Cell, chemical and anatomical views of the Gene Ontology. Mapping to a Roche controlled vocabulary as a test case.

David Osumi-Sutherland[1]*, Others TBA

[1] European Bioinformatics Institute (EMBL-EBI) European Molecular Biology Laboratory Hinxton, Cams, UK

Abstract. The Gene Ontology is part of a network of logically interconnected ontologies including ChEBI, the Cell Ontology and Uberon. These logical interconnections make it possible to query the GO by cell type, chemical or anatomical structure, retrieving relevant GO terms and associated annotations. In this paper we describe the use of such queries to automate mappings from a controlled vocabulary developed by Roche to lists of terms from the GO.

Using OWL-EL queries, we can fully automate mapping for about a third of terms in the Roche vocabulary, with another third having 5 or less GO terms requiring manual mapping.

The approach we describe here is not limited to mapping external vocabularies on to the GO. It could be used to provide chemical, cell or anatomically focussed ways of grouping GO annotations and of performing enrichment analyses. It could also be used for more sophisticated, combinatorial queries of the GO and its annotations.

Keywords: OWL, OWL-EL, controlled vocabulary

1 Introduction

The Gene Ontology is widely used to annotate and group gene products according to their subcellular location (e.g., endoplasmic reticulum), molecular function (e.g., enzyme activity) and their wider role in cellular, developmental and physiological processes (e.g., signal transduction) [1]). The logical structure of the ontology is used to group genes annotated with related terms and for term enrichment, a technique for determining the over- or under-representation of general classes of gene products in experimental datasets [9]. Grouping and term enrichment typically only use logical relationships *within* each of the 3 sub-ontologies of the GO - cellular component, molecular function and biological process.

In recent years, GO has switched its underlying formalization to Web Ontology Language (OWL2) (`http://www.w3.org/TR/owl2-primer/`), and has dramatically increased the number of logical axioms (Mungall et al., 2014). This new axiomatisation includes many new relationship types, relationships between terms in different GO sub-ontologies and extensive logical links to terms from external ontologies including the cell ontology [7], the chemical ontology ChEBI [4] and the Uberon multi-species anatomy ontology [3]. For example, the chemical participants in over 12,000 processes or functions are specified in the GO via axioms referencing chemical entities defined by ChEBI [5]. Over 8000 GO classes have some direct or indirect logical link to a term from the Cell ontology or Uberon. These record, for example, the location of cellular components (e.g., the acrosome and its parts are present only in sperm), cell types that are the sole location of some process (e.g., 'natural killer cell degranulation' only occurs in natural killer cells), and the products of developmental processes (e.g., bone is a product of 'bone morphogenesis').

Axiomatisation of the GO is limited to the EL profile of OWL [8]. This allows GO infrastructure to take advantage of fast, scalable OWL-EL reasoners such as ELK [6] to leverage the classifications in external ontologies to automate classification in GO, and to ensure that classification and querying of the GO will not become intractable as the ontology grows. There is also great potential for using this axiomatisation to provide new, biologically meaningful systems for grouping annotations and term enrichment. For example, we might want to group all annotations to genes involved in processes occurring in T-cells or in the pancreas, or to group annotations to genes involved in the processes involving nitric oxide. In this paper we describe an implementation of this strategy in support of a use case from the pharmaceutical company Roche.

Roche uses a controlled vocabulary internally (from here referred to as RCV). RCV consists of around 360 undefined terms, each of which is mapped to a set of terms from the GO. RCV includes terms named for biological processes and, more rarely, for molecular functions and cellular components. It also includes many terms named for types of cell, chemical, anatomical structure and taxonomic group. Prior to this work, mappings from RCV to the GO were made manually, based on the lexical content of the names of GO terms and the biological knowledge of those doing the mapping. As the GO evolved, it became increasingly impractical for Roche to keep this mapping complete and up-to-date via manual mapping.

Here we describe the development and testing of an automated mapping between GO and RCV, making use of OWL-EL reasoning and a standard system for specifying OWL design patterns.

