SOFTWARE TOOL

KEGGscape: a Cytoscape app for pathway data integration [v1; ref status: indexed, http://f1000r.es/3qe]

Kozo Nishida[1,2], Keiichiro Ono[3], Shigehiko Kanaya[4], Koichi Takahashi[1]

[1]Laboratory for Biochemical Simulation, RIKEN Quantitative Biology Center, Osaka, 565-0874, Japan
[2]JST, National Bioscience Database Center (NBDC), Tokyo, 102-0081, Japan
[3]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA
[4]Graduate School of Information Science, Nara Institute of Science and Technology, Nara, 630-0101, Japan

v1 **First published:** 01 Jul 2014, **3**:144 (doi: 10.12688/f1000research.4524.1)
Latest published: 01 Jul 2014, **3**:144 (doi: 10.12688/f1000research.4524.1)

Abstract

In this paper, we present KEGGscape a pathway data integration and visualization app for Cytoscape (http://apps.cytoscape.org/apps/keggscape). KEGG is a comprehensive public biological database that contains large collection of human curated pathways. KEGGscape utilizes the database to reproduce the corresponding hand-drawn pathway diagrams with as much detail as possible in Cytoscape. Further, it allows users to import pathway data sets to visualize biologist-friendly diagrams using the Cytoscape core visualization function (Visual Style) and the ability to perform pathway analysis with a variety of Cytoscape apps. From the analyzed data, users can create complex and interactive visualizations which cannot be done in the KEGG PATHWAY web application. Experimental data with Affymetrix E. coli chips are used as an example to demonstrate how users can integrate pathways, annotations, and experimental data sets to create complex visualizations that clarify biological systems using KEGGscape and other Cytoscape apps.

Open Peer Review

Invited Referee Responses

	1	2	3
version 1 published 01 Jul 2014	? report	? report	✓ report

1 **Matthew DeJongh**, Hope College USA

2 **Egon Willighagen**, Maastricht University Netherlands

3 **Melissa Cline**, University of California, Santa Cruz USA

Latest Comments

No Comments Yet

This article is included in the Cytoscape App Collection

Corresponding authors: Kozo Nishida (knishida@riken.jp), Keiichiro Ono (kono@ucsd.edu)

How to cite this article: Nishida K, Ono K, Kanaya S and Takahashi K. **KEGGscape: a Cytoscape app for pathway data integration [v1; ref status: indexed, http://f1000r.es/3qe]** *F1000Research* 2014, **3**:144 (doi: 10.12688/f1000research.4524.1)

Grant information: This work was supported by National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 01 Jul 2014, **3**:144 (doi: 10.12688/f1000research.4524.1)
First indexed: 06 Aug 2014, **3**:144 (doi: 10.12688/f1000research.4524.1)

Introduction

Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg)[1] is a widely used biological database of high-level biological functions. It contains pathway data sets that have comprehensive annotations and high quality human-curated, hand-drawn diagrams. Most biological pathway databases store data as machine-readable graph topologies, which leave much of the details about how the diagrams were drawn excluded from the data files. This is a problem when third-party developers want to reproduce the pathway diagrams in their applications. In contrast, the KEGG PATHWAY database stores graphics information in machine-readable KEGG Markup Language (KGML, http://www.kegg.jp/kegg/xml) format. Thus, in these pathway diagrams, biological entities, such as enzymes or compounds, are manually laid-out and the diagrams are easy to understand for biologists.

The KEGG PATHWAY database is deployed as a web application using static bitmap images for pathway diagrams, and user-provided date is integrated with KEGG Mapper (http://www.genome.jp/kegg/mapper.html). Furthermore, KEGG Atlas (http://www.genome.jp/kegg/atlas.html) provides a comprehensive network view of global metabolic pathways. Recent improvements to KEGG Atlas, such as Pathway Projector[2] and iPath2[3], have made it possible to perform basic data integration and visualization like mapping the expression values to node graphics. However, despite these features, it is difficult to integrate external data sets and create custom visualization. Furthermore, they are limited to those on existing desktop pathway analysis applications. To ameliorate these problems, several projects for integrating a user's own models onto the KEGG pathways have therefore been developed (CytoSEED Cytoscape app[4], KEGGtranslator[5]).

