KEGGtranslator: visualizing and translating the KEGG pathway database

Clemens Wrzodek [1,*], Andreas Dräger [1] and Andreas Zell [1*]

[1]Center for Bioinformatics Tübingen (ZBIT), University of Tübingen, Sand 1, 72076 Tübingen, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: The KEGG PATHWAY database is today a widely used service for pathway based information. It contains manually drawn pathway maps with information about the genes, reactions and relations contained within a pathway. To model these pathways, KEGG is using an own XML-format. Parsers and translators are needed to process the pathway maps for usage in other applications and algorithms.

We have developed KEGGtranslator, which is an easy-to-use stand-alone application that can translate KGML formatted XML-files to multiple output formats. Unlike other translators, KEGGtranslator supports a plethora of output formats, is able to complete the information in translated documents (i.e. MIRIAM annotations) beyond the scope of the XML-document and amends missing components to fragmentary reactions within the pathway to allow simulations on those.

Availability: KEGGtranslator is freely available as a webstart application and for download at `http://www.ra.cs.uni-tuebingen.de/software/KEGGtranslator/`. KGML files can be downloaded within the application of manually from `ftp://ftp.genome.jp/pub/kegg/xml/kgml`.

Contact: Clemens.Wrzodek@uni-tuebingen.de

1 INTRODUCTION

Many academic researchers, who want to use pathway based information are using the KEGG PATHWAY database [6]. The database, established in 1995, contains maps for various pathways and is strongly related to the other KEGG databases, by assigning all elements KEGG identifier only. These maps are visualized on the web and can be downloaded without charge (for academics) as xml-files that are using the KEGG Markup Language (KGML) [2].

However, the content of these KGML-formatted xml-files is limited. Gene names are often not very readable and elements are not well annotated. The reason for this is that KEGG releases those xml-files primarily for graphical visualizations of the pathway. For creating fully functional models of the pathway, the content of the xml-file is not sufficient.

KEGGtranslator reads and completes the content of the XML-file with live-annotation of all genes and reactions using the KEGG API [1]. Minor deficiencies are corrected (i.e. the name of a gene), new information is being added (i.e. multiple MIRIAM identifier for each gene and reaction, or SBO terms describing the function) and some crucial deficiencies are addressed:

KEGG uses colors in the visualization of a pathway to annotate organism specific orthologous genes. Nodes in green represent biological entities that occur in the current organism. Nodes in white represent biological entities, that correspond to genes that are occurring in this pathway in other species, but not in the current one. Translating all those nodes into new models, without caring for the node color, would lead to a model, containing invalid genes in the pathway.

Another major deficiency are the reactions. The XML-files only contain those parts of the reaction, that are needed for the graphical representation of the pathway. But they do not always contain the complete chemical equation (see figure 1). KEGGtranslator is able to lookup each reaction and amend the missing components to reactions. This leads to more complete and functional correct pathway models.

To our knowledge, these major deficiencies are not being corrected by other KEGG converters.

2 TRANSLATION OF KGML-FILES

In the first step of a translation, KEGGtranslator reads the XML-file and puts all contained elements into an internal data structure. In addition to the XML-file, the KEGG database is queries via the KEGG API for each id in the document (pathway id, entries, reactions, relations, substrates, products, etc.). This completes the sparse XML-document with comprehensive information. For example, multiple synonyms, identifier of many external databases (Ensembl, EntrezGene, UniProt, ChEBI, Gene Ontology, DrugBank, PDBeChem, and many more) are being assigned to genes.

After this initial step, various preprocessing operations are performed on the pathway. The user may choose if he wants to remove white nodes (genes, that are not occurring in the current organism), remove orphans (nodes without any reactions or relations) and if he wants to let KEGGtranslator autocomplete reactions.

After these preprocessing steps, KEGGtranslator branches between two different conversion modes for the actual translation: a functional translation (SBML) and a graphical translation (e.g. GraphML, GML). Depending on the chosen output format, KEGG-translator determines the way to translate the KGML document.

