Connecting Ontologies for the Representation of Biological Pathways

Anna Maria Masci[1], Mikhail Levin[2], Alan Ruttenberg[3,*], Lindsay G. Cowell[2,*]

[1]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, USA
[2]Department of Clinical Sciences, University of Texas Southwestern Medical Center at Dallas, TX, USA
[3]School of Dental Medicine, University at Buffalo, NY, USA and Creative Commons, Mountain View, CA, USA
*These authors contributed equally

Abstract. Significant effort has been put into the creation of a multitude of large, publicly available pathway databases. Most make their content available in at least one of several standard representation formats, but there are limitations to existing pathway representation formats, including underutilization of a common set of biomedical ontologies. To address this limitation, we developed an approach to representing biological pathways that relies on the use of ontologies from the Open Biomedical Ontologies (OBO) Foundry, including the Relation Ontology (RO), and adheres to the logical principles of ontology development advocated by the Foundry. To demonstrate the utility of this representation approach, we have curated comprehensive pathway representations for the signal transduction pathways initiated by seven of the mouse Toll-like receptors (TLR). Current efforts include the development of approaches for utilizing these representations for pathway analysis.

Keywords: OBO Foundry, ontology, signal transduction pathway, semantic web, OWL

1 Introduction

Biological pathways are central to biology. As a consequence, significant effort has been put into the creation of a multitude of large, publicly available pathway databases, for example Reactome (http://www.reactome.org), Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/), and Pathway Commons (http://www.pathwaycommons.org/pc/). These databases provide tremendous value to the community by making vast quantities of pathway information available both for download and for web browsing.

While each of the pathway databases has its own internal representation, most make their content available in at least one of several standard representation formats, such as BioPax (http://www.biopax.org). The availability of pathway data in standard, machine-readable formats is extremely important given the frequent need to integrate pathway data obtained from different databases and to incorporate the data into custom-written bioinformatics algorithms.

Despite the widespread use of standard pathway representation formats, there are limitations on the extent to which pathway representations can be integrated and jointly analyzed, arising from underutilization of a common set of biomedical ontologies and the lack of a formalism for their use. Some pathway databases, such as Reactome, utilize common ontologies like the Gene Ontology (GO) (http://www.geneontology.org/) to annotate some (but not all) of their pathway events. Other pathway databases, however, have independently developed their own ontologies. The use of common ontologies for pathway annotation is beneficial, not just for supporting interoperability between different pathway databases, but also for supporting interoperability with other information resources. For example, one can easily use the GO annotation of a Reactome event to query the GO database for a list of genes annotated with the GO term.

Even when common ontologies are used to annotate the molecules or events in a pathway, there are still difficulties in interpreting such pathway annotations, as the relation of a GO term to an annotated event is not clearly defined, and detailed domain knowledge is often needed to discern the exact connection in each case.

which we interpret as referring to parts of proteins (i.e. material entities).

2.2 Representation of Toll-like Receptor Pathways

Creation of the ontology-based representation of TLR pathways involved the following steps. We first created a spreadsheet template to facilitate manual curation of the relevant information into the formal framework. The spreadsheet was designed to provide an intuitive organizational structure for biologist domain experts, to ease the process of entering information into the spreadsheet, and to facilitate the automated translation of the spreadsheet into a computable ontology representation format, such as OBO (http://oboedit.org) or the Web Ontology Language (OWL) (http://www.w3.org/2007/OWL). The template is set up such that each pathway is curated into a single spreadsheet comprised of two main parts, one part containing a list of all entities named in the pathway and one part containing a list of asserted relations that when taken together specify the participants in and structure of the pathway. The list of named entities includes, for each entity, a handle used to refer to the entity within the spreadsheet, the corresponding ontology term, and the unique identifier associated with the term in the OBO Foundry ontology that is the source for the term. The approach to representation is process-centric, and this is reflected in the list of asserted relations, the bulk of which relate a process to the molecules that participate in the process, as described below.

To curate relevant pathway information into spreadsheets, a domain expert reviewed the primary literature to obtain information for pathways initiated by receptors formed from TLR2, TLR3, TLR4, TLR5, TLR7, TLR8, and TLR9. For each entity included in the representation, the curator searched each of the ontologies listed above to identify the most appropriate term and obtain its unique identifier. The appropriateness of a term was determined by reading its definition as well as the definition of nearby terms (e.g. parent, child, sibling terms). When no appropriate term was available, a term request was made to the appropriate ontology. For each term, the name of the source ontology, the term label from the source ontology, and the term's unique identifier from the source ontology were recorded in the templated spreadsheet.

