Measuring Relatedness Between Scientific Entities in Annotation Datasets

Guillermo Palma
Universidad Simón Bolívar
Caracas, Venezuela
gpalma@ldc.usb.ve

Maria-Esther Vidal
Universidad Simón Bolívar
Caracas, Venezuela
mvidal@ldc.usb.ve

Eric Haag
University of Maryland
College Park, USA
ehaag@umd.edu

Louiqa Raschid
University of Maryland
College Park, USA
louiqa@umiacs.umd.edu

Andreas Thor
University of Leipzig
Germany
thor@informatik.uni-leipzig.de

ABSTRACT

Linked Open Data has made available a diversity of scientific collections where scientists have annotated entities in the datasets with controlled vocabulary terms (CV terms) from ontologies. These semantic annotations encode scientific knowledge which is captured in annotation datasets. One can mine these datasets to discover relationships and patterns between entities. Determining the relatedness (or similarity) between entities becomes a building block for graph pattern mining, e.g., identifying drug-drug relationships could depend on the similarity of the diseases (conditions) that are associated with each drug. Diverse similarity metrics have been proposed in the literature, e.g., *i*) string-similarity metrics; *ii*) path-similarity metrics; *iii*) topological-similarity metrics; all measure relatedness in a given taxonomy or ontology. In this paper, we consider a novel annotation similarity metric *AnnSim* that measures the relatedness between two entities in terms of the similarity of their annotations. We model *AnnSim* as a 1-to-1 maximal weighted bipartite match, and we exploit properties of existing solvers to provide an efficient solution. We empirically study the effectiveness of *AnnSim* on real-world datasets of genes and their GO annotations, clinical trials, and a human disease benchmark. Our results suggest that *AnnSim* can provide a deeper understanding of the relatedness of concepts and can provide an explanation of potential novel patterns.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Annotation datasets; topological distance; Annotation similarity; weighted bipartite match.

1. INTRODUCTION

Linked Open Data has made available a diversity of scientific collections. Scientists have annotated entities in the collections with controlled vocabulary terms (CV terms) from ontologies or taxonomies. Annotations describe properties of these concepts. For example, the functions of genes are described using Gene Ontology (GO) CV terms.

Annotations induce an annotation graph where nodes correspond to scientific concepts or ontology terms, and edges represent relationships between concepts. Figure 1 illustrates a portion of the Web of Data that can induce an annotation graph. Consider clinical trials linked to a set of diseases or conditions in the NCI Thesaurus (NCIt). Clinical trials from LinkedCT[1] are represented by blue ovals; they are associated with interventions or drugs (green rectangles) and diseases or conditions (pink rectangles). Both interventions and conditions are then annotated with terms from the NCI Thesaurus (red circles). Some annotations of a drug may correspond to terms in the NCIt that identify the drug, while others may correspond to the diseases or conditions that have been treated with this drug.

Knowledge captured within scientific collections, the annotations and the ontologies are rich and complex. For example, the NCI Thesaurus version 12.05d has 93,788 terms. The LinkedCT dataset *circa* September 2011 includes 142,207 interventions, 167,012 conditions or diseases, and 166,890 links to DBPedia, DrugBank, and Diseasome. Thus, the challenge is to explore these rich and complex datasets and to discover patterns, e.g., patterns of annotations across multiple disease conditions or multiple drug interventions. For gene functional annotations, patterns may involve cross-genome functional annotation, e.g., combining the GO functional annotations of two model organisms such as Arabidopsis thaliana (a plant) and Caenorhabditis elegans (a nematode or worm), to predict new functions.

As a first step to discovering complex patterns, we consider an important building block that determines the relatedness (or similarity) of a pair of scientific concepts, based on their annotations with respect to one or more ontologies. An ex-

[1] `http://linkedct.org/`

ample is identifying the relatedness or similarity of (drug, drug) pairs, based on the annotation evidence of diseases (conditions) from the NCIt. This can lead to discoveries of new targets for existing drugs, or it can predict potential side-effects of drugs.

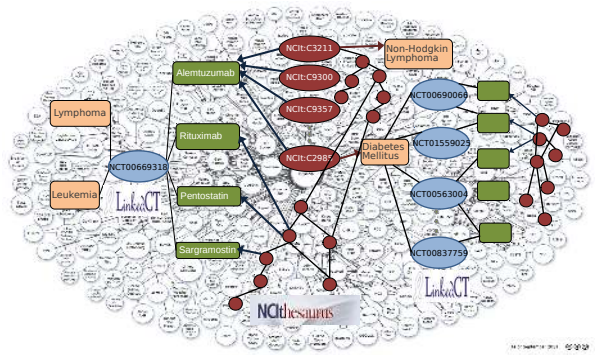


Figure 1: Annotation graph of Clinical Trials from LinkedCT (blue ovals). Interventions are green rectangles; conditions are pink rectangles and CV terms from the NCI Thesaurus are red ovals.

A broad variety of similarity metrics have been proposed in the literature [9, 12, 14, 15, 16, 17, 19, 20, 23, 27, 29]. Existing similarity metrics can be of diverse types: *i*) string-similarity metrics that measure similarity using (approximate) string matching functions (e.g., [11]); *ii*) path-similarity metrics such as *PathSim* and *HeteSim* that compute relatedness based on the paths that connect concepts in a graph (e.g., [23, 27]); and *iii*) topological-similarity metrics that measure relatedness in terms of the closeness of CV terms in a given taxonomy or ontology (e.g., [6, 15, 18]).

EXAMPLE 1.1. *Antineoplastic agents and monoclonal antibodies are two popular and independent intervention regimes that have been successfully applied to treat a large range of cancers. There are 12 drugs that fall within their intersection, and scientists are interested in studying the relationships between these drugs and the corresponding diseases. Consider the two drugs* `Brentuximab vedotin` *and* `Catumaxomab`. *Figure 2 represents a subgraph of the annotation graph of Figure 1. Each path between a pair of conditions, e.g.,* `Carcinoma` *and* `Anaplastic Large Cell Lymphoma` *through the NCI Thesaurus is identified using red circles which represent ontology terms from the NCIt. The count of red circles represents the length of a path. To simplify the figure, we only illustrate the paths from the term* `Carcinoma`.

In this paper, we propose a novel annotation similarity metric *AnnSim*, that measures the relatedness between two entities in terms of the similarity or relatedness of the sets of their annotations. *AnnSim* combines properties of path- and topological-based similarity metrics to decide the relatedness of two scientific concepts. To the best of our knowledge, our research is the first to consider both the shared annotations between a pairs of concepts, as well as the relatedness of the annotations (CV terms) within some ontology, to determine the resulting relatedness of the two concepts.

