Genomics & Informatics Vol. 6(3) 147-152, September 2008

Parsing KEGG XML Files to Find Shared and Duplicate Compounds Contained in Metabolic Pathway Maps: A Graph-Theoretical Perspective

Sung-Hui Kang[1], Myung-Ha Jang[1], Jiyoung Whang[1] and Hyun-Seok Park[1,2]*

[1]Department of Computer Science, Ewha Womans University, Seoul 158-711, Korea, [2]Institute of Bioinformatics, Macrogen Inc., Seoul 153-023, Korea

Abstract

The basic graph layout technique, one of many visualization techniques, deals with the problem of positioning vertices in a way to maximize some measure of desirability in a graph. The technique is becoming critically important for further development of the field of systems biology. However, applying the appropriate automatic graph layout techniques to the genomic scale flow of metabolism requires an understanding of the characteristics and patterns of duplicate and shared vertices, which is crucial for bioinformatics software developers. In this paper, we provide the results of parsing KEGG XML files from a graph-theoretical perspective, for future research in the area of automatic layout techniques in biological pathway domains.

Keywords: drawing algorithm, XML, metabolic pathway, scale-free network

Introduction

The Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway database is a valuable information resource (Kanehisa *et al*., 2000). It contains a metabolic pathway database in the form of wiring diagrams. For example, Fig. 1 shows Map00500, the manually drawn pathway map for starch and sucrose metabolism in KEGG. While this visualization style offers a good pathway presentation, it does not provide the facilities to create and visualize dynamic pathways. On the other hand, Fig. 2 was generated by an automatic layout scheme in KEGG to offer user flexibility for the same diagram in Fig. 1. It is clear that Fig. 2 differs in many aspects from Fig. 1 or the conventional drawings in biochemistry textbooks; the arrangement of its vertices and edges impacts understandability, based on aesthetics. However, a current state-of-the-art survey in this field reveals that research and development in the automatic layout of biological pathway maps are still in their infancy, at least from an aesthetic perspective.

To add more aesthetic value to automatically drawn pathway maps, it is important to understand that different layouts can correspond to the same graph. For example, both graphical representations in Fig. 1 and Fig. 2 are different from the graph itself; i.e., the abstract, non-graphical structure. At the level of software, a different format is needed for quantifying a model to the point where it can be simulated. KEGG offers the KEGG Markup Language, a machine-readable format (KGML: http://www.genome.jp/kegg/xml/) for representing models. KGML enables the automatic drawing of KEGG pathways and provides facilities for computational analysis and modeling of biological networks.

Basically, most of the existing visualization tools that are based on an automatic layout scheme are equipped with a KGML parsing module (Jeong *et al*., 2000; Becker *et al*., 2001; Klucas, 2007), and there are some publications that are related to the comparison and translation of various XML languages and parsers (Funahashi *et al*., 2004; Strömbäck *et al*., 2005; Choi *et al*., 2008). However, because they were mostly aimed at drawing only relatively small-scale drawings that unify?only several pathways, systematic analyses of shared and duplicated compounds between pathway maps were not necessary.

Thus, to the best of our knowledge, KGML analyses tools rarely have been addressed in the literature to draw a large-scale pathway, such as the KEGG Atlas from a graph-theoretical perspective. As a preliminary step in providing automatic graph layout techniques to the genome-scale flow of metabolism, analyzing KEGG XML files is crucial for software developers. Thus, in this paper, we provide shared and duplicate compound information, using our XML analyses tool, to provide valuable information for automatic layout research in the area of systems biology. These kinds of analyses that are based on graph-theoretical perspectives can be extremely useful when drawing a global pathway map in which edge crossing arises as a crucial issue.

*Corresponding author: E-mail neo@ewha.ac.kr
Tel +82-2-3277-2831, Fax +82-2-3277-2306
Accepted 8 September 2008

