Similarity Metrics for Determining Overlap Among Biological Pathways

Lucy L. Wang*, John H. Gennari

Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, USA

ABSTRACT

Systems biology researchers often rely on the use of one or more *pathway resources* for analysis of gene expression data or experimental results. Unfortunately, there is no single, gold-standard pathway knowledge resource, nor are there good ways to merge or combine information from multiple resources. What is needed is clear organization of pathways, whereby represented processes can be enumerated and compared between resources. In this paper, we develop a set of similarity metrics based on (a) pathway participants, (b) pathway names and descriptions, and (c) pathway topological information, which can be used to infer similarity and hierarchical relationships among pathways from different databases. These inferred relationships can be used to derive annotations to the Pathway Ontology or other pathway organizational schemes.

1 INTRODUCTION

Pathway databases provide useful structured knowledge for bioinformaticists and systems biologists, who use pathways to assist in the analysis of gene expression data, build models of physiological processes, and explore the connections between therapeutics and disease. Pathways describe a set of biomolecular reactions and interactions. They can have somewhat arbitrary beginnings and endings, but they aim to capture the details of a biological process or function.

Researchers can choose from a large number of pathway databases and representations. The abundance of choice can lead to confusion, since different databases can offer redundant and sometimes conflicting accounts of the same pathway. Many applications of pathway resources naively combine pathway data sets from multiple resources (e.g. MSigDB, often used as a source of gene sets for gene set enrichment analysis (GSEA), includes gene sets derived from several pathway databases (Liberzon *et al.*, 2015); or ConsensusPathDB, which generates pathway-based networks using pathways from 32 resources (Kamburov *et al.*, 2009)), but both redundancies and conflicts can undermine the output produced by these tools. Results of secondary analysis using pathway databases will change depending on the database chosen (Green and Karp, 2006). Khatri et al discuss annotation inaccuracies in pathway databases as a challenge to pathway analysis (Khatri *et al.*, 2012). A recent publication by Ballouz et al also discusses bias in GSEA due to overlaps between gene sets used for analysis, which are often derived from pathways (Ballouz *et al.*, 2016). These difficulties arise because pathways share membership and content, which necessarily affects analysis performed using overlapping pathways. Instead of using pathways as they are, we propose that individual pathways from different databases should be pre-organized by similarity, and merged based on user needs to generate sets of improved pathways for secondary use.

In this paper, we report on our initial efforts at organizing and determining the overlaps of pathways from seven different well-known resources. Our long-term goals are two-fold: First, we wish to improve secondary analyses by creating a more consistent and custom-tailored set of pathways for use. Second, we aim to develop or improve on a standard nomenclature and organization for pathways. We see this as a significant gap for the development of biological ontologies. Although there are well-established, vetted reference ontologies for the participants of pathway processes (such as Entrez Gene for genes, UniProt for proteins and ChEBI for molecular entities (Maglott *et al.*, 2011; Apweiler *et al.*, 2004; Degtyarenko *et al.*, 2008)), ontologies for higher-level biological process names are lacking, or at least not well-used.

The Gene Ontology includes names for biological functions, but these are mostly at the reaction level, and are not well-organized into coordinated sets of reactions (Ashburner *et al.*, 2000). A better starting point for an organization of pathway knowledge is the Pathway Ontology (PW) (Petri *et al.*, 2014). The PW describes classes of pathways based on biological function. Pathways in the PW are organized using is-a and part-of relationships, where "A part-of B" indicates that A is a subprocess of B. Pathways with instantiation (is-a) or subprocess (part-of) relationships are located closer to their parent pathways in the PW hierarchy.

Current pathway resources (KEGG, Reactome etc.) do not use the PW, but instead may be organized via some custom-tailored ontology or hierarchy. This leads to problems when comparing sets of pathways from resources with disparate ontological structure and organization. When pathways from multiple resources are combined for secondary use, a shared overarching organizational scheme often does not exist. The employment of different ontologies or simply the lack of any initial ontological structure make it challenging to determine pathways that describe similar function.

