Bioinformatics, 2018, 1–3
doi: 10.1093/bioinformatics/bty719
Advance Access Publication Date: 4 September 2018
Applications Note

Databases and ontologies

*meta*Graphite–a new layer of pathway annotation to get metabolite networks

Gabriele Sales*, Enrica Calura and Chiara Romualdi

Department of Biology, University of Padova, Padova 35121, Italy

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on December 22, 2017; revised on July 17, 2018; editorial decision on August 16, 2018; accepted on September 3, 2018

Abstract

Motivation: Metabolomics is an emerging 'omics' science involving the characterization of metabolites and metabolism in biological systems. Few bioinformatic tools have been developed for the visualization, exploration and analysis of metabolomic data within the context of metabolic pathways: some of them became rapidly obsolete and are no longer supported, others are based on a single database. A systematic collection of existing annotations has the potential of considerably boosting the investigation and contextualization of metabolomic measurements.
Results: We have released a major update of our Bioconductor package *graphite* which explicitly tracks small molecules within pathway topologies and their interactions with proteins. The package gathers the information stored in eight major databases, oriented both at genes and at metabolites, across 14 different species. Depending on user preferences, all pathways can be retrieved as gene-only, genemetabolite or metabolite-only networks.
Availability and implementation: The new *graphite* version (1.24) is available on Bioconductor.
Contact: gabriele.sales@unipd.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metabolomics is an emerging 'omics' science involving the comprehensive characterization of metabolites and metabolism in biological systems. Metabolites represent both the downstream output of the genome and the upstream input from the environment. Their study allows the exploration of the link between genes and environment (Wishart, 2016). If data gathering has become easier with the improvement of experimental protocols and high-throughput techniques, the annotation, analysis and interpretation of their results still poses significant challenges. This need has driven the development of metabolite databases, containing information about their biological roles, physiological concentrations, disease associations, chemical reactions, metabolic pathways and reference spectra, as well as computational tools for post-processing and analysis.

Apart from their annotation, it is a matter of fact that the investigation and contextualization of metabolomic data can be considerably enhanced by the integration of other data types. Several databases have been published so far, however very few bioinformatic tools have been proposed for the management, visualization and analysis of metabolomic data within the biological context of metabolic pathways (Cottret *et al.*, 2010; Posma *et al.*, 2014; Rodriguez-Martinez *et al.*, 2017). However, most of the existing tools do not include multiple pathway databases and organisms, forcing the developers to change the analysis strategy at the change of the data settings. Due to the novelty of the field, we moreover consider the use of the R language an important feature to foster new analysis development.

Here, we present a major update of our Bioconductor package *graphite* (Sales *et al.*, 2012) which explicitly tracks small molecules within pathway topologies and their interactions with proteins. *graphite* provides pathway annotation as protein-protein, protein-metabolite and metabolite-metabolite networks according to user preference.

2 Implementation

graphite version 1.24 makes access to metabolite information an opt-in process offering each pathway under three different forms: (i)

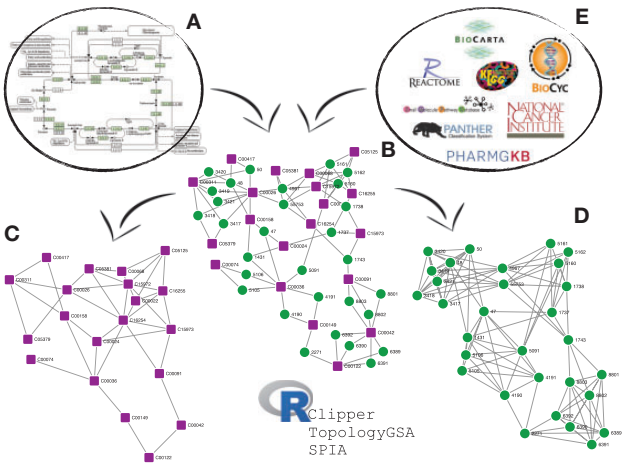

Fig. 1. Human KEGG pathway 'Cytrate Cycle' (**A**) and its corresponding networks in *graphite* (**B**–**D**). *graphite* converts pathway annotation and provides three variants: mixed graph with metabolites and proteins (B), metabolite-only graph (C) and protein-only graph (D) for eight different databases (**E**). Although not shown in the Figure *graphite* preserves protein–protein and protein–metabolite edge directionality according to the database annotation