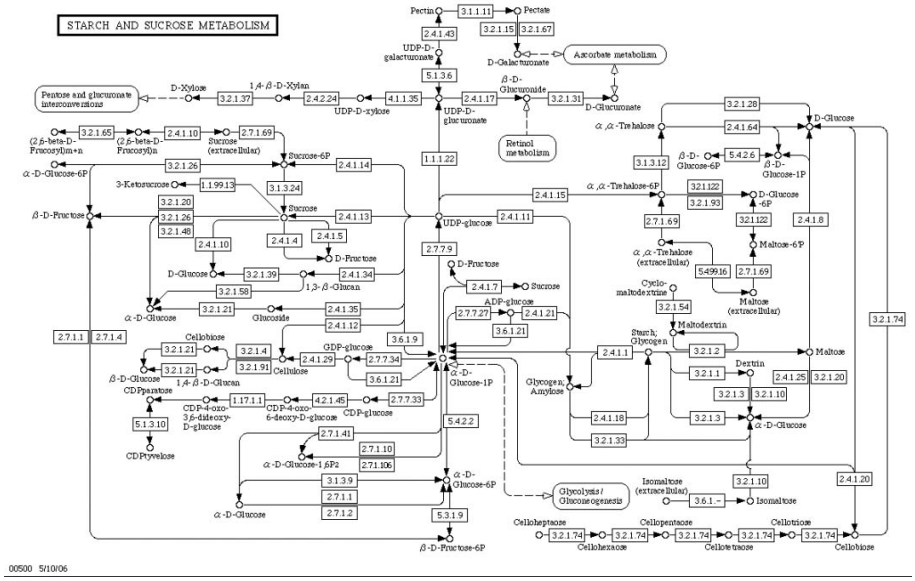

Fig. 1. Map00500 reference pathway shows the manual layout of the KEGG metabolic pathway of starch and sucrose metabolism (Source: http://www.genome.jp/dbget-bin/www_bget?pathway+map00500).

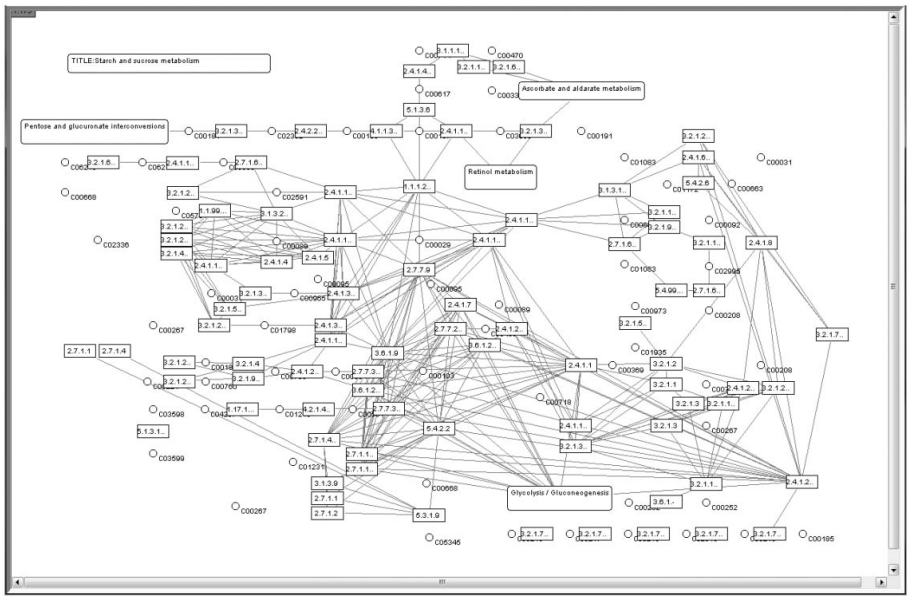

Fig. 2. An automatic layout version of KEGG reference Map-00500 of starch and sucrose metabolism (Source: http://www.genome.jp/kegg-bin/xml/PathwayViewer?-v+0.6.1+map+00500).

Implementation of the KGML Parsing Module

The Extensible Markup Language (XML) is a general-purpose specification for creating custom markup languages. It is classified as an extensible language, and its primary purpose is to help information systems share structured data. It started as a simplified subset of the Standard Generalized Markup Language (SGML). By adding semantic constraints, application languages can be implemented in XML.

The KEGG pathway, one of the representative pathway databases, adopted an XML representation of the metabolic pathway. KEGG PATHWAY is a collection of manually drawn pathway maps that represent our knowledge of molecular interaction and reaction networks (Kanehisa, 2000). The molecular reaction network is the most unique data object in KEGG, which is stored as a collection of pathway maps in the PATHWAY database. As of KEGG Release 47.0+/07-01 of July 08, 94,068 KEGG pathways have been generated from 372 reference pathways. Fig. 3 is an example of a well-formed XML document of KEGG Map00500.

