SOFTWARE TOOL

REVISED

CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows [v2; ref status: indexed, http://f1000r.es/45p]

Lilit Nersisyan[1], Ruben Samsonyan[2], Arsen Arakelyan[1]

[1]Group of Bioinformatics, Institute of Molecular Biology, National Academy of Sciences of the Republic of Armenia, Yerevan, 0014, Armenia
[2]IUNETWORKS LLC, Yerevan, 0025, Armenia

v2 **First published:** 01 Jul 2014, **3**:145 (doi: 10.12688/f1000research.4410.1)
Latest published: 14 Aug 2014, **3**:145 (doi: 10.12688/f1000research.4410.2)

Abstract

The KEGG pathway database is a widely accepted source for biomolecular pathway maps. In this paper we present the CyKEGGParser app (http://apps.cytoscape.org/apps/cykeggparser) for Cytoscape 3 that allows manipulation with KEGG pathway maps. Along with basic functionalities for pathway retrieval, visualization and export in KGML and BioPAX formats, the app provides unique features for computer-assisted adjustment of inconsistencies in KEGG pathway KGML files and generation of tissue- and protein-protein interaction specific pathways. We demonstrate that using biological context-specific KEGG pathways created with CyKEGGParser makes systems biology analysis more sensitive and appropriate compared to original pathways.

This article is included in the Cytoscape App Collection

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 2 published 14 Aug 2014		✓ report
version 1 published 01 Jul 2014	✓ report	? report

1 **Hans Binder**, University of Leipzig Germany

2 **Augustin Luna**, Memorial Sloan-Kettering Cancer Center USA

Discuss this article

Comments (0)

Corresponding author: Arsen Arakelyan (aarakelyan@sci.am)

How to cite this article: Nersisyan L, Samsonyan R and Arakelyan A. **CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows [v2; ref status: indexed, http://f1000r.es/45p]** *F1000Research* 2014, **3**:145 (doi: 10.12688/f1000research.4410.2)

Grant information: This study was funded by research grant from the State Committee of Science of the Ministry of Education and Science of the Republic of Armenia, granted to Arsen Arakelyan (N 13Y-1F0022, PI: AA).

Competing interests: No competing interests were disclosed.

First published: 01 Jul 2014, **3**:145 (doi: 10.12688/f1000research.4410.1)
First indexed: 17 Oct 2014, **3**:145 (doi: 10.12688/f1000research.4410.2)

REVISED **Amendments from Version 1**

The meaning of String confidence score for protein-protein interactions is explained. Information about how String database version is updated has been added.

See referee reports

Introduction

The KEGG pathway database is a widely accepted source for biomolecular pathway maps and has long been considered as the gold standard for pathway-based analyses due to well-formatted human-readable maps supplemented with machine-readable XML files (KGML), quality of curation and comprehensiveness[1]. However, the KEGG pathway database suffers from a number of limitations that reduce the adaptability of the pathways for automated analysis. These include inconsistencies in KGML files supplied with each pathway image, such as absence of event or entity labels (e.g., links to other pathways or biological process labels), reversed directions for some associations, absence of some interactions, and inconsistent representation of compound interactions[2]. Additionally, some features of KEGG pathways such as protein complex nodes and node duplication, enhance graphical representation, but reduce their machine-readability. Another limitation concerns abstractions (generalizations) used in pathway construction: (1) paralogous genes, not always occurring together in the same biological context, are grouped into single nodes, and (2) all the genes are assumed to be expressed and present in the same pathway. Additionally, the sources of information on interactions depicted in pathways differ in quality and the nature of interactions (indirect, physical, regulatory, etc.). Even accounting for these bottlenecks, the KEGG pathway database is still a highly valued resource, and we aimed to develop a tool that would make the best use of the information collected in it.

There is a wide variety of software that manipulate on KEGG pathways, both standalone and Cytoscape 3 apps, such as KEGGscape (http://apps.cytoscape.org/apps/keggscape) for KEGG pathways visualization and data integration, and others. However, none of the available apps addresses inconsistencies in KGML files, and nor do they neither deal with abstractions of KEGG pathways. Herein, we describe CyKEGGParser app for Cytoscape 3 for KEGG pathway retrieval, visualization, adjustment for inconsistencies in computer-assisted manner, context-specific pathway generation, and exporting the pathways in KGML and BioPAX formats. CyKEGGParser is best suited for KEGG signaling pathways.

Implementation

The software is implemented in Java and is available as an app for Cytoscape 3. The general workflow of CyKEGGParser is presented in Figure 1.

Pathway parsing and corrections

The input for parsing is KGML formatted files, either stored locally or downloaded from the web via REST-based KEGG API. The KEGG API can be used for individual downloads for academic use only; bulk download and non-academic usage requires a KEGG FTP subscription and license agreement (http://www.kegg.jp/kegg/legal.html). The pathway selection dialogue provides a list of all KEGG pathways and organisms, however, if pathway KGML does not exist in the database the user will receive a warning message.