2 Results

2.1 Mapping strategy

In the manual mapping between RCV and GO specified by Roche, multiple GO terms are mapped to each RCV term. For the purposes of automated mapping, we interpret the mapped GO terms as subclasses of the class referred to by the RCV term. For each RCV term, we attempted to find an equivalent class expression (a mapping query) that reflected the intended meaning of the RCV term, as judged by the RCV term name and manual mappings and based on discussion with Roche.

Mapping to fully expressive OWL-DL class expressions poses serious problems for scalability: querying and classification of the GO with OWL-DL reasoners is already slow, and may become intractable as the GO grows. For this reason, we chose to restrict mappings to EL class expressions and use the fast, scaleable EL reasoner, ELK to run queries [6]. But this poses a problem for expressiveness: EL lacks various elements of OWL that are potentially useful for mapping queries - most notably disjunction (OR) and negation (NOT).

To compensate partially for the lack of disjunction in OWL-EL, we developed a set of high level object properties for use in queries. For example, we define **occurs_in_OR_has_participant** as a grouping relation, allowing queries for processes that occur in a specified cell, or have that cell as a participant. Similarly, GO does not use a reflexive relation for 'part of', but using one for query purposes means a query for subclasses of "'part of' some X" returns both subclasses and proper parts of class X.

Many RCV terms group processes in which a specified chemical or cell participates, with processes regulating those in which it participates (see Table 1 for example). To support such groupings, we used an OWL property chain axiom (`http://www.w3.org/TR/owl2-primer/#Property_Chains`) to define a relation, **regulates_o_has_participant**, to query for processes that regulate a process in which some specified entity is a participant. We then define a super-property, **participant_OR_reg_participant**, for this new relation and **has_participant**

regulates *o* **has_participant** *subPropertyOf* **participant_OR_reg_participant**
regulates_o_has_participant *subPropertyOf* **participant_OR_reg_participant**
has_participant *subPropertyOf* **participant_OR_reg_participant**

In order to keep the mapping process simple, we added a further restriction: only a single mapping class was specified for each mapping.

The heavy use of OWL Object Property axioms to compensate for loss of expressivity tends to obscure the semantics of mappings. In order to communicate the meanings of mappings clearly, we used a script to generate human readable descriptions for each mapping query. For example, we mapped the RCV term cannabinoid to the ChEBI term cannabinoid (CHEBI:67194) plus pattern **participant_OR_reg_participant**. The automated description of the mapping reads: "A process in which a cannabinoid participates, or that regulates a process in which a cannabinoid participates."

Each mapping was used to generate a mapping table for manual review (an example is shown in table 1), allowing the possibility of blacklisting either automated mappings or manual mappings (as a way of specifying corrections to the original manual mapping).

2.2 Summary of results

We successfully mapped 308/364 RCV terms to mapping queries. As shown in table 2, over a third (104) of the mapping queries were sufficient - meaning that no manual maintenance is required - and a further third of the mappings (148) had 10 or fewer additional manual mappings (figure 1)

Mapping queries found many GO terms that were not in the manual mapping (see figure 2). For a few very general RCV classes (e.g. enzyme), over 1000 new mappings were found. Very few automated mappings were blacklisted - just 70 terms in total. Blacklisting of terms ensured they were removed from the final mapping.

3 Discussion and future directions

This work demonstrates how the logical structure of the GO can be used to achieve biologically meaningful mappings between GO and terms from external controlled vocabularies or ontologies for which there is no corresponding GO term. For example, where the external vocabulary refers to a cell-type, a chemical or an anatomical structure. The mapping system used is fast and scalable,

3.1 Improving the RCV mapping pathway

There is good scope for improving the mapping between RCV and GO so that it is more thoroughly automated. As shown in 3 RCV term names follow patterns that mostly have consistent mappings patterns, but there are some exceptions. More consistency in naming will make the meaning of terms more predictable and make the mapping of new terms more straightforward.

For RCV terms with only a small number of additional manual mappings, it may be worth considering whether the overhead of manual maintenance is worth the effort, especially where these additional mappings could not be achieved by further axiomatisation of the GO. For example, mappings to cell types often include mappings to growth factors acting on those cell types. As these growth factors have much broader functions than action on the cell types for which they are named, GO is unable to add any formal link between factors and cell types.

