An Object Model Based Repository for Biological Pathways Using XML Database Technology

Keyuan Jiang

Department of Computer Information Technology, Purdue University Calumet, U.S.A.
jiang@calumet.purdue.edu

Abstract. With the availability of a growing collection of biological pathway datasets, there is an increasing need for an efficient repository with the powerful search capability to facilitate large scaled analyses of biological pathway. A novel data store was designed based upon the BioPAX object models, and implemented with the emerging XML database technology. Our approach significantly reduces the system complexity while maintaining powerful search functionalities.

Keywords: Biological pathway datasets, BioPAX, XML database, XML datatype.

1 Introduction

Biological pathways represent our current understanding of living organisms at the cellular level. Increasingly, pathway datasets have become an important source for biomedical research, ranging from the elucidation of biological functions using the systems biology approach [1] to the discovery of new pharmaceutical targets [2]. The availability of genomic sequence datasets and advancement of high throughput laboratory techniques have helped accumulation of a large amount of pathway datasets. Today, there exist more than 200 publicly accessible pathway databases [3, 4], each with its own data model and access method, leading to data heterogeneity, incompleteness and inconsistency [5]. The availability of a large amount of pathway datasets is providing an opportunity for large scale analyses [6] which call for a consistent and homogeneous representation and an efficient storage of a variety of pathways and networks for metabolic reactions, signaling transduction, regulatory pathways and protein-protein interactions.

Representing the processes of cellular machinery, biological pathways are human constructs describing molecules involved in particular processes and their relationships. Owing to the nature of pathways, they are typically modeled as directed graphs [5, 7], and such a treatment facilitates the visualization of pathways by human eyes. Internally, pathway software systems use graph data structures to model pathway datasets, and most of them store the datasets in a relational database management system (DBMS) by taking advantage of the mature and efficient query mechanisms and storage built into those database engines. However, mapping graph

data structures onto relational models leads to more complex designs of databases and inefficient queries, and such an approach is not suitable for many applications that analyze pathway datasets in further elucidating biological functions [5]. After reviewing and analyzing the existing data models for storing pathway datasets, Deville et al [5] suggested that the most appropriate data model is the object-oriented model for storing biological pathways.

To overcome the data heterogeneity and adopt the object-oriented data model for biological pathways, we present in this paper our recent work on designing and implementing a storage system for pathway objects based upon the emerging XML database technology. The object-oriented model of pathways is based upon the community-developed BioPAX (Biological Pathways Exchange) standard [8], and each individual object defined in BioPAX is stored in the XML datatype column to provide XML fidelity and leverage the XQuery technology performing sophisticated queries on various BioPAX objects.

2 Design Considerations

Several efforts have been made in order to unify the data format of pathways, notably BioPAX [8], PSI-MI (Proteomics Standard Initiative – Molecular Interactions) [9], SBML (Systems Biology Markup Language) [10], and CellML (Cell Markup Language) [11]. SBML and CellML are primarily for describing quantitative (mathematical) models of cellular networks, and PSI-MI focuses on the molecular interactions. On the other hand, the BioPAX standard is an object-oriented representation of biological pathways with high expressiveness capable of representing various types of interactions and pathways [12]. The BioPAX standard, documented as ontology, defines components of a pathway as classes (entities). Each class is made up of a collection of attributes which may be instances of the BioPAX utility classes. A class may represent a physical object such as an RNA molecule or a process in a pathway such as a biochemical reaction.

The BioPAX standard also defines a special XML format based upon OWL (Web Ontology Language) [13] for representing each entity or utility class. This allows each instance (or object) of a BioPAX class to be represented as an XML document.

For large scaled analyses, it is common to retrieve objects instantiated from the BioPAX classes such as a biochemical reaction, a small molecular, a pathway, an external reference to the identifier in the authoritative data source, each having a biological meaning to some extent. The appropriate granularity of the pathway data model lies at the level of BioPAX objects for a variety of pathway analyses. In our design, BioPAX objects were chosen as the storage unit for data model.

Since there are dozens of BioPAX classes defined in the BioPAX standard, modeling them with the conventional relational model seems awkward. A user or program may need to retrieve the object by merely specifying its unique ID which is its URI made up of the concatenation of the value of `xmlns:base` and that of `rdf:ID` according the W3C specification. This type of retrieval is difficult, if not impossible, to implement with the relational model, because the query needs to determine at the runtime which table to query against.

In our design, instead of using the character datatype in the relational model, we chose the emerging XML datatype, which provides XML fidelity by preserving the InfoSet and XQuery data model, and supports query and update of all aspects of stored XML documents. In addition, a single XML datatype column allows storing multiple categories of XML documents, enabling us to store various BioPAX objects in a single column which facilitates querying objects by their URIs. Leveraging XQuery, a search on an object with one of its synonyms can be readily achieved without creating a linking table to associate objects and their synonyms, simplifying the query statements.

