Ontology-based annotation and integration of pathway databases

Lucy Lu Wang[1,], Mary E. Shimoyama[2], G. Thomas Hayman[2], Jennifer R. Smith[2], Monika Tutaj[2], and John H. Gennari[1]*

[1]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA
[2]Department of Biomedical Engineering, Medical College of Wisconsin, Milwaukee, WI, USA
**Contact: lucylw@uw.edu*

ABSTRACT

Biological pathway alignment is necessary to reduce redundancy in pathway data for secondary analysis, but it is difficult to identify semantically similar pathways to align based on entity membership alone. Annotations to the Pathway Ontology (PW) can be used to identify semantically similar pathways. This paper describes a computationally assisted method for annotating pathways to classes in the PW. An ensemble model using lexical features and ontology matching software was used to derive PW annotations for Reactome pathways. Proposed annotations were reviewed by the authors and PW curatorial team for correctness and inclusion into the PW.

1 INTRODUCTION

Biological pathways provide high-level representations of biological processes; they are important abstractions for modeling complex interactions in the human body. Pathways are used to visualize complex processes, connect and understand disease networks, analyze gene expression data, and bridge physiological models. To provide proper grounding for these various types of analyses, pathways from different databases are often combined, resulting in overlaps in the pathway data set. These overlapping pathway graphs challenge our existing statistical algorithms, generating errors in analysis results. To reduce these errors, we propose organizing pathways under a single ontology, the Pathway Ontology (PW), and normalizing pathways within each ontology class. The normalized pathways would cover a broad range of biological processes while remaining humanly interpretable.

This article describes the motivation behind ontology-based merging of pathway databases, and ongoing work annotating pathways with PW classes. The following terminology is used. *Pathway instances* are individual pathways from databases, and consist of descriptive information, computational representations of biological processes, and associated pathway diagrams. Descriptive information includes names, definitions, and synonyms; computational representations describe entities and relationships, and are available in standard pathway file formats BioPAX, SBML, or GPML. *Pathway databases*, like Reactome or KEGG, are repositories of pathway instances curated by database editors. *Pathway aggregators* like Pathway Commons (Cerami *et al.*, 2011) or ConsensusPathDB (Kamburov *et al.,* 2009) collect data from multiple pathway databases, but do not provide additional organization of pathway instances. Instances are organized under *pathway classes*, e.g., glucose metabolic pathway, or *Wnt* signaling pathway, describing specific processes. A *pathway ontology* is an ontology of pathway classes. An ontology describes how classes are related, through *is-a*, *part-of*, and other relationships.

Most pathway databases have their own organizational scheme, which can be considered a pathway ontology of sorts. However, this is different from *the* Pathway Ontology (PW), an ontological resource created as part of the Rat Genome Database (RGD) that seeks to provide a unified organization for all pathways (Petri *et al.*, 2014). Although pathway instances may already be organized under the schema of their origin database, it is difficult to infer a mapping between these various organizational schema. Instead of matching schema between all databases, we can annotate all pathways to one schema or ontology, such as the Pathway Ontology. Figure 1 shows an example of an existing annotation between a KEGG pathway instance and the PW. With pathways organized under the PW, the relationships between various instances and classes become consistent and clear.

The Pathway Ontology was developed as an effort to catalog and describe the relationships among various biological pathways. The ontology covers broad pathway categories such as metabolic, regulatory, signaling, disease, and drug pathways, and allows for the representation of both subclass and mereological hierarchies via the *is-a* and *part-of* relationships respectively. Pathway from KEGG, the National Cancer Institute's Pathway Interactions Database (PID), and SMPDB, have already been partially or completely annotated with PW classes. However, several large pathway databases, such as Reactome and BioCyc,

containing thousands of pathway instances, have not been annotated in the PW.

Our goal is to use the Pathway Ontology to anchor the organization of pathways derived from different databases. By annotating pathways with the appropriate classes from the PW, we can identify semantically similar pathways. Similar pathways containing content overlap or duplication can then be combined to produce normalized pathway representations for secondary analysis. In this article, we discuss our initial steps toward these goals. As a preliminary example, we discuss mapping the pathways from the Reactome database to classes in the PW.

