Finding Cross Genome Patterns in Annotation Graphs

Joseph Benik[1], Caren Chang[1], Louiqa Raschid[1], Maria-Esther Vidal[2], Guillermo Palma[2], and Andreas Thor[3]

[1] University of Maryland
[2] Universidad Simón Bolívar
[3] University of Leipzig
jvb@umiacs.umd.edu, carenc@umd.edu, louiqa@umiacs.umd.edu, mvidal@ldc.usb.ve, gpalma@ldc.usb.ve, thor@informatik.uni-leipzig.de

Abstract. Annotation graph datasets are a natural representation of scientific knowledge. They are common in the life sciences where concepts such as genes and proteins are annotated with controlled vocabulary terms from ontologies. Scientists are interested in analyzing or mining these annotations, in synergy with the literature, to discover patterns. Further, annotated datasets provide an avenue for scientists to explore shared annotations across genomes to support cross genome discovery. We present a tool, PAnG (Patterns in Annotation Graphs), that is based on a complementary methodology of graph summarization and dense subgraphs. The elements of a graph summary correspond to a pattern and its visualization can provide an explanation of the underlying knowledge. We present and analyze two distance metrics to identify related concepts in ontologies. We present preliminary results using groups of Arabidopsis and C. elegans genes to illustrate the potential benefits of cross genome pattern discovery.

1 Introduction

Arabidopsis thaliana is a flowering plant that is widely used as a model organism and whose genome was completely sequenced in the year 2000. The Arabidopsis Information Resource (TAIR) is a well curated and heavily used portal for accessing Arabidopsis genome information [6,19,21]. TAIR provides a rich synopsis of each gene through links to a variety of data including Gene Ontology (GO) [2,7] and Plant Ontology (PO) [26].

We illustrate annotation datasets using a study of genes involved in photomorphogenesis. The GO-PO annotation graph for gene `CRY2` is in Figure 1. The PO annotations for `CRY2` are on the left side and the GO annotations are on the right. We label this a *tri-partite annotation graph* or *TAG*. Each node of the *TAG* includes the identifier and the label for the Controlled Vocabulary (CV) term. As of September 2011, there were 17 `GO` and 37 `PO` annotations for `CRY2`. The figure illustrates partial annotations. On the right of Figure 1 is a fragment of the relevant `GO` ontology.

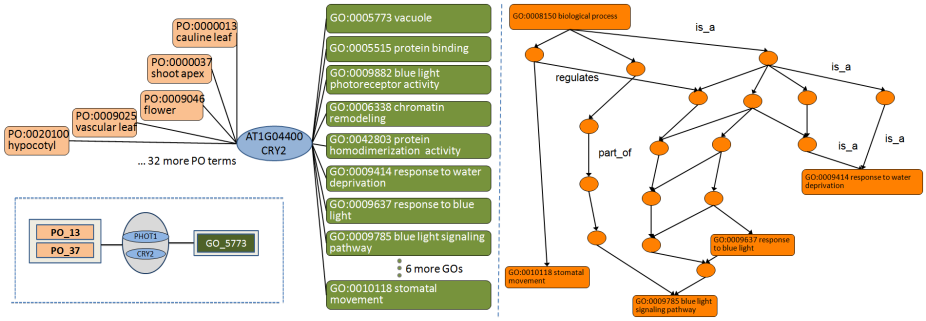


Fig. 1. GO and PO annotations for gene CRY2 (middle); GO fragment (right); Graph Summary(GS) for genes CRY2 and PHOT1 (inset)

Over the past 25 years, knowledge of the Arabidopsis genome has increased exponentially, together with that of other model organisms. This abundance of data has led to an era of comparative genomics, in which genes can be compared across diverse taxa to provide insights into evolutionary similarities as well as key divergences. Already the study of genes in Arabidopsis has helped to inform human research and vice versa. Increasingly, every new genome must be understood in light of previously sequenced and analyzed genomes.

We will consider orthologous genes from three model organisms: Saccharomyces cerevisiae (yeast), Caenorhabditis elegans (nematode) and Drosophila melanogaster (fruit fly). The rationale for including these highly annotated genomes is to synthesize existing knowledge to enhance our understanding of gene function in Arabidopsis (and possibly vice versa). A future aim of our research will be to extend cross genome analysis to include a range of plants, including lower plant species. Currently, a number of these genomes are sparsely annotated. By incorporating the genomes of such plant species, we may help to build knowledge in less well-studied species by bootstrapping to Arabidopsis, while strengthening our overall understanding of plant gene function and evolution.

We recognize that our goals are ambitious and that we have to solve numerous challenges. First, we have to find patterns in annotation graph datasets. On this challenge, we can report some initial success [1,22,27].

Next, we must integrate annotation data across multiple organisms to perform comparative genomics. We must identify a protocol and efficient processes to obtain orthologs or other matching genes from multiple organisms. Potential resources and tools include the Homologene service from NCBI [9], Inparanoid [10], a database that includes animals, and Plaza [18], which is exclusive to plant species. In this paper, we bypass this potentially expensive process and describe a simpler, less expensive protocol. We use shared annotations, and gene and protein families, to harvest Arabidopsis and C. elegans genes and annotations for cross genome analysis. We recognize that this protocol is less accurate at finding orthologous genes and we use it only for proof-of-concept purposes.