In other cases a mapping pattern involving two or more specified classes and a more sophisticated logic would be necessary to obtain a complete mapping. For example, the manual mappings for X metabolism terms are consistently mapping to bot X metabolism and X transport terms in the GO. A more complete mapping to RCV metabolism terms could be achieved using a pattern that named both GO transport and GO metabolic process terms. This could be made scalable with a pathway that combines the results of multiple mapping patterns outside of OWL.

56 terms were not mapped. Some were rejected from the pipeline as they were judged to be duplicates with other RCV terms. The rest were rejected as currently unmappable due to the lack of suitable terms or axiomatisation within the GO at this time. For example, GO currently has no way to group aerobic or anaerobic metabolic processes, although it does reflect the aerobic or anaerobic nature of many metabolic processes in their names and textual definitions. Further formalisation of the GO is likely to improve the number of concepts that can be mapped.

3.2 Alternative views of the GO and its annotations

The mechanisms described here for mapping to external ontologies could also be used for providing alternative views of the GO and its annotations. This is already reflected in some of the newer functionalities of the GO browsing tool AMIGO, which now displayed inferred annotations to cell-types based on axioms in GO recording where processes occur.

3.3 Future work

The system described here was designed to be lightweight and flexible, allowing maximum interaction between the designers of RCV at Roche and GO editors with minimal development overhead.

The pattern-based system used here bears some relationship to the TermGenie system [2] which is already used to generate 80% of new GO terms. One possible approach to fulfilling the needs of external groups for types of classification not included in the GO would be to offer a TermGenie-like system for generating terms that group GO terms in ways that are not currently supported internally by the GO.

4 Methods

All code, mapping tables and results for the pipeline were maintained in a GitHub repository (`https://github.com/GO-ROCHE-COLLAB/Roche_CV_mapping`). As well as providing version control, Github allows nicely formatted display of mapping and results files in an easily editable form (tab separated value (TSV), which can be easily edited via copying and pasting from excel spreadsheets). It also has an integrated ticket system, with an open API. Standard mapping tickets were generated by script for all RCV terms mapped. A standard system of ticket labels allowed tracking of the approval status of all mappings. The archived tickets constitute an audit trail for approval of mappings.

The mapping was specified using a single TSV file in which each line maps an RCV term to a mapping query and a term from GO, ChEBI, CL, Uberon or NCBI taxonomy.

OWL reasoning was carried out via calls to a standard Java API for OWL using the ELK OWL reasoner [6]. The query and processing pipeline was written in Jython, a Python implementation over Java (`http://www.jython.org/`). The pipeline produced a set of results tables, one for each RCV term, in TSV format. These were used for review of mappings by Roche.

Author's contributions

Acknowledgments.

Funding This work was supported by direct funding from Roche. The Gene Ontology is supported by ...

References

1. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, 43(Database issue):D1049–1056, 2015.
2. H. Dietze, T. Z. Berardini, R. E. Foulger, D. P. Hill, J. Lomax, D. Osumi-Sutherland, P. Roncaglia, and C. J. Mungall. TermGenie - a web-application for pattern-based ontology class generation. *J Biomed Semantics*, 5:48, 2014.
3. M. A. Haendel, J. P. Balhoff, F. B. Bastian, D. C. Blackburn, J. A. Blake, Y. Bradford, A. Comte, W. M. Dahdul, T. A. Dececchi, R. E. Druzinsky, T. F. Hayamizu, N. Ibrahim, S. E. Lewis, P. M. Mabee, A. Niknejad, M. Robinson-Rechavi, P. C. Sereno, and C. J. Mungall. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J Biomed Semantics*, 5:21, 2014.
4. J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, 41(Database issue):D456–463, Jan 2013.
5. David P Hill, Nico Adams, Mike Bada, Colin Batchelor, Tanya Z Berardini, Heiko Dietze, Harold J Drabkin, Marcus Ennis, Rebecca E Foulger, Midori A Harris, Janna Hastings, Namrata S Kale, Paula de Matos, Christopher Mungall, Gareth Owen, Paola Roncaglia, Christoph Steinbeck, Steve Turner, and Jane Lomax. Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. *BMC genomics*, 14(1):513, January 2013.
6. Yevgeny Kazakov, Markus Krötzsch, and František Simančík. Elk reasoner: Architecture and evaluation. *CEUR Workshop Proceedings*, 858, 2012.
7. T. F. Meehan, A. M. Masci, A. Abdulla, L. G. Cowell, J. A. Blake, C. J. Mungall, and A. D. Diehl. Logical development of the cell ontology. *BMC Bioinformatics*, 12:6, 2011.
8. C. Mungall, H. Deitze, and D. Osumi-Sutherland. Use of OWL within the Gene Ontology. In C. Maria Keet and V. Tamma, editors, *Proceedings of the 11th International Workshop on OWL: Experiences and Directions (OWLED 2014)*, volume 1265 of *CEUR workshop proceedings*, pages 25–36, 2014.
9. N. H. Shah, T. Cole, and M. A. Musen. Chapter 9: Analyses using disease ontologies. *PLoS Comput. Biol.*, 8(12):e1002827, 2012.

