APPLICATION NOTE Genomics & Informatics Vol. 8(2) 94-96, June 2010

Java DOM Parsers to Convert KGML into SBML and BioPAX Common Exchange Formats

Kyung-Eun Lee, Myung-Ha Jang, Arang Rhie, Chin Ting Thong, Sanduk Yang and Hyun-Seok Park*

Department of Computer Science, Ewha Womans University, Seoul 120-750, Korea

Abstract

Integrating various pathway data collections to create new biological knowledge is a challenge, for which novel computational tools play a key role. For this purpose, we developed the Java-based conversion modules KGML2SBML and KGML2BioPAX to translate KGML (KEGG Markup Language) into a couple of common data exchange formats: SBML (Systems Biology Markup Language) and BioPAX (Biological Pathway Exchange). We hope that our work will be beneficial for other Java developers when they extend their bioinformatics system into SBML- or BioPAX-aware analysis tools. This is part of our ongoing effort to develop an ultimate KEGG-based pathway enrichment analysis system.

Availability: You may download the conversion module from Sourceforge.net (https://sourceforge.net/projects/ngssequencealig/). The feedback from developing the conversion tools will provide valuable feedback for the continued evolution of a universal exchange format.

Keywords: systems biology, visualization, common exchange format

Introduction

Recently, various frameworks for the XML formats of biological network information were proposed (Achard *et al.*, 2001; Finney *et al.*, 2006). Among them, two approaches have gained popularity as promising candidates to be adopted as future standards in the field: SBML and BioPAX (Hucka *et al.*, 2003; Strömbäck *et al.*, 2005). On the other hand, we have been working on a KEGG-based pathway analysis system in past years (Jang, 2010; Ham, 2007). Unfortunately, KGML (Kanehisa and Goto, 2000) is not a representative common pathway exchange format. The KEGG site (http://www.kegg.org/) provides metabolic pathway data, not only in KGML format but also in BioPAX (http://www.biopax.org/) Level 1 and SBML (http://www.sbml.org/) format. Still, the KEGG site itself does not offer conversion tools. We were able to find only two conversion tools, called KEGG2SBML (http://sbml.org/Software/KEGG2SBML) and KEGGconverter (http://www.grissom.gr/Keggconverter). Both KEGG2SBML (Funahashi, 2004) and KEGGconverter (Moutselos *et al*, 2009) deal only with SBML.

Thus, we decided to develop Java-based KGML2SBML and KGML2BioPAX parsers with a graphical user interface. These parsers not only convert KGML into SBML but also convert KGML into BioPAX.

Overview of the BioPAX and SBML standards

BioPAX and SBML are becoming one of the most widely used common data exchange formats that can be widely used to simulate and visualize biological networks (Strömbäck *et al.*, 2005). Both BioPAX (Level 3) and SBML (Level 1) can encode signaling pathways, metabolic pathways, and regulatory pathways, although SBML can represent finer details (Fig. 1). However, they

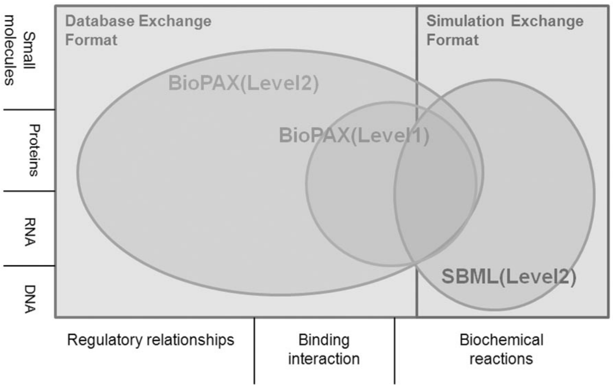


Fig. 1. SBML (Systems Biology Markup Language) for simulation model exchange formats, and BioPAX (Biological Pathway Exchange) for database exchange format. Each SBML Level is intended to describe a biochemical network or a pathway. BioPAX is defined in a number of steps, called levels. Level 1 focuses on metabolic networks, and Level 2 adds molecular interaction networks.

*Corresponding author: E-mail neo@ewha.ac.kr
Tel +82-2-3277-2831, Fax +82-2-3277-2306
Accepted 2 June 2010

are not competing standards. Rather, they are complementary approaches to describe related phenomena.

SBML is targeted at modeling systems of dynamic biochemical reaction networks. BioPAX is an RDF (Resource Description Framework) and OWL-based standard and assumes a static view of the network model. BioPAX does not contain mathematical formulas but provides more detailed information concerning the individual molecules and interactions. If one's aim is simulation and computational analysis, then SBML has better API supports than BioPAX. However, BioPAX is a rich standard when it comes to encoding the interactions, such as genetics.

