HHS Public Access

Author manuscript

Methods Mol Biol. Author manuscript; available in PMC 2019 February 15.

Published in final edited form as:

Methods Mol Biol. 2017 ; 1446: 245–259. doi:10.1007/978-1-4939-3743-1_18.

The Evidence Ontology: Supporting Conclusions & Assertions with Evidence

Marcus C. Chibucos,
Department of Microbiology and Immunology, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA.

Deborah A. Siegele,
Department of Biology, Texas A&M University, College Station, TX 77843, USA

James C. Hu, and
Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843, USA

Michelle Giglio
Department of Medicine, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA.

Abstract

The Evidence Ontology (ECO) is a community resource for describing the various types of evidence that are generated during the course of a scientific study, and which are typically used to support assertions made by researchers. ECO describes multiple evidence types, including evidence resulting from experimental (i.e. wet lab) techniques, evidence arising from computational methods, statements made by authors (whether or not supported by evidence), and inferences drawn by researchers curating the literature. In addition to summarizing the evidence that supports a particular assertion, ECO also offers a means to document whether a computer or human performed the process of making the annotation. Incorporating ECO into an annotation system makes it possible to leverage the structure of the ontology such that associated data can be grouped hierarchically, users can select data associated with particular evidence types, and quality control pipelines can be optimized. Today, over 30 resources, including the Gene Ontology, use the Evidence Ontology to represent both evidence and how annotations are made.

Keywords

annotation; biocuration; conclusion; confidence; evidence; ECO; experiment; inference; literature curation; quality control

The importance of describing evidence in scientific investigations

Scientific investigations routinely produce data from diverse methodologies using a wide range of tools and techniques. Such data generated during the course of a research project contribute to the pool of evidence that ultimately leads a scientific researcher to make a particular inference or draw a given conclusion. Ultimately, the goal of a scientist is to publish the conclusions that are drawn from a given research project in the scientific

literature. Such conclusions typically take the form of assertions, i.e. statements that are believed to be true, about some aspect of biology. The process of biocuration seeks to extract from the literature the **assertion** that summarizes the research finding *in addition to* any relevant **evidence** in support of the finding. Ideally, both of these pieces of information will become integrated into a database in a structured way, so that they are readily accessible to the scientific community (1,2) (Fig. 1).

Recording evidence is essential because: (i) knowing what methodologies were used is central to the scientific method and can impact one's evaluation of the data or results; (ii) associating evidence with data maintained electronically allows for selective data queries and retrieval from even the largest of databases; and (iii) a structured representation of evidence makes automated quality control possible, which is absolutely essential to managing the ever-increasing number and size of biological databases.

Evidence can be associated with assertions in many ways. Manual curation is a common approach, described in Balakrishnan, et al. 2013 (3) and outlined in Fig. 1. However, text mining or other computational methods can also be used to extract biological assertions from the scientific literature (4,5) and assertions can also be made directly via bioinformatic techniques, e.g. assigning of functional annotations as resulting from a functional genome annotation pipeline.

Numerous types of evidence form the bases for assertions that are made by researchers. Laboratory and field experiments are common sources of evidence, but computational (or *in silico*) analysis, whether executed by a person or an unsupervised machine, can also generate the evidence that is used to support assertions about biological function (Fig. 2). In addition, conclusions can be synthesized from investigator speculation or implied by known biology during the literature curation process. We can also consider evidence in the sense of *provenance*. A central goal of biological data repositories is to record in a structured fashion as much information as is known about the origins of a given accession. Yet, sometimes an accession is imported from another database where the source for the annotation at that database is unclear. Even in this case it might be useful for the importing database to note the source of the statement/annotation along with an evidence description of "imported information," indicating that nothing else is known about the evidence for that particular annotation. Thus there are numerous advantages to capturing scientific evidence, from describing specific methodologies to representing chains of custody.

