Proceedings of the First International Symposium on Semantic Mining in Biomedicine (SMBM-2005)
(electronic proceedings: http://CEUR-WS/Vol-148/)

An ontology of chemical entities helps identify dependence relations among Gene Ontology terms

Anita Burgun [1], Olivier Bodenreider [2]

[1] EA 3888, Faculté de Médecine, IFR 140, Université de Rennes I, France
[2] US National Library of Medicine, Bethesda, Maryland, USA
Anita.Burgun@univ-rennes1.fr

The Gene Ontology (GO) is organized in three allegedly independent hierarchies: molecular functions, biological processes, and cellular components. In this paper, we present an approach based on the Chemical Entities of Biological Interest (ChEBI) ontology to identifying dependence relations in GO, especially relations across hierarchies. Our method is based on the identification of the names of ChEBI entities in GO terms. We distinguish between first-order dependence relations between GO terms that share a common chemical name, and second-order dependence relations between GO entities whose names include two chemicals that are hierarchically related. Of the 10,516 entities in ChEBI, 26% were identified in the names of 9,431 GO terms (55% of all GO terms). A total of 771,302 pairs of related GO terms (first-order associations) were computed. Of these, 44% correspond to dependence relations across hierarchies. These results were compared to the 8,714 pairs of GO terms identified as dependent by lexical and statistical methods in a previous study (once restricted to GO terms whose names include a ChEBI entity). Of these, 3,932 (45%) were identified as first-order relations, and 937 (11%) as second-order relations. We show that the two kinds of approaches are complementary. The ChEBI-based is independent of the annotations, allowing even rare dependencies to be identified. Moreover, it takes advantage of the subsumption relations between chemicals in ChEBI, and therefore helps identify second-order dependence relations. This approach can be generalized to other ontologies of chemicals as well as other kinds of ontologies.

INTRODUCTION

Dependence relations are generally not well represented in ontologies, particularly in the Gene Ontology™ (GO) [1] where relations across the three hierarchies (Molecular Function, Biological Process, Cellular Component) are not represented at all. In practice, the annotation of a gene product to one GO hierarchy is independent of its annotation to other hierarchies insofar as the curators of model organism annotation databases are not required to annotate with a term from one hierarchy a gene product annotated with a related term from another hierarchy. However, there exist biological entities that depend on other ones. For example, an implicit dependence relation holds between the molecular function ferric ion binding and the biological process iron homeostasis because the function of binding iron is involved in the process of maintaining a stable concentration of iron at the cell or organism level. As a consequence, if a given gene product is annotated by ferric ion binding, it should most likely also be annotated by iron homeostasis or one of its descendants. Explicit dependence relations are useful for ontology maintenance. Changes made to a given concept should trigger the review of and possibly changes to all concepts to which it has dependence relations. Because they can be used to alert curators to the existence of related concepts, dependence relation would also help produce complete and consistent gene annotation in model organism databases.

Several approaches have been used to identifying and analyzing dependence relations among GO terms. They are based on lexical, statistical, and ontological methods. Ogren et al. have developed a **lexical approach** exploiting the compositional properties of GO terms [2]. They found that 65% of all GO terms contain another GO term as a proper substring. For ex-ample, the molecular function electron transporter activity includes in its name the biological process electron transport. **Ontological approaches** rely on formal ontological principles to formalize the relations expected between biological entities according to general theories specified in upper-level ontologies. For example, the biological process provirus integration is dependent on the cellular component provirus because ontologically processes are dependent on the substances on which they act (i.e., there can be no provirus integration unless there exists a provirus to be integrated). This approach was used by Kumar et al. to analyze de-pendence relations identified by other methods. **Statistical approaches** take

advantage of the knowledge represented in the model organism annotation databases. In the Gene Ontology Annotation Tool project (GOAT) Bada et al. mined the annotation database Gene Ontology Annotation (GOA) for co-occurrence of GO terms. 600,000 associations were obtained by this method, excluding unreliable associations and the hierarchical relations explicitly represented in GO [3]. Kumar applied association rule mining techniques to the TIGR database [4]. In a previous study [5], we combined three approaches: computing similarity in a vector space model, statistical analysis of co-occurrence of GO terms in annotation databases, and association rule mining. We applied them to five annotation databases. A total of 7,665 associations were identified by at least one of these approaches. We compared them to 5,493 lexical relations among GO terms, and we found that only 180[1] associations were identified by both non-lexical and lexical methods. The limited overlap between associations identified by non-lexical and lexical approaches was somewhat unexpected and suggests that the different ap-proaches may complement each other.

