InterOnto – Ranking Inter-Ontology Links

Silke Trißl[1], Philipp Hussels[2], and Ulf Leser[2]

[1] Leibniz Institute for Farm Animal Biology,
Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany
trissl@fbn-dummerstorf.de
[2] Humboldt-Universität zu Berlin, Institute for Computer Science,
Unter den Linden 6, 10099 Berlin, Germany
{hussels, leser}@informatik.hu-berlin.de

Abstract. Entries in biomolecular databases are often annotated with concepts from different ontologies and thereby establish links between pairs of concepts. Such links may reveal meaningful relationships between linked concepts, however they could as well relate concepts by chance.
In this work we present InterOnto, a methodology that allows us to rank concept pairs to identify the most meaningful associations. The novelty of our approach compared to previous works is that we take the entire structure of the involved ontologies into account. This way, our method even finds links that are not present in the annotated data, but may be inferred through subsumed concept pairs.
We have evaluated our methodology both quantitatively and qualitatively. Using real-life data from TAIR we show that our proposed scoring function is able to identify the most representative concept pairs while preventing overgeneralization. In comparison to prior work our method generally yields rankings of equivalent or better quality.

1 Introduction

Entities in many biological databases are frequently annotated with ontology concepts to facilitate researchers in searching, comparing, and browsing the data. Consider for instance the Arabidopsis Information Resource (TAIR) [20]. In TAIR genes are annotated with concepts from the Gene Ontology (GO) [1] and the Plant Ontology (PO) [7]. These annotations are complementary, as GO is a structured vocabulary for the functional description of genes and gene products, while PO concepts are used to describe plant structures and developmental stages.

A side effect of annotating database entries with ontology concepts is that data curators implicitly create links between concepts from different ontologies. Figure 1 shows such implicit links created by GO and PO annotations for the TAIR entry `AT1G15550GA4` (gibberellin 3 β-hydroxylase). Considering these links between concepts of GO and PO a scientist may for example infer that a certain biological process is located in a specific tissue of a plant, or is active in a certain developmental stage. Take the TAIR entry for gibberellin 3 β-hydroxylase in Figure 1. Gibberellins are a family of phytohormones involved in various

developmental processes such as germination, flowering, and stem elongation in *Arabidopsis thaliana* and other vascular plants. Gibberellin 3 β-hydroxylase is an enzyme that catalyses the final biosynthetic reaction to produce bioactive gibberellin 3, an essential step in the signal cascade stimulating germination after exposure to red light [22]. This semantic relationship *red light stimulates germination* could be inferred automatically in Figure 1 by considering the link between the concepts *'response to red light'* and *'germination'*. The problem with ontology links as a means to find associated concepts is that many links are artifacts. Consider again Figure 1. This TAIR entry also links the GO concepts *transcription factor binding* and *cytoplasm* to the PO concepts *root*, *leaf*, and *stem*. None of the six concept pairs derivable from these links is a meaningful association.

Without further preprocessing a researcher would have to look at a vast amount of ontology links to find those that are meaningful. Our sample TAIR entry is annotated with eight terms from GO and 19 terms from PO, creating 152 ontology links. In their daily work, researchers often deal with sets of dozens of genes or other biomolecular entities such as gene families, co-clustered genes in microarray experiments, or interacting proteins in PPI networks. In such a set each gene or protein may contribute unique annotations to the overall set of annotations. The resulting amount of ontology links becomes cumbersome to explore. In this work, we therefore present InterOnto, a method to rank and thus identify meaningful ontology links established by a set of database entries in an automated manner.

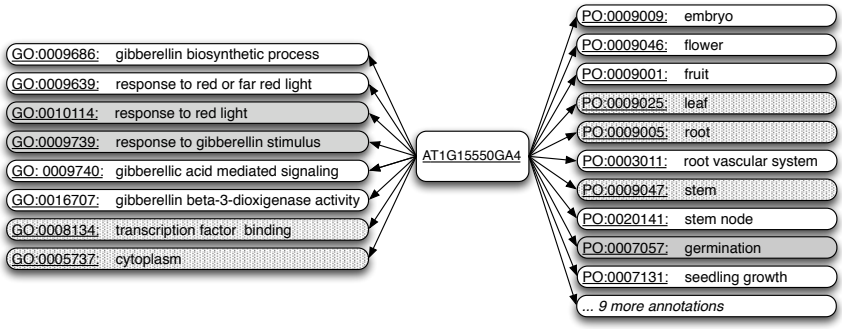
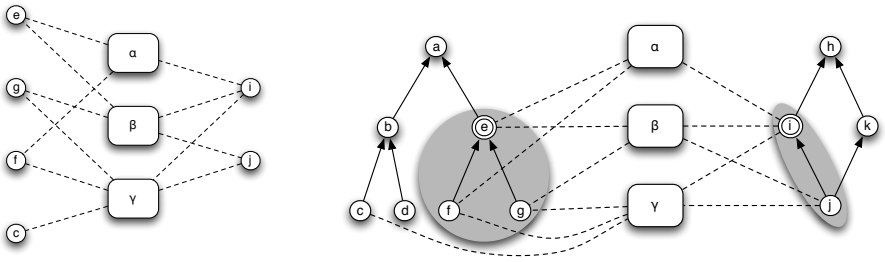


Fig. 1: Sample TAIR entry linking concepts in the Gene and Plant Ontology.