Each KGML file contains entries <pathway>, <entry> and <relation>, which are parsed using Java SAXParser API for reading XML files. The information contained in these entries is kept in Java objects which are instances of Graph, KeggNode and KeggRelation classes. These classes are implemented in CyKEGGParser and are independent of Cytoscape API. All the modifications applied by inconsistency correction algorithms are performed on these objects. Implementation of semi-automatic correction in CyKEGGParser is inherited from KEGGParser and described in detail by Arakelyan and Nersisyan[2].

Once the final Graph with its nodes and edges is created, it is converted into CyNetwork, CyNode and CyEdge objects using CyKEGGParser's KeggNetworkCreator class. During the conversion, all the attributes contained in Graph, KeggNode and KeggEdge objects are set in respective Cytoscape attribute tables. More specifically, we use default CyTables for network, nodes and edges, populating them by creating a new CyColumn for each of the attributes and setting the values in CyRows during iteration over nodes and edges. After CyNetwork, CyNode and CyEdge objects are created, the algorithm iterates through each CyNode, creating a separate view for it and assigning coordinates from respective attributes for X and Y positions. Finally, CyKEGGParser creates attribute-based "kegg_vs" visual style, which is applied on the network

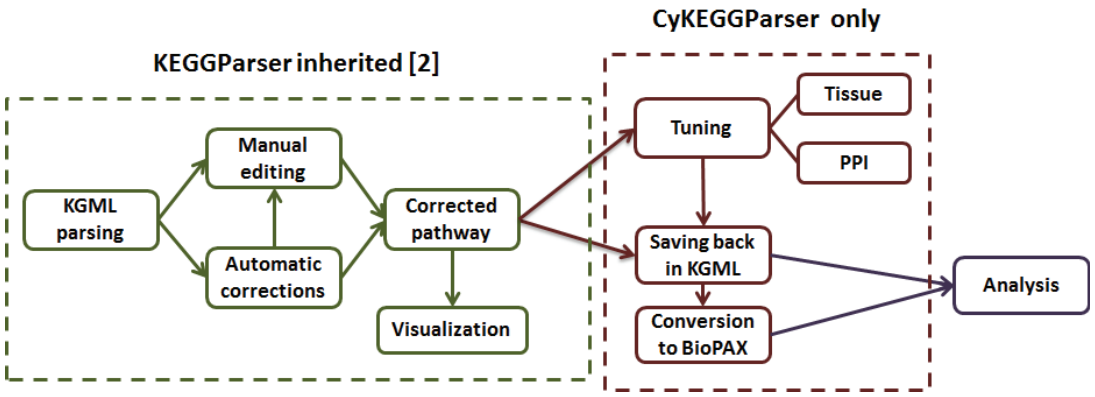

Figure 1. Graphical representation of CyKEGGParser use case.

with *VisualMappingManager* of Cytoscape API. However, any Cytoscape visual style may be applied depending on the user's choice.

All the corrections performed on the network, as well as tuning and saving steps (described below) are tracked in separate log files (see the User Manual provided in the Cytoscape Help menu and at http://molbiol.sci.am/big/apps/cy_kp/jar/CyKEGGParser_User_Manual.pdf).

Limitations for use of KEGG metabolic pathways

KEGG metabolic pathways, along with <relation/> entries, which characterize protein-protein interaction networks (enzyme interactions, in this case), also contain <reaction/> entries, characterizing compound interactions (chemical networks, http://www.kegg.jp/kegg/xml/docs/). Since CyKEGGParser relies on protein-protein interactions (PPI), parsing of metabolic pathways is not always as accurate as it is for signaling pathways. However, if only protein-protein interactions are of concern and if the KGML file contains respective <relation/> entries, CyKEGGParser will parse metabolic pathways similar to signaling ones.

Pathway tuning

Along with the ability to modify the pathways by adding and deleting nodes and edges using Cytoscape-inherent tools, the user may as well customize (or "tune") pathways according to specific biological context: particular tissue or cell type, and experimentally confirmed physical interactions.

Tissue-specific tuning. Tissue-specific tuning is aimed at providing the user with the ability to modify the networks based on genes expressed in a chosen cell/tissue type. Gene expression data for tuning is derived from BioGPS (http://biogps.org/) experiments for human normal and cancer tissues, provided by GeneCards (www.genecards.org), or may be supplied by the user (refer to User Manual for details). Along with specifying the source of data, the user chooses the tissue and specifies gene expression threshold.

The algorithm firstly clones the network preserving all the attributes, except for node and edge identifiers (those should be unique). Then it iterates over all the genes contained in the cloned network nodes, and removes the genes with expression values less than the specified threshold. If a node contains at least one gene that is expressed in current tissue, it remains in the network, otherwise it is removed. Nodes other than of type "gene" are preserved in the network.

PPI based drill-down. In KEGG pathways, node entries represent groups of paralogous genes that have similar functions or interaction profiles[1]. The main incentive of PPI based pathway drill-down is to expand each node into its component genes and connect only those pairs of genes that have been shown to have true physical interactions. Together with tissue-specific tuning, this leads to generation of a "fine-tuned" network, in which all the components occur in the same biological context.