Unfortunately, BioPAX and SBML, the two most widespread standards for storing pathway data, are semantically different. Conversion modules between KGML and these two formats may lead to loss or distortion of data. This seriously limits the automatic integration of models. To obtain sufficient generality in the model, many details have to be omitted in the model hierarchy.

Implementation

We implemented the KGML2SBML and KGML2BioPAX parsers, which convert KGML into SBML Level 2, version 4 documents, and BioPAX. These two parsers are implemented with the Java DOM (Document Object Model) parser and require Java APIs for XML Processing (http://java.sun.com/xml/jaxp/index.html). The mappings between KGML and these two formats are sum-

(a) Mappings between KGML and SBML

(b) Mappings between KGML and BioPAX

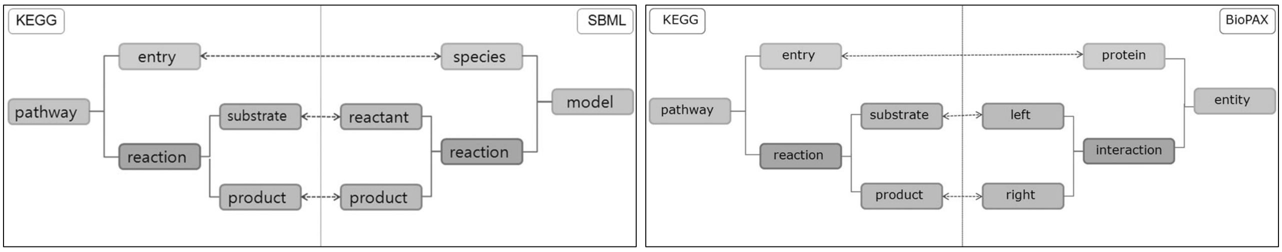
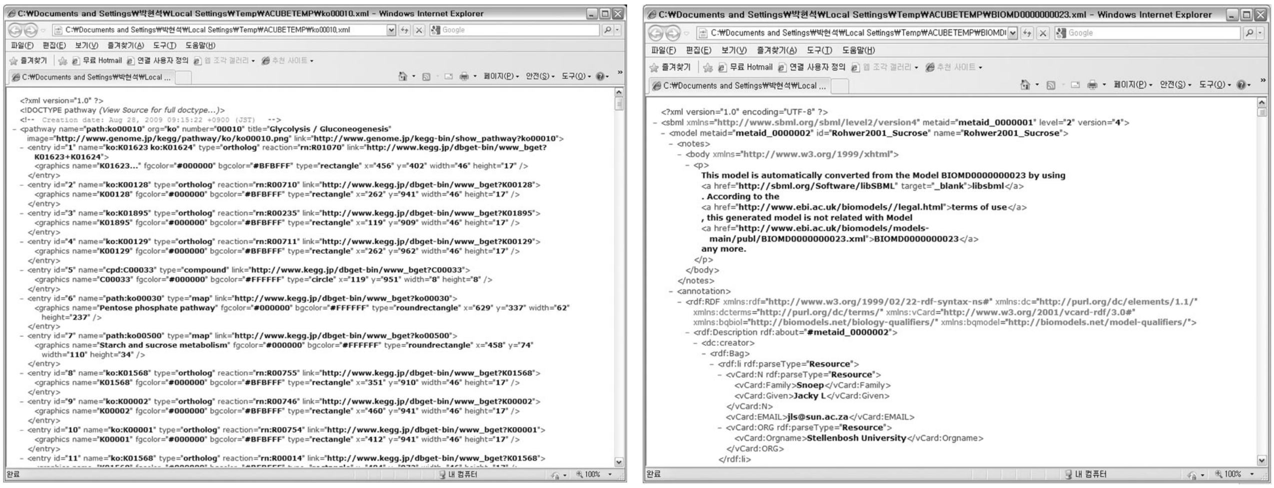


Fig. 2. The KEGG Pathway element is a root element, and one pathway element is specified for one pathway map in KGML. The entry, Relation, and Reaction elements specify the graph information. SBML is defined in terms of the participating Reactants and Products, along with optional modifier Species, an optional Kinetic Law. The basic building blocks of a BioPAX model are the Physical Entity, Interaction, and Pathway classes. Each Interaction describes relationships between Physical Entities.