Scripts to translate the templated spreadsheets into OWL 2 were written in common lisp (ABCL) (http://common-lisp.net/project/armedbear/), calling Java libraries (e.g. APACHE-POI and OWL-API). Imported terms were included using the Minimum Information to Reference an External Ontology Term procedure (http://obi-ontology.org/page/MIREOT).

3 Results

The foundation of our approach to representing biological pathways is the use of terms from OBO Foundry ontologies to name pathway entities, and the assertion of RO type-level relations between the entities to specify the pathway's structure (Figure 1). RO type-level relations (relations between classes) are defined in terms of RO instance-level relations (relations between individuals), and most are defined with an **all-some** structure. Thus, where capital letters indicate types (e.g. `A`, `B`), lower case letters indicate individuals (e.g. `a`, `b`), and *R* and *R** are type- and instance-level relations, respectively, the assertion `A` *R* `B` is interpreted as follows: for **all** individuals `a` of type `A`, there exists **some** individual `b` of type `B` such that `a` *R** `b`. For example, '`nucleus` *part_of* `cell`' is interpreted as: for any individual `nucleus n`, there exists some individual `cell c` such that `n` *part_of** `c`. We have also used a relation submitted for inclusion to RO (*realizes*) which is defined with an **all-only** structure interpreted as: for all individuals `a` of type `A`, if `a` *R** `b` then `b` *is_a* `B`. Phrased another way: only individuals of type `B` can stand in relation *R** to individuals of type `A`. Triples such as `A` *R* `B` are translated to OWL class expressions that encode the intended interpretation.

We formed a set of high-level triples that specify the types of assertions used in our approach to pathway representation. Each high-level triple specifies the RO type-level relation and the types of entities it joins. Specific versions of these triples are used to build each specific pathway representation. In the description below, RO relations are shown in italics, and terms referring to types of entities include a subscript indicating the ontology from which the term was taken. For example, in the high-level triple templates, <protein>$_{\text{PRO}}$ indicates that a term for a type of protein is taken from PRO. In a specific triple, 'TLR4$_{\text{PRO}}$' indicates that the term 'TLR4' is taken from PRO. In our representation the unique identifier from the source ontology identifies all terms, but for ease of presentation we use labels in what follows.

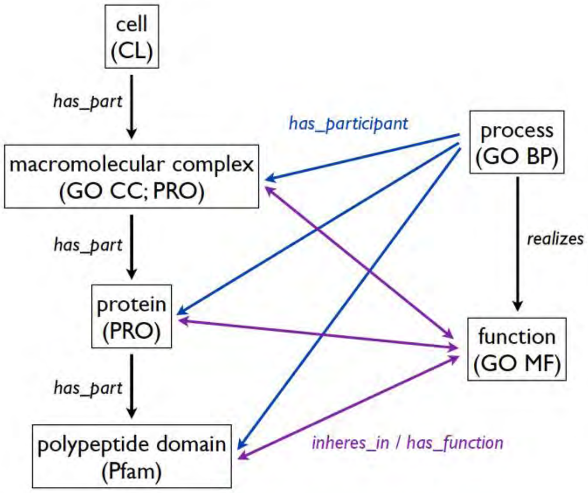

Figure 1: Ontology-based representation biological pathways. Each box corresponds to a type of entity. Shown in parentheses are abbreviated names for the ontologies from which terms for entities of that type are imported. Arrows between boxes represent relations between the entity types. Blue arrows represent the *has_participant* relation. Purple arrows represent the *inheres_in* and *has_function* relations. Black arrows are as labeled.

3.1 Relations between Independent Continuants

The RO relation *has_part* as defined between types of independent continuants [2] is used to relate macromolecular complexes to their component proteins and to relate proteins to protein domains using the assertions

<macromolecular complex>$_{\text{GOCC/PRO}}$[1] *has_part* <protein>$_{\text{PRO}}$

<protein>$_{\text{PRO}}$ *has_part* <polypeptide domain>$_{\text{Pfam}}$[2].

Thus, we make no distinction between the type of part-whole relationship that obtains between complexes and their components and that which obtains between proteins and their domains. The RO *has_part* relation is defined with sufficient generality that it holds in both cases.

The TLR4 pathway representation includes these specific triples using the *has_part* relation:

TLR4:MD2$_{\text{PRO}}$ *has_part* TLR4$_{\text{PRO}}$

TLR4:MD2$_{\text{PRO}}$ *has_part* MD2$_{\text{PRO}}$

TLR$_{\text{PRO}}$ *has_part* TIR domain$_{\text{Pfam}}$.

