Bioinformatics, 38(5), 2022, 1473–1476
https://doi.org/10.1093/bioinformatics/btab793
Advance Access Publication Date: 23 November 2021
Applications Note

Data and text mining

SBGNview: towards data analysis, integration and visualization on all pathways

Xiaoxi Dong[1,2], Kovidh Vegesna[1,2], Cory Brouwer[1,2] and Weijun Luo [1,2,3,*]

[1]Department of Bioinformatics and Genomics, College of Computing and Informatics, UNC Charlotte, Charlotte, NC 28223, USA, [2]UNC Charlotte Bioinformatics Service Division, North Carolina Research Campus, Kannapolis, NC 28081, USA and [3]Novant Health, Charlotte, NC 28207, USA

*To whom correspondence should be addressed.
Associate Editor: Anthony Mathelier

Received on July 17, 2021; revised on November 6, 2021; editorial decision on November 16, 2021; accepted on November 17, 2021

Abstract

Summary: Pathway analysis is widely used in genomics and omics research, but the data visualization has been highly limited in function, pathway coverage and data format. Here, we develop SBGNview a comprehensive R package to address these needs. By adopting the standard SBGN format, SBGNview greatly extend the coverage of pathway-based analysis and data visualization to essentially all major pathway databases beyond KEGG, including 5200 reference pathways and over 3000 species. In addition, SBGNview substantially extends or exceeds current tools (esp. Pathview) in both design and function, including standard input format (SBGN), high-quality output graphics (SVG format) convenient for both interpretation and further update, and flexible and open-end workflow for iterative editing and interactive visualization (Highlighter module). In addition to pathway analysis and data visualization, SBGNview provides essential infrastructure for SBGN data manipulation and processing.

Availability and implementation: The data underlying this article are available as part of the SBGNview package is available on both GitHub and Bioconductor: https://github.com/datapplab/SBGNview, https://bioconductor.org/packages/SBGNview.

Contact: luo_weijun@yahoo.com

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

1 Introduction

Pathway analysis has become a prevalent analytical approach in genomics and omics studies. Numerous analysis methods (Khatri *et al.*, 2012; Luo *et al.*, 2009) and databases (Caspi *et al.*, 2020; Jassal *et al.*, 2020; Jewison *et al.*, 2014; Kanehisa *et al.*, 2017; Mi *et al.*, 2017; Rodchenkov *et al.*, 2020) have been developed. But tools for pathway analysis visualization are sparse, with Pathview (Luo *et al.*, 2017; Luo and Brouwer, 2013) and Cytoscape (Smoot *et al.*, 2011) as two representatives. Cytoscape renders pathways as networks efficiently, which tend to lose the context information and are hard to interpret. Pathview (Luo *et al.*, 2017; Luo and Brouwer, 2013), a tool we developed previously, can map various biological data and generate highly interpretable pathway graphs with biological context (Luo *et al.*, 2018), but only supports KEGG (Kanehisa *et al.*, 2017) and its pathway data format (KGML). Major pathway databases like KEGG (Kanehisa and Sato, 2020) and Reactome (Sidiropoulos *et al.*, 2017) also provide integrated yet limited functions for data analysis and visualization. Existing data visualization tools largely focus on individual pathway databases or small collections (overviewed in Supplementary Table S3). There has not been a tool for data integration and visualization supporting pathways cross major databases and standard formats.

Here, we present a new tool set—SBGNview as a systematic solution to this pressing need. SBGNview is built on Systems Biology Graphical Notation (SBGN) (Le Novere *et al.*, 2009), a set of high quality, standard graphical languages for representing biological processes and interactions. SBGN has been widely adopted and supported by major pathway databases, collections and pathway curation or editing packages (https://SBGN.github.io/software). Unfortunately, SBGN pathways so far have limited usage in data integration and visualization. We developed SBGNview to fill in this gap. Consequentially, all major pathway databases adopting SBGN format are now open to pathway analysis and data visualization in a consistent and robust way, including Reactome (Jassal *et al.*, 2020), PANTHER (Mi *et al.*, 2017), SMPDB (Jewison *et al.*, 2014), MetaCyc (Caspi *et al.*, 2020), MetaCrop (Schreiber *et al.*, 2012) and Pathway Commons (Rodchenkov *et al.*, 2020). This is much broader and deeper collection of reference pathways than KEGG alone, in terms of pathway number, categories, resolution and details. In addition to functions in data integration and visualization, SBGNview provides a comprehensive tool set for SBGN-based pathway

analysis, as well as SBGN data search, retrieval and processing. Note, SBGNview includes functions essential for SBGN pathway analysis, including gene (or compound, molecule) sets extraction, ID conversion and pathway mapping, but not algorithms or statistical tests (e.g. GSEA, GAGE, etc.).

