Systems Biology

BioPAX-Parser: parsing and enrichment analysis of BioPAX pathways

Giuseppe Agapito [1,3,*], Chiara Pastrello [4], Pietro Hiram Guzzi [2,3], Igor Jurisica [4,5] and Mario Cannataro [2,3]

[1] Department of Legal, Economic and Social Sciences, Magna Graecia University of Catanzaro, Catanzaro, Italy
[2] Department of Medical and Surgical Sciences, Magna Graecia University of Catanzaro, Catanzaro, Italy
[3] Data Analytics Research Center, Magna Graecia University of Catanzaro, Catanzaro, Italy
[4] Krembil Research Institute, University Health Network, Toronto, Canada
[5] Departments of Medical Biophysics and Computer Science, University of Toronto, Canada.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Biological pathways are fundamental for learning about healthy and disease states. Many existing formats support automatic software analysis of biological pathways, for examples BioPAX (Biological Pathway Exchange). Although some algorithms are available as web application or standalone tools, no general graphical application for the parsing of BioPAX pathway data exists. Also, very few tools can perform Pathway Enrichment Analysis (PEA) using pathway encoded in the BioPAX format. To fill this gap we introduce *BiP*, an automatic and graphical software tool aimed at performing the parsing and accessing of BioPAX pathway data, along with pathway enrichment analysis by using information coming from pathways encoded in BioPAX.

Availability: BiP is freely available for academic and non-profit organizations at https://gitlab.com/giuseppeagapito/bip under the LGPL 2.1, the GNU Lesser General Public License.

Contact: agapito@unicz.it

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

1 Introduction

Biological Pathways are human representations of coordinated molecular actions within a cell, that can include genes, proteins and small molecules. They can be roughly classified into three categories: Signalling Pathways, Metabolic Pathways, and Regulatory Pathways. Pathways can be represented using structured file formats (i.e., Extensible Markup Language (XML), Web Ontology Language (OWL), Resource Description Framework (RDF), etc.) or unstructured text files. The most used structured file format is the Biological Pathway exchange (BioPAX) categorized in -LEVELS 1,2,3 (Demir *et al.* (2010)). To efficiently manage BioPAX files, the BioPAX consortium has made available the *Paxtools* library. *Paxtools* is available as Java software library (Demir *et al.* (2013)) and as R package *PaxtoolsR* (Luna *et al.* (2015)). Both versions can write, merge, validate, and convert BioPAX files to other formats. Both Paxtools versions can be embedded in a software tool as a library, allowing programmers to use the code and algorithms in their software. Besides, Paxtools algorithms can be run from the command line. To take full advantages of all the resources provided by Paxtools the user is required to have some programming skills either in Java or R. Currently, no software enables non-developers in accessing specific pathway information. To fill this gap we present BioPAX-Parser (BiP), a software tool to simply and efficiently perform access to the pathway information encoded in BioPAX Level 3 (coming from any available public or private database). BiP comes with graphic commands that allow users to explore the proteins/genes in BioPAX pathway, and to annotate the identified proteins/genes by automatically retrieving information from the UniProt database. In addition, BiP allows users to perform Pathway Enrichment Analysis (PEA). PEA is obtained by computing the Hypergeometric function to identify relevant pathways, starting from a list of genes of interest (Subramanian *et al.* (2005)). Paxtools cannot perform PEA exploiting the information present in the BioPAX files. On the other hand, tools for PEA are available as stand-alone as well as web tool applications. Stand-alone tools include TPEA (Yang *et al.* (2017)), SPIA (Tarca *et al.* (2009)), PathVisio (Kutmon *et al.* (2015)), and CePa (Gu and Wang (2013)), whereas ComPath (Domingo-Fernández *et al.* (2018)), ReactomeAnalyzer (Fabregat *et al.* (2015)), DavidAnalyzer (Huang *et al.* (2007)), and pathDIP (Rahmati *et al.* (2016)) belong to the web tools category. The listed stand-alone tools are distributed as R-packages, whose use requires some programming skill, and in their current release can perform PEA only using data coming from the KEGG database. Web applications are easy to use but pose some limitation on the maximum number of genes/proteins that can be used in a single experiment, as well as allow users to work only with preselected databases and do not allow users to upload their own pathway data. Besides, both software categories can perform PEA only by using data coming from a single database, with the exception of pathDIP. In the current version, pathDIP can perform PEA aggregating data coming from 22 different databases, but cannot manage BioPAX data directly. PathVisio can perform PEA using BioPAX files (through additional plug-in), but cannot aggregate data coming from multiple databases. Hence, to perform PEA using data coming from different data sources compliant with BioPAX format, it is necessary to use more than a software tool. BiP instead, allows users to perform PEA using multiple BioPAX data sources in the same analysis, simply loading an input list of biological entities (e.g., genes, proteins). The automatic mapping between genes and proteins symbols identifiers, is done using a conversion BridgeDB (API) (van Iersel *et al.* (2010)) allowing BiP to scan BioPAX data using Ensembl, NCBI Gene, Uniprot. BiP can automatically download BioPAX data from the PatwayCommons database (Cerami *et al.* (2011)), that can then be used locally by the users to perform multiple PEA.