PPI data, retrieved from the String database (http://string-db.org/), have been loaded in an internal MySQL database, located at the server of Bioinformatics Group of the Institute of Molecular Biology NAS RA (http://molbiol.sci.am/big/). The user can choose the source of interactions from the list of databases (GRID, DIP, KEGG, MINT and PDB), as well as set interaction confidence score threshold, which is computed based on various evidence channels, adjusted for probability of randomly observing an interaction[3]. The interactions are manually updated in the local My-SQL database and the version of String used is mentioned on the Tuning dialogue.

The algorithm initially creates a new network, copying all the nodes and node attributes from the former one. Afterwards, it drills down the new network through expanding each node of "gene" type into separate nodes for each member gene. Furthermore, the algorithm iterates over all the pairs of interacting nodes, and connects those members for which there is physical interaction in the corresponding PPI database. Attributes of newly assigned edges are copied from the former network table. After the drill down, duplicated nodes are combined into single ones, and isolated nodes are removed from the network.

Saving

CyKEGGParser provides the functionality of saving the processed pathways back in valid KGML format, so that the modified pathways may be used outside of Cytoscape. All the modifications done to the network are saved in the attributes specific to KGML format. In addition, CyKEGGParser uses KEGGTranslator[4] binary file, embedded in the app package, for KGML conversion to BioPAX2 and BioPAX3 formats (see User Manual for details).

Results and discussion

Parsing and tuning of B Cell Receptor Signaling Pathway with CyKEGGParser

We have taken KEGG B Cell Receptor Signaling Pathway as an example to demonstrate CyKEGGParser functionality and its applicability in pathway-based biology research. B cells are important players in humoral immunity, and their main function is dependent on the B Cell Receptor Signaling Pathway, which is initiated by antigen binding to B cell receptor. We have tuned the B Cell Receptor Signaling Pathway based on BioGPS tissue-specific gene expression data in CD19 B and CD4 T cells (see *Implementation:, Pathway tuning* section), and compared pathway topologies in each case.

Parsing and corrections. Figure 2 shows the pathway parsed with CyKEGGParser with automatic correction options applied. These include three cases of protein-compound-protein (PCP) interaction processing, reversing binding interaction directions of seven edges and processing of two group nodes.

Tissue-specific tuning. We performed B Cell Receptor Signaling Pathway tuning in CD19 B cells and CD4 T cells. Gene expression threshold was set to 25 percentile of gene expression values in the dataset. After tuning, from the 57 nodes available in the original pathway, 54 nodes remained in B cells and 52 nodes remained in T cells. Two nodes, namely, LYN, and CD19 are missing in the B Cell Receptor Signaling Pathway tuned in T cells (Figure 3). Due to their topological importance in signal propagation from the receptors to the target nodes, absence of these two nodes leads to almost complete deactivation of the entire pathway in T cells.

