BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 4 2013, pages 518–519
doi:10.1093/bioinformatics/bts730

Data and text mining Advance Access publication January 3, 2013

KEGGParser: parsing and editing KEGG pathway maps in Matlab

Arsen Arakelyan* and Lilit Nersisyan
Group of Bioinformatics, Institute of Molecular Biology, National Academy of Sciences of the Republic of Armenia, Yerevan 0014, Armenia

Associate Editor: Mario Albrecht

ABSTRACT

Summary: KEGG pathway database is a collection of manually drawn pathway maps accompanied with KGML format files intended for use in automatic analysis. KGML files, however, do not contain the required information for complete reproduction of all the events indicated in the static image of a pathway map. Several parsers and editors of KEGG pathways exist for processing KGML files. We introduce KEGGParser—a MATLAB based tool for KEGG pathway parsing, semiautomatic fixing, editing, visualization and analysis in MATLAB environment. It also works with Scilab.

Availability and implementation: The source code is available at http://www.mathworks.com/matlabcentral/fileexchange/37561.

Contact: aarakelyan@sci.am

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 24, 2012; revised on December 17, 2012; accepted on December 23, 2012

1 INTRODUCTION

KEGG pathway database is a collection of manually drawn pathway maps representing current knowledge on molecular interaction and reaction networks, accompanied with KGML (KEGG Markup language) files for automatic computational analyses and modelling of metabolic and signalling networks (Kanehisa *et al*., 2010). Pathways in KGML are represented as graph objects with entry elements (gene products, compounds, pathways) as nodes and relations between elements as edges.

However, in most cases, KGML files do not fully correspond to the static pathway images. Inconsistencies may include absence of event or entity labels, reversed directions for some associations, absence of some interactions, ambiguous definitions of group nodes, compounds and their interactions. Investigation of five randomly chosen KEGG pathways (tight junction, ubiquitin-mediated proteolysis, Toll-like receptor signalling, autoimmune thyroid disease and homologous recombination) revealed that they contain on average 13, 26, 3 and 10 inconsistencies concerned with labelling, missing interactions, directionality and group node definitions, respectively (see Supplementary Data, 'Why to edit KEGG pathways after KGML parsing' section). These inconsistencies may have significant distorting effects on pathway flows. Thus, preprocessing of information contained in a KGML file is needed before it can be used in automatic analysis.

Several KGML parsers have been developed recently, such as KGML-ED (Klukas and Schreiber, 2007), KEGGgraph (Zhang and Wiemann, 2009), KEGGReader plugin for Cytoscape (http://cytoscape.github.com/kgmlreader/) and KEGGTranslator (Wrzodek *et al*., 2011). Although some of these applications have the appropriate functionality for retrieving, parsing, visualizing and editing KGML formatted files, they all suffer from the aforementioned limitations of those files. Moreover, some of the softwares have additional drawbacks, e.g. KEGGReader has reduced functionality for non-metabolic networks and KEGGgraph and KEGGTranslator do not completely support edge and node editing operations. In general, whichever program is chosen, there is always a need for saving manually edited pathways in an appropriate format suitable for further automated analyses of pathway data. As each program provides certain file formats for pathway maps, which cannot be used in MATLAB for automated pathway analysis, a pathway editor is also needed to be developed in MATLAB environment. In this article, we introduce KEGGParser (pathway semiautomatic parser/editor), which is based solely on functions contained in MATLAB, its Bioinformatics toolbox version 3.x and Image processing toolbox 2.x, and it does not require any third-party package. This tool is aimed at enriching MATLAB's Bioinformatics toolbox functionality, which is focused on solving the most contemporary bioinformatics tasks. In addition, we have also ported KEGGParser into Scilab 5.4, a free open-source alternative for MATLAB (included in the source files archive).

2 APPLICATION GENERAL PIPELINE

The main workflow of KEGGParser consists of the following steps:

Pathway retrieval

There are four ways of loading pathway maps into KEGGParser: (i) from locally downloaded KGML files; (ii, iii) from previously parsed local maps and map collections saved as mat-formatted files; (iv) directly from KEGG website by defining the map URL.

Pathway parsing

After retrieval, the KGML file undergoes initial parsing, resulting in creation of a pathway biograph object.

The user may request automatic correction of protein–compound–protein interactions, group nodes and binding directions, to restore correct flows in the graph. These operations are done based on the data stored in the KGML. A use case for automatic correction is given in Section 3 and is described in detail in the Supplementary Data ('Automatic fixing of inconsistencies'

*To whom correspondence should be addressed.

section). Various subtypes of interactions defined by KEGG are generalized based on their action type: activation, inhibition and binding.

Pathway editing

In addition to automatic correction, the following operations may be performed manually in KEGGParser after curation of the pathway: adding/deleting nodes and edges, adding labels and reversing edge directions. These are necessary steps to be undertaken, as there is not always sufficient information inherent in KGML files for performing all the corrections automatically.

