Looking into Reactome through Biopax Lens

Laleh Kazemzadeh *, Helena Deus*, Michel Dumontier† and Frank Barry‡
* Digital Enterprise Research Institute
National University of Ireland, Galway,
Email : laleh.kazemzadeh@deri.org
† Department of Biology
Ottawa Institute of Systems Biology
Ottawa, Canada
Email: michel_dumontier@carleton.ca
‡ National Centre for Biomedical Engineering Science
National University of Ireland, Galway
Email:frank.barry@nuigalway.ie

***Abstract*—In order to understand cell behavior under different conditions, the computational simulation of biological pathways is of great interest. Hence, to simulate a biological pathway computationally, extensive knowledge of protein-protein interactions (PPIs) in the pathway is required, along with the information about the generic flow of the pathway components i.e. biological reactions, which comprise the concerned pathway.**

The popularity of Semantic Web technologies in tackling the integrative bioinformatics challenges has increased, with various approaches used to aggregate and correlate data from different sources. However the integration of publicly available pathway databases, to determine the different PPIs and hence effectively simulate the cell behavior, has still various obstacles. In this paper, we present a semantic approach in pathway-wise analysis of protein-protein interactions (PPIs) using Biopax standards focusing particularly on Reactome database. We have identified the PPIs involved in a given pathway by the hierarchical extraction of its components (complexes, proteins, small molecules). We have developed a visualization tool which automatically generates a visual representation of the directed graph of PPIs in any specified pathway. Our approach provides intuitive inference of the data by flattening the nested pathways in Reactome and their components instead of wrapping each layer of data in the shell of outer pathway. We have also discussed that the representation of a pathway in Biopax standard format is highly complex and even contains redundant information. Hence tools are needed in order to facilitate the navigation and analysis of pathway datasets, which have been structured in Biopax format.

I. Introduction

The functionality of the human body is tightly regulated by biological pathways. Basic building blocks of these pathways are proteins, which act in an orchestra in order to keep the regulation of pathways intact. Therefore understating the dynamic of these pathways is directly dependent on understanding how the proteins involved in a pathway interact with each other. Interaction between two proteins might be of different types e.g. activation, inhibition, and methylation. Analyzing biological data from a pathway perspective can result in valuable information about the process of disease and suggest new drug discovery methods that target mis-regulation in specific pathways, thus enabling a much more precise targeting of diseases. However, computationally representing a pathway is not a trivial exercise due to the various types of components and interactions; regulation of pathways requires a cascade of events and interactions between genes, proteins and small molecules.

In addition, there is significant cross-talk between pathways, which highlight the fact that pathways are not isolated but are made up of a network of components. As such treating them as a system as opposed to an enclosed and self-contained pathway, can support a more realistic investigation.

II. State of the Art

A large number of tools and applications, vocabularies and ontologies aimed at computationally modeling biological pathways currently exist with enough precision to enable realistic simulations of its processes and determination of mechanism of action of various molecular compounds; examples include the systems biology markup language (SBML) [1] and the Proteomics Standards Initiative-Molecular Interaction (PSI-MI)[1]. These models and data format are also devised to deepen and broaden our understanding of pathways. A few models also keep track of semantics, i.e. they attempt to precisely and unambiguously describe each compound and each interaction such that they can be interpreted by applications and thus be integrated with other models. Biological Pathway Exchange (Biopax) [2] is one such data format. Biopax is a standard format for representing pathways and molecular interactions within and between pathways which has been developed with the aim of facilitating the process of collecting, indexing and sharing data [2]. Several databases hosting pathway and protein interaction information, such as Reactome[2] and Pathway Commons [3], are already available in this format. Information retrieved from expert-curated databases like Reactome is highly valuable for scientific advancement since they are the most accurate training data sets. However, because they rely on human curation, they suffer from limited coverage in the amount of interactions available. Integrating such data