The first idea to rank concept pairs that comes to mind is to count links connecting the same concepts. Certainly, the more database entries link the same concepts, the more confident we are that these links represent a meaningful relationship. The problem with this approach is that data curators describe database entries as precisely as possible, using very specific concepts. Database entries may therefore be similar and yet not share a single annotation. In these

cases, the simple counting approach fails. Consider the situation depicted in Figure 2a. The three selected entries are annotated with terms from two different ontologies O_1 and O_2. These annotations establish 12 ontology links connecting 8 distinct pairs of concepts. The pairs *(e,i)*, *(g,i)*, *(g,j)* and *(f,i)* are linked twice, while the other pairs are linked once.



(a) Linked Concepts (b) Linked Subgraphs

Fig. 2: Set of database entries linking concepts from two ontologies O_1 and O_2.

However, the picture changes once we take not only the frequency of links, but also the relationships between ontology concepts into account. Figure 2b shows the same data plus ancestor concepts in the respective ontologies. It is evident that almost all links connect the same subgraphs in O_1 and O_2. The goal of our work is to find such pairs of strongly interlinked subgraphs and represent them with a single concept pair, i.e., the root concepts of the subgraphs. Identifying these root concepts is a challenging task. Just counting the number of links originating in a subgraph may result in overgeneralization. In Figure 2b the subgraph induced by the more general concept a contains two additional links compared to the subgraph induced by e. As result, contrary to our intuition, a would be selected as representative concept. Intuitively, the method to identify root concepts of strongly interlinked subgraphs should find a balance between the number of links in a subgraph and the generality of the root concept. One option would be to count the number of edges between the concept and the root of the ontology, i.e., the depth of a concept. This measure would assume that each edge in the ontology has the same semantic distance. As this is usually not the case, we propose to use the information content of a concept for our newly developed method InterOnto.

The remainder of this work is composed as follows. We present in Section 2 related work on identifying concept pairs from ontologies. In Section 3 we establish the basics required for our method InterOnto, which we present in Section 4. In Section 5 we present and discuss the results produced by InterOnto. Finally, in Section 6 we conclude the paper.

2 Related Work

In the first step of InterOnto we identify ontology concepts that represent a set of biological entries best. This is related to identifying significant GO concepts for a set of genes. For a review on that topic see Huang *et al.* [5]. In a different application field Brauer *et al.* [3] assign ontology concepts to text documents for ontology-supported document retrieval. To identify the most relevant concepts for a document they also consider the ontology structure and propagate scores from successor concepts upwards, which significantly improves the performance.

Extensive work has been done on finding mappings between concepts in different ontologies. Two surveys [4, 9] present an overview on ontology mapping. Several approaches rely merely on concept properties, such as concept names, synonyms, or parent and child relationships [14, 16]. Other approaches use instance-based ontology mapping to identify semantically equivalent concepts in different ontologies. [2, 15] use association rule mining to find pairs of related GO terms given a set of entries in a database annotated with GO terms. Both approaches ignore the structure of the ontologies.

Several paper present functions to compute a similarity score for a concept pair (o_1, o_2) based on the number of instances with which o_1, or o_2, or both together are annotated with. In [10] Kirsten *et al.* present four different functions. They propagate scores to parent and grandparent concepts. In contrast to our method this approach does not consider the information loss caused through subsumption and arbitrarily limits the propagation of ontology links to only two levels. Isaac *et al.* [6] present and experimentally compare five different functions. Tan *et al.* [21] present another function to find mappings between ontology concepts based on co-occurance in text documents. The functions in both studies may be extended to also account for ontology structure by adding the number of instances with which a descendant concept of o_1 or o_2 is annotated with.

The most similar work to our approach is LSLink [13]. LSLink uses the measures *support* and *confidence* to rank concept pairs based on links induced by selected database entries. Intuitively, the *support* for a concept pair (o_1, o_2) is a measure that gives an estimate if the number of links between o_1 and o_2 in the selected subset is statistically significant with respect to the underlying data. The *confidence* provides a measure to estimate if the number of links between o_1 and o_2 in the selected subset occurs by chance with respect to the annotations frequencies of both individual terms in the subset. Both scores are highest for a link whose concepts are annotated just once to the underlying data. In [12] the authors extend their approach from [13] by boosting the score of a parent concept by the scores of their child concepts to improve the ranking. This extended version of LSLink is similar to our approach InterOnto, but we do not restrict the score propagation to only one level.

In [18] Saha *et al.* mine the tripartite graph induced by selected database entries and their annotated ontology concepts for the densest subgraphs. To compute these subgraphs they not only consider the links themselves, but also the distance, i.e., the number of edges between concepts in an ontology, up to

a certain threshold. This approach has two drawbacks. First, the edge count approach assumes that each edge represents the same semantic distance, which may not be true in many ontologies. Thus, we use the information content of a concept. Second, all relationships in a mined densest subgraph are equally important, which may still leave a researcher with a huge amount of links to explore. In contrast, we present a ranked list of links.

3 Basic Concepts

Our scoring function for mining meaningful associations from ontology links incorporates information on annotation frequencies as well as ontology structure. In this section, we briefly introduce the data and ontology model. In addition, we present a measure for semantic similarity of ontology concepts. Finally, we formalize the notion of ontology links.