Visualization

The visual representation of the pathways created by KEGGParser maximally resembles static images provided by KEGG, preserving node sizes, shapes and their relative positions. In addition, MATLAB built-in graph visualization layouts ('hierarchical', 'radial' and 'equilibrium') are also available as alternatives. Moreover, graph representation of a pathway allows using a wide range of graph-based calculations supported by MATLAB (e.g. finding shortest paths, strongly and weakly connected nodes, minimal spanning trees, maximal flows and so forth). Finally, with the use of MATLAB Compiler toolbox, it is possible to compile KEGGParser into a stand-alone application or incorporate it as a module in other software.

Saving

Pathway graphs may be saved in MATLAB-specific file format (.mat) before and/or after editing.

3 APPLICATION EXAMPLE

To demonstrate capabilities of KEGGParser, we describe here parsing and editing KEGG Chemokine signalling pathway, which contains almost all types of inconsistencies described in Section 1.

During automatic corrections of inconsistencies by KEGGParser, the direction of the 'binding' interaction between 'chemokine receptor' (CXCR6) and 'β-arrestin' (ARRB1) nodes was corrected (Fig. 1A). It is important to take the directions of edges into account, to reserve the correct flow of information in the graph, even though in terms of molecular interactions, binding events are non-directional. Furthermore, the 'group' node formed by 'PKCζ'' (PRKCZ), 'TIAM1' (TIAM2) and 'Par3' (PARD3) nodes was reformatted, whereas before the correction, the 'group' node represented an empty container, and its subnodes were independent, without any incoming and outgoing edges (Fig. 1B). Finally, the protein–compound–protein interaction between 'PLCβ' (PCLB1) and 'PKC' (PRKCB) nodes through 'DAG' (C00165) compound node was fixed (Fig. 1C).

In addition, using 'add node' and 'add edge' commands, we manually restored the pathway branch leading to 'receptor internalization' through binding of 'chemokine receptor' node to 'β-arrestin' node (Fig. 1A), as well as the missing compound–compound interaction leading to activation of 'Ca^{2+}' (C00076) node by 'IP_3' (inositol–trisphosphate, C01245) node (Fig. 1C). The latter also fixed the signal flows from 'PLCβ' node to 'regulation of actin cytoskeleton' node through 'RAP1' node and 'deregulation, chemotaxis and NO induction' functional events through 'PKC' (Supplementary Data, Supplementary Fig. S1).

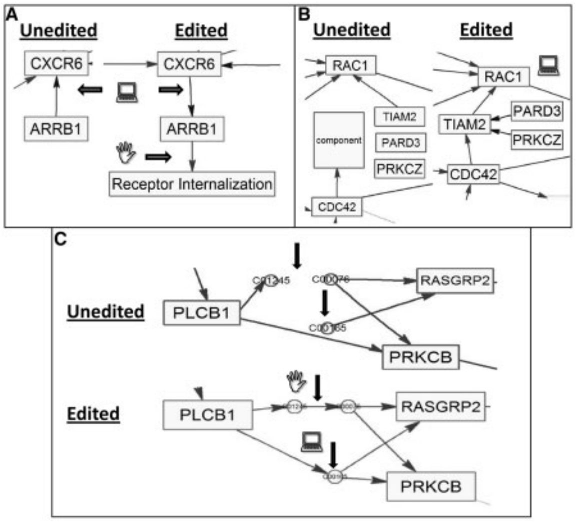

Fig. 1. Example of parsing and editing the chemokine signalling pathway with KEGGParser. (**A**) The flow direction is fixed and the functional event node is added; (**B**) The flows disrupted by the presence of the 'group' node are fixed; and (**C**) The missing edges are added. Arrows indicate fixed inconsistencies. Hand and computer symbols are used to indicate manual and automatic correction, respectively

Thus, the pathway graph resembles the static image and lacks any flow distortions that would potentially lead to wrong results during further analyses.

An additional example demonstrating the usefulness of KEGGParser in graph theory-related analyses is provided in the corresponding section of the Supplementary Data.

4 CONCLUSION

We have created KEGGParser — a tool for parsing, editing and visualizing KEGG pathway maps, implemented in MATLAB. In addition to providing all the functionality available in other contemporary KGML parsers, this tool also performs automatic corrections of inconsistencies between KGML files and static pathway images, and it is a valuable aid for MATLAB-based analysis in bioinformatics research.

Funding: Armenian National Science and Education Fund (ANSEF), NY, USA (NS-molbio 2319 to A.A.).

Conflict of Interest: none declared.

REFERENCES

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Klukas,C. and Schreiber,F. (2007) Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics*, **23**, 344–350.

Wrzodek,C. *et al.* (2011) KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. *Bioinformatics*, **27**, 2314–2315.

Zhang,J.D. and Wiemann,S. (2009) KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics*, **25**, 1470–1471.