There is no well-accepted standard measure of similarity among pathways. Other researchers have discussed similarity of pathway names, gene and molecular membership, and functional annotations as potential indicators of pathway content similarity (Grego *et al.*, 2010; Belinky *et al.*, 2015). In this paper, we assess these sorts of similarity metrics. We focus on three aspects of similarity: pathway names and descriptions, entity membership, and pathway topology.

We approach this problem as one analogous to record linkage and deduplication, where data from similar or identical records can be combined to yield better information (Christen, 2011). Instead of combining data records, we are identifying overlapping, subset, or duplicate portions of pathway representations from different pathway databases. The linkage process we apply to pathways consists of several steps: (1) data extraction and cleaning, (2) entity normalization, (3) indexing to yield pairs of pathways for

*To whom correspondence should be addressed: lucylw@uw.edu

comparison, (4) generation of similarity metrics, and (5) evaluation of output. This paper focuses on describing the procedures used to complete steps 1-4, and shows some initial examples of step 5.

We first extract and clean pathway data from seven well-known public pathway databases. We take advantage of Biological Pathway Exchange (BioPAX), an standard format supported natively by many pathway databases (Demir *et al.*, 2010), and meta-resources like PathwayCommons that provide standardized pathway data (Cerami *et al.*, 2011). Entity normalization involves identifying objects from different databases that reference the same biological entity. Indexing is an initial reduction of the number of in-depth pairwise pathway comparisons that need to be made. Metrics such as graph edit distance are computationally expensive, making exhaustive pairwise pathway comparisons cost- and time-prohibitive. Within the reduced pairwise comparisons, we can generate and evaluate similarity metrics such as entity membership overlap and graph similarity.

2 METHODS & RESULTS

To measure similarity among pathways, we use a combination of entity membership, pathway name and description, and topological similarity metrics. Entity membership overlap has been used in previous efforts to combine pathways into functionally similar superpaths (Vivar *et al.*, 2013; Belinky *et al.*, 2015). These superpaths can be used to generate gene sets with little to no redundancy. However, we hypothesize that there are differences between pathway overlap and pathway subset relationships that may benefit from more detailed investigation. Likewise, graph alignment methods have been used to compare pathways between species to discover evolutionarily conserved modules (Peregrn-Alvarez *et al.*, 2009; Muto *et al.*, 2013). In our case, we are interested in using these techniques to identify areas of similarity and differences between pathway representations.

2.1 Pathway data extraction

The dataset we used includes pathway data from each of seven resources: HumanCyc, the Kyoto Encyclopedia of Genes and Genomes (KEGG), the National Cancer Institute's Pathway Interactions Database (NCI PID), Panther Pathways, Reactome, Small Molecule Pathway Database (SMPDB), and WikiPathways (Romero *et al.*, 2004; Kanehisa and Goto, 2000; Schaefer *et al.*, 2009; Thomas *et al.*, 2003; Croft *et al.*, 2013; Frolkis *et al.*, 2010; Kutmon *et al.*, 2016) Of these, HumanCyc (v20), Panther (v3.4.1), Reactome (v59), and SMPDB (version published June 5, 2016) were acquired in BioPAX format from the resources directly, KEGG and NCI PID were downloaded in BioPAX format from Pathway Commons (PC8), and WikiPathways was exported in Graphical Pathway Markup Language (GPML) format on December 10, 2016. A total of 4,441 pathways were extracted, and the pathway counts per resource are given in Table (1).

For each pathway within these resources, we extracted the pathway name, any comments or descriptions of the pathway content, the set of entities participating in the pathway, the relationships between those entities, as well as any subpathway relationships (similar to the part-of relationship described in PW). Pathway entities are physical entities (protein, complex, molecule, DNA, RNA etc.) that participate (as a reactant, product, or modifier) in a reaction explicitly described as part of the pathway.

Resource	Number of pathways
HumanCyc	242
KEGG	122
NCI PID	745
Panther	177
Reactome	2080
SMPDB	724
WikiPathways	351
Total	4441

Table 1. Pathway counts per resource

The pathway is represented as an undirected graph where nodes represent physical entities, or concepts such as reactions, and edges represent relationships. Subpathway relationships were retained to assist in the exploration of subset relationships between pathways.

