Supporting database annotations and beyond with the Evidence & Conclusion Ontology (ECO)

Marcus C. Chibucos[1]✉, Suvarna Nadendla[1], James B. Munro[1], Elvira Mitraka[1], Dustin Olley[1], Nicole A. Vasilevsky[2], Matthew H. Brush[2], Michelle Giglio[1]

[1]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD United States of America
[2]Ontology Development Group, Library, Oregon Health & Science University, Portland, OR United States of America

✉Corresponding author: mchibucos@som.umaryland.edu; (410) 705-0885; 801 W. Baltimore St., Baltimore, MD, 21201

***Abstract*—The Evidence & Conclusion Ontology (ECO) is a community standard for summarizing evidence in scientific research in a controlled, structured way. Annotations at the world's most frequented biological databases (e.g. model organisms, UniProt, Gene Ontology) are supported using ECO terms. ECO describes evidence derived from experimental and computational methods, author statements curated from the literature, inferences drawn by curators, and other types of evidence. Here, we describe recent ECO developments and collaborations, most notably: (i) a new ECO website containing user documentation, up-to-date news, and visualization tools; (ii) improvements to the ontology structure; (iii) implementing logic via an ongoing collaboration with the Ontology for Biomedical Investigations (OBI); (iv) addition of numerous experimental evidence types; and (v) addition of new evidence classes describing computationally derived evidence. Due to its utility, popularity, and simplicity, ECO is now expanding into realms beyond the protein annotation community, for example the biodiversity and phenotype communities. As ECO continues to grow as a resource, we are seeking new users and new use cases, with the hope that ECO will continue to be a broadly used and easy-to-implement community standard for representing evidence in diverse biological applications. Feel free to visit two ECO-sponsored workshops at ICBO 2016 to learn more: 1. "An introduction to the Evidence and Conclusion Ontology and representing evidence in scientific research" and 2. "OBI-ECO Interactions & Evidence".**

Keywords—annotation; biodiversity; biomedical investigation; conclusion; confidence; curation; evidence; experimental evidence; inference; provenance; sequence similarity.

I. Introduction

The Evidence & Conclusion Ontology (ECO) [1] summarizes types of scientific evidence associated with biological research. Evidence can arise from laboratory experiments, computational methods, manual literature curation, or other means. Researchers, biocurators, and database managers use this evidence to justify their conclusions and support resulting assertions, for example stating that a given protein has a particular function.

Summarizing evidence with ECO allows projects such as the UniProt-Gene Ontology Annotation (UniProt- GOA) project [2] to manage large volumes of annotations in a convenient fashion, as both data management and query applications are supported by systematically describing evidence. Because ECO terms are ontology terms, they contain standard definitions and are networked using defined relationships. Thus, associating research data with descriptions of evidence using ECO can allow, for example, faceted queries of large datasets and implementations of customized quality control mechanisms.

II. Essentials of ECO

A. Basic ECO structure

As depicted in Fig. 1, ECO comprises two high-level classes, 'evidence' (ECO:0000000) & 'assertion method' (ECO:0000217). The definition of 'evidence' is "a type of information that is used to support an assertion" and 'assertion method' is defined as "a means by which a statement is made about an entity" [1]. Together 'evidence' and 'assertion method' can be combined to describe both the support for an assertion and whether the assertion was generated by manual or automatic means. ECO terms descend mainly from the 'evidence' hierarchy. However, 'evidence' leaf terms are related to the 'assertion method' terms by the 'used_in' relationship. Thus, one can assert not only what evidence is used to support a particular assertion, but also whether the assertion was made by a human being or a computer (Fig. 1).

B. Traditional uses of ECO

Some traditional example applications of ECO are found in uses by the Gene Ontology [3]: (a) hierarchical ECO classes are used to support structured data queries; (b) when a protein is annotated based on sequence similarity to another annotated protein, the identity of that protein must be recorded in the annotation file along with the evidence from ECO; (c) quality control assessment can be enforced by only allowing certain annotations to terms from a given ontology to be supported by particular evidence types—lest such annotations be flagged for review; and (d) circular annotations based on computational predictions alone can be determined, and thus avoided. In the ways described above, ECO has been used by many databases (e.g. UniProt, model organisms, Gene Ontology, et cetera) to support protein annotations. However, ECO has additional uses.

C. Recent ECO term development

A growing number of resources/applications use ECO (more than 40 of which we are aware). ECO has recently expanded its

This work is supported by funding from the National Science Foundation Division of Biological Infrastructure under award number 1458400.