A key element in finding patterns is identifying related ontological concepts. A fragment of the GO ontology is shown on the right of Figure 1. We postulate that ontology terms that are located in proximity to each other in the ontology are more related. In addition, terms which are located along branches of the tree with greater depth and/or breadth potentially reside in areas where the ontological concepts are defined at a more granular level of functional or descriptive detail. Finally, pairs of terms within the same proximity and that are (both) more distant from the root of the tree may be more related. We propose a metric d_{tax} for taxonomic distance and compare to d_{ps} [16], a state-of-the-art metric. Figure 1 also illustrates different types of relationships in GO including *part_of*, *is_a* and *regulates*. While these relationship types are important in determining relatedness, we have not used these features in our current work. The contributions of this paper are as follows:

- We present the concept of tripartite annotation graphs (TAG) and our tool PAnG (Patterns in Annotation Graphs) to identify patterns. PAnG relies on dense subgraphs and graph summarization methods.
- Using sample datasets of groups of genes from Arabidopsis and C. elegans that share similar gene function, we show some preliminary results of validating PAnG for cross genome analysis.
- We study the properties of metrics d_{tax} and d_{ps} for a subset of GO terms and demonstrate that d_{tax} is better able to discriminate between taxonomically close terms.

This paper is organized as follows: Section 2 presents an overview of PAnG including dense subgraphs and graph summarization. Section 3 considers groups of genes from Arabidopsis and C. elegans, with shared function and GO annotation, to explore the potential benefits of cross genome pattern discovery. Section 4 presents the two distance (similarity) metrics d_{tax} and d_{ps} and compares their properties on several subsets of GO biological process (GO-BP) terms.

2 Overview of PAnG

Figure 2 illustrates the overall workflow of PAnG. The input is a tripartite annotated graph G, and the output is a graph summary. Our workflow consists of two steps. The first step is optional and deals with the identification of dense subgraphs, i.e., highly connected subgraphs of G that are (almost) cliques. The goal is to identify interesting regions of the graph by extracting a relevant subgraph.

Next, graph summarization transforms the graph into an equivalent compact graph representation. Graph summaries are made up of the following elements: (1) supernodes; (2) superedges; (3) deletion and addition edges (corrections). The left inset of Figure 1 shows a fragment of a graph summary obtained from the analysis of photomorphogeneis genes in Arabidopsis. There is a supernode with the two genes PHOT1 and CRY2 and another supernode with two PO terms. There is a superedge between these two supernodes reflecting that the two genes are both annotated with the two PO terms. Both genes are also annotated with

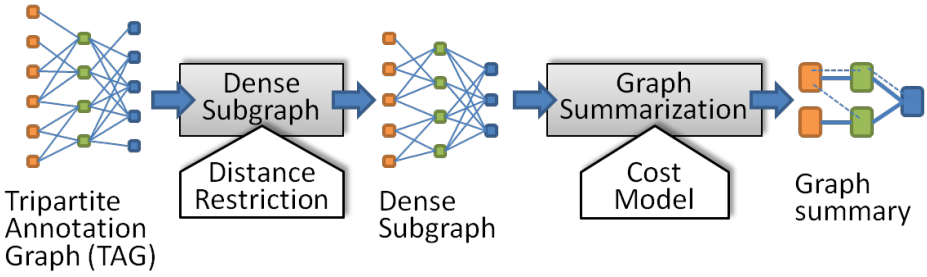


Fig. 2. The original TAG can be subject to an optional filter step to identify dense subgraphs. The PAnG tool employs graph summarization to identify patterns.

the GO term. We note that while this appears to be a very simple pattern, the association of these two genes and their PO and GO terms annotations represented an as yet unknown and potential interaction between phototropins (`PHOT1`) and chryptochromes (`CRY2`) [22].

The summary reflects the basic pattern (structure) of the graph and is accompanied by a list of corrections, i.e., deletions and additions, that express differences between the graph and its simplified pattern. For example, a deletion reflects that a gene *does not have* a particular annotation that is shared by other genes within the supernode.

A graph summary has several advantages. First, it gives a better understanding of the structure of the underlying graph and is good for visualization. Second, the summary captures semantic knowledge not only about individual nodes and their connections but also about groups of related nodes. Third, the corrections, in particular deletions, are intuitive indicators for future edge prediction.

Our approach is not limited to TAGs. A k-partite layered graph can be first converted to a more general (bi-partite) graph. Our experience is that when presented with patterns, a bi-partite graph that combines terms from multiple ontologies into one layer may not convey the same intuitive meaning to a scientist. With more than 3 layers, however, the patterns become more difficult to comprehend.

2.1 Dense subgraphs

Given an initial tripartite graph, a challenge is to find interesting regions of the graph, i.e., candidate subgraphs, that can lead to valuable patterns. We commence with the premise that an area of the graph that is rich or dense with annotation is an interesting region to identify candidate subgraphs. For example, for a set of genes, if each is annotated with a set of GO terms and/or a set of PO terms, then the set of genes and GO terms, or the set of genes and PO terms, form a clique. We thus exploit cliques, or dense subgraphs (DSG) representing cliques with missing edges. Density is a measure of connectedness. It is the ratio of the number of induced edges to the number of vertices in the subgraph.

Even though there are an exponential number of subgraphs, a subgraph of maximum density can be found in polynomial time [13,8,5]. In contrast, the maximum clique problem to find the subgraph of largest size having all possible edges is NP-hard; it is even NP hard to obtain any non-trivial approximation. Finding densest subgraphs with additional size constraints is NP hard [12]; yet, they are more amenable to approximation than the maximum clique problem.

An annotation graph is a tripartite graph $G = ((A, B, C), (X, Y))$. PAnG employs our approach in [22] and thus first transforms the tripartite graph G in a weighted bipartite graph $G' = (A, C, E)$ where each edge $e = (a, c) \in E$ is labeled with the number of nodes $b \in B$ that have links to both a and c. We then compute a densest bipartite subgraph G_2 by choosing subsets of A and C to maximize the density of the subgraph. Finally, we build the dense tripartite graph G_3 out of the G_2 by adding all intermediate nodes $b \in B$ that are connected to at least one node of G_2.