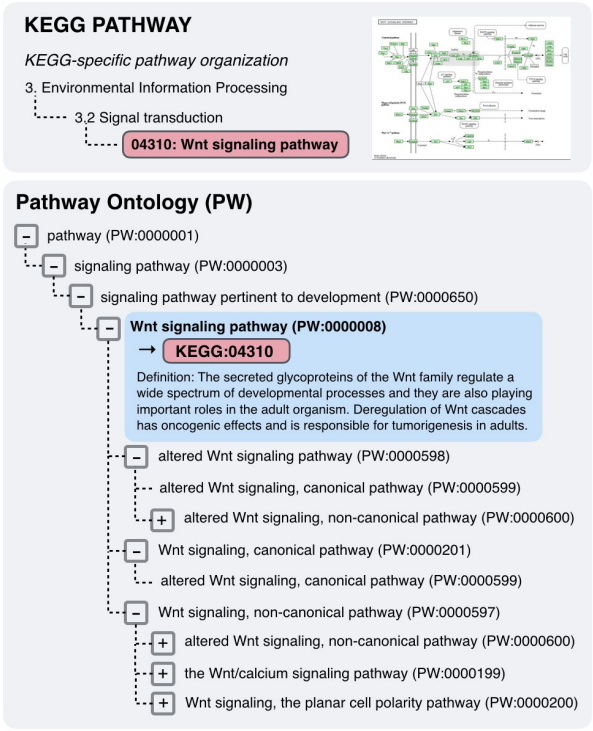

Fig. 1. The *Wnt* signaling pathway in KEGG is annotated to a class in PW. KEGG:04310 is organized under a KEGG-specific organizational scheme (*above*), which does not correspond to the PW *Wnt* signaling class hierarchy (*below*).

2 PATHWAY CLASSIFICATION MODEL

Reactome is a large and popular pathway database hosting pathways from several model organisms (Croft *et al.*, 2013). It contains over 2,200 human pathways. To map pathways from Reactome to the PW, we combine computational and manual approaches. Due to the large number of pathways in Reactome, manual review of each pathway instance-class pair is untenable. Computational techniques are necessary to reduce the number of annotations requiring curator review. All analysis described below uses version 7.42 of the Pathway Ontology released 2016, Oct 21, and version 59 of Reactome. Source code and documentation can be found at https://github.com/lucylw/pathhier.

Figure 2 shows the workflow for annotating pathway instances. This workflow can be applied to any pathway database, but is initially optimized for Reactome. A classification model produces preliminary annotations between Reactome pathway instance and PW classes. These preliminary annotations are then manually reviewed by a group of three PW curators from RGD (GTH, JRS, MT). For each source pathway instance, the curators determine which PW class(es) is the best fit for annotation.

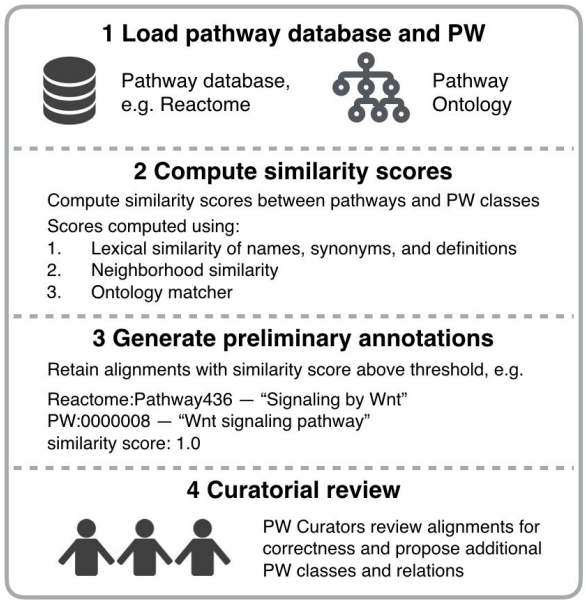

Fig. 2. Workflow for annotation.

For each pathway in the human Reactome BioPAX export, we extract the pathway's name, synonyms, definition, relationships to other pathways, cross-reference identifiers, and associated entities. We consider associated entities as those that are direct participants in pathway reactions, or those that modify or control any of the reactions. We next apply a candidate selection module that uses simple lexical features to choose candidate PW classes for each Reactome pathway. For each candidate pair, a matching algorithm based on lexical similarity and parent-child relations is used to generate a similarity score. Our algorithm primarily uses word token Jaccard indices weighted by the inverse document frequency (*idf*) of tokens. The Jaccard index is a measure of set similarity, and the *idf* is a measure of the rarity of a word in a corpus of text. Together, they give a measure of unique word token overlaps between two strings.