The argument for an ontology of evidence

Due to the diversity of ways that exist to describe the multitude of scientific research methodologies, a means of representing evidence in a descriptive but structured way is required in order to maximize utility. The most efficient way to achieve this is to use an ontology, a controlled vocabulary where each term is well-defined and linked to other terms via defined relationships (6,7). In an ontological framework, evidence descriptions are represented not as free text, but rather as networked ontology classes where each child term is more specific (granular) than its parent. High-level descriptions of types of evidence (such as "experimental evidence") are contained in more basal classes closest to the root class

evidence. Increasingly specific terms that are grouped under the more general classes describe particular sub-types of evidence (such as “chromatography evidence”). The most specific terms, the so-called “leaf nodes” that contain no child terms, represent the most granular types of evidence generated during the course of a scientific investigation (for example “thin layer chromatography evidence”). **The Evidence Ontology (ECO)** (http://evidenceontology.org) was created to enable the structured description of experimental, computational, and other evidence types to support the assertions captured by scientific databases (8).

A brief history of the Evidence Ontology

As described throughout this book, the Gene Ontology (GO) uses controlled vocabularies to capture functional information about gene products. The need to systematically document evidence while curating annotations was recognized from the inception of the GO (9) and a set of “evidence codes” was created for this purpose (http://www.geneontology.org/page/guide-go-evidence-codes). In time it was realized that a better-structured and more comprehensive way to represent evidence was required. Thus, the set of initially created GO codes, along with terms created by two model organism databases, FlyBase (10) and The *Arabidopsis* Information Resource (11), evolved into the first version of ECO, the “Evidence Code Ontology”. Since then, the use of ECO by other resources has continued to grow and the ontology has shifted its focus beyond GO in order to become a generalized ontology for the capture of evidence information. The official name of ECO is now the “Evidence and Conclusion Ontology”. ECO is presently being developed to define and broaden its scope, normalize its content, and enhance interoperability with related resources. The GO remains an active user and participant in developing ECO. It is anticipated that soon the three letter GO evidence codes to which so many are accustomed will be replaced by ECO term identifiers.

Evidence Ontology structure & content

Evidence terms descend from the root class “evidence”, which is defined as “a type of information that is used to support an assertion” (Fig. 3). Most evidence terms are either experimental or computational in nature, e.g. “chromatography evidence” or “sequence similarity evidence”, respectively (Fig. 3). However, ECO also comprises other types of evidence, such as “curator inference” and “author statement”. In addition to describing evidence, ECO can also describe the means by which assertions are made, i.e. by a human or a machine. ECO calls this the “assertion method” and defines it as “a means by which a statement is made about an entity” (Fig. 1C, Fig. 2B, Fig. 3). For example, whether a curator makes an annotation after reading about a result in a scientific paper or after manually evaluating pairwise sequence alignment results, ECO can express that a manual curation method was used (3,8). Conversely, if an algorithm was used to assign a predicted function to a protein, ECO can express that an automated computational method was used. The current version of ECO comprises 630 terms that describe “evidence”, “assertion method”, or “evidence × assertion method” cross products. ECO architecture was recently described in Chibucos, et al., 2014 (8).

Fundamentals of evidence-based GO annotation

Creating an association between a GO term and a gene product is the fundamental essence of the GO annotation process. Documenting the evidence for any given GO annotation is a critical component of this annotation process, and an annotation would be incomplete without the requisite evidence. In fact, evidence capture by the GO requires both a "GO evidence code" that describes in detail the type of work or analysis that was performed in support of the annotation, as well as a citation for the reference from which the evidence was derived. Curators go to great lengths to understand and properly apply the correct "evidence code" to a given annotation, and an online guide exists to explain the often-subtle distinctions between multiple related evidence types (http://geneontology.org/page/guide-go-evidence-codes).

