement).
modify	Date of last modification	N/A, GenMAPP specific	
remarks	Any descriptive details the author wishes to include on the graphic. References, credits, limited to 50 characters Place all BioPAX attributes listed into GenMAPP remarks attribute on import of maps from a BioPAX format On export of GenMAPP maps, place GenMAPP remarks	comment	Comment on the data
		organism	An organism (e.g., *Homo sapiens.*
		synonyms	One or more synonyms for the name of this entity
		xref	Values of this slot define external cross-references from this entity to entities in external databases.
boardwidth	Width in twips of the virtual DraftingBoard	N/A, GenMAPP specific	
boardheight	Height in twips of the virtual DraftingBoard	N/A, GenMAPP specific	
windowWidth	Width in twips of the DraftingBoard window	N/A, GenMAPP specific	
windowHeight	Height in twips of the DraftingBoard window	N/A, GenMAPP specific	
notes	These will not appear on the graphic	N/A, GenMAPP specific	
InfoboxLeft	Layout of GenMAPP infobox	N/A, GenMAPP specific	
InfoboxTop	Layout of GenMAPP infobox	N/A, GenMAPP specific	
LegendLeft	Layout of GenMAPP legend	N/A, GenMAPP specific	

GenMAPP pathway attribute	**GenMAPP V2 attribute definition**	**BioPAX**	**BioPAX Level 1.0 attribute definition**
LegendTop	Layout of GenMAPP legend	N/A, GenMAPP specific	
No match with GenMAPP V2	**GenMAPP V2 has no interaction type**	**pathway-components**	**A list of interactions and/or steps in this pathway/network**

Table 3: GenMAPP 2.0 map objects mapped to the BioPAX exchange format 1.0

GenMAPP GeneProduct object attribute and definition	**GMML GeneProduct attribute**	**BioPAX PhysicalEntity Protein, RNA, or Complex attribute**
ID Gene ID stated as primary by user.	**Name**	**name**
System Characters to identify the gene ID system of the gene ID. See *SystemCode* in *Systems* table in Gene Database formats. The program allows only 2 characters; the third is reserved for expansion.	**Data-Source** Allowable systems (such as Swiss-Prot) could be itemized in DTD and GenMAPP import / export program could translate to GenMAPP code	**data-source**
Label Text for genes, labels, proteins, etc. This is text that appears on the map and may be different than gene name.	**Short-Name**	**short-name** A short label. Short enough to be used in a visualization application to label a graphical element
Head Heading for the Backpage display for this object	**BackpageHead**	N/A, this is a GenMAPP display concept and not a pathway property
Remarks Remarks will show on GenMAPP Backpage for this gene	**Comment**	**comment**
Links Links to other stuff, internal or internet.	**Xref**	**Xref** An external cross-reference
Notes Appears only in database	**Notes**	N/A, this is an internal GenMAPP concept
CenterX	**CenterX**	N/A, this is a display property not a pathway property
CenterY	**CenterY**	N/A, this is a display property not a pathway property
Width Object width in twips	**Width**	N/A, this is a display property not a pathway property
Image Image to display within the object.	**Image**	N/A, this is a display property not a pathway property

