Chapter 18

The Evidence and Conclusion Ontology (ECO): Supporting GO Annotations

Marcus C. Chibucos, Deborah A. Siegele, James C. Hu, and Michelle Giglio

Abstract

The Evidence and Conclusion 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 a 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 and Conclusion Ontology to represent both evidence and how annotations are made.

Key words Annotation, Biocuration, Conclusion, Confidence, Evidence, ECO, Experiment, Inference, Literature curation, Quality control

1 Describing Evidence in Scientific Investigations

1.1 Importance of Documenting Evidence

Investigations in the life sciences 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, one 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

Christophe Dessimoz and Nives Škunca (eds.), *The Gene Ontology Handbook*, Methods in Molecular Biology, vol. 1446,
DOI 10.1007/978-1-4939-3743-1_18,

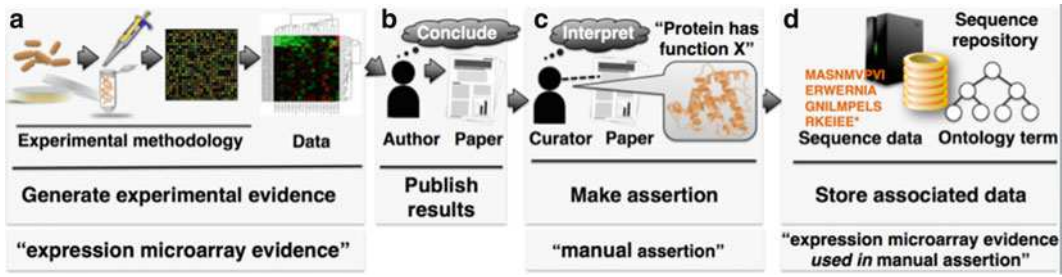


Fig. 1 Representing experimental methods and conclusions in a biological database. (**a**) An experiment is performed that generates data. (**b**) A researcher interprets methods and data, and draws conclusions that are published in a scientific journal and indexed in PubMed, for example. (**c**) A biocurator reads that paper, interprets the results presented therein, and makes an assertion. (**d**) The assertion is represented by associating an ontology term with the item being studied and stored along with other data, for example a protein sequence, at a biological database. (General summaries and related ECO classes are depicted along the bottom.)

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: (1) knowing what methodologies were used is central to the scientific method and can impact one's evaluation of the data or results; (2) associating evidence with data maintained electronically allows for selective data queries and retrieval from even the largest of databases; and (3) 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.

1.2 Multiple Types of Evidence and Ways of Associating Evidence with Assertions

Evidence can be associated with assertions in many ways. Manual curation is a common approach [3, 4], outlined in Fig. 1. However, text mining or other computational methods can also be used to extract biological assertions from the scientific literature [5, 6], and assertions can also be made directly via bioinformatic techniques [7], 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 *provenance*, a concept related to and sometimes conflated with evidence. 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

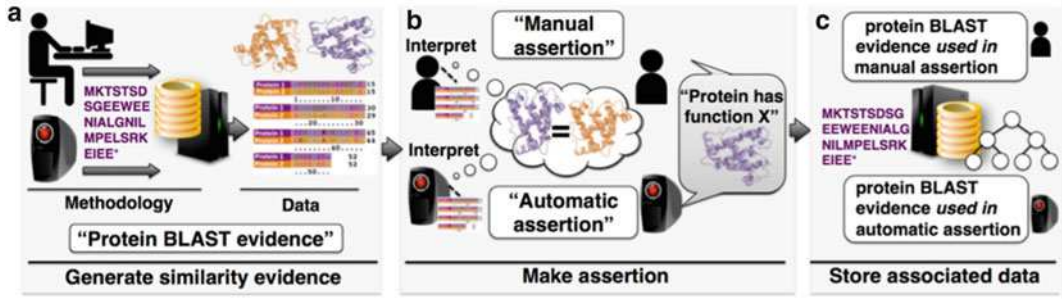


Fig. 2 Computational evidence and assertion. (**a**) A human or computer performs an analysis, for example comparing the sequence of a protein of unknown function to sequences at a database. A protein of known function is returned as a hit with corresponding alignment. (**b**) The alignment is analyzed and the protein sequences are deemed to 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.)

unclear. Even in this case it might be useful for the importing database to note the source of the statement/annotation along with a description of "imported information," indicating that nothing else is known about the evidence or provenance of that particular annotation. Thus there are numerous advantages to capturing scientific evidence and provenance, from describing specific methodologies to representing chains of custody.

