Semantic representation of Gene Ontology terms by using Gene Regulation Ontology

Jung-jae Kim[1,*], Vivian Lee[2] and Dietrich Rebholz-Schuhmann[2]
[1]School of Computer Engineering, Nanyang Technological University, Singapore;
[2]European Bioinformatics Institute, Cambridge, U.K.

Abstract. Gene Ontology (GO) has been developed to provide concepts for the functional annotation of biological entities. This development has already contributed to significant biomedical research results. Nonetheless, GO could have provided even stronger support to biomedical text mining, if it delivered domain-independent logical definitions of its concepts.
We present a method that extracts the semantic structures of GO terms by using the Gene Regulation Ontology (GRO). The method annotates substrings of GO terms with, if any, corresponding concepts of OBO ontologies and then converts the syntactic structures of GO terms into GRO-based semantic structures. We show that the semantic structures can be used to deduce implied relations from GO terms.

1. Introduction

Gene Ontology (GO) provides a controlled vocabulary for describing the functions and properties of gene products. The big size and steady growth of the ontology leads us to the need of automated aids for maintenance tasks such as consistency checking (Verspoor et al., 2009) and functional annotation (Camon et al., 2005). One of the obstacles in the maintenance of GO is that GO terms are long and have complex syntactical structures (Ogren et al., 2005). One solution to this problem would be to decompose GO terms into basic concepts and then combine the concepts into a compositional structure that represents relations between the involved basic concepts.

The cross-product extensions of GO (Mungall et al., 2009) is the ongoing work for such a solution that has formal descriptions on the internal structures of GO terms, called *logical definitions*. Each definition is an intersection of participant-role relations. The participants are in turn characterized as other concepts of OBO ontologies, including GO, Cell Ontology, ChEBI, and Sequence Ontology. They have used roles that are formally defined in the OBO Relation Ontology.

However, the cross-products have several issues to deal with. First, many of their relation types, or roles (e.g. results_in_transport_of), are not event-independent. This unnecessarily constrains the usage of the cross-products for text mining systems as they mostly assume event-independent roles (e.g. has_agent, has_patient) for the purpose of

the uniform identification of various event types from text (Carreras & Màrquez, 2005; Kim et al., 2009). Second, Mungall and colleagues (2009) reported that no reasoner is capable to reason over all cross-product sets and all referenced ontologies.

We propose to use basic relations to represent the semantic structure of GO terms, including agent-patient relations, part-of relations, space-related relations, time-related relations, and biological relations. In fact, we utilize Gene Regulation Ontology (GRO) as the framework for the semantic structures. GRO is a conceptual model for the domain of gene regulation that defines the basic concepts and relations of the domain (Beisswanger et al., 2008). The concepts of GRO are cross-linked to OBO ontologies.

We implemented a system that first represents substrings of a GO term, which correspond to existing OBO ontology concepts, with the cross-linked GRO concepts, and then associates the GRO concepts with each other through the basic relations of GRO. The resultant semantic structures of GO terms are nested type-value frames like the cross-products of (Mungall et al., 2009). Note that in contrast to the GO cross-products, it is possible for a reasoner to reason over all the semantic structures represented with GRO.

Furthermore, we have successfully applied the semantic structures of GO terms for a task of semantic similarity analysis. We show that the analysis leads us to the discovery of implied relations between concepts.

2. Methods

We present a method of identifying logical definitions of GO terms. It has three steps: term recognition, parsing, and pattern matching. The term recognition step is to label substrings of GO terms with appropriate GRO concepts. The GRO concepts labeled are used as the base units for pattern matching. The parsing step is to identify the syntactic structures of GO terms. The pattern matching is applied to the syntactic structures to generate the semantic structures. The pattern matching method is based on our rule-based system for event extraction from text (Hahn et al., 2009).

2.1. Term Recognition

First, we have used SwissProt for recognizing gene/protein names and Enzyme Nomenclature for enzyme names in GO terms. Once the names are located, they are labeled with the GRO concepts Gene and Enzyme, respectively.

Second, we have used OBO ontologies, including Sequence Ontology, ChEBI, MeSH, and GO, for recognizing ontological concepts in GO terms. The terms from ChEBI are labeled with the GRO concept Chemical.

For other ontologies, we have constructed mapping tables linking their terms to GRO concepts. These mappings are is-a relations. In fact, many of the mappings are equivalence relations, while there are non-equivalence relations such as the mapping from the GO concept “biological process” to the GRO concept “Process”. We have mapped to GRO only the GO concepts that are not descendants of “regulation of gene expression” (GO:0010468). The mappings are available at the project homepage (http://www.ntu.edu.sg/home/jungjae.kim/GO2GRO/ .).