2.2 Entity normalization

Entity normalization is the process of identifying equivalent or similar entities from different pathway resources. Pathways from all seven resources annotate their entities using external reference identifiers, e.g., Entrez and UniProt identifiers for proteins, ChEBI identifiers for molecules, etc. These cross-reference identifiers offer a starting point for entity normalization. However, based on previous observations, these identifiers alone do not do a sufficient job of aligning like entities between databases (Wang *et al.*, 2016). One main issue is the existence of synonymous identifiers (e.g., secondary accession identifiers) and related identifiers (e.g., ChEBI conjugate acids/bases) in cross-reference databases, i.e., a single entity can reference two or more identifiers. Another issue is the existence of multiple cross-reference databases for a particular class of biological entities. We must therefore normalize entities both within and among different cross-reference identifier databases.

We generated an identifier normalization dictionary starting with the cross-reference identifiers given in each pathway database. If a single entity references two or more identifiers, for example, the protein "Tryptophan 5-hydroxylase 1" in HumanCyc references both Entrez:AAA67050 and UniProt:P17752, then we infer synonymy between these two identifiers regardless of their given synonymy in Entrez and UniProt. Further synonyms are derived from UniProt and ChEBI services. We queried for secondary accession numbers of all UniProt identifiers, and for secondary accession numbers, conjugate acids and bases, and tautomers of all ChEBI identifiers referenced in our dataset. Lastly, we supplemented this normalization dictionary using BridgeDB, a service for mapping identifiers across different cross-reference databases (van Iersel *et al.*, 2010). For identifiers extracted from our pathways, we queried synonyms from BridgeDB (Ensembl and UniProt for proteins, ChEBI and PubChem for molecules), which were used to derive further equivalences between different identifiers.

For the following entity membership comparisons, we normalized the entities in each pathway based on their cross-reference identifiers along with the additional synonyms we derived from BridgeDB, UniProt, and ChEBI. Although this improved the number of entities matched among resources compared to using naive cross-reference identifier matching alone, there were still many entities for which the appropriate normalization could not be obtained. Further normalization of both entities and relationships was explored using graph alignment techniques, which are discussed in section 2.5.

2.3 Indexing using pathway names and descriptions

Determining groups of similar pathways is a problem akin to that of record linkage and deduplication. Indexing techniques are used in record linkage problems to determine likely pairs of similar records, which can then be compared in depth (Christen, 2011). Dividing the data into blocks (blocking) and only comparing within blocks is an effective way to reduce computational cost. Due to the time and resource cost of computing graph edit distance, we employ blocking of pathway representations based on name and description similarity. This reduces the total number of pairwise comparisons from around 10 million (4,441 choose 2) to a much smaller number based on the number and sizes of blocks generated.

Using pathway name similarity as a measure for pathway content similarity has not been very successful in the past (Belinky *et al.*, 2015). Very few pathways share identical names across resources, and those with identical or similar names usually vary significantly on content, as measured by entity membership. However, even though pathway name alone is not a good proxy for content, it, along with a free-text description of the pathway, should yield blocks of pathways with higher within-block similarity than random chance. That is, we believe that names and descriptions offer some information about the content of a pathway representation.

Of the 4,441 pathways in our data set, only 2,627 had analyzable pathway descriptions. The remainder had either no pathway description, or a pathway description containing meta-information on the writing, editing, or reviewing of the pathway. Some resources, such as NCI PID and SMPDB, had no descriptions of any of the pathways encoded in their BioPAX exports. We therefore analyzed the pathway names and descriptions separately, generating two sets of pathway blocks. We describe the work done for pathway names; the process was repeated for pathway descriptions.