2 The Evidence and Conclusion Ontology (ECO)

2.1 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 [8, 9]. 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 [10]. 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 and Conclusion 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 [11].

2.2 A Brief History of ECO

As described throughout this book, the Gene Ontology (GO) uses terms organized into controlled vocabularies, and the relationships among these terms, to capture functional information about gene products. The need to systematically document evidence while curating annotations was recognized from the inception of the GO [12] and a set of "evidence codes" was created for this purpose [13]. 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 [14] and The *Arabidopsis* Information Resource [15], 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.

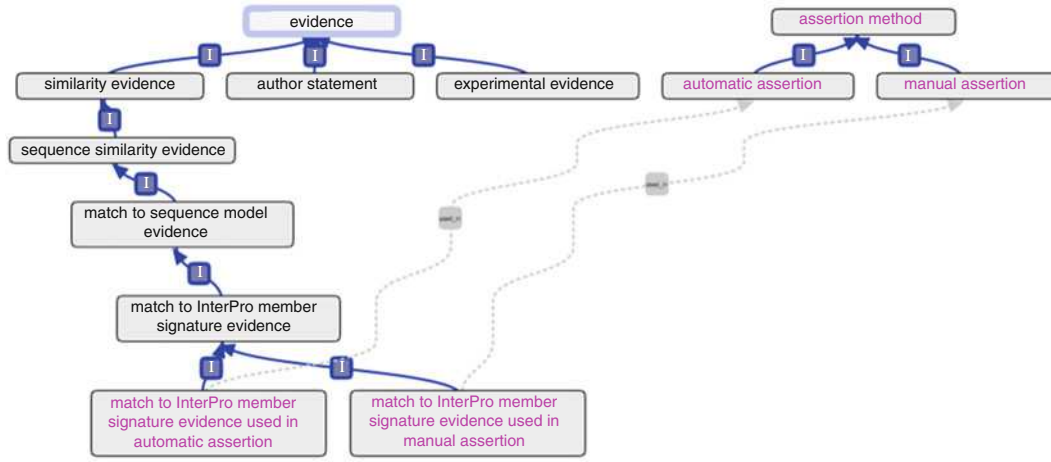


Fig. 3 Simplified representation of ECO, depicting general structure. ECO comprises two root classes along with their respective hierarchies, *evidence* (terms in *black*) and *assertion method* (terms in *pink*). A given type of evidence can be applied to (*used_in*; *dotted lines*) automatic assertion or manual assertion, which necessitated the creation of ECO leaf nodes that are *evidence* x *assertion method* cross products. For simplicity, most ECO classes are not displayed in the figure, including, for example, five of eight direct subclasses of *evidence* or three of four types of *similarity evidence* and so on

2.3 ECO Structure and 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" (Figs. 1c and 2b). For example, whether a curator makes an annotation after reading about an experimental 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. Thus "assertion method" forms a second root class with two branches: "manual assertion" and "automatic assertion" (Fig. 3).

The current version of ECO comprises 630 terms that describe "evidence", "assertion method", or "evidence x assertion method" cross products. Ontology architecture of ECO was recently described in Chibucos et al. [11].

2.3.1 Extending ECO Beyond GO

Recent development efforts of ECO have emphasized meeting the needs of a larger research community; *see* for example [11, 16], 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 resources other than the GO, such verbiage has been removed, while retaining the essence of the term's meaning and applicability to GO. As 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 would not be complete without mention of the GO evidence code IEA or "inferred from electronic annotation". IEA is used to connote that an annotation was assigned through automated computational means, e.g., transferring annotations from one protein to another. Because IEA describes how an annotation was *assigned*, rather than the specific type of supporting evidence, this term belongs as a subclass of "assertion method". As described above, "assertion method" has two child terms, "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.

3 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) [4, 13].