We tokenize the pathway names and stem the resulting word tokens using the Python nltk library. We then compute the *idf* score of each token as $log(N / df)$,

where N is the total number of pathways, and *df* the document frequency, in this case the number of pathway names in which the token occurs. For each candidate pair, we compute a weighted Jaccard index between the name of the pathway instance and the name of the PW class. In computing the weighted Jaccard, each token, instead of being of equal weight to other tokens, is weighted by its *idf*. Where the normal Jaccard index is computed as:

$$J = \frac{A \cap B}{A \cup B}$$

with A as the set of tokens in the pathway instance name and B the set of tokens in the PW class name, the weighted Jaccard index is calculated as:

$$J_{weighted} = \frac{\sum_{t \epsilon A \cap B} idf(t)}{\sum_{t \epsilon A \cup B} idf(t)}$$

Here, the function *idf(t)* retrieves the *idf* score of word token *t*. This yields a weighted Jaccard index between 0.0 and 1.0, where 1.0 indicates that the two sets of tokens are equivalent. In addition to computing this index for pathway names, we also apply it to tokens that make up pathway definitions, as provided by both Reactome and the PW. These similarity scores are used by the model to rank instance-class pairs for annotation.

We also compute a similarity score using OntoEmma (https://github.com/allenai/ontoemma), an ontology matcher. Ontology matchers are designed to generate alignments between the class trees of two ontologies, and take advantage of the entity descriptions and relationships between classes during matching. The organizational schema of a pathway database is comparable to an ontology, so we employ an ontology matcher to generate alignments. OntoEmma uses lexical features computed on the names and definitions of entities, as well as basic parent-child relationships to generate a similarity score. Instance-class matches generated by OntoEmma are combined with those generated by our classification model.

We thresholded these similarity scores to produce a preferred list of annotations. All pathway instance-class pairs with similarity scores above a 0.2 threshold are retained for manual review. This threshold was selected to achieve a balance of high recall without subjecting curators to too many spurious matches. The list of proposed annotations is then reviewed by PW curators.

3 EXAMPLE ANNOTATIONS

The classification model generated 5,227 candidate annotations between Reactome pathway instances and PW classes with similarity score above the 0.2 threshold. Initial review by curators produced some correspondences between Reactome pathways and PW classes, as well as observations for improving the classification model.

Table 1 shows example annotations for the top 10 Reactome pathways that were selected by the classification model in relation to PW:0000008, the class "Wnt signaling pathway." Each proposed instance-class pair was evaluated by curators, who proposed the annotations shown in the right column to classes in the *Wnt* signaling class hierarchy (Figure 1). The first line in the table suggests equivalence between Reactome:Pathway436 and PW:0000008. Other proposed matches are related to PW:0000008. In some cases, the existing properties used in the PW may be insufficient for conveying granular relationships, and new properties (shown in red) were proposed. Two matches produced by the algorithm (Pathway729 and Pathway340) were incorrect, and did not lead to any annotations to *Wnt* signaling pathways.

The classification model was unable to match some Reactome pathways to PW classes. These pathways differ too much from existing PW classes, either due to lexical/syntactic differences in names and definitions between database and ontology, or because they belong to classes yet to be included in the PW. These pathways represent opportunities to add new classes to the PW.

4 DISCUSSION

We have shown some initial annotations of Reactome pathways to PW classes using a computationally assisted method for selecting potential annotations using lexical and topological similarity. The goal is to reduce the burden on curators. Continuing, we plan to complete annotations for Reactome and evaluate the quality of the completed annotations.

To increase the number of correct annotations, especially among pathways for which lexical match is currently insufficient, we intend to leverage existing PW annotations to databases like KEGG, NCI PID, and SMPDB. These annotations are the result of prior curatorial efforts, and can be used to bootstrap a supervised learning model using pathway entity membership and pathway-subpathway relationships as similarity features. In that case, we may encounter issues involving entity disambiguation.

To flexibly accommodate knowledge from various pathway resources, we may need to extend the coverage of the PW. If we can develop a sufficiently comprehensive and inclusive ontology, then when we annotate pathways to classes in that ontology, we can better leverage the expressiveness of pathway classes and descriptions to recognize synonymy and other relationships between pathways. As we complete annotations for Reactome and other pathway databases of interest, we will use our results to update the PW

Table 1. Top 10 Reactome pathways associated with PW:0000008 "Wnt signaling pathway." Annotations to existing PW classes are proposed by curators; newly proposed properties are shown in red.