One of the first things to do when dealing with XML

```
<?xml version="1.0" ?>
<!DOCTYPE pathway (View Source for full doctype...)>
<!-- Creation date: Jul 30 2008 19:30:24 +0900 (JST) -->
- <pathway name="path:map00500" org="map" number="00500" title="Starch and sucrose metabolism"
    image="http://www.genome.jp/kegg/pathway/map/map00500.gif" link="http://www.genome.jp/dbget-bin/show_pathway?
    map00500">
  - <entry id="1" name="cpd:C00092" type="compound" link="http://www.genome.jp/dbget-bin/www_bget?compound+C00092">
      <graphics name="C00092" fgcolor="#000000" bgcolor="#FFFFFF" type="circle" x="982" y="293" width="8" height="8" />
    </entry>
  - <entry id="2" name="ec:3.2.1.21" type="enzyme" reaction="rn:R02887" link="http://www.genome.jp/dbget-bin/www_bget?
    enzyme+3.2.1.21">
      <graphics name="3.2.1.21" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="236" y="520" width="45" height="17" />
    </entry>
  - <entry id="3" name="ec:3.2.1.21" type="enzyme" reaction="rn:R00026" link="http://www.genome.jp/dbget-bin/www_bget?
    enzyme+3.2.1.21">
      <graphics name="3.2.1.21" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="236" y="493" width="45" height="17" />
    </entry>
  - <entry id="4" name="ec:2.7.1.1" type="enzyme" reaction="rn:R03920" link="http://www.genome.jp/dbget-bin/www_bget?
    enzyme+2.7.1.1">
      <graphics name="2.7.1.1" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="96" y="477" width="45" height="17" />
    </entry>
  - <entry id="5" name="ec:3.2.1.122" type="enzyme" reaction="rn:R00837" link="http://www.genome.jp/dbget-bin/www_bget?
    enzyme+3.2.1.122">
      <graphics name="3.2.1.122" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="919" y="282" width="45" height="17" />
    </entry>
  - <entry id="6" name="path:map00500" type="map" link="http://www.genome.jp/dbget-bin/get_linkdb?pathway+map00500">
      <graphics name="TITLE:Starch and sucrose metabolism" fgcolor="#000000" bgcolor="#FFFFFF" type="roundrectangle"
        x="222" y="73" width="286" height="27" />
```

Fig. 3. KEGG xml file of Map00500: Each object is identified by the KEGG object identifier, consisting of a five-digit number prefixed by an upper-case letter, such as C00092 (Chemical compound: 00092) and R00026 (Reaction: 00026), or prefixed by a two-to-four-letter code for PATHWAY, such as map00500 (Source: ftp://ftp.genome.jp/pub/kegg/xml/map/map00500.xml).

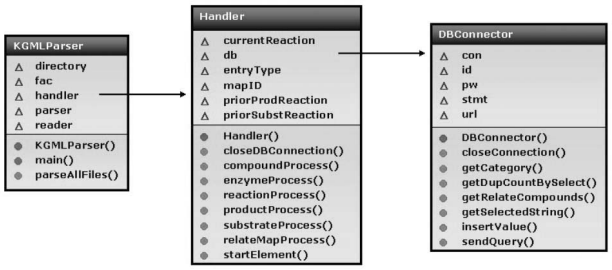

Fig. 4. KGML Parser and its class diagram: The system was developed in the Eclipse platform, and it was implemented using the Java SDK 1.5, Java.xml package, and the MySQL database, with the Tomcat application server for future use.