We treat each pathway name as a document, and cluster them into topics. This is accomplished by calculating the term frequency-inverse document frequency (tf-idf) statistic for each word in each pathway name. The tf-idf is a measure of the significance of a word to a group of documents, and is often used for term-weighting in topic modeling. The statistic is given by equation (1), where the statistic for $w_{i,d}$ (word i in document d) is the product of the term frequency $tf_{i,d}$ (how often the word i appears in document d), and the log inverse of the document frequency df_i (the number of documents in the corpus that contain word i) divided by N, the total number of documents in the corpus.

$$w_{i,d} = tf_{i,d} \times \log\left(\frac{N}{df_i}\right) \tag{1}$$

After computing tf-idf scores for all words in all pathway names, we performed k-means clustering on the tf-idf score vectors to generate blocks of similar pathway names. K-means clustering is a centroid-based unsupervised clustering method that yields k clusters from the input data where each data point belongs to the cluster with the closest mean. Due to the imbalance in the number of pathways from each resource, we expect many singleton clusters, those with only one member. We initially employed the elbow method to select the number of resulting clusters k, but because no clear elbow was seen in within-cluster variance as we increased the cluster count, we were unable to determine an ideal k experimentally. Instead, we calculated a theoretical minimum k based on the number of pathways in each resource. Assuming all pathways from all resources have matches in all other resources, the theoretical minimum number of clusters is 2,080. However, assuming that some pathways may have multiple matches in other resources and possibly within the resource itself, we selected a k of 1,040, 0.5 times the theoretical minimum, as a conservative choice. Similarly, for pathway descriptions, we computed a theoretical minimum of 2,053 clusters, and used a k of 1,026.

For N total pathways, blocking reduces the number of pairwise comparisons from $C(N, 2)$ (N choose 2) to $\sum_{i=1}^{k} C(c_i, 2)$ where c_i is the size of the i-th cluster and $\sum_{i=1}^{k} c_i = N$. With $N = 4441$, the number of exhaustive comparisons is around 9.9 million. Clustering on pathway names resulted in 1,040 clusters, of which 584 were singleton clusters. This reduced the number of pairwise comparisons to around 250,000. Clusters ranged from those consisting of pathways that share an identical name, such as the cluster for "Glycolysis," consisting of pathways from HumanCyc, Panther, Reactome, and SMPDB, to those that show a common theme, such as a cluster of 11 pathways with names such as "G2/M DNA damage checkpoint," "Mitotic G2-G2/M phases," and "response to G2/M transition DNA damage checkpoint signal." Some clusters contained unrelated pathways with shared words in their names, which may have clustered together because k was artificially lowered.

For pathway descriptions, exhaustive analysis yields around 3.5 million pairwise comparisons. Clustering on descriptions resulted in 1,026 clusters, of which 849 were singleton clusters, thereby reducing the number of pairwise comparisons to around 150,000. Inspection reveals clusters where pathways explore similar themes, such as a large cluster that includes pathways dealing with DNA synthesis and repair, and another that deals with pathways of fatty acid metabolism.

2.4 Using entity membership overlap to determine the validity of clustering results

We can test the validity of clustering on pathway names and descriptions using the independent measure of pathway entity membership overlap. Entity membership overlap can be represented using the Jaccard index, a measure of set similarity, defined for two sets (S_1 and S_2) as the ratio of the size of their intersection over the size of their union (equation (2)).

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{2}$$

For each pathway, its entity membership is represented as the set $P_i = \{e_1, e_2, ...e_n\}$. The Jaccard index is computed between a pair of pathways i and j as $J(P_i, P_j)$. From k-means clustering results on pathway names, we performed pairwise comparisons of all pathways within each cluster. The average pairwise within-cluster Jaccard index (APWJ) was 0.021. We generate an expected APWJ distribution using the following bootstrapping method. We randomly sample the data into clusters corresponding to the cluster sizes of our k-means output. Within these generated clusters, we compute the APWJ. We randomly sample 10,000 times to generate an expected distribution for APWJ. This expected distribution is Gaussianly distributed with mean 0.013 and sigma 7.0e-4 (Figure (1A)). The APWJ of our pathway name k-means clustering results falls more than 11 standard deviations away from the mean of this expected distribution. This indicates that our pathway clusters show significantly higher entity overlap than clusters generated

through random sampling of the data. In other words, pathway name is effective at blocking the data into content-similar groups.