Cytoscape[6,7] is a de-facto standard software platform for biological network analysis and visualization. One of its advantages is its large collection of apps for a variety of biological problem domains, such as Gene Ontology term enrichment analysis (BiNGO[8]) and statistical network analysis (CentiScaPe[9]), which are also mostly open source software. Additionally, Cytoscape has a flexible network visualization function and is optimized for large-scale network analysis. There are several applications dedicated to biological pathway analysis (Vanted[10], VisANT[11]) that support KGML by default. Although Cytoscape does not have a built-in function to load biological pathways, if this task is done with a separate app, users can take advantage of its large-scale network analysis features, variety of analysis apps, and data visualization of biologists-friendly human curated pathways.

The goal of our new Cytoscape app, KEGGscape, is to bridge the flexibility of fully-featured network analysis platforms with the high-quality pathway diagrams available in the KEGG PATHWAY web application. KEGGscape, a successor of KGMLReader (http://apps.cytoscape.org/apps/kgmlreader) for Cytoscape 2 series, is an app that imports KEGG pathway diagrams from KGML files and provides a new way to use KEGG pathway diagrams as data integration blueprints in cooperation with Cytoscape core features and an existing variety of apps. KEGGscape is completely re-designed for the new Cytoscape 3 API and supports signaling pathways in addition to metabolic pathways, including the global metabolic pathways used in KEGG Atlas (Figure 1). In this paper, we present a basic design and implementation of KEGGscape and an example workflow utilizing KGML files, and experimental data to create information-rich pathway visualizations for clarifying omics-scale data sets. KGMLReader is the first open-source Cytoscape app that reads the graphics details of KGML files, and KEGGscape was designed to use standard Cytoscape features only. These feature enable users to use KEGG pathways with other data sets easily.

Implementation

KEGGscape is a Cytoscape 3 app written in Java programming language and is designed to load pathway data files in KGML format.

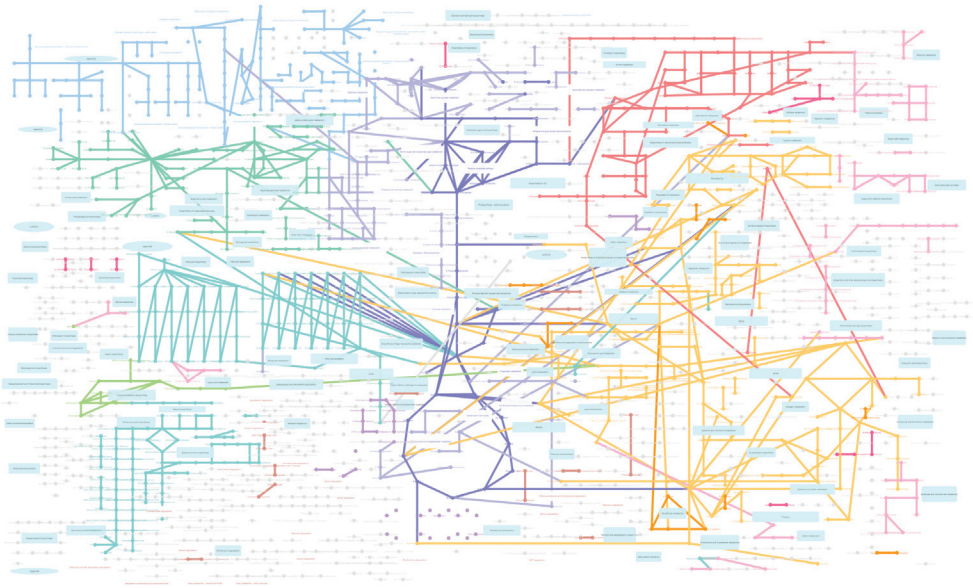

Figure 1. A KEGG Global Metabolic Pathway generated with the KEGGscape app.