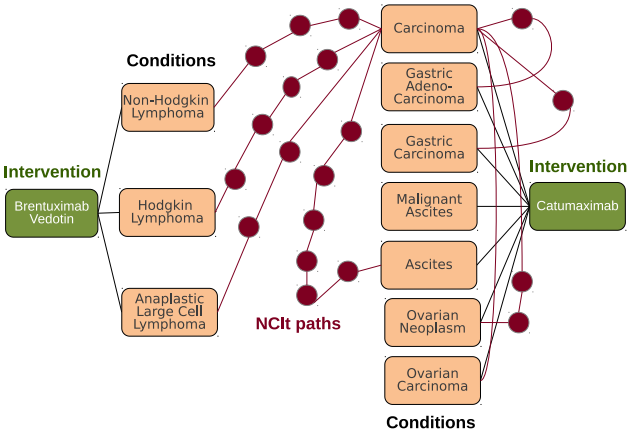


Figure 2: Annotation subgraph representing the annotations of `Brentuximab vedotin` and `Catumaxomab`. Interventions are green rectangles; conditions are pink rectangles; ontology terms in the NCIt are red ovals.

A naive implementation of *AnnSim* would require us to compute the topological similarity of all pairs of ontology terms, or the cartesian product, over the two sets of NCIt ontology terms that annotate each of the pair of concepts. However, many of these pairs of terms may be unrelated. We model *AnnSim* as a 1-to-1 maximal weighted bipartite matching, and we exploit properties of existing solvers to provide an efficient solution.

We empirically study the effectiveness of *AnnSim* on real-world datasets of genes and their GO annotations, evidence from clinical trials, and a well known human disease benchmark. We compare the quality of *AnnSim* with respect to existing similarity metrics including d_{tax} [6], d_{ps} [18], *HeteSim* [23].

We use the sequence-based similarity for genes based on the normalized Smith and Waterman scores [25] computed by BLAST[2], further normalized as suggested in [8], as ground truth for genes. We also use the evolutionary phylogenetic tree for a family of related genes as ground truth.

The contributions of this paper can be summarized as follows:

- The formalization of an annotation-based similarity metric *AnnSim* that defines the relatedness of two concepts in terms of the sets of their annotations. An implementation relies on an existing 1-to-1 maximal weighted bipartite graph matching solver.
- An empirical study that validates *AnnSim* using a variety of ground truth datasets including human curation as well as sequence based and phylogenetic evidence.
- Our results suggest that *AnnSim* can provide a deeper understanding of the relatedness of concepts, and in some cases it can also provide an explanation of patterns.

[2] `http://blast.ncbi.nlm.nih.gov/`

This paper is organized as follows: Section 2 summarizes related work. Section 3 gives the preliminary knowledge of this work and illustrates the performance of existing approaches in a real-world example. Section 4 presents our approach. Experimental results are reported in Section 5. Finally, we conclude in Section 6 with an outlook to future work.

2. RELATED WORK

A key element in finding patterns is identifying related concepts. Similarity metrics or distance metrics can be used to measure relatedness; we briefly describe some of the existing metrics.

The first class of metrics are string-similarity; they compare the names or labels of the concepts using string comparison functions based on edit distances or other functions that compare strings. The broadly used string distance metrics either reflect the number of edit operations that have to be performed on two strings to covert one in the other (e.g., the Levenstein distance), or they count the number and order of common characters between two strings (e.g., Jaro-Winkler [11]).

The next are path-similarity metrics that compute relatedness based on the paths that connect the concepts within some appropriate graph. Nodes in the paths can be all of the same abstract types (e.g., PathSim [27]) or they can be heterogeneous (*HeteSim* [23]). Furthermore, topological-similarity metrics extend the concept of path-similarity and they look at relationships within an ontology or taxonomy that is itself designed to capture relationships (e.g., nan[15], d_{ps} [18] and d_{tax}[6]).

Smith and Waterman [26] propose an algorithm to identify sequence alignment in sequences of nucleotides or amino-acids. BLAST[3] and FASTA[4] propose some restrictions to the sequence entries to speed up the alignment computation process, potentially at the cost of reducing quality. We use a normalized sequence based similarity score as ground truth.

Ontology matching (OM) tries to identify correspondences between semantically related entities of different ontologies [3, 24]. Advanced OM techniques utilize ontology structure [2]. Instance-based techniques (e.g., [28]) may also make use of annotations. OM and *AnnSim* have shared objectives and differences. The results of OM, i.e., the sets of correspondences (also called mappings or alignments) are primarily used for data integration, e.g., ontology merging or query rewriting. In contrast, *AnnSim* is interested in applying the metric to exploring patterns in families (graphs). More important, *AnnSim* has a focus on the entire annotation evidence. Thus, a mismatch of annotations must reduce *AnnSim*. Such nuances may not apply in general to OM.

The problem of 1-1 weighted maximal bipartite match has been tested on different domains, e.g, semantic equivalence between two sentences and measuring similarity between shapes for object recognition [4, 7, 22]. These approaches clearly show the benefits of solving this matching problem. *AnnSim* differs from the prior research in that we consider the relatedness of the sets of annotations. Further, we use an ontology structure to determine ontological relatedness. We extend the Dice coefficient to measure set agreement between the sets of annotations in the 1-1 weighted maximal bipartite match; the *AnnSim* score will be penalized if one of the concepts is associated with a large number of annotations while only a small number of annotations participate in the match.

Finally, we note that the value of any annotation-based similarity metric will naturally depend on the accuracy and comprehensiveness of the underlying annotation. Since AnnSim considers the graph structure of the ontology, it has the potential to be robust in the presence of missing or incomplete annotations, or similar yet not identical annotations.

3. SELECTED SIMILARITY METRICS

We present two taxonomic distance metrics from the literature: d_{tax} [6] and d_{ps} [18]. Both metrics define the distance of two nodes in terms of the depth of the nodes to the root of the ontology, and the distance to the their lowest common ancestor (LCA). These concepts are defined as follows:

Given a directed graph G, the *depth* of a vertex x in G is the length of the longest path from the root of G to x.
Given a directed graph G, the *lowest common ancestor* [5] of two vertices x and y, is the vertex of greatest depth in G that is an ancestor of both x and y.
Let $d(x, y)$ be the number of edges in the shortest path between vertices x and y in a given ontology. Also let $lca(x, y)$ be the lowest common ancestor of vertices x and y.

The intuition behind the d_{ps} metric is to capture the ability to represent the taxonomic distance between two vertices with respect to the depth of the common ancestor of these two vertices. Extending on this idea, d_{tax}[6] tries to assign low(er) values of taxonomic distance to pairs of vertices that are (1) at greater depth in the taxonomy and (2) are closer to their lowest common ancestor. A value close to 0.0 means that the two vertices are close to the leaves and both are close to their lowest common ancestor. A value close to 1.0 represents that both vertices are general or that the lowest common ancestor is close to the root of the taxonomy.

The distance metric d_{tax} is as follows where *root* is the root node in the ontology:

$$d_{tax}(x, y) = \frac{d(lca(x, y), x) + d(lca(x, y), y)}{d(root, x) + d(root, y)} \quad (1)$$

The distance metric d_{ps} is defined as follows:

$$d_{ps}(x, y) = 1 - \frac{d(root, lca(x, y))}{d(root, lca(x, y)) + d(lca(x, y), x) + d(lca(x, y), y)} \quad (2)$$

For the pair of drugs `Brentuximab vedotin` and `Catumaxomab`, we could locate these drugs within the NCIt and directly use either of the distance metrics and compute similarity values, $(1 - d_{tax})$ or $(1 - d_{ps})$. The similarity values are 0.60 and 0.43, respectively. However, unlike the proposed *AnnSim* metric, this similarity between the pair of drugs does not exploit the knowledge of their annotations,

[3] `http://blast.ncbi.nlm.nih.gov/`
[4] `http://www.ebi.ac.uk/Tools/sss/fasta/`

i.e., the diseases to which these drugs have been applied.