The goal of this study is to investigate how an ontology external to GO can help identify and analyze relations among GO terms. Our hypothesis is that two GO terms whose names include the name of a given chemical entity are in dependence relation. For example, the function of a molecule transporting potassium (e.g., potassium ion transporter activity), the cellular component involved in potassium transport (e.g., potassium ion-transporting ATPase complex) and the biological process of potassium transport (e.g., potassium ion transport) have in common that their names include that of the ion being transported (potassium). From an ontological perspective, this biological process, this cellular component and this molecular function are all dependent on the chemical entity as no one would exist without the potassium existing. Additionally, the biological process is dependent on the molecular function because any changes in the function would influence the process; and the function and the process are both dependent on the cellular component. The objective of this study is precisely to identify pairs of GO terms related by this kind of dependence relations. Furthermore, there are cases in which two GO terms include the names of chemical entities that are not identical, but rather stand in a hierarchical relation. For example, the molecular function cation channel activity involves cation and the biological process potassium ion transport involves potassium. We show that an ontology of chemicals in which potassium is represented as a kind of cation contributes to identify this kind of dependence relations automatically.

[1] In [5], we reported 230 associations identified by both non-lexical and evaluation methods. Here we have re-stricted evaluation methods to the sole lexical approach and the overlap between lexical and non-lexical ap-proaches has been reduced to 180 terms.

MATERIALS

Gene Ontology.

The Gene Ontology (GO) is a controlled vocabulary developed by the Gene Ontology Consortium for the annotation of gene products in model organisms. GO names were extracted from the OBO file downloaded on December 29, 2004 comprising 17,250 GO terms. Both preferred names (*name* field) and synonyms (*exact_synonym* field) are used in this study. A total of 22,525 names were extracted from the file (5,275 synonyms in addition to one preferred name for each GO term). GO is organized in three separate hierarchies for molecular functions (9,180 terms), biological processes (11,558 terms) and cellular components (1,787 terms). In each GO hierarchy, an entity may have more than one parent. For example, names for the molecular function identified by GO:0005249 include the preferred name *voltage-gated potassium channel activity* and three synonyms: *voltage gated potassium channel activity*, *voltage-dependent potassium channel activity* and *voltage-sensitive potassium channel*. *Voltage-gated potassium channel activity*; this entity has two parents in GO: *potassium channel activity* [GO:0005267] and *voltage-gated ion channel activity* [GO:0005244].

ChEBI

The Chemical Entities of Biological Interest (ChEBI) is "a freely available dictionary of 'small molecular entities' (i.e., atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc.); ChEBI entities are either products of nature or synthetic products used to intervene in the processes of living organisms." ChEBI is developed at the European Bioinformatics Institute (EBI). ChEBI names were extracted from the OBO file dated December 22, 2004. Both preferred names (*name* field) and synonyms (*synonym* field) are used in this study. A total of 27,097 names were extracted from the file (13,709 synonyms in addition to one preferred name for each of the 10,516 entities). ChEBI entities are organized in a subsumption (*IS_A*) hierarchy where each entity may have more than one parent (polyhierarchy). 11,872 links (i.e., pairs of related entities) were extracted from the file. For example, names for the ChEBI entity identified by CHEBI:26216 include the preferred name *potassium* and two synonyms: *kalium* and *K*. The hierarchy where potassium is located includes one parent (*alkali metals*) and one child (*potassium(1+)*).

Dependence relations from the PSB study.