C:₩Documents and Settings₩박현석₩Local Settings₩Temp₩ACUBETEMP₩ko00010.xml - Windows Internet Explorer
C:₩Documents and Settings₩박현석₩Local Settings₩Temp₩ACUBETEMP₩ko00010.xml
Google
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
즐겨찾기 무료 Hotmail 연결 사용자 정의 웹 조각 갤러리 추천 사이트
C:₩Documents and Settings₩박현석₩Local ...
페이지(P) 안전(S) 도구(O)
완료 내 컴퓨터 100%
C:₩Documents and Settings₩박현석₩Local Settings₩Temp₩ACUBETEMP₩BIOMD0000000023.xml - Windows Internet Explorer
C:₩Documents and Settings₩박현석₩Local Settings₩Temp₩ACUBETEMP₩BIOMD
Google
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
즐겨찾기 무료 Hotmail 연결 사용자 정의 웹 조각 갤러리 추천 사이트
C:₩Documents and Settings₩박현석₩Local ...
페이지(P) 안전(S) 도구(O)
완료 내 컴퓨터 100%

(a) The KEGG Markup Language (KGML) is an exchange format of the KEGG graph objects, especially the KEGG pathway maps. The detailed format is described in http://www.genome.jp/kegg/xml/.

```
  <?xml version="1.0" ?>
  <!DOCTYPE pathway (View Source for full doctype...)>
  <!-- Creation date: Aug 28, 2009 09:15:22 +0900 (JST) -->
- <pathway name="path:ko00010" org="ko" number="00010" title="Glycolysis / Gluconeogenesis"
   image="http://www.genome.jp/kegg/pathway/ko/ko00010.png" link="http://www.genome.jp/kegg-bin/show_pathway?ko00010">
 - <entry id="1" name="ko:K01623 ko:K01624" type="ortholog" reaction="rn:R01070" link="http://www.kegg.jp/dbget-bin/www_bget?
     K01623+K01624">
     <graphics name="K01623..." fgcolor="#000000" bgcolor="#BFBFFF" type="rectangle" x="456" y="402" width="46" height="17" />
   </entry>
 - <entry id="2" name="ko:K00128" type="ortholog" reaction="rn:R00710" link="http://www.kegg.jp/dbget-bin/www_bget?K00128">
     <graphics name="K00128" fgcolor="#000000" bgcolor="#BFBFFF" type="rectangle" x="262" y="941" width="46" height="17" />
   </entry>
 - <entry id="3" name="ko:K01895" type="ortholog" reaction="rn:R00235" link="http://www.kegg.jp/dbget-bin/www_bget?K01895">
     <graphics name="K01895" fgcolor="#000000" bgcolor="#BFBFFF" type="rectangle" x="119" y="909" width="46" height="17" />
   </entry>
 - <entry id="4" name="ko:K00129" type="ortholog" reaction="rn:R00711" link="http://www.kegg.jp/dbget-bin/www_bget?K00129">
     <graphics name="K00129" fgcolor="#000000" bgcolor="#BFBFFF" type="rectangle" x="262" y="962" width="46" height="17" />
   </entry>
 - <entry id="5" name="cpd:C00033" type="compound" link="http://www.kegg.jp/dbget-bin/www_bget?C00033">
     <graphics name="C00033" fgcolor="#000000" bgcolor="#FFFFFF" type="circle" x="119" y="951" width="8" height="8" />
   </entry>
 - <entry id="6" name="path:ko00030" type="map" link="http://www.kegg.jp/dbget-bin/www_bget?ko00030">
     <graphics name="Pentose phosphate pathway" fgcolor="#000000" bgcolor="#FFFFFF" type="roundrectangle" x="629" y="337" width="62"
       height="237" />
   </entry>
 - <entry id="7" name="path:ko00500" type="map" link="http://www.kegg.jp/dbget-bin/www_bget?ko00500">
     <graphics name="Starch and sucrose metabolism" fgcolor="#000000" bgcolor="#FFFFFF" type="roundrectangle" x="458" y="74"
       width="110" height="34" />
   </entry>
 - <entry id="8" name="ko:K01568" type="ortholog" reaction="rn:R00755" link="http://www.kegg.jp/dbget-bin/www_bget?K01568">
     <graphics name="K01568" fgcolor="#000000" bgcolor="#BFBFFF" type="rectangle" x="351" y="910" width="46" height="17" />
   </entry>
 - <entry id="9" name="ko:K00002" type="ortholog" reaction="rn:R00746" link="http://www.kegg.jp/dbget-bin/www_bget?K00002">
     <graphics name="K00002" fgcolor="#000000" bgcolor="#BFBFFF" type="rectangle" x="460" y="941" width="46" height="17" />
   </entry>
 - <entry id="10" name="ko:K00001" type="ortholog" reaction="rn:R00754" link="http://www.kegg.jp/dbget-bin/www_bget?K00001">
     <graphics name="K00001" fgcolor="#000000" bgcolor="#BFBFFF" type="rectangle" x="412" y="941" width="46" height="17" />
   </entry>
 - <entry id="11" name="ko:K01568" type="ortholog" reaction="rn:R00014" link="http://www.kegg.jp/dbget-bin/www_bget?K01568">
```

(b) A definition of a model in SBML consists of lists of one or more of these various components: *Compartment, Species, Reaction, Parameter, Unit definition*, and *Rule*. The detailed format is described in http://sbml.org/Documents/Specifications.