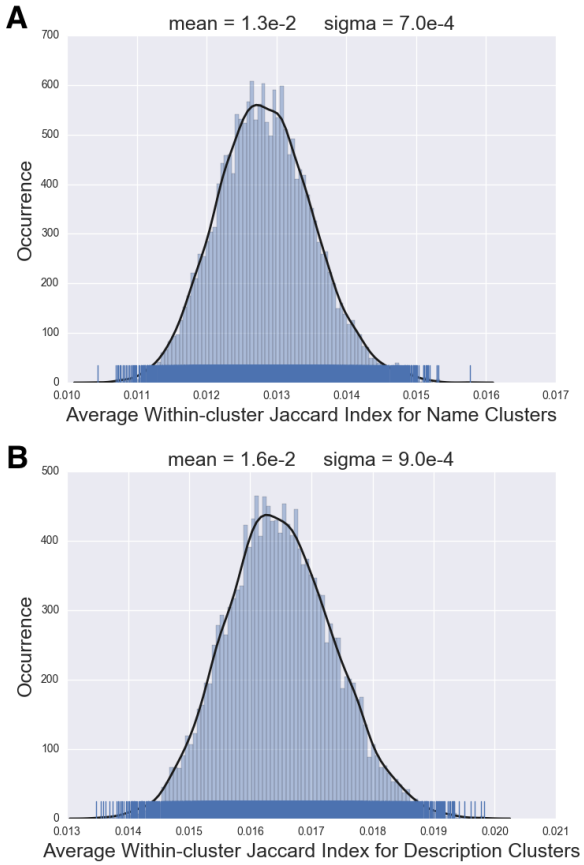

Fig. 1. The average within-cluster Jaccard index for 10,000 random clusterings of pathway names (A) and pathway descriptions (B).

The same procedure was followed for pathway descriptions. The bootstrapped APWJ distribution for our data had mean 0.016 and sigma 9.0e-4 (Figure (1B)). The APWJ for k-means clustering results was 0.027, more than 12 standard deviations away from the expected value for random clusters, indicating significant content overlap in our clusters compared to random.

2.5 Employing graph edit distance

Graph edit distance (GED) is a measure of similarity between two graphs. The measure is based on the number of node and edge insertions, substitutions, or deletions necessary to transform one graph into another. The measure is calculated by performing a global graph alignment between two graphs, and then calculating the number of transformations necessary.

In our case, we prematch entities between two pathway graph representations, which reduces the computational complexity of performing a global graph alignment. We used the GEDEVO software tool from the Computational Systems Biology Group of the Max Planck Institute for Informatics in Saarbruecken (Ibragimov *et al.*, 2013). This tool takes two graph representations as input, and calculates their GED along with a global graph alignment. The GED is normalized to between 0 and 1, with 1 indicating a perfect topological match between the two graphs. A higher GED score indicates improved topological matching, which by itself does not guarantee accurate entity matching. Two graphs with the same topology and completely different entity memberships will have a high GED score, so the GED is only useful in the context of high entity Jaccard index. GEDEVO also does not penalize having graphs of different sizes, and extra nodes and edges can remain unmatched.

Because there is no gold standard entity alignment among pathway representations, evaluations of the goodness of the graph alignments produced could only be done manually. The global alignment showed promise in cases where a good portion of all entities were prematched. Otherwise, the alignment did not offer usable entity alignments. For example, figure (2) shows an alignment between the Reactome pathway "Phenylalanine and tyrosine catabolism" and the HumanCyc pathway "tyrosine degradation I," two pathways that clustered together based on pathway name tf-idf scores. The entity memberships of the pathways show overlap (Jaccard index = 0.43). In this case, the entities in the HumanCyc pathway are actually a subset of the entities in the Reactome pathway, so we expect a potential part-of relationship between the two pathways. In figure (2) we observe this subset relationship. We recognize that visualization of complex pathway information is an open and largely unsolved problem outside of our scope; Figures (2) and (3) are hand-drawn.