[*]to whom correspondence should be addressed

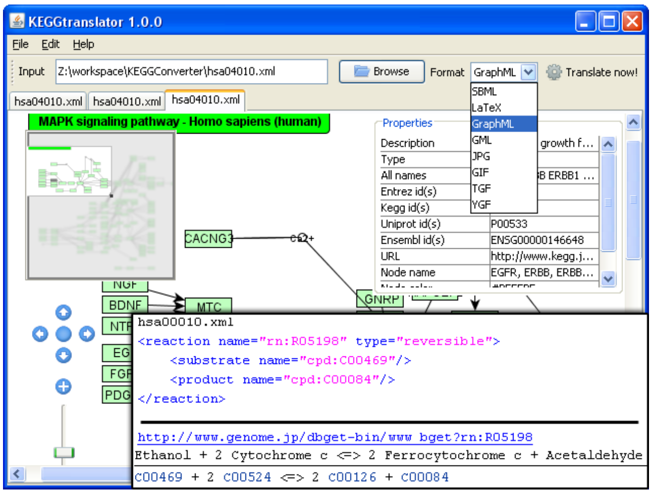
The functional translation is performed by converting the KGML code to a jSBML data structure (TODO:JSBMLCitation). The focus here is to generate valid and specification-conform SBML (Level 2 Version 4) code that eases e.g. a dynamic simulation of the pathway. Each entry (pathway references, genes, compounds, enzymes, reactions, reaction-modifiers, etc.) is being assigned multiple MIRIAM-URNs and an SBO-Term, which describes best the function of the element. Additionally, notes are being assigned to each element with human-readable names and synonyms, a description of the element and links to pictures and further information. The user may also choose to add graphical information by putting CellDesigner annotations to the model. The key difference in this mode, to the graphical translation, is that the focus lies on reactions in KGML documents. Besides the already mentioned completion of reactions, each enzymatic modifier is correctly assigned to the reaction by building so called modifierSpeciesReferences. The reversibility of the reaction is annotated and a picture of the reaction equation is being put into the notes. As a final step, we have integrated the SBML2LaTeX [4] tool into KEGGtranslator, which allows users to automatically generate a LaTeX or PDF-report to document the SBML-code of the translated pathway. Kinetics may be added by using the SBMLSqueezer [3] tool after the translation.

In graphical translations, results can be saved as GraphML, GML or YGF and finally as images of type JPG, GIF, or TGF. In this mode, the KGML data structure is being converted to a yFiles [8] data structure. The focus here lies on the visualization of the pathway. Relations are being translated by inserting arrows with the appropriate style, which is given in the KGML document. For example, dashed arrows without heads represent bindings or associations and a dotted arrow with a simply head illustrates an indirect effect. Please see the KGML specification [2] for a complete list. As in the functional translation, GraphML allows to define custom annotation elements. KEGGtranslator makes use of those, by putting several identifiers (e.g. EntrezGene or Ensembl) and descriptions to the single nodes. From the KGML-document, the shape of the node is translated as well as the colors and labels. Links to descriptive HTML-pages are being setup and hierarchical group nodes are being created for defined compounds. All these features lead to a nice graphical representation of the pathway that provides as many information about the elements as possible.

3 DISCUSSION

KEGGtranslator is a stand-alone application with a graphical user interface, that runs on every operating system that has a java virtual machine. There are tools for converting KGML to SBML (e.g. KEGGconverter [7] and KEGG2SBML [5]) and for converting KGML to graph structures in R (KEGGgraph [9]). But, to our knowledge, no stand-alone KEGG converter exists that is able to translate KGML formatted files to various output formats. In contrast to all other KEGG converting tools, KEGGtranslator comes with important functionalities like the autocompletion of reactions or the annotation of each element of the translated file, using various identifiers.

Because of the variety of output formats, the pathways converted with KEGGtranslator can easily be opened in other applications. Further processing steps can be done quickly because a node or edge has many more information and identifier that just a name. And dynamic simulations on pathways can be done accurate with complete reactions.



Input: Z:\workspace\KEGGConverter\hsa04010.xml

Format: GraphML

Properties	
Description	growth f...
Type	
All names	B ERBB1 ...
Entrez id(s)	
Kegg id(s)	
Uniprot id(s)	P00533
Ensembl id(s)	ENSG00000146648
URL	http://www.kegg.j...
Node name	EGFR, ERBB, ERBB...

