

Annotation Comparison Examples

This page gives examples of sentences from papers along with modified versions that show when a sentence can and cannot be annotated. This comparison helps to show the differences in wording that signal an annotation or the lack of one.

Example 1

PMID:16840531: “*The lack of induction of the *Synechocystis recA*, *lexA* and *crhR* genes following DNA damage suggests these gene products are not required during the cellular response to DNA damage.*”

- ECO:0000008
- **Name:** Expression Pattern Evidence
- **Term Confidence:** Medium
From "induction".
- **Assertion Strength:** Medium
From "suggests".
- **Category:** Biological Process
The assertion is about the cellular response to DNA damage.
- **Sentence Pair:** No
- **Negative Assertion:** Yes
The assertion states that the gene products are NOT required..

Compare the above sentence with this one below and note the different wording makes the following not annotatable:

“*The *Synechocystis recA*, *lexA* and *crhR* genes were not induced following DNA damage*”

→ Biological Process (weak) from: “following DNA damage”

→ ECO:0000008 from: “induced”

→ No assertion, just the ECO:0000008 readout. The assertion needs to be inferred from context. Note because genes are in the sentence (*recA*, *lexA*, and *chrR*), the category cannot be Phenotype.

Example 2

PMID:17129387: “*The presence of differentially expressed proteins in the two strains, together with results of previous studies [28,29] demonstrated the role of CcpA as global regulator in *L. plantarum*.*”

- ECO:0000010
- **Name:** Protein Expression Evidence
- **Term Confidence:** Medium
From "differentially expressed proteins".
- **Assertion Strength:** High
From "demonstrated".
- **Category:** Biological Process
The assertion is about the CcpA as a global regulator, which means it is regulating gene expression, a biological process.
- **Sentence Pair:** No
- **Negative Assertion:** No

Compare the above sentence with this one below and note the different wording makes the following not annotatable:

"The presence of differentially expressed proteins in the two strains, together with results of previous studies [28,29] demonstrated increased activity of CcpA."

→ Category -- unclear what "activity" is referring to. Maybe Molecular Function

→ ECO:0000010 from: "differentially expressed proteins"

→ No assertion, just the ECO:0000010 readout.

Example 3

PMID:17129387: "Indeed, the GroEL protein appears to be more abundant in the LM3 wild type strain compared to the LM3-2 mutant strain, suggesting the involvement of the CcpA protein in the positive regulation of its expression."

- ECO:0000046 and ECO:0000015 (two evidence types for one assertion = 2 annotations)
- **Name:** Protein Expression Level Evidence (ECO:0000046)
- **Term Confidence:** Medium
From "protein appears to be more abundant" -- the sentence is talking about the quantity of a protein.
- **Name:** Mutant Phenotype Evidence (ECO:0000015)
- **Term Confidence:** High
From "in the LM3 wild type strain compared to the LM3-2 mutant strain" -- both the wild type and the mutant are mentioned and compared, hence "High" confidence.
- **Assertion Strength:** Medium
From "suggesting".
- **Category:** Biological Process
From "positive regulation of its expression" -- regulation of gene expression, a biological process.
- **Sentence Pair:** No
- **Negative Assertion:** No

Compare the above sentence with this one below and note the change in the evidence confidence for ECO:0000015 (only):

“Indeed, the GroEL protein appears to be more abundant in the LM3 mutant strain, suggesting the involvement of the CcpA protein in the positive regulation of its expression.”

- ECO:0000046 and ECO:0000015 (two evidence types for one assertion = 2 annotations)
- **Name:** Protein Expression Level Evidence (ECO:0000046)
- **Term Confidence:** Medium
From "protein appears to be more abundant" -- the sentence is talking about the quantity of a protein.
- **Name:** Mutant Phenotype Evidence (ECO:0000015)
- **Term Confidence:** Low
From "in the LM3 mutant strain" -- there is no mention of the wild type or comparison with it, hence "Low" confidence.
- **Assertion Strength:** Medium
From "suggesting".
- **Category:** Biological Process
From "positive regulation of its expression" -- regulation of gene expression, a biological process.
- **Sentence Pair:** No
- **Negative Assertion:** No

Now compare the first sentence of this example with the next sentence below and note the change to make the sentence not annotatable:

“Indeed, the GroEL protein appears to be more abundant in the LM3 wild type strain compared to the LM3-2 mutant strain”

→ no assertion; self-regulation is implicit here, but must be inferred from the fact that the mutant strain has less GroEL

Example 4

PMID:17129387: *“The sequence analysis of the groESL promoter region revealed the presence of the CIRCE element of 9 bp separated by 9 bp spacer, with the DNA sequence TTAGCACTT-N9-GAGTGCTAA, starting four nucleotide residues downstream of the start site of transcription.”*

- ECO:0000044
- **Name:** Sequence Similarity Evidence
- **Term Confidence:** Low
From "sequence analysis of the groESL promoter region".
- **Assertion Strength:** High
From "revealed the presence of".
- **Category:** Sequence Feature
From "CIRCE element".
- **Sentence Pair:** No
- **Negative Assertion:** No

Compare the above sentence with this one below and note the different wording makes the following not annotatable:

“The sequence analysis of the groESL promoter region revealed the sequence TTAGCACTT-N9-GAGTGCTAA”

→ not clear that there is a particular sequence feature here, just a bunch of nucleotides; so a readout, not an assertion

→ one could try to infer that the N9 is indicative of a model (bases not conserved, vs the consensus), but this needs to be inferred, it's not explicit, so no annotation.

