PathMe: Merging and exploring mechanistic pathway knowledge

Daniel Domingo-Fernández[1,2,*], Sarah Mubeen[1,2], Josep Marín-Llaó[1], Charles Tapley Hoyt[1,2], and Martin Hofmann-Apitius[1,2]

1. Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53754, Germany
2. Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany

***Corresponding Author**: Domingo-Fernández, D., Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany. Telephone details: +49 2241 14-2269. Fax: +49 2241 14-2656

Abstract

The complexity of representing biological systems is compounded by an ever-expanding body of knowledge emerging from multi-omics experiments. A number of pathway databases have facilitated pathway-centric approaches that assist in the interpretation of molecular signatures yielded by these experiments. However, the lack of interoperability between pathway databases has hindered the ability to harmonize these resources and to exploit their consolidated knowledge. Such a unification of pathway knowledge is imperative in enhancing the comprehension and modelling of biological abstractions. Here, we present PathMe, a Python package that transforms pathway knowledge from three major pathway databases into a unified abstraction using Biological Expression Language as the pivotal, integrative schema. PathMe is complemented by a novel web application which allows users to comprehensively explore pathway cross-talks and compare areas of consensus and discrepancies.

Availability: PathMe's source code is available at https://github.com/ComPath/PathMe under the Apache 2.0 license. We provide a freely accessible deployment of the PathMe Viewer at https://pathme.scai.fraunhofer.de/.

Introduction

Pathway databases have emerged as comprehensive resources to support the interpretation of data-driven approaches yielded by genome-scale experiments. While they have embraced standard file formats and schemata in order to facilitate the exchange of pathway knowledge, each has chosen a different one, such as SBML or BioPAX (Hucka *et al.*, 2003; Demir *et al.*, 2010), based on their respective subjects and scopes. Therefore, resources such as Pathway Commons and graphite (Cerami *et al.*, 2011; Sales *et al.*, 2018) have been created with the primary intention to integrate pathway knowledge from multiple databases.

While they have succeeded in accumulating and increasing the availability of database content, there has not yet been a systematic evaluation that investigates the degree of overlap or the amount of agreements/discrepancies in related or equivalent pathways from different databases. Previous comprehensive comparisons of database content were restricted to single or small sets of pathways because of the considerable amount of manual intervention (e.g., entity/relationship normalization, image reconstruction, etc.) required to shed light on the degree of overlap of equivalent pathways (Stobbe *et al.,* 2011; Chowdhury and Sarkar, 2015). Conversely, conducting a systematic comparison would involve harmonization of entities and biological interactions across databases and minimizing pathway information loss whilst accommodating databases into an interoperable schema (i.e., retain most of the different biological abstractions that each database offers in the transformation process). Finally, continual updates in pathway definitions introduced by novel findings (Wadi *et al.*, 2016) motivate an approach that can be reproduced regularly to evaluate how pathway knowledge changes.

Here, we introduce PathMe, an extensible package that harmonizes multiple databases using Biological Expression Language (BEL) as a common interoperable schema and enables pathway knowledge evaluation and exploration powered by a web application.

Implementation

2.1 Integrating knowledge across pathway databases

Integrating pathway knowledge from multiple databases first requires transforming the content of each database into a common underlying schema. While multiple triple-based formats can be used to formalize pathways in system biology, we adopted BEL as the pivotal unifying schema since it provides a reasonable trade-off between expressivity and standardized organization. Subsequently, we implemented parsers for each of three major databases: KEGG, Reactome, and WikiPathways (Kanehisa *et al*., 2016; Fabregat *et al.*, 2017; Slenter *et al.*, 2017) that acquire the pathway information and serialize it to BEL.

Because the main goal of PathMe is to enable a direct comparison and exploration of pathways from different databases, the parsers harmonize molecular entities to a standard identifier as well as each interaction type into its corresponding BEL relationship. For example, HGNC identifiers are prioritized for genes/proteins as the software is primarily concerned with human pathways while in the absence of HGNC identifiers or for those gene products coming from other species, other identifiers are used in their place. The Supplementary Information describes and presents statistics about the harmonization process in BEL for each of the three databases.

2.2 Exploring pathway cross-talks and contradictions

Along with the software handling the integration of pathway knowledge, we implemented a web application (the PathMe Viewer) for querying, browsing, and navigating content.

