GO faster ChEBI with reasonable biochemistry

Duncan Hull

Manchester Interdisciplinary Biocentre, School of Chemistry, 131 Princess Street,
University of Manchester, Manchester, M1 7DN, UK
firstname.surname@manchester.ac.uk,
http://duncan.hull.name

1 Introduction

Many new ontologies have been developed in recent years with the aim of facilitating data integration in both the chemical and life sciences. One such ontology is Chemical Entities of Biological Interest (ChEBI) [4]. As the name suggests, this ontology describes biologically interesting chemical entities - which includes small molecules such as aspirin. Currently, ChEBI does not make use of description logic but ongoing revisions to the ontology [2] have created new opportunities for more extensive reasoning over ChEBI in the future. This also raises some challenging problems which require attention. This paper describes some of these problems, structured as follows: Section 2 introduces and describes ChEBI in more detail. This is followed by a discussion of some of the issues ChEBI currently faces in its maintenance and development in section 3, which also outlines and discusses potential solutions. Finally, section 4 draws some conclusions and points to future work.

2 Chemical Entities of Biological Interest

Each molecular entity in ChEBI has several representations of its chemical structure, describing what atoms are connected and how: 1) The **S**implified **M**olecular **I**nput **L**ine **E**ntry **S**pecification (SMILES) string and 2) the **In**ternational **Ch**emical **I**dentifier (InChI) and a two dimensional structure diagram, see [6] for figure. The latest release (version 49) contains 15 833 annotated entities in its database which makes ChEBI fairly small when compared to similar databases. PubChem for example contains more than 38 million substances and 18 million unique structures [10]. The automation of routine and mundane curation tasks would leave curators with time to to do more skilled work [5]. A full description of ChEBI is outside the scope of this paper, further details can be found in [4]. Having briefly introduced ChEBI, the next section looks at some of the challenges and opportunities for reasoning with the ChEBI ontology.

3 Refining ChEBI: necessary and sufficient

As with many ontologies, ChEBI has potential for refinement and improvement - some of these issues have become apparent in the REFINE project, introduced in

section 3.1, others are more general. Where appropriate, solutions discussed here that use the Web Ontology Language (OWL) are demonstrated with examples using the Manchester syntax. Currently, ChEBI defines all classes using *necessary* conditions, none are defined using *necessary and sufficient* conditions, which limits the use of reasoning over the ontology and means that many parent-child relations have to be maintained manually, rather than automatically inferred by reasoners. So for example, `organic molecular entities` (CHEBI:25700) are defined textually as "a molecular entity that contains carbon". The machine-understandable OWL version of ChEBI defines this class as:

```
Class: OrganicMolecularEntity

    SubClassOf:
        MolecularEntity
```

This states that to be a member of the `OrganicMolecularEntity` class, there is only one necessary (`SubClassOf`) condition, that the entity concerned is also `MolecularEntity`, but unlike the textual definition makes no mention of carbon. From this information, a reasoner will not be able to classify entities as subclasses (or not) of `OrganicMolecularEntity`. Instead these links have to be asserted manually by a curator. Currently, ChEBI asserts eight immediate "children" of this class shown below. Maintaining these parent-child links by hand is a tedious and error-prone task. An alternative definition of Organic Molecular Entity that makes more use of reasoning, by using necessary and sufficient (`EquivalentTo`) conditions would be:

```
Class: OrganicMolecularEntity

    EquivalentTo:
    MolecularEntity
    that hasPart some CarbonAtom and hasPart some HydrogenAtom
```

This would say more about what an OrganicMolecularEntity was, and would allow a reasoner to infer and classify cholesterol and methane (ChEBI:16183) as organic but carbon dioxide as inorganic, because although it contains carbon, it does not have any hydrogen atoms. This is a fairly trivial example, but chemistry has many of these kinds of rules which could be encoded as "defined" rather than "primitive" classes. Adding defined classes would make ChEBI easier to maintain, so instead of asserting a multiple-inheritance hierarchy by hand, curators manage a single-inheritance hierarchy, and let a reasoner infer the rest. Currently, this is something the ChEBI developers plan to do in future releases[1].