2 Main Features

2.1 SBGNview greatly extends the coverage of pathway-based analysis and data visualization

SBGNview naturally supports all pathway databases and collections that adopt SBGN standards and its format (SBGN-ML) (Rougny *et al.*, 2019; van Iersel *et al.*, 2012), and make them accessible for pathway analysis and data visualization (Table 1 and Supplementary Table S1). SBGNview provides two tiers of support for SBGN pathway data. Tier 1, deep and direct support (i.e. diagram optimization and ID mapping) to a core collection of pathway data from five major pathway databases including Reactome (Jassal *et al.*, 2020), PANTHER (Mi *et al.*, 2017), SMPDB (Jewison *et al.*, 2014), MetaCyc (Caspi *et al.*, 2020) and MetaCrop (Schreiber *et al.*, 2012). These databases together covers 5200 reference pathways and over 3000 species (Table 1). This is a much broader and deeper collection of reference pathways than KEGG alone, especially in the domains of major research species, crops, small molecules and metabolic pathways. For example, Reactome alone covers 1746 reference pathways (versus 430 in KEGG, hence broader overall), MetaCyc covers 2518 metabolic pathways (versus 91 in KEGG, hence deeper in this domain) (Table 1). In addition to the 3000 species from the original databases, SBGNview can also map reference pathways to 6190 KEGG species via KEGG Orthology (Kanehisa *et al.*, 2016). SBGNview provides diagram optimization and ID mapping on these pathway databases. In other words, we have prepared a full collection of high-quality SBGN pathways from these databases, which can be directly used in data analysis. The original heterogeneous pathway diagrams have been consistently and extensively modified in graph layout, format, space usage and multi-level details (Fig. 1 versus Supplementary Fig. S2). These changes are essential for data visualization, analysis and comparison. In addition, glyph IDs of the original SBGN-ML files are largely databases specific, not common molecule IDs. We generated ID mapping between SBGN-ML glyph IDs and common molecule IDs, so that omics or molecular data can be easily mapped to these pathways. To be specific, we downloaded source mapping between glyph ID to the primary external IDs (UniProt, Entrez or ChEBI, etc.) from the source databases, which were then be mapped to ~20 other common IDs using Pathview's mapping functions (details in SBGNview & Pathview package documentation). Tier 2, generic support to all other pathway databases, collections and users' custom pathway data in SBGN format (Supplementary Table S1). These data can be used in SBGNview, except diagrams may not be optimized for data visualization, and users need to provide ID mapping or make sure the glyph IDs are commonly used molecule IDs. Note the collection of pathways in databases with Tier 1 support is extensive, and should cover most of the user needs in pathway analysis and data visualization.

2.2 SBGNview extends the architecture and workflow of Pathview with numerous unique features

SBGNview has a similar design as the Pathview package (Luo and Brouwer, 2013) in four functional modules: Downloader, Parser, Mapper and Viewer (Supplementary Fig. S1). These modules are not only integral components of SBGNview, but they are also useful for SBGN data input/output, processing, parsing and rendering in general. In addition to these Pathview-like modules, SBGNview workflow has a unique part, the Highlighter module. Highlighter provides a post-rendering modification mechanism, which can highlight or modify any part (s) of the initial rendering of the SBGN graph (SBGNview object in Supplementary Fig. S1), including any subsets of nodes, arcs, paths and their attributes specified by the user (Fig. 1A–C and Supplementary Fig. S3). SBGNview README (https://github.com/datapplab/SBGNview) showcases the usage of the Higlighter functions, with details in the function documentation and main vignette. The updated rendering (SBGNview object) can be further modified by Highlighter iteratively, and can also be re-fed into Viewer for updated graphics output (Supplementary Fig. S1), or saved in RData or SVG formats for external tools. In other words, the Highlighter makes SBGNview workflow an open-ended process, suitable for iterative editing/updating and interactive visualization.

SBGNview also differs from Pathview (Luo and Brouwer, 2013) in both input and output. For input pathway data, SBGNview supports the standard SBGN format adopted by multiple major databases, while Pathview supports KGML used only by KEGG (Supplementary Table S2). Pathview has two discrete output styles, KEGG view and Graphviz view. The former is a raster image with all context information (tissue/cell types, etc.) labeled by KEGG curator. The latter is a vector image with no context information. SBGNview has one unified output style (with different file formats, i.e. SVG, PDF, PNG and PS), a high-resolution vector image with comprehensive meta-data and context information defined with SBGN-ML. Even more importantly, the primary output format SVG, as XML-based graphics, can be automatic or manual updated by editing the XML source.

