Cross-Product Extensions of the Gene Ontology

Christopher J. Mungall[1], Michael Bada[2], Tanya Z. Berardini[3], Jennifer Deegan[4], Amelia Ireland[1,4], Midori A. Harris[4], David P. Hill[5], Jane Lomax[4]

[1]Lawrence Berkeley National Laboratory, Mail Stop 64R0121, Berkeley, CA 94720, USA; [2]University of Colorado Denver, Department of Pharmacology, Aurora, CO 80206; [3]Carnegie Institute for Science, 260 Panama St, Stanford, CA 94555; [4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; [5]The Jackson Laboratory, Bar Harbor, ME, USA

Abstract

The Gene Ontology is being normalized and extended to include computable logical definitions. These definitions are partitioned into mutually exclusive cross-product sets, many of which reference other OBO Foundry ontologies. The results can be used to reason over the ontology, and to make cross-ontology queries.

Introduction

The Gene Ontology (GO)[1] was conceived of as a means of providing structured annotations for genes and gene products, in terms of molecular function (MF), biological process (BP) and cellular component (CC). The current version of the GO has nearly 27,000 terms and 47,000 relationships. As the GO evolves, the relational graph becomes more tangled, which poses a problem for ontology maintenance, correctness and visualization. It has long been recognized that a normalized approach to ontology development helps with re-use, maintainability and evolution[2,3,4]. The OBO Foundry[5] was initiated in part to provide a means of normalizing the GO, such that for example the GO definition of "oocyte differentiation" could reference the term "oocyte" in the OBO Cell ontology (CL), and an automated reasoner tool could be used to classify this as a kind of "germ cell differentiation", based on the CL classification. This is also an example of a 're-use' pattern, common in software engineering.

Almost all of the terms in the GO have textual definitions, crafted for the human users of the GO. When these textual definitions are rendered in a computable from, we can leverage reasoner technology to automate the more tedious and error-prone aspects of ontology maintenance. We can also use these computable definitions to make cross-ontology queries and better visualize the ontology.

Logical Definitions and Cross Products

We provide computable logical definitions for terms using genus-differentia constructs, of the form "an X is a G that D". Here X is the term we are defining, G is the genus (more general term), and D is the differentia, a collection of characteristics that serve to discriminate instances of X from other instances of G. The differentiae are specified as relationships to other terms, using relations from the Relations Ontology[6]. In OBO Format (the native means of representing the GO) these are specified using `intersection_of` tags, which list the necessary and sufficient conditions for a term. For example:

```
  [Term]
  id: GO:0032543
  name: mitochondrial translation
  intersection_of: GO:0006412 ! translation
  intersection_of: occurs_in GO:0005739 !
mitochondrion
```

In OWL Manchester Syntax, this is written as an equivalence axiom between the class mitochondrial translation and the description **`translation`** `and` **`occurs_in`** `some` **`mitochondrion`**.

In the example above, the logical definition for the process references a GO cellular component term. Often we will want to reference other OBO ontologies, and this introduces multiple dependencies. We therefore partition the full set of logical definitions for GO into cross-product mapping. A cross-product of two ontologies A x B is the set of biologically meaningful terms that can be constructed by extending A using terms from B as differentia. The GO term in the example above would be mapped to a definition in the BP x CC cross product.

Each cross-product mapping is maintained as an individual resource, independent of the others (see Table 1). Currently they are optional add-ons to the GO. We distinguish between intra-GO cross products and inter-GO cross products, the latter consisting of logical definitions that reference an OBO ontology not under the management of the Gene Ontology Consortium.

A subset of the intra-GO cross products are the self-cross products: terms that can be defined solely by using terms in the same ontology.

Biological processes

The **BP x CC** cross-product set includes definitions for biological process terms that have cellular component terms as differentia. Sometimes we need to specify the subcellular location in which a process occurs, in which case we use an *occurs_in* relation. Sometimes we are describing the output of a process, such as when a cellular component is assembled or disassembled. The GO also has a rich set of subcellular transport terms, in which case the logical definition needs to be precise about the origin, destination and route of the transported entity.

