Exposé for the diploma thesis: Exploiting Links between Biological Ontologies

Philipp Hussels
Chair for Knowledge Management in Bioinformatics
Department of Computer Science
Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany
hussels@informatik.hu-berlin.de

Advisor: Silke Trissl, Prof. Ulf Leser

1 Introduction

Entities in many public biological data bases are annotated with terms from agreed-upon vocabularies. Usually these vocabularies are not plain lists of terms, but ontologies containing concepts and the relationships between them. Since most biological ontologies focus on specific sub-domains of biological research, entities such as proteins are frequently annotated with terms from different ontologies. Thus, one biological entry may relate concepts from multiple ontologies. By exploring these links researchers might gain useful information and even reveal unknown interrelations.

Consider for example the Arabidopsis Information Resource (TAIR) [6], where entries are annotated with terms from the Gene Ontology [1] as well as the Plant Ontology [2]. GO terms describe the molecular function of a gene product, the biological process it is involved in, or the cellular component of it's action. PO terms in contrast describe structure or developmental stage of plants. While most existing approaches towards ontology mapping focus on alignment, i.e. finding near synonym relationships between ontologies, exploiting ontology links allows to find mappings of arbitrary meaning. From a link contained in TAIR a scientist might infer that a certain biological process is located in a specific part of a plant or regulates the growth of this part the plant. Gaining knowledge from ontology links is not trivial though. An entry in TAIR might be annotated with several terms from GO and PO, resulting in a large amount of concept pairs to explore. Since annotations may describe different characteristics of an entity, many of these links might not be particularly meaningful.

The first step towards knowledge inference from ontology links is to identify those pairs of linked concepts that most likely represent some biological fact. This study aims at defining a measure for estimating the meaningfulness of inter-ontology links and the application of this measure in an algorithm that computes the top-k concept pairs given a set of annotated biological entities.

2 Problem Description

To illustrate the problem consider for instance the TAIR entry for gibberellin 3 β-hydroxylase as shown in Figure 1. This entry is annotated with eight terms from the Gene Ontology and 19 terms from the Plant Ontology, immediately indicating 152 potential associations. Even more concept pairs can be generated from this link through ontology inference. GO and PO like many other ontologies can be viewed as hierarchical graphs. Through semantic subsumption introduced by is-a type relationships an entity annotated with a certain concept is implicitly annotated with all ancestor concepts in the hierarchy. This multiplies the number of concept pairs derivable from a linking entry. Furthermore biologists usually select sets of entries from a data source like TAIR. Although proteins in this set presumably share many annotations every single protein might contribute unique concept pairs. Exploring the resulting amount of con-

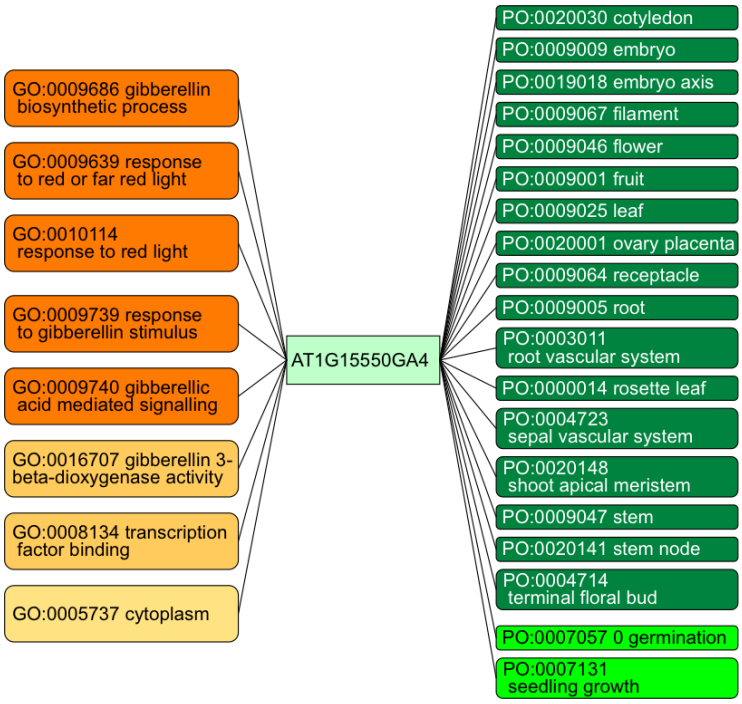

Figure 1: Terms from Gene Ontology and Plant Ontology linked through the TAIR entry for gibberellin 3 β-hydroxylase.