In an ontology (see right inset of Figure 1), nodes from PO and GO are hierarchically arranged to reflect their relationships (e.g., is-a or part-of). The PAnG tool allows users to include restrictions on the ontology terms in the DSG. The simplest restriction is a *distance restriction* that specifies the maximal path length between pairs of nodes in set A (C). To this end, PAnG employs a distance metric d_A (d_C) and computes the densest subgraph G_3 that ensures that all node pairs of A (C) are within a given distance τ_A (τ_C). Furthermore, the user can filter the ontology by the relationship type, i.e., only node pairs that are in a specific relationship are considered for distance computation. The current version of PAnG [1] uses the simple *shortest path length* between a pair of terms as the distance metric. In this paper, we evaluate more sophisticated distance metrics in Section 4.

2.2 Graph summarization

PAnG generates graph summaries for representing patterns. A summary of a tripartite annotation graph is also a graph. While there are many methods to summarize graphs, we focus on the graph summarization (GS) approach of [15]. Their graph summary is an aggregate graph comprised of a signature and corrections. It is the first application of minimum description length (MDL) principles to graph summarization and has the added benefit of providing intuitive coarse-level summaries that are well suited for visualization and link prediction.

A graph summary (GS) of a graph $G = ((A, B, C), (X, Y))$ consists of a graph **signature** $\Sigma(G)$ and a set of **corrections** $\Delta(G)$. The graph signature is defined as follows: $\Sigma(G) = ((S_{AC}, S_B), S_{XY})$. The sets S_{AC} and S_B are a disjoint partitioning of $A \cup C$ and B, respectively, that cover all elements of these sets. Each element of S_{AC} or S_B is a **supernode** and consists of one or more nodes of the original graph. An element of S_{XY} is a **superedge** and it represents edges between supernodes, i.e., $S_{XY} \subseteq S_{AC} \times S_B$. The **corrections** are the sets of edge additions and deletions $\Delta(G) = (S_{add}, S_{del})$. All edge additions are edges of the original graph G, i.e., $S_{add} \subseteq X \cup Y$. Deletions are edges between nodes of G that do not have an edge in the original graph, i.e., $S_{Del} \subseteq ((A \cup C) \times B) - (X \cup Y)$.

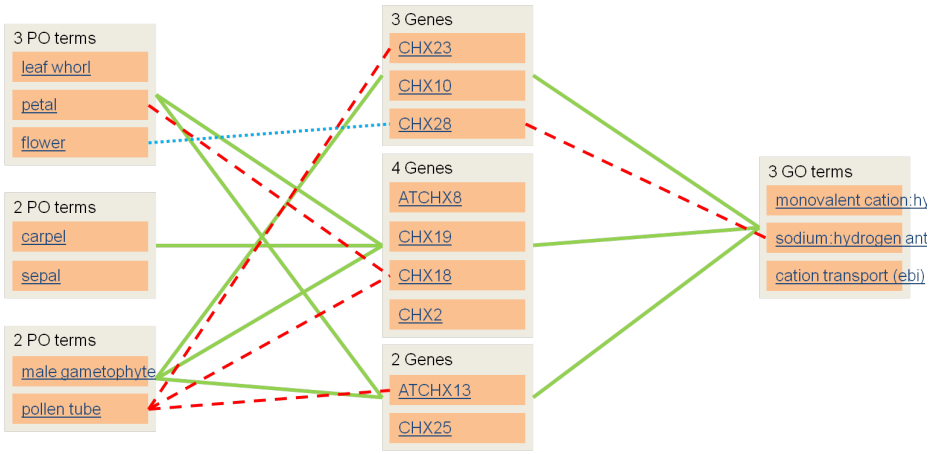


Fig. 3. Screenshot of a graph summary as generated by PAnG. Superedges are represented by green solid lines. Corrections include red dashed (deletion) and blue dotted (addition) lines, respectively.

Graph summarization is based on a two-part minimum description length encoding. The complexity of the original GS problem is currently unknown. However, if nodes are allowed to belong to more than one super node (i.e., overlapping supernodes), the problem reduces to finding the maximum clique in a graph, which is NP-hard. We use a greedy agglomerative clustering heuristic. The possible summaries of a graph will depend on the cost model used for an MDL encoding. In general, the cost model assigns weights to the number of superedges, deletions, and additions, respectively. Graph summarization looks for a graph summary with a minimal cost. Currently PAnG employs a simple cost model that gives equal weight to supernodes, superedges, and corrections.

2.3 Example Graph Summary

The Sze Laboratory at the University of Maryland is studying 20 `CHX` (Cation/H+ Exchanger) genes within the `CPA2` family [23]; Figure 3 shows a dense subgraph (DSG) of the following 9 genes: `CHX2`, `CHX10`, `CHX18`, `CHX19`, `CHX23`, `CHX25`, `CHX28`, `ATCH8`, and `ATCH13`. A *supernode* (shaded rectangle) groups together either genes or GO terms or PO terms.[4] For example, there are 3 gene supernodes; the top supernode includes 3 genes, `CHX23`, `CHX10`, and `CHX28` while the middle supernode has 4 genes, `ATCHX8`, `CHX19`, `CHX18`, and `CHX2`. Figure 3 also includes 3 PO supernodes of 3, 3, and 2 terms, respectively; this summary also contains 1 GO supernode with 3 terms.

A *superedge* is a thick edge in the figure and occurs between 2 supernodes; it represents that all nodes in both the supernodes are connected to each other.

[4] Supernodes can also group dissimilar terms if desired, e.g., GO and PO terms.

