Molecular symmetry and specialization of atomic connectivity by class-based reasoning of chemical structure

Michel Dumontier

Department of Biology, Carleton University, 1125 Colonely By Drive, Ottawa, Ontario, Canada K1S5B6; michel_dumontier@carleton.ca

Abstract
Chemical biology and drug discovery seek to uncover the relationship between chemical structure and function. In the context of the emerging life science semantic web, we have previously investigated multiple strategies for the representation and reasoning of chemical structure, functional groups and chemical attributes using RDF, OWL, SWRL and so-called Description Graphs. Here, we continue our investigation on the representation of molecular structure using class-based approach to infer molecular symmetry and specialization of atomic connectivity. This work provides new design patterns towards representing and reasoning about structured objects.

Keywords. OWL, chemical, graph, representation, design pattern, chemoinformatics

1 Introduction

Chemical biology and drug discovery seek to uncover the relationship between chemical structure and function. Quantitative structure-activity relationships correlate so-called descriptors or aspects of chemical structure with activity or reactivity, in the hopes of identifying other reactive molecules in the absense of experimental results. Hundreds of so-called descriptors are now being used for QSAR studies, and efforts have been made to capture these under a common ontology of chemical information [1]. In this ontology, descriptors can be associated with the structural parts or qualities that they pertain to [2], thereby enhancing the potential for enrichment analyses over the emerging life science semantic web [3, 4].

In the context of building an emerging semantic web for the life sciences, we previously investigated multiple strategies for the representation of chemical structure for the purpose of semantic annotation, classification and question answering. While a class-based representation [5] was used to describe and classify chemicals by their functional groups (chemical substructures), we were only able to represent molecules

containing cycles as instance data which could be queried using SPARQL or a rule-based formalism such as SWRL. Such an instance-based description also implied an inability to compare specific molecules in terms of their structural descriptions. More recent work [5] investigated the use of Description Graphs as a means to do chemical classification at the class level, albeit with significant limitations [7].

In this work, we pursued a class-based axiomatic representation of chemical structure using OWL, which captures object structures containing cycles. Using automated reasoning, we infer molecular symmetry and specialization of atomic connectivity, an important aspect of reasoning over functional groups. This work provides new design patterns to generate insight into reasoning over structured objects in OWL.

2 Methods

We considered the six molecules illustrated in Figure 1, as they include linear, forked and cyclical molecules.

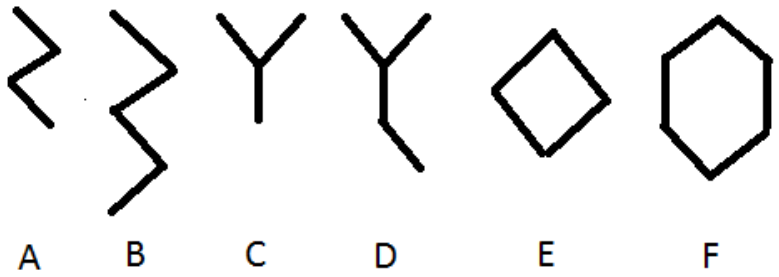


Figure 1: A) butane, B) pentane, C) iso-butane, D) iso-pentane, E) cyclobutane and F) cyclohexane.

Non-hydrogen molecular connectivity tables specified in SDF files were converted to OWL a PHP-based SDF file parser and our PHP-based OWL API. The ontology is available at http://goo.gl/3qif3. Scripts available on demand. Ontologies were reasoned about and queried using Protege 4.2 (build 269) with HermiT and FaCT++, and the built-in explanation workbench.

3 Results

3.1 Formalization

At the core of this work is the formlization of the relations that hold between a molecule, its atomic parts, and the connectivity between these atoms. We first create a named entity that exactly represents a fully connected atom, but which is not associated

with any particular molecule. The general pattern includes specifying a qualified cardinality restriction to the target atom:

`fully connected atom M`
equivalentTo
`atom type`
and `has bond with` exactly 1 `fully connected atom N`
and ...

where `atom type` refers to a specific kind of atom (e.g. carbon atom), `has bond with` uses a more specific relation (e.g. `has single bond with`, `has double bond with`, `has triple bond with` and `has aromatic bond with`) and N specifies the target atom. The pattern connections between M and other atoms are captured here.

This molecule-independent atomic connectivity is then used as a base description for specifying molecule-associated atoms. The following equivalent class axiom indicates the molecule that it is an intrinsic part of and what this atom is necessarily connected to:

`atom X from molecule A`
equivalentTo
`fully connected atom M`
and `is component part of` some `molecule Y`

In addition, we can formalize the definition of a molecule in terms of its fully connected atomic parts:

`molecule A`
equivalentTo
`molecule`
and `has component part` some `atom X from molecule A`
and ...