[1]http://www.psidev.info
[2]http://www.reactome.org

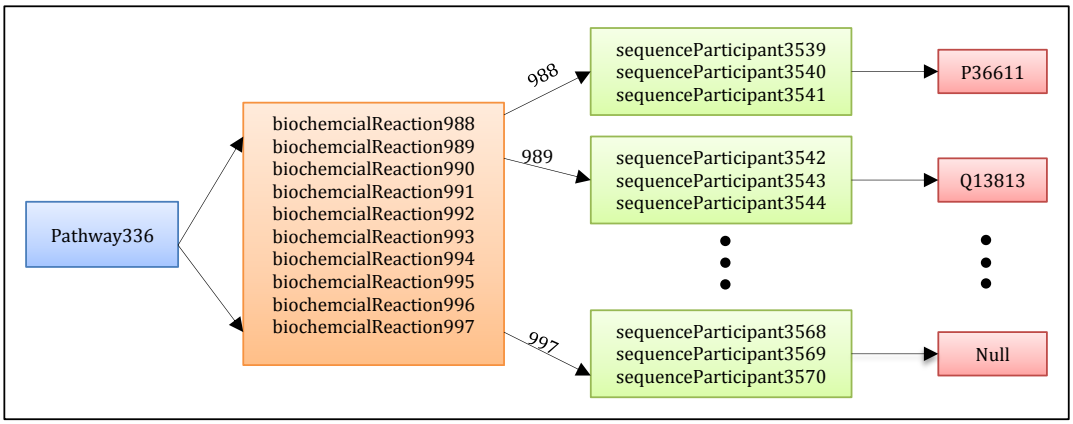

Fig. 1: Example of redundancy and incompleteness of data represented in Biopax level2 taken from caspase-mediated cleavage of cytoskeletal pathway. Blue box indicates the sample pathway, orange boxes represent list of biochemical reactions associated to this pathway, green boxes show sequence participant at left and right of each biochemical reaction, red boxes depict the unique Uniprot ID for each protein which each left and right of a biochemical reaction points to.

warehouses in one standard format will improve the coverage and highlight the role of Biopax in standardization. There is an enormous potential in using the information represented in Biopax format to realistically address biological questions, for example, the metabolic effects of a compound in the cell or how certain alterations in the metabolic network can be at the root cause of diseases or drug resistance. The discovery and confirmation of a biologically meaningful molecular interaction often requires the analysis of enormous amount of heterogeneous data which are typically deposited in local databases and isolated from each other. Therefore, considerable amount of molecular interactions are "hidden" in this data, which can only be exposed once these results are integrated and recurrence of patterns indicative of interactions analyzed. The data integration challenges in life science have motivated the researchers to adapt the new integration technologies offered by Semantic Web and Linked Data. Semantic Web technologies can provide a bridge between the datasets, enabling the discovery of links, which are often not obvious. These bridges are often standard vocabularies and ontologies developed toward improvements in knowledge discovery that lead to the next challenge: the representation, application and acceptance of these standard vocabularies by the domain experts. The motivational scenario for the work presented here is the extraction of all the molecular components that act in a particular biological process as described by Biopax in its various data sources. We have chosen Biopax firstly because it has been adapted by several databases, which provide information in signalling pathways and secondly becasue it faciliates data integration from other sources containing protein information.

Biopax has been developed to capture various aspects of signalling, regulatory and metabolic pathways. However in order to provide a descriptive solution and to cover all details in the description of pathways, some complexity needed to be introduced. In Biopax each pathway is constructed in the form of nested pathways which partially, but not fully, illustrate the overlaps between several pathways. Furthermore, each biochemical reaction is described as a function of the "left" and "right" hand side of the stoichiometric equation. Fig. 1. illustrates an example of data complexity and redundancy in representing biochemical reactions involved in pathway336 (caspase-mediated cleavage of cytoskeletal). As it is mentioned before each biochemical reaction has left and right components each of which refers to unique and separate sequence participant. However, each of these sequence participants points to the same protein ID from UniProt database. In other word, both left and right of a given biochemical reaction point to the same protein and this increases the redundancy of the data. The aim of our work is to devise a tool that aggregates information from this data e.g. the protein interactions and components of protein complexes in pathways. This will allow us to easily identify common interaction between various components (proteins, complexes, etc.) across pathways, abstracting from the complexity of pathway representation in Biopax. The data analysis tools made available by Reactome are unable to provide this inner-pathways analysis unless pathways are nested or siblings.

III. Methods

One typical way of querying a pathway or interaction between two proteins from different online databases is through browsing their webpage. As easy as it seems, it is time consuming and cumbersome to go through all the databases available manually. Instead we can query the PPIs directly from the raw data provided by the databases like Reactome and other such pathway databases. We propose an approach to overcome such problems which is explained below.