3.1 Data Model and Ontology Structure

We view a data source as a comprehensive set of database entries that describe the same kind of biological entity, e.g., genes, proteins, or diseases. We assume that a particular subset of these entries, the *user dataset*, is of interest to a researcher. A user data set may for example consist of genes that were over-expressed in a microarray experiment, regulated by the same transcription factor, or associated to the same disease. Entries may be *annotated* with concepts from an *ontology*.

An ontology $O(V, E)$ is a directed acyclic graph with vertex set $V(O)$ and edge set $E(O)$. Each vertex represents a concept and edges represent *is-a* relationships between concepts. Given a concept c we use the notation $desc(c)$ and $anc(c)$ to denote the sets of descendant and ancestor concepts of c. The notations $desc'(c)$ and $anc'(c)$ are short forms for $\{c\} \cup desc(c)$ and $\{c\} \cup anc(c)$.

Given a data source S and an ontology $O(V, E)$. An annotation is an ordered pair $(s, c) \in S \times V(O)$ of a database entry and an ontology concept.

The goal of our work is to mine meaningful concept associations from sets of ontology links. In the simplest case an ontology link is established by a single entry s that is annotated with concepts from two different ontologies. Let us assume $c_1 \in V(O_1)$ and $c_2 \in V(O_2)$ are two concepts from ontologies O_1 and O_2, respectively. If annotations (s, c_1) and (s, c_2) exist, we say s establishes link (c_1, s, c_2) between the concepts c_1 and c_2.

Definition 1 (Ontology Link). *Let S be a data source and $O_1(V, E)$ and $O_2(V, E)$ be two ontologies. An ontology link $(c_1, s, c_2) \in V(O_1) \times S \times V(O_2)$ is a 3-tuple, with (c_1, s) and (c_2, s) being annotations.*

3.2 Measure for Semantic Similarity

For various practical applications it is necessary to know how similar or dissimilar two ontology concepts are [8]. Several measures have been developed to determine the semantic similarity or distance between concepts in an ontology.

The simplest distance measure between two ontology concepts is the *edge count distance* [11] where the distance is the minimum number of edges between the two concepts. The implicit assumption underlying the edge count distance is that edges in *is-a* ontologies are equidistant in terms of semantic distance. In real-world ontologies this is usually not the case. Thus, several authors [8, 17] proposed to use the *self-information* of a concept to evaluate semantic similarity. The self-information is quantified as the negative logarithm of the *relative annotation frequency* $f(\hat{c})$.

Definition 2 (Self-information of a concept). *Let S be a data source, $O(V, E)$ be an ontology, and $\mathcal{A} \subset S \times V(O)$ be a set of annotations. The information content $I(\hat{c})$ of $\hat{c} \in V(O)$ regarding S is given in Equation 1.*

$$I(\hat{c}) = -\log_2(f(\hat{c})) \tag{1}$$

$f(\hat{c})$ is the relative annotation frequency of $\hat{c}$ as given in Equation 2.

$$f(\hat{c}) = \frac{|\{s \in S : \exists (s, c) \in \mathcal{A} \wedge c \in desc'(\hat{c})\}|}{|S|} \tag{2}$$

The relative annotation frequency $f(\hat{c})$ of a concept $\hat{c}$ as given in Equation 2 is the relative frequency of entries with which the concept itself or one of its sub-concepts c is annotated.

Resnik [17] suggested to use the information content of the most informative common ancestor of two ontology concepts c_1 and c_2 as measure of semantic similarity of two concepts. This score is called *shared information content*.

Definition 3 (Shared information content). *Let $c_1, c_2 \in V(O)$ be two concepts in an ontology O. Let $C = anc'(c_1) \cap anc'(c_2)$ denote the set of common ancestor concepts of c_1 and c_2. The shared information content σ of c_1 and c_2 is:*

$$\sigma(c_1, c_2) = \max_{\hat{c} \in C} I(\hat{c}) \tag{3}$$

Note, the measure σ does not produce values between 0 and 1. The scores range from 0 for the root concept to a maximum value for a given set of annotations. A concept obtains this maximum value if it is present in the least number of annotations, which usually means in exactly one annotation. This fact is important for understanding the results. An advantage of this measure is that it can naturally be extended to determine the similarity of concept sets of arbitrary size, as long as these concepts share a common ancestor. In Section 4 we show how this property is particularly useful for our application.

4 InterOnto – Linking Ontologies Using Evidence

In the following sections we present InterOnto, a methodology to rank pairs of ontology concepts based on how likely they represent meaningful information to

a researcher. We show how to incorporate information encoded in the structure of *is-a* ontologies to improve the rankings.

Consider Figure 2b again. When we simply count the links we get scores for the leaf concept pairs (f, j) and (g, j) of 1 and 2, respectively. However, an entry annotated with a particular ontology concept is implicitly also annotated with all its ancestor concepts. Thus, the scores for concept pairs (f, i), (g, i), and (e, j) are 3, 4, and 4 when counting the number of links in the subgraphs. The pair (e, i) outscores these pairs with 10 supporting links. Eventually, this approach would choose the concept pair (a, i) as it has the highest score of 12. This is counterintuitive as from a users perspective the concept pair (e, i) would represent the selected subset best. The reason for choosing (a, i) over (e, i) is, that the counting approach does not take the loss of specificity caused by subsumption into account.

