Pathway Knowledge Base: Integrating BioPAX compliant pathway knowledgebases

Nikesh Kotecha, Kyle Bruck, William Lu and Nigam Shah
Stanford University, Stanford CA

The interaction between genes, proteins, and small molecules is central to understanding biological processes and complex biological states such as disease. A common way of enumerating and communicating these interactions is via "pathway" diagrams. The complexity and abstraction represented in a pathway is decided by its author, generally a biologist attempting to convey the results of his research and educate others about the interaction between a set of genes, proteins, and small molecules. There are over 200 data sources that provide biological process information in the form of pathways. The lack of a standard pathway representation impedes users' abilities to integrate information from multiple pathway data sources in order to ensure a comprehensive understanding of the biological process or pathway that he is studying.

BioPAX is an emerging format for sharing pathways that aims to provide a standard for representing metabolic, biochemical, transcription regulation, protein synthesis, and signal transduction pathways. Currently, leading pathway resources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), BioCyc, and Reactome make their data available in BioPAX format and BioPAX compatible viewers are available in pathway analysis tools such as PATIKA and Cytoscape.

However, integration of these BioPAX-compliant data sources, especially leading resources such as KEGG, Reactome and BioCyc, into one resource has yet to happen. Much of the early focus has been on translating and exporting data from existing resources into the BioPAX format rather than building an integrated resource. Such an integrated pathway knowledgebase will allow for qualitative modeling and reasoning about large biological systems using computational tools.

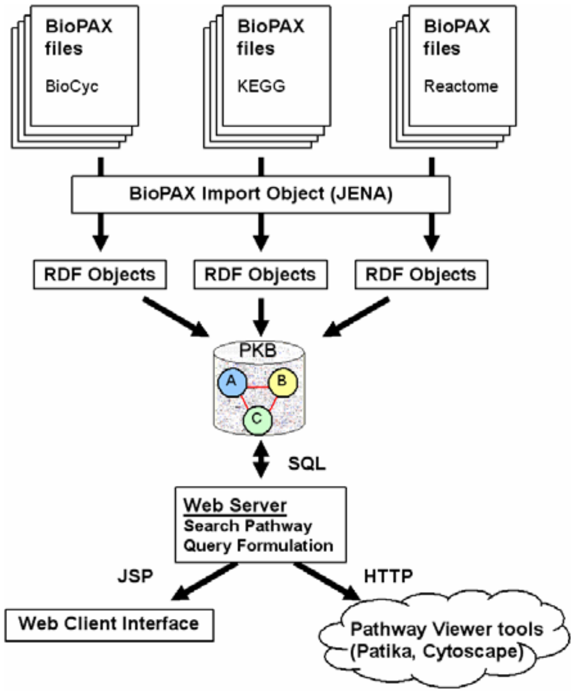

Figure 1 BioPAX objects are converted to RDF objects and stored in PKB. Information is accessed via SQL and presented via the PKB client interface.

We have integrated data from KEGG, Reactome and BioCyc using BioPAX to build the Pathway Knowledge Base (PKB, http://pkb.stanford.edu). Our current implementation contains over 2000 pathways from three species (Homo sapiens, S. cerevisiae and E. coli) and stores pathway data as resource description framework (RDF) triples. PKB has over 1,101,551 triples in the database which represent 2067 pathways, 7007 reactions and 9061 publication references. 979 of these pathways are human, 489 E. coli, and 599 are yeast.

We will present the technical challenges in building PKB and discuss the benefits and limitations of our approach: storing and querying BioPAX compliant pathway data in an Oracle RDF store. We will describe the cross-species, cross-database queries enabled by PKB as well as discuss novel analyses enabled by PKB such as the ability to systematically compare multiple pathway sources for differences and the opportunity to write "Knowledgebase proof-readers" to verify the integrity of data from multiple resources at once.