The coevolution of ontologies and knowledge-based analytics in bioinformatics

Robert Hoehndorf

Extended Abstract

Developing high quality ontologies is expensive, and, like most infrastructure components of the life sciences, ontologies have evolved in response to specific needs and requirements of the biomedical community. At the same time, new tools utilizing ontologies emerged to enable or improve analysis of biological data. In my talk, I will explore how bio-ontologies have evolved in response to a changing bioinformatics environment and how bioinformatics tools and methods evolved in response to changing ontologies; my main aim will be to characterize the current changes in bioinformatics through large-scale application of machine learning, and how ontologies have to change to accommodate these changes.

The Gene Ontology (GO), the first bio-ontology that was and still is widely used, emerged as a consequence of breakthroughs in gene and genome sequencing and the resulting understanding of how many genes are conserved in different organisms [2]. This novel understanding, combined with the rapid change of knowledge in the field of molecular biology, necessitated the development of the GO, to keep track of the changing knowledge in the field and simultaneously provide a means to describe our knowledge of gene and protein functions. Using the GO for describing protein functions solved many challenges. A form of deductive inference ("true path rule") allowed capturing the most specific information about a protein as possible while still allowing inference of more general information, and use of a taxonomy allowed knowledge to evolve by gradually adding more specific functions to a protein without invalidating previous assertions.

Today, some of the most exciting developments (and challenges) in bio-ontologies still occur in fields where novel experimental techniques are leading to a radical change of our understanding of biological phenomena. For example, recently, our understanding of cell types has changed drastically, resulting from single cell sequencing technologies and the resulting detailed information available about cell types and their relations; ontologies of cell types had to change accordingly [15], and cell ontologies are now one of the most active areas of bio-ontology development (as evidenced, for example, by the regular CELLS workshop co-lated with the International Conference on Biomedical Ontologies).

Yet, what the early development of the GO (and similar ontologies) has shown is that the development and evolution of ontologies in life sciences is not a one-way road and only determined by changes in experimental techniques; rather, the availability of ontologies has also led to novel computational analysis methods, and ontologies will change in response to the emergence of novel methods. Two methods are particularly noteworthy here, ontology enrichment analysis and semantic similarity measures. Both techniques are some of the most widely used computational analysis methods involving ontologies. An ontology enrichment analysis uses an ontology together with its annotations in order to determine whether there is a function that is statistically enriched in a set of genes or gene products [8, 24, 16]. Ontology-based semantic similarity measures utilize the knowledge contained in ontologies (in particular within the formal axioms) to define measures of similarities between ontology classes, sets of classes, instances of classes, or entities annotated with (sets of) classes [9]. Semantic similarity was first used to query for and retrieve "semantically" related proteins [13], and later extended to find other entities with some association using a "guilt by association" approach.

My key take away message from these methods is that bioinformatics has developed a set of computational methods that crucially relied on ontologies providing accurate results. Both enrichment analysis and semantic similarity require that inferences in ontologies, in particular inferences about annotated genes or gene products (the "true path rule" in GO and more elaborate versions of this rule), are *accurate* (accurate in the sense that they are biologically correct and experimentally verifiable). Early ontologies did not always produce accurate inferences [20, 19, 21, 5], and finding these incorrect inferences has, arguably, led to one of the most active periods for ontology development and quality improvement, where the community applied and developed methods inspired, among others, by philosophy [23], linguistics [3], and logics [10].

With further improvement in experimental methods, in particular the emergence of high throughput sequencing methods, the demands on ontologies rose further, both in terms of their accuracy as well as in their detail and discriminatory power. Ontologies now had to cope with Big Data, and manually building ontologies would no longer scale in many domains. In this time, ontology design patterns [14], upper ontologies [18, 22], more and elaborate ontology design principles and community standards allowed ontologies to "scale up" both to capturing Big Data and more detailed nuances in biological phenomena. The new problem arose that our tools (reasoners and ontology editors such as Protege) no longer scaled to the new size and complexity of ontologies. The solution was to switch to different tools like Elk [11], and apply modularization techniques such as MIREOT [7]; while these work in solving the problem of scalability to Big Data, they have also hidden (and lost) some information; automated reasoners such as Elk only consider a tiny subset of the language we use to formalize ontologies, and modularization techniques can hide inconsistencies and therefore allow inconsistencies to increase [17].