Many BP terms can be defined using a BP term in the differentia. For example, the different phases of the cell cycle can be subtyped according to whether they are part of mitosis or meiosis. These definitions are grouped into the **BP x BP** set.

GO includes 3 broad categories of regulatory processes – regulation of molecular function, regulation of biological process, and regulation of biological quality – these comprise 3 distinct cross-product sets. The first two are intra-GO; the latter references terms from the PATO ontology of biological qualities[7], together with anatomical ontologies.

The cross products make use of 3 new relations introduced into GO – *regulates*, *negatively_regulates* and *positively_regulates*.

We have created a separate cross-product set for the more complex multi-organism interaction regulation terms. The logical definitions we provide here are necessarily a simplification, as we must go beyond the current expressive capabilities of OBO or OWL in order to represent inter-organism interactions.

Anatomy

GO has many terms describing development at the cellular and gross anatomical level. There are also non-development terms that nevertheless reference types of anatomical entity – for example, "muscle contraction".

We use the species-neutral OBO Cell ontology (CL) [8] for defining terms such as "oocyte differentiation" in the **BP x CL** set. Gross anatomy proves more of a challenge because the main OBO gross anatomy ontologies are specific to a species or taxon. We therefore extracted the implicit anatomical ontology embedded in the GO and used this together with alignments to existing anatomy ontologies to seed a multi-species anatomy ontology called Uberon, which is used in the definition of terms such as "muscle contraction". These definitions are part of the **BP x Uberon** set. Uberon covers only animals; plant development terms are in **BP x PO** (plant anatomy ontology). There are also individual species-species extensions such as **BP x Fly_anatomy**.

Molecules and proteins

Molecular and chemical entities are represented in the CHEBI ontology[9], with proteins represented in PRO[10]. We use these in 3 cross-product sets, **{BP,MF} x CHEBI**[11] and **BP x PRO**. The Protein Ontology is still relatively new, so this last set is currently relatively small. We also intend to work with the PRO curators to make an **CC x PRO** set.

Cellular components

Many of the terms in CC can be assigned logical definitions based on parthood relations to other components – for example, "nuclear chromosome" is a chromosome that is *part_of* a nucleus. For other definitions in **CC x CC**, we introduce additional spatial relations, such as *surrounds* and *surrounded by*.

The GO CC ontology has many terms representing complexes, some of which are defined by their constituent parts, others by function. The latter have logical definitions in the **CC x MF** cross-product.

Some cell component terms are differentiated by the cell type of which they are a part – for example, a sarcoplasm *is a* cytoplasm that is *part_of* a muscle cell. We map GO terms such as sarcoplasm to the BP x CL set, most of which use the *part_of* relation. For others, such as neuromuscular junction we use adjacency relations.

Reasoning

The current set of logical definitions can be used by a variety of different reasoners. We use the OBO-Edit[12] reasoner, because it is integrated within the normal editing environment for the GO, and provides incremental reasoning support.

We have not found any reasoner that is capable of reasoning over the union of the GO plus all cross-product sets plus all referenced ontologies. However, we are able to reason over individual cross-product sets and their referenced ontologies individually.

We use reasoning primarily for ontology maintenance, to compute and check the subsumption hierarchy. The GO regulation hierarchy in particular has benefited from this work, with over 2000 missing links added to GO, which could potentially improve the results of term enrichment analyses. We use the reasoner in what we call 'repair mode' – we invoke the reasoner to spot mistakes and fill in missing links in the ontology, always asserting links that can be automatically computed. This ensures that editors can edit the ontology without invoking the reasoner over the union of all logical definitions. This stands in contrast to how the reasoner is used in SO and the Fly anatomy ontology. We also use the reasoner to make inferences about the source ontologies[13].