4.1 Finding Representative Concepts

Consider the situation depicted on the left side of Figure 3. Intuitively, we can identify two distinct groups of two and three annotations, as highlighted in grey. To represent these groups, one would probably choose the concepts a and d as representative concepts. Counting the number of annotations of a concept and its successor concepts would rank concept a highest, as it has five annotations. The counting approach does not consider the loss of specificity when moving up the ontology. To model this loss, we propose Equation 4 to assign a *similarity based score* to a concept $\hat{c}$ with respect to the set of annotations $\mathcal{A}$ present in the user dataset.

Definition 4 (Similarity based scoring function). *Let S' be the user selected dataset of data source S, $O(V, E)$ an ontology, and $\mathcal{A}' \subset S' \times V(O)$ a set of annotations. The similarity based score of a concept $\hat{c} \in V(O)$ is given by:*

$$score_\sigma(\hat{c}) = \sum_{c \in desc'(\hat{c})} |\{(s, c) : (s, c) \in \mathcal{A}'\}| \cdot \sigma(c, \hat{c}) \tag{4}$$

If we omit factor $\sigma(c, \hat{c})$ in Equation 4 we obtain the number of annotations in the subgraph of $\hat{c}$. The factor $\sigma(c, \hat{c})$ describes the similarity between concepts c and $\hat{c}$, i.e., the more similar the concepts are the higher the similarity value.

In Equation 4 the contribution of a specific annotation c to the overall score of a concept $\hat{c}$ is the shared information content of c and $\hat{c}$, $\sigma(c, \hat{c})$. According to Definition 3 this score equals $\max_{\hat{c} \in C} I(\hat{c})$. Since $\hat{c}$ is per definition the most specific common ancestor of all c we may simply use the self-information of $\hat{c}$, $I(\hat{c})$. This allows us to transform Equation 4 to Equation 5.

$$score_\sigma(\hat{c}) = |\{(s, c) : (s, c) \in \mathcal{A}' \wedge c \in desc'(\hat{c})\}| \cdot I(\hat{c}) \tag{5}$$

In Equation 5 the score of a given concept depends on its self-information and the number of annotations in which it is present in the user dataset. Consider the example in Figure 3 again. Let us assume that each depicted concept

is annotated to one entry in the underlying data source and the total number of annotations for Ontology 1 and 2 is 50 each. Thus, for concepts d and a we have an information content of $I(d) = -\log_2 \frac{3}{50} = 4.06$ and $I(a) = -\log_2 \frac{14}{50} = 1.84$. Calculating $score(\hat{c})$ for all concepts results in the highest score for concept d with $3 \cdot 4.06 = 12.2$, followed by c with $3 \cdot 3.64 = 10.9$ and a with $5 \cdot 1.84 = 9.2$. This example as well as our experimental evaluation shows that $score_\sigma$ allows us to identify representative concepts, but overcomes the problem of overgeneralization; the more general a representative concept is, the more annotations must support it to yield a good score, as per definition $I(\hat{c}) \leq I(c), \forall \hat{c} \in anc(c)$.

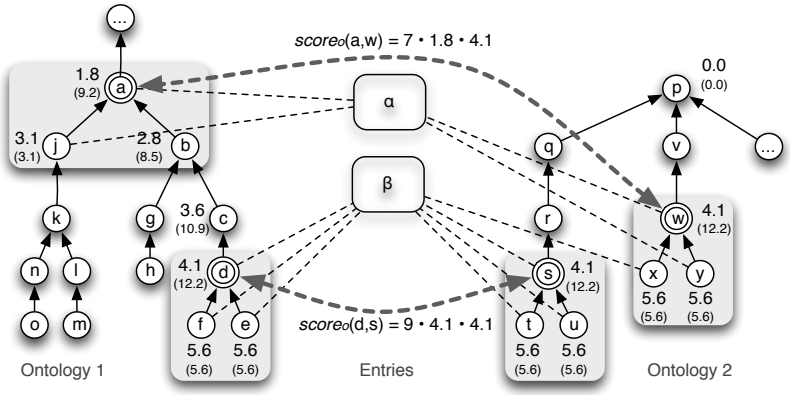


Fig. 3: Scores $score_\sigma(\hat{c_1}, \hat{c_2})$ for concept pairs. The numbers on nodes represent the information content $I(\hat{c})$ of a concept, while the numbers in brackets represent $score_\sigma(\hat{c})$.

4.2 Scoring Representative Concept Pairs

We now get back to the original problem of mining meaningful associations from ontology links. An initial idea would be to produce the cartesian product of concepts in O_1 and O_2 and rank the concept pairs based on the product of $score_\sigma(\hat{c_1})$ and $score_\sigma(\hat{c_2})$. However, this approach does not yield a meaningful result as it does not consider the number of links established through entries in the user dataset. Consider the example shown in Figure 3. The three most informative concepts are d, s, and w. The cartesian product would yield the same score for concept pairs (d, s) and (d, w). However, intuitively we would rank (d, s) higher as it is supported by nine links in total (one direct link over entry β and eight through descendant concepts of d and s over β). In comparison, the for concept pair (d, w) we find three supporting links, namely (d, β, x), (f, β, x), and (e, β, x).

The example shows that simply multiplying the similarity scores does not yield the desired result to find representative concept pairs. We may not just

consider annotations (s, c_1) and (s, c_2) separately, but we have to consider the actual links between concepts in O_1 and O_2 established through entries in the user dataset. Equation 6 provides the similarity based scoring function for pairs of ontology concepts (c_1, c_2).