KGML is an XML file format designed by the KEGG project and contains the topology of pathways and visual representations of all elements in the diagram. KGML has formal specification as a DTD (Document Type Definition) file, which enables the use of unmarshaller (https://jaxb.java.net) for converting XML elements directly into Java objects. This conversion creates two types of data: pathway topology and its graphical representations. Pathway topology and its properties are converted into CyNetwork and CyTable objects, which are the standard data model in Cytoscape 3. In KGML, all graphical information, such as the color of enzymes or shape of compounds is stored under <graphics> tag. Instead of setting the graphics details of nodes and edges directly from this information, Cytoscape generates Visual Style, which is a collection of default visual properties and visual mapping function, for each pathway based on the information under this tag. KEGGscape follows a standard CyNetworkReader design guideline, which enables Cytoscape to detect KGML files automatically.

Workflow

Figure 2 shows an example of a pathway analysis workflow with KEGGscape. To take advantage of the flexible visualization and analysis features in Cytoscape, users need to import as much information as possible for the pathways they want to analyze. Although Cytoscape is a powerful tool for biological data integration, it is not the best platform for data preparation or cleansing. Users can instead prepare annotations and experimental data sets for the pathway using tools of their choice, such as R (Bioconductor[12]), Python, or Excel. Once the data files are ready, Cytoscape can read them into an on-memory session and visualize the data on the KEGG pathways. Imported data sets only use standard Cytoscape data objects, and users can then access all of the standard Cytoscape features to create custom pathway visualizations. An actual workflow will be presented in a later section.

Limitations

Although KEGGscape can read all information of the pathways saved in KGML files, some of the pathway visualizations in Cytoscape look slightly different from the original hand-drawn diagrams available on the KEGG website. The cause of this issue is missing graphics information in the KGML files. Figure 3 is a side-by-side comparison of the same pathway visualization (human MAPK signaling pathway; KEGG ID: hsa04010). The original diagram (left) contains several background visual annotations that are not visible in the visualization created by Cytoscape (right). The hand-drawn compartmental annotations are not encoded in KGML files, which means they cannot be reproduced by KEGGscape.

Results

As an example workflow, we integrated and visualized a KEGG pathway and gene expression profile using KEGGscape and external tools. In this example, the differentially expressed genes between two groups, mutants and controls, in a global expression profile are mapped on the KEGG pathway, as too are the t-test results.

Data preparation

To perform this pathway analysis in Cytoscape, we used Bioconductor (http://www.bioconductor.org/) to prepare the gene expression matrix data. We normalized Affymetrix GeneChip data by the robust multi-array average (RMA) method with the Bioconductor packages ecoliLeucine[13] and affy[14]. The leucine regulatory protein (Lrp) is a DNA binding protein and known as a leucine responsive global regulator[15]. The p-value for each probeset between four lrp mutant strains and four control chips was calculated by rowttest method in genefilter package[16]. From this calculation, we obtained a list of genes that are differentially expressed (p-value < 0.05). We sent these probeset identifiers to KEGG Mapper and picked the highest hit, which was the glycine, serine and thereonine metabolic pathway (KEGG ID: eco00260) for visualization.

Visualization

To create a visualization using all the data sets, we imported the KGML file of eco00260 and the p-value matrix file prepared in the previous data preparation to Cytoscape 3, and merged the matrices with a custom Python script (Figure 4). Because Cytoscape does

