BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 21 2014, pages 3139–3141
doi:10.1093/bioinformatics/btu501

Data and text mining Advance Access publication July 28, 2014

NetPathMiner: R/Bioconductor package for network path mining through gene expression

Ahmed Mohamed[1,*], Timothy Hancock[1,2], Canh Hao Nguyen[1] and Hiroshi Mamitsuka[1]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Japan and [2]Department of Computing and Information Systems, The University of Melbourne, Victoria, Australia

Associate Editor: Jonathan Wren

ABSTRACT

Summary: NetPathMiner is a general framework for mining, from genome-scale networks, paths that are related to specific experimental conditions. NetPathMiner interfaces with various input formats including KGML, SBML and BioPAX files and allows for manipulation of networks in three different forms: metabolic, reaction and gene representations. NetPathMiner ranks the obtained paths and applies Markov model-based clustering and classification methods to the ranked paths for easy interpretation. NetPathMiner also provides static and interactive visualizations of networks and paths to aid manual investigation.

Availability: The package is available through Bioconductor and from Github at http://github.com/ahmohamed/NetPathMiner

Contact: mohamed@kuicr.kyoto-u.ac.jp

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

Received on April 14, 2014; revised on June 30, 2014; accepted on July 17, 2014

1 INTRODUCTION

Mining subnetworks from genome-scale biological networks is an important step in biological data analysis because as their size and complexity increase, manual analysis becomes infeasible. Various methods for subnetwork mining from experimental data are available, where output subnetworks are expressed as sets of nodes (Vert and Kanehisa, 2003), frequent subgraphs (Chechik *et al.*, 2008) or linear paths (Hancock *et al.*, 2010). We focus on linear paths, which are particularly useful by carrying an intuitive meaning, as in metabolic reaction paths and signaling cascades.

Currently, software for mining paths from networks is hampered by two main challenges: (i) The lack of a universal parsing tool for different pathway types and database file formats. In fact, pathways can be diverse, such as metabolic or signaling, representing chemical conversions or protein interactions, respectively. Also, pathway information in databases can be accessed through different file formats, with each format representing pathway structure and annotation differently. (ii) The lack of effective visualizations for (a) different network types and (b) up to thousands of paths.

Currently available tools are either specific to particular pathway types or file formats. An earlier approach, *PathRanker* (Hancock *et al.*, 2010), is limited to mining paths from metabolic networks constructed from KGML files, providing static visualization on one network representation only. Tools like *Pathview* (Luo and Brouwer, 2013) and *rBiopaxParser* (Kramer *et al.*, 2013) also construct networks from only one file format.

We present NetPathMiner, an R package, to overcome the two main challenges. Figure 1 describes the process flow implemented in our package (see next section and Supplementary Materials for details). NetPathMiner implements a flexible module-based process flow for network path mining and visualization, which can be fully integrated with user-customized functions. NetPathMiner enables fast processing for all KGML, SBML and BioPAX formats into genome-scale metabolic or signaling networks (See Supplementary Figures S1 and S2 in detail). NetPathMiner also provides path highlighting over three different representations and interactive visualizations, making the analysis of even thousands of output paths possible.

2 MAIN FEATURES

2.1 Pathway file processing (Step 1 in Figure 1)

Table 1 summarizes key differences among file formats supported by NetPathMiner. To alleviate these differences, NetPathMiner processes files into either one of the two outputs, regardless of input formats: (i) Metabolic networks, in metabolic representation, as a bipartite graph with metabolites and reactions parts, and edges representing production or consumption of metabolites. (ii) Signaling networks, in gene representation where vertices and edges represent genes and their interactions. NetPathMiner can merge multiple files into a single network.

Network objects and annotations. We chose igraph (Csardi and Nepusz, 2006) to represent all constructed network objects in R to efficiently handle large graphs commonly encountered in biology and to allow NetPathMiner to integrate with other network analysis tools. NetPathMiner stores extracted annotations according to MIRIAM guidelines and implements an attribute fetcher using BridgeDb web service (van Iersel *et al.*, 2010) to convert between different annotations.

2.2 Network representations (Step 2)

For metabolic representation, NetPathMiner generates two other representations: (i) Reaction representation, where edges represent successive reactions, (ii) Gene representation, where vertices represent genes and edges represent participation in successive reactions. Multiple representations allow investigation of both chemical and genetic components of metabolic activity.

[*] To whom correspondence should be addressed.

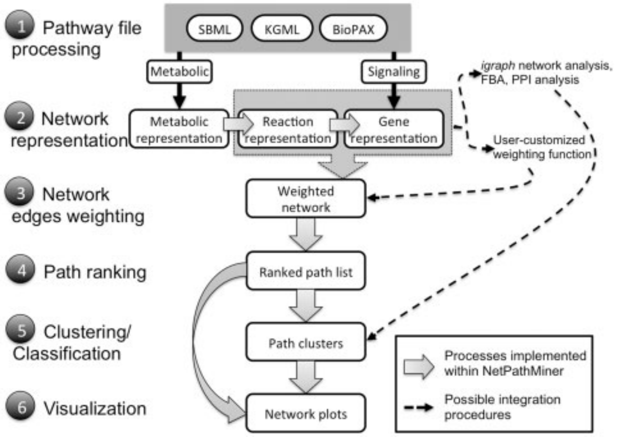
1 Pathway file processing
SBML
KGML
BioPAX
Metabolic
Signaling
igraph network analysis, FBA, PPI analysis
2 Network representation
Metabolic representation
Reaction representation
Gene representation
User-customized weighting function
3 Network edges weighting
Weighted network
4 Path ranking
Ranked path list
5 Clustering/ Classification
Path clusters
6 Visualization
Network plots
Processes implemented within NetPathMiner
Possible integration procedures