We are still exploring uses of the cross-product sets beyond ontology construction and maintenance. This includes improved visualization, enhancing term enrichment analyses, annotation inferences and using the CHEBI cross-products to harmonize pathway database representations and GO metabolic processes.

Post-composition

The GO does not pre-compose terms for all biologically meaningful compositions of terms, as this would lead to a large, unwieldy ontology. The guiding principle is to generate compositional terms where the differentia is important to the biology. We are simultaneously exploring an approach whereby annotators can extend GO terms on-the-fly, i.e. selecting compositions from the cross-product at annotation time. For example, an annotator can select the GO term 'mitochondrial membrane' for a cellular component annotation and extend this using a differentia '*part_of* Purkinje cell', with the differentia term coming from CL. This is logically equivalent to annotating to a term 'mitochondrial membrane of Purkinje cell', but avoids bloating the ontology with the full set of biologically instantiable terms in the CC x CL cross-product.

			XP Name	Size	Examples
Intra-GO			* Biological process	606	**S phase of mitotic cell cycle = S phase** and *part_of* **mitosis**
	regulation		Biological process X self (regulates)	3529	**Regulation of neuroblast proliferation = biological regulation** and *regulates* **neuroblast proliferation**
			Biological process X self (multi-organism)	374	**modulation of intracellular transport in other organism during symbiotic interaction = interspecies interaction between organisms** and *regulates* **intracellular transport** and *during* **symbiosis** and *regulates_process_in* **external organism**
			Biological process X MF (regulates)	201	**Regulation of protein kinase activity = biological regulation** and *regulates* **protein kinase activity**
			Biological process X cellular component	476	**Mitochondrial translation = translation** and *occurs_in* **mitochondrion**
			Biological process X SO	61	**group I intron catabolic process = catabolic process** and *has_input* **group I intron**
			* Cellular component X self	682	**Acrosomal membrane = membrane** and *surrounds* **acrosome**
			Cellular component X molecular function	173	**histone deacetylase complex = protein complex** and *has_function* **histone deacetylase activity**
			* Molecular function X self (regulates)	104	**Lipase activator activity = molecular function** and *regulates* **lipase activity**
			Molecular function X cellular component	48	**Microtubule motor activity = motor activity** and *results_in_movement_along* **microtubule**
Inter-GO	Anatomy		Biological process X cell	544	**Oocyte differentiation = cell differentiation** and *results_in_acquisition_of_features_of* **oocyte**
			Biological process X Uberon	583	**Neural plate formation = anatomical structure formation** and *results_in_formation_of* **neural plate**
			Biological process X quality {X anatomy}	31	**Regulation of cell volume = biological regulation** and *regulates* **(volume** and *quality_of* **cell)**
			Molecular function X Uberon	9	**Structural constituent of bone = structural molecule activity** and *inheres_in* **bone**
			Cellular component X cell	28	**Neuromuscular junction = synpase** and *adjacent_to* **motor neuron axon** and *adjacent_to* **contractile fiber**
	Molecule	CHEBI	Biological process X CHEBI	3077	**L-cysteine catabolic process to taurine = catabolic process** and *has_input* **L-cysteine** and *has_output* **taurine**
			Molecular function X CHEBI	315	**nitrate reductase activity = oxidoreductase activity** and *reduces* **nitrate**
		PRO	Biological process X PRO	37	**Interleukin-1 biosynthesis = biosynthetic process** and *has_output* **interleukin-1**

Table 1. *GO logical definitions are partitioned into mutually exclusive cross-product sets. Examples are shown from each of the sets. The second column shows the number of existing GO terms that have been mapped to a logical definition in each set. Asterisks (*) denote self cross-products. In total 10878 terms have been mapped, 41% of all terms in the ontology.*

Conclusions

The extended collection of cross-product resources described here represents a significant advance in the evolution of the GO and its integration with other OBO ontologies. The use of these logical definitions, in conjunction with a reasoner has substantially increased the quality of the GO and eased the more prosaic aspects of ontology maintenance. We are still exploring application beyond the ontology itself.