For example, the superedge between the middle PO supernode with 2 PO terms `carpel` and `sepal`, and the middle gene supernode with 4 genes indicates that all 4 genes are each annotated with both PO terms.

We use the supernodes and superedges to explain the patterns. The top gene supernode (with 3 genes) has 1 superedge to the bottom PO supernode (with 2 terms). In contrast, the middle gene supernode (4 genes) has 3 superedges to each of the 3 PO supernodes. The bottom gene supernode (2 genes) has 2 superedges. Thus, the pattern distinguishes the 4 genes in the middle gene supernode, each annotated with 7 PO terms, from the 3 genes in the top gene supernode with the least number of PO annotations. `CHX18` in the middle gene supernode is an outlier as will be discussed. 8 genes (except `CHX28`) are also annotated with all 3 GO terms in the GO supernode; thus, the gene function behavior of these 8 genes is identical with respect to these 3 GO terms. Sze confirmed the consistency of these patterns with results reported in [24].

Finally, the summary in Figure 3 illustrates *deletion* edges; these are broken edges in the figure and represent a deviation of behavior. A *deletion* reflects that a gene *does not have* a particular GO or PO annotation that is shared by the other genes (within the supernode). For example, `CHX18` (middle supernode) is *not annotated* with PO terms `petal` or `pollen tube`; this is consistent with tissue localization results in [24]. `CHX28` (top supernode) is not annotated with GO term `sodium:hydrogen`.... While this gene has not been studied, the patterns appears consistent with function based on phylogenetic tree analysis [23].

3 Preliminary Cross Genome Validation

3.1 Data Collection Protocol and Statistics

Ideally, PAnG would use tools such as Inparanoid [10], Plaza [18], and Homologene [9] to find all known homologs of a given gene, in some alternate organism. For our proof-of-concept prototype, we apply a simpler protocol to identify genes with shared annotations. Collaborators Sze and Haag identified families of Arabidopsis or C elegans genes, respectively, as genes of interest. We then used GO terms describing their function to retrieve corresponding genes in a sister organism.

- **At_8** and **Ce_9**
  At_8: Eight Arabidopsis genes in families labeled `NHX` or `SOS`; responsible for ion transport; seven are members of a sodium proton exchanger family.
  Ce_9: Nine C. elegans genes in families labeled `nhx` or `pbo`; all are members of a sodium proton exchanger family.
- **At_37** and **Ce_53**
  At_37: 37 Arabidopsis genes. We started with a collection of 19 genes, identified by Sze, as occurring in family number(s) 212, 277, and 469; all are putative heavy ion transporting P2A-type ATPase genes [4,25]. We expanded this set to include all genes in families labeled `ACE`, `ECA`, `HMA`, `RAN`, and `PAA`.
  Ce_53: 53 C. elegans genes that are annotated with terms `ion transport` and/or `divalent cations`.

Dataset	Genes	Annotation	unique GO Terms	Biol. Proc.	Cell. Comp.	Molec. Funct.
At_8	8	117	28	17	6	5
Ce_9	9	91	25	20	3	2
Overlap			6	4	0	2

Table 1. Statistics for datasets At_8 and Ce_9. The 4 overlapping biological processes are cation transport, regulation of pH, sodium ion transport, and transmembrane transport. The 2 overlapping molecular functions are sodium:hydrogen antiporter activity and solute:hydrogen antiporter activity.

Dataset	Genes	Annotation	unique GO Terms	Biol. Proc.	Cell. Comp.	Molec. Funct.
At_37	455	37	106	55	29	22
Ce_53	53	685	48	21	9	18
Overlap			17	8	6	4

Table 2. Statistics for datasets At_37 and Ce_53. The 8 overlapping biological processes are ATP catabolic process, ATP biosynthetic process, transport, cation transport, metabolic process, response to metal ion, response to manganese ion, and manganese ion homeostasis. The 6 overlapping cellular components are intracellular, nucleus, cytoplasm, vacuolar membrane, membrane, and integral to membrane. The 4 overlapping molecular functions are zinc ion binding, coupled to transmembrane movement of ions, phosphorylative mechanism, and ATPase activity,

We next report on the number of (distinct) GO terms associated with each dataset, as well as the overlap, along the three GO branches, biological process (GO-BP), molecular function (GO-MF) and cellular component (GO-CC) in Table 1 and Table 2, respectively. We also report on the total number of annotations since multiple genes in the dataset could be annotated with the same GO term [5].

3.2 Cross Genome Validation Using GS and DSG+GS Summaries

Figure 4 shows a graph summary (GS) for 8 genes in **At_8**. GO-BP terms are on the right and GO-MF and GO-CC on the left. Two genes supernodes include (`NHX2`, `NHX6`), and (`NHX3`, `NHX4`, `NHX5`), respectively. All the genes are annotated with the 3 GO-BP terms in the top GO-BP supernode. They do not appear to share many other GO-BP terms. Similarly, the 8 genes do not appear to share many GO-MF or GO-CC terms. We note that there is a deletion edge indicating that while `NHX2` is associated with both `sodium hydrogen antiporter` and `sodium ion transmembrane transporter` function, `NHX6` which shares many functions with `NHX2` and is in the same gene supernode, is not annotated with `sodium ion transmembrane transporter` function.

[5] We further note that there are cases where a single gene is annotated more than once with the same GO term; this occurs when there is dissimilar annotation evidence from multiple sources.