Fig. 2 shows an overall view of the steps, which were taken in our approach in order to identify the protein-protein interactions pathway-wise. We downloaded the protein-protein interaction file for Homo sapiens from Reactome webpage in Biopax format. This data was uploaded to our Sesame

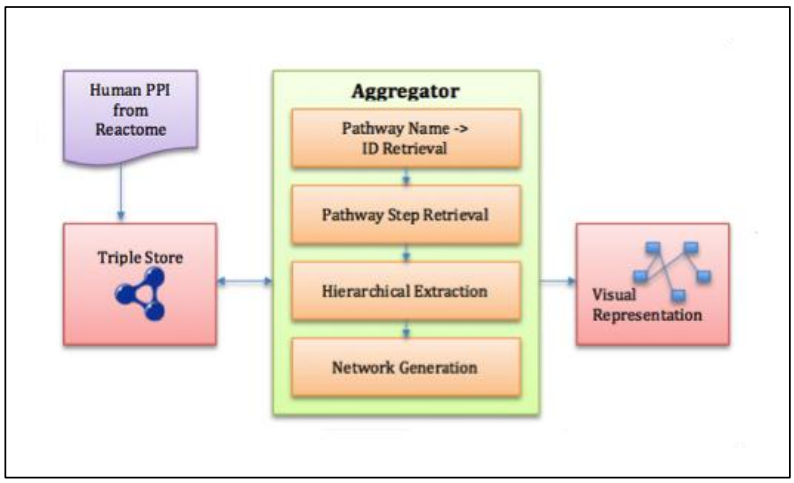

Fig. 2: Overall view of the proposed method.

server[3] in the form of triples. The Aggregator module has been developed in order to extract the components involved in a pathway and break down the pathway to the level of complexes, proteins and molecules.

The system provides a list of selectable pathways compatible with the pathways names used in Reactome. The ID of the selected pathway e.g. Apoptosis or Programed Cell Death (PCD) is retrieved from the triple store by the ID Retrieval module. The Pathway Step Retrieval retrieves the list of inner pathways (pathway-steps) forming the selected pathway. Each of these pathways is segregated hierarchically in the Extraction module.

The extracted data from Pathway Step contains bundle of relational information explaining reactions, complex blocks, proteins and small molecules forming complexes. Network Generator constructs a model in the final stage from the data extracted in the previous step. This model is then fed to the network visualizer, which renders and displays the relational graph between components of the pathway. In this model, the relation between each entity, complex, protein and molecule in the pathway is illustrated in a directed graph where nodes represent the entities, pathways, proteins and molecules and edges represent the connections between source and target nodes or the higher level and lower level components in a pathway tree.

The interaction Aggregator is written in PHP using ARC2[4] package in order to query the Reactome triples. The force-directed graph is generated by the Data Driven Documents (d3)[5], library written in Javascripts.

IV. Results

Raw material in our approach is an input .owl file, which contains the information of any pathway in Biopax. Applying our method we were able to generate a pathway wise PPIs network which is shown and discussed below.

Fig. 3 shows a small part of the network visualization generated by our tool for the Apoptosis pathway. The generated network contains 60 interactions between 40 pathways, representing nested pathways in Reactome, and 87 proteins involved in inner pathways of Apoptosis. Here we show the interaction between pathway336 and pathway335, which are caspase-mediated cleavage of cytoskeletal proteins and apoptotic cleavage of cellular proteins pathways respectively. These two pathways are part of outer pathways of Apoptotic execution phase and Apoptosis, which are not shown here.

The number of identified proteins in pathway336 is 8, while the number of reported proteins for the same pathway in Reactome database is 32. The reason for these differences is that some of the reported proteins in Reactome point to the same unique protein identifier. As an example protein P08670, Vimentin, has been mentioned 7 times. Likewise Q151149 and the rest of identified proteins have been reported 3 times. Our algorithm was not able to identify 3 proteins (caspase 3,6,7) in the list of 32 proteins reported in Reactome database due to incompleteness of the original data which was downloaded from the Reactome webpage.

Of great interest in pathway anlysis is identification of protein hubs. Protein hubs are those proteins with high degree of connectivity and more likely to be essential in the cell. Example of such a protein is shown in Fig. 4. Protein Q14790 (caspase 8), appears to be involved in the following pathways: Fasl/DD95L signaling (pathway309), TNF signaling (pathway310), Trail signaling (pathway311), Formation of caspase 8 (pathway312), Activation of pro-caspase 8 (pathway313) and Apoptotic execution (pathway 334). Knowing the protein ID or name and assuming the protein of interest is involved in different pathways we are able to retrieve the same information from Reactome search tool, however it does not give us the intuitiveness of the visualization. Querying the same protein, casapse 8, in Reactome returns more hits than the number of

[3]http://hcls.deri.org:8080/openrdf-workbench/repositories/
[4]https://github.com/semsol/arc2/wiki
[5]http://d3js.org/