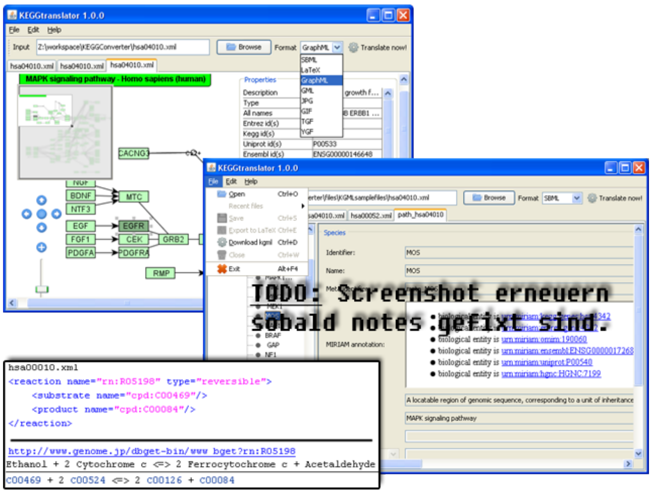
```
hsa00010.xml
<reaction name="rn:R05198" type="reversible">
    <substrate name="cpd:C00469"/>
    <product name="cpd:C00084"/>
</reaction>
```

http://www.genome.jp/dbget-bin/www bget?rn:R05198

Ethanol + 2 Cytochrome c <=> 2 Ferrocytochrome c + Acetaldehyde

C00469 + 2 C00524 <=> 2 C00126 + C00084

Fig. 1. Screenshot of KEGGtranslator. The list of all output formats is shown in the upper right corner. On the center, a translated GraphML pathway is displayed. The lower right corner demonstrates the need for autocompleting reactions: on the upper half one can see the KGML-file with only one substrate and product. On the lower half, the complete reaction equation is shown. As one can see, one substrate and product is missing in the XML-document.



Input: Z:\workspace\KEGGConverter\hsa04010.xml

Format: GraphML

Properties	
Description	growth f...
Type	
All names	B ERBB1 ...
Entrez id(s)	
Kegg id(s)	
Uniprot id(s)	P00533
Ensembl id(s)	ENSG00000146648

...erter\files\KGMLsamplefiles\hsa04010.xml

Format: SBML

Identifier: MOS

Name: MOS

MIRIAM annotation:

- biological entity is urn:miriam:omim:190060
- biological entity is urn:miriam:ensembl:ENSG0000017268
- biological entity is urn:miriam:uniprot:P00540
- biological entity is urn:miriam:hgnc:HGNC:7199

A locatable region of genomic sequence, corresponding to a unit of inheritance

MAPK signaling pathway

TODO: Screenshot erneuern sobald notes gefixt sind.

```
hsa00010.xml
<reaction name="rn:R05198" type="reversible">
    <substrate name="cpd:C00469"/>
    <product name="cpd:C00084"/>
</reaction>
```

http://www.genome.jp/dbget-bin/www bget?rn:R05198

Ethanol + 2 Cytochrome c <=> 2 Ferrocytochrome c + Acetaldehyde

C00469 + 2 C00524 <=> 2 C00126 + C00084

Fig. 2. Alternative picture for Fig. 1. We have to choose one of those.

ACKNOWLEDGEMENT

Text Text Text Text Text Text Text Text.

Funding: Text Text Text Text Text Text Text Text.

REFERENCES

[1]*KEGG API [http://www.genome.jp/kegg/soap/].*

[2]*KEGG Markup Language [http://www.genome.jp/kegg/xml/docs/].*

[3]A. Dräger, N. Hassis, J. Supper, A. Schröder, and A. Zell. SBMLsqueezer: a CellDesigner plug-in to generate kinetic rate equations for biochemical networks. *BMC Systems Biology*, 2(1):39, Apr. 2008.

[4]A. Dräger, H. Planatscher, D. M. Wouamba, A. Schröder, M. Hucka, L. Endler, M. Golebiewski, W. Müller, and A. Zell. SBML2LaTeX: Conversion of SBML files into human-readable reports. *Bioinformatics*, 25(11):1455–1456, Apr. 2009.

[5]A. Funahashi, A. Jouraku, and H. Kitano. Converting kegg pathway database to sbml. *8th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004).*

[6]M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

[7]K. Moutselos, I. Kanaris, A. Chatziioannou, I. Maglogiannis, and F. N. Kolisis. Keggconverter: a tool for the in-silico modelling of metabolic networks of the kegg pathways database. *BMC Bioinformatics*, 10:324, 2009.

[8]R. Wiese, M. Eiglsperger, and M. Kaufmann. yfiles: Visualization and automatic layout of graphs. *Proceedings of the 9th International Symposium on Graph Drawing (GD 2001)*, pages 453–454., 2001.

[9]J. D. Zhang and S. Wiemann. Kegggraph: a graph approach to kegg pathway in r and bioconductor. *Bioinformatics*, 25(11):1470–1471, Jun 2009.