The online GO documentation contains specific usage notes and caveats concerning evidence, such as "although evidence codes do reflect the type of work or analysis described in the cited reference which supports the GO term to gene product association, they are not necessarily a classification of types of experiments/analyses. Note that these evidence codes are intended for use in conjunction with GO terms, and should not be considered in isolation from the terms." Thus, GO evidence codes are useful in themselves because they represent much more detailed descriptions of evidence types; however, they are maintained as a list with a rather shallow hierarchical structure that lacks the depth of a formal ontology. Furthermore, as conceptualized by the GO, the evidence codes must be considered within the context of the GO term being annotated to. Finally, the GO notes that all GO codes except for IEA (Inferred from Electronic Annotation) are to be assigned by a curator. Despite these caveats, it should be emphasized that each GO evidence code maps to an ECO term, as ECO maintains database cross references to the GO codes. GO codes therefore represent a subset of the Evidence Ontology. Since independent development of ECO was undertaken, a number of terms equivalent to GO evidence codes have already been instantiated in ECO, e.g. IBA, IBD, IKR, IRD, and ECO will continue to develop terms for the GO as they are needed.

Extending ECO beyond GO

Recent development efforts of ECO have emphasized meeting the needs of a larger research community, see for example (12–14), while still capturing the needed information for GO annotation, such as by adding comments and synonyms to a term. Many high-level ECO term definitions were written with explicit GO usage notes contained therein because ECO originated during early efforts of the GO. However, in order to increase overall usability of ECO by other resources than the GO, such verbiage has been removed, while retaining the essence of the term's meaning and applicability to GO. As the ECO has been developed, more and more granular terms have been created to represent increasingly complex laboratory, computational, and even inferential techniques.

A discussion of ECO and GO evidence codes would not be complete without mention of the GO evidence code IEA or "inferred from electronic assertion". IEA is used to connote that an annotation was assigned through automated computational means, i.e. transferring

database annotations to another database. To ECO, because IEA describes how an annotation was *assigned*, rather than the specific type of supporting evidence, a second root class was created in ECO called “assertion method” that conveys how an annotation was generated. “Assertion method” has two subclasses, “manual assertion” and “automatic assertion”, with the latter being equivalent to IEA. Now it is possible to more accurately model evidence and the annotation process using ECO.

Aside from rewording definitions and creating a second root class, the biggest conceptual modification of ECO is reflected by removal of the prefix “inferred from” from every term name (see the GO codes for a sense of how ECO terms were previously labeled). This was done because ECO considers not just inferences made during the curation process, per se, but other aspects of evidence documentation, such as what research methodologies were performed.

Benefits & applications of ECO to the GO

Despite broadening the scope of ECO to support projects outside of GO annotation and some conceptual shifts, the ECO is proving more useful to the GO than ever before. As the GO records evidence to support comprehensive high-quality annotations of gene products, it can derive benefits simply not possible without an ontology. There are currently over 365 million annotations in the GO repository linked to an evidence term, and these can be queried and maintained better with the help of an ontology by leveraging its hierarchical structure.

One of the most direct applications for using an ontology of evidence is *selective data query*, i.e. to query a database for records associated with a particular evidence type. For example, searching for “thin layer chromatography evidence” (a leaf term with no subclasses) would return only the records associated with that evidence type and no others. But *grouping annotations* is also possible with this approach. A query for “chromatography evidence” will return data associated not only with “chromatography evidence” but also its more specific subtypes including “thin layer chromatography evidence” and “high performance liquid chromatography evidence”.

But there are further benefits to be derived from an ontology of evidence beyond simple structured queries. For example:

1. The GO curatorial process uses evidence to support computable rules about the kinds of information that must be associated with different evidence types. For example, one rule states that annotation of a protein based on alignment with another protein requires that the identity of the matching protein be captured, along with the evidence type “protein alignment evidence”. If such an evidence type were missing, this would flag the annotation for review.
2. The GO uses evidence as a quality control mechanism for annotation consistency. For example, expression pattern evidence is restricted to annotations for terms from the “biological process” ontology. Annotations to terms from

either of the other two GO ontologies (“molecular function” or “cellular component”) would be flagged as suspect.

3. Evidence is used to prevent circular annotations based solely on computational predictions. Chains of evidence are computationally evaluated to ensure that inferential annotations are linked to experimental evidence. For example, annotations supported by “sequence alignment evidence” require the inclusion of a database identifier for a match gene product that is itself linked to an annotation supported by experimental evidence.