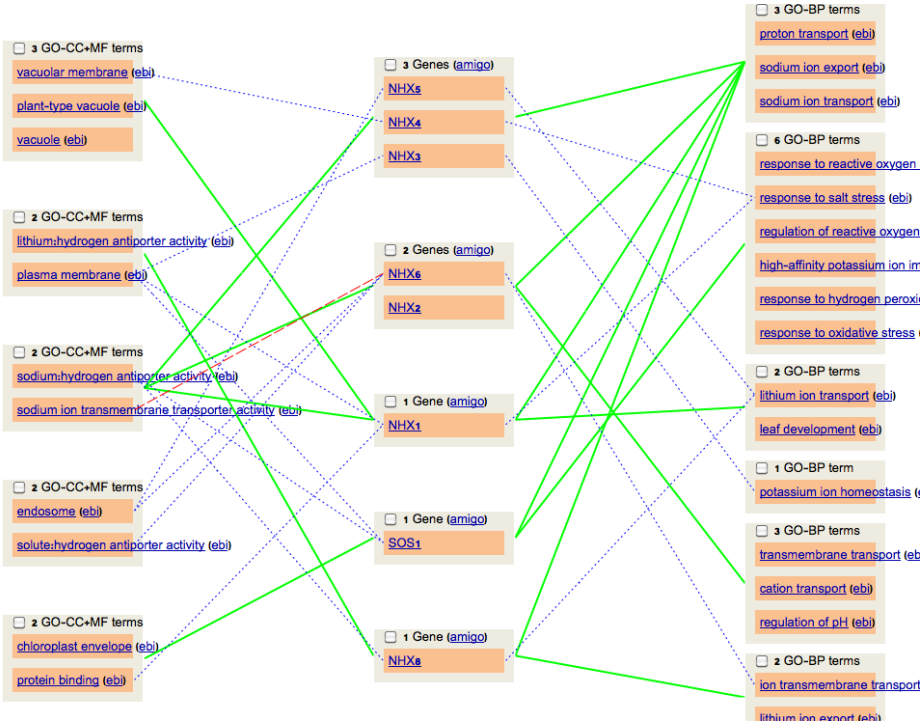


Fig. 4. Graph Summary (GS) of GO annotations for 8 genes in **At_8**; GO-BP on the right and GO-MF and GO-CC on the left.

Figure 5 shows the graph summary (GS) for 9 genes in **Ce_9**. GO-BP terms are on the right and GO-MF and GO-CC on the left. All 9 genes share the three GO-MF terms in the supernode on the upper left. They also share 4 GO-BP terms grouped into a supernode on the right in the middle of the figure. 8 of the genes are gouped into a single gene supernode, with `nhx-2` being the outsider. This is because `nhx-2` appears to be much more richly annotated with an additional 11 GO-BP terms. Only `nhx-1` shares a single GO-BP term in that group, `positive regulation of growth rate`. However, these 8 genes in the supernode do not appear to share many additional GO-BP annotations. For example, only `nhx-1`, `nhx-4` and `nhx-6` share the GO-BP annotation `embro development ending in birth`.

Finally, we combine the 8 Arabidopsis and the 9 C. elegans genes. We then identify a dense subgraph (DSG) with no distance restrictions. Figure 6 illustrates the benefit of creating a DSG prior to applying the graph summary (GS); DSG+GS identifies a *single* gene supernode that includes the 9 C. elegans genes and 2 Arabidopsis genes, `NHX2` and `NHX6`. These 2 Arabidopsis genes are included since both are annotated with the 4 GO-BP terms that annotated all 9 C. elegans genes as well as one (two) GO-MF terms. We note that the same 4

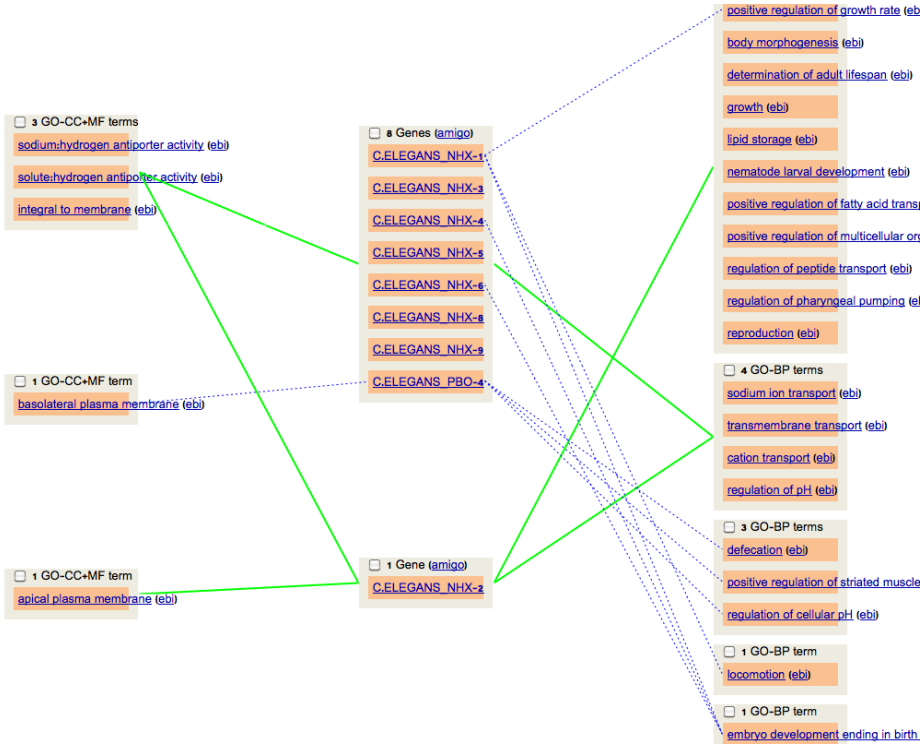


Fig. 5. Graph Summary (GS) of GO annotations for 9 genes in **Ce_9**; GO-BP on the right and GO-MF and GO-CC on the left.