As a result, a switch took place within the bio-ontologies community and the focus was no longer only on "ontologies" as formal artifacts capturing domain knowledge accurately, but rather on constructing "knowledge graphs" in which the focus is on linking information in some (vaguely) meaningful manner. The tendency to focus more on "knowledge graphs" instead of ontologies was by no means universal but certainly noticeable and still ongoing today. The move was motivated by the desire to focus on "relatedness" instead of precision, and find ways to integrate (i.e., link) large amounts of resources, in particular in the biomedical domain; the resources that are linked were often not ontologies but (medical) terminologies, so that ontological precision may have been an obstacle to successful integration. At the same time, and further motivating the focus on knowledge graphs instead of ontologies, novel knowledge graph analytics approaches emerged, in particular machine learning methods that would operate directly on graphs or knowledge graphs [25, 1], and graph neural networks that can exploit the knowledge graphs for various tasks [26]. In particular, knowledge graph embedding methods have been adopted widely within the bioinformatics community to exploit information in knowledge graphs for predictive or analytical tasks. Several knowledge graph embedding methods have been developed [25], but some of the most popular are based on the principle that, if the fact $r(a, b)$ is in the knowledge graph, then $\vec{a} + \vec{r} \approx \vec{b}$ (where $\vec{a}$ etc. are the "embedding" vectors of some dimension that "represent" a, r, and b in a distributed manner) [4]. The advantage of these embedding methods is their interpretability, simplicity, and almost universal applicability.

The role of ontologies in graph-based machine learning methods (such as knowledge graph embeddings, or graph neural networks) is to provide a source of nodes, and the formal axioms in the ontologies provides a source of relatedness (edges) that make up the resulting graph [6]. Yet, many aspects that have been considered crucial in developing ontologies are lost, specifically all benefits arising from semantics, both logical and ontological [12]: the ability for complex queries; ensured consistency; and deductive inference. In particular deductive inference (which is required both for complex queries and determining consistency) is crucial for exploring the knowledge ontologies contain beyond what has been explicitly asserted, but this ability for deductive inference is largely lost in graph-based methods.

Before ontologies (considered here as artifacts which explicitly and formally specify a conceptualization of a domain using a logic-based language) can become relevant in machine learning in bioinformatics, methods that can utilize the semantics of ontologies need to first be developed, because very few such methods exist in the field of AI; and it is even more of a challenge to tune such methods to the specific peculiarities of bio-ontologies which have distinct properties when compared to ontologies used in other domains, in particular computer science.

Some new methods emerged over the past years that apply machine learning methods to bio-ontologies. While some of these methods are simple extensions of learning from graph-structured data or learning from text, more recent approaches aim to explicitly address the missing formal semantics in machine learning models. These neuro-symbolic methods can produce deductive inferences directly, either by implementing a deduction system using neural approaches or by generating model structures using neural approaches. Establishing this correspondence between classical semantics and neural networks enables novel applications and demands on ontologies, but also opens novel opportunities, both for bioinformatics and AI. In bioinformatics, these methods allow machine learning to utilize the vast and rich knowledge contained in bio-ontologies thereby endowing the machine learning models with domain knowledge (and the ability to explore the knowledge more deeply than would be possible using only knowledge graphs), which can be used to provide access to the results of over a hundred years of experiments that are now contained in ontologies and knowledge bases. One of the most obvious areas of application are rare diseases where only little training data will ever be available. For AI, bio-ontologies provide a vast und largely underused resource of knowledge with direct implications for health, the environment, and well-being.

References

[1] Mehdi Ali, Charles Tapley Hoyt, Daniel Domingo-Fernández, Jens Lehmann, and Hajira Jabeen. BioKEEN: a library for learning and evaluating biological knowledge graph embeddings. *Bioinformatics*, 35(18):3538–3540, February 2019.

