

Data Integration with BioPAX Pathway Datasets for Automated Navigation

Keyuan Jiang

Department of Computer Information Technology
Purdue University Calumet
Hammond, IN, USA

Abstract - *The emergence of biological pathway datasets in BioPAX format provides a standard way to exchange pathway datasets and new ways to explore and navigate biological data. The lack of a searchable, centralized repository of the BioPAX datasets poses a challenge in integrating internal and external data elements. We have designed a system which uses an XML database along with a layer of adaptors to link internal and external data elements, facilitating the automated navigation of biological data for biomedical research.*

Keywords: Data Integration, Biological Pathway Datasets, Bioinformatics Web Services.

1 Introduction

Biological pathways represent our current understanding of biological processes at the cellular level. Biomedical scientists employ the pathway information in formulating hypotheses, verifying experimental results and sharing research outcomes [1], and in doing so they explore the pathway data along with various other datasets. With the introduction of many new laboratory techniques, more and more data are available for the curation and inference of new biological pathways. The number of publicly accessible biological pathway databases has grown to more than 200 [2], each with its own access method, data schema and storage mechanism. The BioPAX effort was initiated to provide a common format for exchanging biological pathway datasets [3]. Since the release of the BioPAX Ontology Level 1 [4] in 2005, several datasets have been made available by the data providers such as BioCyc [5], Reactome [6], and KEGG [7]. These datasets are in the format of OWL [8], a machine processable format, and primarily pertain to the components (or entities in the BioPAX terms) and relationships of the components in pathways. While it is a major leap forward in exchanging biological pathway data using a common format, it still remains challenging to make full use of the BioPAX datasets to address intriguing biological problem in terms of available software tools and the capability to integrate datasets with external databases.

Given the vast amount of available biomedical data on the Web, the biomedical researchers can query various

datasets via manual navigation using a Web browser. The OWL format of BioPAX datasets provides a way of organizing data pieces relevant to biological pathways based upon the ontology, facilitating the automated navigation and integration by computer programs such as intelligent software agents. Two types of data linkages exist in a BioPAX dataset: internal links and external links. The internal links define the relationships of the elements within a BioPAX data file and they are in the machine processable format where either they are embedded in a pathway instance or they reference to other objects defined in the same dataset through the use of URIs. But the linkages to external are not defined in the same way, leaving the integration of external data to bioinformaticians who develop software that processes the BioPAX datasets.

The BioPAX standard adopts the concept of external references (Xrefs) to associate elements involved in the pathways to the elements existing in the external data sources. Such external data sources are typically the authoritative sources of those entities which have unique identifiers in the source. Given the various access methods and diverse formats of the external data sources, the BioPAX does intentionally specify that the external references (including unificationXrefs and publicationXrefs) are not RDF IDs which are typically URIs that can be linked to the resources on the Web. Although it lacks the support of URIs for external Web resources, the ID and DB elements in an Xref provide a hint to the external data source and object in that data source.

2 Integration Methods

Different types of linkage require different types of methods of integration. Efficient internal linkages require a robust data storage mechanism an efficient queries, whereas external linkages need a mechanism to access remote data programmatically.

2.1 Integration with Internal Links

In BioPAX datasets, pathways and their components are interlinked, and components are defined in relation to their roles in the pathway. Navigating to various components can be achieved by resolving the URIs

assigned to each component. The navigation through the pathway space can cross the boundary of a pathway. For example, a researcher may start navigation with the “glycolysis I” pathway in BioCyc dataset for *E. coli*. He then discovers that one product of the pathway (pyruvate) is involved in a biochemical reaction (which is part of the “mixed acid fermentation” pathway) leading to the production of acetyl CoA which is in turn involved in the “glyoxylate cycle” pathway. Such multiple pathway navigation requires a robust storage mechanism which supports sophisticated queries efficiently.

Due to lack of a central repository and an efficient way of accessing the datasets, currently the BioPAX datasets are offered in two formats: a single file of all pathways in an organism (BioCyc), or one pathway in a single data file (Reactome and KEGG). Neither of these formats of dataset is suitable for automated navigation. For the multiple-pathways-per-file format, the data files can be of significant size. For example, the *E. coli* dataset provided by BioCyc is about 20 MB in size, and can not be processed efficiently by many existing XML parsers using its native format – we observed in our experiments that many parsers could not even load a single dataset, let alone processing the data. For the single-pathway-per-file format, although the small datasets can be processed efficiently by existing XML parsers, it creates a barrier to querying across the multiple pathways, because the associations of pathways can not be determined by examining the file names.

To solve this processability issue while maintaining the internal linkages contained in the same dataset, we flattened the structure of the BioPAX data and decomposed the pathways such that instances of BioPAX classes (objects) are used as the storage unit. All the objects are stored in an XML database called the Pathway Gateway [9] as shown in Figure 1. The decomposition of the pathways is based upon the fact that only a small amount of relevant data is focused on when navigating, and the overloaded trivial information will be filtered out. Such a treatment makes searching robust through the use of the XML database and processing more efficient because only a small amount of data needs to be loaded into the memory for processing. The Gateway provides two types of object data, one being the single level BioPAX objects and the other a complete set of all the objects at all levels involved in particular pathway. For all the internal links within a pathway, the later type of objects contains sufficient data, and when inter-pathway links need to be explored, the former type of objects offers more efficiency.