GO-BP terms and 2 of the 3 GO-MF of Figure 6 were identified in the overlap of Table 1. Thus, the DSG+GS annotation pattern is consistent with the shared annotations and overlap. Further, the DSG+GS annotation pattern provides a more detailed and nuanced understanding compared to the simple data of the overlap.

Based on phylogeny studies, NHX5 and/or NHX6 (intramembrane/Golgi phenotype) are more likely to be close homologs to C. elegans genes. NHX2 is part of the [NHX1-NHX4] group with a vacuolar-localized phenotype; phylogenetic studies show they are typically *plant-specific* genes. Thus, the supernode grouping of NHX2 and NHX6 with the 9 C. elegans genes appears to be partially validated using biological knowledge but requires further study to determine if this grouping may also have resulted from an incomplete annotations of these genes.

4 Using Distance Metrics for Validation

In Section 3, the dense subgraph (DSG) did not consider any distance restriction between the pairs of GO terms in the subgraph. Similar the graph summarization

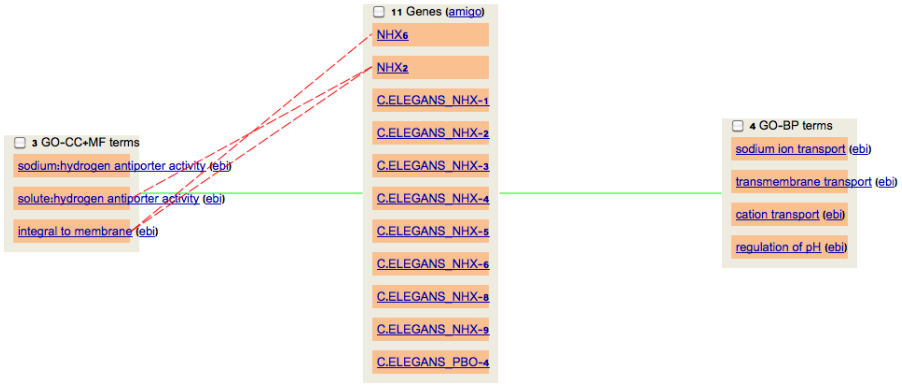


Fig. 6. Graph Summary of a Dense Subgraph (DSG+GS) of GO annotations for 8 genes in **At_8** and 9 genes in **Ce_9**; GO-BP on the right and GO-MF and GO-CC on the left.

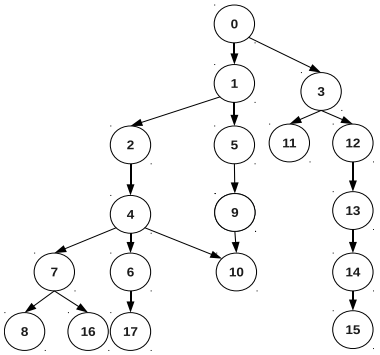
(GS) did not consider any distance restriction when constructing GO supernodes. This has significant limitations since terms in GO reflect concepts, and proximity (parents, siblings and neighbors) reflect relatedness of these concepts.

In this section, we consider two distance metrics that can be applied to taxonomies to measure the relatedness or similarity of concepts. We recognize that relatedness and similarity are not always synonymous. Our proposed metric is labeled d_{tax} and we compare it to d_{ps} [16], a state-of-the-art metric from the literature. We report on experiments performed on several datasets. In the range $[0.0 \ldots 1.0]$, where 1.0 represents no similarity, d_{tax} provides a wider dispersion of values, compared to d_{ps}. This wider dispersion provides d_{tax} with better discrimination of concepts that are not related, and hence more suited to the task of identifying related concepts.

4.1 Distance Metrics

The taxonomic organization of vertices in an ontology, as well as node properties such as descendants and ascendants, have been considered to develop state-of-the-art distance metrics that identify near neighbors, i.e., those that are proximal to each other in the taxonomy [11,14,17,20,28]. Consider the taxonomy of Figure 7(a). A *good* taxonomic distance metric should reflect that while the number of edges (say shortest path length) between a pair of nodes (1, 9) and (4, 17) may both be equal to 2, the taxonomic distances between these pairs should be different. The reasoning is that nodes that are deeper in the hierarchy and farther from the root are more specific.

Taxonomic distance metrics take values in the range $[0.0 \cdots 1.0]$, where 0.0 represents the greatest similarity. A desirable property is that two nodes that are (1) farther from the root and (2) closer to their lowest common ancestor, should be closer in distance. For example, in Figure 7(a), the pair of nodes (8,



(a) Example taxonomy

Pair	d_{tax}	d_{ps}
(0,1)	1.0	1.0
(1,3)	1.0	1.0
(6,7)	0.25	0.40
(7,8)	0.11	0.20
(8,15)	1.0	1.0
(8,16)	0.09	0.17
(8,17)	0.4	0.57
(10,17)	0.33	0.50
(11,12)	0.50	0.66

(b) Distances

Fig. 7. Example of taxonomic distance(s) between pairs of vertices for an example taxonomy.

16) should have a lower taxonomic distance to each other compared to the pair (11, 12), although the path length = 2 for both pairs. This is because the pair (8, 16) is farther from the root compared to (11, 12). The depth of a node from the root and the lowest common ancestor are defined as follows:

Definition 1 (Vertex Depth). *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.*

Definition 2 (Lowest Common Ancestor [3]). *Given a directed graph G, the* **lowest common ancestor** *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 on the longest 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.

We propose a taxonomic distance (d_{tax}) which is defined as follows:

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

where *root* is the root node in the ontology.

