Guidelines for Selecting Gene Ontology (GO) Evidence Sentences

GO evidence sentences (GOES) are the basis for human curators to make associated GO annotations, which typically include three elements: 1) a curatable entity (i.e., gene or gene product), 2) a GO term (e.g., receptor-mediated endocytosis), and 3) a GO evidence code (e.g., Inferred from Mutant Phenotype (IMP)). See examples (ex1 and ex2) in page 2.

Importance of Capturing all GOES Thoroughly in a Paper

Since many learning-based text-mining algorithms rely on both positive and negative training instances, it is important to be as thorough as possible when manually annotating sentences. Evidence for GO annotations may be derived from a single sentence, or multiple continuous, or discontinuous, sentences. Additionally, evidence for a GO annotation can derive from multiple lines of experimentation, leading to multiple sentences in a paper supporting the same annotation. It is therefore important to try to capture all of these relevant sentences to ensure the positive & negative sets are as distinct as possible.

Types of Sentences to Capture

For the GO task, we'd like to capture all sentences positive for GO annotation, which include two different types of GOES:

- **GOES – Experiment**: These sentences describe experimental results and can be used to make a complete GO annotation (i.e., the entity being annotated, GO term, and GO evidence code). The annotation of such sentences is required throughout the paper, including the abstract, and any supporting summary paragraphs such as 'Author summary' or 'Conclusions'.

- **GOES – Summary**: Distinct from statements that describe the details of experimental findings, papers also include many statements that summarize these findings. These summary statements don't necessarily indicate exactly *how* the information was discovered, but often contain concise language about *what* was discovered. Such sentences are helpful to capture because they inform GO term selection in a concise manner despite its lack of information about evidence code selection.

It is not necessary to annotate true negative sentences that would not be used to make a GO annotation as long as all positive sentences (Experiment or Summary type) in the same paragraphs have been annotated exhaustively. That is, when you find and mark up a true positive sentence for GO annotation in a paragraph, please thoroughly read all other sentences in the same paragraph and make sure all true positive sentences are selected. By doing so, all the remaining sentences in the same paragraph that are not selected can be regarded as true negatives.

Examples of Sentences to Annotate

GOES (Experiment) are typically found throughout the Results section (if, nominally, there is one) of a paper. They may include sentences within the main body of the text or sentences that are part of Figure legends. Sometimes the evidence for a GO annotation is succinctly captured within one statement; in other cases, the information required to make a Experiment annotation is spread across multiple sentences. That is, evidence for GO annotations may be derived from a single sentence, or multiple continuous, or discontinuous, sentences.

Examples of **single evidence sentences** from PMC3469465:

Ex1: *On the other hand, the amount of UNC-60B-GFP was reduced and UNC-60A-type mRNAs, UNC60A-RFP and UNC-60A-Experiment, were detected in asd-2 and sup-12 mutants (Figure 2H, lanes 2 and 3), consistent with their colour phenotypes shown in Figure 2C and 2A, respectively.*

This sentence contains information about:

1) the entities to be annotated: *asd-2* and *sup-12*
2) the evidence code to be used, Inferred from Mutant Phenotype (IMP)
3) GO term, regulation of alternative mRNA splicing, via spliceosome (GO:0000381)

Ex2: *The unc-60 reporter worms exhibited weak Red phenotype in the asd-2 (yb1540) background (Figure 2C) and body wall muscle-specific expression of ASD-2b cDNA rescued the colour phenotype (Figure 2D), confirming that asd-2b is involved in the muscle-specific regulation of the unc-60 reporter.*

This sentence contains information about:

1) the entity to be annotated, *asd-2*
2) the evidence code to be used, Inferred from Mutant Phenotype (IMP)
3) GO term, regulation of alternative mRNA splicing, via spliceosome (GO:0000381)

Example of **multiple evidence sentences** from PMC3469465:

To investigate subcellular localization of ASD-2, we raised polyclonal antibodies against recombinant Experiment-length ASD-2b protein and stained wild-type and asd-2 (yb1540) worms with a purified immunoglobulin G (IgG) fraction (Figure 2E, 2F). Nuclei of body wall muscles, which are aligned along the dorsal and ventral periphery, are stained in the wild type (Figure 2E) and not in asd-2 mutant (Figure 2F).

In this case, curation of ASD-2 subcellular localization requires two sentences: the first sentence confirms the entity to be curated and the appropriate evidence code, while the second provides information on the cellular component.

1) entity to be annotated: ASD-2
2) evidence code: Inferred from Direct Assay (IDA)
3) GO term: nucleus (GO:0005634)

GOES (Summary) can commonly occur throughout the Abstract, at the end of paragraphs in the Results sections, as Figure legend titles, and in the last and first paragraphs of the Introduction and Discussion sections, respectively. Note that by summary statements, we mean those statements in a paper that directly refer to experiments performed in that paper, as opposed to experiments published in other papers or speculative statements that the authors make when discussing possible interpretations of their data.

While summary sentences may not allow a curator or text mining application to definitively select a GO evidence code, because of the concise nature of the language, their potential use in text mining is of interest as they may provide a viable alternative to replacing Experimenty manual annotation, perhaps by use of a text mining-specific evidence code.

When annotating summary sentences in the curation tool, please select a GO term and an entity to annotate, but leave the evidence code field blank. This will allow developers to distinguish summary sentences from experimental sentences.

Examples of **summary sentence** from PMC3469465:

From the Abstract:
Taken together, our results demonstrate that muscle-specific splicing factors ASD-2 and SUP-12 cooperatively promote muscle-specific processing of the unc-60 gene, and provide insight into the mechanisms of complex pre-mRNA processing; combinatorial regulation of a single splice site by two tissue-specific splicing regulators determines the binary fate of the entire transcript.

