Integration of Sabio-RK to the Reactome Graph Database for Efficient Gathering of Cell Signaling Pathways

Fabio Montoni[1,2,3,6], **Ronaldo N. de Sousa**[1,2,3], **Marcelo B.L. Junior**[1,3], **Cristiano G.S. Campos**[4], **Vivian M. Constantino**[1,2,3], **Willian Wang** [1], **Cássia S. Sanctos**[1], **Hugo A. Armelin**[1,2,5] **and Marcelo S. Reis**[1,4,6]

[1] Center of Toxins, Immune-Response and Cell Signaling (CeTICS);

[2]Cell Cycle Laboratory, Butantan Institute, Brazil

[3]Bioinformatics Graduate Program, University of São Paulo, Brazil

[4]Laboratory of Artificial Intelligence and Inference in Complex Data (Recod.ai); Institute of Computing, University of Campinas, Brazil

[5]Biochemistry Department, Institute of Chemistry, University of São Paulo, Brazil

[6]Corresponding authors. E-mail: msreis@ic.unicamp.br, montoni@ime.usp.br

Abstract. *Over the years, several tools have been developed with the aim of recreating signaling pathways, allowing the in silico representation of a biological system to be glimpsed from afar, which would improve disease studies. However, despite all the progress, much information needed to create a reliable model is diffused in public repositories with different objectives (e.g., Sabio-RK, which stores kinetic constants, and Reactome, a database for biochemical reactions) and the computational cost for simulating large sections of pathways in an exponential universe of possibilities can be challenging. As an alternative to deal with complex and heavy data, graph databases have been increasingly used to represent biological models. Here, we present a way to combine the stored quantitative information from Sabio-RK into the Reactome Graph Database, while keeping the graph-based structure of the latter. To assess the proposed integration, we implemented it using Python and subsequently verified its correctness through cypher queries. We expect that such integrated database would be a useful tool for cell signaling pathways studies, especially in the designing of computational models of those pathways.*

1. Introduction

Living organisms are endowed with a complexity orchestrated by their biological systems, that in turn are ruled by the dynamic responsive network of signaling pathways. Once those pathways are intertwined, their study is filled with several challenges. The use of computational power to improve these studies have been increasing over the years, and graph databases have been playing an important role on representing biological models [Henkel et al. 2015, Touré et al. 2016, Balaur et al. 2017, Swainston et al. 2017]. Nowadays, Reactome Graph Database [Fabregat et al. 2018] is one of the most prominent biological graph database, that contains millions of biological connections from a number of organisms. However, despite the numerous advantages that this database can provide, it lacks quantitative information on signaling pathway reactions. On the other

hand, Sabio-RK [Wittig et al. 2018] is a repository that provides access to several kinetic data from biochemical reactions, though it lacks the network connections as seen in Reactome. Therefore, the integration of Sabio-RK information into Reactome Graph database would represent an improvement in cell signaling pathway studies, since we would make available in a single place information that previously could not be easily accessed on a large scale. More specificaly, that integration would allow the combination of the advantages in having the access to complex pathways reactions through Reactome Graph Database with the biochemical reactions kinetic data stored on Sabio-RK, which can be used, for instance, for designing of computational dynamic models of signaling pathways.

2. Schema of the proposed integration

In Figure 1), we show the proposed schema for database integration, which takes the Reactome Graph Database and augments it to accommodate Sabio-RK kinetic data.

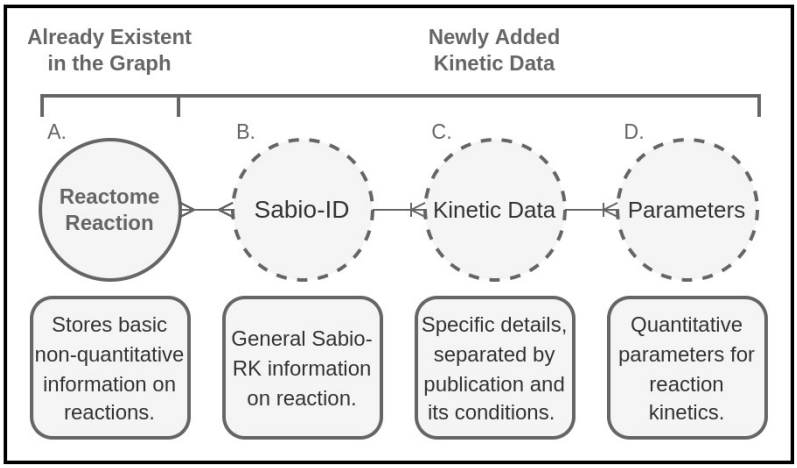

Figure 1. The proposed integration schema. Each circle represents a node from the schema. A: the already stored Reactome reactions. This node is where all Sabio-RK reactions will be related to, in a N:N relationship. B: the general information on Sabio-RK reaction, that will be shared amongst subsequent nodes. This node related in a 1:N relationship with Kinetic Data node. C: the nodes that will store specific information from the kinetic data (e.g., *wildtype*, *disease*). This node relates in a 1:N with Parameters node. D: storage of quantitative parameters (e.g., kinetic constants).

3. Implementation of the integrated database

For the implementation of the proposed database, we coded in Python a program that access through API services both Sabio-RK and Reactome, verifies organisms and reactions in common and, lastly, updates the new Sabio-RK information on Reactome Graph database, in which can be accessed using Neo4J [Balaur et al. 2017]. This implementation is free, open-source and can be downloaded in github.com/Dynamic-Systems-Biology-Group/BreSci-2022.