a protein-only graph (Fig. 1, panel D), just like in previous versions; (ii) a protein–metabolite graph (Fig. 1, panel B), hereafter called *mixed* graph and finally (iii) a metabolite–metabolite graph (Fig. 1, panel C). In all three cases, we start from the same pathway annotations and we extract all the interactions there described preserving both their types and directionality. At this stage *graphite* makes no distinction between protein–protein and protein–metabolite interactions: the software is designed to keep the pathway edges as close as possible to the definition in the original database. It is only later that we optionally introduce context–specific rules as described in Sales *et al.* (2012) to obtain multiple graph variants: we remove all metabolites for the protein graph and, similarly, all the proteins for the metabolite graph. We have designed data storage using a scheme that keeps only one copy of edges in common among the different variants of the same pathway. This significantly helps in reducing the disk footprint of the package.

To deal with identifier ambiguity (e.g. '1001' could either represent human cadherin 3 or a compound Phenethylamine) we changed the way *graphite* represents pathways. The edge table gained two new columns, `src_type` and `dest_type`, with the types of the nodes connected by the edge (e.g. '`ENTREZID`' and '`PUBCHEM`'), thus, the node list, was changed adding a prefix to each entry (e.g. '`ENTREZID: 1001`' or '`PUBCHEM: 1001`'). While the first change is backwards-compatible, the second is unfortunately going to require some manual adjustments to downstream software. All package functions (Table 1) have been updated to handle appropriately the new node types.

The new structure opened the door to an expansion of the information; apart from the six major databases we already tracked previously (Biocarta, HumanCyc, KEGG, NCI/Nature Pathway Interaction Database, Panther and Reactome), we further extended the collection with two resources explicitly dedicated to metabolites: SMPDB and PharmGKB. *graphite* now covers 14 different species, for a total of 21 116 pathways and over 22 million reactions. Pathways are collected and pre-processed at every BioConductor release. This guarantees the synchronization of the provided data with all other BioConductor annotation packages.

Table 1. Main functions of the new *graphite* package

Main functions	Description
`Pathway`	Retrieve pathway annotation
`PathwayGraph`	Convert a pathway into a graphNEL object
`ConvertIdentifiers`	Convert between node identifiers
`CytoscapePlot`	Plot the network in Cytoscape

The conversion of identifiers works both on proteins (for instance, from UniProt to ENSEMBL proteins) and on metabolites. In the latter case it supports the following annotations: CAS, CHEBI, KEGG compounds, KEGG drugs, KEGG glycans and PubChem. *graphite* can now export all its pathways as Cytoscape 3 networks. By default, it introduces different shapes to distinguish protein and metabolite nodes. The user can further customize the graphical style using the node and edge annotations that are automatically transferred between *graphite* and Cytoscape.

3 Application

The area of topological gene set analysis was pioneered by Draghici *et al.* (2007) and Tarca *et al.* (2009) and has subsequently seen a tremendous development. Today, there are over 20 such topological analysis methods [see Mitrea *et al.* (2013) for an extensive review]. Gene set analysis was devised in particular for gene expression data, however the new structure of *graphite* makes it possible to apply the same approaches to metabolic and mixed networks fostering the development of novel methods for data integration. Specifically the current *graphite* version gives easy access, through dedicated functions, to the SPIA (Tarca *et al.*, 2009), topologyGSA (Massa *et al.*, 2010) and clipper (Martini *et al.*, 2013) analyses. We will include additional topological methods when they will become available as part of the Bioconductor platform.

