A Falsification approach to create and check ontology definitions

Citlalli Mejía-Almonte, Julio Collado-Vides
Computational Genomics Program
Center for Genomic Sciences, UNAM
Cuernavaca, México

***Abstract*— One of the most important features of ontological representation of knowledge is the possibility of creating formal definitions that allow automatic reasoning. Reasoning in ontologies is based on symbolic logic representation. This requires that ontological definitions state either necessary conditions or necessary and sufficient conditions. Here we propose a manual approach to review the necessity and sufficiency of ontological definitions that can be used to analyze the most prominent concepts of a domain.**

Keywords—falsification; ontology definition; necessary and sufficient conditions

I. INTRODUCTION

Since the publication of the Gene Ontology, Biomedical ontologies have thrived. As a result, a growing number of ontologies are created to represent all aspects of the biological world. Currently there are 182 ontologies in OntoBee [1] and 716 in BioPortal [2], the OBO foundry [3] and the NCBO [4] ontology repositories respectively. Some of these ontologies are foundational, for they are species-independent models aimed to be reused in or extended by species-specific ontologies. Although categorization of ontologies into species dependent and species independent is not straightforward if authors have not established it in the scope description, we found 57 species-independent, 36 taxonomically restricted (at higher taxonomic ranges), 19 whose scope does not include biological entities, and 63 species-specific ontologies in OntoBee. When authors did not specify taxonomic range, this classification was based on the next criteria: species-independent if the ontology includes classes representing organisms of more than one kingdom, and species-specific if the ontology is human-centric.

This large set of computational models can provide the means for automatic reasoning to generate mechanistic hypothesis for the biomedical research [5]. However, foundational, species-independent ontologies must have formal definitions general enough to support pertinent inferences throughout all kingdoms of life.

Here we present a manual approach to check the suitability of necessity and sufficiency of ontological definitions for the current state of affairs in biological sciences. This allowed us to find out that if we consider natural language definitions of extant foundational ontologies as necessary and sufficient conditions, some prokaryotic instances may be left out.

II. METHODS

Ontological primitive classes are described only by necessary conditions, whereas defined classes are described by necessary and sufficient conditions [6]. Necessary and sufficient conditions are explained in terms of the conditional logical relation. Let A be a class or concept and let P be some property. There are many language items to refer to this [7]:

- A only if P; if A, then P; P is necessary for A; and A is sufficient for P.

Any of these statements means that all instances of A satisfy property P, or that for all objects of the universe, if some satisfies P then it is an instance of A. When this logical condition holds in both directions, that is:

- A is necessary and sufficient condition for B and B is necessary and sufficient condition for A

We say that A means B, or A is equivalent to B. This relation of equivalency is the one we look for to make ontological definitions.

Necessity of P is proved by demonstrating that all instances of A have property P. However, demonstration of necessity is epistemologically impossible in experimental sciences, even assuming an agent with the complete knowledge of the current state of affairs. Thus, we took a falsification approach [8].

- We can disprove sufficiency by finding some object that has property P and does not belong to A.
- We can disprove necessity by finding some instance of A that does not hold property P.

Based on this, we propose the following workflow to analyze necessity and sufficiency of proposed definitions:

- Retrieve definitions from diverse sources such as the literature and extant ontologies.
- Based on the retrieved definitions, generate a list of the commonly used properties to define these concepts.
- Search counter examples for definitions to discard necessity or sufficiency of the defining properties.

- Keep those properties that were not falsified to generate a new definition.

III. RESULTS

As a matter of example, we apply this approach to the definition of bacterial promoter in the sequence ontology (SO) [9]. The following are the two relevant definitions extracted from this ontology in July 2018:

- Promoter: A regulatory_region composed of the TSS(s) and binding sites for TF_complexes of the basal transcription machinery
    - Bacterial RNA-polymerase promoter: A DNA sequence to which bacterial RNA polymerase binds, to begin transcription.