The metric *HeteSim* [23] defines the relatedness of object pairs in terms of the paths that connect the objects in a graph. Paths considered during the computation of this metric are type-path constrained, i.e., they must correspond to instances of a sequence of classes or types named relevance path. *HeteSim(s,t | P)* measures how likely s and t will meet at the same node when s follows along the path that respects the relevance path P and t goes against the path P. *HeteSim* is as follows:

DEFINITION 3.1 (*HeteSim* [23]). *Given two objects s and t ($s \in R_1$ and $t \in R_l$) and a relevance path $p = R_1 \circ R_2 \circ \cdots \circ R_l$,*

$$HeteSim(s,t \mid R_1 \circ R_2 \circ \cdots \circ R_l) = \frac{1}{\mid O(s \mid R_1) \mid\mid I(t \mid R_l)}$$

$$\sum_{i=1}^{|O(s|R_1)|} \sum_{j=1}^{|I(s|R_l)|} HeteSim(O_i(s \mid R_1), I_j(t \mid R_l) \mid R_2 \circ \cdots \circ R_{l-1})$$

where $O(s \mid R_i)$ and $I(s \mid R_j)$ correspond to the out-neighbors and in-neighbors of s based on relations R_i and R_j, respectively, and $O_t(s \mid R_i)$ and $I_k(s \mid R_j)$ represent the t-th and k-th elements in the out-neighbors and in-neighbors of s based on relations R_i and R_j, respectively.

For example, given the annotation graph of Figure 2, and paths of type (*Drug, NCIt, NCIt, Drug*), the value of *HeteSim*(Brentuximab vedotin, Catumaxomab) has a value of 0.0; this is because *HeteSim* only considers an exact match between the NCIt annotations of each drug. We note that *HeteSim* could be extended to further consider paths through the NCIt, i.e., these will be paths outside the annotation dataset.

4. ANNOTATION SIMILARITY METRIC FOR ANNOTATION GRAPHS

An annotation graph *G=(V,E)* is a particular graph comprised of two type of nodes in V: scientific concepts and terms from an ontology. Edges in G can be between scientific concepts and ontology terms.

Given two concepts c_1 and c_2 from an annotation graph *G=(V,E)*, we define an annotation similarity metric, *AnnSim*, based on their sets of annotations, A_1 and A_2, respectively. We assume that we know the pairwise similarity between elements of A_1 and elements of A_2, i.e., $sim(a_1, a_2) \in [0, 1]$ for all $a_1 \in A_1$ and $a_2 \in A_2$.
These relationships between terms in A_1 and A_2 can be represented as a weighted bipartite graph BG with two node sets A_1 and A_2.
An edge between $a_1 \in A_1$ and $a_2 \in A_2$ has a weight $sim(a_1, a_2)$, where $sim(a_1, a_2)$ is computed using a taxonomic distance metric.

The computation of *AnnSim* first requires building a bipartite graph BG with the links in the cartesian product between the set of annotations of two scientific terms, computing all pairwise similarities, and then determining the 1-to-1 maximal weighted bipartite graph match. The time complexity of this process is $O(n^4)$, where n is the number of nodes in the ontology; the cost of computing the topological similarity values of each one of the n^2 links is $O(n^2)$.

To achieve an efficient implementation of the *AnnSim* metric on BG, we reduce the bipartite graph BG to a *1-to-1 maximal weighted bipartite graph MWBG.*

DEFINITION 4.1. *[21] A* **1-to-1 maximal weighted bipartite graph matching** *MWBG=($A_1 \cup A_2$, WEr) for a weighted bipartite graph BG=($A_1 \cup A_2$, WE) is as follows:*

- *$WEr \in WE$, i.e., MWBG is a sub-graph of BG.*
- *the sum of the weights of the edges in WEr is maximized, i.e.,*

$$max \sum_{(a_1,a_2) \in WE} sim(a_1, a_2)$$

- *for each node in $A_1 \cup A_2$ there is only one incident edge in WEr, i.e.,*
  - $\sum_{i=1}^{|A_1|}(a_i, a_j) = 1, \forall j = 1 \cdots \mid A_2 \mid$
  - $\sum_{j=1}^{|A_2|}(a_i, a_j) = 1, \forall i = 1 \cdots \mid A_1 \mid$

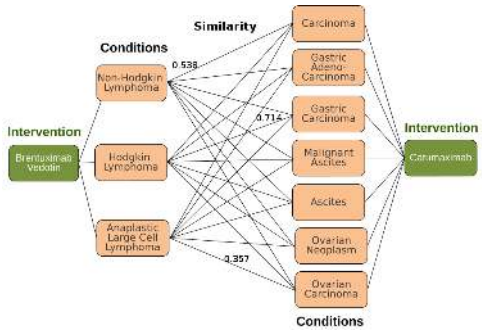
EXAMPLE 4.1. *Consider the two drugs* `Brentuximab vedotin` *and* `Catumaxomab`*. Figure 3 represents the 1-to-1 maximal weighted bipartite graph match produced by the BlossomIV solver [10].*

DEFINITION 4.2 (*AnnSim* ANNOTATION SIMILARITY). *Consider two concepts c_1 and c_2 annotated with the set of terms A_1 and A_2 in an annotation graph AG. Let BG=($A_1 \cup A_2$, WE) be a weighted bipartite graph for set of terms A_1 and A_2. Let MWBG=($A_1 \cup A_2$, WEr) be 1-to-1 maximal weighted bipartite graph matching for BG. The annotation similarity of c_1 and c_2 is defined as follows:*

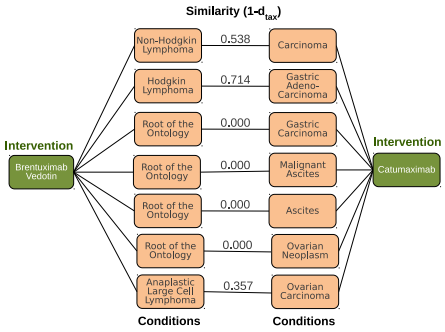
$$AnnSim(c_1, c_2) = \frac{2 \cdot \sum_{(a_1,a_2) \in WEr} sim(a_1, a_2)}{|A_1| + |A_2|}$$

The above definition is in the style of the well-known dice coefficient. The maximal similarity of 1.0 is achieved if and only if both annotation sets have the same cardinality ($|A_1| = |A_2|$) and all edge weights equal 1. Further, *AnnSim* penalizes (large) differences in the cardinality of A_1 and A_2. We apply an exact solution to the problem of computing the *1-to-1 maximal weighted bipartite graph MWBG* from a weighted bipartite graph BG using the BlossomIV solver [10]. Considering the 1-to-1 Maximal Weighted Bipartite Graph Matching for anticancer drugs Brentuximab vedotin and Catumaxomab in Figure 3. We can observe that *AnnSim* (Brentuximab vedotin,Catumaxomab) is 0.324 representing certain grade of similarity between these two drugs.