GENMAPP DTD FOR XML IMPORT AND EXPORT

To further gauge compatibility between GenMAPP map objects and the BioPAX ontology and to explore the possibility of importing and exporting GenMAPP maps in XML, I developed an XML Document Type Definition and incorporated, to the extent possible, the BioPAX ontology. When there was a choice between using a GenMAPP and a BioPAX term, I used BioPAX for a clearer transfer. The DTD, attached as Appendix C, will be used with GenMAPP 2.0 as an interim measure to import and export maps as XML. The DTD reflects the decision to export maps in a form that GenMAPP software requires to accurately reconstruct the maps on import. The DTD describes an XML (referred to as GenMAPP Markup Language or GMML) designed to export and import GenMAPP maps, including both pathway and other map information. The top-level element is called a **pathway.** Pathway elements in GMML have some attributes, such as Name and Organism, which map directly to BioPAX pathway attributes. Other GenMAPP pathway element attributes have no match in BioPAX but are available in GMML to preserve the information on a GenMAPP map. Unlike the BioPAX version of a pathway class that is constructed solely from the BioPAX classes interaction and physicalEntity, the GenMAPP pathway element consists of any number of **GeneProducts, Lines, Labels, Shapes, SmallMoleculeShapes, CellShapes, CellComponentShapes,** and **ProteinComplexShapes**. These GMML categories encompass all the possible objects that can be on a GenMAPP map, and all that end with the word "shape" are shapes on the map and not gene objects in the GenMAPP sense. A GMML **GeneProduct** corresponds to a GenMAPP gene object. A **GeneProduct** is further subclassified into protein, complex, or RNA through the type attribute and thus can be used to import and export BioPAX protein, complex, and RNA classes. The GMML category **Shape** includes the GenMAPP arc, oval, brace, and rectangle. The GMML **SmallMoleculeShape** includes all GenMAPP receptor and ligand icons. Biological receptors are generally proteins. However, I chose not to classify receptor icons as GeneProducts of type protein since GenMAPP genes have special meaning, and a GenMAPP receptor icon is only a shape. The GMML category **CellComponentShape** includes all GenMAPP organ shapes as well as the ribosome shape.

The gene is the fundamental unit behind the GenMAPP software. It is used to display gene expression data on maps and provides the mechanism to tie each gene to its "Backpage" information. A new window displays a GenMAPP Backpage when a user double clicks on a specific gene. A Backpage contains additional information about the gene from its data source (e.g., Swiss-Prot) and other sources (e.g., LocusLink or GeneOntology) as well as appropriate web links. Therefore, preservation of gene information during import and export of maps is critical. In GMML, a GenMAPP gene can be a protein, RNA, or complex. Several of the BioPAX protein, RNA, and complex attributes describe a GeneProduct in GMML, as shown in Table 3.

IV. CONCLUSIONS AND RECOMMENDATIONS

Feasibility of importing existing external metabolic maps into GenMAPP. Close to 700 KEGG metabolic maps were successfully imported into GenMAPP. These converted maps are of sufficient quality to be used as a starting point for some GenMAPP users but will require some degree of editing once displayed.

The KEGG to GenMAPP conversion provided an ideal example to show the challenges and opportunities of this project. KEGG exported its catalog of maps into a language based on XML (KGML) and provided a DTD and some additional documentation for KGML users. However, the definitions in the documentation were not always rigorous enough to use for map conversion. A DTD does not require detailed definition of terms. For example, while the KGML contained x-y coordinates for each node in the KGML, the KGML documentation did not provide units for these coordinates and KEGG did not respond to queries on this topic. The KGML did not contain all the relationships on the hand-drawn maps, causing the converted maps to have some objects that were not connected to anything else on the map. Also, it was not clear why the organism specific KGML had fewer relationships than the reference KGML, and if, in fact, it was appropriate to use the reference maps as a starting point for organism specific maps. The alternative was to generate organism specific maps directly from the organism KGML files, but these maps were sketchy at best, and did not correspond to maps displayed at KEGG. It is not clear why this relationship information is not in the KGML. Several inquiries to Genome Net did not result in additional information on this.

The KGML provided graphical coordinates for objects, but not relationships between objects. Layout of relationships on the converted maps was problematic. Lines representing a relationship between two objects sometimes cross over other objects and lines. This is because the algorithm used for drawing these relationships was from the edge of one object to the edge of the other. Map layout software that would solve this problem was beyond the scope of this project. This makes it difficult to use KGML for GenMAPP maps. The KEGG site has a KGML viewer that displays maps directly from the KGML. The layout of these maps is radically differently from the hand-drawn maps and also lacks the relationship information from the hand drawn maps.