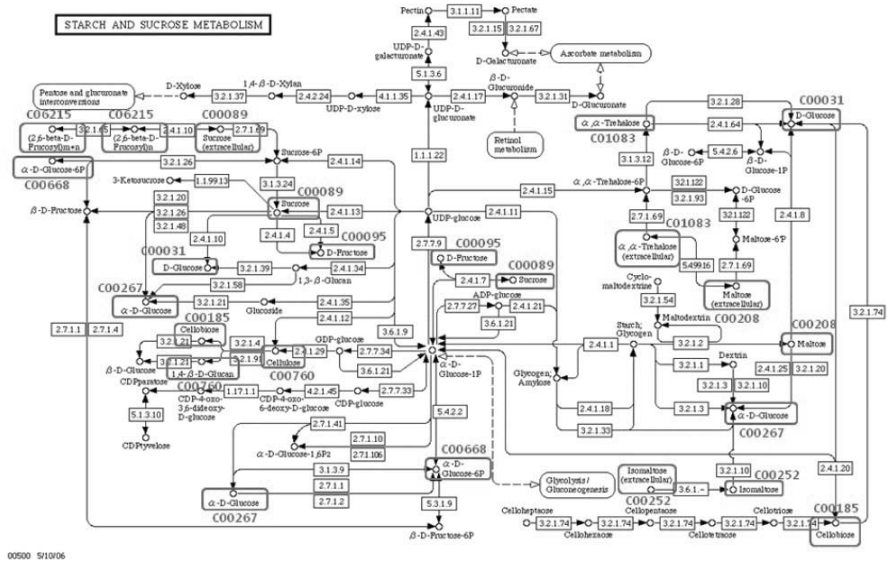

Fig. 5. Duplicate compounds in Map00500.

programmatically is to take an XML document and parse it. As the document is parsed, the data in the document become available to the application that is using the parser. The KGML Parser module-i.e., an XML processor-is used to read XML documents and provide access to their content and structure. The parsing module of KEGG XML files has been implemented using the Eclipse platform and has been deployed to Sun's J2SE Reference implementation. Three classes, KGMLParser, Handler, and DBConnector, represent the domain model of the system, as in Fig. 4, based on the MVC model. One of the biggest challenges in this line of research is to de-

velop automated and tractable techniques for ensuring the static-type safety of programs, which involves basic reasoning tasks that involve very complex constructions.

Results of Parsing XML Files

Analyzing KEGG XML files using our KGML parser gives us valuable information on automatic layout algorithms for the global layout of the metabolic pathway. Especially, the distribution of duplicate compounds and shared compounds will give us some insight into global layout schemes.

Duplicate Compounds in a Single Pathway

In the KEGG pathway database, each process or metabolite is associated with a reaction number. Because Reaction appears at most once in a single pathway graph, the reaction number that is associated with it would be sufficient as its identification in the layout information. However, a compound could be appearing in several places in a single pathway. For example, C06215 appears twice in the Starch and Sucrose Metabolism Map, as in Fig. 1 (Fig. 5 for all of the duplicate compounds). We counted the number of compounds and enzymes that had the same ecNum and rID, and they are represented by 4112 and 5194 vertices, respectively. Thus, we have a total of 9306 vertices in the KEGG metabolic network, while 1818 vertices, around 9.25%, are shared by multiple pathways. Among the compounds, there are 528 compounds that appear more than once in a pathway. Also, 333 enzymes ap-

Table 1. Duplicate compounds list in Map00500

mapID	Title	Duplicate compound	# of Duplication
path:map00500	Starch and sucrose metabolism	cpd:C00031	2
path:map00500	Starch and sucrose metabolism	cpd:C00089	3
path:map00500	Starch and sucrose metabolism	cpd:C00095	2
path:map00500	Starch and sucrose metabolism	cpd:C00185	2
path:map00500	Starch and sucrose metabolism	cpd:C00208	2
path:map00500	Starch and sucrose metabolism	cpd:C00252	2
path:map00500	Starch and sucrose metabolism	cpd:C00267	3
path:map00500	Starch and sucrose metabolism	cpd:C00668	2
path:map00500	Starch and sucrose metabolism	cpd:C00760	2
path:map00500	Starch and sucrose metabolism	cpd:C01083	2
path:map00500	Starch and sucrose metabolism	cpd:C06215	2