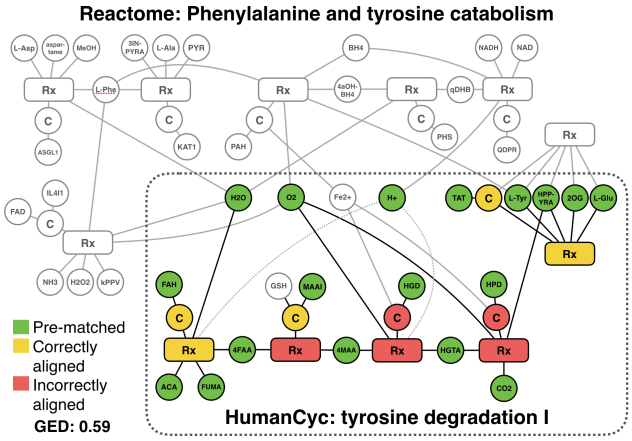

Fig. 2. Graph alignment results of two metabolic pathways with a subset relationship. Entities found in both pathways are outlined in black. Gray lines and circles are those relationships and entities found only in Reactome; dotted gray lines show relationships only found in HumanCyc. All reactions are labeled 'Rx'; all complexes are labeled 'C'; and abbreviations for proteins and molecules are taken from Reactome. Green entities are prematched between the two pathways on cross-reference identifiers. Yellow entities are correctly aligned by GEDEVO, and red entities are incorrectly aligned by GEDEVO and manually aligned by inspecting entity names and types.

Terminology in the Pathway Ontology can be used to describe the relationship between this pair of pathways. Both pathways are examples of PW:0001074, named "hydrophobic amino acid metabolic pathway." The HumanCyc pathway is an instance of the PW leaf node PW:0001284, or "tyrosine degradation pathway." The Reactome pathway consists of elements of both the PW leaf node "phenylalanine degradation pathway" (PW:0001288) and the PW

leaf node “tyrosine degradation pathway” (PW:0001284). Using the organization of the PW, we find that the Reactome pathway could potentially be broken down into two constituent parts, the conversion of phenylalanine into tyrosine, and the subsequent degradation of tyrosine.

A simple example of the benefit of a unifying ontology such as PW is that it would eliminate mismatches due to simple synonyms such as “catabolism” (used by Reactome) and “degradation” (used by HumanCyc). More interestingly, an ontology may allow for a more careful distinction of the relationships between pathways, for example, by drawing attention to the tyrosine degradation pathway being a subprocess of the phenylalanine degradation pathway. Although biologically, it may make sense to combine these into one pathway, as Reactome does, the duplication of the tyrosine degradation subprocess may be problematic for secondary use.

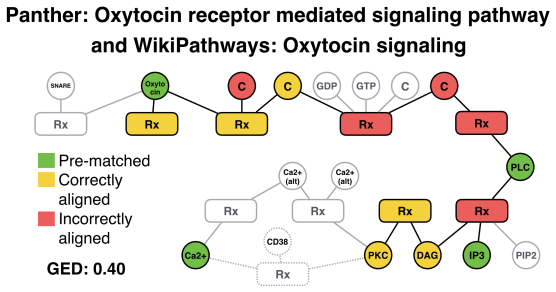

Fig. 3. Graph alignment of two signaling pathways describing the same process, where a majority of reactions are shown in both pathways. Entities found in both pathways are outlined in black. Gray lines and circles are those relationships and entities found only in Panther; dotted gray lines show relationships only found in WikiPathways. All reactions are labeled ‘Rx’; all complexes are labeled ’C’; all abbreviations for proteins and molecules are taken from Panther.

Figure (3) shows another example, this time of two overlapping pathways, Panther Pathway’s “Oxytocin receptor mediated signaling pathway” and WikiPathways’ “Oxytocin signaling.” The entity memberships of these two pathways show good overlap (Jaccard index = 0.40). Both of these pathways describe the same process, which is denoted by the PW leaf node PW:0000494, or “oxytocin signaling pathway.” There were a few prematched entities, and the graph alignment produced by GEDEVO was able to infer several additional entity matches between the two pathways. However, the performance seems less good compared to the previous example due to greater differences in representation between the two pathways.

3 DISCUSSION

Improving the way we discuss and measure similarity among pathway representations will have many repercussions for secondary use of pathway resources. Instead of using all pathways available for pathway analysis, we could eliminate redundant pathways and increase the power of our analysis results. We could also better organize pathways, thereby making clear where overlap and subprocess relationships occur. Thus, our work builds from the Pathway Ontology, and we aim to infer similarity and hierarchical relationships among pathways across resources.