Table 1. Example mapping table.

name	ID	manual	auto	checked	blacklisted	is_obsolete
regulation of endocannabinoid signaling pathway	GO_2000124	1	1	1	0	0
cannabinoid signaling pathway	GO_0038171	1	1	1	0	0
endocannabinoid signaling pathway	GO_0071926	0	1	0	0	0
cannabinoid receptor activity	GO_0004949	0	1	0	0	0
cannabinoid biosynthetic process	GO_1901696	0	1	0	0	0

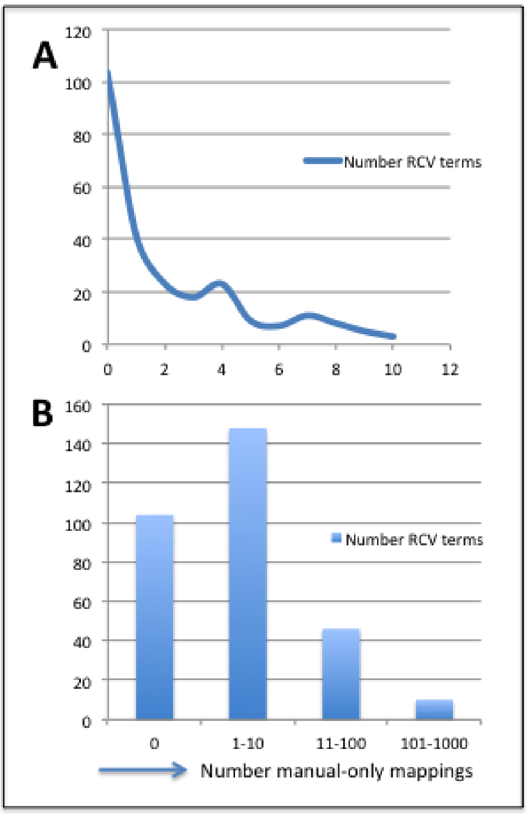


Fig. 1. Distrution of manual only mappings. A: Number of terms (Y-axis) vs number of manual mappings (X-axis) (cut-off at 20 manual mappings). 1B: Distribution of manual only mapping: 104 terms have no manual-mappings at all. A further 148 have between 1 and 10. The largest number of manual-only mappings for a single RCV terms was 323

Table 2. Summary of mappings.*Auto sufficient:* Number of RCV terms for which no manual mapping required; *Manual only*: Number of GO terms only in manual mapping, *Auto only*: Number of GO terms only in automated mapping; *Manual blacklist* Number of blacklisted manually mapped terms; *Auto blacklist*: blacklisted terms in automated mapping. Average and median values are per RCV term.