THEOREM 4.1 (PROPERTIES OF *AnnSim*). *Consider two concepts c_1 and c_2 annotated with the set A_1 and A_2 in agraph AG then:*



(a) Weighted bipartite graph for `Brentuximab vedotin` and `Catumaxomab`



(b) 1-to-1 Maximal weighted bipartite graph for `Brentuximab vedotin` and `Catumaxomab`

Figure 3: Bipartite graphs for drugs `Brentuximab vedotin` and `Catumaxomab`. For legibility only the value of the highest matching edges are shown in Figure 3(a).

- **Symmetry**: $AnnSim(c_1, c_2) = AnnSim(c_2, c_1)$.
- **Self-maximum**: $AnnSim(c_1, c_2) \in [0, 1]$.
- **Time complexity**: *polynomial in the size of AG.*

5. EXPERIMENTAL EVALUATION

While scientists use annotations widely, *AnnSim* is novel to our work. Thus, there is no prior *gold standard* that can be used to evaluate the quality of *AnnSim*. Further, there are few established ground truth datasets or alternate metrics to use as a baseline; thus, our evaluation is somewhat indirect out of necessity. We provide details of the datasets and our protocol to construct ground truth datasets for evaluation. We then present evaluation results.

5.1 Datasets and Evaluation Roadmap

Dataset 1: 30 pairs of diseases from the Mayo Clinic Benchmark; each pair is coded for similarity from 1.0 (least similar) to 4.0 (most similar). The coding was performed by 3 physicians (**Phy**) and 10 medical coders from the Mayo Clinic (**Cod**) [15, 17]. Diseases were annotated with NCI Thesaurus version 12.05d. Dataset 1 is used to compare $(1 - d_{tax})$ and $(1 - d_{ps})$ using SNOMED and MeSH.

Dataset 2: 12 anticancer drugs in the intersection of monoclonal antibodies and antineoplastic agents: Alemtuzumab, Bevacizumab, Brentuximab vedotin, Cetuximab, Catumaxomab, Edrecolomab, Gemtuzumab, Ipilimumab, Ofatumumab, Panitumumab, Rituximab, and Trastuzumab. The drugs were associated with conditions or diseases in clinical trials in LinkedCT *circa* September 2011 and each disease was linked to its corresponding term in the NCI Thesaurus version 12.05d. The number of annotations varies from 1 to 100+. Dataset 2 is used to compare *AnnSim* with $(1 - d_{tax})$, $(1 - d_{ps})$, and *HeteSim*. We recognize that *HeteSim* performs poorly because it is not designed to consider terms that are close to each other in the ontology as related. However, we use this baseline since it is the only metric that can consider paths between heterogeneous nodes.

Dataset 3: 10 families of Arabidopsis thaliana transporter genes [1]; 20 genes were selected for each family. 10 sets of 20 genes were also randomly chosen across all transporter Arabidopsis transporter gene families to create a control dataset in Dataset 3.

Dataset 4: Families of genes from Caenorhabditis elegans, e.g., `actins`. Genes in Datasets 3 and 4 were annotated with GO circa April 2013. Annotations were obtained from the portals TAIR [5] and WormBase[6].

For genes in Datasets 3 and 4, we compare *AnnSim* with *SeqSim*, a sequence based similarity for genes based on the normalized Smith and Waterman scores [25] computed by BLAST[7], further normalized as suggested in [8]:

$$SeqSim(g, g') = \frac{SW(g, g')}{\sqrt{SW(g', g')}\sqrt{SW(g, g)}}$$

where, g and g' are genes and SW is the pairwise or reflexive Smith and Waterman score. We also construct the phylogenetic (evolutionary) tree for the Arabidopsis families of Dataset 3 and compute d_{tax} and d_{ps} against these trees.

5.2 Effectiveness in Dataset 1

The goal of the experiment is to tune the performance of $(1 - d_{tax})$ and $(1 - d_{ps})$ with respect to multiple ontologies. This will reveal if *AnnSim* scores will be stable across different taxonomic metrics and ontologies.

We annotated the 30 diseases of Dataset 1 with their corresponding terms in SNOMED, MeSH and the NCI Thesaurus. The scores determined by $(1-d_{tax})$ and $(1-d_{ps})$ are compared to the human ground truth evaluation of physicians and coders. Table 1 reports on this comparison. Additionally, Table 2 reports on the Normalized Discounted Cumulative Gain [13] (nDCG) between the ranking of the results using $(1 - d_{tax})$ and $(1 - d_{ps})$, and the ground truth from a physician panel or a coder panel. The nDCG correlations take values between 0.0 and 1.0, where a value close to 1.0 represents a high correlation of the ranking induced by the similarity metric and the one in the ground truth.

Given the order of the pairs of diseases induced by the values

[5] `http://www.arabidopsis.org/`
[6] `http://www.wormbase.org/#012-3-6`
[7] `http://blast.ncbi.nlm.nih.gov/`

Table 1: Similarity Dataset 1: $(1-d_{tax})$ and $(1-d_{ps})$ for SNOMED, MeSH, and NCIt. Empty Cells(-) represent terms that do not appear in the ontology. Values highlighted in bold show high correlation between the relevance given by the physician, coder and the metrics.