From our clustering results, we can observe several different relationships between pathways. Some pathways describe similar processes, and show good entity overlap, such as the example given in figure (3). These overlapping pathways (A and B) are both instances of the same pathway class C (if the PW were adopted, the class would be “oxytocin signaling pathway”). Other pathways show a subset relationship as in figure (2), where one pathway can be described as a subprocess of the other pathway, exemplifying the part-of relationship. A third case is possible, but not illustrated, where one pathway is both a subset of another pathway and describes the same overall process. This could happen if modelers use different levels of granularity. The subset entities would then be interleaved through the larger pathway as opposed to forming a tightly connected subnetwork as in the subprocess case. This would still be an example of sibling relationships to some parent class, with the siblings differing in granularity. All three cases: overlap, subprocess, and granularity subset, can be discovered using a combination of entity membership and graph metrics.

Identifying these relationships is an important step to reducing redundancy in pathway data for secondary use. Overlapping pathways could be reduced to a single pathway representation. Pathways containing subprocesses could be modularized into several non-overlapping parts, or subpathways. For example, the Reactome pathway “Phenylalanine and tyrosine catabolism” from figure (2) could be broken down into two subprocesses, the conversion of phenylalanine into tyrosine (the gray entities from the figure), and the degradation of tyrosine (the colored entities from the figure). PW terms could be used to help identify these relationships between pathways. The PW is-a relationship describes both overlap and granularity subset relationships, and the PW part-of relationship describes subprocess relationships.

Our initial results are two-fold. First, we demonstrate that useful similarity information can be gathered from pathway names and descriptions. Second, we propose that further similarity information can be derived by combining a set of measures: pathway names, entity membership, and graph edit distance. We demonstrate this second point with some initial proof-of-concept examples (Figures (2) and (3)). We also advocate the use of an organizing ontology such as the Pathway Ontology to help identify pathway overlap and subprocess relationships.

3.1 Limitations & Future Work

There are several notable points of potential improvement in the procedures described in this paper. Pathway names and descriptions were used to cluster pathways using tf-idf scores. Stemming and lemmatization could be employed to derive better clustering results. Stemming and lemmatization is the process of reducing words to their base form; for example, metabolism, metabolic, and metabolite all share the same word stem. Prefix and suffix analysis can also help discover similar classes of words, especially chemical species, whose types can be derived from suffixes, like -oses (sugars) and -ases (proteins). Especially for pathway names, for which few words are present, stemming and suffixing could greatly improve our measure of name similarity. Additionally, tf-idf scores do not represent syntactic or semantic information, causing similar phrases with different key words to cluster together incorrectly. Using a greater variety of lexical features could help offset this weakness.

Another major challenge was entity normalization. In many cases, we discovered synonymous entities in two resources that did not share cross-reference identifiers. This could be helped by extending our entity normalization dictionary. More synonyms can be derived from third-party reference databases, although our usage of BridgeDB identifier mapping services already does this to some degree. We can also infer entity equivalence or synonymy using other information, such as the entity's name, or the reactions in which it participates. The calculation of a global graph alignment is one way to derive potential synonyms. The graph alignment algorithm employed by GEDEVO does not take into account features such as node name or type, which may help in identifying more synonyms. The inference of identifier synonymy through alternative means could also potentially be used to identify missing cross-reference identifiers in reference ontologies.

Lastly, we hope to provide a platform for exploring the overlaps among these pathways and to allow for the generation of pathway data sets with reduced redundancies among member pathways. Such an interface would allow the user to search for pathways from multiple sources, adjust the degree to which similar pathways should be merged into superpathways or broken down into non-overlapping subpathways, and export the resulting pathways for secondary use. For example, the user could generate unique gene sets for GSEA or other types of pathway-based enrichment analysis, or create novel explanatory pathways using the non-overlapping segments of existing pathways. A user interface could also leverage the work of the Pathway Ontology for organizing or annotating pathways from different databases. An evaluation of the usefulness and correctness of identified overlaps and similarities between pathways can be conducted formally through a qualitative assessment of biologists and their interactions with various merged pathway representations through this proposed platform.