It was time consuming to do this transformation based on the information available from KEGG. We now have almost 700 metabolic maps converted from KEGG in GenMAPP, but the process of conversion made a need for rigorous data definition quite clear. In addition to a standard exchange format we need a rigorous definition of terms, semantic relationships, and characteristics of data. If an exchange format is to be successful, it must be documented beyond the scope required for a DTD.

The BioPAX Pathways Exchange Format. The BioPAX Pathways Exchange format is an appropriate future standard data exchange format for GenMAPP maps with some modifications**.** BioPAX is a promising development in the solution for the pathways exchange problem. Future deployment of GenMAPP maps in BioPAX is a direction that this study would support**.** I assessed the BioPAX pathway exchange format and found that GenMAPP maps have two categories of data that can be mapped to BioPAX. The first is general pathway information. The second category is the non-graphic attributes of GenMAPP gene objects. A third category will be added when GenMAPP adds an interaction type to its maps. All other GenMAPP objects, such as the cell, ribosome, organelle, line, arrow, and label are not mapped to the BioPAX pathways exchange format.

Proposed GenMAPP pathway exchange format. An exchange format for GenMAPP maps should independently provide for both exchange of pathway data and additional objects on the map. It must also provide a clear separation of these categories of data in the exchange format. GenMAPP should also provide rigorously defined terminology along with an exchange format to prevent misunderstanding of terms by external sources. This exchange format will change over time as GenMAPP adds additional features and adds an interaction type to maps.

GenMAPP should implement an exchange format in several phases. The first phase involves importing and exporting GenMAPP *pathway maps* as XML, using the GMML and DTD defined in this document. This provides a short-range opportunity for importing and exporting maps as XML. However, GenMAPP should first review and clarify all definition of map data and objects for the biological pathways user community[s]. Conversion of KEGG maps without rigorous definition of terms was difficult and error-prone. An OWL ontology would be one rigorous way to define GenMAPP map terms even if it initially was not used as an exchange format. The GenMAPP

[s] The GenMAPP V2 software has a glossary but this is not sufficient for the purposes of map exchange.

pathway map ontology should be available to anyone who wants to exchange pathways with GenMAPP. GMML allows for some separation of pathways and other map data since different elements categorize map objects, and at this time only GenMAPP Pathway and GeneProduct objects have any correlation to BioPAX. In addition, all graphics coordinates in GMML are in a separate element from map objects. If only pathway data is required for export, external programs can import only those attributes of objects that represent pathway data. This will be a step in the right direction and would provide a straightforward way to exchange maps with the pathway community in the near future. External pathway sources that wish to exchange pathway information with GenMAPP would need to understand and accept this format for input into their software. As GenMAPP evolves, GMML would need to evolve with it and with revisions could still be used to export and import pathway maps after a GenMAPP interaction type is developed.

The second phase involves using the BioPAX exchange standard for GenMAPP *pathway data*. GenMAPP must develop an interaction type for its map data for full implementation of the BioPAX pathway data exchange format. Discussions within GenMAPP are under way at this time to iron out the specific details of an interaction type implementation. GenMAPP staff involved in these discussions must thoroughly review the BioPAX ontology interaction class, subclasses, and attributes since compatibility with BioPAX can only be advantageous. In addition, GenMAPP needs to write the appropriate software to import and export pathways in BioPAX format. An OWL ontology can include instances of classes and is therefore the BioPAX exchange format. Software would need to export pathways data as an OWL file and import pathways data from an OWL file. This approach would address importing and exporting pathway data and would be completely compatible with BioPAX. The BioPAX comment fields are potentially useful for encoding layout information.

In summary, a phased implementation for the GenMAPP pathways exchange format is recommended. This format can include both pathway data and map layout information, with the initial implementation in GMML with use of BioPAX terminology where possible. The second phase should wait until GenMAPP implements an interaction type for maps. At that time it will be appropriate to move to the BioPAX ontology.