The GO gene association file (GAF) format contains required columns for both evidence code and reference. Each GO evidence code maps directly to an ECO term. ECO maintains database cross references to the GO codes for easy mapping between systems. GO codes therefore represent a subset of the Evidence and Conclusion Ontology. Since independent development of ECO was undertaken, a number of new GO evidence codes have been created, e.g., IBA, IBD, IKR, IRD. Equivalent terms have been instantiated in ECO (Fig. 4a), which will continue to develop such terms for the GO.

3.1 ECO Terms Versus GO Codes

Although GO evidence codes are useful in themselves because they represent detailed descriptions of evidence types, they are maintained as a controlled vocabulary with a shallow hierarchical structure that lacks the advantages of a formal ontology like ECO. Further, the full set of terms within ECO provides the ability to capture more breadth and depth of evidence information than the GO evidence codes do. Additionally, as the field of biocuration evolves and the kinds of evidence being curated from the literature continue to grow both more detailed and nuanced, the number of two- and three-letter acronyms (e.g., IEA, IMP, EXP, and ISS) available for new terms will hit an upper limit (there are only 676 possibilities using all 26 two-letter combinations, as the first letter of the three-letter GO codes often stands for "inferred"). In fact, ECO developers have already received requests from different users to develop new, but unrelated, terms that had the same suggested three-letter acronyms. For all of these reasons, there are discussions underway about transitioning GO evidence storage to use ECO terms rather than GO evidence codes. Such a shift would combine the

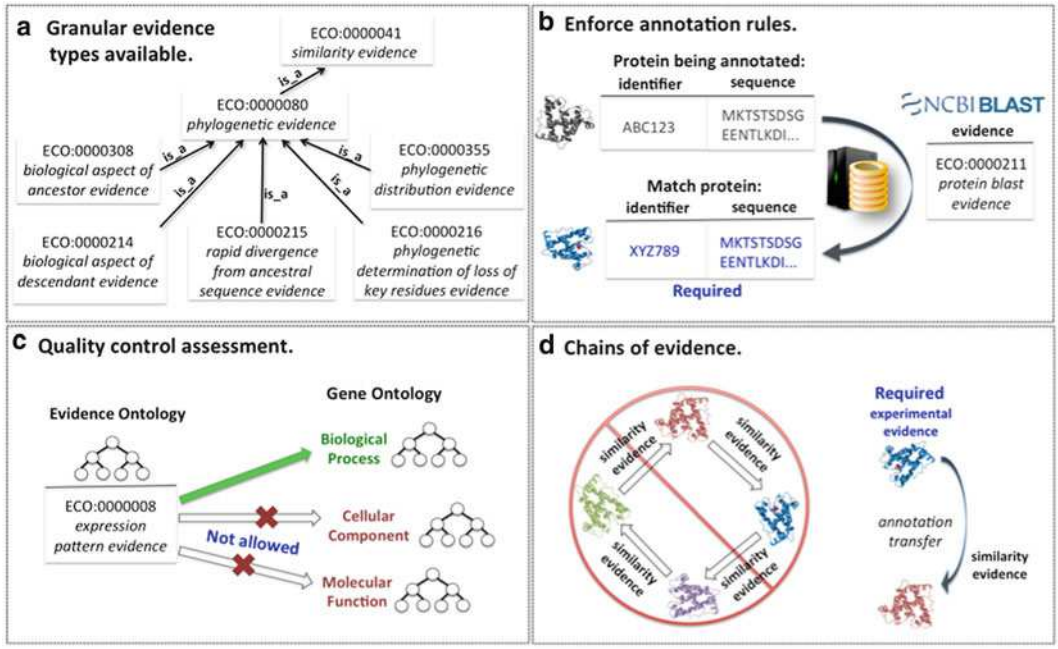


Fig. 4 Applications of ECO to GO. (**a**) ECO evidence classes are hierarchical such that broader classes parent more granular ones; depicted here are evidence types that support a phylogenetic tree-based approach for generating manually reviewed, homology-based annotations. (**b**) When a protein is 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. (**c**) Quality control assessment: Expression pattern evidence is only allowable for annotations to the GO Biological Process ontology. (**d**) 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

advantages of both systems and would still provide a mechanism for filtering evidence annotations by the previous codes if desired. If ECO terms were to be fully adopted by GO, the GAF format would change to require "ECO term" instead of "evidence code." Since most GO evidence codes have a one-to-one mapping to ECO terms (while the remainder, i.e., IEA, IGC, ISS, map, in conjunction with various GO standard references [http://purl.obolibrary.org/obo/eco/gaf-eco-mapping.txt], to specific ECO terms), GO data depositors could use a straightforward replacement based on the mappings. Other resources outside of GO have modeled their annotation capture systems on the GAF format. For example, the Ontology of Microbial Phenotypes [17] uses a modified version of the GO GAF, but employs ECO terms instead of GO evidence codes. The full use of ECO terms by the GO would enhance the integration of data derived from such diverse sources.