3.2 Conclusion

Understanding similarities and redundancies among pathway representations is critical for improving the quality of secondary analyses performed using pathway resources. Associations among different pathways can be deduced by studying the features of each individual pathway, such as its name, description, entity membership, and topological structure. A hierarchical organizational structure such as the Pathway Ontology is a useful way to organize pathways. Here, we have shown that an analysis of a combination of features (names, entities and graph topology) could be used to infer similarity and relationship information between pathways. Our goal is to provide an umbrella organizational structure across multiple pathway databases that will make it easier for researchers to use pathways with appropriate content and granularity.

ACKNOWLEDGEMENTS

This study was supported in part by the National Library of Medicine (NLM) Training Grant T15LM007442.

REFERENCES

Apweiler, R., Bairoch, A., and et al, C. H. W. (2004). Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, **32**(Database issue), D115–119.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, **25**, 25–29.

Ballouz, S., Pavlidis, P., and Gillis, J. (2016). Using predictive specificity to determine when gene set analysis is biologically meaningful. *Nucleic Acids Research*, **45**.

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

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

Christen, P. (2011). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, **24**, 1537–1555.

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

Degtyarenko, K., de Matos, P., and et al, M. E. (2008). Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, **36**(Database issue), D344–350.

Demir, E., Cary, M., and et al, S. P. (2010). The biopax community standard for pathway data sharing. *Nature Biotechnology*, **28**(9), 935–42.

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., Liu, P., Gautam, B., Ly, S., Guo, A. C., Xia, J., Liang, Y., Shrivastava, S., and Wishart, D. S. (2010). Smpdb: The small molecule pathway database. *Nucleic Acids Research*, **38**, D480–D487.

Green, M. L. and Karp, P. D. (2006). The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Research*, **34**, 3687–3697.

Grego, T., Ferreira, J. D., Pesquita, C., Bastos, H., Vila Vicosa, D., Freire, J., and Couto, F. M. (2010). Chemical and metabolic pathway semantic similarity. *FC-DI - Technical Reports*.

Ibragimov, R., Malek, M., Guo, J., and Baumbach, J. (2013). Gedevo: An evolutionary graph edit distance algorithm for biological network alignment. *GCB*, **2013**, 68–79.

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.

Kanehisa, M. and Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27–30.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Comput Biol*, **8**(2), e1002375.

Kutmon, M., Riutta, A., and et al, N. N. (2016). Wikipathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res*, **44**(D1), D488–D494.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database (msigdb) hallmark gene set collection. *Cell Systems*, **1**, 417–425.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2011). Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research*, **39**, D52–57.

Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., and Kanehisa, M. (2013). Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions. *J. Chem. Inf. Model.*, **53**(3), 613–622.

Peregrn-Alvarez, J. M., Sanford, C., and Parkinson, J. (2009). The conservation and evolutionary modularity of metabolism. *Genome Biology*, **10**, R63.

Petri, V., Jayaraman, P., Tutaj, M., Hayman, G. T., Smith, J. R., Pons, J. D., Laulederkind, S. J. F., Lowry, T. F., Nigam, R., Wang, S.-J., Shimoyama, M., Dwinell, M. R., Munzenmaier, D. H., Worthey, E. W., and Jacob, H. J. (2014). The pathway ontology updates and applications. *Journal of Biomedical Semantics*, **5**.

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., and Karp, P. D. (2004). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, **6**(R2), 1–17.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). Article navigation pid: the pathway interaction database. *Nucleic Acids Research*, **37**, D674–D679.

Thomas, P. D., Campbell, M. J., and et al, A. K. (2003). Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, **13**, 2129–2141.

van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B. R., and Evelo, C. T. (2010). The bridgedb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, **11**.

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**, 414–422.

Wang, L. L., Gennari, J. H., and Abernethy, N. F. (2016). An analysis of differences in biological pathway resources. *Proceedings of the Joint International Conference on Biological Ontology and BioCreative*, **2016**.