Like any other XML data, BioPAX compliant datasets are hierarchical in nature. When defined, objects can be embedded in the composing object, and such embedding can be multiple levels in depth. In order to store various types of objects in our data store, the XML-based hierarchy was flattened by rearranging the BioPAX compliant data such that embedded objects were replaced with references and moved out of the composing object. In other words, all the objects are organized in a single layer. When the composing object along with all its components is retrieved, all its components are retrieved by specifying their IDs and placed in the same result set.

3 Implementation

Our pathway repository was implemented using Microsoft SQL Server 2005 which supports the XML datatype, XML indexing, and XQuery. In our database, there is a single XML column to store all types BioPAX objects and a primary key column for object unique IDs: their URIs. All the queries were created as stored procedures in order to achieve the best performance.

Three different strategies were employed to create queries. 1) For simple queries such as one that only involves a URI, standard SQL statements are executed. 2) If a query deals with the internal part of a BioPAX object – for example, a synonym, an XQuery statement is issued. 3) The most sophisticated queries involve retrieving all the components that make up a particular object (e.g., all the objects making up a particular pathway). This will require multiple queries to accomplish. In this case, we use C# to create CLR (Common Language Runtime) stored procedures which demonstrate high flexibility and performance in manipulating the data stored in a SQL Server 2005.

To provide public and platform-neutral access to our pathway repository by other software applications, we developed a SOAP-based Web service called PathwayGateway service which furnishes a number of predefined access methods. In addition, the same set of methods is available as HTTP POST and GET methods. More detailed documentation on these methods as well as the WSDL [16] file can be found at <http://jlab.calumet.purdue.edu/theGateway>. We have imported into our repository a number of pathway datasets retrieved from KEGG [15] and BioCyc Websites. Currently our system contains the datasets of the following species: homo sapiens, Escherischia coli (K-12), and Escherischia coli (O157:H7).

4 Discussions

Our approach of employing XML database technology to store BioPAX objects provides a much simplified data model in storing biological pathways, significantly reduces the complexity of system design and the implementation, without any compromise on the query power. Leveraging the XML database technology, three different query strategies are explored to provide the best performance and flexibility.

The availability of a large quantity of pathway datasets kept in an efficient data store will promote future researches in large scaled analyses of biological pathways and development of a new breed of software applications. Examples of future directions may include (a) the comparative study of pathways through similarity searches, and (b) the dynamic visualization and exploration of pathways by parsing the XML-based pathway objects at the runtime.

Acknowledgments. The author wishes to thank Microsoft Research for its support to this project, Christopher Nash and Qian Wu for their efforts in implementing the system, and reviewers for their constructive comments.

References

1. Cary, M.P., Bader, G.D. and Sander, C.: Pathway information for systems biology. *FEBS Lett.* **579** (2005) 1815-20
2. Fishman, M.C. and Porter, J.A.: Pharmaceuticals: a new grammar for drug discovery. *Nature.* **437** (2005) 491-3
3. Pathguide: the Pathway Resource List [<http://pathguide.org>]
4. Bader, G.D., Cary, M.P. and Sander, C.: Pathguide: a pathway resource list. *Nucleic Acids Res.* **34** (2006) D504-6
5. Deville, Y., et al.: An overview of data models for the analysis of biochemical pathways. *Brief Bioinform.* **4** (2003) 246-59
6. Sharan, R. and Ideker, T.: Modeling cellular machinery through biological network comparison. *Nat Biotechnol.* **24** (2006) 427-33
7. Schaefer, C.F.: Pathway databases. *Ann N Y Acad Sci.* **1020** (2004) 77-91
8. Luciano, J.S.: PAX of mind for pathway researchers. *Drug Disc Today.* **10** (2005) 937-42
9. Hermjakob, H., et al.: The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat Biotechnol.* **22** (2004) 177-83
10. Hucka, M., et al.: The systems biology markup language (SBML). *Bioinformatics.* **19** (2003) 524-31
11. Lloyd, C.M., Halstead, M.D. and Nielsen, P.F.: CellML: its future, present and past. *Prog Biophys Mol Biol*, **85** (2004) 433-50
12. Stromback, L., et al.: Representing, storing and accessing molecular interaction data: a review of models and tools. *Brief Bioinform.* **7** (2006) 331-8
13. W3C: Web Ontology Language (OWL). [<http://www.w3.org/2004/OWL/>].
14. BioCyc: BioCyc Database Collection. [<http://biocyc.org/>].
15. KEGG: Kyoto Encyclopedia of Genes and Genomes. [<http://www.genome.jp/kegg/>].
16. W3C: Web Service Description Language. [<http://www.w3.org/TR/wsdl>].