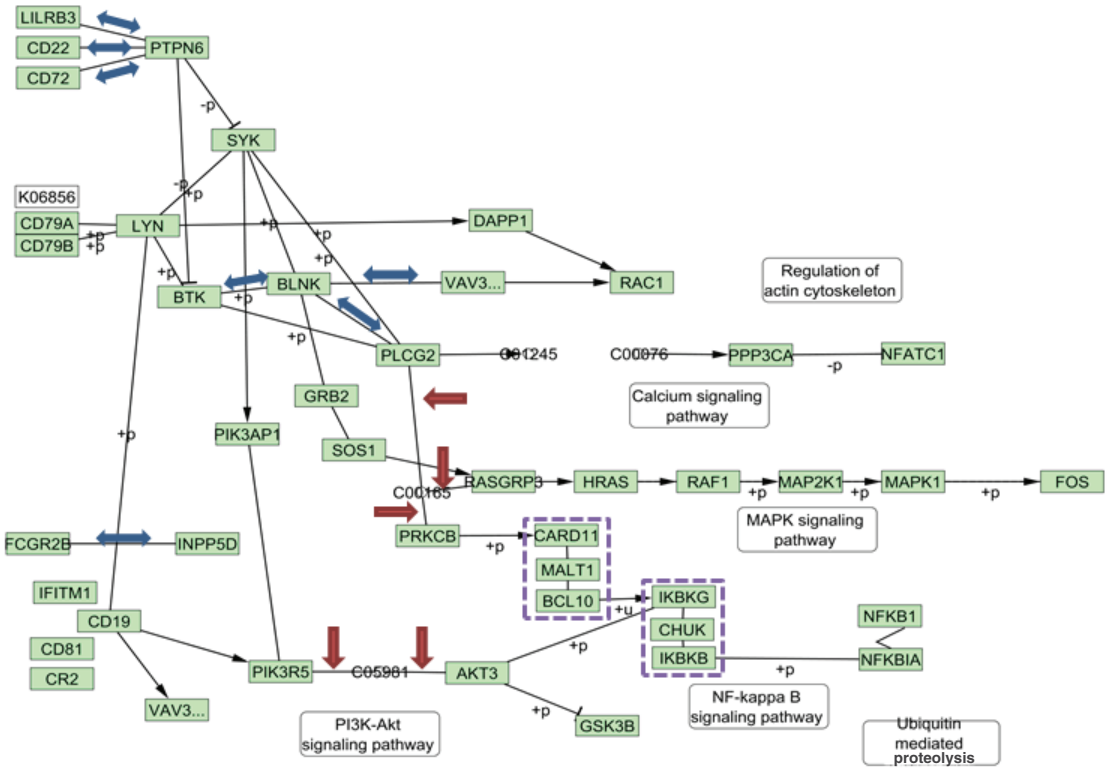

Figure 2. Visualization of KEGG B Cell Receptor Signaling Pathway after parsing and automatic correction. Red arrows indicate edges created by PCP corrections, blue double arrows indicate reversed interactions, and violet dashed rectangles indicate processed group nodes.

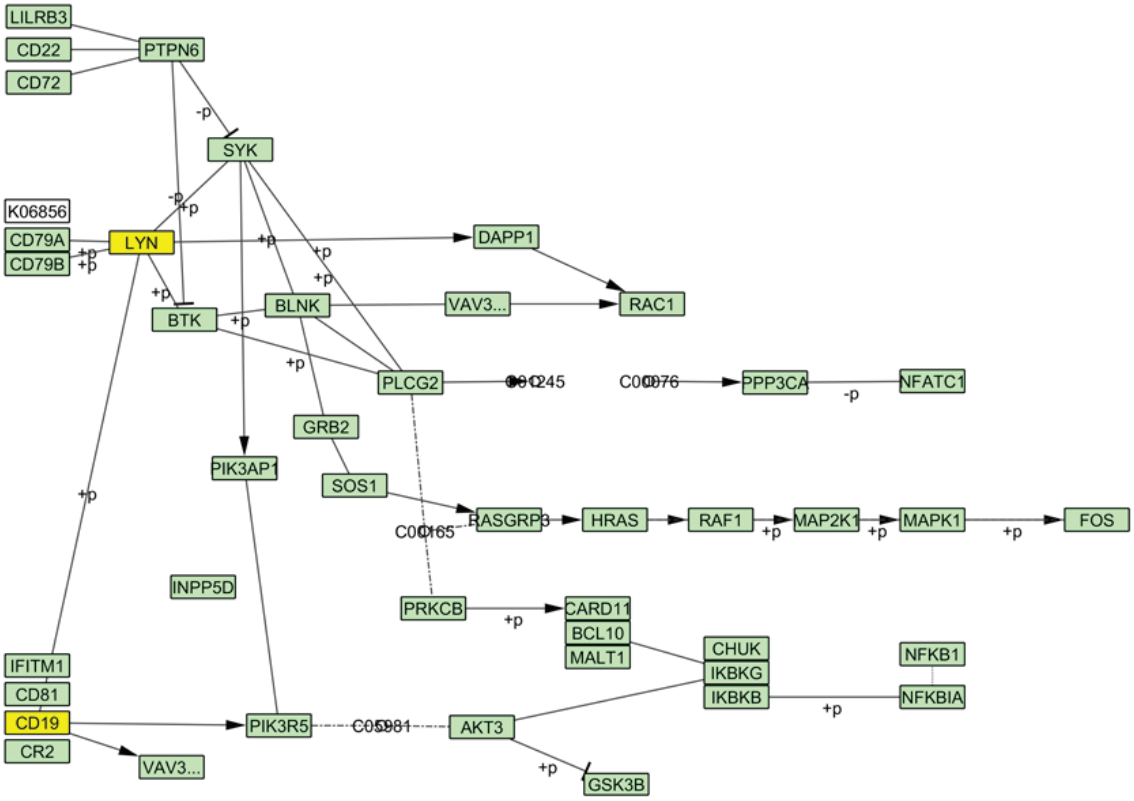

Figure 3. KEGG B cell signaling pathway tuned in CD19 B cells and CD4 T cells. Tuning was performed with 25 percentile threshold of gene expression values for each tissue. Highlighted in yellow are the nodes ("LYN" and "CD19") not present in the pathway tuned in CD4 T cells.

Protein-protein interaction based tuning. The CD19 B cell tissue-specific version of the pathway was further tuned based on PPI. All the database sources (GRID, MINT, KEGG, DIP, PDB) were chosen and 0.8 confidence score threshold was set. Comparison of the PPI-tuned and the original networks showed that the node "VAV3…", which contains three genes, VAV1, VAV2 and VAV3, was duplicated in the original pathway, but remained only in one place in the tuned network (Figure 4). Moreover, of the three VAV member genes only VAV1 interacts with CD19 and BLNK, transducing the signal to rac1 and rac2 nodes. This observation is in accordance with a previously published study indicating VAV1 as the only player in B Cell Receptor Signaling Pathway[5].