These triples assert that TLR4:MD2 protein complexes have the proteins TLR4 and MD2 as part, and TLR proteins have TIR domains as part. Note that macromolecular complexes may have non-protein parts, which we do not currently specify.

To assert the relationship between cell types and their cell surface receptors (which may be proteins or complexes), we use the *has_plasma_membrane_part* relation used in the CL:

<cell>$_{\text{CL}}$ *has_plasma_membrane_part* <macromolecular complex>$_{\text{GOCC/PRO}}$

<cell>$_{\text{CL}}$ *has_plasma_membrane_part* <protein>$_{\text{PRO}}$.

For example,

dermal dendritic cell$_{\text{CL}}$

has_plasma_membrane_part TLR4$_{\text{PRO}}$.

has_plasma_membrane_part is defined in terms of the RO *has_part* relation and the GOCC term 'plasma membrane' [10] and is currently a candidate relation submitted to the RO.

3.2 Relations Between Independent Continuants and Processes

Fundamentally, pathways are collections of interconnected processes, linked through the requirement of one process for a participant produced by another process. Thus, relations between processes and their participants are central to our representation approach. We use the RO relation *has_participant* [2] and a proposed subrelation *has_output_participant* (http://www.berkeleybop.org/ontologies/obo-all/ro_proposed/ro_proposed.obo.html) to relate processes to the molecules that participate in them and are produced by them, creating assertions of the form

<process>$_{\text{GOBP}}$ *has_participant* <macromolecular complex>$_{\text{GOCC/PRO}}$

[1] The term 'macromolecular complex' is a GOCC term. Terms for specific complexes are taken from either GOCC or PRO.

[2] The term 'polypeptide domain' is a Sequence Ontology term (http://www.sequenceontology.org/). Terms for specific types of domains are taken from Pfam via a prototype translation to OWL available on request.

<process>$_{\text{GOBP}}$ *has_participant* <protein>$_{\text{PRO}}$

<process>$_{\text{GOBP}}$ *has_output_participant* <macromolecular complex>$_{\text{GOCC/PRO}}$

<process>$_{\text{GOBP}}$ *has_output_participant* <protein>$_{\text{PRO}}$.

For example,

TLR4:MD2 complex assembly$_{\text{GOBP}}$ *has_participant* TLR4$_{\text{PRO}}$

TLR4:MD2 complex assembly$_{\text{GOBP}}$ *has_participant* MD2$_{\text{PRO}}$

TLR4:MD2 complex assembly$_{\text{GOBP}}$ *has_output_participant* TLR4:MD2$_{\text{PRO}}$

which represents participation of TLR4 and MD2 in a process by which the TLR4:MD2 complex is formed.

To distinguish types of process participants, we relate the participants in a process to the functions they manifest in that process. To do so, we utilize relations submitted to RO and defined on the basis of the treatment of functions in the Basic Formal Ontology (BFO), the upper-level ontology for OBO Foundry ontologies. According to this treatment, functions are dispositions to participate in processes that belong to independent continuants and are manifested, or realized, when a continuant participates in a process of the relevant type. Thus, we have the following set of triples relating proteins to functions

<protein>$_{\text{PRO}}$ *has_function* <function>$_{\text{GOMF}}$

<function>$_{\text{GOMF}}$ *inheres_in* <protein>$_{\text{PRO}}$

along with similar triples for protein domains and macromolecular complexes, and these triples relating functions and processes

<function>$_{\text{GOMF}}$ *realized_in* <process>$_{\text{GOBP}}$

<process>$_{\text{GOBP}}$ *realizes* <function>$_{\text{GOMF}}$.

For example, the triple

TIR domain$_{\text{Pfam}}$ *has_function* TIR domain binding$_{\text{GOMF}}$

asserts that TIR domains are capable of binding to other TIR domains. Similarly,

phosphorylation$_{\text{GOBP}}$ *realizes* kinase activity$_{\text{GOMF}}$

asserts that phosphorylation processes are processes in which kinase functions are realized.

Under our approach, the full specification of a process involves assertions that combine the *realizes* relation with the *inheres_in* relation. For example, phosphorylation processes in which dual specificity mitogen-activated protein kinase kinase 3 serves as the kinase have the assertion

<process>$_{\text{GOBP}}$ *realizes* (kinase activity$_{\text{GOMF}}$ AND (*inheres_in* dual specificity mitogen-activated protein kinase kinase 3$_{\text{PRO}}$)).