2.3 SBGNview provides a comprehensive tool set for SBGN pathway analysis, data processing and visualization

SBGNview integrates three primary functions in working with SBGN pathways: data visualization, data integration and pathway analysis workflow. For Data visualization, SBGNview offers extensive graphics control on all elements (glyphs, complex, compartments, arcs, reactions, processes, text, etc.) and their attributes (line type, width, shape, strike, fill, color, labels, size, position, even margins, etc.) in the pathway diagrams. The Highlighter module in SBGNview provides extra mechanism of graphics control, targeting specific graphic attributes of selected sets, subsets or classes of pathway components (proteins/enzymes, interactions, reactions, processes) or graph elements (glyphs, arcs and paths) (Supplementary Fig. S3). This is especially useful for sub-pathway level analysis, visualization and interpretation. For data integration, by connecting to Bioconductor (Gentleman *et al.*, 2004) and KEGG (Kanehisa *et al.*, 2017) annotation resources, SBGNview can map, integrate

Table 1. SBGN pathway databases directly supported by SBGNview (Tier 1 support)

Database	SBGN data	Pathways[a]	Species	Description
Reactome	Pathway Commons	1746	16	Pathway database for major research species
SMPDB	Pathway Commons	725	1	Human small molecule pathways
PANTHER	Pathway Commons	149	132	Signaling pathways of multiple species
MetaCyc	MetaCyc	2518	2980	Metabolic pathways from all domains of life
MetaCrop	MetaCrop	62	9	Crop metabolism pathways
KEGG[b]	KEGG	430	6190	Pathway database for most sequenced species

[a]Ortholog pathways across multiple species counted as the same pathway, similar to reference pathway in KEGG.
[b]KEGG statistics for comparison. All data were collected in 2020.

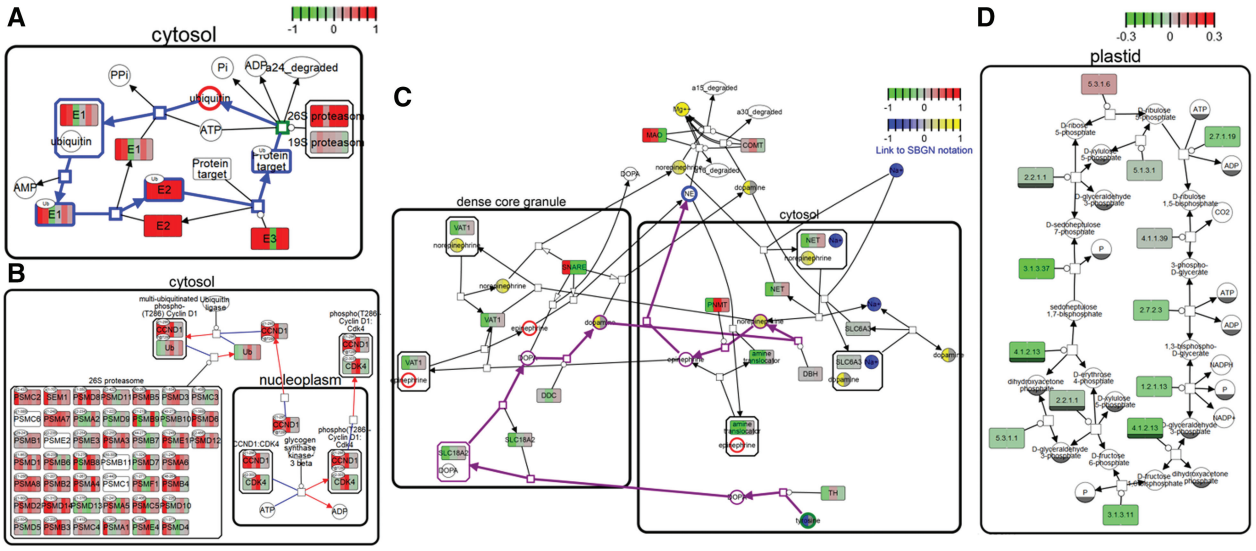

Fig. 1. Example SBGNview graphs: (**A**) ubiquitin proteasome pathway (PANTHER pathway P00060) and (**B**) ubiquitin-dependent degradation of Cyclin D1 pathway (Reactome pathway R-HSA-69229) upregulated in breast cancer (GEO GSE16873); (**C**) both gene data (GEO GSE16873) and compound data (simulated with Pathview) with highlighted nodes and path on the Adrenaline and noradrenaline biosynthesis pathway (PANTHER pathway P00001); (**D**) Calvin cycle (MetaCrop pathway) downregulated in Arabidopsis pen3 mutants (GEO GSE3220). As in Pathiew, color keys represent the gene or compound expression/abundance levels or changes. In SVG outputs, the URLs below color keys link to SBGN legend cards. We also showcase the highlight functions of SBGNview here. In (A), the ubiquitin cycle in the pathway is highlighted in blue, with ubiquitin marked in red. In (B), different types of reactions are marked by arcs highlighting, blue arcs for consumption and red arcs for production. In (C), Nodes highlighted in red are epinephrine, the shortest path between tyrosine (start node in green) and epinephrine (end node in blue) is highlighted in purple