In a previous study [5], which we will refer to as the 'PSB study', we computed dependence relations among GO terms using various approaches. The first approach (baseline) consisted of identifying GO names present in other GO names as proper substring (e.g., the string *potassium ion transport* is included in the string *potassium ion transporter activity*). The other approaches are based on the association of GO terms in annotation databases and draw of various techniques including similarity in a vector space model, statistical analysis of co-occurrence and association rule mining. A total of 44,969 dependence relations (pairs of GO terms) have been identified during this study. Among them, 12,978[2] associations were dependence relations across GO hierarchies. Examples of associations identified by the various approaches are presented in Table 1.

Association		**VSM**	**COC**	**ARM**	**LEX**
MF:	*potassium channel activity* [GO:0005267]	**X**	**X**	**X**	
BP:	*potassium ion transport* [GO:0006813]				
MF:	*chemokine activity* [GO:0008009]		**X**	**X**	
BP:	*immune response* [GO:0006955]				
CC:	*hemoglobin complex* [GO:0005833]	**X**	**X**		
BP:	*oxygen transport* [GO:0015671]				
MF:	*taste receptor activity* [GO:0008527]	**X**		**X**	
BP:	*perception of taste* [GO:0050909]				
MF:	*metal ion transporter activity* [GO:0046873]	**X**		**X**	**X**
BP:	*metal ion transport* [GO:0030001]				

Table 1 – Examples of associations identified by similarity in a vector space model (VSM), analysis of co-occurring GO terms (COC), association rule mining (ARM) and lexical methods (LEX) in the PSB study

[2] In [5], we reported 13,398 associations across GO hierarchies identified by non-lexical or evaluation methods. Here we have restricted evaluation methods to the sole lexical approach and the total number of associations identified has been reduced to 12,978 terms.

METHODS

The methods of this study can be summarized as follows. First, we identify names of ChEBI entities included in GO terms, creating a bipartite graph including ChEBI and GO. We then use these relations between a given ChEBI entity and the GO terms in which it is included to compute co-occurrence relations among the GO terms in this set, called first-order associations. Using the transitive closure on ChEBI subsumption relations, we identify second-order associations among GO terms, where the ChEBI entities included in GO names are not the same, but stand in a hierarchical relation. Finally, we evaluate pairs of co-occurring GO terms obtained to the pairs of GO terms from the PSB study.

Identifying ChEBI entities in GO.

As noted by Ogren [2], the names of many GO terms include names of other GO terms as a proper substring. Analogously, the names of ChEBI entities are part of many GO terms. For example, the entity potassium is present in 43 GO terms including potassium-uptake-ATPase activity and regulation of potassium transport. Every ChEBI name is searched for in every GO name. ChEBI names of less than three characters are ignored. These names often correspond to chemical symbols (e.g., K, symbol of potassium) and may be ambiguous with words in English (e.g., As – symbol of arsenic – and the preposition as). As the names of ChEBI en-tities may be capitalized, the comparison between ChEBI and GO strings is rendered cased-insensitive. In order to avoid infelicitous matches, the name of a ChEBI entity is required to be not simply a substring, but a lexical item. In practice, the characters surrounding the name of the ChEBI entity in a GO name must be word boundaries (i.e., space, hyphen, punctuation, etc.). For example, the ChEBI entity carbon is identified in the GO name carbon-oxygen lyase activity, but not in carbonic anhydrase activity. Finally, we performed a limited normalization of the ChEBI names, principally to allow the names of classes of entities – often in plural form (e.g., cations, acids, esters, nitrates, etc.) to match names of entities derived from these classes, often present in singular form as in GO names. In practice, we complemented the list of synonyms provided by ChEBI by adding, if necessary, the singular form for the name of a plural class (e.g., ester for esters). 2,872 such synonyms were added to ChEBI[3] .

[3] As we simply removed the trailing *s* from ChEBI names, some inaccurate names were generated (e.g., *phosphoru* and *mustard ga*). Such incomplete names will not mach any lexical items in GO names and, beside slowing down slightly the matching process,

Identifying sets of GO terms related to a given ChEBI entity

First order relations. For each ChEBI entity, we computed the set of GO terms whose names include one of the names for this entity. Then, we computed the associations between each pair of GO terms in the set. All GO terms in a set have the property of being linked through their names to the same ChEBI entity (Fig. 1). For this reason, we call these associations among GO terms *first-order associations*. For example, the ChEBI entity *uronic acid* (CHEBI:27252) is identified in three GO terms. The set of GO terms related to this entity is shown in Table 2.

