Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies 2005:17-18.

# *Linking the Gene Ontology to other biological ontologies*

*Olivier Bodenreider[1,*] and Anita Burgun[2]*

*[1] National Library of Medicine, Bethesda, USA*
*[2]EA 3888, IFR 140, Université Rennes 1, France*

**ABSTRACT**

The entities described in the Gene Ontology, (i.e., molecular functions, cellular components and biological processes), often make reference (in their names) to other entities, either from GO or from other ontologies, such as ontologies of chemical entities, cell types and organisms. We developed a method for mapping terms from the Open Biomedical Ontology (OBO) family to GO. We show that 55% of the 17,250 GO terms include in their names the name of some chemical entity (ChEBI). Our findings are consistent with that of other studies. Additionally, our study provides a quantification of the relations between GO terms and terms from other ontologies.

## 1 INTRODUCTION

Several approaches have been used to identifying relations among terms form the Gene Ontology (GO[1]) [1]. The lexical approach developed by Ogren et al. exploits the compositional properties of GO terms, i.e., GO terms nested within other GO terms [2]. They found that 65% of all GO terms contain another GO term as a proper substring. For example, the molecular function *electron transporter activity* includes in its name the biological process *electron transport*.

The goal of this study is slightly different: it is to investigate the degree to which GO terms are related to terms from ontologies external to GO. In particular, we are interested to make explicit the relations existing between GO terms and terms from other ontologies of the Open Biomedical Ontology (OBO) family[2]. OBO includes ontologies such as ChEBI (Chemical entities of biological interest), InterPro (protein families, domains and functional sites) and Plant ontology (plant structures and growth/developmental stages).

Related to this study is Obol [3], a language created for representing relations embedded in the names of GO entities, with the objective of facilitating the maintenance of the ontology. The work most closely related to ours is the GONG project [4], an attempt to convert GO into a description logics formalism. In addition to GO terms themselves, GONG also used entities from the KEGG database as a reference for the enzymes referenced in GO. The objective of this study is to generalize such cross-references (i.e., between GO and other ontologies) to all entities represented in OBO ontologies.

As suggested by Smith & al. [5], 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)*). In a previous study [6], we investigated the relations between GO and ChEBI. This paper proposes to generalize the method developed for ChEBI to other members of the OBO family.

## 2 LINKING GO TO CHEBI

The first phase of this project consisted to link GO terms to chemical entities from the Chemical Entities of Biological Interest (ChEBI). 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). For example, names for the ChEBI entity identified by CHEBI:26216 include the preferred name *potassium* and two synonyms: *kalium* and *K*.

### 2.1 Methods

Every ChEBI name is searched for in every GO name (Figure 1). 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 entities 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, punc-

* To whom correspondence should be addressed.

[1] http://www.geneontology.org/

[2] http://obo.sourceforge.net/

tuation, 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].

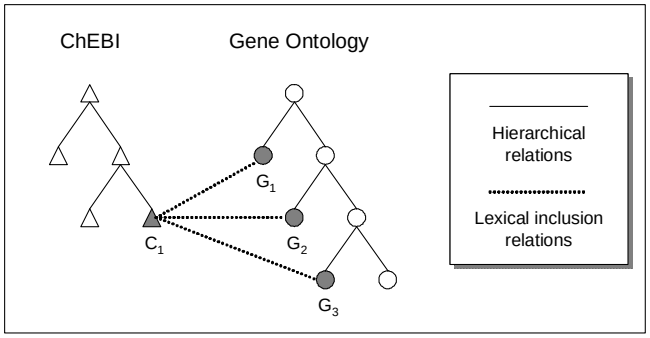


Figure 1 – Lexical inclusion relations between ChEBI terms and GO terms.

### 2.2 Results

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.

## 3 GENERALIZATION TO OTHER BIOLOGICAL ONTOLOGIES

In addition to updating the results of an earlier mapping between GO and ChEBI, this study proposes to apply the method developed for ChEBI to the other members of the OBO family. All terms from the 23 other ontologies (apart from GO and ChEBI) will be mapped to GO and quantitative results will be reported.

These results will contribute to quantifying the relations existing between entities in GO and in the other OBO ontologies. This work can be understood as a first step towards the generalization of Obol to the other OBO ontologies [4]. In addition, as shown in [6], such relations can be used to suggest dependence relations among GO terms.

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

## 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. 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 Biocomput 2003:624-35.
4. Mungall, C. Obol: integrating language and meaning in bio-ontologies. Comparative and Functional Genomics. 2004;5:6-7, 509-520.
5. Smith B, Williams J, Schulze-Kremer S. The ontology of the gene ontology. AMIA Annu Symp Proc. 2003;:609-13.
6. Burgun A, Bodenreider O. An ontology of chemical entities helps identify dependence relations among Gene Ontology terms. First Symposium on Semantic Mining in Biomedicine 2005:(in press).