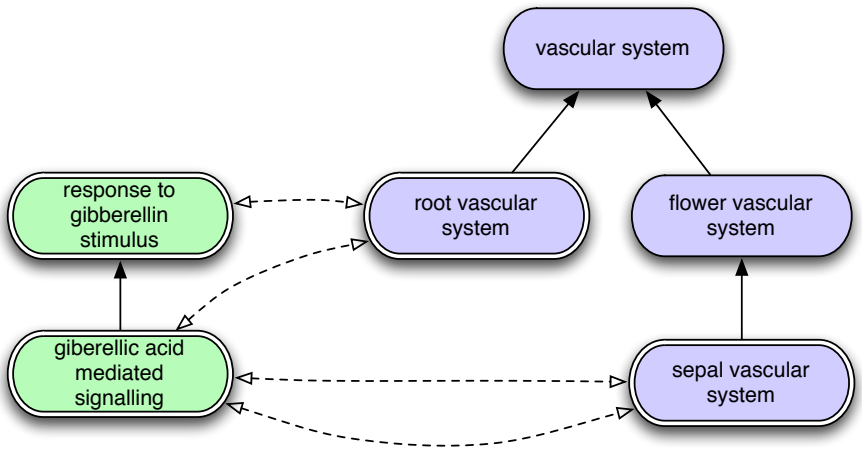

Figure 2: Mapping between subgraphs selected from GO and PO by terms annotated to gibberellin 3 β-hydroxylase

cept pairs manually in order to gain information is not feasible for a scientist. However, not all concept pairs - even if representing meaningful associations - are of interest to researchers.

Figure 2 shows sub graphs from GO and PO. Double border concept nodes are explicitly annotated to the above mentioned TAIR entry. As indicated with dotted lines, four term pairs could be generated from the chosen subset. But all these pairs represent the same biological fact, namely that gibberelic acid, a growth regulator in vascular plants, is distributed through the vascular system. It would thus be desirable to relate the sub graphs through a single - the 'best' - concept pair. The intuitive choice is to use the least general super concepts of both structures resulting in the concept pair *('response to gibberellin stimulus', 'vascular system')*. This pair actually represents the same biological fact and is easier to interpret than the cross product of - more specific - annotated terms. Replacing multiple annotations with common super concepts, i.e., choosing representative concepts for graph structures is a feasible approach. Finding appropriate rooted subtrees is not trivial however. Consider the sub graph from PO in Figure 3. This structure can not be represented by a single concept. Subsuming the two root concepts *'embryo'* and *'vascular system'* we would end up with a concept too general to be regarded valuable information. If furthermore the term *'cotyledon vascular system'* was actually annotated to the sample TAIR entry, the concepts *'embryo'* and *'vascular system'* would no longer adequately represent the sub graph as a whole. The information that both concepts are related regarding this entity would be lost. Developing a metric to reflect the balance between generalization and information loss will be the first part of this thesis.

The second step after condensing annotated ontology data is to find the 'best' matching pairs of representative concepts. Consider again Figure 1. The term pairs *('gibberellic acid mediated signalling','vascular system')* and *('gibberellic acid mediated signalling','germination')* represent meaningful associations as gibberelic acid *regulates* seedling growth and *is_distributed_through* the vascular system. A similar relationship exists for the term pair *('response to red light','germination')* as red light *stimulates* germination. No binary relationship however can be inferred from the term pair *('response to red light','vascular system')* for example. Primary objective of this study is to derive meaningful relationships between concepts from data source statistics and ontology information. Obviously concepts linked through many entities in the underlying data source are likely to be semantically related. This is particularly true if the number of links significantly exceeds the expectancy based on random distribution. However, such pairs could be considered common knowledge and are thus unlikely to reveal unknown interrelations. More sophisticated heuristics are required to distinguish rare but meaningful from false associations. Within the scope of this thesis I will develop a scoring scheme to assess the meaningfulness of concept pairs based on the scores assigned to representative concepts, data source statistics and ontology information.

