BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 4 2013, pages 520–522
doi:10.1093/bioinformatics/bts710

Databases and ontologies

Advance Access publication December 28, 2012

rBiopaxParser—an R package to parse, modify and visualize BioPAX data

Frank Kramer[1,*], Michaela Bayerlová[1], Florian Klemm[2], Annalen Bleckmann[1,2] and Tim Beißbarth[1]

[1]Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32 and [2]Department of Hematology/Oncology, University Medical Center Göttingen, Robert-Koch-Straße 40, 37073 Göttingen, Germany

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Biological pathway data, stored in structured databases, is a useful source of knowledge for a wide range of bioinformatics algorithms and tools. The Biological Pathway Exchange (BioPAX) language has been established as a standard to store and annotate pathway information. However, use of these data within statistical analyses can be tedious. On the other hand, the statistical computing environment R has become the standard for bioinformatics analysis of large-scale genomics data. With this package, we hope to enable R users to work with BioPAX data and make use of the always increasing amount of biological pathway knowledge within data analysis methods.

Results: rBiopaxParser is a software package that provides a comprehensive set of functions for parsing, viewing and modifying BioPAX pathway data within R. These functions enable the user to access and modify specific parts of the BioPAX model. Furthermore, it allows to generate and layout regulatory graphs of controlling interactions and to visualize BioPAX pathways.

Availability: rBiopaxParser is an open-source R package and has been submitted to Bioconductor.

Contact: frank.kramer@med.uni-goettingen.de

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

Received on August 31, 2012; revised on November 22, 2012; accepted on December 16, 2012

1 INTRODUCTION

The past years have seen an enormous increase in biological knowledge about cellular signalling and regulatory pathways, which is stored in numerous databases (Bader *et al.*, 2006). However, the fragmentation of the available pathway data led to the need for a standardized language to ease the knowledge exchange between scientists. The Biological Pathway Exchange (BioPAX) language is an ontology that models biological pathway concepts and their relationships (Demir *et al.*, 2010). Implemented in the Web Ontology Language (OWL), an RDF/XML-based format, it allows the user to encode pathway knowledge in a well-documented and standardized way.

The integration of biological knowledge stored in these databases into high level analysis methods for genomics experiments is a fundamental issue. Bioinformaticians can integrate pathway knowledge, collected and curated by the scientific community, into statistical analysis as previous knowledge. Examples for methods for high-level data analysis that require previous pathway knowledge are Gene Set Enrichment Analysis (Beißbarth, 2006; Geistlinger *et al.*, 2011), specialized classification algorithms for personalized medicine (Johannes *et al.*, 2010) or network reconstruction algorithms (Bender *et al.*, 2011; Fröhlich *et al.*, 2009). Many of these algorithms are implemented using the R Project for Statistical Computing. Although the packages *NCIgraph*, *KEGGgraph* and *graphite* on Bioconductor (Gentleman *et al.*, 2004; Zhang and Wiemann, 2009) offer exports of signalling networks from specific databases, no packages exist that allow the user to conveniently work with BioPAX data in R.

We created the R package rBiopaxParser to ease use and integration of pathway data in the BioPAX format within R. Furthermore, the use of R scripts allows reproducible handling and manipulation of pathway data as well as the integration of these data into high-level statistical analysis routines.

2 BIOPAX

The BioPAX ontology includes classes for the annotation of molecules, molecular interactions and pathways. The language definition as well as further information, manuals, tools and examples can be found at http://www.biopax.org. A number of online databases provide users with a data export in the BioPAX format, often free of charge, for example, the National Cancer Institute (NCI), which offers exports of the popular databases Pathway Interaction Database (Schaefer *et al.*, 2009), Biocarta (Nishimura, 2001) and Reactome (Croft *et al.*, 2010).

3 FEATURES

The core functions of our package are designed to parse BioPAX data into R and to export it back to the OWL format. Many convenience functions help the user to view and modify specific parts of the BioPAX model. The rBiopaxParser currently supports BioPAX Level 2 and 3. More detailed documentation and examples can be found in the package vignettes and manual.

*To whom correspondence should be addressed.