4 Benefits of ECO and Applications for the GO

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" (at present 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 (Fig. 4). For example:

1. 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 [18]. This phylogenetic annotation methodology necessitated a new set of evidence terms to capture the inference process (Fig. 4a). Currently over 150,000 annotations are associated with these new terms and the number continues to grow.
2. 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" (Fig. 4b). If such an evidence type were missing, this would flag the annotation for review.
3. 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 (Fig. 4c).
4. Evidence is used to prevent circular annotations based solely on computational predictions (Fig. 4d). 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 the matching gene

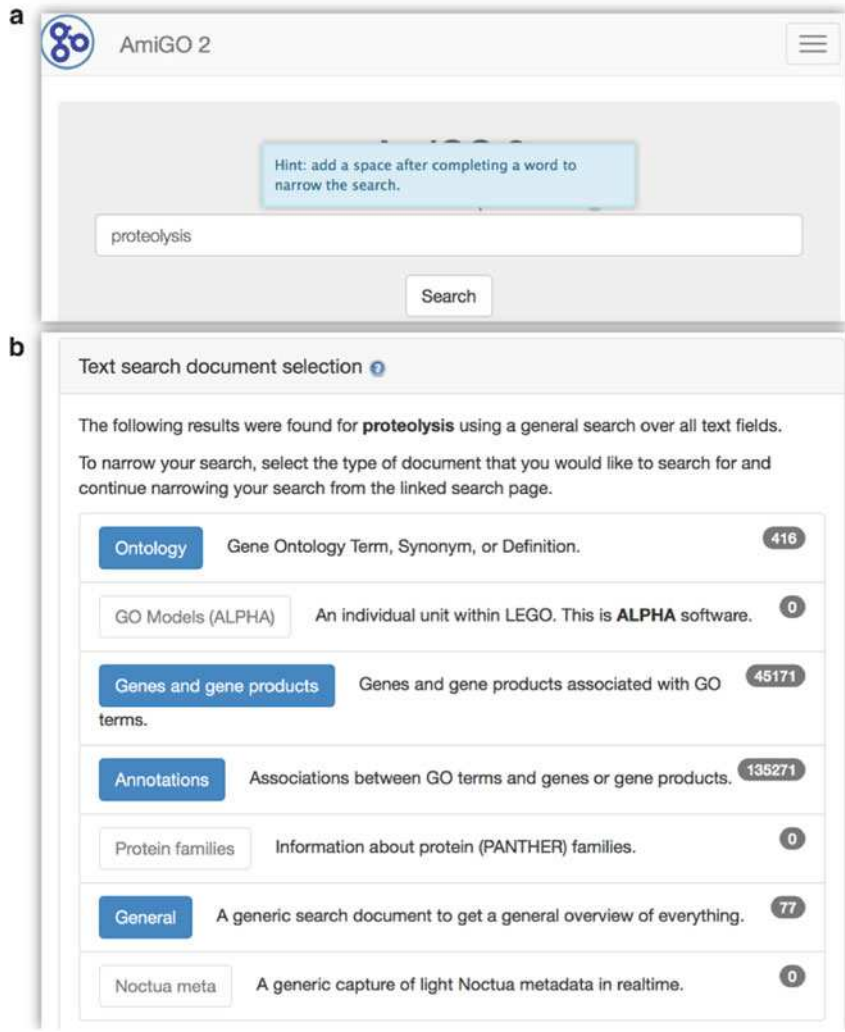


Fig. 5 AmiGO 2 query and results. (**a**) User has typed "proteolysis" into the search box. (**b**) Number of hits (*right gray box*) shown for each document category (*blue boxed text*). Clicking on "Annotations" will open a new page with more detailed results

product that is itself linked to an annotation supported by experimental evidence.