Effects of tissue-specific tuning on activity of cell signaling pathways

To further demonstrate necessity of tissue-specific tuning for assessment of pathway activity changes, we compared pathway flows in original and tuned KEGG Calcium Signaling Pathways with three gene expression datasets (norm vs B05 and B01) in CD14 monocytes, Adipocytes, and Cardiac myocytes (see Supplementary Material for details). For calculations, we have used the Pathway Scoring Application for Cytoscape[6]. The simulations show that pathway tuning increases the sensitivity of the pathway for signal flow analysis and thus the ability of the method to detect differentially expressed gene-related changes (Figure 5).

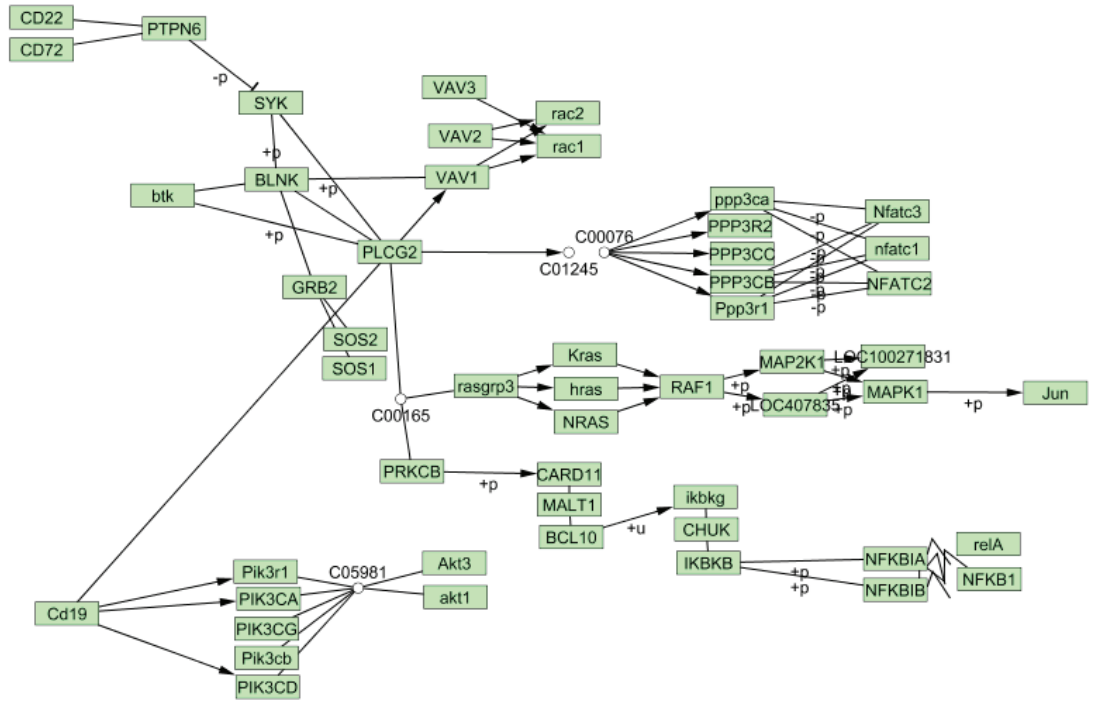

Figure 4. KEGG B cell signaling pathway after tissue specific and PPI-based tuning in CD19 B cells. Tissue-based tuning was performed with 25 percentile gene expression threshold. The confidence score for PPI-based tuning was set to 0.8 and all database sources were included.

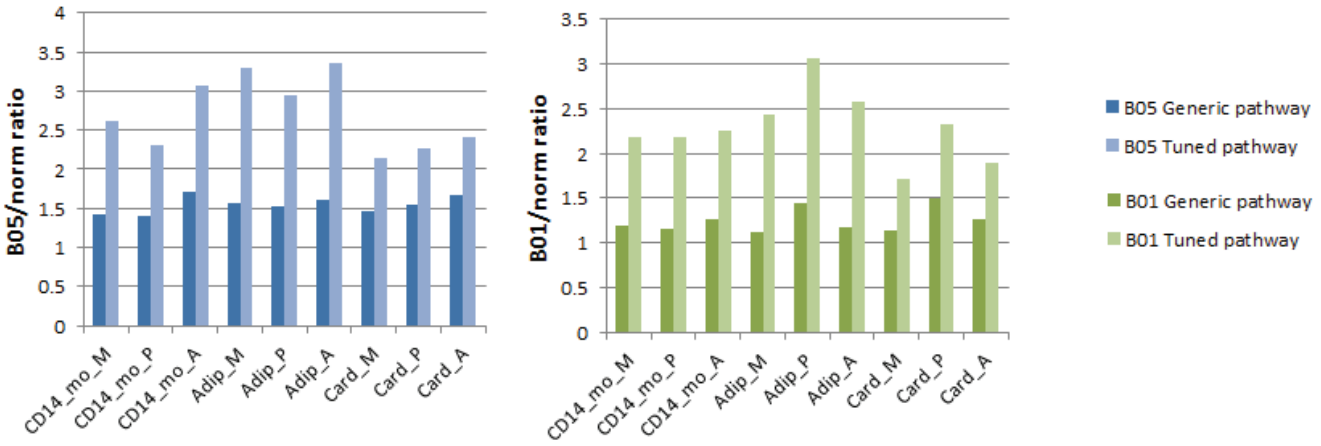

Figure 5. PSA score ratios of Calcium Signaling Pathway computed with simulated data. Target nodes are: M-"MAPK signaling pathway", P-"Phosphatidylinositol signaling system", A-"Apoptosis". Tissues are: "CD14_mo"-CD14 monocytes, "Adip"-Adipocytes, "Card"-Cardiac myocytes.

