Pattern search in BioPAX models

Özgün Babur, Bülent Arman Aksoy, Igor Rodchenkov, Selçuk Onur Sümü, Chris Sander, and Emek Demir

1Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, Box 460, New York, NY 10065, USA
2Tri-Institutional Training Program in Computational Biology and Medicine, New York, USA
3Banting and Best Department of Medical Research, The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada

ABSTRACT

Motivation: BioPAX is a standard language for representing complex cellular processes, including metabolic networks, signal transduction and gene regulation. Due to the inherent complexity of a BioPAX model, searching for a specific type of subnetwork can be non-trivial and difficult.

Results: We developed an open source and extensible framework for defining and searching graph patterns in BioPAX models. We demonstrate its use with a sample pattern that captures directed signaling relations between proteins. We provide search results for the pattern obtained from the Pathway Commons database and compare these results with the current data in signaling databases SPIKE and SignaLink. Results show that a pattern search in public pathway data can identify substantial amount of signaling relations that do not exist in signaling databases.

Availability: BioPAX-pattern software was developed in Java. Source code and documentation is freely available at http://code.google.com/p/biopax-pattern under Lesser GNU Public License.

Contact: pattersearch@cbio.mskcc.org

1 INTRODUCTION

BioPAX is a community standard for pathway representation developed by a broad community of researchers working on pathways and related resources (Demir et al., 2010). It can represent metabolic and signaling pathways, molecular and genetic interactions and gene regulation networks. Currently, there are more than 30 pathway resources that support BioPAX representation, listed in Pathguide (Bader et al., 2006).

The detailed structure of BioPAX allows it to cover a wide spectrum of biological phenomena; however, its complexity also creates a significant barrier for researchers to effectively use it. Paxtools is a Java library that was developed to address this need. It exposes BioPAX models as Java objects and provides an array of utility methods (Demir et al., 2010) to query its contents. Although Paxtools facilitates simple searches, coding more complex queries that require evaluating links between multiple objects can be tedious and error-prone. What is often needed is a way to efficiently define and run such queries as graph patterns.

Existing graph searching tools for biological networks is limited to simple binary graphs that use nodes for molecules and edges for the interactions between (Giugno and Shasha, 2002; Berg and Lässig, 2004; Ferro et al., 2007). Generic RDF/OWL tools such as SPARQL and OWL-Reasoners allow defining more complex searches thus is more suitable for searching rich BioPAX pathways. Unfortunately, there are still many biologically relevant patterns that can not be captured by RDF/OWL tools (see Supp. Doc).

Here we present a framework and software tool that allows users to specify complex patterns using the rich BioPAX ontology and search any BioPAX level 3 model for those patterns. We demonstrate the tool with a pattern sample that we query in Pathway Commons (Cerami et al., 2011) database and provide the search result as Supplementary Data. The tool can be used as a standalone application or a library to seamlessly integrate pattern searches into the existing software.

2 METHODS

We define a pattern as a fixed number of BioPAX elements that satisfy a list of constraints. Constraints are re-usable objects that can be mapped to elements in the pattern.

Consider a case where we want to detect reactions that post-translationally modify a protein – an important question for proteomic data analysis. An example is shown in Figure 1a, where RAF1 is activated by phosphorylation and translocation. Figure 1b contains a diagram representing a sample pattern that captures such a relation. This pattern is composed of 4 BioPAX elements and 5 constraints. The pattern includes an EntityReference (ER), which has at least two different PhysicalEntity associated (PE1 and PE2)1, and a Conversion (Conv) that has PE1 and PE2 as participant on different sides.

Constraints that can enumerate matching objects of last element for a given set of assigned prior elements are called generative constraints. For example in Figure 1c, the first constraint can search for member PhysicalInfinity objects, given that the EntityReference is already assigned, thus it is generative. However, the last constraint checks for inequality and needs both of its mapped elements already assigned to an object, thus it is not generative.

1 EntityReference provides mapping to a specific entry in a reference molecule database, like UniProt. PhysicalEntity, on the other hand, is a specific modification state at a specific cellular location of that entity.
Users can run a graph search using constructed patterns on any level 3 BioPAX model. Previous levels of BioPAX can be auto-converted to level 3 using Paxtools. Running times of the searches depend on the pattern to be searched. The search method is iterative, i.e. it evaluates candidate members for the slots in the pattern linearly and outputs every possible matching.

It is also possible to pre-assign some elements in the pattern to conduct more specific searches. An example is provided in the Suppl. Doc. Patterns can be easily extended, combined and re-used allowing users to define increasingly complex biological processes such as signaling cascades as patterns and share them with other researchers.

3 RESULTS AND DISCUSSION

We demonstrate the use of pattern searches with a biologically interesting example in the Pathway Commons database, which is an aggregate of public pathway databases that provide their data in BioPAX. The example pattern detects pairs of molecules where the first one is controlling an interaction that the second one participates in. This is useful for capturing signal transduction in the cell. Unlike the simple example in Figure 1, this pattern also handles homology relations and molecular complexes. The pattern and the search result are given in the Suppl. Doc.

Searching this pattern in Pathway Commons returns 32599 relations between 4695 proteins. We compared these results with the relations that we could get from public signaling databases SPIKE (Paz et al., 2011) and Signalink (Fazekas et al., 2013). Even though our results are comparable to other signaling databases in size, only about 10% of data is overlapping (see Suppl. Doc). This shows that there can be a lot to gain by searching patterns in public detailed models even though databases specialize for the information of interest. The project web site contains other examples and documentation on how to use.

Funding: This research was supported by NIH grants U41HG006623 and GM103504.

REFERENCES