Yet another application of ECO for the GO has been realized in the UniProt-Gene Ontology Annotation (UniProt-GOA) project. Arguably, UniProt is the most comprehensive and best-curated protein database available to the research community. ECO terms have replaced the original UniProtKB [19] evidence types and are available in UniProtKB XML [11]. Novel ways of mapping and extending ontologies have been discussed with ECO and the GO Consortium to ensure appropriate development for UniProtKB annotation. The

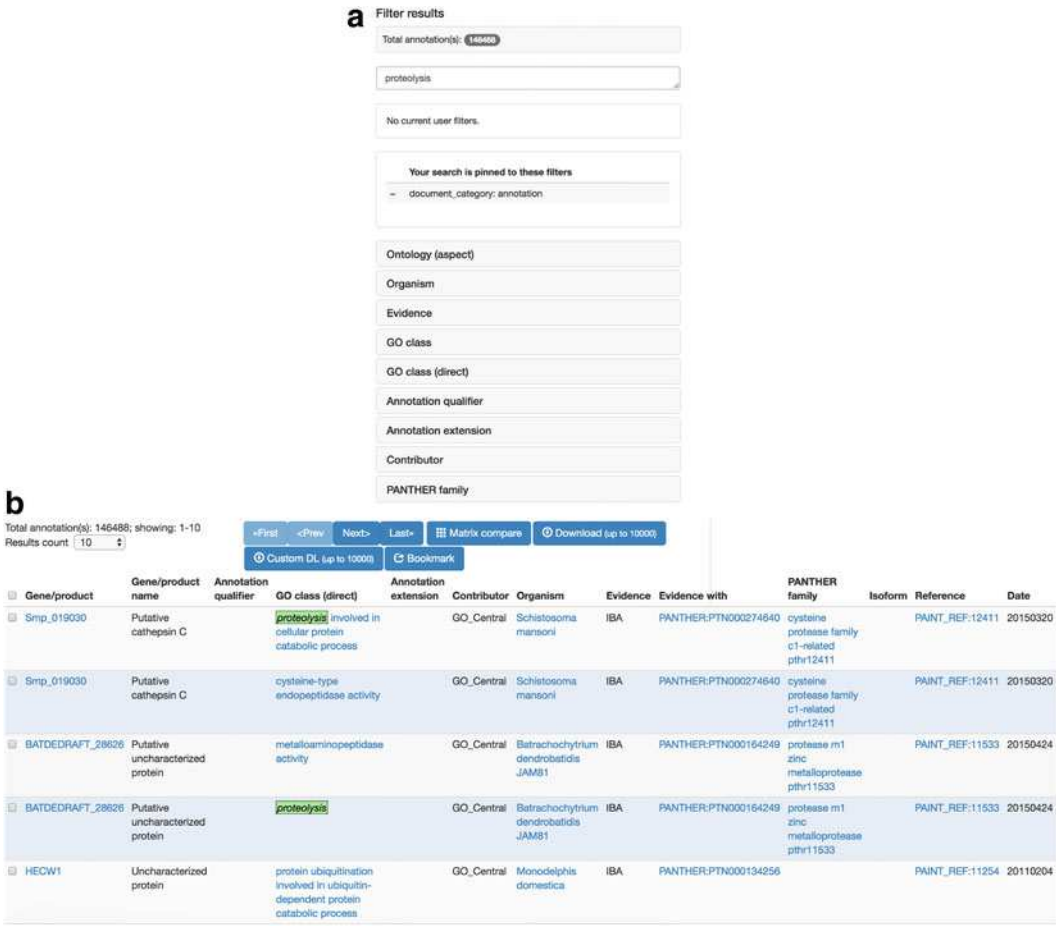


Fig. 6 Annotation hits to a query search. (**a**) To the left of the search results, the user has an opportunity to click on filters. (**b**) To the right, each annotation row is shown for a given protein

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 [20]. 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 ECO terms. Because ECO 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 ECO term. Thus, UniProt annotations can be grouped by leveraging the structure of ECO.