3.2 Molecular Symmetry

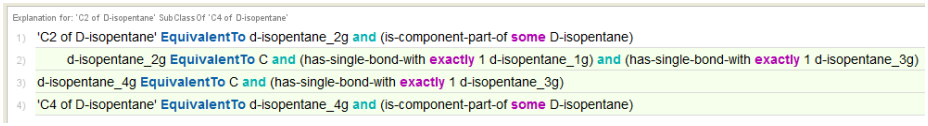
Given the equivalent class axioms specified above, an OWL reasoner will identify equivalent atoms if and only if they are connected to exactly the same atom, and no more. On reasoning, we discover equivalence between atoms 2,3,4 of iso-butane, atoms 4,5 of iso-pentane and atoms 1, 3 of cyclobutane and 2,4 of cyclobutane. In the case of isobutane and isopentane, the equivalence occurs in the three terminal single bonded carbon atoms that are attached to a common atom. For instance, the explanation of why atoms 2 and 3 of iso-butane are equivalent is provided through the explanation workbench:

In the case of cyclobutane, we see that atoms 1 and 3 share the same connectivity while atoms 2 and 4 share the same connectivity. So, while they appear to be equally

connected to carbon atoms, they are not equal in sense of specific connectivity and nor is the symmetry equal along points of common connectivity. However, despite the cycle, no equivalence is detected in the cyclohexane molecule.

3.3 Molecular Specialization

Since our representation does not include hydrogen atoms, then certain atoms will have fewer connected atoms than normal. For instance, terminal methyl carbons might normally be connected to one carbon and three hydrogens, but in this representation, there would only be one connection to a carbon. Thus, an atom would be more specialized than another if it contains all the connections of the parent and more. In our sampe dataset, we obtain atomic specializations for butane (C3 subclassof C1; C2 subclassof C4), pentane (C3 subclassof C1; C3 subclassof C5) iso-pentane (C3 subclass of C1; C2 subclass of C4; C2 subclassof C5). In the case of where C2 is a subclass of both C4 and C5 in isopentane, the explanation lies in the fact that both C4 and C5 are only connected to C3, C2 is connected to C3 and C1, as as such is a more specific kind of C4/C5 atom with respect to its connectivity.



In the case of cyclohexane, no specialization is observed given that there is exact shared connectivity nor different in the number of connected atoms.

4 Discussion

In this preliminary work, we investigated a representation of molecular structure for the purpose of class-based reasoning about atomic connectivity. Our design pattern involved the declaration of a `fully-connected atom` class, which captures the bond connectivity of atoms in a molecule, in which molecular atoms can be further described. Our representation makes it possible to analyze structures for the presence of symmetry within acyclic or cyclic molecules, with the caveat that atoms must have a common connectivity. Moreover, with a subset of common connectivity, pairs of atoms may exhibit subclass specialization.

Specified using equivalent class axioms involving exact cardinality restrictions, the representations falls outside of OWL-EL but within OWL-DL (ALEQ). Given that we only currently investigate a single molecule at a time, we expect good performance from reasoners even for significantly larger and more highly connected molecules.

An ongoing challenge lies in the representation of cyclic structures or so-called structured objects. While our representation is able to describe cyclic structures at the class-level, we note that the representation is incomplete and would results in unintended models. Problematically,

some models would only consist of infinite chains of carbon atoms. Thus, it's trivial for 4 carbon butane to satisfy a DL query asking for 10 connected carbon atoms. The system is highly underconstrained and may produce erroneous answers to queries.

Despite these difficulties, a key goal remains the inference of equivalent atoms in a cyclic structure such as cyclobutane and cyclohexane. An alternative to the approach taken here may lie in the generation of identifiers (URIs) for atoms based solely on their descriptions, and has been described in our prior work [6]. We expect that the combination of unique structural identifiers with the molecule-free atomic descriptions may be sufficient to infer equivalent atoms and identify structurally equivalent molecules.

5 Acknowledgements

We would like to thank the anonymous reviewers for providing valuable feedback which improved the quality of this manuscript. Reseach and travel is supported by a NSERC Discovery Grant.

6 References

[1] J. Hastings, L. Chepelev, E. Willighagen, N. Adams, C. Steinbeck, and M. Dumontier, The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. PLoS One, 2011. **6**(10): p. e25513.

[2] M. Konyk, A. De Leon, and M. Dumontier. Chemical Knowledge for the Semantic Web. in *DILS*. 2008. Evry, France: Springer.

[3] M.A. Nolin, M. Dumontier, F. Belleau, and J. Corbeil, Building an HIV data mashup using Bio2RDF. Briefings in Bioinformatics, 2012. **13**(1): p. 98-106.

[4] B. Chen, X. Dong, D. Jiao, H. Wang, Q. Zhu, Y. Ding, and D.J. Wild, Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC bioinformatics, 2010. **11**: p. 255.

[5] J. Hastings, M. Dumontier, D. Hull, M. Horridge, C. Steinbeck, U. Sattler, R. Stevens, T. Horne, and K. Britz, Representing chemicals using OWL, description graphs and rules, in *7th International Workshop of OWL: Experiences and Directions*: San Francisco, California, USA. p. 10.

[6] L. Chepelev and M. Dumontier, Increasingly accurate biochemical knowledge representation with precise, structure-based chemical identifiers., in *Proceedings of the International Workshop on Bio-ontologies.* 2009: Stockholm, Sweden.