Those mappings of *is-a* relations are used to recognize the synonyms of GRO concepts and also hyponyms to GRO concepts. If an ontological term is located in a GO term and if this term or one of its ancestor terms is mapped to a GRO concept, then it is labeled with the GRO concept. If many ancestors of an ontological term are mapped to GRO, we use only the GRO concept mapped to the nearest ancestor according to the ontology concept hierarchy.

2.2. Parsing

Our method for analyzing the semantic structures of GO terms works by matching syntactic patterns to the syntactic structures of GO terms. We have adopted Enju, a HPSG parser (Sagae et al., 2007), to identify the syntactic structures of GO terms. We converted the predicate-argument structures produced by Enju into dependency structures.

We integrate the GRO labels annotated by the term recognition module into the dependency structures. If an ontological term consists of only one word, we simply annotate its semantic labels onto the node of a dependency structure which corresponds to the word. If a term has more than one word, we merge the corresponding nodes into a single node and then annotate the semantic labels onto the node.

2.3. Pattern Matching

A GO term often has a complex structure which involves different participants and cascading relations between the participants, where a participant can be an event again. To deal with the cascaded structure, our method compositionally matches multiple patterns to a single GO term (see Hahn et al., 2009, for details). Figure 1 depicts the semantic analysis for the GO term "positive regulation of gene expression". The numbers 1, 2, and 3 in the figure indicate the order of pattern matching in a cascaded approach.

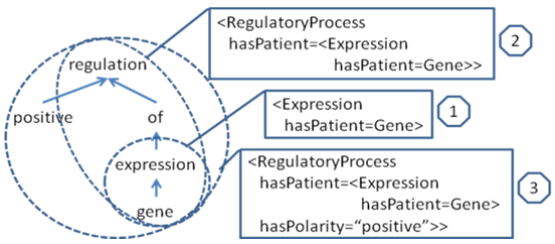

Fig. 1. Pattern matching for the GO term "positive regulation of gene expression"

3. Evaluation

3.1. Input GO terms

For evaluation, we have focused on the domain of gene regulation. The version of Gene Ontology published on January 29, 2010 contains 428 concepts under the concept “regulation of gene expression”, designated ROGE. If a GO term includes phrases such as “DNA-dependent” and “gene-specific”, we ignore the phrases since they are mainly used to distinguish the GO term from others that indicate similar processes but with different mechanisms (e.g. "RNA-dependent", "mating-type specific"). By ignoring the phrases, we treat the GO term as default, compared to the others with different mechanisms.

3.2. Pattern Construction

We have manually constructed 123 patterns to analyze the semantics of the GO terms. If a pattern encodes a biologically ambiguous relation, we represent the relation with the GRO concept Process which includes all possible biological relations. For instance, the relation between A and B of the pattern “A, B” can be any biological relation (e.g. ‘binding’, ‘regulation’). We represent it as <Process agent=A patient=B>. In fact, this semantic representation does not explicitly express the fact that it is a child concept of A. The ideal representation would be <A involvedIn=B>, where the involvement in B is regarded as an attribute of A. However, we have discarded this option since the property ‘involvedIn’ is too artificial and has not been defined in any other ontologies. It is still possible to deduce the ideal semantics from the proposed semantics by inference.

3.3. Experiment Results

We have applied our method to 428 GO terms under ROGE and successfully obtained correct semantic structures for 321 GO terms (75%). The mappings were evaluated by the authors, including one scientific curator who is experienced in annotations for GO and text mining competitions (e.g. BioCreative). The correct semantic structures are available at the project homepage.

3.3.1. Error analysis

Table 1 shows a summary of terms that did not correctly map to our representation. The categories of errors are not mutually exclusive such that the incorrect semantic structure of a GO term may fall into multiple categories. For simplicity, however, we chose only one error type for each incorrect result, roughly preferring higher category to lower one in Table 1.

Table 1. The results of error analysis

Fail to recognize	Count
Word	40 (9%)
‘host’, ‘symbiont’, ‘mating-type’	[24]
Adjectives (e.g. ‘small’, ‘other’)	[14]
Nouns (e.g. ‘integration’)	[2]
Named entity	32 (7%)
Gene/Protein	[15]
Chemical, Sequence	[6]
Other	[11]
Comma	9 (2%)
Parse Error	24 (6%)
Other	2 (1%)
Total	107 (25%)

Many of the incorrect semantics result from the low coverage of the term recognition module. For example, the module does not recognize gene/protein family names (e.g. “survival gene”, “gap gene”).

