Ontologies in Metabolomics

Schlegel DR, Ruttenberg A and Elkin PL*

Department of Biomedical Informatics, University at Buffalo, Buffalo, Newyork, USA

Introduction

Metabolomics studies the structure, function and relationships between biological and chemical entities. As we move toward systems biology we need to be certain that we are representing this knowledge consistently between studies and between laboratories. The application of ontology to metabolomics can improve the consistency of study data and can help link data using relationships that extend the computational capacity of the study data and enrich that knowledge source with a myriad of nationally available data to help fuel hypothesis driven laboratory based research.

Ontologies have been successfully used to map databases to each other so that they may be used more effectively. There are several databases used in metabolomics research, including general databases such as KEGG LIGAND [1,2], which contains information about chemical compounds, reactions, and enzymes relevant to life, and MetaCyc [3], a database of metabolic pathways; and organism-specific databases such as those for *E. coli* : the *E. coli* metabolome database (ECMDB, [4]) and EcoCyc [5]. Some of these databases (namely, MetaCyc and EcoCyc) also have ontological components.

Ontologies also bring formal structure to domain terminologies, adding confidence in their utility. At the core of this standardization is a hierarchically organized controlled vocabulary. The vocabulary is derived from the prevailing term usage within the domain, and is generally mediated among, and agreed upon by, the field's practitioners. The organization into a hierarchy allows for subtype-super type reasoning (i.e., subsumption), which is important for recognizing trends within classes of entities. Ontologies contain other relations among terms, representing scientifically vetted domain knowledge which may be used to understand what typifies terms and differentiates them from others. This standardization eliminates duplication of knowledge (as is often found in resources such as PubChem [6]), and if experiments are modeled on ontologies, leads to reproducible research.

Ontologies have had an extremely positive effect on several subfields within biology and biomedicine, partly since many of the most common ontologies evolve in tandem as part of the OBO foundry [7]. The ontologies most relevant to the field of metabolomics are all from the OBO foundry: the Chemical Entities of Biological Interest ontology (ChEBI), [8], the Gene Ontology (GO), [9], the Plant Ontology (PO), [10], and the Ontology for Biomedical Investigations (OBI) [11].

Within metabolomics, these ontologies are being used with greater frequency and effectiveness. Perhaps the most obvious place to apply ontologies is in tools and data we use with our tools. The BiNChE web tool uses ChEBI for enrichment analysis of raw metabolomics data, in order to allow scientists to better understand and sort through those data [12,13]. Enrichment analysis is also done using the PO in studies of plant metabolomics [14]. OntoMaton [15] is a tool allowing users to mark-up Google Spreadsheets documents with ontological data, and is used in the MetaboLights [16] database of metabolomics studies. The mzTab data format is used to communicate mass spectrometry data to a wider audience, through being usable in common tools like Excel and R. It uses ontological terms from GO and others to annotate terms for easy computational interaction with other data [17].

Ontologies have been used to find synergies across articles within the bibleome [18]. Marker Discovery databases indexed by ontologies such as GO and SNOMED CT have been used to try and identify the basic mechanism of Oncogenesis [1,13]. Ontologies have great promise as methods for identifying common meaning across the many ways researchers might represent a unit of knowledge.

Researchers are also using ontologies as curated knowledge to provide guidance for their studies. In studying LSD1's contribution to carcinogenesis through chromatin regulation, the authors of [19] used categories of biological processes specified in GO in performing signal pathways analysis. Terms from the PO have also been used for pathway analysis, in the study of heretosis in *Arabidopsis thaliana* [20].

In the future, as adoption of ontologies grows, the advantages to the research community will also continue to grow. For example, with OBI, research studies may be modeled directly on ontology (see Example Use 3 in [11]). As this becomes prevalent, we will be able to organize and find studies with specific sets of properties very quickly. This will allow us to easily identify studies which use the best practices and techniques, separating them from lesser work.

