*Bioinformatics*, 34(4), 2018, 703–704
doi: 10.1093/bioinformatics/btx660
Advance Access Publication Date: 18 October 2017
Applications Note

OXFORD

Systems biology

# PyBEL: a computational framework for Biological Expression Language

**Charles Tapley Hoyt[1,2,*], Andrej Konotopez[1] and Christian Ebeling[1]**

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany and [2]Department of Life Science Informatics, Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53113, Germany

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on July 5, 2017; revised on September 28, 2017; editorial decision on October 16, 2017; accepted on October 17, 2017

## Abstract

**Summary:** Biological Expression Language (BEL) assembles knowledge networks from biological relations across multiple modes and scales. Here, we present PyBEL; a software package for parsing, validating, converting, storing, querying, and visualizing networks encoded in BEL.
**Availability and implementation:** PyBEL is implemented in platform-independent, universal Python code. Its source is distributed under the Apache 2.0 License at https://github.com/pybel.
**Contact:** charles.hoyt@scai.fraunhofer.de
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Among the most popular modeling and data exchange languages in systems biology are currently the Biological Pathways Exchange (BioPAX), Systems Biology Markup Language (SBML) and Biological Expression Language (BEL). BioPAX captures metabolic, signaling, molecular, gene-regulatory, and genetic interaction networks (Hucka *et al.*, 2003); SBML accommodates mathematical models of biochemical networks, cellular signaling, and metabolic pathways (Demir *et al.*, 2010); and BEL assembles qualitative causal and correlative relations between biological entities across multiple modes and scales, with full provenance information including namespace references, relation provenance (citation and evidence), and biological context-specific relation metadata (anatomy, cell, disease etc.) (Slater, 2014).

Although there exist several software packages for BioPAX and SBML, the ecosystem of open-source software for BEL is much more limited. An assessment of previous software (see Supplementary Table S3) shows there is an unmet need for easily installable, stable, facile software that parses modern BEL and provides programmatic access to a data container that enables the resulting network to be extended, queried, manipulated, analyzed, and visualized. Furthermore, a converter between common data formats is needed to enable re-usability and interoperability between general and BEL-specific software for network analysis and visualization.

Here, we present PyBEL; a software package designed to fulfill each of these needs.

## 2 Software architecture

The PyBEL software package consists of five main components: (i) network data container, (ii) parser and validator, (iii) network database manager, (iv) data converter and (v) network visualizer.

Although a graph refers to an abstraction for a set of objects (i.e. nodes) and their relations (i.e. edges), its instantiation in a real-world application is often called a network. We provide an implementation of a directed multigraph (i.e. a graph whose edges have directionality and any given pair of nodes may have multiple edges) that maps the biological entities and concepts in the subjects and objects of BEL relations to nodes in a network and their relations, with corresponding metadata, to edges. We extended the MultiDiGraph class from `NetworkX` (http://networkx.github.io) to enable users direct access to their suite of network algorithms and static visualizations to support their further development into biologically meaningful analyses.

The parser performs tokenization, lexical analysis, parsing, and validation on each of the three sections of BEL documents (see Supplementary Figs S1 and S2). Callbacks are used to annotate the entries in the document metadata section to a network instance, download and store the resources referenced in the definitions section,

maintain a list of current annotations from SET statements, and parse BEL relations to populate a network instance with the corresponding nodes, edges, and their metadata from the current internal state. Although relations' syntax is implicitly validated, the semantics of their subjects' and objects' identifiers are validated against the references from the definitions section. Finally, feedback is provided to users to support thoughtful re-curation, which could lead to more robust knowledge assemblies and enable more reproducible science.

Namespaces and networks are cached with a relational database to improve the speed of validation and access to data. Although relational databases lack the faculty for applying network algorithms, they provide indexing functionality that enables complicated queries and filters over the nodes, edges, and metadata of increasingly large collections of networks. For example, this could help identify intersections and potential cross-talk between disease-specific networks.

We implemented lossless converters for common file formats including Node-Link JSON, JGIF, CX, and binary as well as for database formats including SQL, Neo4J, and NDEx. We also provide lossy exporters to Excel, CSV, SIF, XGMML, and GSEA to facilitate usage in other programs. Notably, we have deferred implementing a RDF (Resource Description Framework) converter until improvements are made to the existing BEL to RDF mapping and its documentation (https://wiki.openbel.org). Future work will also include converters for BioPAX and SBML. See Supplementary Tables S1 and S2 for more detailed descriptions of each format.