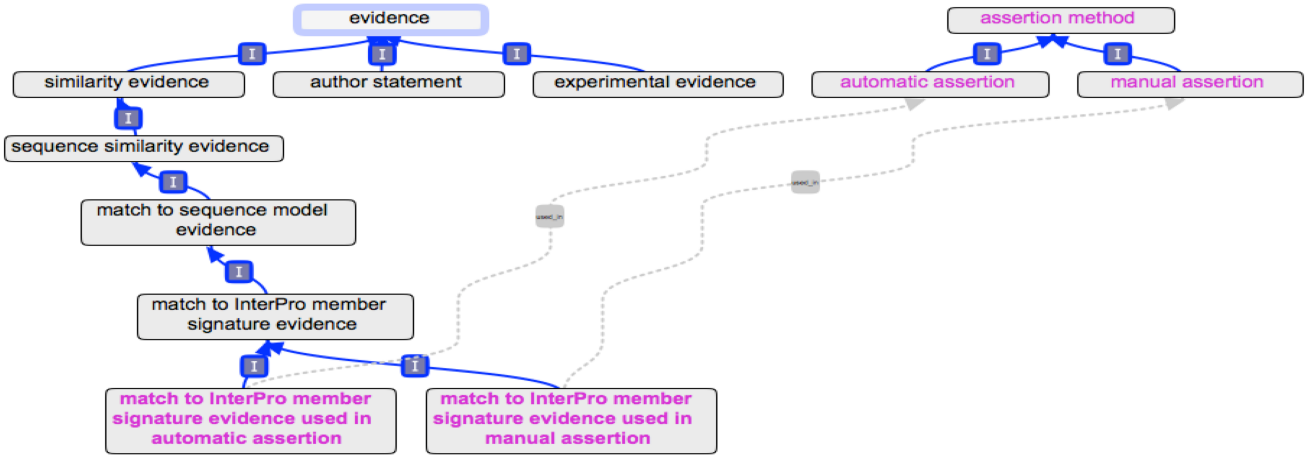

Fig. 1. ECO root classes and combinatorial terms. Leaf terms depicted are logically defined as the 'evidence' parent class ('match to InterPro member signature evidence') related to the 'assertion method' class via the 'used_in' relationship (gray boxes).

evidence representation through collaborations with many groups, for example: IntAct [4] (biological system reconstruction), CollecTF [5] (motif prediction), Ontology of Microbial Phenotypes [6] (microbial assays), Planteome (http://planteome.org; genotype-phenotype associations), Gene Ontology [3] (logical inference & synapse research techniques), SwissProt [7] (diverse experimental assays), and UniProt [2,7] (detection techniques).

III. The Future of ECO

A. *Increasing the logic within ECO*

In May 2016, 14 people met in person at the Institute for Genome Sciences in Baltimore, MD, while approximately seven others joined remotely, to discuss modeling scientific research evidence [8]. An objective of the meeting, titled "OBI-ECO Baltimore 2016: Evidence," was to devise strategies for cross-ontology coordination between ECO and the Ontology for Biomedical Investigations (OBI) [9]. One decided outcome of the meeting was to logically define ECO 'experimental evidence' classes using OBI classes. This work has been under way, and a cataloging of issues and areas for development in both ontologies has been undertaken. Followup discussions and a review of this ongoing work will take place at ICBO 2016 at workshop W08 titled "OBI-ECO Interactions & Evidence" and participation by any interested users is welcome.

B. *Beyond protein annotation*

Although ECO was originally created circa 2000 to support gene product annotation by the Gene Ontology, today ECO is used by many groups concerned with evidence, and even provenance, in scientific research. While numerous experimental and computational evidence types have been added to ECO on behalf of a number of resources (see above and www.evidenceontology.org), the ECO user base and diversity of applications continues to increase.

Some examples of new/potential ECO users include WikiData (https://www.wikidata.org), the deep sea community (https://github.com/geneontology/deep_sea), the biodiversity and phenotype communities, and the Disease Ontology [10]. Specific examples of these will be addressed at the ICBO 2016 workshop titled "An introduction to the Evidence and Conclusion Ontology and representing evidence in scientific research" (workshop W11) and new users and adopters are especially encouraged to attend to learn more.

Acknowledgment

The authors acknowledge the Ontology for Biomedical Investigations (OBI) Consortium and, in particular, Bjoern Peters for ongoing collaboration with ECO. We thank Christian J. Stoeckert, Jr. and Jie Zheng for co-organizing the ICBO 2016 workshop W08 titled "OBI-ECO Interactions & Evidence."

References

[1] M.C. Chibucos, C.J. Mungall, R. Balakrishnan, K.R. Christie, R.P. Huntley, O. White, J.A. Blake, S.E. Lewis, and M. Giglio, "Standardized description of scientific evidence using the Evidence Ontology (ECO)," Database (Oxford), v.2014:bau075, 2014.

[2] E.C. Dimmer, R.P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M.J. Martinet, et al., "The UniProt-GO Annotation database in 2011," Nucleic Acids Res., 40, D565–D570, 2012.

[3] The Gene Ontology Consortium, "Gene Ontology Consortium: going forward," Nucleic Acids Res., **43**(Database issue):D1049-1056, 2015.

[4] B.H.M. Meldal, O. Forner-Martinez, M.C. Costanzo, J. Dana, J. Demeter, M. Dumousseau, et al., "The complex portal – an encyclopaedia of macromolecular complexes," Nucleic Acids Res., nar.gku975, 2014.

[5] S. Kılıç, D.M. Sagitova, S. Wolfish, B. Bely, M. Courtot, S. Ciufo, et al., "From data repositories to submission portals: rethinking the role of domain-specific databases in CollecTF," Database, v.2016:baw055 2016.

[6] M.C. Chibucos, A.E. Zweifel, J. Herrera, W. Meza, S. Eslamfam, P. Uetz, et al., "An ontology for microbial phenotypes," BMC Microbiology, 14(1):294, 2014.

[7] The Uniprot Consortium, "UniProt: a hub for protein information," Nucleic Acids Res., **43**(Database issue):D204-212, 2015.

[8] The OBI Consortium, et al., "Cross-community ontological modeling of scientific evidence," unpublished.

[9] A. Bandrowski, R. Brinkman, M. Brochhausen, M.H. Brush, B. Bug, M.C. Chibucos, et al., "The Ontology for Biomedical Investigations," PLoS One, 11(4):e0154556, 2016.

[10] W.A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, et al., "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," Nucleic Acids Res., Oct 27. pii: gku1011, 2014