Medical Terms	Phy	Cod	SNOMED		MeSH		NCIt	
			$1-d_{tax}$	$1-d_{ps}$	$1-d_{tax}$	$1-d_{ps}$	$1-d_{tax}$	$1-d_{ps}$
Renal Insufficiency - Kidney Failure	**4.00**	**4.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Heart - Myocardium	**3.30**	3.00	**0.77**	0.64	**0.80**	0.67	0.20	0.11
Stroke - Infarction	**3.00**	2.80	0.31	0.31	**0.80**	0.67	**0.87**	**0.78**
Abortion - Miscarriage	3.00	**3.30**	**0.89**	**0.80**	0.00	0.00	**0.92**	**0.86**
Delusions - Schizophrenia	3.00	2.20	0.00	0.00	0.00	0.00	0.80	0.67
Congestive heart failure - Pulmonary edema	3.00	**1.40**	0.50	0.46	0.00	0.00	0.59	**0.42**
Metastasis - Adenocarcinoma	**2.70**	**1.80**	**0.83**	**0.71**	**0.25**	0.14	0.00	0.00
Calcification Stenosis	2.70	**2.00**	**0.55**	0.38	0.00	0.00	0.40	0.25
Diarrhea - Stomach cramps	2.30	**1.30**	0.29	**0.17**	0.75	0.63	0.42	0.30
Mitral Stenosis - Atrial Fibrillation	**2.30**	1.30	**0.63**	0.46	0.50	0.33	0.53	0.36
Chronic obstructive pulmonary disease - Lung infiltrates	**2.30**	1.90	0.70	**0.63**	-	-	0.13	0.07
Rheumatoid Arthritis - Lupus	**2.00**	**1.00**	**0.50**	0.33	**0.00**	0.11	0.86	0.75
Brain tumor - Intracranial hemorrhage	2.00	**1.30**	0.63	0.57	0.63	0.50	**0.17**	0.09
Carpal Tunnel Syndrome - Osteoarthritis	2.00	**1.00**	0.33	0.33	**0.00**	**0.00**	0.33	0.20
Diabetes Mellitus - Hypertension	**2.00**	**1.00**	0.64	**0.50**	**0.00**	**0.00**	0.17	0.09
Acne - Syringe	2.00	**1.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Antibiotic - Allergy	1.70	**1.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Cortisone - Total knee replacement	1.70	**1.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Pulmonary Embolism - Myocardial Infarction	**1.70**	1.20	0.36	**0.42**	0.29	0.29	0.63	**0.46**
Pulmonary Fibrosis - Lung Cancer	1.70	1.40	0.75	0.63	0.67	0.50	0.60	0.50
Cholangiocarcinoma - Colonoscopy	1.30	**1.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Lymphoid hyperplasia - Laryngeal cancer	1.30	**1.00**	0.43	0.33	**0.00**	**0.00**	0.36	0.22
Multiple Sclerosis - Psychosis	**1.00**	**1.00**	0.44	0.29	**0.00**	**0.00**	0.33	0.20
Appendicitis - Osteoporosis	**1.00**	**1.00**	0.31	0.31	**0.00**	**0.00**	0.50	0.36
Rectal polyp - Aorta	**1.00**	**1.00**	**0.00**	**0.00**	-	-	**0.00**	**0.00**
Xerostomia - Liver Cirrhosis, Alcoholic	**1.00**	**1.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.14	0.08
Peptic Ulcer - Myopia	**1.00**	**1.00**	0.23	0.29	**0.00**	**0.00**	0.15	0.08
Depression- Cellulitis	**1.00**	**1.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.31	0.18
Varicose vein - Entire knee meniscus	**1.00**	**1.00**	0.13	0.07	-	-	**0.00**	**0.00**
Hyperlipidemia - Metastasis	**1.00**	**1.00**	0.33	0.20	**0.00**	**0.00**	**0.00**	**0.00**

of $(1-d_{tax})$ and $(1-d_{ps})$, a high value of nDCG of a given pair highly ranked by the physicians (or coders) indicates that the pair appears at the top of the ranking list. A low value of nDCG reflects that the relevant pair appears at the bottom of the ranking list.

We can observe that both $(1-d_{tax})$ and $(1-d_{ps})$ have similar values of nDCG across SNOMED, Mesh and NCI Thesaurus, for both physicians and coders. This reveals that both metrics are successful at computing high values of similarity for the pairs that were also ranked highly by the physicians and coders. These values also suggest that both metrics have similar performance.

To summarize, the two metrics to compare taxonomic relatedness perform well across multiple ontologies, and their performance is matched.

5.3 Effectiveness in Dataset 2

Our objective is to compare the performance of *AnnSim* with existing similarity metrics to determine relatedness between the drugs in this family. We consider both topological metrics $(1-d_{tax})$, $(1-d_{ps})$ and *HeteSim*. Intuitively, *HeteSim* would detect that two drugs are similar if they have many (identical) diseases in common. *HeteSim* will perform poorly when drugs do not treat identical diseases. In contrast, *AnnSim* also considers diseases that are not identical but are similar based on the topology of the NCI Thesaurus. Finally, $(1-d_{tax})$ and $(1-d_{ps})$ only consider the topology of the drug terms in the NCI Thesaurus and will ignore the annotation evidence.

Table 3 reports on the values of these four similarity metrics when `Alemtuzumab` is compared to the eleven other drugs in the dataset. We can observe that *HeteSim* consistently assigns very low values of similarity. Although all these drugs are used to treat different types of cancers, Alemtuzumab shares only a small number of identical diseases with the rest of the 11 drugs and this confuses *HeteSim*.

AnnSim, however, assigns higher values because is able to detect that many of the diseases treated with Alemtuzumab share similar topological properties in NCIt with the diseases treated by the rest of the drugs.

What is notable is that the taxonomic metrics $(1-d_{tax})$ and $(1-d_{ps})$ only consider the topology of the drug terms in the NCIt and they ignore the annotation evidence. Thus, they return uniformly high similarity scores. The column AnnotCount of Table 4 summarizes the number of annotations for each drug; it is clear that there is a wide variation in the diseases that are treated by these drugs. Hence, the inability to exploit the annotation evidence does not allow the taxonomic metrics to differentiate between these drugs.

Table 4 summarizes the pairwise scores for the four metrics for each drug, compared to the other 11 drugs. For each drug, the score is used to rank the other 11 drugs. Then $SRank_1$ is the Spearman's correlation for *AnnSim* and $(1-d_{tax})$ and $SRank_2$ is the correlation for *AnnSim* and $(1-d_{ps})$. We observe that *HeteSim* consistently assigns very low values of similarity. *AnnSim* again assigns higher values overall. Values of $SRank_1$ and $SRank_2$ are higher than 0.5,

Table 2: Normalized Discounted Cumulative Gain (nDCG) of $(1 - d_{tax})$ and $(1 - d_{ps})$

Metric	SNOMED		MeSH		NCI	
	Physician	**Coder**	**Physician**	**Coder**	**Physician**	**Coder**
$1 - d_{tax}$	0.837	0.961	0.977	0.957	0.959	0.959
$1 - d_{ps}$	0.966	0.963	0.976	0.987	0.959	0.959

Table 3: Pairwise comparison of Alemtuzumab with the rest of the 11 drugs. *HeteSim* **assumes perfect matching between annotations and assigns low similarity values.**

Pair drug	AnnSim	$1 - d_{tax}$	$1 - d_{ps}$	HeteSim
Alemtuzumab - Bevacizumab	0.263	0.670	0.500	0.001
Alemtuzumab - Brentuximab vedotin	0.140	0.364	0.222	0.000
Alemtuzumab - Catumaxomab	0.199	0.364	0.222	0.000
Alemtuzumab - Cetuximab	0.359	0.727	0.571	0.000
Alemtuzumab - Edrecolomab	0.037	0.727	0.571	0.000
Alemtuzumab - Gemtuzumab	0.046	0.500	0.333	0.000
Alemtuzumab - Ipilimumab	0.482	0.727	0.571	0.005
Alemtuzumab - Ofatumumab	0.468	0.727	0.571	0.002
Alemtuzumab - Panitumumab	0.422	0.727	0.571	0.000
Alemtuzumab - Rituximab	0.409	0.727	0.571	0.002
Alemtuzumab - Trastuzumab	0.319	0.727	0.571	0.000
Average	**0.286**	**0.635**	**0.479**	**0.001**

suggesting that *the annotation evidence is consistent with the topological relationships of the drugs* in the NCIt.