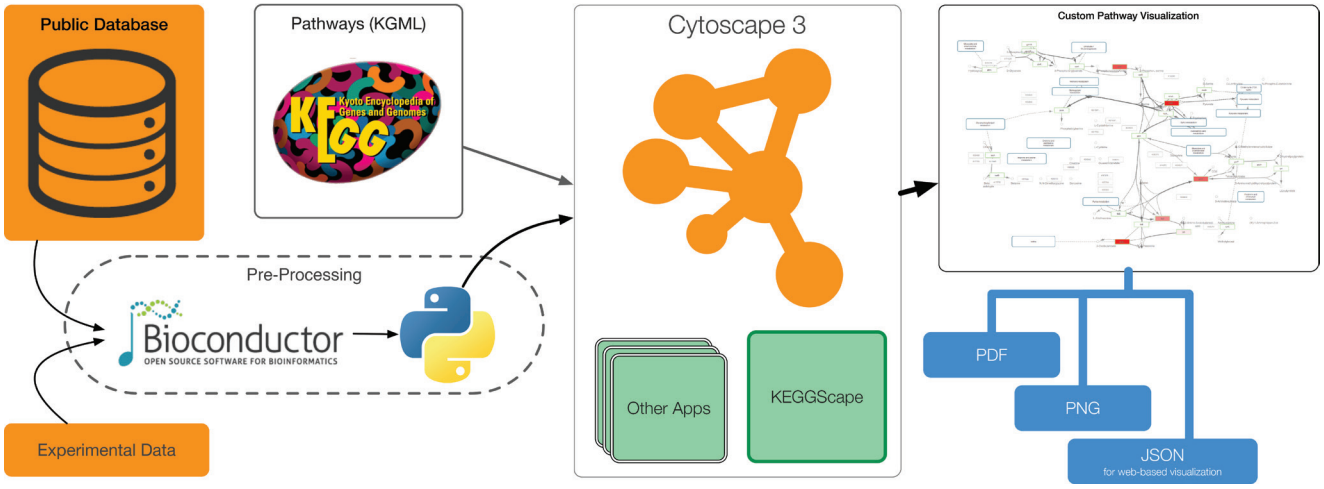

Figure 2. Basic pathway analysis workflow with KEGGscape. Cytoscape with KEGGscape can be used as a part of larger workflows to publish integrated pathway visualizations as vector graphics, bitmap images, or JSON for web-based visualization using Cytoscape.js (http://cytoscape.github.io/cytoscape.js/).

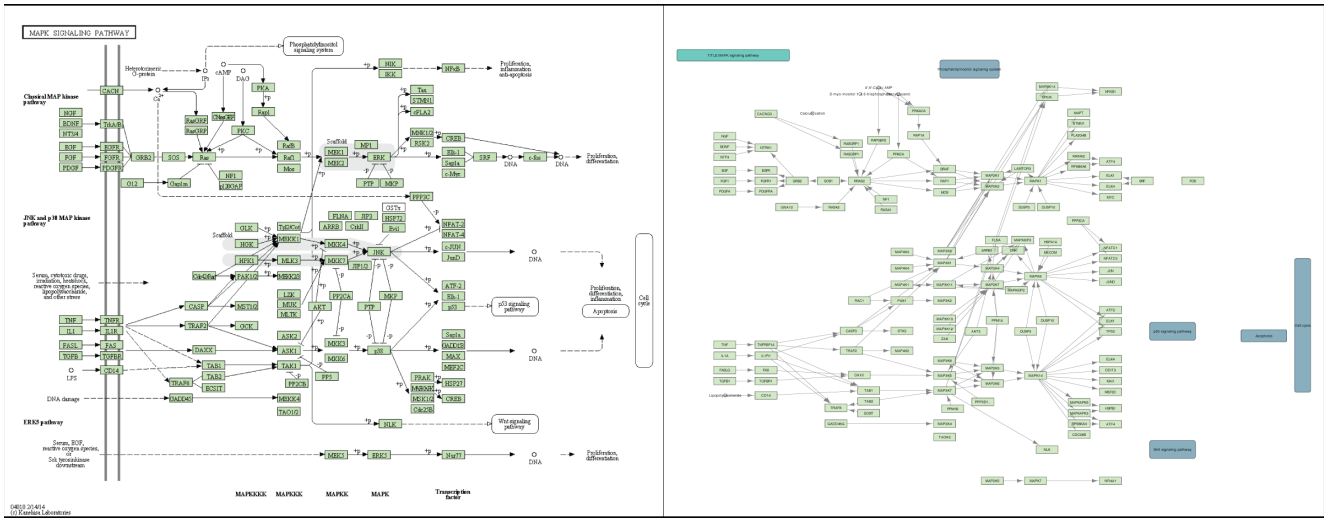

Figure 3. Comparison of the original diagram and Cytoscape visualization for the human MAPK signaling pathway (KEGG hsa04010).