Definition 5 (Similarity based scoring function). *Let S' be the user selected datset of data source S and O_1 and O_2 two ontologies. Let furthermore $\mathcal{L}' \subset V(O_1) \times S' \times V(O_2)$ denote a set of ontology links between them. The similarity based score of a concept pair $(\hat{c}_1, \hat{c}_2)$ with $\hat{c}_1 \in V(O_1)$ and $\hat{c}_2 \in V(O_2)$ is given by Equation 6.*

$$score_\sigma(\hat{c_1}, \hat{c_2}) = |\{(c_1, s, c_2) : (c_1, s, c_2) \in \mathcal{L}' \wedge c_1 \in desc'(\hat{c_1}) \wedge c_2 \in desc'(\hat{c_2})\}| \cdot I(\hat{c_1}) \cdot I(\hat{c_2}) \tag{6}$$

4.3 Eliminating Redundant Representative Concept Pairs

While incorporating hierarchy information may improve the result set, it may also introduce a significant number of redundant concept pairs. Consider for example the concept pair (c, s) in Figure 3. This pair represents the same set of links as the pair (d, s), but is less specific and therefore receives a lower score. When evaluating $score_\sigma(c_1, c_2)$ in Section 5 we thus eliminate pairs that do not add information to the overall set of concept pairs. For every set of links we only keep the most informative representative concept pair, i.e., the pair that receives the highest score.

5 Evaluation

We tested our methodology on several real world data sets, which we present in Section 5.1. We compare the top-k concept pairs produced by our method InterOnto with the top-k concept pairs resulting from the LSLink measures *support* and *confidence*.

5.1 Input Data

In the following subsections we give an overview of the ontologies, biological data sources, and data sets used for evaluation purposes.

Ontologies. In our experiments we used the Gene Ontology (GO) [1] and the Plant Ontology (PO)[7]. The Gene Ontology contains three different sub-ontologies, namely *'Molecular Function'*, *'Biological Process'*, and *'Cellular Component'*. The Plant Ontology contains two different sub-ontologies, which are *'Plant Structure'* and *'Plant Growth and Development Stage'*. In our experiments we consider these different sub-ontologies as independent ontologies, as no *'is-a'* relationships between concepts in these sub-ontologies exist.

Biological Data Sources. We used TAIR [20] as data source. In TAIR an entry may be annotated with concepts from GO and PO. We found that 52,766 entries in TAIR are annotated with GO concepts and 19,883 entries with PO concepts. In total we found 145,627 GO-TAIR annotations and 514,567 PO-TAIR annotations, resulting in 3,361,887 distinct GO-PO concept pairs.

Test Data Sets. For our quantitative evaluation we selected 20 gene sets from TAIR that constitute gene families, shown in Table 1. The genes in these gene families fulfill a similar role in the organism. Thus, for manual inspection we are able to evaluate if high ranked concepts and concept pairs are meaningful.

id	Gene Family Name	Genes	Annotated[%]	Annotations	Unique
1	Core Cell Cycle Genes	61	98	303	102
2	basic Helix-Loop-Helix (bHLH) Transcription Factor	162	98	565	103
3	Plant Cell Wall Biosynthesis Families	31	97	167	36
4	Cytoplasmic ribosomal protein gene family	248	95	1090	48
5	Lipid Metabolism Gene Families	98	94	280	98
6	Chloroplast and Mitochondria gene families	50	94	222	53
7	Primary Pumps (ATPases) Gene Families	81	89	253	104
8	Monosaccharide transporter-like gene family	53	100	222	40
9	Acyl Lipid Metabolism Family	610	92	1802	425
10	Kinesins	61	98	122	40
11	zinc finger-homeobox gene family	17	94	45	14
12	Glycosyltransferase Gene Families	280	98	1032	171
13	ABC Superfamily	126	97	307	96
14	Heat Shock Transcription Factors	21	100	77	20
15	Protein synthesis factors	95	99	321	64
16	Inorganic Solute Cotransporters	83	100	323	86
17	Ion Channel Families	59	100	240	58
18	Phosphoribosyltransferases (PRT)	15	100	58	23
19	Glycoside Hydrolase Gene Families	307	98	664	158
20	Response Regulator	32	100	197	38

Table 1: Selected gene families from the TAIR database.

5.2 Evaluation Method

Relating concepts from orthogonal ontologies is a relatively new research area. To our knowledge no gold standard exists, with which we could compare our results. One option to assess the quality of our ranking is by domain experts. But this assessment is very time consuming and may be subjective. We thus decided to use existing inter-ontology mappings as basis for our evaluation.

The problem with existing inter-ontology mappings is that they are usually not very comprehensive. For instance, the mapping between PO and GO provided by the OBO Foundry [19] consists of only 137 relationships between the sub-ontologies *Biological Process* and *Plant Structure*. A notable exception to this lack of coverage are the inter-ontology mappings in GO itself, where the sub-ontologies *Molecular Function* and *Biological Process* are richly inter-linked. In our snapshot of GO we found 517 relationships of type *biological process regulates molecular function* and 206 relationships of type *molecular function is part*

of biological process. At first glance, this number seems sufficient for assessing the ability of methods to rediscover those links.