Using the search box found on the main page, users can submit a query to explore a specific pathway or a set of pathways. The result of the query leads to a visualization that renders the corresponding network, powered by multiple, built-in functionalities enabling users to navigate through it (Figure 1)**.** For instance, when multiple pathways are selected, a novel boundary visualization facilitates the exploration of pathway cross-talks (i.e. the interaction of pathways through their sharing of common entities). Furthermore, search and mining tools enable navigation of the resulting network as well as the identification of contradictory and consensus relationships across pathways. Finally, networks can be exported to multiple formats such as BEL, GraphML, or JSON to be used in other analytical software.

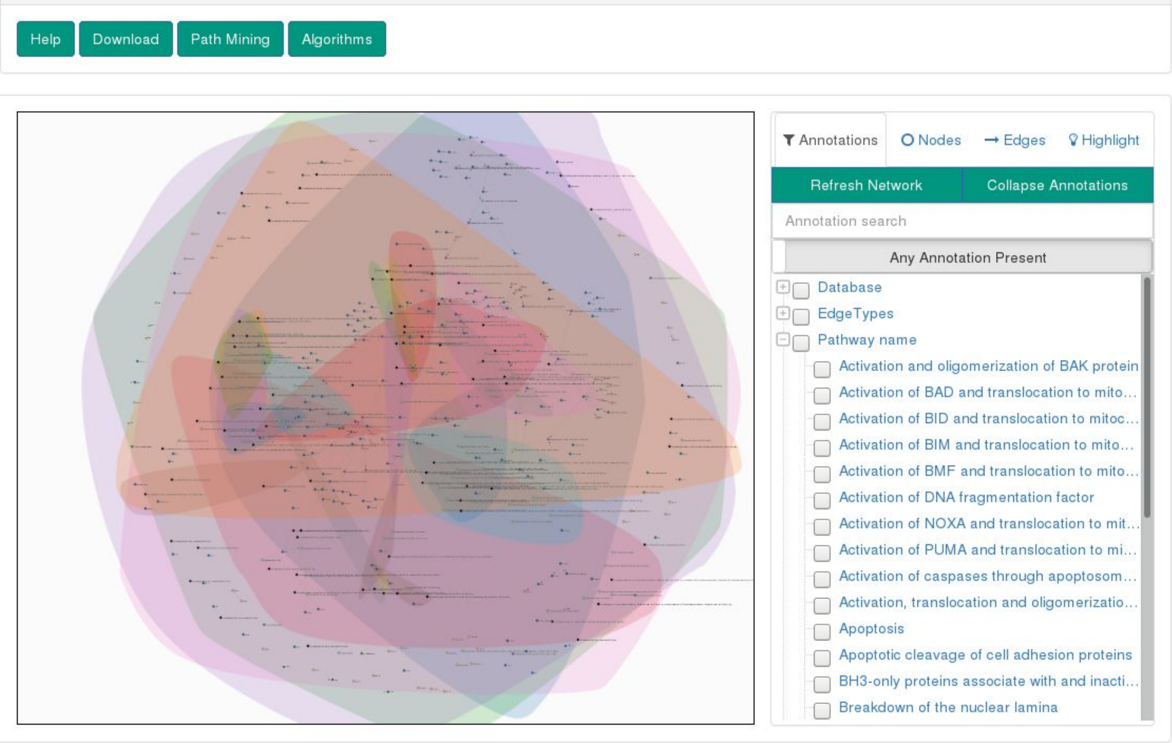

Help
Download
Path Mining
Algorithms
Annotations
Nodes
Edges
Highlight
Refresh Network
Collapse Annotations
Annotation search
Any Annotation Present
Database
EdgeTypes
Pathway name
Activation and oligomerization of BAK protein
Activation of BAD and translocation to mito…
Activation of BID and translocation to mitoc…
Activation of BIM and translocation to mito…
Activation of BMF and translocation to mito…
Activation of DNA fragmentation factor
Activation of NOXA and translocation to mit…
Activation of PUMA and translocation to mi…
Activation of caspases through apoptosom…
Activation, translocation and oligomerizatio…
Apoptosis
Apoptotic cleavage of cell adhesion proteins
BH3-only proteins associate with and inacti…
Breakdown of the nuclear lamina

Figure 1. The merged *Apoptosis* network from KEGG, Reactome, and WikiPathways visualized in the PathMe Viewer.

