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

Frank Kramer¹,*, Michaela Bayerlová¹, Florian Klemm², Annalen Bleckmann¹,² and Tim Beißbarth¹

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

Associate Editor: Dr. 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 this 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

1 INTRODUCTION

The last years have seen an enormous increase in biological knowledge about cellular signaling 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 prior knowledge. Examples for methods for high level data analysis that require prior 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 (Fröhlich et al. 2009; Bender et al. 2011). 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 signaling 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 utilization of R scripts allows reproducible handling and manipulation of pathway data as well as the integration of this 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.

3.1 Downloading & 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.

To whom correspondence should be addressed.

© The Author (2012). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
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 & 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 (See Figure 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 preprocessing and parsing of BioPAX input files, as well as the generation of BioPAX output, is accomplished via the `XML` package. The regulatory graphs which 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: This work was supported by the Deutsche Forschungsgemeinschaft (clinical research group KFO179 and research group FOR942) as well as from the German Ministry of Education and Research (BMBF) grant MetaSys from the platform e:Bio.

REFERENCES