2.2 Integration with External Data Sources

External references (Xrefs) in BioPAX contain the identifiers of the external data elements and the data sources where the data elements are kept. Based upon this, we designed a mapping between the DB names and the methods of access to the corresponding external resources, and a layer of adaptors to unify the access methods to the various data sources residing on the Web. The external data sources have different access methods. Some are Web services which make possible the automated navigation and integration, while others are URLs to the query methods, requiring parsing the query results.

A single BioPAX data file may contain a number of different external data sources and a significant number of data elements. For instance, the BioPAX pathway data file of *E. coli* provided by BioCyc contains 15 different external data sources and 12,780 data elements. The linkage to various data sources have to employ different access mechanisms provided. Several data hosts provide a programmable interface via Web services such as NCBI eUtilities [10], EBI Web Services [11] and KEGG API [12]. Others such as Gene Ontology offer URLs for retrieving data elements from their data repository. A mapping is created to associate the external DB information in Xrefs with the access methods of corresponding external data sources.

The system architecture of the data integration is shown in Figure 1. To navigate to an internal data element from a BioPAX dataset, the agent queries the Pathway Gateway using the URI of the element via the database API or a Web service interface. When linking to an external data element, the agent parses the source database name and the identifier of the data element embedded in the external reference. The source database name is then used to look up the mapping to determine which adaptor to call. The call to the selected adaptor along with the data element identifier will retrieve the data from the external data source. The resultant data which are parsed and formatted by the adaptor are forwarded to the data aggregator. Various pieces of data are presented to the agent which may further process them or present the aggregated result to the end user.

3 Discussions

While the BioPAX datasets contain much of the information needed for the construction of pathways, there exists a need for integration of internal and external data elements with the BioPAX datasets for several potential applications. Several types of applications can be

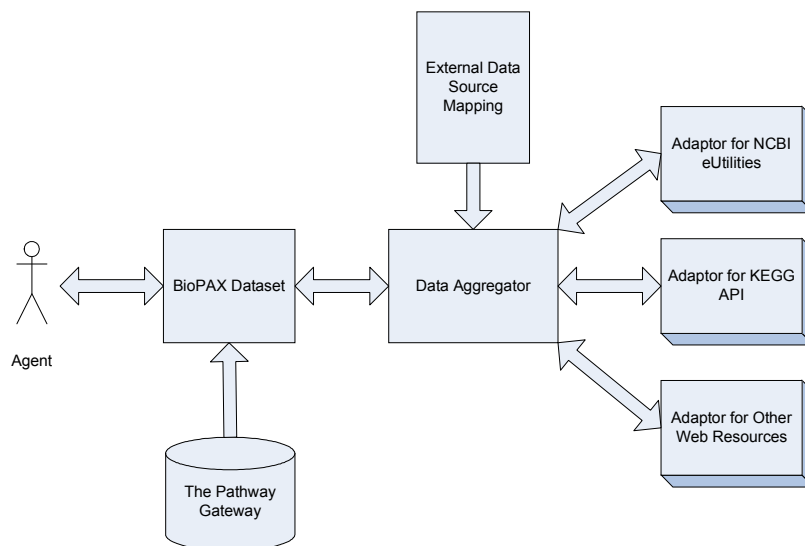


Figure 1. Integration with the BioPAX Datasets.

developed based upon the integration mechanism presented in this paper. One type of application is to link pathways derived from the different sources. Different data hosts provide different coverage of biological pathways. The integration across the datasets from different sources can be achieved by combining using external references and querying the datasets stored in the Pathway Gateway.

Another type of application is to further aggregate information from various sources and present it to the scientist. For instance, the abstracts provided by the PubMed are typically not contained in the BioPAX datasets, but can be readily retrieved through the external integration presented in this paper. The NCBI Entrez Utility Web Service supports querying against the PubMed database for abstract search. The beauty of this approach is that the linkage is created dynamically at request unlike many Web sites where all the (URL) links are hard-coded, making software application more adaptable.

4 Acknowledgement

This project was support in part by a Microsoft Research eScience Program Grant (Award #: 12703). The author wishes to thank Chris Nash for his assistance in developing the Pathway Gateway components for this project.

5 References

[1] P. Saraiya, K. Duca and C. North: "Visualization of Biological Pathways: An Ethnographic Study and Systems Evaluation", Information Visualization, Vol. 4, No. 3, 2005.

[2] PathGuide: the Pathway Resource List, [<http://cbio.mskcc.org/prl/>]

[3] J.S. Luciano: PAX of mind for pathway researchers. Drug Discov Today, 10(13):937-42, 2005.

[4] BioPAX Ontology Level 1 document: [<http://www.biopax.org/release/biopax-level1-documentation.pdf>]

[5] The BioCyc datasets: [<http://biocyc.org/open-reg.shtml>]

[6] The Reactome datasets: [<http://banon.cshl.org/download/>]

[7] The KEGG datasets: [<ftp://ftp.genome.jp/pub/db/community/biopax/>]

[8] W3C: Web Ontology Language (OWL): [<http://www.w3.org/2004/OWL/>]

[9] K. Jiang and C. Nash: "Application of XML Technologies to Biological Pathway Datasets." Submitted.

[10] NCBI: Entrez Utilities Web Service [http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html]

[11] EBI: Web Services at EBI: [<http://www.ebi.ac.uk/Tools/webservices/index.html>]

[12] KEGG: KEGG API - SOAP/WSDL interface for the KEGG system: [<http://www.genome.jp/kegg/soap/>]