I used the GenMAPP pathway format as a vehicle to explore the world of biological pathways. I found a virtual “Tower of Babel” with inconsistent formats that are difficult to exchange. I identified some promising solutions to this problem and devised a plan to allow GenMAPP to play a constructive role in our ultimate goal of creating a publicly accessible and comprehensive map of biology.

V. REFERENCES

(2004). Extensible Markup Language (XML) 1.1 W3C Recommendation, T. Bray, Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F., Cowan, J., ed.

Andersson, O., Armstrong, P.,Axelsson, H.,Berjon, R., Bezaire, B, Bowler, J., Brown, C., Bultrowicz, M., Capin, t., Capsimalis, M., Carlander, M., Cederquist, J., Christopoulis, C., Cohn, R., Cole,L, and others (2003). Scalable Vector Graphics (SVG) 1.1 Specification.

Bader, G. D. (2004). Email to the author regarding formal standard . L. Ferrante, ed.

Bader, G. D. (2004). Email to the author regarding SBML, L. Ferrante, ed.

Bader, G. D., Betel, D., and Hogue, C. W. (2003). BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res *31*, 248-250.

BioPAX-Workgroup (2004). BioPAX- Biological Pathways Exchange Language Level 1, Version 1.0 Documentation.
http://www.biopax.org/Download.html

BioPAX-Workgroup (2004). BioPAX Level 1 Ontology. http://www.biopax.org/Download.html

BRAINSYS, I. (unknown). Graphlet: the GML File Format. http://www.infosun.fmi.uni-passau.de/Graphlet/GML/index.html

Daconta, M., Obrst, L. J., and Smith, K. T. (2003). The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management).

Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet *31*, 19-20.

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol *4*, R7.

Gauges, R., Rost, U., Sahle, S., and Wegner, K. (2004). Including layout Information in SBML files Version 2.1. http://projects.eml.org/bcb/sbml/level2/20040630/SBMLLayoutExtension-20040630.pdf

GenomeNet (2004). README for KGML v0.4. http://www.genome.jp/kegg/docs/xml/KGML_v0.4_.readme.html

GenomeNet (undated). KEGG Markup Language Manual. http://www.genome.ad.jp/kegg/docs/xml/

Goldfarb, C. F., and Prescod, P. (1998). The XML handbook (Upper Saddle River, NJ, Prentice Hall PTR).

Hendler, J., Berners-Lee, T. and and Miller, E. (2002). Integrating Applications on the Semantic Web. Journal of the Institute of Electrical Engineers of Japan *122*, 676-680.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A*., et al.* (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics *19*, 524-531.

Joshi-Tope G., V. I., Gopinathrao G., Matthews L., Schmidt E., Gillespie M., D'Eustachio P., Jassal B., Lewis S., Wu G., Birney E., and Stein L. (2003). The Genome Knowledgebase: A Resource for Biologists and Bioinformaticists. In Cold Spring Harbor Symposia on Quantitative Biology (Cold Spring Harbor Laboratory Press).

Kanehisa, M. (1997). A database for post-genome analysis. Trends Genet *13*, 375-376.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). The KEGG databases at GenomeNet. Nucleic Acids Res *30*, 42-46.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res *32 Database issue*, D277-280.

Karp P.D., A. M., Collado-Vides J., Ingraham J., Paulsen I.T., Saier M.H. Jr. (2004). The E. coli EcoCyc Database: No Longer Just a Metabolic Pathway Database. ASM News *70*, 25-30.

Schaefer, C. F. (2004). Pathway databases. Ann N Y Acad Sci *1020*, 77-91.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Wittig, U., and De Beuckelaer, A. (2001). Analysis and comparison of metabolic pathway databases. Brief Bioinform *2*, 126-142.