3.1 Refining metabolism, alignment and ambiguity

A modular and orthogonal ontology are essential requirements because ChEBI is used to describe the components of biochemical pathways (series of reactions)

[1] `http://chebi.wiki.sourceforge.net/New+ChEBI+Ontology`

and is not used only in isolation. So for example, glucose (ChEBI:17234) is part of a pathway called glycolysis (GO:0006096) which involves several entities described by both ChEBI and the Gene Ontology. Aligning these two ontologies has been problematic [1], because of the need for abductive rather than deductive reasoning to identify non-alignments. Working with metabolic models, the REFINE project[2] has been using ChEBI to mine PubMed and other bibliographic databases using curated models of a biochemical pathways from the biomodels database. The aim is to link these models, via text-mining, to the primary quantitive and qualitative evidence in the literature that a given reaction, or series of reactions, actually exists. This is made possible through the use of ChEBI, which has been used to annotate pathways in the biomodels database, however some models are annotated inconsistently. For example, a model describing glycolysis in yeast contains a reference to ChEBI:26055, an old identifier which redirects to ChEBI:44897 (Phosophenolpyruvic acid) see [6] for figure. Cross references from this ChEBI record refer to KEGG:C00074 (phosphoenolpyruvate) which then links back to two different entities in ChEBI via annotations ChEBI:18021 (phosphoenolpyruvate) ChEBI:44897 (Phosophenolpyruvic acid)[3]. These kinds of contradictory annotations are very common and it may be possible to highlight inconsistencies through the use of functional properties in OWL, although this still leaves their resolution unsolved.

3.2 Representing and Searching for Structures

The ability to search for chemical structures is an essential requirement for chemical databases. Currently, ChEBI supports three methods (similarity, substructure and identity) using techniques that are well established in cheminformatics [7]. A key requirement for searching is the ability to represent cyclic structures, such as benzene (ChEBI:16716) which has a "ring" structure, shown as part of Figure 1. This is a challenge for OWL, because most reasoners work by constructing tree-like graphs, which do not always lend themselves to representing and reasoning about circular structures. However, description graphs [8] and SWRL make this possible, and it would be an interesting exercise to see if semantic techniques could improve on established methods[7]. It may even be the case that reasoning can succeed where conventional cheminformatics has failed, due to the inherent problems of dealing with "semantically bleached strings" (InChI's).

4 Conclusions and Future work

Several previous attempts to build "chemical semantic webs" [9, 3] have concentrated on the use RDF, rather than OWL, to represent metadata, and have made little or no use of reasoning. This paper has briefly shown where OWL can

[2] REFINE project: `http://dbkgroup.org/refine`
[3] see `http://pod.cs.man.ac.uk/srp/infotech.mov` for details

help by making more use of necessary and sufficient conditions, allowing reasoning about chemical structure, highlighting ambiguous inconsistencies although problems with Gene Ontology (GO) alignment still remain.

5 Acknowledgements

This work has been funded by the BBSRC grant reference BB/E004431/1 as part of the REFINE project devised by Douglas Kell and run by Sophia Ananiadou. The author would also like to thank all the ChEBI team, The OBO consortium, Paul Dobson, Colin Batchelor and Steve Pettifer. A more complete version of this paper can be found at [6].

References

1. Michael Bada and Lawrence Hunter. Identification of obo nonalignments and its implications for obo enrichment. *Bioinformatics (Oxford, England)*, 24(12):1448–1455, May 2008.
2. Colin Batchelor. An upper-level ontology for chemistry. In *The 5th International Conference on Formal Ontology in Information Systems (FOIS 2008)*, October 2008.
3. O. Casher and H. S. Rzepa. Semanticeye: A semantic web application to rationalize and enhance chemical electronic publishing. *J. Chem. Inf. Model.*, 46(6):2396–2411, November 2006.
4. Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan Mcnaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, October 2007.
5. Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P. Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, Simon Twigger, Owen White, and Yon. Big data: The future of biocuration. *Nature*, 455(7209):47–50, 2008.
6. Duncan Hull. Go faster chebi with reasonable biochemistry. *Nature Precedings*, September 2008. http://dx.doi.org/10101/npre.2008.2329.1.
7. Andrew R. Leach and Valerie J. Gillet. *An Introduction to Chemoinformatics*. Kluwer Academic Publishers, rev. ed edition, September 2007.
8. Boris Motik, Bernardo C. Grau, Ian Horrocks, and Ulrike Sattler. Representing structured objects using description graphs. AAAI Press, 2008.
9. K. R. Taylor, R. J. Gledhill, J. W. Essex, J. G. Frey, S. W. Harris, and D. C. De Roure. Bringing chemical data onto the semantic web. *J. Chem. Inf. Model.*, 46(3):939–952, May 2006.
10. A. J. Williams. A perspective of publicly accessible/open-access chemistry databases. *Drug discovery today*, 13(11-12):495–501, June 2008.