We have not addressed GO terms that contain ‘host’ or ‘symbiont’, in the sense that these GO terms are likely to be expressed in text, not with the two words, but with specific species names whose roles as host and symbiont are well described in the context. Furthermore, our system does not yet properly deal with such adjectives as ‘small’ and ‘other’.

It also fails when GO terms have commas with a particular usage. For instance, the lectin pathway in GO:0001868 (regulation of complement activation, lectin pathway) is one of the pathways for complement activation, and our system does not recognize this is-a relation due to the lack of such domain knowledge.

Enju is one of the state-of-the-art parsers in the biomedical domain, but it still produces incorrect results in parsing GO terms. One notorious example is “modification by virus of host polysomes” (GO:0046783), where the prepositional phrase “of host polysomes” modifies ‘modification’. The parser incorrectly identifies the head of the prepositional phrase as ‘virus’. This is an ambiguous example which cannot be resolved without deep domain knowledge.

For the rest of the paper, we have used only the correctly recognized semantic structures of the 321 GO terms.

3.3.2. Comparison with GO cross-products

We compared the GRO-based semantic representation of GO terms with the GO cross-products (published on 15 January 2010). Our method successfully identifies the semantic structure of 321 GO terms (75%), while the cross-products have logical definitions for 63 GO terms (15%). Let us compare the following representations of GO:0060967 (negative regulation of gene silencing by RNA):

(1) <RegulatoryProcess hasPatient = GeneSilencing
hasPolarity=”negative” agent = RNA>

(2) a. intersection_of: GO:0060969 ! negative regulation of gene silencing

b. intersection_of: OBO_REL:mediated_by CHEBI:33697 ! ribonucleic acids

The cross-product of a GO term usually includes the is-a relation with its parent term (e.g. GO:0060969) as in (2a). The logical definition of a GO term is complete only when the parent term is also logically defined. According to this notion, the cross-products have complete logical definitions for only 15 GO terms under ROGE (4%).

The difference between the two representations lies in the difference between the two relation types (i.e. hasAgent, mediated_by), while the other parts (i.e. gene silencing, RNA) are in essence identical to each other.

The relation types used by our method can be roughly classified into six groups: 1) agent-patient relations (i.e. hasAgent, hasPatient), 2) part-of relations (i.e. partOf, hasPart), 3) space-related relations (i.e. startFromLocus, locatedIn, actsOn, moveFrom), 4) time-related relations (i.e. precededBy), 5) biological relations (i.e. fromSpecies, encodedIn), and 6) other attributes (i.e. hasPolarity, hasQuality).

Among them, only "encodedIn" involves an event in itself, where it represents the relation between an 'unregistered' protein and the gene encoding the protein. First order logic can be used to replace this relation type with the ontology concept of "protein encoding" while linking the protein and the gene to the concept by using variables. Our method does not support variables, and we leave this issue to future improvements.

Both our method and the cross-products have used concepts from OBO ontologies. In fact, the two methods share five ChEBI concepts. The other three ChEBI concepts and two Sequence Ontology concepts used by the cross-products are actually defined in GRO, and our method uses the corresponding GRO concepts. Therefore, the two methods share all the concepts from ChEBI and Sequence Ontology.

However, the two methods hardly share concepts from the other sources. First, the cross-products have not used MeSH terms. Though MeSH is not a formal ontology, we utilize it in order to increase the coverage of text mining applications based on our results. Second, while we ignored 'host' and 'symbiont', the cross-products have defined them in their own concept repository. Third, the two methods share only seven out of 73 GO concepts. This is because our method uses GO concepts outside ROGE and makes use of additional sources for entities and concepts, while the cross-products usually represent is-a relations between parent and child GO concepts.

In summary, our method mostly utilizes event-independent relation types, while the cross-products often not. The two methods share many concepts from ChEBI and Sequence Ontology, but do not share other ontology concepts. Our method has established many more cross-links between the GO concepts under ROGE and those outside ROGE than the GO cross-products.

