Latest from the Data Integration Frontier: Using Pathway Data to Build and Annotate Systems Biology Models

Oliver Ruebenacker[1], Ion I. Moraru[1], and Michael L. Blinov[1]

***Short Abstract* —As the biomedical research community accumulates data at a rapidly growing pace, the key question becomes whether one can easily use the data from different research domains. We present a software solution illustrating an important test case: using pathway data to build, extend and annotate kinetic models, such as expressed in Systems Biology Markup Language (SBML). While building models with the Virtual Cell (VCell), the Systems Biology Linker (SyBiL) gives VCell users access to pathway data describing thousands of biochemical interactions and their participants. Queries are submitted from VCell to Pathway Commons via SyBiL and results are imported back to VCell, where the information is used to create or annotate model components. Both lexical and semantic queries are supported. For a lexical query, a search string (e.g. name or keyword) is used to identify potentially relevant substances and pathways. Once a relevant substance or pathway is identified, a semantic query can identify related substances and pathways, and offer the user to pick any substance or interaction for inclusion in the model. Data describing which substances participate in which interaction can be used to create new model components, while annotations are used to store any other data of interest, including links to external data.**

***Keywords* — kinetic modeling; pathway databases; SBML; BioPAX; systems biology; protein-protein interactions.**

Biomedical research is at a stage where huge amount of data is accumulating at an accelerating pace. However efficient ways of communicating sets of data have been established primarily within, and not so much between, different research domains. In particular, transferring data from the pathway community to the modeling community yields valuable benefits, but it remains a challenge.

The pathway community needs to extract data from publications, link, categorize and catalog it, and allow researchers to query and visualize it. Therefore, the pathway community adopted Biological Pathways Exchange (BioPAX), [1], a format based on an OWL ontology supporting these tasks.

The modeling community needs to build mathematical models, which involves identifying quantifiable entities, and creating a mathematical model for simulation. For this purpose, XML-based formats have been developed, and used both as exchange formats, such as SBML [2], and as specific formats, such as Virtual Cell Markup Language (VCML).

As a result, BioPAX and SBML/VCML are structurally and semantically very different, which makes mapping one to one ambiguous or may not correspond to modeling assumptions.

We designed the SyBiL [3] as a part of the VCell modeling and simulation environment [4] to give modelers access to pathway data describing tens of thousands of biochemical interactions and their participants. Pathway data from public databases are queried and retrieved as BioPAX files through Pathway Commons. The data is used to create or annotate model components, to build a new model or to extend an existing one.

SyBiL addresses difficulties in the translation from BioPAX to SBML and VCML by providing a flexible conversion mechanism, which is based on the OWL-based BioPAX-SBML bridging ontology Systems Biological Pathway Exchange (SBPAX), that we have recently developed [5]. SBPAX allows mapping of all common elements to OWL objects and representing of all relevant relationships. Since SBML and VCML allow attaching Resource Description Framework (RDF) data to the mapped model elements, BioPAX, and SBPAX can be included as RDF annotations.

SyBiL uses various types of queries. A lexical query submits a search string and returns components (e.g. substance, reaction) that can be used to extend or annotate a model.

A semantic query submits a component and obtains components linked to it. The user can query for entire pathways and select from them, or query for a component's neighbors - reactions and substances directly connected to it. By retrieving the neighbors of the neighbors and so forth, SyBiL can gradually build a larger network. SyBiL can also query for relationships between two given components, e.g. identifying the shortest chain of neighbors connecting the two. This search can exclude chains passing through highly connected nodes (e.g. ATP).

By bringing the power of the Semantic Web to the modeling community, we turn models into a part of an interlinked and ever growing web of biomedical data.

References

[1] Luciano JS. (2005) *Drug Discov Today* **10**, 937-42.
[2] Hucka M, *et al.* (2003) *Bioinformatics* **19**, 524-531.
[3] Ruebenacker O, *et al*. (2007) *Proc. of the 2007 IEEE BIBE*: 339-348.
[4] Moraru II,. *et al*. (2008) IET Syst Biol. 2(5):352-62.
[5] Ruebenacker O, *et al*., submitted to IET Systems Biology.

This work has been supported by NIH grants RR022232 and GM076570.

[1]Center for Cell Analysis and Modeling, University of Connecticut Health Center, Farmington, CT. E-mail: oruebenacker@uchc.edu, moraru@uchc.edu, blinov@uchc.edu.