Networks can be exported for visualization in Cytoscape or uploaded to NDEx (Pratt *et al.*, 2015) to take advantage of its viewer and simple query interface. Alternatively, we provide an interactive network explorer tailored to BEL networks (appropriate node coloring, metadata pop-ups etc.) that can be directly embedded as HTML in email, Jupyter Notebook, or a web application. It has already been used to produce visualizations in the NeuroMMSig Web Service (Domingo-Fernández *et al.*, 2017). Supplementary Figures S3–S5 present these visualizations side-by-side. In addition to their programmatic interfaces, the parser, storage, conversion, and visualization features are exposed via a command line tool.

## 3 Case study

The PyBEL suite includes functions for querying and mutating networks with which it implements state-of-the-art algorithms for over-representation analysis, functional class scoring, and pathway topological analysis of BEL networks such as Reverse Causal Reasoning (Catlett *et al.*, 2013). Figure 1 presents a case study in which a novel heat diffusion work flow was used to assess the observed impact on biological processes from differential gene expression in Alzheimer's disease (AD). Technical documentation is included in the Supplementary Material.

## 4 Discussion

Even after its v2.0 update, BEL does not yet explicitly specify many concepts in molecular biology such as epigenetic information (Irin *et al.*, 2015). The inevitability of language evolution prompted us to develop the parser in modules so that new syntax could be proposed and implemented quickly. As a proof of concept, a syntax extension for gene modifications is included in the package by default.

Historically, BEL has used a custom namespace file format, but the creation and maintenance of biological terminologies has tended towards using OWL (Web Ontology Language). Furthermore, many domains (e.g. SNPs) are growing too large to enumerate during semantic integration and validation. The modular architecture of the parser enables easy implementation of new definition file formats, external validation services, or even alternative namespace definition schemes to address these issues.

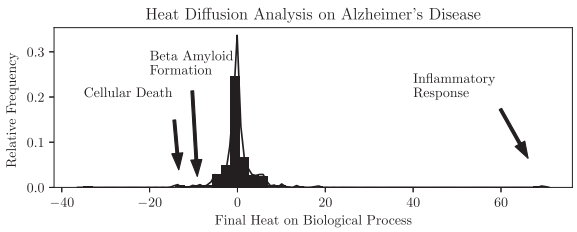


**Fig. 1.** Plotted is the distribution of the final heat on biological processes from the NeuroMMSig AD Knowledge Assembly (Domingo-Fernández *et al.*, 2017) following heat diffusion analysis with a differential gene expression experiment from the brains of AD patients (E-GEOD-5281, Liang *et al.*, 2007). The significant down-regulation of biological processes related to inflammatory response (heat = 69) and up-regulation of cellular death (heat = −13) and beta-amyloid formation (heat = −9) match common clinical observations and serve as a validation for this approach

Although BEL is often used to formalize knowledge curated from unstructured sources, our software also enables the integration of knowledge from structured sources. For example, existing solutions for resolving equivalences across namespaces rely on the creation and hosting of extensive lookup tables. Alternatively, the parser could be extended with a dedicated syntax and draw equivalencies directly from OWL.

Finally, we plan to present this software as a web service to enable a wider audience of researchers across disciplines to validate, explore, and analyze their BEL networks.

## Acknowledgements

We thank Sumit Madan and Scott Colby for their advice and feedback.

## Funding

This work was supported by the European Union/European Federation of Pharmaceutical Industries and Associations (EFPIA) Innovative Medicines Initiative Joint Undertaking under AETIONOMY [grant number 115568], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution.

*Conflict of Interest*: none declared.

## References

Catlett,N. *et al.* (2013) Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics*, **14**, 340.

Demir,E. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 1308–1308.

Domingo-Fernández,D. *et al.* (2017) Multimodal Mechanistic Signatures for Neurodegenerative Diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics (Oxford, England)*, btx399.

Hucka,M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, **19**, 524–531.

Irin,A.K. *et al.* (2015) Computational Modelling Approaches on Epigenetic Factors in Neurodegenerative and Autoimmune Diseases and Their Mechanistic Analysis. *J. Immunol. Res.*, **2015**, doi:10.1155/2015/737168.

Liang,W.S. *et al.* (2007) Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol. Genomics*, **28**, 311–322.

Pratt,D. *et al.* (2015) NDEx, the network data exchange. *Cell Syst.*, **1**, 302–305.

Slater,T. (2014) Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discov. Today*, **19**, 193–198.