Bacterial RNA-polymerase promoter is a subclass of promoter. Thus, the list of properties that define a Bacterial RNA-polymerase promoter is:

- has part some TSS
- has part some basal TF binding sites
- initiates some transcription
- binds some RNA polymerase

If we assume that basal transcription factor (TF), which is a term most commonly used in the domain of eukaryotic gene regulation, is equivalent to the most common sense in which transcription factor term is used in the domain of prokaryotic gene regulation, then "has part basal TF binding site" is not a necessary condition, since we can find counter examples in constitutive promoter sequences [10] that transcribe without the need of any transcription factor, and promoters of endosymbionts, whose reduced genome has been found to have lost most of the regulation by means of transcription factors [11]. On the other hand, from the biological point of view the closest to those "basal TFs" would be sigma factors. In this case, definition is correct and just have to be more explicitly specified in the definition.

A. Automatic logical consistency check is not suitable to detect these lack of generality

We are aware that logical consistency is one of the main applications of automatic reasoning [12]. However, the necessity of a restriction is more an issue of ontological commitment [13] that would be dropping out some class instances, owing to the lack of generality of definitions.

That is, if, in the first assumption scenario (i.e., basal transcription factors are bacterial transcription factors), we reuse the current conceptualization of SO and then create an instance or a subclass representing a specific promoter lacking the TF binding site constraint, either no logical inconsistency will rise owing to the open world assumption [6] or the reasoner will fail to infer the subsuming relation and we are going lose track of this entity as a promoter.

We are currently applying this approach to generate an ontology on prokaryotic gene regulation. In the process, we are reviewing the applicability of definitions of the existing ontologies. This step-by-step workflow can ease up the involvement of domain-experts in the generation of logically-sound ontological definitions based on ontological realism. However, we have not planned any training session to help other groups to check their ontological definitions.

IV. LIMITATIONS

This approach can be useful to apply OBO principle of maintenance [3]. However, as it requires huge human effort, we believe it could be applied in a top-down approach to check for the necessity and sufficiency of the most general or prominent concepts of a domain.

ACKNOWLEDGMENT

C.M.A. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, UNAM, receives fellowship 576333 from CONACYT and received financial aid from Programa de Apoyos para Estudios de Posgrado (PAEP) for this conference.

REFERENCES

[1] Xiang Z, Mungall C, Ruttenberg A, He Y. Ontobee: A Linked Data Server and Browser for Ontology Terms. *Proceedings of the 2nd International Conference on Biomedical Ontologies (ICBO)*, July 28-30, 2011, Buffalo, NY, USA. Pages 279-281. URL: http://ceur-ws.org/Vol-833/paper48.pdf.

[2] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., & Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, *37*(suppl_2), W170-W173.

[3] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., ... & Leontis, N. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, *25*(11), 1251.

[4] Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M. A., ... & NCBO team. (2011). The national center for biomedical ontology. *Journal of the American Medical Informatics Association*, *19*(2), 190-195.

[5] Hunter, L. E. (2018). Mechanistic hypothesis generation in molecular biology: A grand challenge for knowledge-based reasoning.

[6] Horridge, M., Knublauch, H., Rector, A., Stevens, R., & Wroe, C. (2004). A practical guide to building OWL ontologies using the Protégé-OWL plugin and CO-ODE tools edition 1.0. *University of Manchester*.

[7] Devlin, K. Introduction to Mathematical Thinking. [Week 2: Equivalence] MOOC offered by Stanford University through Coursera. Retrieved June 30th, 2018 from https://www.coursera.org/learn/mathematical-thinking/lecture/A5msF/lecture-4-equivalence

[8] Popper, Karl. *The logic of scientific discovery*. Routledge, 2005.

[9] Mungall, Christopher J., Colin Batchelor, and Karen Eilbeck. "Evolution of the Sequence Ontology terms and relationships." *Journal of biomedical informatics* 44.1 (2011): 87-93.

[10] Liang, S-T., et al. "Activities of constitutive promoters in Escherichia coli1." *Journal of molecular biology* 292.1 (1999): 19-37.

[11] Miravet-Verde, S., Lloréns-Rico, V., & Serrano, L. (2017). Alternative transcriptional regulation in genome-reduced bacteria. *Current opinion in microbiology*, *39*, 89-95.

[12] Hunter, L. E. Knowledge-based biomedical Data Science. *Data Science*, (Preprint), 1-7.

[13] Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology?. In *Handbook on ontologies* (pp. 1-17). Springer, Berlin, Heidelberg