We compare d_{tax} to a state-of-the-art distance metric d_{ps} [16] which is defined as follows:

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

The intuition behind the d_{ps} metric proposed by Pekar and Staab [16] is that it captures the ability to represent the taxonomic distance between two vertices

Dataset	d_{tax}		d_{ps}	
	average	std_dev	average	std_dev
At_8 Arabidopis	0.936	0.207	0.964	0.133
Ce_9 C. elegans	0.868	0.275	0.925	0.182
At_8 ∩ **Ce_9**	0.872	0.270	0.925	0.185
At_8 ∪ **Ce_9**	0.962	0.159	0.977	0.106
At_37 Arabidopis	0.817	0.324	0.877	0.246
Ce_53 C. elegans	0.947	0.187	0.974	0.103
At_37 ∩ **Ce_53**	0.820	0.316	0.881	0.220
At_37 ∪ **Ce_53**	0.965	0.153	0.980	0.097

Table 3. Average and Variance of Taxonomic Distance(s) between GO-BP

with respect to the depth of the common ancestor of these two vertices. Our proposed d_{tax} distance metric is 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. Although d_{tax} distance metric satisfies theoretical distance properties, i.e., zero law, symmetry and triangle inequality, we do not focus on the formalization of these properties in this paper. In contrast, we show an empirical analysis of d_{tax} and how it compares to $d_{ps}$when both metrics are used to measure the relatedness or similarity of taxonomic concepts.

4.2 Properties of the Distance Metrics

Figure 7(b) illustrates the values assigned by both d_{tax} and d_{ps} to vertices in the taxonomy shown in Figure 7(a). In general, both metrics are able to assign values close to 0.0 to pairs of vertices separated by a small number of edges, and a value close to 1.0 to pairs of vertices separated by a large number of edges, e.g., (0, 1) and (8, 1). However, consider the pairs (10, 17) and (11, 12); d_{tax} is able to distinguish that both pairs have different taxonomic properties, i.e., the ratio $\frac{d_{tax}(10,17)}{d_{tax}(11,12)}$ is 0.6. Note that a value of 1.0 for this ratio implies that the taxonomic distances are judged to be similar. However, d_{ps} is not able to identify that these two pairs have different taxonomic properties. It assigns values such that the ratio $\frac{d_{ps}(10,17)}{d_{ps}(11,12)}$ is 0.75, i.e., closer to 1.0.

Next we report on the *distribution* of the pairwise distance d_{tax} and d_{ps} for several datasets (Table 3). We focus on the GO-BP terms and report on average and standard deviation. For **At_8** and **Ce_9**, the GO-BP terms in the intersection were more closely related compared to the individual datasets. The average distance for d_{tax} is also observed to be lower than the average for d_{ps}. Similarly, **Ar_37** had many pairs that were very close while there were also very distance pairs. We observe that the average is lower while there is also higher variance in the values.

To understand the discrimination capability of both metrics we "bucketize" pairs of GO-BP terms in U_1 and U_2 based on the length of the shortest path

	Path Length	#Pairs	d_{tax}		d_{ps}	
			average	std_dev	average	std_dev
At_8 ∪ Ce_9	1	51	0.15	0.09	0.31	0.12
	2	396	0.79	0.35	0.85	0.27
	3	2217	0.95	0.18	0.97	0.11
	4	4527	0.98	0.11	0.99	0.06
	5	1850	0.99	0.07	1.00	0.03
	6	254	1.0	0.0	1.0	0.0
	7	21	1.0	0.0	1.0	0.0
At_37 ∪ Ce_53	1	7	0.13	0.09	0.26	0.17
	2	63	0.74	0.35	0.84	0.23
	3	210	0.97	0.14	0.99	0.07
	4	372	0.98	0.10	0.99	0.04
	5	304	1.0	0.04	1.0	0.02
	6	103	1.0	0.03	1.0	0.02
	7	21	1.0	0.0	1.0	0.0

Table 4. Taxonomic Distance(s) between GO-BP for **At_8** ∪ **Ce_9** and **At_37** ∪ **Ce_53** (bucketized by path length).

between them. Table 4 reports on the number of pairs, and the average and standard deviation for both metrics.

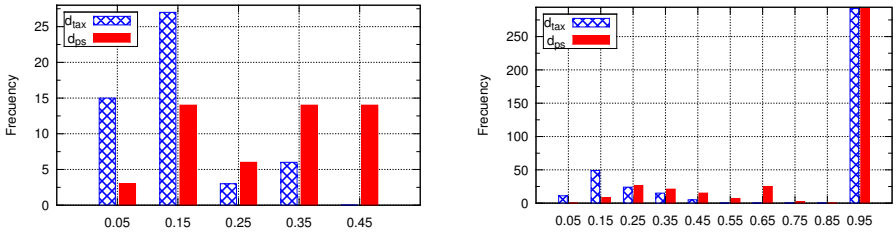
We can observe that the d_{tax} average for path length = 1 and 2 are 0.15 and 0.79 whereas the values for d_{ps} are 0.31 and 0.85, respectively. This reflects that d_{tax} is more sensitive to path length, and is able to discriminate better than d_{ps}, when vertices are connected by a small number of edges. However, for distant vertices both metrics exhibit similar behavior.

Finally, we report on the *distribution* of values for d_{tax} and d_{ps}, for GO-BP terms in U_1, for path length = 1 and 2, in Figure 8. For path length = 1, 15 pairs have a value of 0.05 and 30 pairs have a value of 0.15, for d_{tax}. In contrast, 3 pairs have a value of 0.05 and 14 pairs have a value of 0.15, for d_{ps}.

To summarize, d_{tax} appears to be more sensitive in capturing the range of (dis)similarity or distance between pairs of terms. In contrast, d_{ps} appears to compress the distance distribution. Thus, d_{tax} appears to be more useful in differentiating closer pairs from more distant pairs.