Reactome ID	Name	Proposed PW annotation
Pathway436	Signaling by Wnt	*has_exact_synonym* PW:0000008
Pathway702	Signaling by WNT in cancer	*part_of* OR *has_related_synonym* PW:0000598
Pathway441	TCF dependent signaling in response to WNT	*part_of* OR *has_related_synonym* PW:0000201
Pathway452	Beta-catenin independent WNT signaling	*has_exact_synonym* PW:0000597
Pathway720	RNF mutants show enhanced WNT signaling and proliferation	*part_of* OR *has_related_synonym* OR *regulates* PW:0000599
Pathway442	WNT mediated activation of DVL	*part_of* OR *has_related_synonym* PW:0000201 *part_of* OR has_related_synonym PW:0000597
Pathway706	AXIN mutants destabilize the destruction complex, activating WNT signaling	*part_of* OR has_related_synonym OR *regulates* PW:0000599
Pathway729	Oncogenic MAPK signaling	NONE
Pathway437	WNT ligand biogenesis and trafficking	*has_related_synonym* OR *upstream_of PW:0000201* *has_related_synonym* OR *upstream_of PW:0000597* *has_related_synonym* OR *upstream_of PW:0000599* *has_related_synonym* OR *upstream_of PW:0000600*
Pathway440	Repression of WNT target genes	*has_related_synonym* OR *downstream_of* PW:0000201 *has_related_synonym* OR *downstream_of* PW:0000597 *has_related_synonym* OR *downstream_of* PW:0000599 *has_related_synonym* OR *downstream_of* PW:0000600

framework. Following annotation, we aim to normalize pathways within each PW class to reduce redundancy.

Unlike statistical approaches such as PathCards (Belinky *et al.*, 2015) or ReCiPa (Vivar *et al.*, 2013), we retain better interpretability of normalized pathways by observing class boundaries in the PW. As described in previous publications addressing pathway data integration, we face many similar challenges, such as the usage of different organizational schemes by different databases, incomplete or inconsistent description of pathway-subpathway relationships, and differences in identifier and semantic choices in representing pathway data among the various source databases (Bauermehren *et al.*, 2009; Vivar *et al*., 2013; Belinky *et al.*, 2015; Wang *et al*., 2016). Using a unifying ontology for organization at the pathway level ameliorates the first two of these challenges. To address the third, we plan to expand on methods described in prior work combining techniques in entity disambiguation and global graph alignment (Wang *et al.*, 2017).

Conclusion

Pathway representations are critical for modeling and understanding the physiological processes underlying both normal and disease health states, but a lack of understanding of the relationships between pathways of different provenance undermine their collective usability. Combining data from different pathway databases under a unifying ontology could address many of these issues. We demonstrate preliminary work constructing a computationally-assisted pipeline for annotating Reactome pathways to classes in the Pathway Ontology. We are working to improve the quality and quantity of proposed annotations using feedback from curators. Following the completion of this annotation phase, we will proceed by aligning pathways under each PW class, generating normalized pathway representations. Merging pathways along ontological class lines will produce better and cleaner pathways for use in secondary statistical analysis.

REFERENCES

Bauer-Mehren, A., Furlong, L. I., and Sanz, F. (2009). Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular Systems Biology*, **5**(1), 290.

Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., and Lancet, D. (2015). PathCards: multi-source consolidation of human biological pathways. *Database* (Oxford), 2015.

Cerami, E. G., Gross, B. E., and al, E. D. e. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res*, **39**(Database issue), D685–690.

Croft, D., Mundo, A. F., and al, R. H. e. (2013). The Reactome pathway knowledgebase. *Nucleic Acids Res*, **42**(Database issue), D472–477.

Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009). ConsensusPathDB – a database for integrating human functional interaction networks. *Nucleic Acids Res*, **37**(Database issue), D623–628.

Petri, V., Jayaraman, P., Tutaj, M., Hayman, G. T., Smith, J. R., De Pons, J., et al. (2014). The pathway ontology -- updates and applications. *J Biomed Semantics*, **5**(7).

Vivar, J. C., Pemu, P., McPherson, R., and Ghosh, S. (2013). Redundancy Control in Pathway Databases (ReCiPa): An Application for Improving Gene-Set Enrichment Analysis in Omics Studies and Big Data Biology. *OMICS*, **17**(8), 414–422.

Wang, L. L. and Gennari, J. H. (2017). Similarity Metrics for Determining Overlap Among Biological Pathways. In Proceedings of the International Conference on Biomedical Ontology, Newcastle upon Tyne, UK.

Wang, L. L., Gennari, J. H., and Abernethy, N. F. (2016). An Analysis of Differences In Biological Pathway Resources. In Proceedings of the International Conference on Biological Ontology and BioCreative, Corvallis, OR.