Dataset: Simulation Data Sets for CyKEGGParser

http://dx.doi.org/10.5256/f1000research.4410.d29107

Dataset 1: "PSA_scores_for_CalciumSignalingPathway.csv"

Description: Pathway scoring application scores for human Calcium signaling pathway, computed with gene expression data for CD14 Monocytes, Adipocytes and Cardiac myocytes with normal BioGPS gene expression data, and simulated B01 and B05 datasets. These data is presented in Figure 5 of the manuscript.

Dataset 2: "CalciumSignalingPathway_gene_expression_data.csv"

Description: Gene expression data for genes belonging to KEGG Calcium signaling pathway from BioGPS experiments for normal human CD14 Monocytes, Adipocytes and Cardiac Mycocytes, and from two simulated datasets (B01 and B05). B05 and B01 datasets were generated from the normal tissue gene expression data, and by randomly assigning two-fold changes to genes based on Bernoulli distribution with probabilities 0.5 (B05) and 0.1 (B01), respectively.

Conclusion

We have developed CyKEGGParser app for Cytoscape 3 that allows for import, correction, visualization, and tuning of KEGG pathways. Although KGML-based pathway import in Cytoscape has also been addressed by KGMLReader (http://apps.cytoscape.org/apps/kgmlreader) and KEGGscape (http://apps.cytoscape.org/apps/keggscape), semi-automatic correction and tuning-based enhancement of pathway specificity are unique and valuable features of CyKEGGParser. With this functionality we aim to maximize the effectiveness and sensitivity of gene expression-based systems biology analyses based on KEGG pathways.

Software availability

App website: http://apps.cytoscape.org/apps/cykeggparser

Source code: https://github.com/lilit-nersisyan/cykeggparser

License: GNU Public License 3.0: https://www.gnu.org/licenses/lgpl.html

Author contributions

LN performed software design and development, testing and analyses, and manuscript preparation, RS implemented PPI database generation and integration, AA performed software design, algorithm development, and manuscript preparation. All the authors have read and approved the final manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This study was funded by research grant from the State Committee of Science of the Ministry of Education and Science of the Republic of Armenia, granted to Arsen Arakelyan (N 13Y-1F0022, PI: AA).

Acknowledgements

We would like to acknowledge the GeneCards database for kindly providing normal and cancer tissue gene expression datasets.

Supplementary Material

Gene expression data generation

We have analyzed KEGG Calcium Signaling Pathway with three gene expression datasets. As normal state (norm), we have taken BioGPS normal gene expression data for three tissues: CD14 monocytes, cardiac myocytes and adipocytes. For simulation of diseased states, we have taken the genes belonging to Calcium signaling pathway and randomly assigned two-fold change in a set of genes based on Bernoulli distribution with probabilities 0.5 (B05) and 0.1 (B01), respectively. In this way we have come up with one "diseased state" (B05) containing 50 and the other (B01) containing 8 differentially expressed genes (the data is provided in supplementary file "CalciumSignalingPathway_gene_expression_data.csv").

Next we have tuned the pathway in CD14 monocytes, cardiac myocytes and adipocytes. For each cell type, the pathway was tuned with an arbitrary threshold of 6.5 corresponding to 27–33 percentiles of gene expression values in the three tissues.

References

1. Kanehisa M, Goto S, Furumichi M, *et al.*: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res.* 2010; **38**(Database issue): D355–60.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Arakelyan A, Nersisyan L: **KEGGParser: parsing and editing KEGG pathway maps in Matlab.** *Bioinformatics.* 2013; **29**(4): 518–9.
PubMed Abstract | Publisher Full Text
3. von Mering C, Jensen LJ, Snel B: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Res.* 2005; **33**(Database Issue): D433–7.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Wrzodek C, Dräger A, Zell A: **KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats.** *Bioinformatics.* 2011; **27**(16): 2314–5.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Packard TA, Cambier JC: **B lymphocyte antigen receptor signaling: initiation, amplification, and regulation.** *F1000Prime Rep.* 2013; **5**: 40.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Isik Z, Ersahin T, Atalay V, *et al.*: **A signal transduction score flow algorithm for cyclic cellular pathway analysis, which combines transcriptome and ChIP-seq data.** *Mol Biosyst.* 2012; **8**(12): 3224–31.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Referee Status: ✓ ✓