4.1 Exercise

Once the reader has gained a basic understanding of ECO and its connection to GO, we can perform the following simple exercise

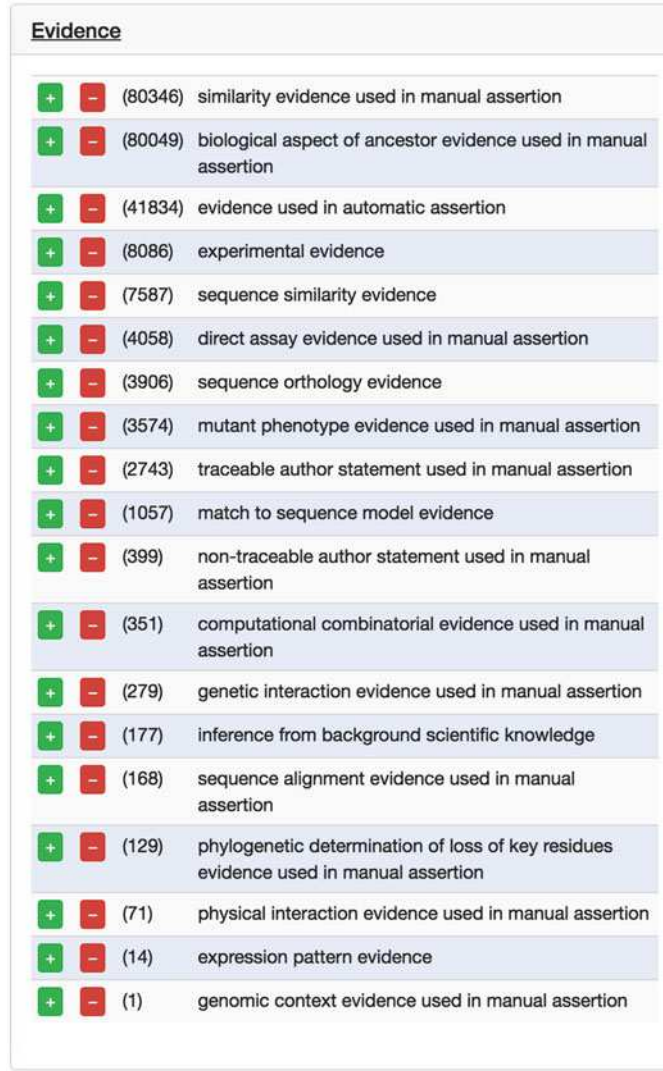
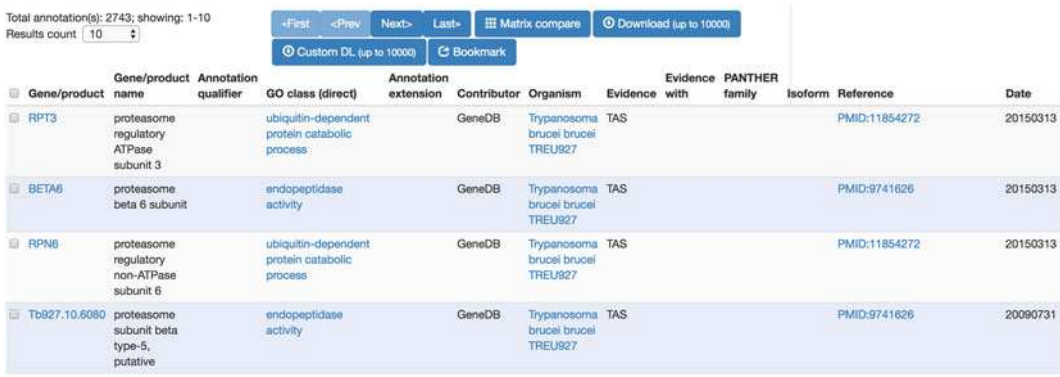


Fig. 7 Selected ECO terms in use by the GO Consortium that are related to the present query. The number of annotations supported by a given evidence type is shown in *parentheses*

that displays a faceted query using ECO in AmiGO 2 (http://amigo2.geneontology.org/amigo).