Despite these numbers we still face the problem that many established mappings are fairly generic. Consider for example the term pair *('transferase activity, transferring glycosyl groups', 'polysaccharide biosynthetic process')*. In fact glycosyl transferases are key enzymes in the synthesis of polysaccharides, but no valid path in GO confirms this fact. The closest established relationship, as depicted in Figure 4, is the fairly generic association *'catalytic activity' part-of 'metabolic process'*.

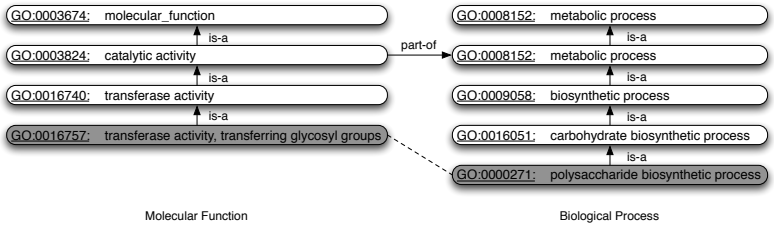


Fig. 4: Example for a potentially true positive association, PTP (dotted line).

To also use such generic relationships for assessing more specific concept pairs, we introduce *Potentially True Positive associations* (PTPs). A potentially true positive association between two ontology concepts exists, if the concepts themselves or any of their ancestor concepts have an established inter-ontology mapping.

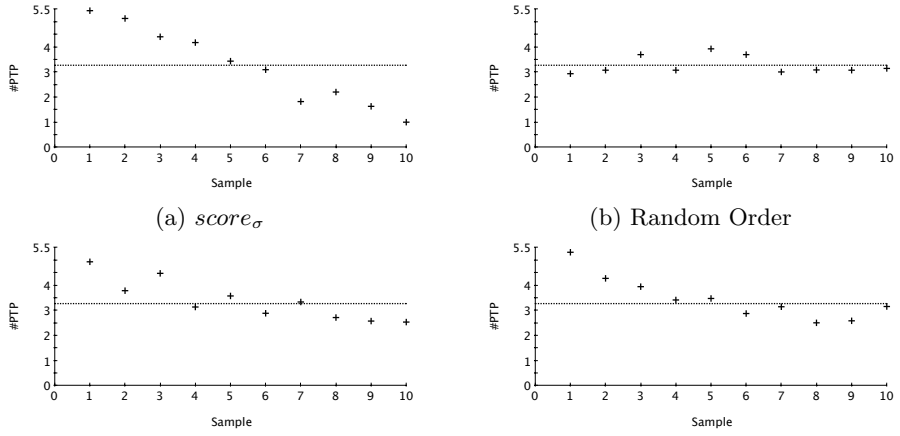
5.3 Ranking Representative Concept Pairs

We now evaluate the quality of proposed inter-ontology concept pairs. We validate the top-k concept pairs found by our method InterOnto and by LSLink against existing inter-ontology mappings. This allows us to compare both methods quantitatively.

We used the 20 gene families from TAIR to compute sets of inter-ontology links using our own scoring function $score_\sigma$ and the *confidence* and *support* scores defined by the LSLink methodology [13]. As baseline we used randomly ordered ontology links of each set. To compare the different result sets we partitioned each ranking into 10 equally sized sublists and extracted the first 10 entries from each sublist, starting with the top-10 concept pairs. Figure 5 shows the average number of PTPs for each of the 10 sublists averaged over all 20 gene families.

The most notable finding is that on average the top-10 lists for all three scoring functions contain considerably more PTPs than subsequent or random samples. This effect is stronger for $score_\sigma$ and *confidence* than for the *support* score. Since the PTP heuristic is based on well known relationships the different

results for *support* and *confidence* are actually in accordance with our expectations. The overall correlation of rank and number of PTPs is clearly the strongest for $score_\sigma$. Using linear regression we yield a fit line with an incline of -0.5 and a regression factor of 0.97. For *confidence* and support we measure an incline of -0.242 and -0.236 and regression factors of 0.77 and 0.74, respectively.



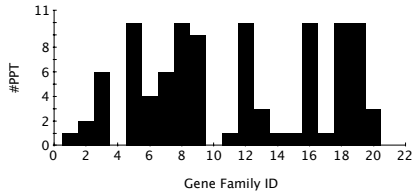
(a) $score_\sigma$ (b) Random Order

(c) *support* (d) $confidence$

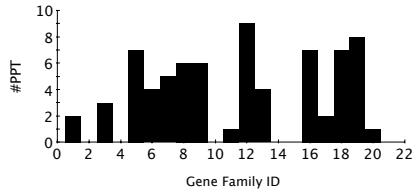
Fig. 5: Number of PTPs in top-10 concept pairs from 10 equally sized partitions. Samples are numbered in ascending order of partition by rank. Values are averaged over 20 different sets of links between GO Molecular Function and GO Biological Process, corresponding to the 20 TAIR gene families. Dotted lines mark the average number of PTPs among random concept pairs.

The plots in Figure 5 do in principle confirm our expectations, although we would have expected a higher drop between the top-10 links and subsequent or random samples. One reason for not observing that drop are overly general mappings in the Gene Ontology, e.g., *'enzyme regulator activity' regulates 'catalytic activity'*. Those mappings may detect meaningless PTPs and thus increase the score for subequent or random samples. Another factor influencing the results is that some meaningful relationships are not modeled at all. Considering the results for the different gene families from TAIR depicted in Figure 6 confirms our assumption. The plots show a strong variation in number of PTPs, with some top-10 lists not containing a single PTP for all three scores. Take for example the gene family '4 - Kinesins'. Kinesins are motor proteins that move along microtubules. Our top-10 list contains meaningful term pairs such as *('microtubule motor activity','microtubule based movement')*. Yet, none of these pairs is a PTP.