BP:uronic acid metabolism	[GO:0006063]
MF:uronic acid transporter activity	[GO:0015133]
BP:uronic acid transport	[GO:0015735]

Table 2 – Set of GO terms related to the ChEBI entity uronic acid

The following three pairs of GO terms are computed from the set:

- GO:0006063-GO:0015133
- GO:0006063-GO:0015735
- GO:0015133-GO:0015735

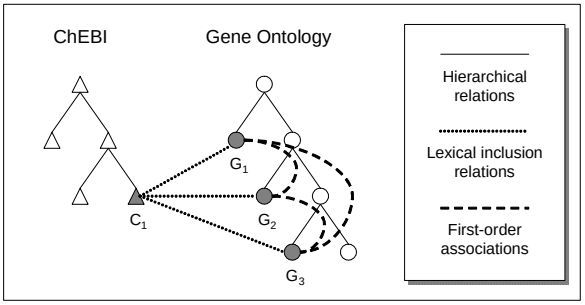

Figure 1 –First-order associations

Second order relations. The transitive closure of ChEBI subsumption relations was computed using Warshall's algorithm. In the resulting structure, a given entity is related not only to its direct parents as it is the case in the original ChEBI file, but to each of its ancestors all the way up to the root of ChEBI hierarchy.

Unlike first-order associations where GO terms share a given ChEBI entity, we define second-order associations among GO terms associations in which the ChEBI entities included in the GO terms stand in an ancestor-descendant relation in the sense of the transitive closure on ChEBI relations presented above (Fig. 2). For example, the molecular function cation channel activity [GO:0005261] and the biological process potassium ion transport [GO:0006813] would not qualify for a first-order association. However, as cations [CHEBI:23058] subsumes potassium ion [CHEBI:29103] in ChEBI (CHEBI: 29103 is_a CHEBI:25414[4] is_a CHEBI:23058), there is a second-order association between cation channel activity and potassium ion transport.

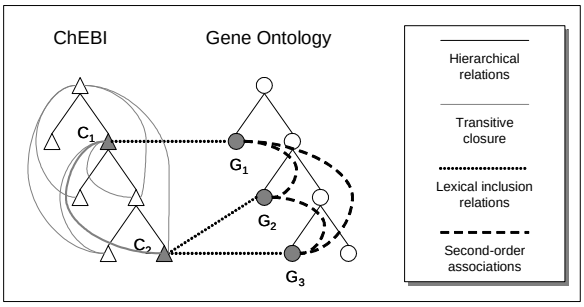

Figure 2 –Second-order associations

Evaluation

In order to evaluate the pairs of GO terms, we compared the pairs of dependent GO terms obtained using lexical and statistical methods in the PSB study to the pairs of GO terms shar-ing a common ChEBI entity in their names. None of the studies takes in consideration the direction of the dependence relation. Thus the comparison simply consists of creating the intersection of the sets of pairs obtained in each study. In the comparison, we distinguish between first- and second-order associations.

RESULTS

Identifying ChEBI entities in GO

Of the 10,516 entities in ChEBI, 2,700 (26%) were identified in the names of 9,431 GO terms. In other words, 55% of the 17,250 GO terms include in their names the name of some ChEBI entity. These name inclusion relations resulted in 20,497 associations between a ChEBI entity and a GO term.

Identifying sets of GO terms related to a given ChEBI entity

One set of GO terms was created for every ChEBI entity whose name is included in GO names, resulting in 2,700 sets. The cardinality of these sets ranges from 1 (1,096 such singletons, e.g, for *2-Chloro-4,6-dihydroxy-1,3,5-triazine*) to 590 (for *acids*). Other examples of large sets are related to entities such as

this overgeneration has no detrimental consequences on the identification of ChEBI entities in GO names.