User types "proteolysis" into the query box (Fig. 5a) and sees a number of hits returned (Fig. 5b). Next, after clicking on "Annotations" in the blue rectangle, the user sees all the annotation-related terms that had hits to "proteolysis" (Fig. 6a, b), split into two parts here for easier viewing. Clicking on "Evidence" in the filter box (Fig. 6a) will expand it to display all constituent evidence types (Fig. 7). Clicking on


Total annotation(s): 2743; showing: 1-10
Results count 10
«First <Prev Next> Last» Matrix compare Download (up to 10000)
Custom DL (up to 10000) Bookmark

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Isoform	Reference	Date
RPT3	proteasome regulatory ATPase subunit 3		ubiquitin-dependent protein catabolic process		GeneDB	Trypanosoma brucei brucei TREU927	TAS				PMID:11854272	20150313
BETA6	proteasome beta 6 subunit		endopeptidase activity		GeneDB	Trypanosoma brucei brucei TREU927	TAS				PMID:9741626	20150313
RPN6	proteasome regulatory non-ATPase subunit 6		ubiquitin-dependent protein catabolic process		GeneDB	Trypanosoma brucei brucei TREU927	TAS				PMID:11854272	20150313
Tb927.10.6080	proteasome subunit beta type-5, putative		endopeptidase activity		GeneDB	Trypanosoma brucei brucei TREU927	TAS				PMID:9741626	20090731

Fig. 8 Filtering on evidence. After filtering on "traceable author statement used in manual assertion", only annotations supported by that evidence type are displayed, shown as "TAS" in the "Evidence" column. Number of annotations associated with that evidence type is shown at the *top left*

"traceable author statement used in manual assertion" will open a subset of the results that match that more restrictive filter (Fig. 8). The evidence filter box now says "Nothing to filter" (Fig. 9).

5 The Future of ECO

What else can an ontology of evidence do? One aspect of active exploration for ECO is the evaluation of confidence or quality of evidence. Work has begun [21] 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 use ECO to describe the *quality* of the evidence supporting an annotation in addition to the *type* of evidence that supports the annotation.

In summary, the Evidence and Conclusion 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, and the future promises both additional new applications of ECO as well as advancements to current ones.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Award Number 1458400 and National

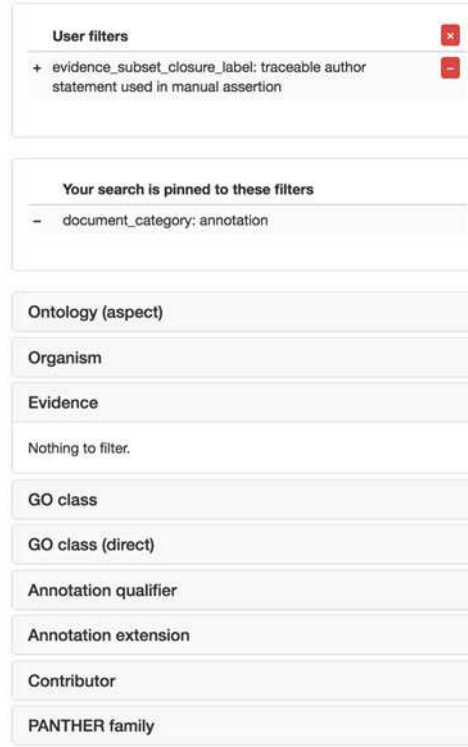


Fig. 9 Filter box fully selected. Increasingly granular evidence filters have been applied until there is nothing left to filter

Institutes of Health/National Institute of General Medical Sciences under Grant Number 2R01 GM089636. Open Access charges were funded by the University College London Library, the Swiss Institute of Bioinformatics, the Agassiz Foundation, and the Foundation for the University of Lausanne.