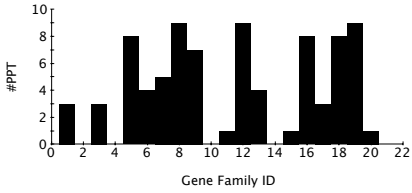
The results of our analysis suggest that our approach generally yields rankings of higher quality than those produced by LSLink. We further studied how the actual top-k concept pairs differ. We determined the relative overlap of top-



(a) $score_\sigma$



(b) *support*

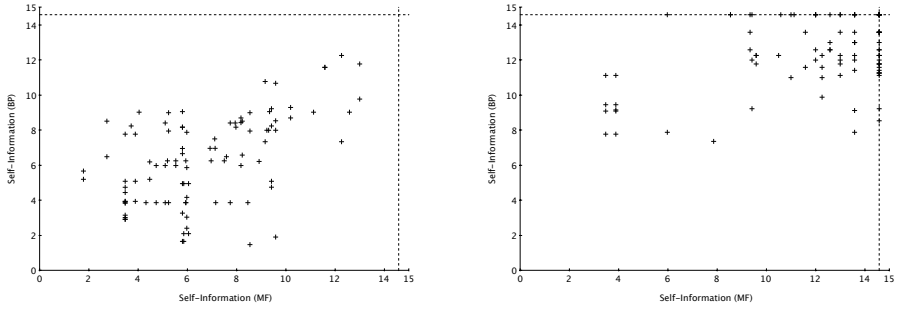


(c) *confidence*

Fig. 6: Number of potential true positive (PTP) associations in top-10 concept pairs for different link sets, generated using different scoring functions.

5 and top-10 lists for all three functions. We found that the lists produced by LSLink's *confidence* and *support* scores overlap to a much higher degree with each other than with results produced using $score_\sigma$ (data not shown).

Manual analysis of sample results showed that we may attribute these differences to the impact of hierarchy information. To quantify the differences we determined the information content of associated concepts and visualized them in Figure 7 in separate 2-dimensional plots for the top-k concept pairs computed by each of the three scoring functions (plot for *confidence* similar to plot for *support* and thus omitted). For easier interpretation we added dashed lines to depict the maximum values for self-information. Note, a concept obtains the maximum value only if it is annotated in this scenario to exactly one database entry. The plots for the top-5 concept pairs produced by the measures *confidence* and *support* (Figure 7b) show that many of the top-ranked pairs contain at least one concept with maximum self-information. This is counter intuitive, since such associations are based only on a single ontology link, which may be an artifact or an error in annotated data. In contrast, Figure 7a shows that our approach does not yield such low evidence concept pairs. We may assume though that if such associations are indeed meaningful, several similar links exist. In this case, our method should return a more general concept pair with lower self-information score that subsumes these highly specific links. Our analysis shows that in InterOnto on average 54% of the top-5 and 62% of top-10 concept pairs are inferred through subsumption and have thus not been present in the original links or can be detected without hierarchy-aware methods. This way, InterOnto may be more robust to errors in annotations compared to LSLink. Another desirable property



(a) Top-5 concept pairs $score_{\sigma}$ (b) Top-5 concept pairs *support*

Fig. 7: Self-Information of paired GO and PO concepts in top-5 lists generated using different measures. Evaluation was performed on 20 distinct sets of links corresponding to the 20 TAIR gene families listed in Table 1.

of InterOnto is that it is more likely to associate concepts of similar specificity, as the majority of points are distributed along the graphs main diagonal.

As concrete example we use the TAIR gene family 16 'Inorganic Solute Co-transporters'. Table 2 shows the top-5 concept pairs for all three functions. Notably, all top-5 lists contain only biologically valid associations. The crucial difference is that the pairs ranked high by LSLink methods all refer to the transport of specific substances through the cell membrane, while our approach returns pairs that refer to ion transport through the cell membrane in general. In other words, LSLink returns specific examples of the information contained in the link sets, while InterOnto summarizes this information, characterizing the overall dataset based on evidence provided by a large number of similar links.

6 Conclusion

We introduced a new scoring function to rank concept associations from a set of ontology links. In contrast to existing work our approach considers not only ontology concepts linked directly, but also the hierarchy of ontologies in a systematic manner. Our results show that incorporating hierarchy information allows the identification of more descriptive, yet not overgeneral, concept pairs compared to methods that do not incorporate hierarchy information, such as LSLink. Based on our experiments we believe that our method performs well and should be useful for researchers. For a thorough evaluation of our and other methods linking ontology concepts a gold standard for the quality of concept pairs would be useful.