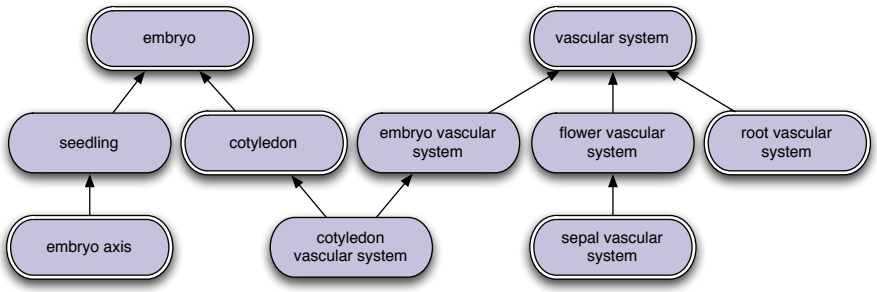

Figure 3: Sub graph from the plant ontology. Double border nodes are annotated to the TAIR entry for gibberellin 3 β-hydroxylase

3 Approach

As implicitly described in Section 2 finding the top-k concept pairs regarding a set of entities from a biological data source can be separated into two subsequent steps, namely finding representative concepts within ontologies and pairing representatives from different ontologies in an appropriate manner. Both steps require preprocessing. The following sub sections elaborate all three steps including solution statements to be evaluated and refined during the course of this thesis.

3.1 Identifying representative concepts

The annotations of biological entities select sub graphs from ontologies; consisting of the annotated concepts themselves and their node-to-root paths. The first major goal of this thesis is to develop a method to identify (preferably small) subsets of concepts that adequately represent subgraphs within ontologies. To achieve this goal an evaluation metric for representative concepts needs to be defined. For the sake of simplicity this metric will be designed assuming a single linking entity. In a second step set orientation will be achieved by either generalization or aggregation. The following paragraphs describe two possible solutions for the single link case.

3.1.1 Generic Metric

A simple approach to measure a concept's suitability as representative is to count the descendent concepts present in annotations. Intuitively, every such descendant provides evidence that the concept in question characterizes the annotated entity. Simple counting however results in overgeneralization. Any common ancestor of the concepts at hand - and in particular the ontology's root concept - would be assigned the maximum score. Introducing a decreasing factor $0 < \epsilon < 1$ to penalize every step of abstraction solves this problem in a straightforward manner. Instead of counting ancestor nodes equally each node A is assigned a value $e = \epsilon^d$, where d is the minimum number of edges traversed from A to the representative concept.

3.1.2 Metric based on Information Content

The generic metric described in Subsection 3.1.1 uses a static factor to penalize subsumption. The implicit assumption underlying this approach is that path length between concept nodes accurately reflects semantic distance. But, as explained in Section 1, biological ontologies conceptualize different sub domains of biological research. Core concepts of such domains are modeled in more detail, i.e., edges between descendant concepts are 'shorter' than for concepts included for the sake of completeness. More important, the actual distance between concepts is a matter of context. Quantitive information on semantic distance can be derived from data source statistics. In [3] the authors give an overview of different frequency-based semantic similarity measures. The basic idea underlying all these measures is that, given a data source, the information content of a concept is inversely related to it's number of occurrences in annotations. As sub concepts logically imply their super concepts, differences in information content reflect semantic differences between ancestor and descendant concepts. One objective of this study is to apply a frequency-based distance measure to improve the metric described in Subsection 3.1.1.

3.2 Relating representative concepts

The second major goal of this thesis is to find those pairs of linked representative concepts that semantically relate to each other. Again, data source statistic can be utilized to mine the strength of relationships. As elaborated in Section 2 frequently co-occurring concepts likely represent semantic relationships. Furthermore ontology information can be utilized to avoid loosing rare but meaningful and thus particularly interesting pairs. Besides considering 'close relatives' e.g. siblings of meaningful related concepts, domain specific relationships might be leveraged to extend the coverage. In the second part of my thesis I will develop a method to evaluate the strength of relationships between linked representative concepts based on the score assigned to these concepts, ontology information and data source statistics.