APPENDIX A: KEGG Markup Language v0.2 DTD draft specification

http://www.genome.ad.jp/kegg/KGML/KGML_v0.2/DraftSpecification.html

KEGG Markup Language v0.2 DTD draft specification

Structure of KGML documents (About the tree structure of the elements)

- **pathway** (root)
  - **entry** (0..*)
    - **graphics** (0..1)
  - **reaction** (0..*)
    - **substrate** (1..*)
    - **product** (1..*)
  - **relation** (0..*)

KGML data types

- **number.type**
  Positive integer
- **string.type**
  Character string
- **id.type**
  Identification number applied to each entry
- **idref.type**
  ID numbers reference
- **url.type**
  URL form
- **keggid.type**
  Character string of KEGGID form [db]:[acc]
- **maporg.type**
  Alphabet of three characters string : organism prefix or "map"
- **mapnumber.type**
  Five-digit number : map number
- **entry-type.type**
  It is in any of the following.
  (enzyme|product|products|ortholog|compound|map)
- **reaction-type.type**
  It is in any of the following.
  (reversible|irreversible)
- **relation-type.type**
  It is in any of the following.
  (ECrel|PPrel|GErel)
- **graphics-type.type**
  It is in any of the following.
  (rectangle|circle|roundrectangle)
- **graphics-color.type**
  The form is a numerical RGB specification.

- ex) #FFFFFF (this is white.)

About the attributes of each element

- **pathway**

NAME	Type	Comment	Default
name	keggid.type	the KEGGID of this pathway map	REQUIRED
number	mapnumber.type	the map number of this pathway map	REQUIRED
org	maporg.type	the organism prefix that this pathway map expresses ("map" in case of reference map)	REQUIRED
title	string.type	the title of this pathway map	IMPLIED
image	url.type	the URL of this pathway map's image	IMPLIED
link	url.type	the URL which relates to this pathway map	IMPLIED

- **entry**

Name	Type	Comment	Default
id	id.type	the ID of this entry in this map (Positive number)	REQUIRED
name	keggid.type	the KEGGID of this entry	REQUIRED
type	entry-type.type	the type of this entry	REQUIRED
link	url.type	the URL which relates to this entry	IMPLIED
reaction	keggid.type	the KEGGID of REACTION DB that this entry has	IMPLIED
map	idref.type	the ID of map where this entry appears(This attribute appears only to the appearing entry in other maps.)	IMPLIED

- **graphics**

Name	Type	Comment	Default
name	string.type	the label of this graphics object	IMPLIED
x	number.type	the X axis position of this graphics object	IMPLIED
y	number.type	the Y axis position of this graphics object	IMPLIED
type	graphics-type.type	the shape of this graphics object	rectangle
width	number.type	the width of this graphics object	45
height	number.type	the height of this graphics object	17
fgcolor	graphics-color.type	the foreground color used by this graphics object	#000000
bgcolor	graphics-color.type	the background color used by this graphics object	#FFFFFF

- **reaction**

Name	Type	Comment	Default
name	keggid.type	the KEGGID of this reaction	REQUIRED
type	reaction-type.type	the type of this reaction	REQUIRED

- **substrate**

Name	Type	Comment	Default
name	keggid.type	the KEGGID of this substrate	REQUIRED

- **product**

Name	Type	Comment	Default
name	keggid.type	the KEGGID of this reaction	REQUIRED

- **relation**

Name	Type	Comment	Default
entry1	idref.type	the first entry which define this relation	REQUIRED
entry2	idref.type	the second entry which define this relation	REQUIRED
compound	idref.type	the entry of compound which connects between entries	IMPLIED
type	relation-type.type	the type of this relation	REQUIRED

[KGML]