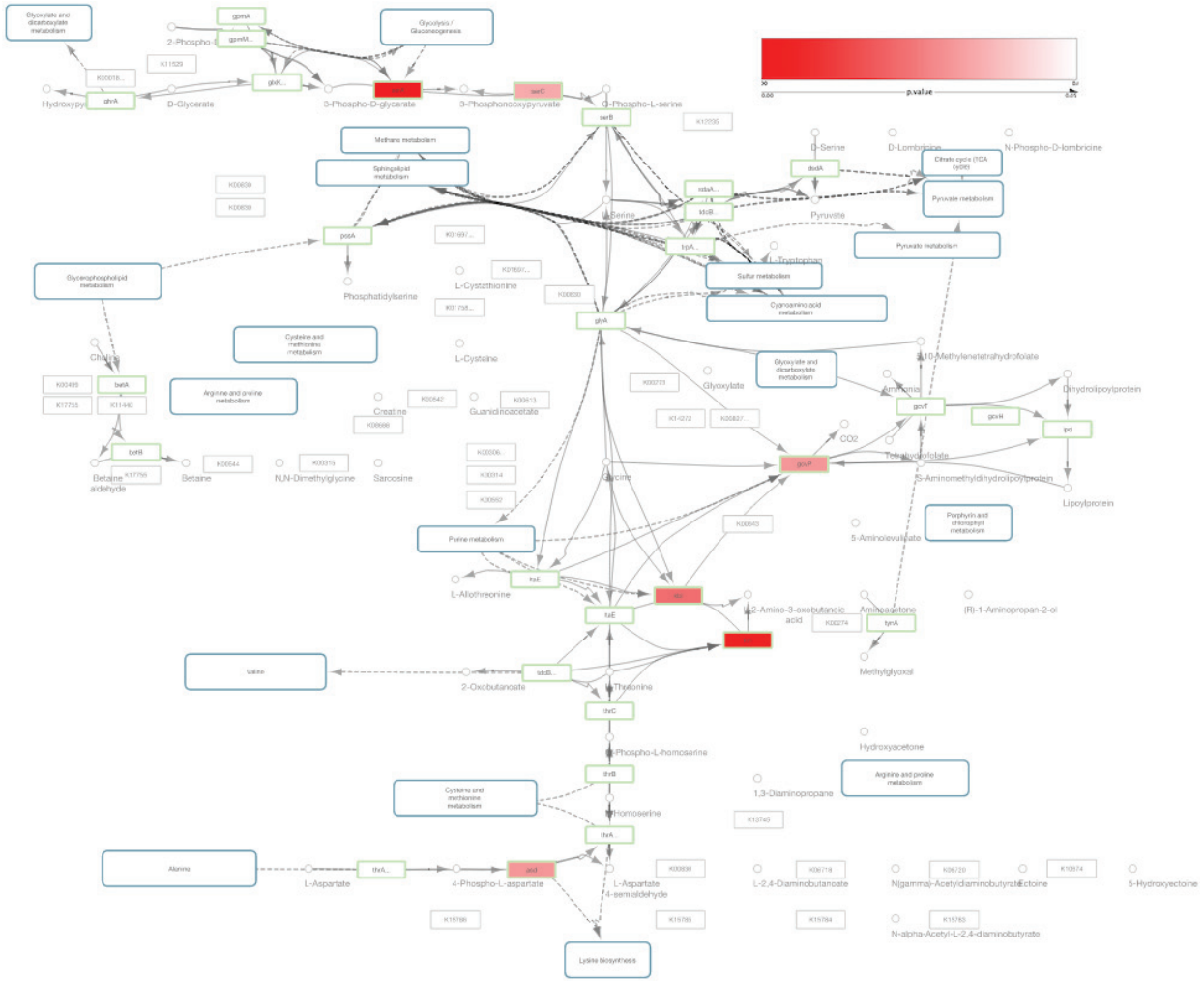

Figure 4. KEGG pathway visualization integrated with gene expression data for the glycine, serine and threonine metabolism pathway of Escherichia coli K-12 MG1655 (KEGG eco00260). Green border nodes are KO (KEGG Orthology) annotated nodes. Red colored nodes include differentially expressed genes (p-value < 0.05).

not support fuzzy key matching, we used our Python script to append a key column to the p-value matrix to utilize the Cytoscape table merge tool.

The node table in Cytoscape for the imported KGML had KEGG gene annotations. The gene IDs for each enzyme node were used as keys for merging the KGML node table and p-value matrix. In this Figure 4, node colors in the original KEGG pathway were mapped to node border colors and p-values were mapped to node color gradient (red to white) to visualize the significantly expressed genes.

Conclusions

In this paper, we presented the design and implementation of KEGGscape and an example analysis workflow integrating global gene expression profiles and KEGG pathways using KEGGscape and two external tools, Bioconductor and Python. The workflow demonstrates how users can integrate omics data in an interactive pathway diagram.

Future plan

Current workflow can map arbitrary omics data onto interactive KEGG pathway diagrams, but it requires some manual editing to create informative visualizations. To minimize the manual process in the workflow, we plan to implement a collection of utility Python scripts to manipulate networks and Visual Styles via RESTful API, which will be published as a part of the Cytoscape 3.2.0 release. This set of Python scripts works to merge pathway related table metadata (omics profiles, non-KEGG pathway metadata) from external platforms like R and Cytoscape to automate common tasks in the visualization process.

Software availability

The app website: http://apps.cytoscape.org/apps/keggscape

Latest source code: https://github.com/idekerlab/KEGGscape

Source code as at the time of publication: https://github.com/F1000Research/KEGGscape/releases/tag/V1.0

Archived source code as at the time of publication: http://dx.doi.org/10.5281/zenodo.10560[17]

License: Apache License Version 2.0

Author contributions

SK guided the GeneChip data preparation and the statistical tests, KT guided the total system design of the project. KN designed and implemented the software under KO's supervision and performed sample analysis and visualization. KO contributed to the writing of this article.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors would like to thank the Google Summer of Code program and Biohackathon attendees for helpful suggestions at the early stage of KEGGscape development, and Peter Karagiannis for reading this article and for useful comments.