3.3 Pre-computations

The major focus of this thesis is to find the most meaningful, i.e., the top-k relationships between concepts of different ontologies, given a set of linking entities. Instead of computing and ranking all possible concept pairs, we intend to consider only the most promising concepts for pairing. A crucial step in finding representative concepts is to identify the subgraphs selected from an ontology by annotations. Assuming every ontology graph is stored in two relational tables *nodes(id, term)* and *edges(node_id, parent_id)* this task can be dramatically simplified by pre-calculating the transitive closure *tc(node_id, ancestor_id, distance)*. This way candidate concept scores as defined by the metric described in Subsection 3.1.1 can be calculated using the following pseudo SQL statement: *SELECT ancestor_id, sum(*$0.5^{distance}$*) FROM tc WHERE node_id IN (annotations) GROUP BY ancestor_id*. As ontologies are small and grow moderately compared to other graph data sources (like biological pathway databases for example) pre-calculating the transitive closures is non-critical concerning both runtime and memory/disk-consumption.

Besides ontology information data source statistics are needed both to compute the semantic distance between concepts within an ontology and to measure the strength of relationships between concepts in different ontologies. As described in Subsection 3.1.2 the semantic distance between ancestor and dependent concepts depends on the frequencies of annotations, i.e., the number of occurrences of both concepts in annotations. In a relational database this information can be gathered by a single scan on the table that links data source entries (fact table) to ontology concepts. Measuring the strength of inter-ontology relationships additionally requires concept pair frequencies to be pre-calculated. Although pairing concepts requires joins between the fact table, link tables and ontology tables, this operation is not expected to be performance critical and can be expressed in a single SQL statement.

4 Evaluation

Evaluating the performance of the top-k concept pair algorithm requires manual assessment of sample data, as to the best of my knowledge no gold standard exists for relating concepts through annotated data. While in [5] and [4] the authors present a methodology to calculate support and confidence scores for concept pairs associated through chains of linked entities, their approach only considers annotated concepts and their parent nodes. Ignoring further ancestor nodes this method does not qualify as benchmark. Manually created global ontology mappings, such as those maintained by the Open Biomedical Ontology Foundry [7], can be explored to find the concept pairs relevant to a set of entities. A concept pair is relevant if the mapped concepts are linked through at least one entity in the data set at hand. Note that ancestors of explicitly linked concepts have to be considered as well. Regarding manually curated mappings as ground truth we can calculate sensitivity and selectivity measures to evaluate the performance of the algorithm.

References

[1] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25:25–29, May 2000.

[2] S. Avraham, C.W. Tung, K. Ilic, P. Jaiswal, E.A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S.Y. Rhee, M.M. Sachs, M. Schaeffer, L. Stein, P. Stevens, L. Vincent, F. Zapata, and D. Ware. The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res.*, 36:D449–454, Jan 2008.

[3] Francisco M. Couto, Mario J. Silva, and Pedro M. Coutinho. Measuring semantic similarity between gene ontology terms. *Data & Knowledge Engineering*, 61(1):137–152, April 2007.

[4] Woei-Jyh Lee, Louiqa Raschid, Hassan Sayyadi, and Padmini Srinivasan. Exploiting ontology structure and patterns of annotation to mine significant associations between pairs of controlled vocabulary terms. In Amos Bairoch, Sarah Cohen Boulakia, and Christine Froidevaux, editors, *DILS*, volume 5109 of *Lecture Notes in Computer Science*, pages 44–60. Springer, 2008.

[5] Woei-Jyh Lee, Louiqa Raschid, Padmini Srinivasan, Nigam Shah, Daniel L. Rubin, and Natasha Fridman Noy. Using annotations from controlled vocabularies to find meaningful associations. In Sarah Cohen Boulakia and Val Tannen, editors, *DILS*, volume 4544 of *Lecture Notes in Computer Science*, pages 247–263. Springer, 2007.

[6] S.Y. Rhee, W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L.A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D.C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, 31:224–228, Jan 2003.

[7] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255, November 2007.