As an example, we report the results obtained combining gene expression and metabolite data on mixed metabolic pathways using *clipper* (Martini *et al.*, 2013) on breast cancer data (Terunuma *et al.*, 2014) with the KEGG and SMPDB databases. Terunuma *et al.* (2014) showed that the oncometabolite 2-hydroxyglutarate (2HG) preferentially accumulates in ER-negative breast tumours, that it associates with an increased activity of MYC pathway, a global increasing of DNA methylation and a poor prognosis.

The analysis conducted here via *graphite* highlights the 'Alanine, aspartate and glutamate metabolism' and 'Arginine biosynthesis' pathways. These enclose most of the genes and metabolites discussed in Terunuma *et al.* (2014) that follow the abundance pattern of 2HG such as N-acetyl amino acids, especially the mitochondrial N-acetyl-aspartate (NAA), the glutaminase (GLS1) protein and the aspartoacylase (ASPA), which is generally reduced in breast tumors but most significantly within the ER-negative tumor group which suggests a possible cause for increased tumor-associated NAA. Another *graphite* pathway worthy of note is 'D-Glutamine and D-glutamate metabolism' that contains L-Glutamine (CAS: 56-85-9) and L-Glutamic Acid (CAS: 56-86-0), two of the most altered metabolites between ER+ and ER- breast tumours. Finally using the SMPDB database we obtained 190 significant pathways among which we found 'The oncogenic action of 2-hydroxyglutarate' that is exactly the pathway describing the role of 2HG in tumour environment.

For more details, see the Supplementary File and the Vignette available at the BioConductor web page.

4 Conclusions

We released a major update of the Bioconductor package *graphite*. In this new version metabolites are explicitly tracked during the BioPax/KGML parsing step, providing two additional networks types: metabolite–only and mixed (protein–metabolite) networks. The number of supported databases has been expanded taken into consideration two resources explicitly dedicated to metabolites: SMPDB and PharmGKB. Ambiguity between mixed IDs (proteins and metabolites) has been addressed by changing slightly the appropriate data structures. At the time of this writing, *graphite* is one of the most comprehensive resources for protein and metabolite pathways in the R environment. The software belongs to top 5% of BioConductor packages by popularity with more than 10 000 downloads in 2017 alone. We believe the new layer about metabolites we have introduced will greatly improve the extraction of useful information from biological measurements and will expand the applicability of our package.

Acknowledgement

The authors thank Dr. Keisuke Ito, M.D., Ph.D. and Massimo Bonora, Ph.D for their critical discussions; Salvatore Milite for his feedback on metabolomic data analyses.

Funding

This work was supported by Italian Association for Cancer Research (IG17185 to CR).

Conflict of Interest: none declared.

References

Cottret,L. *et al*. (2010) Metexplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res*., **38**, W132–W137.

Draghici,S. *et al*. (2007) A systems biology approach for pathway level analysis. *Genome Res*., **17**, 1537–1545.

Martini,P. *et al*. (2013) Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Res*., **41**, e19.

Massa,M.S. *et al*. (2010) Gene set analysis exploiting the topology of a pathway. *BMC Syst. Biol*., **4**, 121.

Mitrea,C. *et al*. (2013) Methods and approaches in the topology-based analysis of biological pathways. *Front. Physiol*., **4**, 278.

Posma,J.M. *et al*. (2014) Metabonetworks, an interactive matlab-based toolbox for creating, customizing and exploring sub-networks from Kegg. *Bioinformatics*, **30**, 893–895.

Rodriguez-Martinez,A. *et al*. (2017) Metabosignal: a network-based approach for topological analysis of metabotype regulation via metabolic and signaling pathways. *Bioinformatics*, **33**, 773–775.

Sales,G. *et al*. (2012) Graphite–a bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, **13**, 20–22292714. PMID:

Tarca,A.L. *et al*. (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.

Terunuma,A. *et al*. (2014) Myc-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. *J. Clin. Investig*., **124**, 398–412.

Wishart,D.S. (2016) Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov*., **15**, 473.