References

1. Kanehisa M, Goto S, Sato Y, *et al.*: **Data, information, knowledge and principle: back to metabolism in KEGG.** *Nucleic Acids Res.* 2014; **42**(Database issue): D199–D205.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Kono N, Arakawa K, R Ogawa, *et al.*: **Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API.** *PLoS One.* 2009; **4**(11): e7710.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Yamada T, Letunic I, Okuda S, *et al.*: **ipath2.0: interactive pathway explorer.** *Nucleic Acids Res.* 2011; **39**(Web Server issue): W412–W415.
PubMed Abstract | Publisher Full Text | Free Full Text
4. DeJongh M, Bockstege B, Frybarger P, *et al.*: **CytoSEED: a Cytoscape plugin for viewing, manipulating and analyzing metabolic models created by the model SEED.** *Bioinformatics.* 2012; **28**(6): 891–892.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Wrzodek C, Dräger A, Zell A: **KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats.** *Bioinformatics.* 2011; **27**(16): 2314–2315.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Shannon P, Markiel A, Ozier O, *et al.*: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003; **13**(11): 2498–2504.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Smoot ME, Ono K, Ruscheinski J, *et al.*: **Cytoscape 2.8: new features for data integration and network visualization.** *Bioinformatics.* 2011; **27**(3): 431–432.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.** *Bioinformatics.* 2005; **21**(16): 3448–3449.
PubMed Abstract | Publisher Full Text
9. Scardoni G, Petterlini M, Laudanna C: **Analyzing biological network parameters with CentiScaPe.** *Bioinformatics.* 2009; **25**(21): 2857–2859.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Rohn H, Junker A, Hartmann A, *et al.*: **VANTED v2: a framework for systems biology applications.** *BMC Syst Biol.* 2012; **6**(1): 139.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Hu Z, Chang YC, Wang Y, *et al.*: **VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies.** *Nucleic Acids Res.* 2013; **41**(Web Server issue): W225–W231.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Gentleman RC, Carey VJ, Bates DM, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol.* 2004; **5**(10): R80.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Gautier L: **ecoliLeucine**: Experimental data with Affymetrix E. coli chips, 2007. R package version 1.5.0.
Reference Source
14. Gautier L, Cope L, Bolstad BM, *et al.*: **affy--analysis of Affymetrix GeneChip data**