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <sbml xmlns="http://www.sbml.org/sbml/level2/version4" metaid="metaid_0000001" level="2" version="4">
 - <model metaid="metaid_0000002" id="Rohwer2001_Sucrose" name="Rohwer2001_Sucrose">
  - <notes>
   - <body xmlns="http://www.w3.org/1999/xhtml">
    - <p>
        This model is automatically converted from the Model BIOMD0000000023 by using
        <a href="http://sbml.org/Software/libSBML" target="_blank">libsbml</a>
        . According to the
        <a href="http://www.ebi.ac.uk/biomodels//legal.html">terms of use</a>
        , this generated model is not related with Model
        <a href="http://www.ebi.ac.uk/biomodels/models-
          main/publ/BIOMD0000000023.xml">BIOMD0000000023</a>
        any more.
      </p>
     </body>
    </notes>
  - <annotation>
   - <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dc="http://purl.org/dc/elements/1.1/"
       xmlns:dcterms="http://purl.org/dc/terms/" xmlns:vCard="http://www.w3.org/2001/vcard-rdf/3.0#"
       xmlns:bqbiol="http://biomodels.net/biology-qualifiers/" xmlns:bqmodel="http://biomodels.net/model-qualifiers/">
    - <rdf:Description rdf:about="#metaid_0000002">
     - <dc:creator>
      - <rdf:Bag>
       - <rdf:li rdf:parseType="Resource">
        - <vCard:N rdf:parseType="Resource">
            <vCard:Family>Snoep</vCard:Family>
            <vCard:Given>Jacky L</vCard:Given>
          </vCard:N>
          <vCard:EMAIL>jls@sun.ac.za</vCard:EMAIL>
        - <vCard:ORG rdf:parseType="Resource">
            <vCard:Orgname>Stellenbosh University</vCard:Orgname>
          </vCard:ORG>
         </rdf:li>
```

Fig. 3. An exemplary conversion of KGML into SBML.

marized in Fig. 2. KGML2SBML uses the LIGAND database of KEGG as input to generate SBML documents. For example, *KineticLaw* in reaction and specific names of compounds can be added in a converted SBML file with the use of KEGG Ligand information.

An exemplary conversion between KGML and SBML is shown in Fig. 3. Conversion between KGML and BioPAX has been done in a similar fashion. The online SBML Validator (http://sbml.org/Facilities/Validator) has been used to check the syntax and internal consistency for converted SBML documents. Also, BioPAX Validator (http://www.biopax.org/biopax-validator/index.html) has been used to check converted BioPAX documents and fix syntatic and semantic errors. All converted SBML and BioPAX documents are available (https://sourceforge.net/projects/ngssequencealig/).

Conclusion

Most current simulation and analysis packages already support SBML and BioPAX, and more are in the process of being extended to support them. Also, most of the top journals in the bioinformatics field encourage authors to prepare models of biochemical reaction networks using common exchange formats and to deposit the model into the BioModels database. For this reason, we developed simple Java utility tools for converting KGML to SBML and BioPAX. Making the databases available in common pathway exchange formats would be a first step toward building an open source pathway analysis resource.

Acknowledgments

This software was developed thanks to generous funding from the SMBA Industrial-Educational Cooperation Project (000358780109) of the Korean government.

References

Achard, F., Vaysseix, G., and Barillot, E. (2001). XML, bioinformatics and data integration. *Bioinformatics* 17, 115-125.

Finney, A., Hucka, M., Bornstein, B.J., Keating, S.M., Shapiro, B.E., Matthews, J., Kovitz, B.L., Schilstra, M.J., Funahashi, A., Doyle, J.C., and Kitano, H. (2006). Software infrastructure for effective communication and reuse of computational models. Systems modeling in cell biology: from concepts to nuts and bolts. *MIT Press.* pp. 369-378.

Funahashi, A., Jyoraku, A., and Kitano, H. (2004). Converting the KEGG Pathway Database to SBML. *5th Int. Conf. Syst. Biol. (ICSB2004).*

Ham, S.I., Song, E.H., Yang, S.D., Thong, C.T., Rhie, A., Galbadrakh, B., Lee, K.E., Park, H.S., and Lee, S.H. (2007). J2.5dPathway: a 2.5D visualization tool to display selected nodes in biological pathways, in parallel planes. *Genomics Inform.* 7, 171-174.

Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., and Kitano, H. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524-531.

Jang, M., Whang, J.W., Lewis, E., and Park, H.S. (2010). An interpretation of biological metabolites and their reactions based on relation degree of compound pairs in KEGG XML Files. *JSW.* 5, 187-194.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* 28, 27-30.

Moutselos, K., Kanaris, I., Chatziioannou, A., Maglogiannis, I., and Kolisis, F. (2009). KEGG converter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database. *BMC Bioinformatics* 10, 324.

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