5 Summary and Conclusions

We present a tool, PAnG (Patterns in Annotation Graphs), that is based on a complementary methodology of graph summarization and dense subgraphs. Our collaborators (Heven Sze an Arabidopsis specialist and Eric Haag who studies C. elegans) helped us to validate potential cross genome annotation patterns using gene families with shared function in the two organisms. We demonstrate that a proposed metric for taxonomic distance d_{tax} is better able to discriminate among pairs of GO terms. In future work, we plan large scale cross genome experiments



(a) Path Length = 1 (b) Path Length = 2

Fig. 8. Frequency Distributions of distance t_{tax} and d_{ps} for pairs of GO-BP terms in dataset **At_8** ∪ **Ce_9** with Path Length = 1 and 2.

and human validation of both the annotation patterns and the relatedness of pairs of GO terms. Further, we plan to study the properties of the proposed distance metric in ontologies as MeSH [6] and Plant Ontology [7].

References

1. Anderson, P., Thor, A., Benik, J., Raschid, L., Vidal, M.E.: Pang - finding patterns in annotation graphs. In: Proceedings of the ACM Conference on the Management of Data (SIGMOD) (2012)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: tool for the unification of biology. NATGENET 25(1), 25–29 (May 2000)
3. Bender, M.A., Farach-Colton, M., Pemmasani, G., Skiena, S., Sumazin, P.: Lowest common ancestors in trees and directed acyclic graphs. Journal of Algorithms 57(2), 75–94 (2005)
4. Bock, K., Honys, D., Ward, J., Padmanaban, S., Nawrocki, E., Hirschi, K., Twell, D., Sze, H.: Integrating membrane transport with male gametophyte development and function through transcriptomics. Plant Physiology 140(4), 1151–1168 (2006)
5. Charikar, M.: Greedy approximation algorithms for finding dense components in a graph. In: APPROX. pp. 84–95 (2000)
6. Garcia-Hernandez, M., Berardini, T.Z., Chen, G., Crist, D., Doyle, A., Huala, E., Knee, E., Lambrecht, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Rhee, S.Y., Scholl, R., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J., Zhang, P.: TAIR: a resource for integrated Arabidopsis data. Functional and Integrative Genomics 2(6), 239 (2002)
7. Gene Ontology Consortium: The gene ontology project in 2008. Nucleic Acids Res. 36(Database Issue), D440–D444 (2008)

[6] `http://www.nlm.nih.gov/mesh/`
[7] `http://www.plantontology.org/`

8. Goldberg, A.V.: Finding a maximum density subgraph. Tech. Rep. UCB/CSD-84-171, EECS Department, University of California, Berkeley (1984), http://www.eecs.berkeley.edu/Pubs/TechRpts/1984/5956.html
9. Homologene. http://www.ncbi.nlm.nih.gov/homologene
10. Inparanoid. http://inparanoid.sbc.su.se/cgi-bin/index.cgi
11. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. CoRR cmp-lg/9709008 (1997)
12. Khuller, S., Saha, B.: On Finding Dense Subgraphs. In: International Colloquium on Automata, Languages and Programming (ICALP). pp. 597–608 (2009)
13. Lawler, E.: Combinatorial optimization - networks and matroids. Holt, Rinehart and Winston, New York (1976)
14. Lin, D.: An information-theoretic definition of similarity. In: ICML. pp. 296–304 (1998)
15. Navlakha, S., Rastogi, R., Shrivastava, N.: Graph summarization with bounded error. In: Proc. of Conference on Management of Data (SIGMOD) (2008)
16. Pekar, V., Staab, S.: Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In: COLING (2002)
17. Pesquita, C., Faria, D., Falcão, A., Lord, P., Couto, F.: Semantic similarity in biomedical ontologies. PLoS Computational Biology 5(7), e1000443 (2009)
18. Inparanoid. http://bioinformatics.psb.ugent.be/plaza/
19. Reiser, L., Rhee, S.Y.: Using The Arabidopsis Information Resource (TAIR) to Find Information About Arabidopsis Genes. Current Protocols in Bioinformatics, JWS (2005)
20. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: IJCAI. pp. 448–453 (1995)
21. Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J., Zhang, P.: The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to arabidopsis biology, research materials and community. NUCLEICACIDSRES 31(1), 224–228 (1 January 2003)
22. Saha, B., Hoch, A., Khuller, S., Raschid, L., Zhang, X.N.: Dense subgraphs with restrictions and applications to gene annotation graphs. In: Conference on Research on Computational Molecular Biology (RECOMB) (2010)
23. Sze, H., Chang, C., Raschid, L.: Go and po annotations for cation/h+ exchangers. Personal Communication (2011)
24. Sze, H., Padmanaban, S., Cellier, F., Honys, D., Cheng, N., Bock, K., Conejero, G., Li, X., Twell, D., Ward, J., Hirschi, K.: Expression pattern of a novel gene family, atchx, highlights their potential roles in osmotic adjustment and k+ homeostasis in pollen biology. Plant Physiology 1(136), 2532–2547 (2004)
25. List of arabidopsis thaliana transporter genes on sze lab page. http://www.clfs.umd.edu/CBMG/faculty/sze/lab/AtTransporters.html
26. The Plant Ontology Consortium: The plant ontology consortium and plant ontologies. Comparative and Functional Genomics 3(2), 137–142 (2002), http://dx.doi.org/10.1002/cfg.154
27. Thor, A., Anderson, P., Raschid, L., Navlakha, S., Saha, B., Khuller, S., Zhang, X.N.: Link prediction for annotation graphs using graph summarization. In: Proceedings of the International Semantic Web Conference (2011)
28. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.F.: A new method to measure the semantic similarity of go terms. Bioinformatics 23(10), 1274–1281 (2007)