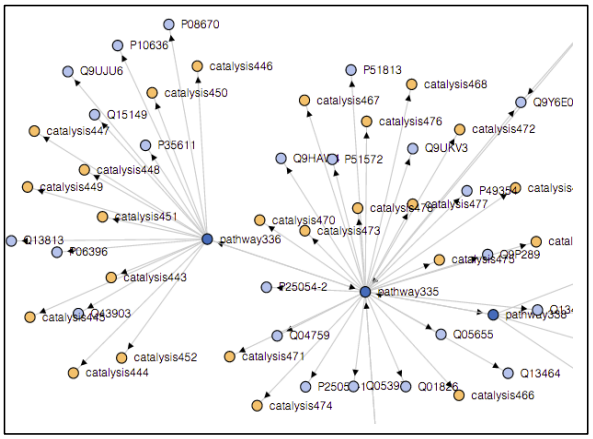

Fig. 3: Directed graph generated by the network visualizer. Graph shows the interaction between and within two pathways. Pathways and proteins are shown with their unique IDs. Each edge represents the connection between pair of source and target nodes. Dark Blue: pathways, light blue: proteins, orange: catalysis.

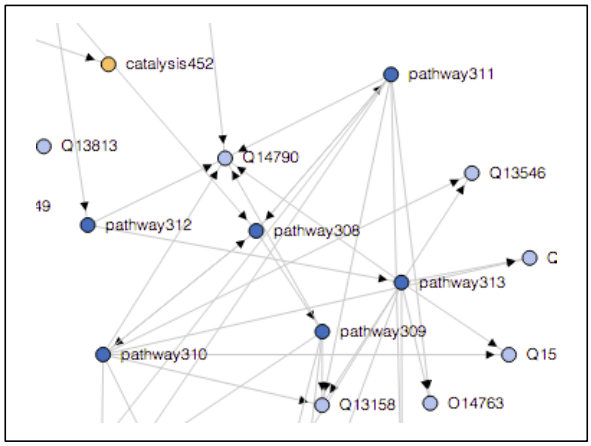

Fig. 4: Protein hub connecting six inner pathways in the Apoptosis pathway.

pathways we discussed here since we limited the search only to the Apoptosis pathway and not all the pathways exist in Reactome.

V. Conclusion

In this work we were able to extract PPI associated with any given pathway. Our visualization provides a better representation of elements involved in a pathway since it is capable of retrieving and representing data while conserving the hierarchy in which data was originally represented. Our aim was to highlight the PPIs in the pathways hence we represented only pathways and proteins in the deepest level of each pathway step of an outer pathway. However the data retrieved from the triple store by Aggregator contains more information about each pathway than only its components (e.g. pathway name) and with the current structure of our tool it is possible to add an extra layer of data to the Network Generator and create a visual representation of the extended network including e.g. protein complexes or type of interactions which, if added, the system will be more infromative. Our tool is compatible with Biopax level 2 thus it may not generate the same expected result when it is provided with a data file in Biopax level 3. Moreover, during the course of this work we have observed and analyzed Biopax format in detail. Some of the classes and properties introduced in Biopax appear unnecessary but also raise the level of complexity in the pathway representation and pathway analysis. Some of these complexity issues have been addressed and improved in later release of Biopax but pathways represented in Biopax level 2 suffers from this unnecessary complexity. In this work we tried to diminish the amount of redundant data by omitting the biochemical reaction, left and right step in each pathway step and showing only the proteins involved in a single pathway at the most inner level.

VI. Future Work

Future work will be the integration of pathways and interactions from other databases like BioGrid [4], MINT [5], HPRD [6] and the expansion of the query and visualization in such a way that two or more pathways from different sources can be queried and the common interactions highlighted. Furthermore, identified interactions will be ranked based on the number of occurrence in the databases and the literature.

Acknowledgment

This work has been funded by Program for Research in Third Level Institutions (PRTLI) Cycle 5, which is co-funded by the European Regional Development Fund (ERDF).

References

[1] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, and J. Hofmeyr, *he Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models, Bioinformatics, vol. 19, pp. 524–531, 2003*

[2] E. Demir, M. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, J. Luciano, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, and K. Kumaran, *The BioPAX community standard for pathway data sharing, Nature Biotechnology, vol. 28, pp. 935–942, 2010*

[3] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander, *Pathway Commons, a web resource for biological pathway data, Nucl. Acids Res., 2010*

[4] C. Stark, B.J Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and and M. Tyers, *BioGRID: a general repository for interaction datasets, Nucleic Acid Re., no. 1, pp. 535–9, 2006*

[5] A. Ceol, A. A. Chatr, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni, *MINT, the molecular interaction database: 2009 update, Nucleic Acids Res., vol. 38,Database, 2010*

[6] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, and C. J. H. Kishore, *Human Protein Reference Database - 2009 Update, Nucleic Acids Research., no. 37, 2009*