4. Application

We have compared the semantic structures of parent GO terms with those of children. Structural comparison has advantages over string comparison (cf. Verspoor et al., 2009). Let us consider the following pairs of GO terms:

(3) Parent: regulation of transcription by carbon catabolites
Child: regulation of transcription by glucose

(4) Parent: regulation of transcription from RNA polymerase II promoter
Child: regulation of transcription involved in G1 phase of mitotic cell cycle

The example (3) shows an is-a relation, where the difference between the parent term and the child term lies in the pair of "carbon catabolites" and "glucose". We can thus deduce from the example that glucose is a carbon catabolite. However, we cannot deduce from the parent term of (4) any relation between "RNA polymerase II promoter" and "G1 phase of mitotic cell cycle", not only because of the difference between 'from' and "involved in", but also because the two prepositional phrases modify 'transcription' and 'regulation', respectively. We cannot understand this difference without the semantic structures of those GO terms.

Table 2 shows the most frequent cases of such differences. The chemical names found in GO terms are replaced with their corresponding identifiers of ChEBI. For example, "CHEBI:17234" indicates glucose.

Table 2. The most frequent cases of structural difference between parent and child GO terms

Unique part in parent	Unique part in child	Count
	hasPolarity="negative"	63
	hasPolarity="positive"	58
Multiple attributes are mismatched		47
	partOf=Mitosis	12
	hasAgent=Stress	9
has Agent= CarbonCatabolite	hasAgent= <Chemical name="CHEBI:17234">	6

The third row of the table means that only 47 pairs of parents and children out of 488 pairs have differences in multiple attributes. It means that in most cases (90%) the child differs from its parent only in one attribute. In fact, in many cases, the child has an additional attribute that is not included in the parent. By using the structural differences, we can deduce relations between concepts. For instance, we can deduce an is-a relation between CarbonCatabolite and 'glucose' (CHEBI:17234), which is explicitly expressed neither in Gene Ontology nor in ChEBI. Another example is the part-of relation between Translation and TranslationInitiation, which is already expressed in Gene Ontology.

We have enumerated the newly identified relations in Table 3. The first three cases are of is-a relations. The last case has not been expressed in Gene Ontology due to the lack of the relationship type that can specify the relation between 'transcription' and "transcription factor activity".

Table 3. Newly discovered relations from GO terms

Unique part in parent	Unique part in child
hasAgent= CarbonCatabolite	hasAgent=<Chemical name= "CHEBI:17234">
hasAgent= CarbonCatabolite	hasAgent=<Chemical name= "CHEBI:28260">
hasAgent=Stress	hasAgent=Starvation
hasPatient=Transcription	hasPatient=TranscriptionFactorActivity

5. Conclusion

Our results reveal that the GRO-based representation can be better used for text mining than GO cross-products because of the usages of event-independent relations. This leads us to future plans: 1) using the GRO-based representation in recognizing GO terms in text, considering them as complex events, and 2) developing GRO-like ontologies to recognize other ontology terms in text, ultimately aiming at constructing the Semantic Web for biomedical literature.

References

Beisswanger, E., Lee, V., Kim, J.J. *et al*. (2008) Gene Regulation Ontology (GRO): design principles and use cases. *Studies in health technology and informatics*, **136**, 9–14.

Camon, E.B., Barrell, D.G., Dimmer, E.C. *et al*. (2005) An evaluation of go annotation retrieval for Biocreative and GOA. *BMC Bioinformatics*, **6** (Suppl 1), S17.

Carreras, X. and Màrquez, L. (2005) *Introduction to the CoNLL-2005 shared task: Semantic role labeling*. In CoNLL-2005.

Hahn, U., Tomanek, K., Buyko, E. *et al*. (2009) *How feasible and robust is the automatic extraction of gene regulation events? A cross-method evaluation under lab and real-life conditions*. In BioNLP, 37-45.

Kim, J.D., Ohta, T., Pyysalo, S. *et al*. (2009) *Overview of BioNLP'09 shared task on event extraction.* In BioNLP: Shared Task, 1–9.

Mungall, C.J., Bada, M., Berardini, T.Z. *et al*. (2010) Cross-product extensions of the Gene Ontology. *Journal of Biomedical Informatics*, 2010 Feb 10.

Ogren, P.V., Cohen, K.B., Hunter, L. (2005) *Implications of compositionality in the Gene Ontology for its curation and usage*. In Pacific Symposium on Biocomputing, 174–85.

Sagae, K., Miyao, Y., Tsujii, J. (2007) *HPSG parsing with shallow dependency constraints*. In 45th Annual Meeting of the Association of Computational Linguistics, 624–631.

Verspoor, K., Dvorkin, D., Cohen, K.B. *et al*. (2009) Ontology quality assurance through analysis of term transformations. *Bioinformatics*, **25**(12), i77–84.