2 BiP Software Tool

BiP (BioPAX Parser) is a graphical software tool fully developed using the cross-platform language *Java 8*, which is available for many operating systems. BiP can be used to acquire information from pathways encoded in BioPAX format. The BioPAX files reader in BiP is developed using the *RDF-Reader* component available within the *Jena*[1] library version 3.1.0. We chose to use a general RDF reader instead of the Paxtools API, because we don't need to use the reasoning to infer types, and we wanted to speed up the loading and extraction process from BioPAX files. Pathway data in BiP are mapped using an hash-function making it possible to reduce the memory consumption, using a linked data structure emulating a graph-based representation. This solution makes it more efficient to go across the pathway improving the automatic pathway data management during the enrichment analysis as well as to obtain reliable pathway information, i.e., the proteins or biochemical reaction list of a specific pathway. All functionalities provided by BiP are available as graphical-commands, that can be used by non-developers to obtain information about the components stored into the BioPAX file, e.g., to identify the proteins/genes belonging to a specific pathway and/or to a biochemical reaction. The BiP user guide describing how to use BiP functionalities, is available in Supplementary materials. In addition, BiP can be used to compute PEA employing data coming from several databases that use BioPAX-Level3 format, enabling BiP users to take advantage of the growing number of databases complying only with the BioPAX-Level3 standard. PEA in BiP is performed through a statistical engine implemented in Java able to compute the Hypergeometric test to determine if the genes of interest are enriched in a specific pathway. To reduce the number of false positives, BiP also implements two methods of multiple test adjustment, *False Discovery Rate* (*FDR*) and *Bonferroni*, to correct the computed *p-value*. BiP allows users to perform PEA by using BioPAX pathway data manually downloaded from web databases, i.e., Reactome, PathwayCommons, or to analyze their own BioPAX pathway, as well as it provides an automatic procedure for the collection and storage of BioPAX data coming from the PathwayCommons database (data-collector module). BiP supports multi-threaded processes for fast parallel computation on multicore processors. The comparison of BiP versus other tools is performed by considering the availability of GUI (Graphical User Interface), the required Programming Skills (PS), the support of PEA, the capability to identify, show, and annotate the proteins/genes involved in a single pathway (SPPA), the capability to identify, show, and annotate all the protein in a BioPAX file (FPA), the possibility to perform PEA using multiple databases encoding data in BioPAX (PEAMDBs), and finally if there are limits of use (LofU) i.e., a limited number of genes to use in a single experiment, or a limited number of analysis per day. Table 1 summarizes the functionalities available in BiP versus the other PEA software tools.

Table 1. Summary of the features available in the pathway enrichment tools.

Tools	GUI	PS	PEA	SPPA	FPA	PEAMDBs	LofU
BiP	Yes	No	Yes	Yes	Yes	Yes	No
Paxtool-R	No	Yes	No	No	No	No	No
CePa	No	Yes	Yes	No	No	No	No
SPIA	No	Yes	Yes	No	No	No	No
TPEA	No	Yes	Yes	No	No	No	No
PathVisio	Yes	No	Yes	No	No	Yes	No
David	Yes	No	Yes	Yes	No	No	Yes
Reactome	Yes	No	Yes	Yes	Yes	No	No
ComPath	Yes	No	Yes	Yes	No	No	No
pathDIP	Yes	No	Yes	Yes	No	Yes	No

GUI = Graphical User Interface, PS = Programming Skills, PEA = Pathway Enrichment Analysis, SPPA = Single pathway proteins analysis, FPA= Full Proteins analysis, PEAMDBs = PEA from Multiple databases, LofU= Limits of use

3 Conclusion

In summary, we presented BioPAX-Parser (BiP), a software tool to parse, access pathways data and to perform PEA using data encoded in BioPAX. BiP is the first available tool that allows users to i) work with BioPAX files without needing extensive programming knowledge; and ii) perform PEA using data from multiple different BioPAX pathway databases.

References

Cerami, E. G. *et al.* (2011). Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, **39**(Database issue), D685–D690.

Demir, E. *et al.* (2010). The biopax community standard for pathway data sharing. *Nature Biotechnology*, **28**(9), 935–942.

Demir, E. *et al.* (2013). Using biological pathway data with paxtools. *PLoS computational biology*, **9**(9), 1–5.

Domingo-Fernández, D. *et al.* (2018). Compath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *npj Systems Biology and Applications*, **4**(1), 43.

Fabregat, A. *et al.* (2015). The Reactome pathway Knowledgebase. *Nucleic Acids Research*, **44**(D1), D481–D487.

Gu, Z. and Wang, J. (2013). CePa: an R package for finding significant pathways weighted by multiple network centralities. *Bioinformatics*, **29**(5), 658–660.

[1] https://jena.apache.org

Huang, D. W. *et al.* (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, **35**(Web Server issue), W169–W175.

Kutmon, M. *et al.* (2015). Pathvisio 3: An extendable pathway analysis toolbox. *PLOS Computational Biology*, **11**(2), 1–13.

Luna, A. *et al.* (2015). Paxtoolsr: pathway analysis in r using pathway commons. *Bioinformatics*, **32**(8), 1262–1264.

Rahmati, S. *et al.* (2016). pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis. *Nucleic Acids Research*, **45**(D1), D419–D426.

Subramanian, A. *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Tarca, A. L. *et al.* (2009). A novel signaling pathway impact analysis. *Bioinformatics (Oxford, England)*, **25**(1), 75–82.

van Iersel, M. P. *et al.* (2010). The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, **11**(1), 5.

Yang, Q. *et al.* (2017). Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. *Briefings in Bioinformatics*, **20**(1), 168–177.