Version 2

Referee Report 17 October 2014

doi:10.5256/f1000research.5389.r5832

Augustin Luna
Memorial Sloan-Kettering Cancer Center, New York, NY, USA

The authors have addressed my concerns and the information they provide will be helpful for future users. With respect to KEGG to BioPAX conversion, users of this Cytoscape plugin and make heavy use of the BioPAX output should now be aware from my comments here that the BioPAX export can be validated at the following URL:

http://biopax.baderlab.org/validator/check.html

in the case that they have any concerns of the validity of the output.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 06 August 2014

doi:10.5256/f1000research.4720.r5509

Augustin Luna
Memorial Sloan-Kettering Cancer Center, New York, NY, USA

The authors Lilit Nersisyan, Rouben Samsonyan and Arsen Arakelyan present the Cytoscape plugin/app, CyKEGGParser. The tool provides the functionality to parse KEGG KGML files and edit interactions based on expression data and external protein-protein interaction networks. This tool would be of interest to any researcher wishing to overlay experimental results onto curated pathways.

Issues:

1: Currently, the diagrams rendered in CyKEGGParser lack some elements of the original diagrams. Here is the link to the original "*B Cell Receptor Signaling Pathway*"

http://www.genome.jp/kegg-bin/show_pathway?hsa04662

The original KEGG diagram contains vertical lines indicating transmembrane proteins that are missing from the CyKEGGParser version. There is also a missing interaction between RAC1 and "*Regulation of actin cytoskeleton*". Are both of the elements missing from KGML or are there limitations in the capabilities of Cytoscape to render elements, such as the vertical membrane lines?

Is there any resource (by the authors or others) that keeps track issues with the KGML files, if this is the result of missing information in the KGML?

2: In the PPI drill-down option, a confidence score threshold is provided, but it would be useful to direct readers to more information about this score and how to select appropriate values.

What version of STRING do the authors use? Is this data automatically updated in their internal MySQL database?

3: Currently, there seems to be a problem with the BioPAX export. I believe the problem is in the way KEGGTranslator is called in that lacks the --format option (I found the call to KEGGTranslator in the parsing.log file; this was tested on OS X 10.8.5). If this is indeed the problem, it should be easy to fix. It would be good if the authors validated the resulting BioPAX output from their KGML edits to see if any other issues arise.

http://biopax.baderlab.org/validator/check.html

Missing ID cross-references (XRefs) is a likely error, but one that the authors might not be able to fix if this information is missing from the KGML file.

4: "ppi" in "*Set ppi threshold*" should be capitalized in the Protein-protein interaction Settings window of the "*Pathway tuning settings*".

There are some other minor issues in the dialog. Some of the panels do not seem large enough to accommodate the presented text. On the "*Gene Expression Settings*" with no file the text appears as "*No file selecte*"

CyKEGGParser seems stable and it performs the key function of helping users edit KEGG pathways, but it may require additional steps by users that want figures that mimic the aesthetics of the original KEGG diagrams (this may be unavoidable due to missing layout information in KGML and/or Cytoscape's inability to render all the KGML elements).

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 07 Aug 2014

Lilit Nersisyan, Institute of Molecular Biology NAS RA, Armenia

Thank you for reviewing our paper and your comments, the points you have mentioned are valuable for the app and comprehensiveness of the paper. Here is the detailed response and the description of changes we have made:

1: Currently, the diagrams rendered in CyKEGGParser lack some elements of the original diagrams. Here is the link to the original "B Cell Receptor Signaling Pathway"

http://www.genome.jp/kegg-bin/show_pathway?hsa04662

The original KEGG diagram contains vertical lines indicating transmembrane proteins that are missing from the CyKEGGParser version. There is also a missing interaction between RAC1 and "Regulation of actin cytoskeleton". Are both of the elements missing from KGML or are there limitations in the capabilities of Cytoscape to render elements, such as the vertical membrane lines?

Is there any resource (by the authors or others) that keeps track issues with the KGML files, if this is the result of missing information in the KGML?

KEGG pathway map images, besides containing information about pathway nodes and their interactions, are also decorated with various graphical notations for visualization of spatial distribution of pathway components (i.e. cells, cell and nuclear membranes, organelles, and structural proteins). Since a KGML file solely includes nodes, relations between genes (gene products), compounds and maps (in some metabolic pathways only) and reactions, it is not possible to reconstruct any graphical decoration from a KGML file.

The same applies to present for protein – pathway and compound – pathway interactions, such is the case of *RAC1* and “Regulation of actin cytoskeleton”. In some cases, pathway map nodes are presented to indicate that some nodes and edges are a part of another pathway. In such cases, pathway nodes should not be linked to any node via interactions. In other cases, however, the link from a node to pathway node indicates functional relationship, which is not included in KGML files. In these cases, users can add these interactions manually and save them back in KGML format using CyKEGGParser’s functionality.

Unfortunately, there is no resource that keeps track of the elements that are missing in KGML files, nor is it possible for the app to “guess” what is missing while parsing the actual KGML file.

2: In the PPI drill-down option, a confidence score threshold is provided, but it would be useful to direct readers to more information about this score and how to select appropriate values.