[4] CHEBI:25414 monocations

phosphate, *DNA* and *Coenzyme A*. The median cardinality of the 1,604 non-singletons is 5. From these sets, 771,302 pairs of related GO terms (first-order associations) were computed. Of these pairs of GO terms, 340,527 (44%) correspond to dependence relations across hierarchies. Most of these relations (86%) hold between biological processes and molecular functions. For obvious reasons, we did not attempt to compute all second-order associations. Instead, we checked their existence as necessary. The transitive closure of ChEBI hierarchical relations yielded 79,835 relations (i.e., almost seven times as many as the 11,872 original hierarchical relations in ChEBI).

Evaluation

Of the 44,969 pairs of GO terms identified as dependent in the PSB study – including asso-ciations both within and across hierarchies, 8,714 correspond to GO terms whose names include a ChEBI entity. Of these, 4,869 (56%) were also identified as pairs of dependent GO terms in this study. 3,932 (45%) were identified as first-order relations in this study. Additionally, 937 (11%) were identified as second-order relations. In other words, 19% of the dependence relations (937/4,869) correspond to second-order relations.

Of the 771,302 pairs of GO terms identified as related by a first-order relation via ChEBI, less than 1% were also identified in the PSB study.

Examples of pairs of dependent GO terms identified by either methods and by both methods are given in Table 3.

Association		**PSB**	**ChEBI (1st order)**	**ChEBI (2nd order)**
MF:	*potassium channel activity* [GO:0005267]	**X**	**X**	
BP:	*potassium ion transport* [GO:0006813]			
MF:	*cation channel activity* [GO:0005261]	**X**		**X**
BP:	*potassium ion transport* [GO:0006813]			
CC:	*hemoglobin complex* [GO:0005833]	**X**		
BP:	*oxygen transport* [GO:0015671]			
MF:	*tRNA binding* [GO:0000049]		**X**	
BP:	*glycine-tRNA ligase comple* [GO:0009345]			
MF:	*carboxylesterase activity* [GO:0004091]			**X**
BP:	*lipoic acid metabolism* [GO:0000273]			

Table 3 – Examples of associations identified in the PSB study, using ChEBI (first- and second-order associations) or both

DISCUSSION

Advantages and limitations

By exploiting not only the terminological component of ChEBI (i.e., the names of chemical entities), but also its structure (i.e., the hierarchical relations among chemical entities), our method identifies second-order associations (pairs of GO terms whose names include chemical entities standing in a hierarchical relation) in addition to the first-order associations (pairs of GO terms whose tnames share a chemical entity). The contribution of subsumption in ChEBI to identifying dependence relations in GO corresponds to 19% of the relations (proportion of second-order relations in this study).

Although no systematic evaluation of the dependence relations obtained has been performed, we noted the presence of false positives inherent to lexical approaches. For example, a synonym for electron in ChEBI is beta (for beta-particle), which is present in many GO terms with a different meaning (e.g., beta-amyloid binding). These errors may result in overgenerating a large number of relations due to the combinatorial process of creating co-occurrences. For example, only 46 of the 353 GO terms linked to electron actually include electron, resulting in some 47,000 inaccurate relations (10% of the total). By imposing constraints on the matching of ChEBI names in GO names, we tried to limit erroneous matches. The constraint on lexical items (matching is limited to complete lexical items rather than substrings) prevented, for example, the chemical name cation [CHEBI:23058]. from being erroneously matched to the biological process DNA replication [GO:0006260]. While preventing erroneous matches, these limitations also prevent valid matches from being identified, corresponding to false negatives. For example, because of constraints on lexical items, the chemical name imidazolone [CHEBI:27850] was not matched to the molecular function imidazolonepropionase activity [GO:0050480]. The threshold of three characters for the minimum length of ChEBI strings to be searched for in GO, was selected as a trade-off between false positives and missed matches. A threshold of four would, for example, eliminate matches involving CoA, present in 225 GO terms. Moreover, normalization applied to the ChEBI names is limited and only allows the plural names of chemicals to match their singular form in GO terms.

Finally, the primary relations identified by this method are relations between ChEBI entities and GO terms. These relations are generally participation relations [4] and the exact nature of the relation is most often easy to determine. Computed from these primary relations are the dependence relations among GO terms. The nature of relationship linking GO terms related to the same chemical entity is often more difficult to determine automatically.