[2] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, Michael J. Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie I. Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[3] Michael Bada and Lawrence Hunter. Enrichment of OBO ontologies. *Journal of Biomedical Informatics*, 40(3):300–315, June 2007.

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[5] Werner Ceusters, Peter Elkin, and Barry Smith. Referent tracking: The problem of negative findings. *Stud Health Technol Inform*, 2006.

[6] Jiaoyan Chen, Pan Hu, Ernesto Jimenez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. OWL2Vec*: embedding of OWL ontologies. *Machine Learning*, June 2021.

[7] Mélanie Courtot, Nick Juty, Christian Knüpfer, Dagmar Waltemath, Anna Zhukova, Andreas Dräger, Michel Dumontier, Andrew Finney, Martin Golebiewski, Janna Hastings, Stefan Hoops, Sarah Keating, Douglas B. Kell, Samuel Kerrien, James Lawson, Allyson Lister, James Lu, Rainer Machne, Pedro Mendes, Matthew Pocock, Nicolas Rodriguez, Alice Villeger, Darren J. Wilkinson, Sarala Wimalaratne, Camille Laibe, Michael

Hucka, and Nicolas Le Novère. Controlled vocabularies and semantics in systems biology. *Molecular systems biology*, 7(1), October 2011.

[8] Scott W Doniger, Nathan Salomonis, Kam D Dahlquist, Karen Vranizan, Steven C Lawlor, and Bruce R Conklin. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology*, 4(1):R7, 2003.

[9] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, 30(5):740–742, 2014.

[10] Robert Hoehndorf, Frank Loebe, Janet Kelso, and Heinrich Herre. Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. *BMC Bioinform.*, 8, 2007.

[11] Yevgeny Kazakov, Markus Krötzsch, and Frantisek Simancik. The incredible elk. *Journal of Automated Reasoning*, 53(1):1–61, 2014.

[12] Frank Loebe and Heinrich Herre. Formal semantics and ontologies - towards an ontological account of formal semantics. In Carola Eschenbach and Michael Grüninger, editors, *Formal Ontology in Information Systems, Proceedings of the Fifth International Conference, FOIS 2008, Saarbrücken, Germany, October 31st - November 3rd, 2008*, volume 183 of *Frontiers in Artificial Intelligence and Applications*, pages 49–62. IOS Press, 2008.

[13] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

[14] David Osumi-Sutherland, Melanie Courtot, James P. Balhoff, and Christopher Mungall. Dead simple OWL design patterns. 8(1), June 2017.

[15] David Osumi-Sutherland, Chuan Xu, Maria Keays, Adam P. Levine, Peter V. Kharchenko, Aviv Regev, Ed Lein, and Sarah A. Teichmann. Cell type ontologies of the human cell atlas. *Nature Cell Biology*, 23(11):1129–1135, November 2021.

[16] Mark D Robinson, Jörg Grigull, Naveed Mohammad, and Timothy R Hughes. FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, 3(1):35, 2002.

[17] Luke T. Slater, Georgios V. Gkoutos, and Robert Hoehndorf. Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies. *BMC Medical Informatics and Decision Making*, 20(S10), December 2020.

[18] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome Biol*, 6(5):R46, 2005.

[19] B. Smith and C. Rosse. The role of foundational relations in the alignment of biomedical ontologies. *Medinfo*, 11(Pt 1):444–448, 2004.

[20] B. Smith, J. Williams, and S. Schulze-Kremer. The ontology of the gene ontology. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 609–613, 2003.

[21] Barry Smith. Against fantology. In M. E. Reicher and J. C. Marek, editors, *Experience and Analysis. Proceedings of the 27th International Wittgenstein Symposium.*, volume 6, pages 153–170, 2005.

[22] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe R. Serra, Alan Ruttenberg, Susanna A. Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255, 2007.

[23] Barry Smith and Werner Ceusters. Ontological realism: A methodology for coordinated evolution of scientific ontologies. *Appl. Ontol.*, 5:139–188, August 2010.

[24] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[25] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

[26] Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in Genetics*, 12, July 2021.