The robust ecosystem of ontologies in biomedicine already may be used to help researchers in conducting their individual research, but when combined with other technologies may be used as the basis of vast knowledge bases derived from the work of the entire community. Using natural language processing based techniques (e.g., [21]); one can understand the content of articles and build ontology-based indexes. These indexes might reveal new research questions. For example, should studies on several subtypes of a metabolite reveal a certain property, and then it may be the case that the parent metabolite has this property as well. Reasoning systems such as those based on description logics or other subsumption or hybrid reasoners (e.g., [22,23]) may be able to identify these new research questions automatically. As ontologies become more engrained in our research, it is the belief of the authors that there will be an incredible growth in understanding and the movement toward systems biology.

References

1. Elkin PL, Frankel A, Liebow-Liebling EH, Elkin JR, Tuttle MS, et al. (2012) Bio prospecting the bibleome: Adding evidence to support the inflammatory basis of cancer. Metabolomics 2:112.

2. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M (2002) LIGAND: database of chemical compounds and re-actions in biological pathways. Nucleic Acids Res 30: 402-404.

*Corresponding author: Elkin PL, Department of Biomedical Informatics, University at Buffalo, Buffalo, New York, USA, Tel: 716 888 4854; E-mail: elkinp@buffalo.edu

3. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krumme-nacker M, et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36: D623-D631.

4. Guo AC, Jewison T, Wilson M, Liu Y, Knox C, et al. (2013) Ecmdb: the *E. coli* metabolome database. Nucleic Acids Res 41: D625-D630.

5. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, et al. (2005) EcoCyc: A comprehensive database resource for *Escherichia coli*. Nucleic Acids Res 33: D334-D337.

6. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) Pubchem: Integrated platform of small molecules and biological activities. Annu Rep Comp Chem 4: 217–241.

7. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25: 1251-1255.

8. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, et al. (2008) ChEBI: A database and ontology for chemical entities of biological interest. Nucleic Acids Res 36: D344-D350.

9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: Tool for the unification of biology. Nat Genet 25: 25-29.

10. Plant Ontology Consortium (2002).The Plant Ontology™ Consortium and Plant Ontologies. Comp Funct Genomics 3: 137-142.

11. Brinkman RR, Courtot M, Derom D, Fostel J, He Y, et al. (2010) Modeling biomedical experimental processes with obi. J Biomedical Semantics 1: S7.

12. Moreno P, Beisken S, Harsha B, Muthukrishnan V, Tudose I, et al. (2015) BiNChE: A web tool and library for chemical enrichment analysis based on the ChEBI ontology. BMC bioinformatics 16: 56.

13. Elkin PL, Tuttle MS, Brown SH (2012) Completing the Metabolome. Metabolomics 2: e115.

14. Shameer K, Naika MB, Mathew OK, Sowdhamini R (2014) Poeas: Automated plant phenomic analysis using plant ontology. Bioinform Biol Insights 8: 209-214.

15. Maguire E, Gonza´lez-Beltra´n A, Whetzel PL, Sansone SA, Rocca-Serra P (2012) Ontomaton: a bioportal powered ontology widget for google spreadsheets. Bioinformatics 29: 525-527.

16. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, et al. (2012) MetaboLights an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res 4: D781-D786.

17. Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, et al. (2014). The mzTab data exchange for- mat: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol Cell Proteomics 13: 2765-2775.

18. Elkin PL, Tuttle MS, Trusko BE, Brown SH (2009) Bio Prospecting: novel marker discovery obtained by mining the bibleome. BMC bioinformatics 10: S9.

19. Hayami S, Kelly JD, Cho HS, Yoshimatsu M, Unoki M, et al. (2011) Overexpression of lsd1 contributes to human carcinogenesis through chromatin regulation in various cancers. Int J Cancer 128: 574-586.

20. Andorf S, Meyer RC, Selbig J, Altmann T , Repsil- ber D (2012) Integration of a systems biological network analysis and QTL results for biomass heterosis in *Arabidopsis thaliana*. PLoS One 7: e49951.

21. Elkin PL, Trusko BE, Koppel R, Speroff T, Mohrer D (2010) Secondary use of clinical data. Stud Health Technol Inform 155: 14-29.

22. Sansone SA, Schober D, Atherton H, Fiehn O, Jenkins H, et al. (2007) Metabolomics standards initiative: ontology working group work in progress. Metabolomics 3: 249-256.

23. Schlegel DR, Shapiro SC (2015) Inference graphs: Combining natural deduction and subsumption inference in a concurrent reasoner. AAAI 15: 579-585.