APPENDIX B: Document Type Definition for KEGG Markup Language v0.2

http://www.genome.ad.jp/kegg/KGML/KGML_v0.2/KGML_v0.2_.dtd

```
<!-- DTD for KEGG Markup Language v0.2 -->

<!-- Positive number type -->
<!ENTITY % number.type    "NMTOKEN">

<!-- String type -->
<!ENTITY % string.type    "CDATA">

<!-- ID type -->
<!ENTITY % id.type        "%number.type;">

<!-- IDREF type -->
<!ENTITY % idref.type     "%number.type;">

<!-- URL type -->
<!ENTITY % url.type       "%string.type;">

<!-- KEGGID type
   KEGG ID form : "[db]:[accession]"
-->
<!ENTITY % keggid.type    "%string.type;">

<!-- MAPORG type
   organism prefix or "map" : The alphabet of two or three characters
-->
<!ENTITY % maporg.type    "%string.type;">

<!-- MAPNUMBER type
   map number : The five-digit number
-->
<!ENTITY % mapnumber.type "%string.type;">

<!-- Type of Entry -->
<!ENTITY % entry-type.type "(enzyme|product|products|ortholog|compound|map)">

<!-- Type of Reaction -->
<!ENTITY % reaction-type.type "(reversible|irreversible)">

<!-- Type of Relation -->
<!ENTITY % relation-type.type "(ECrel|PPrel|GErel)">

<!-- Type of graphics shape -->
<!ENTITY % graphics-type.type "(rectangle|circle|roundrectangle)">

<!-- graphics-color type
    this type is a string that represents the color to be used by the Graphic object.
    The color is a numerical RGB specification.
    ex) #FFFFFF
-->
<!ENTITY % graphics-color.type "%string.type;">

<!ELEMENT pathway (entry*, reaction*, relation*)>
<!ATTLIST pathway name        %keggid.type;         #REQUIRED>
<!ATTLIST pathway number    %mapnumber.type;   #REQUIRED>
<!ATTLIST pathway org          %maporg.type;        #REQUIRED>
```

```
<!ATTLIST pathway title          %string.type;          #IMPLIED>
<!ATTLIST pathway image        %url.type;             #IMPLIED>
<!ATTLIST pathway link          %url.type;             #IMPLIED>

<!ELEMENT entry    (graphics?)>
<!ATTLIST entry    id           %id.type;              #REQUIRED>
<!ATTLIST entry    name         %keggid.type;          #REQUIRED>
<!ATTLIST entry    type         %entry-type.type;      #REQUIRED>
<!ATTLIST entry    link         %url.type;             #IMPLIED>
<!ATTLIST entry    reaction     %keggid.type;          #IMPLIED>
<!ATTLIST entry    map          %idref.type;           #IMPLIED>  <!-- If the entry has attribute of map ,
                                                        it is a entry on other pathwaymap. -->

<!ELEMENT reaction (substrate+, product+)>
<!ATTLIST reaction name     %keggid.type;             #REQUIRED>
<!ATTLIST reaction type     %reaction-type.type;      #REQUIRED>

<!ELEMENT substrate EMPTY>
<!ATTLIST substrate name     %keggid.type;            #REQUIRED>
<!ELEMENT product   EMPTY>
<!ATTLIST product   name     %keggid.type;            #REQUIRED>

<!ELEMENT relation EMPTY>
<!ATTLIST relation entry1   %idref.type;              #REQUIRED> <!-- This attribute value indicates
                                                        attribute of ID defined in the entry.  -->
<!ATTLIST relation entry2   %idref.type;              #REQUIRED> <!-- This attribute value indicates
                                                        attribute of ID defined in the entry.  -->
<!ATTLIST relation compound %idref.type;              #IMPLIED>  <!-- This attribute value indicates
                                                        attribute of ID defined in the entry.  -->
<!ATTLIST relation type     %relation-type.type;      #REQUIRED>

<!ELEMENT graphics EMPTY>
<!ATTLIST graphics name     %string.type;         #IMPLIED   >
<!ATTLIST graphics x        %number.type;         #IMPLIED   >
<!ATTLIST graphics y        %number.type;         #IMPLIED   >
<!ATTLIST graphics type     %graphics-type.type;   "rectangle">
<!ATTLIST graphics width    %number.type;          "45"       >
<!ATTLIST graphics height   %number.type;          "17"       >
<!ATTLIST graphics fgcolor  %graphics-color.type;  "#000000"  >
<!ATTLIST graphics bgcolor  %graphics-color.type;  "#FFFFFF"  >
```