Fig. 1. NetPathMiner process flow chart

Table 1. Differences between major pathway file formats

Features	KGML	SBML	BioPAX
Number of pathways per file	One	One	One or more
Are metabolic reactions distinct?	Yes	No	Yes
Transport reactions	No	Yes[a]	Yes
Reaction kinetics	No	Yes	No
Cellular location	No	Yes	Yes
MIRIAM annotations	No	Yes	Yes
Databases	KEGG	Reactome, Biomodels, Recon X	PID, Reactome, BioCyc, Biocarta, WikiPathways

[a]Transport reactions are detected indirectly from reaction description.

2.3 Weighting networks and path ranking (Steps 3, 4)

Given a network in reaction or gene representations, edge weights can be computed by Pearson correlation, or a user-defined function, of expression profiles of adjacent genes. If gene expression is measured for different categories (control/treatment), edge weights can be computed for each category separately. From a weighted network, NetPathMiner ranks node sequences by two methods: (i) 'shortest path' (Hancock *et al.*, 2010) returning a list of k-maximum weighted paths, (ii) '*P*-value' (Hancock *et al.*, 2012) returning the paths of which the sum of edge weights is significantly higher than that of random paths. Path ranking can be used independently or as a part of the entire process.

2.4 Clustering and classification of paths (Step 5)

NetPathMiner provides machine learning methods to summarize output paths. Using Markov mixture models, paths can be clustered or classified according to their structure or association with a response label, respectively. Both methods are adopted from *PathRanker* (Hancock *et al.*, 2010).

2.5 Visualization (Step 6)

NetPathMiner provides both static and interactive visualizations of ranked paths using annotation information and machine learning techniques, making manual investigation easier. Figure 2a-c show a visualization example of different graph representations using the output of the last step, allowing users to examine metabolic regulation at different biological system levels. Visualization function matches vertices across all input representations and plots them using the same layout. To make visualization of a huge number of paths clearer, NetPathMiner assigns the same color to all paths in each obtained cluster, and assigns the same color to vertices within the same cellular compartment (not shown). Figure 2d and e show vertices in each path as well as probability of each path belonging to clusters. NetPathMiner also supports interactive visualization in Cytoscape by either exporting networks in GML format or using *RCytoscape* (Shannon *et al.*, 2013), which allows thorough investigation of vertex annotations and full customization of network colors and layout.

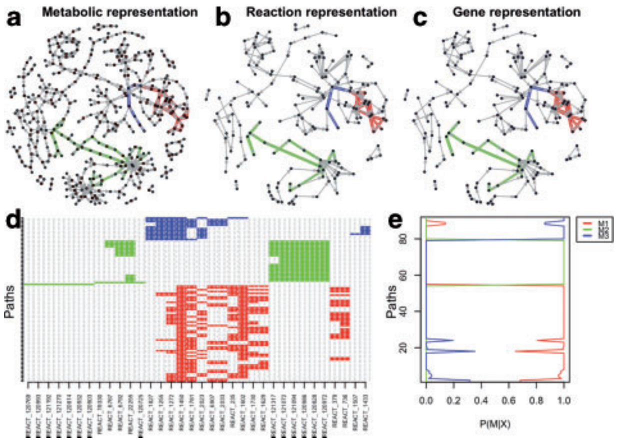

Fig. 2. NetPathMiner path visualization. Top 100 paths on Reactome's carbohydrate metabolism network, grouped into three clusters (red, blue, green). (**a–c**) Metabolic, reaction and gene representations. (**d**) Paths (rows) and their reactions (columns). (**e**) Probability that a top path belongs to a cluster

3 CONCLUSION

We present NetPathMiner, an easy-to-use R package for network path mining. NetPathMiner constructs genome-scale networks from KGML, SBML and BioPAX files, overcoming the limitations of current tools. NetPathMiner summarizes output paths using machine learning methods and provides different visualizations, easing manual investigations. All modules in NetPathMiner can be used independently, particularly the *pathway file processing* module, and replaced with user-customized modules. Future developments include supporting other network representations, such as rule-based modeling formats, as inputs and outputs for broader integration.

Funding: This work is partially supported by KAKENHI (2430054) of MEXT, Japan.

Conflict of Interest: none declared.

REFERENCES

Chechik,G. *et al.* (2008) Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nat. Biotechnol.*, **26**, 1251–1259.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *Int. J. Complex Sys.*, **1695**.

Hancock,T. *et al.* (2010) Mining metabolic pathways through gene expression. *Bioinformatics*, **26**, 2128–2135.

Hancock,T. *et al.* (2012) Identifying neighborhoods of coordinated gene expression and metabolite profiles. *PLoS One*, **7**, e31345.

Kramer,F. *et al.* (2013) rBiopaxParser—an R package to parse, modify and visualize biopax data. *Bioinformatics*, **29**, 520–522.

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

Shannon,P.T. *et al.* (2013) RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics*, **14**, 217.

van Iersel,M.P. *et al.* (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC bioinformatics*, **11**, 5.

Vert,J. and Kanehisa,M. (2003) Extracting active pathways from gene expression data. *Bioinformatics*, **19** (**Suppl. 2**), ii238–ii244.