3.3 Relations between Processes

The RO *has_part* relation as defined between types of occurrents [2] is used to relate complex processes or sets of processes to their component processes:

<process>$_{\text{GOBP}}$ *has_part* <process>$_{\text{GOBP}}$.

For example,

TLR4 signaling pathway$_{\text{GOBP}}$ *has_part* I-kappaB phosphorylation$_{\text{GOBP}}$.

Note that we do not specify any order to the processes in a pathway. These can be inferred from the participant assertions.

4 Discussion

We have developed an ontology-based approach to the representation of biological pathways with the goal of enhancing interoperability among pathway representations as well as between pathway and other information resources. Key features of our approach include (i) the use of terms from OBO Foundry ontologies to designate each entity in a pathway, rather than just as annotations, (ii) the use of RO relations to structure the pathway, and (iii) use of the same logical formalism as is used in developing OBO Foundry ontologies.

We anticipate several benefits from this approach. The use of terms from common ontologies to name pathway entities significantly reduces the ambiguity that can exist between different pathway representations regarding the named entities, thereby facilitating their integration into a single network for analysis. The use of common ontologies also facilitates the integration of pathways with other kinds of ontology-annotated data.

The use of ontological relations to specify the structure of pathways provides for the direct integration of pathways with ontologies and creation of a unified network that includes the ontology and pathway relations. We anticipate that such a unified network will support the use of the ontology hierarchies to further ease the difficulties of integrating heterogeneous pathway representations.

The ability to incorporate into pathway representations relations asserted in ontologies could reduce the effort of curating pathways, as many of the needed relations are being incorporated into ontologies. For example, the developers of PRO are adding *has_part* relations to PRO, and the CL developers are adding *has_plasma_membrane _part* assertions to CL.

We have encoded the TLR pathways using OWL and are currently developing algorithms that use the relationships encoded in the ontologies for pathway analysis. We have already seen benefits from using OWL for consistency checking to detect curation errors, and we anticipate significant benefit from its application to the detection of inconsistencies in integrated pathway representations. We are also utilizing OWL reasoning to support pathway queries and are currently evaluating the advantages it offers over the keyword querying available through most pathway resources.

The primary barrier we faced in applying this approach to the representation of the TLR pathways was the absence of software. The creation of representational artifacts built from portions of multiple ontologies would be improved by software that allows a user to:

- query a specific set of ontologies, select terms for import, and import specific pieces of information about the term;
- submit term requests to a specific ontology when a needed term cannot be found; and
- assert relations between the imported terms.

The availability of such software would allow this approach to be widely applied.

We see two possible disadvantages to this approach. Curation into the representation framework we describe may take longer or be less intuitive for domain experts than alternative representation frameworks. We believe that any disadvantage in this regard can be addressed through the development of curation software. A second possible disadvantage is a loss of expressivity through the exclusion of domain-specific relations. We believe that the opportunity to directly integrate portions of ontologies with pathways and compute over the integrated resource provides a benefit that outweighs any possible loss of expressivity. If, however, such a loss of expressivity did present a significant disadvantage, the ontology-based core representations could be enhanced with domain-specific information that may not be accessible to all analysis tools.

Acknowledgments

This work was supported by an NIAID-funded R01 (AI077706) and a Burroughs Wellcome Fund Career Award to LGC.

References

1. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nat Biotechnol, 2007. **25**(11): p. 1251-5.
2. Smith, B., et al., *Relations in biomedical ontologies.* Genome Biol, 2005. **6**(5): p. R46.
3. Natale, D.A., et al., *Framework for a protein ontology.* BMC Bioinformatics, 2007. **8 Suppl 9**.
4. Meehan, T.F., et al., *Logical development of the cell ontology.* BMC Bioinformatics, 2011. **12**: 6.
5. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D211-22.
6. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
7. Bard, J., S.Y. Rhee, and M. Ashburner, *An ontology for cell types.* Genome Biol, 2005. **6**(2): p. R21.
8. Diehl, A.D., et al., *Hematopoietic cell types: Prototype for a revised cell ontology.* J Biomed Inform, 2010.
9. Rosse, C. and J.L. Mejino, Jr., *A reference ontology for biomedical informatics: the Foundational Model of Anatomy.* J Biomed Inform, 2003. **36**(6): p. 478-500.
10. Masci, A.M., *et al.*, *An improved ontological representation of dendritic cells as a paradigm for all cell types.* BMC Bioinformatics, 2009. **10**: 70.