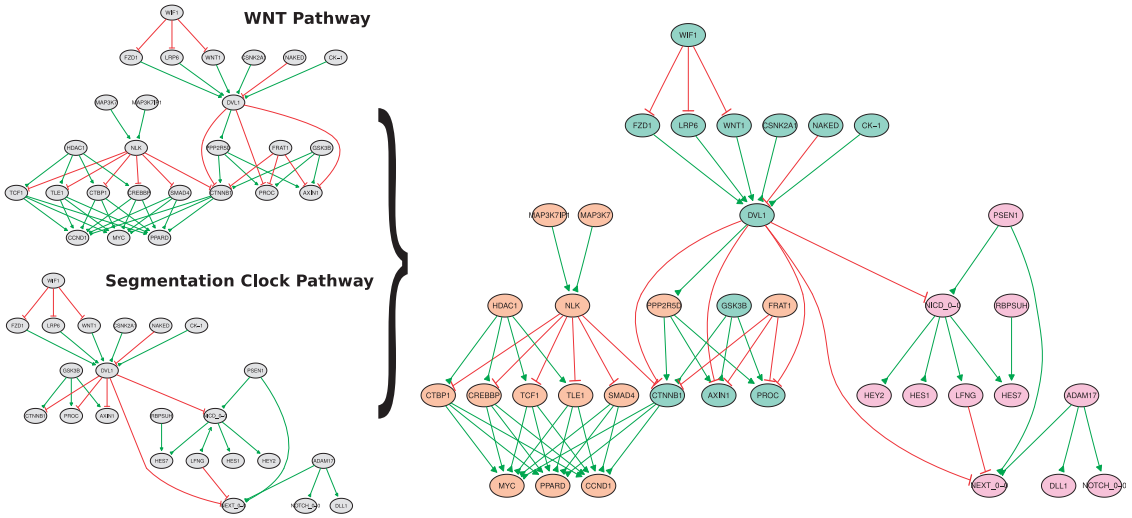

Fig. 1. Merged regulatory network of the WNT and the segmentation clock pathway parsed from the NCI Biocarta export. Green nodes are overlapping nodes found in both pathways; orange nodes are only present in the WNT pathway; and pink nodes are only present in the segmentation clock pathway

3.1 Downloading and parsing BioPAX

The function *downloadBiopaxData* has been created to directly download from Web resources like the NCI databases. The function *readBiopax* allows the user to parse arbitrary valid BioPAX models from the file system, which are then available within R.

3.2 Internal data structure

After successful parsing, the BioPAX data will be stored as a *data.frame* in an R object of class *biopax*. All instances and their properties are stored within this *data.frame*. The inheritance structure of BioPAX classes and the list of properties available for each class are also stored within the package. Functions to check for super- and subclass relationships as well as properties are available. The *biopax* object can be passed on to various functions to modify and extract information. A detailed report of the internal data representation can be found in the Supplementary Material.

3.3 Accessing and modifying BioPAX data

The parsed BioPAX data can be easily accessed with the help of functions like *listPathways*, *listPathwayComponents* and *selectInstances*. The biopax model can be directly modified by functions to add or remove pathways, interactions and molecules. Furthermore, function *mergePathway* allows the user to specify two existing pathways that should be merged into a single new pathway. Advanced users can access the BioPAX data directly and edit it to their liking. Finally, the modified BioPAX models can be exported in the BioPAX OWL format using *writeBiopax*.

3.4 Regulatory graphs and visualization

Encoded BioPAX data can include many biological processes like translocation, modification or transcription. However, it can be desirable to focus only on regulatory knowledge about molecules activating or inhibiting each other. The function *pathway2regulatoryGraph* achieves this by generating a graph from all regulatory interactions within a certain pathway. Additional functions for lay-outing, merging or intersecting graphs are also implemented (Fig. 1).

4 IMPLEMENTATION

The rBiopaxParser package takes advantage of the data processing and visualization tools that the R projects provide. Initial data downloading capabilities for NCI data exports are implemented using the *RCurl* package, giving the user access to large amounts of data. The pre-processing and parsing of BioPAX input files, as well as the generation of BioPAX output, are accomplished via the *XML* package. The regulatory graphs that can be generated from the internal BioPAX model are generated and visualized using the *graph* and *Rgraphviz* packages (Carey *et al.*, 2005).

5 CONCLUSION

The rBiopaxParser is a freely available R package that allows the user to parse, modify and visualize BioPAX models. Regulatory graphs can be extracted and used for further analyses. With this R package, we also hope to enable users to apply newly developed algorithms to the always increasing amount of knowledge about biological pathways. By implementing a BioPAX parser, we offer a tool for R users to ease the task of integrating existing pathway resources in statistical analyses and encourage the use of a standardized way to encode pathway knowledge.

Funding: Deutsche Forschungs-gemeinschaft (clinical research group KFO179 and research group FOR942); German Ministry of Education and Research (BMBF) grant MetastaSys from the platform e:Bio.

Conflict of Interest: none declared.

REFERENCES

Bader,G.D. *et al.* (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.*, **34**, D504–D506.

Beißbarth,T. (2006) Interpreting experimental results using gene ontologies. *Methods Enzymol.*, **411**, 340–352.

Bender,C. *et al.* (2011) Inferring signaling networks from longitudinal data using sampling based approaches in the R-package 'ddepn'. *BMC Bioinformatics*, **12**, 291.

Carey,V.J. (2005) Network structures and algorithms in Bioconductor. *Bioinformatics*, **21**, 135–136.

Croft,D. *et al.* (2010) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.

Demir,E. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 935–942.

Fröhlich,H. *et al.* (2008) Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data. *Bioinformatics*, **24**, 2650–2656.

Geistlinger,L. *et al.* (2011) From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, **27**, i366–i373.

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

Johannes,M. *et al.* (2010) Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics*, **26**, 2136–2144.

Nishimura,D. (2001) BioCarta. *Biotech. Software Internet Rep.*, **2**, 117–120.

Schaefer,C.F. *et al.* (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.

Zhang,J.D. and Wiemann,S. (2009) KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics*, **25**, 1470–1471.