1) entities to annotate: ASD-2 and SUP-12
2) GO term: regulation of alternative mRNA splicing, via spliceosome (GO:0000381)

From the last paragraph of the Introduction:
We provide genetic and biochemical evidence that SUP-12 and another muscle-specific splicing regulator Alternative-Splicing-Defective-2 (ASD-2), a member of the signal transduction and activation of RNA (STAR) family of RNA-binding proteins [34], cooperatively repress excision of the first intron through specific binding to the intron.

1) entities to annotate: ASD-2 and SUP-12
2) GO term: regulation of alternative mRNA splicing, via spliceosome (GO:0000381), pre-mRNA intronic binding (GO:0097157), ribonucleoprotein complex (GO:0030529)

From the end of a paragraph in the Results section:
We therefore concluded that ASD-2b and SUP-12 can weakly interact with each other and that unc-60 intron 1A RNA promotes the formation of the stable ASD-2b/SUP-12/RNA ternary complex by providing juxtaposed CUAAC repeats and UGUGUG stretch that are specifically recognized by ASD-2b and SUP-12, respectively.

1) entities to annotate: ASD-2b and SUP-12 (include unc-60 for GO:0030529 only)
2) GO terms: protein binding (GO:0005515), ribonucleoprotein complex (GO:0030529), pre-mRNA intronic binding (GO:0097157)

From Figure legend titles:
ASD-2 and SUP-12 regulate muscle-specific processing of the unc-60 reporter in body wall muscles.

1) entities to annotate: ASD-2 and SUP-12
2) GO term: regulation of alternative mRNA splicing, via spliceosome (GO:0000381)

From the first paragraph of the Discussion:
In muscles (Figure 7B), ASD-2b and SUP-12 cooperatively bind to CUAAC repeats and UGUGUG stretch, respectively, in intron 1A to repress excision of intron 1A and weakly of intron 2A during transcription of the UNC-60A region.

1) entities to annotate: ASD-2b and SUP-12 (include unc-60 for GO:0030529 only)
2) GO terms: ribonucleoprotein complex (GO:0030529), pre-mRNA intronic binding (GO:0097157), regulation of alternative mRNA splicing, via spliceosome (GO:0000381)

Finally, please note that despite all our best efforts, there will always be variation in the depth of annotation between curators and organisms. For instance, there may be gray areas where some curators will select a sentence relating to a phenotype as a GO sentence, while others not.

FAQs

Q1) If a gene product is referred to in a sentence using a general term or phrase, should we still annotate it?

A1) Yes, if the general term or phrase can be linked to specific gene products studied in the paper. In these cases, when you make your annotation, please include, if you can, one or more sentences from the paper that identify the specific gene products.

Example (PMC3192830):
From Author Summary: Loss of BLM and any one of the nucleases results in severe genome instability, reduced cell proliferation, and, ultimately, death of the animal.

In this case, curators should include a sentence to indicate what 'nucleases' refers to. For example, from the abstract: “We present a comparison of phenotypes occurring in double mutants that lack DmBLM and either MUS81, GEN, or MUS312, including chromosome instability and deficiencies in cell proliferation.”

Q2) If a sentence refers to a complex should we annotate to the complex?

A2) No, since the GO is not currently annotating to protein complexes, we will not include them in this exercise.

Example: PP2A (Protein Phosphatase 2A) and its regulatory subunit, Twins (Tws), are required for centriole duplication

This sentence can be annotated for the Twins gene product, but not for the PP2A complex. If, in other sentences, an annotation can be made to a specific subunit of a complex, those should be included.

Q3) If a sentence refers to a gene product using an anaphora such as 'it', do we need to include a sentence that identifies 'it'?

A3) Yes, if there is an additional sentence that can be used to identify the gene product referred to as 'it', then please include that sentence.

Example: Its localization at the meiotic spindle, mitotic centrosomes, and metaphase microtubules are all dependent on the presence of microtubules.

In this case, curators need to supply another sentence to indicate what 'Its' refers to. The sentence in the paper preceding the sentence above could be used: “In conclusion, consistent with its role in chromosome segregation, LIN-5 is localized at the spindle apparatus in meiosis and mitosis.”

Q4) Some sentences in the Abstract and Discussion section contain all information necessary to make a GO annotation, i.e., the gene product, the GO term, and the GO evidence code. Should these sentences still be annotated as Experimental sentences?

A4) Yes, the content, or language, of the sentence is more important than its location in the paper. If you can confidently assign a complete GO annotation from a sentence present somewhere other than the Results section, then please do so.

Q5) If a paper contains sentences that support GO annotations for a species other than what our group typically annotates, should we include them?

A5) Curators can use their discretion here. If you feel confident that you can make the correct annotation for another species, then curators should feel free to do so. If, however, the sentences refer to biology outside your normal domain of expertise, you may skip these annotations. This will mean that there will be a number of false-negative sentences in the dataset due to curators skipping annotations for species that aren't necessarily the main focus of the paper (for example a worm gene turning up in a fly paper).

Q6) Should we annotate sentences that refer to Supplemental Information?

A6) If the sentence that refers to Supplemental Information has the complete information for a GO annotation but simply refers to a Figure or Table in Supplemental Information, you should annotate it.

If some of the information for the annotation is actually contained in the Supplemental documentation however, you shouldn't go to those documents to get the additional sentences.

A general rule of thumb in these cases: if the text you need to annotate is on the screen for that paper, then you can make the annotation. If you need to go to a different document to get all the information then, for the purposes of this task, don't make the annotation.