4. To amplify the benefits of experimental knowledge that curators capture, the GO Consortium is using a phylogenetic tree-based approach to generate manually reviewed, homology-based annotations for a range of species (15). This phylogenetic annotation methodology necessitated a new set of evidence terms to capture the inference process. Currently over 150,000 annotations are associated with these new terms and the number continues to grow.

Yet another application of ECO for the GO has been realized in the UniProt-GOA project. Arguably, UniProt is the most comprehensive and best-curated protein database available to the research community. EO terms have replaced the original UniProtKB (16) evidence types and are available in UniProtKB XML (8). Novel ways of mapping and extending ontologies have been discussed with EO and the Gene Ontology Consortium to ensure appropriate development for UniProtKB annotation. The UniProt-Gene Ontology Annotation (UniProt-GOA) project provides >169 million manual and electronic evidence-based associations between GO terms and 26.5 million UniProtKB proteins covering >411,000 taxa (17). Of these, manual annotation provides 1.4 million annotations to ~260,000 proteins. Since 2010, UniProt-GOA has supplied GO annotations in a Gene Product Association Data (GPAD) file format, which allows inclusion of EO terms. Because EO terms are cross referenced to corresponding GO codes, even if evidence for annotations was supplied to UniProt as GO codes, the GPAD file will display the appropriate equivalent EO term. Thus, UniProt annotations can be grouped by leveraging the structure of EO.

The future of ECO & GO

In summary, an evidence ontology can be used to support faceted queries of data, to establish computable rules about required types of evidence, as a quality control check for annotation consistency, and as a mechanism to prevent circular annotations rooted only in computational predictions. GO is already benefitting from these applications of ECO.

But what else can an evidence ontology do? One area of active exploration for ECO is in the area of confidence or quality of evidence. Work has begun (1) (http://wiki.isbsib.ch/biocuration/Quality_codes) to develop a mechanism to incorporate quality information into ECO or, as needed, to create a standalone system. It might one day be possible to describe using ECO the *quality* of the evidence supporting an annotation in addition to the *type* of evidence that supports the annotation.

Numerous other applications of ECO to the Gene Ontology are already being realized, and the future promises both additional new applications of ECO as well as advancements to current ones.

References

1. Gaudet P, Arighi C, Bastian F, et al. (2012) Recent advances in biocuration: meeting report from the Fifth International Biocuration Conference. Database (Oxford), 2012, bas036. [PubMed: 23110974]
2. Burge S, Attwood TK, Bateman A, et al. (2012) Biocurators and biocuration: surveying the 21st century challenges. Database (Oxford), 2012, bar059. [PubMed: 22434828]
3. Balakrishnan R, Harris MA, Huntley R, Van Auken K, Cherry JM (2013) A guide to best practices for Gene Ontology (GO) manual annotation. Database (Oxford), 2013, bat054. [PubMed: 23842463]
4. Arighi CN, Carterette B, Cohen KB, et al. (2013) An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. Database (Oxford), 2013, bas056. [PubMed: 23327936]
5. Altman RB, Bergman CM, Blake J, et al. (2008) Text mining for biology--the way forward: opinions from leading scientists. Genome Biol, 9 Suppl 2, S7.
6. Smith B (2003) Ontology In Floridi L (ed.), Blackwell Guide to the Philosophy of Computing and Information. Blackwell, Oxford, pp. 155–166.
7. Smith B (2008) Ontology (Science) In Eschenbach C and Grüninger M (eds.), Formal Ontology in Information Systems. IOS Press, pp. 21–35.
8. Chibucos MC, Mungall CJ, Balakrishnan R, et al. (2014) Standardized description of scientific evidence using the Evidence Ontology (ECO). Database (Oxford), 2014, bau075. [PubMed: 25052702]
9. Ashburner M, Ball CA, Blake JA, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics, 25, 25–9. [PubMed: 10802651]
10. The FlyBase Consortium. (2002) The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic acids research, 30, 106–8. [PubMed: 11752267]
11. Huala E, Dickerman AW, Garcia-Hernandez M, et al. (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic acids research, 29, 102–5. [PubMed: 11125061]
12. Kilic S, White ER, Sagitova DM, Cornish JP, Erill I (2014) CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. Nucleic Acids Res, 42, D156–60. [PubMed: 24234444]
13. Chibucos MC, Zweifel AE, Herrera JC, et al. (2014) An ontology for microbial phenotypes. BMC Microbiol, 14, 294. [PubMed: 25433798]
14. Stockinger H, Altenhoff AM, Arnold K, et al. (2014) Fifteen years SIB Swiss Institute of Bioinformatics: life science databases, tools and support. Nucleic Acids Res, 42, W436–41. [PubMed: 24792157]
15. Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput Biol, 5, e1000431. [PubMed: 19578431]
16. UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res, 42, D191–8. [PubMed: 24253303]
17. Dimmer EC, Huntley RP, Alam-Faruque Y, et al. (2012) The UniProt-GO Annotation database in 2011. Nucleic acids research, 40, D565–70. [PubMed: 22123736]