STAT	RCV_name	Auto sufficient	Manual only	Auto only	Manual blacklist	Auto blacklist
SUM	308	104	4133	44012	81	70
AVERAGE	-	-	13.42	142.9	0.26	0.23
MEDIAN		-	2.0	25.0	0.0	0.0

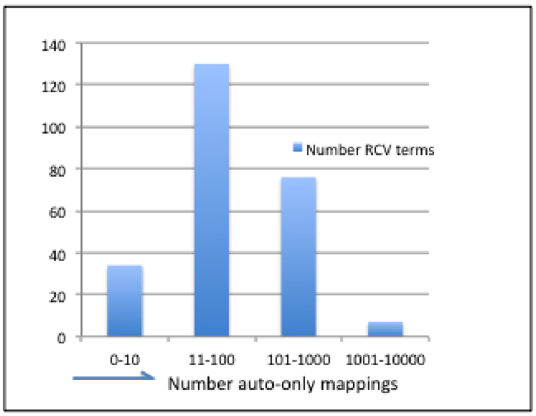
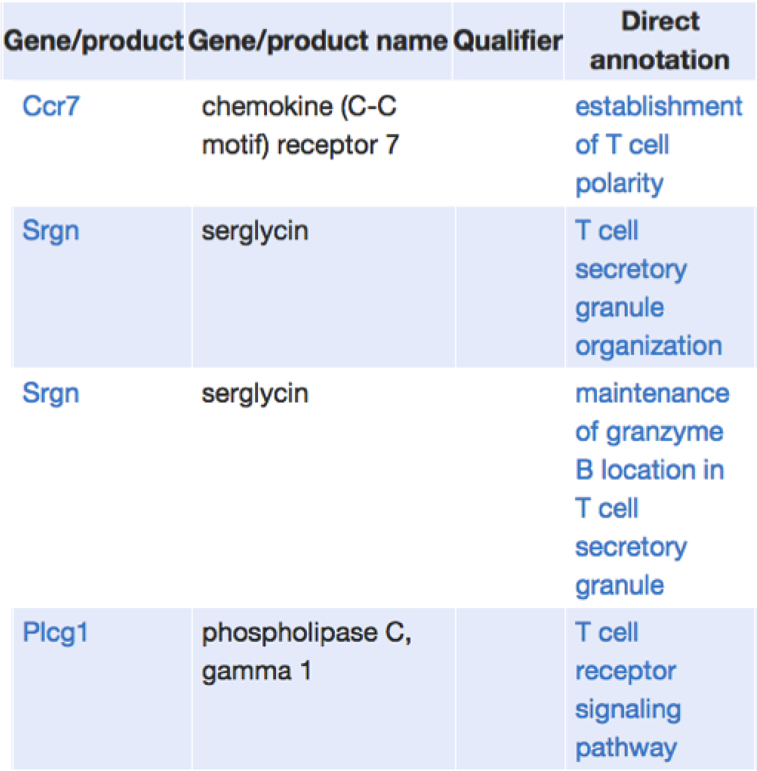


Fig. 2. Distribution of auto only mappings. X axis = number of auto-only mappings. Y axis = Number of RCV terms. Most mapping queries found under 100 additional (auto only) mappings, but over 75 found between 101 and 1000 and a few mapping queries found between 1001 and 10000 new mappings.

Table 3. Common mapping queries

RCV term name pattern	**Most common mapping query (relation + target class ontology)**	**No. cases different mapping query used**
chemical (e.g. cannabinoid)	participant_OR_reg_participant + ChEBI (used 22 times)	7
chemical metabolism (e.g.	Equivalence + GO (used 34 times)	6
cell type (e.g. T cell)	parts_participation_location + CL (covers parts, participant processes and processes occurring in cell type. used 30 times)	1
X development (e.g. bone development)	is_a_OR_part_of_OR_regulates + GO (used 41 times)	6
X pathway/signalling (e.g. BMP signalling)	is_a_OR_part_of_OR_regulates + GO (used 21 times)	6



Gene/product	Gene/product name	Qualifier	Direct annotation
Ccr7	chemokine (C-C motif) receptor 7		establishment of T cell polarity
Srgn	serglycin		T cell secretory granule organization
Srgn	serglycin		maintenance of granzyme B location in T cell secretory granule
Plcg1	phospholipase C, gamma 1		T cell receptor signaling pathway

Fig. 3. The AmiGO GO browsing tool displayes annotations to cells. The T cell page on AmiGO displays annotations to processes that only occur in T-cells. From http://amigo.geneontology.org/amigo/term/CL:0000084