We note on a couple of outlier cases. Both `Edrecolomab` and `Gemtuzumab` have a single annotation, `Colorectal Carcinoma` and `Acute Myeloid Leukemia`, respectively. While these diseases are different, the drugs have very similar and low values for *AnnSim*. We note that the drugs have high values for the taxonomic metrics; e.g., (1 - d_{tax}(*Colorectal Carcinoma*, *Acute Myeloid Leukemia*)) is equal to 0.714. Since d_{tax} meets the triangle inequality property [6], any disease that is similar to one disease will also be similar to the other. We further note that the $SRank_1$ and $SRank_2$ have a negative score for `Edrecolomab` but the score is closer to 0.5 for `Gemtuzumab`. This reflects that further work is needed to tune these metrics to consider outliers.

Details of the 12 drugs in Dataset 2 as well as their annotations and pairwise values of *AnnSim* can be found at `http://pang.umiacs.umd.edu/AEDdemo.html`.

5.4 Effectiveness in Dataset 3

The goal of the experiment is to compare the performance of *AnnSim* with both *SeqSim* and topological similarity on the phylogenetic (evolutionary) tree. We consider both family and random (control) datasets. We compute pairwise values for *AnnSim*, as well as *SeqSim* based on the normalized Smith and Waterman scores [25] computed by BLAST[8], further normalized as suggested in [8]. For each gene, we rank the 19 other genes in the family/set using *AnnSim* and *SeqSim*. We then calculate the Spearman's Rho correlation coefficient $SRank_3$ for the 2 rankings. We expect to have a greater measure of relatedness in families over the random datasets.

We also determine the phylogenetic (evolutionary) tree for each of the families. We then compute d_{tax} and d_{ps} for these trees, and the Spearman's correlation, $SRank_4$ and $SRank_5$ with *AnnSim*. We do not compute these metrics for the control genes since they are not closely related using phylogeny.

[8]`http://blast.ncbi.nlm.nih.gov/`

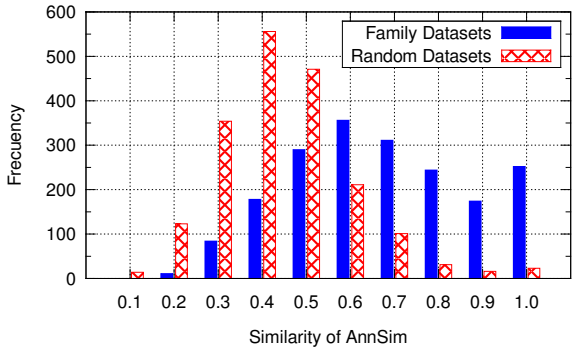


Figure 4: Annotation Similarity Distribution for Family and Control (Random) Datasets

Figure 4 shows the annotation similarity distributions for families and control (random) groups. Table 5 reports on the (average, standard deviation) for *SeqSim*, *AnnSim* and correlation $SRank_3$ for the families and the control (random) groups.

Our evaluation **strongly confirms** that *AnnSim* is a useful metric. *AnnSim* assigns higher values to genes in families compared to genes in the control datasets as can be seen from the distribution of Figure 4. From Table 5, the average value for families is 0.62, while the same value for the control datasets is 0.40. The average Spearman's correlation coefficient $SRank_3$ comparing the rank based on sequence similarity and *AnnSim* is 0.20 for families and 0.01 for the control dataset. All of these values were significant at the 95% confidence interval.

Figure 5 shows the distribution of annotation similarity at the disaggregated level for 4 of the 10 Arabidopsis transporter families. We see that *cpa2* and *mate* and *mfs* have the highest values of annotation similarity, while *vic* and *p-atpase* show the lowest annotation similarity.

Table 4: Average similarity and Standard Deviation (avg; std) when each is compared with 11 other drugs (Antineoplastic Agents and Monoclonal Antibodies)

Drug	Annot Count	$AnnSim$	$(1 - d_{tax})$	$(1 - d_{ps})$	$HeteSim$	$SRank_1$	$SRank_2$
Alemtuzumab	39	(0.286; 0.161)	(0.635; 0.150)	(0.479; 0.146)	(0.001; 0.002)	0.625	0.625
Bevacizumab	136	(0.206; 0.173)	(0.636; 0.152)	(0.479; 0.146)	(0.002; 0.002)	0.505	0.543
Brentuximab vedotin	3	(0.206; 0.125)	(0.433; 0.093)	(0.284; 0.091)	(0.002; 0.007)	0.752	0.752
Catumaxomab	7	(0.244; 0.106)	(0.416; 0.066)	(0.269; 0.061)	(0.002; 0.003)	0.348	0.339
Cetuximab	50	(**0.303**; 0.189)	(0.691; 0.163)	(0.547; 0.171)	(0.003; 0.004)	0.523	0.507
Edrecolomab	1	(**0.157**; 0.211)	(0.691; 0.162)	(0.547; 0.171)	(0.004; 0.014)	-0.318	-0.318
Gemtuzumab	1	(**0.157**; 0.219)	(0.539; 0.045)	(0.375; 0.046)	(0.000 0.000)	0.511	0.466
Ipilimumab	22	(0.363; 0.208)	(0.691; 0.163)	(0.547; 0.171)	(0.004; 0.003)	0.502	0.502
Ofatumumab	18	(**0.302**; 0.159)	(0.692; 0.162)	(0.547; 0.171)	(0.003; 0.007)	0.382	0.411
Panitumumab	22	(0.358; 0.212)	(0.692; 0.162)	(0.547; 0.171)	(0.007; 0.014)	0.514	0.525
Rituximab	100	(0.222; 0.169)	(0.691; 0.163)	(0.547; 0.171)	(0.001; 0.001)	0.311	0.311
Trastuzumab	18	(**0.304**; 0.175)	(0.692; 0.162)	(0.547; 0.171)	(0.002; 0.003)	0.350	0.364
Average	**34.750**	**(0.259; 0.176)**	**(0.625; 0.137)**	**(0.476; 0.141)**	**(0.003; 0.005)**	**0.417**	**0.419**

.

Table 5: Average and Standard Deviation (avg; std) of Sequence Similarity (*SeqSim*), AnnSim and Spearman's Rank ($SRank_3$) for Families and Control (Random) Dataset 3.

(a) Dataset 3: Families

Familes	*SeqSim*	*AnnSim*	$SRank_3$
aaap	(0.093; 0.023)	(0.654; 0.102)	0.025
abc	(0.088; 0.026)	(0.573; 0,110)	0.222
cpa2	(0.051; 0.015)	(0.788; 0.077)	0.220
dmt	(0.045; 0.013)	(0.542; 0.116)	0.146
f-atpase	(0.091; 0.023)	(0.540; 0.060)	0.231
mate	(0.093; 0.020)	(0.857; 0.071)	0.134
mfs	(0.074; 0.022)	(0.724; 0.090)	0.016
mip	(0.097; 0.044)	(0.607; 0.044)	0.234
p-atpase	(0.142; 0.046)	(0.502; 0.064)	0.322
vic	(0.075; 0.017)	(0.462; 0.064)	0.392
Average	(0.085; 0.025)	(0.625; 0.080)	0.194