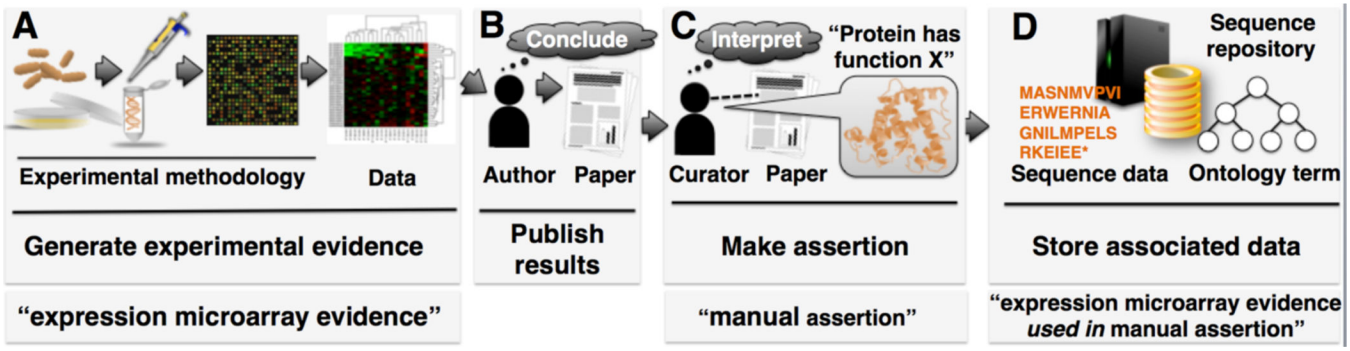

Figure 1. Representing experimental methods and conclusions in a biological database.

(**A**) An experiment is performed that generates data. (**B**) A researcher interprets methods & data, and draws conclusions that are published in a scientific journal. (**C**) A biological curator reads that paper, interprets the results presented therein, and makes an assertion. (**D**) The assertion is represented by an ontology term and stored along with the protein sequence and other data at a biological database. (General evidence and assertion summaries are depicted at the bottom.)