at the probe level. *Bioinformatics.* 2004; **20**(3): 307–315.
PubMed Abstract | Publisher Full Text

15. Hung SP, Baldi P, Hatfield GW: **Global gene expression profiling in Escherichia coli K12. The effects of Leucine-responsive regulatory protein.** *J Biol Chem.* 2002; **277**(43): 40309–40323.
PubMed Abstract | Publisher Full Text

16. Gentleman R, Carey V, Huber W, *et al.*: **genefilter: methods for filtering genes from high-throughput experiments**. R package version 1.47.5.
Reference Source

17. Nishida K, Ono K, Kanaya S, *et al.*: **F1000Research/KEGGscape.** *ZENODO.* 2014.
Data Source

Open Peer Review

Current Referee Status:

Referee Responses for Version 1

Melissa Cline
Center for Biomolecular Science & Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA

Approved: 06 August 2014

Referee Report: 06 August 2014
doi:10.5256/f1000research.4838.r5318

KEGG is one of the foremost sources of metabolic pathway data. Cytoscape is the *de facto* standard pathway visualization platform. Cytoscape-based visualization of KEGG pathways has always been cumbersome and limited. KEGGscape gets over many of these limitations by streamlining the process, and improving the translation of the graphical elements of KEGG pathways. This makes KEGGscape a very useful resource for the scientific community.

The manuscript is clear and informative. There are only two changes I would ask for:

- First, the authors should mention that KGML files are currently freely available on the KEGG website. This will be news to some readers of this paper, since KEGG has gone through various distribution and licensing models.

- Second, the authors should expand their discussion of merging gene expression data with KGML diagrams. The power of network visualization is in pathway data integration. If there are special challenges in integrating with KEGG pathways, these should be brought forth. The mention of custom python scripts and fuzzy ID matching suggests that this integration task can involve special challenges.

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.

Egon Willighagen
Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, Netherlands

Approved with reservations: 22 July 2014

Referee Report: 22 July 2014
doi:10.5256/f1000research.4838.r5317

The paper describes an interesting and relevant Cytoscape plugin; it is well written (there are a few typos), but I think it can be improved in places. An important reservation is that I cannot determine at this

moment how common or serious the data loss in KGML reading is (see below). That is it should be clarified whether this paper is a first report, a work in progress or a finished project..

Title and Abstract

I find the title appropriate, but it may be promising more than what the app actually does. As far as I can tell, the app is not a general solution, but one specific for KEGG. The abstract is short on setting the state of the domain and, like the Introduction, does not refer to existing research in fields beyond their own. In this way, the abstract is less than clear what unsolved problem was resolved.

Introduction

The Introduction sufficiently explains the setting of the plugin. One thing that strikes me is the lack of references to related work. For example, it mentions "*most biological pathway databases*" without citing or even naming them, which I find suboptimal as they critique those databases and to which the authors contrast their solution. Unlike the Introduction currently seems to suggest, KGML is not the only pathway format that tracks graphical information (e.g. GPML used by PathVisio). The first paragraph currently suggests it solves a problem that others have been tackling too. (P.S. under what license is the KGML specification available?).