(b) Dataset 3: Control

Random	*SeqSim*	*AnnSim*	$SRank_3$
#1	(0.044; 0.012)	(0.347;0.066)	0.061
#2	(0.041; 0.012)	(0.418; 0.083)	0.004
#3	(0.042; 0.010)	(0.367; 0.081)	-0.139
#4	(0.050; 0.008)	(0.418; 0.065)	0.003
#5	(0.052; 0.017)	(0.450; 0.089)	0.069
#6	(0.048; 0.018)	(0.358; 0.066)	-0.039
#7	(0.040; 0.013)	(0.418; 0.095)	0.089
#8	(0.043; 0.012)	(0.378; 0.075)	-0.053
#9	(0.047; 0.013)	(0.427; 0.079)	0.061
#10	(0.033; 0.011)	(0.408; 0.062)	0.084
Average	(0.044; 0.013)	(0.399; 0.076)	0.014

Table 6 compares *AnnSim* with $(1 - d_{tax})$ and $(1 - d_{ps})$ computed over the phylogenetic tree for each of the families. We observe that the values of $(1 - d_{tax})$ and $(1 - d_{ps})$ are quite low. This is a results of the binary topology of the phylogenetic trees. Note that in binary trees only one pair of terms appear in the same branch. Thus, only one of the genes in the family is considered to be similar and the other genes will have very low similarity. Unfortunately, both metrics suffer from this behavior.

We further note that the Spearman's correlations $SRank_4$ and $SRank_5$ with $(1 - d_{tax})$ and $(1 - d_{ps})$ are identical, even though the values for $(1 - d_{tax})$ and $(1 - d_{ps})$ appear to be different. This too is due to the behavior of these two metrics in the binary phylogenetic tree.

In summary, our results show that *AnnSim* validates the hypothesis that genes from families will have a higher distribution of similar annotations, and will have higher correlation with sequence based *SeqSim*, compared to control datasets.

Table 7: Pairs of (average, standard deviation) of *SeqSim* normalized sequence similarity, *AnnSim* annotation similarity for actins; Spearman's Rank $SRank_6$ compares rankings of *SeqSim* and *AnnSim*.

Family	*SeqSim*	*AnnSim*	$SRank_6$
Actins	(0.185; 0.093)	(0.520; 0.132)	0.078

5.5 Effectiveness in Dataset 4

We use a case study of manual curation and the phylogenetic evidence to explore *AnnSim* on a family of Caenorhabditis elegans genes. Actins are long cytoplasmic proteins that are slow-evolving, and present in all eukaryotes. For animals (and some plants) they are usually present in a small family of 5-10 copies. Their function is to allow the cell to control its shape, usually working together with an unrelated protein called myosin. Muscle cells are mostly filled with dense arrays of actin and myosin, but all cells have at least a small amount of actin.

Figure 6 shows the distribution of pair-wise annotation similarity for the actins. Figure 7 reports on the (average, standard deviation) for normalized sequence similarity (*SeqSim*) and annotation similarity (*AnnSim*) for actins. We also report on the the Spearman's correlation coefficient between the two rankings induced by these metrics, ($SRank_6$).

We observe that actins have high values of annotation similarity; this confirms that *AnnSim* assigns higher values to genes in functionally related families. However, we note that the *SeqSim* values appear to be very low. This is explained by the results of the clustering to be described next. All values were significant at the 95% confidence interval.

Figure 7 illustrates the clustering of actins based on the values of *AnnSim*; the threshold for edges was 0.5 to appear in the figure. The clustering and resulting communities using *AnnSim* is consistent with the phylogenetic evidence. Based on the phylogenetic evidence, act-1, act-2, act-4 and act-5 are closely related to each other. act-3 is a redundant variant splice form. The arx-n genes are *actin related* genes. Of the 7 genes, only arx-1 and arx-2 can be aligned with the act-n genes. The other 5 genes arx-3 through arx-7 do not share sequence similarity with the act-n genes. Finally, ani-1, ani-2 and ani-3 are *actin binding* genes. None of them

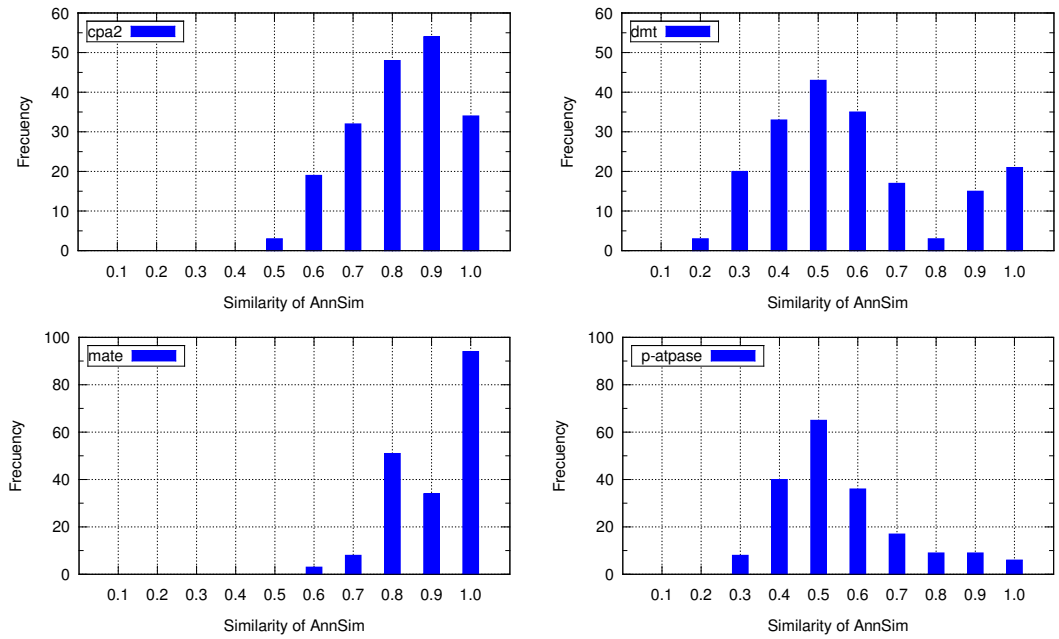


Figure 5: Distribution of Annotation Similarity for 4 Arabidopsis Families

Table 6: Average similarity 10 Arabidopsis thaliana families;$(1 - d_{tax})$ and $(1 - d_{ps})$ where computed on the phylogenetic tree of the genes that comprise each family in Dataset 3.

Family	$(1 - d_{tax})$	$(1 - d_{ps})$	**AnnSim**	$SRank_4$	$SRank_5$
abc	(0.135; 0.038)	(0.091; 0.027)	(0.573; 0.110)	0.399	0.399
cpa2	(0.154; 0.042)	(0.108; 0.031)	(0.788; 0.077)	0.307	0.307
dmt	(0.156; 0.052)	(0.105; 0.036)	(0.542; 0.116)	0.491	0.491
f-atpase	(0.174; 0.060)	(0.121; 0.045)	(0.540; 0.060)	0.567	0.567
mate	(0.147; 0.046)	(0.099; 0.033)	(0.857; 0.071)	0.185	0.185
mfs	(0.167; 0.055)	(0.114; 0.038)	(0.724; 0.090)	0.069	0.069
mip	(0.186; 0.070)	(0.126; 0.049)	(0.607; 0.044)	0.407	0.407
p-atpase	(0.206; 0.064)	(0.144; 0.049)	(0.502; 0.064)	0.387	0.387
vic	(0.168; 0.056)	(0.116; 0.039)	(0.462; 0.064)	0.327	0.327
Average	**(0.166; 0.054)**	**(0.114; 0.039)**	**(0.622; 0.077)**	**0.349**	**0.349**

share sequence similarity with the act-n genes.