3.1. Data insertion into the integrated database

To insert the new information into the existing graph, the Python program filters the data and uses Neo4J package to be able to insert new nodes in the graph location. This way,

the user can access not only the structure, but also quantitative information that was not available through simple cypher queries on Neo4J. As of March 2022, in total Sabio-RK has 1192 organisms, while Reactome has 84. All Reactome organisms are present in Sabio-RK database. From these 84 organisms, 46 can be included into an integrated one. (Table 1).

Organism Name	Sabio-RK Reactions	Reactions available to Reactome Graph DB Association
Homo sapiens	13,543	2,107
Mus musculus	1,922	354
Others	24,755	4,923
Total	**40,220**	**7,384**

Table 1. Number of Sabio-RK reactions from the 46 common organisms that can be associated with Reactome Graph Database to generate a new graph.

In Table 1, we show the number of Sabio-RK reactions that can be associated with Reactome Graph Database from organisms in common. From all of these, about 15% of the total reactions from *Homo sapiens*, 17% of the total reactions from *Mus musculus* and 11% for the others. This means that even being reactions from common organisms, not all of them can be included in an integrated database. This is due the requirement of the *Reactome Reaction ID* for a correct association of Reactome reactions within Sabio-RK, which is done through MetaNetX [Moretti et al. 2021]. Despite the initial integration coverage can be considered low, is worth to note that with time, the number of Sabio-RK reactions that have an *Reactome Reaction ID* associated tends to increase over the time, and consequently, allowing more reactions to be added to Reactome Graph Database.

3.2. Proof-of-concept of the integrated database

Once the integrated database was populated, we proceeded to verify its consistency using cypher queries. This first query allows the user to verify the information stored into the integrated database. All nodes from all reactions are displayed.

Cypher Query 1

```
// SELECT ALL SABIO-RK REACTIONS
MATCH (n:SabioRkReaction)-[:KinecticDatafor]-(k),
(n)-[:GeneralReactionFor]-(r), (k)-[:ParameterInfo]->(p)
RETURN n, k, r, p
```

This second query allows the user to search for a specific *Sabio-RK ID* and returns all reactions and parameters related to it.

Cypher Query 2

```
// SELECT SABIO-RK REACTION BY ID NUMBER EXAMPLE
MATCH (n:SabioRkReaction{SabioReactionID: 594})-[:KinecticDatafor]-(k),
(n)-[:GeneralReactionFor]-(r)
RETURN n, k, r
```

In the aforementioned queries, n is the Sabio-RK reaction ID, k is Kinectic Data, r is Reactome Reaction, and p is Parameters node.

4. Conclusion

Here, we proposed a way to augment the Reactome Graph Database, integrating it with the Sabio-RK repository. The resulting integrated databased stores information from both biochemical reactions (from Reactome) and also parameters related to reaction kinetics (from Sabio-RK). To assess the proposed integration, we inserted data from those two databases into the integrated one and executed cypher queries. We expect that this work would be a useful tool for cell signaling pathway studies, since it stores in a single place all relevant information for the designing of pathway models.

Acknowledgements

This work was supported by scholarships from CAPES, CNPq, BECAS Santander and also by grants 13/07467-1, 19/21619-5, 19/24580-2, 20/10329-3, 20/08555-5 and 21/04355-4, São Paulo Research Foundation (FAPESP).

References

Balaur, I., Mazein, A., Saqi, M., Lysenko, A., Rawlings, C. J., and Auffray, C. (2017). Recon2neo4j: applying graph database technologies for managing comprehensive genome-scale networks. *Bioinformatics*, 33(7):1096–1098.

Fabregat, A., Korninger, F., Viteri, G., Sidiropoulos, K., Marin-Garcia, P., Ping, P., Wu, G., Stein, L., D'Eustachio, P., and Hermjakob, H. (2018). Reactome graph database: Efficient access to complex pathway data. *PLoS computational biology*, 14(1):e1005968.

Henkel, R., Wolkenhauer, O., and Waltemath, D. (2015). Combining computational models, semantic annotations and simulation experiments in a graph database. *Database*, 2015.

Moretti, S., Tran, V. D. T., Mehl, F., Ibberson, M., and Pagni, M. (2021). Metanetx/mnxref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. *Nucleic Acids Research*, 49(D1):D570–D574.

Swainston, N., Batista-Navarro, R., Carbonell, P., Dobson, P. D., Dunstan, M., Jervis, A. J., Vinaixa, M., Williams, A. R., Ananiadou, S., Faulon, J.-L., et al. (2017). biochem4j: Integrated and extensible biochemical knowledge through graph databases. *PloS one*, 12(7):e0179130.

Touré, V., Mazein, A., Waltemath, D., Balaur, I., Saqi, M., Henkel, R., Pellet, J., and Auffray, C. (2016). Ston: exploring biological pathways using the sbgn standard and graph databases. *BMC bioinformatics*, 17(1):1–9.

Wittig, U., Rey, M., Weidemann, A., Kania, R., and Müller, W. (2018). Sabio-rk: an updated resource for manually curated biochemical reaction kinetics. *Nucleic acids research*, 46(D1):D656–D660.