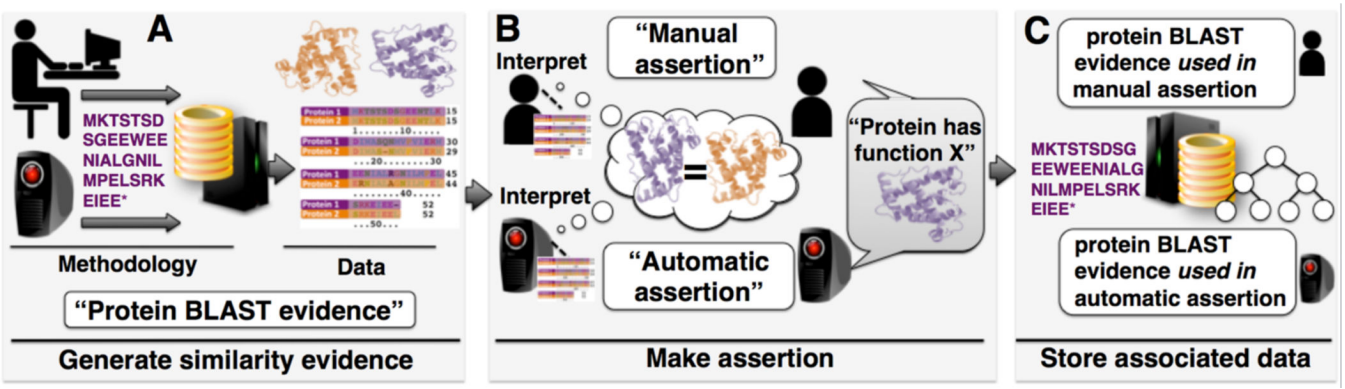

Figure 2. Computational evidence and assertion.

(A) A human or computer performs an analysis, for example comparing the sequence of a protein of unknown function to sequences at a database. A protein of known function is returned as a hit. (B) The alignment is analyzed and the protein sequences share enough similarity to be considered homologs (related through common evolutionary descent). The query protein is assigned the same function as the database protein. (C) This information is stored at a sequence repository along with other data and metadata. (Text in white boxes depicts evidence and assertion methods used in this process.)

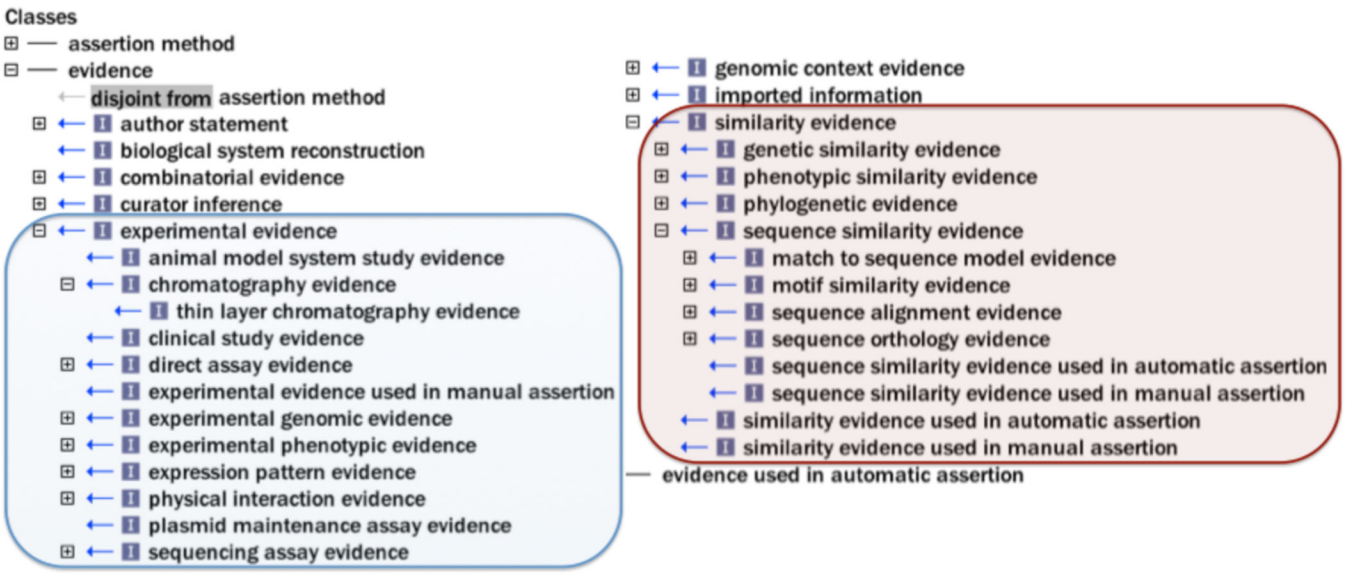

Figure 3. High-level Evidence Ontology (ECO) classes.

ECO comprises two root nodes, “evidence” and “assertion method”. “Experimental evidence” and selected subclasses are highlighted in blue. “Similarity evidence” and selected subclasses are highlighted in red.