Example 5

PMID:1712938: “On the contrary, transcription of the *dnaK* operon was induced 16-fold in the LM3 wild type strain (Fig. 3B, lanes 1 and 2) and only 8-fold in the LM3-2 mutant strain upon exposure to a heat shock (Fig. 3B, lanes 3 and 4), as measured by a quantitative analysis by PhosphorImager.”

→ ECO:0000008 (expression pattern evidence) from: “ was induced 16-fold...upon exposure to a heat shock...quantitative”. Note the phosphorimager is not a separate evidence type but the method for reading the fold changes for the expression pattern evidence.

→ Biological Process (response to heat) from: “upon exposure to a heat shock”

→ ECO:0000015 mutant phenotype evidence from “LM3 wild type strain...the LM3-2 mutant strain”

→ **no assertion; it's a readout**

→ You need to infer that the mutant corresponds to a gene here. No explicit actor is mentioned. So again, no annotation here.

Compare the above with the following sentence that **could be annotated**:

“On the contrary, involvement of gene ABC in heat-shock response was shown by 16-fold induction of the *dnaK* operon in the LM3 wild type strain (Fig. 3B, lanes 1 and 2) and only 8-fold in the LM3-2 mutant strain upon exposure to a heat shock (Fig. 3B, lanes 3 and 4), as measured by a quantitative analysis by PhosphorImager.”

An assertion now is evident -- “involvement of gene ABC in heat-shock response” -- there is an explicit actor, namely “gene ABC”.

- ECO:0000008 and ECO:0000015 (two evidence types for one assertion = 2 annotations)
- **Name:** Expression Pattern Evidence (ECO:0000008)
- **Term Confidence:** High
From the various fold inductions.
- **Name:** Mutant Phenotype Evidence (ECO:0000015)
- **Term Confidence:** High
From “in the LM3 wild type strain... and in the LM3-2 mutant strain” -- both the wild type and the mutant are mentioned and the fold changes compared, hence “High” confidence.

- **Assertion Strength:** High
From "involvement...was shown".
- **Category:** Biological Process
Response to heat from "exposure to a heat shock" -- this is a biological process.
- **Sentence Pair:** No
- **Negative Assertion:** No

Example 6

PMID 23861975: "On the basis of primary sequence and structural features, PA3699 and PA4135 can be assigned to the TetR-like and MarR-like family of transcriptional regulators, respectively [22]."

We make the following annotation:

- ECO:0000044
- **Name:** Sequence Similarity Evidence
- **Term Confidence:** High
From "On the basis of primary sequence and structural features".
- **Assertion Strength:** Medium
From " can be assigned".
- **Category:** Taxonomy/Phylogeny
"the TetR-like and MarR-like family". Assignment to gene families is giving an evolutionary relationship.
- **Sentence Pair:** No
- **Negative Assertion:** No

Note that a second annotation cannot be made because it must be inferred. Saying that the families are "transcriptional regulators" implies that PA3699 and PA4135 are also involved in regulation of transcription, but the sentence doesn't clearly state this.

Compare the above with the following sentence that **could be annotated for Biological Process too:**

"On the basis of primary sequence and structural features, PA3699 and PA4135 can be assigned to the TetR-like and MarR-like family of transcriptional regulators, respectively, thus demonstrating the role of PA3699 and PA4135 in regulation of transcription."

The second annotation would be:

- ECO:0000044
- **Name:** Sequence Similarity Evidence
- **Term Confidence:** High
From "On the basis of primary sequence and structural features".
- **Assertion Strength:** High
From "thus demonstrating the role".
- **Category:** Biological Process
From "in regulation of transcription".
- **Sentence Pair:** No
- **Negative Assertion:** No

Example 7

PMID:22984476: “As further determined by DNase I footprinting (Fig. 7d), His-AphA protected a single region from 110 to 76 bp upstream of opaR against DNase I digestion in a dose-dependent manner .”

→ ECO:0001810 (DNase footprinting evidence) from: “DNase I footprinting”.

→ **no assertion; it’s a readout**

→ You need to infer that there is binding from the word "protected". So again, no annotation here.

Compare the above with the following sentence that **could be annotated**:

“As further determined by DNase I footprinting (Fig. 7d), His-AphA bound to a single region from 110 to 76 bp upstream of opaR against DNase I digestion in a dose-dependent manner.”

- ECO:0001810
- **Name:** DNase Footprinting evidence (ECO:0001810)
- **Term Confidence:** High
From "DNase I footprinting".
- **Assertion Strength:** High
From "bound" -- for binding assertions, the word "bound" is a strong statement.
- **Category:** Molecular Function
"bound to" indicates binding, a molecular function.
- **Sentence Pair:** No
- **Negative Assertion:** No

Also compare the format of the above sentence with "protected" to this different sentence below that has the phrase "was able to bind". This next sentence can be annotated because it explicitly says there is binding.

“A 334 bp promoter-proximal DNA region of opaR was subjected to EMSA with purified His-AphA protein (Fig. 7c). The results showed that His-AphA was able to bind to the DNA fragment in a dose-dependent manner in vitro .”

- ECO:0000096
- **Name:** Electrophoretic mobility shift evidence (ECO:0000096)
- **Term Confidence:** High
From "EMSA".
- **Assertion Strength:** High
From "was able to bind" -- for binding assertions, the words "was able to bind" is a strong statement.
- **Category:** Molecular Function
"was able to bind" indicates binding, a molecular function.
- **Sentence Pair:** Yes
- **Negative Assertion:** No