The Introduction focuses on technical issues of the integration. It could also say a bit about identifier mapping and the role of licensing in data integration. That is, given the enormous impact of KEGG it can be wondered why this kind of integration with KEGG has not been done yet, because people have done this for other data sets for years (both open source and proprietary). Or, if it has, it should be cited.

Minor comments:

- Check the hyphenation of KEGGtranslator (KEGG-translator versus KEG-Gtranslator).
- The comment that Cytoscape does not have a built-in function to load biological pathways probably refers to formats, not the pathways themselves.
- Since the app focuses on KGML support, some reference to apps supporting other formats (BioPax, SGML) seems appropriate.

Figures

The figures are too hard to read, both in print (where it is impossible) but also in the PDF. This should really be addressed.

Implementation

The Implementation sections can be improved: the section is short on details on the KEGGscape source code, the design of the app and the KGML reader in particular, build instructions, development model, testing (tests seem absent in the code repository).

I also note issues with the KGML readers. For example, connections clear on the KEGG website are missing and misplaced when read into Cytoscape (e.g. between glycine and sarcosine, and the link between Purine metabolism, see Fig. 4) and labels are often misleadingly placed and often unconnected (a limitation of KEGG), though accurately copied from the KGML, it seems. The first data-loss clearly indicates a problem with the reader. These things must be discussed in the Limitations subsection, in my

opinion (and/or returned to in the Future Plan at the end).

Minor comments:

- on-memory -> in-memory?
- <graphics> tag -> <graphics> element (a tag does not really have something under it in XML, attributes at best; an element does).
- What is "Visual Style"?
- Under what license is the DTD available?

Results

The Results section seems to accurately describe what they did, but not how they did this. I think something like a methods section (or an Open Notebook) that explains how the steps are performed (in more detail) would be helpful. However, in the end I was able to figure out how to import a KGML file (which is not under Import).

Another example is reference to a custom Python script without description. In fact, thinking about this again, for an app that does data integration, I would actually expect the app to take care of data processing as much as possible. While I understand that general preparation starting from raw data is best done in other tools, this Python script does not seem to do more than format handling of some kind. Is that correct? If so, why is this not done by the app? Fortunately, this observation was made by the authors too, and they return to it in the Future plan subsection.

Minor comments:

- probeset -> probe set
- thereonine -> threonine

Conclusion

The Conclusion section does not sufficiently summarize the paper: it should reiterate what the problem is that is solved and what makes KEGGscape unique. Although I do not want to imply that I think this paper is not relevant (it is)!

Also the limitations should be returned to here (those already listed in the paper, and those missing), particularly in relation to the "Future plan". The authors' focus should really be on getting the KGML reading performing well: data loss, and data corruption are serious issues. A decent testing toolkit as Maven supports, may be a good start here, but is currently missing from the source code repository.

Minor comments:

- check the hyphenation of KEGGscape (KEGG-scape versus KEG-Gscape)

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: I am contributing to the open source WikiPathways, BridgeDb, and PathVisio projects. I have no competing financial interest.

Matthew DeJongh
Department of Computer Science, Hope College, Holland, MI, USA

Approved with reservations: 08 July 2014

Referee Report: 08 July 2014
doi:10.5256/f1000research.4838.r5316

KEGGscape is valuable Cytoscape app that enables flexible visualization and manipulation of KEGG pathway maps. The capability for visualizing data sets (e.g., gene expression levels) on pathways in Cytoscape is particularly appealing. However, the example presented in the paper does not contain sufficient information to enable a reader to duplicate the process. First, the authors do not cite the origin of the gene expression dataset they use in their example. Second, as far as I can tell, there is no tutorial provided with KEGGscape. At a minimum, the authors should provide instructions for opening a KGML file and loading a dataset.

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.