Complementarity among approaches

Statistical approaches to identifying dependence relations among GO terms rely on the knowledge represented in the annotation databases. However, only about 30% of all GO terms are used in 70% of the annotations. For this reason, the frequency of many valid associations represented in the annotation databases may not be sufficient for them to be deemed significant by statistical techniques. With ontology-driven methods such as the one we presented, dependence relations are extracted regardless of their presence or frequency in annotation databases. Conversely, the approach presented in this paper presupposes the existence of external ontologies to link to while statistical approaches may be used even when no external ontologies are available.

This study confirmed the benefit of combining several approaches to identifying dependence relations in GO also when both annotation databases and external ontologies are available. An example of this complementarity is the ability of statistical methods to detect complex associations in biological pathways, for example, the dependence between hemoglobin complex and oxygen transport, while the ChEBI-based approach consistently and systematically identifies associations, for example among all GO terms involving tRNA. All approaches can be applied to enriching GO with relations both across hierarchies and within. For example, our ChEBI-based approach identified a second-order association between the two molecular functions carbohydrate transporter activity [GO:0015144] and maltose porter activity [GO:0015581].

While the limited overlap between approaches reflects their complementarity, it also limits the possibility of evaluating them against the each other. In theory, cross validation is possible when lexical, statistical and ontological methods are combined: dependence relations identified by several methods are expected to be valid. In practice, only few associations are captured by different methods therefore contribution of cross validation is limited. The relations identified require manual validation. The false positives identified in the review of a limited number would need to be filtered out prior to starting the manual validation.

Future directions and challenges

As mentioned by Wroe et al [6], who used an ontology derived from MeSH, this study confirmed the interest of using an ontology of chemicals to infer new relations between GO terms. A richer representation of chemical entities in ChEBI would help link to GO and identify additional dependence relations among GO terms. Not surprisingly for a resource that has been released less than one year ago, the coverage of ChEBI remains limited. Its content is only partially curated at this time, with many chemical entities that have not been classified yet. Meanwhile, since ChEBI records cross-reference to other chemical entity repositories (e.g., CAS registry number), additional information (e.g., synonyms) can be extracted from external resources (e.g., PubChem) also referencing these identifiers.

This study confirmed the necessity of formally linking molecular functions, cellular com-ponents and biological processes in GO to an ontology of chemical entities. The presence of ChEBI names in GO terms only represents an implicit link that must be formalized. A step further, the fact that GO entities of different kinds (e.g., a molecular function and a biological process) may be related to the same chemical entity, represents a dependence relation that must also be formalized.

More generally, as suggested by B. Smith [7], GO entities must be linked to entities in external ontologies such as cell types (e.g., alpha-beta T-cell activation) and organisms (e.g., light-harvesting complex (sensu Viridiplantae)). Our approach is not specific to chemical entities and could be applied to other external ontologies.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.
2 Ogren P.V., Cohen K.B., Acquaah-Mensah G.K., Eberlein J,. Hunter L. The compositional structure of Gene Ontology terms. Pac Symp Biocomput 2004, 214-25

3. Bada M., Turi D., McEntire R, Stevens R. Using reasoning to guide annotation with Gene Ontology terms in GOAT. SIGMOD Record 33(2004). http://www.acm.org/sigmod/record/issues/0406/04.Bada_Turi_McEntire_Stevens.pdf
4. Kumar A., Smith B, Borgelt C. Dependence relationships between Gene Ontology terms based on TIGR gene product annotations. Proceedings of the 3rd International Workshop on Computational Terminology (CompuTerm 2004), 31-38 (2004)
5. Bodenreider O, Aubry M, Burgun A. Non-lexical approaches to identifying associa-tive relations in the Gene Ontology. Pac Symp Biocomput 2005: 91-102
6. Wroe CJ, Stevens R, Goble CA, Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. Pac Symp Biocom-put 2003:624-35
7. Smith B, Williams J, Schulze-Kremer S. The ontology of the gene ontology. AMIA Annu Symp Proc. 2003;:609-13.