If we examine Figure 7, we observe that the *actin binding* ani-1 through ani-3 are not connected to the other genes. The 5 actins form a community. They are connected to arx-2 which is one of the 2 *actin related* genes that share sequence similarity with the actins. The 7 *actin related* arx-n genes also form a community.

To summarize, Dataset 4 is a **strong validation** of *AnnSim* using the phylogenetic evidence. The results from Datasets 2 and 4 suggest that *AnnSim* can also be used to explore and explain deeper and more nuanced relationships among genes or drug families.

6. CONCLUSIONS AND FUTURE WORK

We have proposed an annotation similarity metric *AnnSim* to determine the relatedness of two concepts based on the topological similarity of their sets of annotations. *AnnSim* is defined as a *1-to-1 maximal weighted bipartite graph match*. We have performed an extensive evaluation using multiple datasets and ground truth. We note that the *1-to-1 maximal weighted bipartite graph match* has many limitations since it ignores unmatched terms and does not consider groups of matching terms. In future work, we will explore extensions to *n-m weighted bipartite graphs*.

7. REFERENCES

[1] Classified transporter families in arabidopsis. `http://www.clfs.umd.edu/CBMG/faculty/sze/lab/AtTransporters.html`.

[2] D. Aumueller, H. H. Do, S. Massmann, and E. Rahm. Schema and ontology matching with coma++. In *SIGMOD Conference*, pages 906–908, 2005.

[3] Z. Bellahsene, A. Bonifati, and E. Rahm, editors. *Schema Matching and Mapping*. Springer, 2011.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509–522, 2002.

[5] M. A. Bender, M. Farach-Colton, G. Pemmasani, S. Skiena, and P. Sumazin. Lowest common ancestors in trees and directed acyclic graphs. *Journal of Algorithms*, 57(2):75–94, 2005.

[6] J. Benik, C. Chang, L. Raschid, M. E. Vidal, G. Palma, and A. Thor. Finding cross genome patterns in annotation graphs. In *Proceedings of Data Integration in the Life Sciences (DILS)*, 2012.

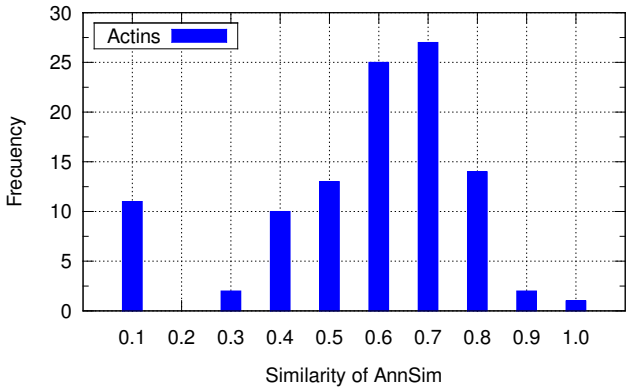


Figure 6: Annotation Similarity Distribution for Actins

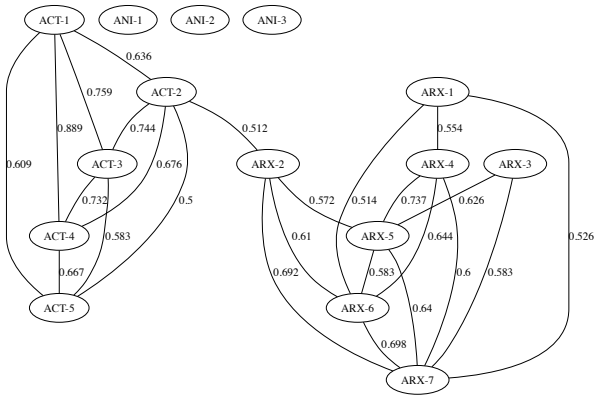


Figure 7: Clustering of the genes of Actins based on the values of AnnSim. An edge between two nodes represents that the AnnSim similarity of these two nodes is greater than 0.5.

[7] S. Bhagwani, S. Satapathy, and H. Karnick. Semantic textual similarity using maximal weighted bipartite graph matching. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 579–585. Association for Computational Linguistics, 2012.

[8] K. Bleakley and Y. Yamanishi. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, 2009.

[9] C. Chen, S. Hsieh, Y. Weng, W. Chang, and F. Lai. Semantic similarity measure in biomedical domain leverage web search engine. *Proc.IEEE Eng Med Biol Soc*, pages 4436–4439, 2010.

[10] W. Cook and A. Rohe. Blossom iv: Code for minimum weight perfect matchings. `http://www2.isye.gatech.edu/~wcook/software.html`.

[11] M. A. Jaro. Probabilistic linkage of large public health data files. *Statistics in Medicine*, pages 491–498, 1995.

[12] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008, 1997.

[13] J. K. Kalervo Jarvelin. Cumulated gain-based evaluation of ir techniques. *JACM Transactions on Information Systems*, 20(4):422■446, 2002.

[14] D. Lin. An information-theoretic definition of similarity. In *ICML*, pages 296–304, 1998.

[15] B. McInnes, T. Pedersen, and S. Pakhomov. Umls-interface and umls-similarity : Open source software for measuring paths and semantic similarity. *Proceedings of the AMIA Symposium*, pages 431–435, 2009.

[16] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, and G. Melton. Semantic similarity and relatedness between clinical terms: An experimental study. *Proceedings of the AMIA Symposium*, pages 572–576, 2010.

[17] T. Pedersen, S. Pakhomov, S. Patwardhan, and C. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007.

[18] V. Pekar and S. Staab. Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In *COLING*, 2002.

[19] C. Pesquita, D. Faria, A. Falcão, P. Lord, and F. Couto. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7):e1000443, 2009.

[20] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995.

[21] J. Schwartz, A. Steger, and A. Weißl. Fast algorithms for weighted bipartite matching. In *WEA*, pages 476–487, 2005.

[22] Y. Shavitt, E. Weinsberg, and U. Weinsberg. Estimating peer similarity using distance of shared files. In *International workshop on peer-to-peer systems (IPTPS)*, volume 104, 2010.

[23] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu. Relevance search in heterogeneous networks. In *EDBT*, pages 180–191, 2012.

[24] P. Shvaiko and J. Euzenat. Ontology matching: State of the art and future challenges. *IEEE Trans. Knowl. Data Eng.*, 25(1):158–176, 2013.

[25] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981.

[26] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, pages 195–197, 1981.

[27] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003, 2011.

[28] A. Thor, T. Kirsten, and E. Rahm. Instance-based matching of hierarchical ontologies. In *BTW*, pages 436–448, 2007.

[29] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.