This work also highlights the importance and necessity of the OBO Foundry effort, particularly with respect to efforts to create single orthogonal well-partitioned ontologies each representing a distinct domain of biology.

Methods and availability

In contrast to some ontology development efforts, in which computable definitions are assigned when terms are created, we have been working retrospectively, constructing logical descriptions for pre-existing terms. To help us with this task we use Obol[14], which heuristically generates proposed logical definitions based using ontology-specific grammars. Ontology editors then vet the definitions, often substantially.

The full extended GO can be obtained on the GO wiki: http://wiki.geneontology.org/index.php/Category:Cross_Products

Comments and contributions are welcome.

Acknowledgments

This work is supported by the NHGRI, via the Gene Ontology Consortium, HG002273.

References

1. Ashburner, M.; Ball, C. A.; Blake, J.; Butler, H.; Cherry, J.; Corradi, J.; Dolinski, K.; Eppig, J.; Harris, M.; Hill, D.; Lewis, S.; Marshall, B.; Mungall, C.; Reiser, L.; Rhee, S.; Richardson, J.; Richter, J.; Ringwald, M.; Rubin, G.; Sherlock, G. & Yoon, J. Creating the gene ontology resource: design and implementation. *Genome Res,* 2001*, 11*, 1425-1433
2. Hill, D. P.; Blake, J. A.; Richardson, J. E. & Ringwald, M. Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies*. Genome Res,* 2002*, 12*, 1982-91
3. Wroe, C. J.; Stevens, R.; Goble, C. A. & Ashburner, M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput,* 2003, 624-35
4. Rector, A. L. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. *Proceedings of the 2nd international conference on Knowledge capture, ACM,* 2003, 121-128
5. Smith, B.; Ashburner, M.; Rosse, C.; Bard, J.; Bug, W.; Ceusters, W.; Goldberg, L. J.; Eilbeck, K.; Ireland, A.; Mungall, C. J.; Consortium, T. O.; Leontis, N.; Rocca-Serra, P.; Ruttenberg, A.; Sansone, S.-A.; Scheuermann, R. H.; Shah, N.; Whetzel, P. L. & Lewis, S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol,* 2007*, 25*, 1251-1255
6. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C: Relations in biomedical ontologies. *Genome biology* 2005, 6(5):R46.
7. Gkoutos, G. V.; Green, E. C.; Mallon, A. M.; Hancock, J. M. & Davidson, D. Using ontologies to describe mouse phenotypes*. Genome Biol,* 2005*, 6*, R8
8. Bard, J.; Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biol,* 2005*, 6*, R21
9. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M: ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research* 2008, 36(Database issue):D344-350.
10. Natale DA, Arighi CN, Barker WC, Blake J, Chang TC, Hu Z, Liu H, Smith B, Wu CH: Framework for a protein ontology. *BMC bioinformatics* 2007, 8 Suppl 9:S1.
11. Bada, M. & Hunter, L. Identification of OBO nonalignments and its implications for OBO enrichment. *Bioinformatics,* 2008*, 24*, 1448-1455
12. Day-Richter, J.; Harris, M. A.; Haendel, M.; Lewis, S. OBO-Edit--an ontology editor for biologists. *Bioinformatics,* 2007*, 23*, 2198-2200
13. Bada, M.; Mungall, C. & Hunter, L. A Call for an Abductive Reasoning Feature in OWL-Reasoning Tools toward Ontology Quality Control. *5th OWL Experiences and Directions Workshop (OWLED 2008),* 2008

14. Mungall, C. J. Obol: Integrating Language and Meaning in Bio-Ontologies. *Comparative and Functional Genomics,* 2004*, 5(7)*, 509-520