What version of STRING do the authors use? Is this data automatically updated in their internal MySQL database?

To clarify for the meaning of the confidence score, we have added the following paragraph in the User Manual and directed the user to String database’s help page for more details:

“In String, the confidence score is derived by combining evidence about protein-protein interactions from various sources, adjusted for probability of randomly observing the interaction. More information about confidence score meanings and interaction sources can be found at the

Help page of String database:
http://string-db.org/newstring_cgi/show_info_page.pl?UserId=z7Cu7ePjQXnI&sessionId=rD_f8EsHGn
."

In the manuscript body, we have added the following paragraph, where we refer to the String database paper:

"The user can choose the source of interactions from the list of databases (GRID, DIP, KEGG, MINT and PDB), as well as set interaction confidence score threshold, which is computed based on various evidence channels, adjusted for probability of randomly observing an interaction [3]."

The local My-SQL database will be manually updated when new versions of String database are published. Currently, the String version 9.05 is used, and we are populating the database with interactions from version 9.1 at the moment. We have added a label on the tuning dialogue of the app, where the last updated version is seen, and added this information in the User Manual:

"The interactions are manually updated in the local My-SQL database and the version of String used is mentioned on the Tuning dialogue."

3: Currently, there seems to be a problem with the BioPAX export. I believe the problem is in the way KEGGTranslator is called in that lacks the --format option (I found the call to KEGGTranslator in the parsing.log file; this was tested on OS X 10.8.5). If this is indeed the problem, it should be easy to fix. It would be good if the authors validated the resulting BioPAX output from their KGML edits to see if any other issues arise.

http://biopax.baderlab.org/validator/check.html

Missing ID cross-references (XRefs) is a likely error, but one that the authors might not be able to fix if this information is missing from the KGML file.

CyKEGGParser calls KEGGTranslator for BioPAX2 and BioPAX3 conversion, with --format BioPAX_level2 and --format BioPAX_level3 options applied, respectively. Could you, please, send us the log file, to see why the — format option is missing?
KEGGTranslator performs a number of steps in order to retrieve the data missing in KGML files and come up with a valid BioPAX model, including completion of reactions, fixing invalid content of KGML entities and fetching cross-references, as described in http://www.biomedcentral.com/content/pdf/1752-0509-7-15.pdf. As In some cases, however, KEGGTranslator's output is not successfully validated with BioPAX validator, mainly the BioPAX level 2 format. Since CyKEGGParser relies solely on command line calls to KEGGTranslator, there is nothing we can about these cases. However, we have tested whether the successfully validating outputs are also validated after pathway edits performed with CyKEGGParser or by the user in Cytoscape environment. Before calling KEGGTranslator, CyKEGGParser checks for absence of fields required by BioPAX2 and BioPAX3 formats, and adds default values as needed. This ensures that the edits do not induce problems for BioPAX conversion and further validation. The process of saving in KGML format and translating this format into BioPAX is described in the User Manual as follows:

"KGML format saving assures that all the attributes required for BioPAX translation are available. For nodes, these are "entry: id" and "entry: type" attributes: these are assigned the default values

(the next maximum id in the network and "gene" respectively). Node color and coordinates are not required for format conversion, however, if those are missing in the attributes, they will be assigned the values they have in Cytoscape. For interactions, the required attribute is the "type" attribute. If this attribute is missing in the network, it is assigned a value based on source and target node types, as follows: if both nodes are of type "gene" the interaction is assigned the type "PPrel", if either of the nodes is of type "compound", the type "PCrel" is assigned, otherwise the algorithm looks at the required format: in case of KGML and BioPAX_level2, it will assign "maplink" type to those interactions, where one of the nodes is of type "Map", and will give the default value "PPrel" otherwise; in case of BioPAX_level3 the interaction will be removed from the network, since the latter format does not allow for interactions between non-physical entities."

4: "ppi" in "Set ppi threshold" should be capitalized in the Protein-protein interaction Settings window of the "Pathway tuning settings".
There are some other minor issues in the dialog. Some of the panels do not seem large enough to accommodate the presented text. On the "Gene Expression Settings" with no file the text appears as "No file selecte".

We have changed "ppi" to "PPI" in the "Set ppi threshold" of the dialogue, and added more space to file selection label in the "Gene Expression Settings" tab.

Competing Interests: No competing interests were disclosed.

Referee Report 18 July 2014

doi:10.5256/f1000research.4720.r5481

Hans Binder
Interdisciplinary Centre for Bioinformatics, University of Leipzig, Leipzig, Germany

The article *CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows* by Lilit Nersisyan, Rouben Samsonyan and Arsen Arakelyan presents a comprehensive software tool allowing to edit KEGG pathways based on additional information taken, e.g., from tissue expression data and/or protein interaction networks. The functionalities are well described and illustrated using selected pathways in the context of B-cell signalling.

In my opinion, the impact of the contribution is high especially for researchers who are interested in different kinds of biological networks and particularly who aim at interpreting experimental results in terms pathway activities.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.