References

1. Gaudet P, Arighi C, Bastian F, Bateman A, Blake JA, Cherry MJ, D'Eustachio P, Finn R, Giglio M, Hirschman L, Kania R, Klimke W, Martin MJ, Karsch-Mizrachi I, Munoz-Torres M, Natale D, O'Donovan C, Ouellette F, Pruitt KD, Robinson-Rechavi M, Sansone SA, Schofield P, Sutton G, Van Auken K, Vasudevan S, Wu C, Young J, Mazumder R (2012) Recent advances in biocuration: meeting report from the Fifth International Biocuration Conference. Database:bas036. doi:10.1093/database/bas036
2. Burge S, Attwood TK, Bateman A, Berardini TZ, Cherry M, O'Donovan C, Xenarios L, Gaudet P (2012) Biocurators and biocuration: surveying the 21st century challenges. Database:bar059. doi:10.1093/database/bar059
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:bat054. doi:10.1093/database/bat054
4. Poux S, Gaudet P (2016) Best practices in manual annotation with the gene ontology. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 4
5. Arighi CN, Carterette B, Cohen KB, Krallinger M, Wilbur WJ, Fey P, Dodson R, Cooper L, Van Slyke CE, Dahdul W, Mabee P, Li D, Harris B, Gillespie M, Jimenez S, Roberts P, Matthews L, Becker K, Drabkin H, Bello S, Licata L, Chatr-aryamontri A, Schaeffer ML, Park J, Haendel M, Van Auken K, Li Y, Chan J, Muller HM, Cui H, Balhoff JP, Chi-Yang Wu J, Lu Z, Wei CH, Tudor CO, Raja K, Subramani S, Natarajan J, Cejuela JM, Dubey P, Wu C (2013) An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. Database:bas056. doi:10.1093/database/bas056
6. Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A, Gannon F, Grivell L, Hahn U, Hersh W, Hirschman L, Jensen LJ, Krallinger M, Mons B, O'Donoghue SI, Peitsch MC, Rebholz-Schuhmann D, Shatkay H, Valencia A (2008) Text mining for biology--the way forward: opinions from leading scientists. Genome Biol 9(Suppl 2):S7. doi:10.1186/gb-2008-9-s2-s7
7. Cozzetto D, Jones DT (2016) Computational methods for annotation transfers from sequence. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 5
8. Smith B (2003) Ontology. In: Floridi L (ed) Blackwell guide to the philosophy of computing and information. Blackwell, Oxford, pp 155–166
9. Smith B (2008) Ontology (Science). In: Eschenbach C, Grüninger M (eds) Formal ontology in information systems. Ios Press, Amsterdam, pp 21–35
10. Hastings J (2016) Primer on ontologies. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 1
11. Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M (2014) Standardized description of scientific evidence using the Evidence Ontology (ECO). Database:bau075. doi:10.1093/database/bau075
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29. doi:10.1038/75556
13. Gaudet P, Škunca N, Hu JC, Dessimoz C (2016) Primer on the gene ontology. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 3
14. The FlyBase Consortium (2002) The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic Acids Res 30(1):106–108
15. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res 29(1):102–105
16. 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(Database issue):D156–D160. doi:10.1093/nar/gkt1123
17. Chibucos MC, Zweifel AE, Herrera JC, Meza W, Eslamfam S, Uetz P, Siegele DA, Hu JC, Giglio MG (2014) An ontology for microbial

phenotypes. BMC Microbiol 14(1):294. doi:10.1186/s12866-014-0294-3

18. 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(7):e1000431. doi:10.1371/journal.pcbi.1000431
19. UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 42(Database issue):D191–D198. doi:10.1093/nar/gkt1140
20. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, Bely B, Browne P, Mun Chan W, Eberhardt R, Gardner M, Laiho K, Legge D, Magrane M, Pichler K, Poggioli D, Sehra H, Auchincloss A, Axelsen K, Blatter MC, Boutet E, Braconi-Quintaje S, Breuza L, Bridge A, Coudert E, Estreicher A, Famiglietti L, Ferro-Rojas S, Feuermann M, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, James J, Jimenez S, Jungo F, Keller G, Lemercier P, Lieberherr D, Masson P, Moinat M, Pedruzzi I, Poux S, Rivoire C, Roechert B, Schneider M, Stutz A, Sundaram S, Tognolli M, Bougueleret L, Argoud-Puy G, Cusin I, Duek-Roggli P, Xenarios I, Apweiler R (2012) The UniProt-GO Annotation database in 2011. Nucleic Acids Res 40(Database issue):D565–D570. doi:10.1093/nar/gkr1048
21. Bastian FB, Chibucos MC, Gaudet P, Giglio M, Holliday GL, Huang H, Lewis SE, Niknejad A, Orchard S, Poux S, Skunca N, Robinson-Rechavi M (2015) The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. Database:bav043. doi:10.1093/database/bav043