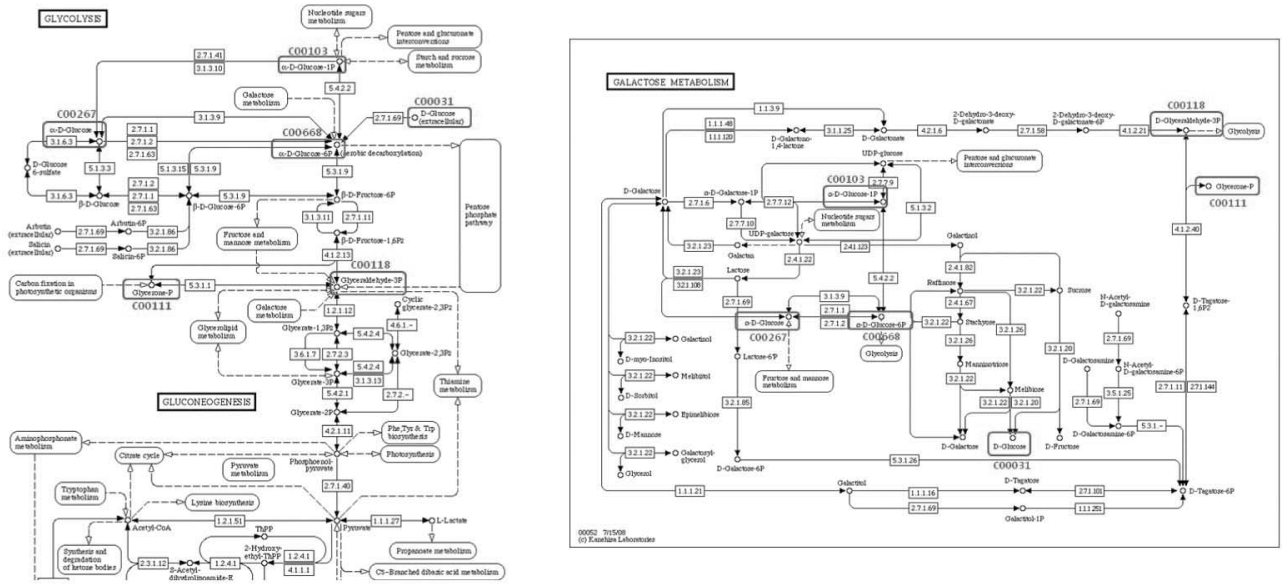

Fig. 6. Shared Compounds Appearing in Two Pathways: Map00010 and Map00052 share six compounds: C00031, C00103, C00111, C00118, C00267, and C00668.

pear more than once in multiple pathways.

For example, Table 1 is the parsing result of Map00500 of starch and sucrose metabolism, corresponding to Fig. 5, and there are 11 compounds that appear more than once: C00031, C00089, C00095, C00185, C00208, C00252, C00267, C00668, C00760, C01083, and C06215.

Shared Compounds between Two Pathways

Parsing information that is related only to a single pathway is not sufficient. Metabolic studies, used to being dedicated to a single pathway, recently have begun to focus on the entire network (Jeong *et al*., 2000). For example, KEGG, the representative biological pathway database, now provides the KEGG Metabolism Atlas (Okuda *et al*., 2008) by manually combining existing metabolic pathway maps.

However, when visualizing a metabolism atlas, based on an automatic layout scheme, analyzing the shared vertices between pathways is critically important for aesthetic visualization of multiple pathways. For example, note that the two metabolic pathways (Map00010 & Map00052: glycolysis and galactose metabolism) share the same vertices, as in Fig. 6. In our initial attempts to display global pathways, these shared vertices are unified. Otherwise, a reaction cascade that is shared by two pathways can be represented twice, appearing in different areas of the drawing. However, when multiple pathways are sharing more than one node, calculating the relative positioning between them with the highest aesthetic value for the shared node becomes a very hard problem.

Thus, information with regard to shared compounds in KEGG pathway map pairs influences a graph layout algorithm. There are 876 vertices, which accounts for 55% of all shared vertices, shared by only two pathways. In other words, among the 528 duplicate compound vertices that appear more than once, 348 compounds appear only in two pathways. Also, among the 333 enzymes that appear more than once, 223 enzymes appear only in two pathways. For example, Fig. 6 shows that six compounds, including C00031, C00103, C00111, C00118, C00267, and C00668, are shared between Map00010 and Map00052.

Table 2 is the complete list of duplicate compounds that appear in the KEGG pathways, generated by our KGML Parser. There are 113 compounds that are duplicated more than once, and among all 1777 pathway pairs that have shared vertices, 1097 (60.73%) of them are sharing only one vertex. Hence, solving the problem in this simple situation would solve a large percentage of layout problems in visualizing graphs. Some of the biggest numbers of shared vertices between two pathway map pairs were: 27 vertices between Map00062 and Map00071, 17 vertices between Map00040 and Map00053, and 14 vertices between Map 00020 and Map00720. This is valuable information to consider, with regard to automatic layout algorithms.