Application

Merging pathway knowledge enables conducting cross-talk analyses for any set of pathways through the PathMe Viewer. As a case scenario, we used PathMe in conjunction with the Viewer to explore the knowledge consolidated from 21 equivalent pathways across the three databases previously curated by Domingo-Fernández *et al.* (Supplementary Table 4). Whilst conducting a cross database pathway comparison previously required either extensive manual curation or harmonization of both entity identifiers and data formats on a case by case basis, this example illustrates how PathMe can be exploited to enable a systematic comparison of equivalent pathways.

To evaluate the degree of overlap between the three representations of each equivalent pathway, we used a variation of the Szymkiewicz–Simpson coefficient calculated for the common molecular nodes between the networks (Supplementary Equation 1).

Each of the 21 equivalent pathways showed partial overlap, except one which did not contain the pathway information required to convert the pathway into BEL in two of its original files. Among the equivalent pathways with the highest degree of similarity, we found well-studied pathways such as ‘Cell cycle’, ‘Toll-like receptor signaling', 'mTOR signaling', Hedgehog signaling', and 'Apoptosis'. Although the three databases represent the most widely studied molecular players in each of these pathways, merging their knowledge assists in filling the gaps between the complex interactions occurring in these pathways. Pathways with low similarity, such as 'TCA Cycle' and 'Sphingolipid Metabolism', indicate the resources captured distinct aspects of the biology within the pathway. Finally, the viewer offers a feature to detect contradictory edges between identical nodes across two databases. Though no such contradictions were found in the equivalent pathways, this feature may be useful to identify contradictory interactions between a set of pathways.

Discussion

Parallel developments of pathway databases during recent decades have resulted in different formalization schemas, hampering the interoperability between these resources and creating data silos. Overcoming this obstacle is instrumental to better understand the mechanisms underlying pathway knowledge. Additionally, while our approach can accommodate multi-scale pathway information from divergent database formats into a singular and standardized schema, a minority of entities and interactions have no discernible equivalencies in BEL and, as such, had to be omitted.

We have presented a framework through which content across multiple pathway databases can be integrated and transformed into a unified schema. Even though PathMe currently only incorporates content from three major pathway databases, its flexibility allows for future inclusion of additional pathway databases. Finally, it holds the capacity to update its content and track developments in pathway knowledge, an issue earlier outlined by Wadi *et al.*.

Funding

This work was supported by the EU/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

1. Cerami, E. G., *et al.* (2011). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res*. 39(Suppl. 1), D685–D690.
2. Stobbe, D., *et al.* (2011). Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC systems biology*, *5*(1), 165.
3. Demir, E., *et al.* (2010). The BioPAX community standard for pathway data sharing. *Nature biotechnology*, *28*(9), 935.
4. Domingo-Fernández, D., *et al.* (2018). ComPath: An ecosystem for exploring, analyzing, and curating mappings across pathway databases. *bioRxiv* 353235.

5. Fabregat, A., *et al.* (2017). The reactome pathway knowledgebase. *Nucleic Acids Res*., 46(D1), D649-D655.

6. Hucka, M., *et al*. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, *19*(4), 524-531.

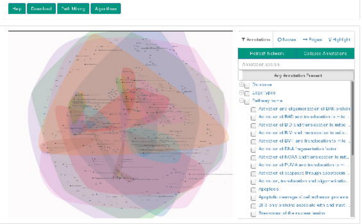
7. Kanehisa, M., *et al.* (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*., 45(D1), D353-D361.

8. Sales G., *et al.* (2018). *meta*Graphite - a new layer of pathway annotation to get metabolite networks, *Bioinformatics*, bty719.

9. Slenter, N., *et al.* (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res*., 46(D1), D661-D667.

10. Stobbe, D., *et al.* (2011). Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC systems biology* 5.1: 165.

11. Wadi, L., *et al.* (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nature methods*, *13*(9), 705.


Help
Download
Path Mining
Algorithms
Annotations
Nodes
Edges
Highlight
Refresh Network
Collapse Annotations
Annotation search
Any Annotation Present
Database
Edge Types
Pathway name
Activation and oligomerization of BAK protein
Activation of BAD and translocation to mito...
Activation of BID and translocation to mitoc...
Activation of BIM and translocation to mito...
Activation of BMF and translocation to mito...
Activation of DNA fragmentation factor
Activation of NOXA and translocation to mit...
Activation of PUMA and translocation to mi...
Activation of caspases through apoptosom...
Activation, translocation and oligomerizatio...
Apoptosis
Apoptotic cleavage of cell adhesion proteins
BH3-only proteins associate with and inact...
Breakdown of the nuclear lamina