#	Molecular Function	Biological Process
	Top-5 pairs for $score_\sigma$	
1	ion transmembrane transporter activity	ion transport
2	substrate-specific transmembrane transporter activity	ion transport
3	transmembrane transporter activity	ion transport
4	ion transmembrane transporter activity	cation transport
5	metal ion transmembrane transporter activity	cation transport
	Top-5 pairs for *support*	
1	molybdate ion transmembrane transporter activity	molybdate ion transport
2	high affinity copper ion transmembrane transporter activity	high-affinity copper ion transport
3	high affinity secondary active ammonium transmembrane transporter activity	ammonium transport
4	ammonium transmembrane transporter activity	ammonium transport
5	ammonium transmembrane transporter activity	methylammonium transport
	Top-5 pairs for *confidence*	
1	molybdate ion transmembrane transporter activity	molybdate ion transport
2	high affinity copper ion transmembrane transporter activity	high-affinity copper ion transport
3	high affinity secondary active ammonium transmembrane transporter activity	ammonium transport
4	low affinity phosphate transmembrane transporter activity	phosphate transport
5	ammonium transmembrane transporter activity	ammonium transport

Table 2: Top-5 concept pairs for TAIR gene family 16 - 'Inorganic Solute Co-transporters'.

Acknowledgement

This work was partly supported by the BMBF supported project Phänomics (grant no. 035539G). We would like to thank Louiqa Rashid for fruitful discussions on the topic of ontology links.

References

1. M. Ashburner, C. Ball, J. Blake, D. Botstein, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000.
2. O. Bodenreider, M. Aubry, and A. Burgun. Non-lexical approaches to identifying associative relations in the gene ontology. *Pac Symp Biocomput*, pages 91–102, 2005.
3. F. Brauer, M. Huber, G. Hackenbroich, U. Leser, *et al.* Graph-Based Concept Identification and Disambiguation for Enterprise Search. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 171–180, 2010. ACM.
4. S. Castano, A. Ferrara, S. Montanelli, and G. Varese. Ontology and Instance Matching. In *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, volume 6050 of *Lecture Notes in Computer Science*, pages 167–195. Springer, 2011.
5. D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1–13, Jan 2009.

6. A. Isaac, L. van der Meij, S. Schlobach, and S. Wang. An Empirical Study of Instance-Based Ontology Matching. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, volume 4825 of *Lecture Notes in Computer Science*, pages 253–266, 2007. Springer.
7. P. Jaiswal, S. Avraham, K. Ilic, E. A. Kellogg, *et al.* Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages. *Comp Funct Genomics*, 6(7-8):388–397, 2005.
8. J. J. Jiang and D. W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *CoRR*, cmp-lg/9709008, 1997.
9. Y. Kalfoglou and M. Schorlemmer. Ontology mapping: the state of the art. *The Knowledge Engineering Review*, 18(1):1 – 31, 2003.
10. T. Kirsten, A. Thor, and E. Rahm. Instance-Based Matching of Large Life Science Ontologies. In *Proceedings of the 4th International Workshop on Data Integration in the Life Sciences (DILS)*, volume 4544 of *Lecture Notes in Computer Science*, pages 172–187, 2007. Springer.
11. J. H. Lee, M.-H. Kim, and Y.-J. Lee. Ranking Documents in Thesaurus-Based Boolean Retrieval Systems. *Inf. Process. Manage.*, 30(1):79–92, 1994.
12. W.-J. Lee, L. Raschid, H. Sayyadi, and P. Srinivasan. Exploiting Ontology Structure and Patterns of Annotation to Mine Significant Associations between Pairs of Controlled Vocabulary Terms. In *Proceedings of the 5th International Workshop on Data Integration in the Life Sciences (DILS)*, volume 5109 of *Lecture Notes in Computer Science*, pages 44 – 60, 2008. Springer.
13. W.-J. Lee, L. Raschid, P. Srinivasan, N. Shah, *et al.* Using Annotations from Controlled Vocabularies to Find Meaningful Associations. In *Proceedings of the 4th International Workshop on Data Integration in the Life Sciences (DILS)*, volume 4544 of *Lecture Notes in Computer Science*, pages 247–263, 2007. Springer.
14. A. Maedche and S. Staab. Computing Similarity between Ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, volume 2473 of *LNAI*, 2002. Springer Verlag.
15. S. Myhre, H. Tveit, T. Mollestad, and A. Laegreid. Additional Gene Ontology structure for improved biological reasoning. *Bioinformatics*, 22(16):2020–2027, Aug 2006.
16. N. Noy and M. Musen. The PROMPT suite: Interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, 59(6):983–1024, 2003.
17. P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 448–453, 1995.
18. B. Saha, A. Hoch, S. Khuller, L. Raschid, and X.-N. Zhang. Dense Subgraphs with Restrictions and Applications to Gene Annotation Graphs. In *Proceedings of the 14th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, volume 6044 of *Lecture Notes in Computer Science*, pages 456–472, 2010. Springer.
19. B. Smith, M. Ashburner, C. Rosse, J. Bard, *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–1255, Nov 2007.
20. D. Swarbreck, C. Wilks, P. Lamesch, T. Z. Berardini, *et al.* The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*, 36(Database issue):D1009–D1014, Jan 2008.

21. H. Tan, V. Jakoniene, P. Lambrix, J. Aberg, and N. Shahmehri. Alignment of Biomedical Ontologies Using Life Science Literature. In *Proceedings of the PAKDD International Workshop on Knowledge Discovery in Life Science Literature (KDLL)*, volume 3886 of *Lecture Notes in Computer Science*, pages 1–17, 2006. Springer.
22. S. Yamaguchi, M. W. Smith, R. G. Brown, Y. Kamiya, and T. Sun. Phytochrome Regulation and Differential Expression of Gibberellin 3β-Hydroxylase Genes in Germinating Arabidopsis Seeds. *Plant Cell*, 10(12):2115–2126, Dec 1998.