Table 2. Complete list of duplicate compounds

mapID	# of Nodes	Duplicate Compounds	Total Duplications
path:map00020	51	2	4
path:map00030	77	1	3
path:map00051	121	5	10
path:map00052	96	2	4
path:map00500	144	11	24
path:map00520	76	1	2
path:map00530	81	3	6
path:map00630	112	1	2
path:map00190	16	2	12
path:map00195	11	1	6
path:map00680	66	2	4
path:map00710	55	6	20
path:map00720	30	1	4
path:map00910	96	2	4
path:map00071	136	2	8
path:map00100	148	1	2
path:map00140	120	1	2
path:map00561	77	2	5
path:map00564	113	3	9
path:map00591	41	2	4
path:map01040	87	16	38
path:map00230	260	9	21
path:map00240	181	5	14
path:map00272	55	3	6
path:map00280	93	3	9
path:map00290	61	1	3
path:map00310	114	2	4
path:map00330	119	1	2
path:map00350	203	4	9
path:map00380	158	1	2
path:map00450	53	1	2
path:map00480	53	2	4
path:map00540	45	1	2
path:map00550	63	11	28
path:map00563	25	2	4
path:map00522	77	5	15
path:map00790	80	1	2
path:map00860	187	2	4
path:map00312	2	1	2
path:map00401	43	2	4
path:map00906	152	1	2
path:map00908	46	3	14
path:map00362	113	1	2
path:map00622	90	2	4
path:map00982	136	2	4
	4,163	133	336

Discussion & Future Direction

Because metabolic studies have been focusing on the notion of the pathway, existing visualization algorithms have been mostly dedicated to visualization of a single pathway, except for a few attempts to deal with a small number of multi-pathways (Jeong *et al*., 2000; Becker *et al*., 2001; Klucas, 2007). Using such algorithms does not enable the study of cascades of reactions that span a large number of pathways. Moreover, directly applying these algorithms to network visualization is not possible, because they would disregard distribution patterns of duplicate compounds that appear in several pathways.

When multi-pathways are sharing more than one node, calculating the relative positioning between them becomes a very hard problem. Our initial attempts to visualize all the metabolic pathways automatically in a single atlas map resulted in a confusing diagram that was difficult to interpret (Song *et al*., 2008). Therefore, building an XML parser to acquire statistical data that are related to duplicate and shared nodes in pathway maps was necessary, presenting the hopes of presenting novel algorithms for multi-pathway maps. However, it seems extremely difficult to present a new generic method to do this for any number of shared nodes. We could approximate the relative positioning of the shared nodes, provided that the shared nodes are near one side of the boundary, when two pathways are sharing multiple nodes, as was suggested by Wang' s algorithm (Wang, 2008). Or, we could just place the shared nodes in the extra space between multi-pathways.

There could be various tactics and algorithms. Unfortunately, the scope of this paper does not address solid algorithms for the automatic layout of multiple metabolic pathways. Rather, it simply shows statistics with regard to biological pathways as a preliminary step in providing automatic layout algorithms in the future. Without the analysis of KGML, visualizing all of the pathways globally in a single atlas map generally would result in a confusing diagram.

Based on these concepts and caveats, there are different graph layout strategies, and it is important to formally specify, implement, and verify such algorithms in the future.

Acknowledgments

We would like to thank Dr. Woong-Yang Park for valuable comments. This work was partially supported by a grant from the Ministry of Information and Communication of Korea.

References

Becker, M.Y., and Rojas, I. (2001). A Graph Layout Algorithm for Drawing Metabolic Pathways, *Bioinformatics* 17, 461-467.

Choi, K.M., and Kim, S. (2008). comparative enzyme analysis and annotation in pathway/subsystem contexts. *BMC Bioinformatics* 9, 145.

Funahashi, A., Jouraku, A., and Kitano, H. (2004). Converting KEGG pathway database to SBML. *8th Annual International Conference on Research in Computational Molecular Biology.*

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabasi, A.L. (2000). The Large-scale Organization of Metabolic Networks. *Nature* 407, 651-654.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28, 27-30.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*. 36, W423-W426.

Klukas, C., and Schreiber, F. (2007). Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics* 23, 344-50.

Song, E.H., Kim, M.K., and Lee, S.H. (2006). A Metabolic Pathway Drawing Algorithm for Reducing the Number of Edge Crossings. *Genomics & Informatics* 4, 118-124.

Song, E.H., Ham, S.I., Yang, S.D., Rhie, A., Park, H.S., and Lee, S.H. (2008). J2pathway: A Global Metabolic Pathway Viewer with Node Abstracting Features. *Genomics & Informatics* 6, 118-124.

Strömbäck, L., and Lambri, P. (2005). Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics* 21, 4401-4407.

Wang, Y. (2008). Familiar Layouts Generation for Metabolic Pathway Graph Visualization. MS Thesis, Case Western Reserve University, Computing and Information Science.