and render a large variety of biological data on SBGN pathways, including any data that can be mapped to gene (genomes, metagenomes, transcripts, proteins) or compound (metabolites, drugs, small molecules) IDs. It supports most common gene or compound ID types, thousands of reference pathways and species (Table 1). Like Pathview, SBGNview can be easily integrated with a wide variety of existing tools (in R or portable in R) for omics data analysis and pathway analysis. In one package vignette, we demonstrated a complete pathway analysis workflow using GAGE (Luo *et al.*, 2009) + SBGNview. In Supplementary Note S2, we showcased SBGNview's functions in pathway analysis and data visualization using an example analysis of microarray data.

SBGNview also provides general infrastructure for SBGN pathway data processing. KEGG (Kanehisa *et al.*, 2017) hosts a centralized REST API for pathway data search, retrieval, extraction and downloading. Pathview takes the advantage of the KEGG API for these tasks. Unfortunately, there has been no such facility for SBGN pathway data largely because the data come from many heterogeneous sources. SBGNview fills in this gap as the first tool focusing on SBGN pathway-based data analysis. The five functional modules of SBGNview are generally useful for SBGN pathway data processing hence an important infrastructure. In addition, we provide a high-quality collection of SBGN pathway data with open access via GitHub (SBGNhub: https://github.com/datapplab/SBGNhub). We also provide functions, data and supportive package (SBGNview.data: https://github.com/datapplab/SBGNview.data) for SBGN pathway search, retrieval, data mapping and extraction. For pathway analysis, SBGNview has functions for pathway gene/molecule sets extraction, ID conversion and searching (but not testing algorithms), as shown in Supplementary Note S2 and in package vignettes.

3 Conclusion

SBGNview maps, integrates and renders a wide range of biological data to SBGN pathways. By adopting the standard SBGN format, SBGNview greatly extends the coverage of pathway-based analysis and data visualization to essentially all major pathways beyond KEGG. SBGNview extends the proven design of Pathview with multiple unique features, including standard and widely supported input and output formats, high-resolution graphics accessible to both machine and human, and flexible, iterative and open-ended workflow. While Pathview primarily focuses on data visualization, SBGNview provides a complete workflow and comprehensive tool set for SBGN-based pathway analysis, data visualization and processing.

Funding

This work was supported by the National Science Foundation [ABI-1565030 to W.L.].

Conflict of Interest: none declared.

References

Caspi,R. *et al.* (2020) The MetaCyc database of metabolic pathways and enzymes – a 2019 update. *Nucleic Acids Res.*, **48**, D445–D453.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Jassal,B. *et al.* (2020) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.

Jewison,T. *et al.* (2014) SMPDB 2.0: big improvements to the small molecule pathway database. *Nucleic Acids Res.*, **42**, D478–484.

Kanehisa,M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

Kanehisa,M. and Sato,Y. (2020) KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci.*, **29**, 28–35.

Kanehisa,M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–462.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.

Le Novere,N. *et al.* (2009) The systems biology graphical notation. *Nat. Biotechnol.*, **27**, 735–741.

Luo,W. and Brouwer,C. (2013) Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, **29**, 1830–1831.

Luo,W. *et al.* (2009) GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, **10**, 161.

Luo,W. *et al.* (2017) Pathview Web: user friendly pathway visualization and data integration. *Nucleic Acids Res*, **45**, W501–W508.

Luo,W. *et al.* (2018) Systematic reconstruction of autism biology from massive genetic mutation profiles. *Sci. Adv.*, **4**, e1701799.

Mi,H. *et al.* (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**, D183–D189.

Rodchenkov,I. *et al.* (2020) Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res.*, **48**, D489–D497.

Rougny,A. *et al.* (2019) Systems biology graphical notation: process Description language Level 1 Version 2.0. *J. Integr. Bioinf.*, **16**. https://academic.oup.com/database/article/doi/10.1093/database/baaa017/5821574?login=true

Schreiber,F. *et al.* (2012) MetaCrop 2.0: managing and exploring information about crop plant metabolism. *Nucleic Acids Res.*, **40**, D1173–1177.

Sidiropoulos,K. *et al.* (2017) Reactome enhanced pathway visualization. *Bioinformatics*, **33**, 3461–3467.

Smoot,M.E. *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

van Iersel,M.P. *et al.* (2012) Software support for SBGN maps: SBGN-ML and LibSBGN. *Bioinformatics*, **28**, 2016–2021.