Measuring inter-annotator agreement in GO annotations

Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, Maslen J, Binns D, Apweiler R. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 2005; 6 Suppl 1:S17. PMID: 15960829.

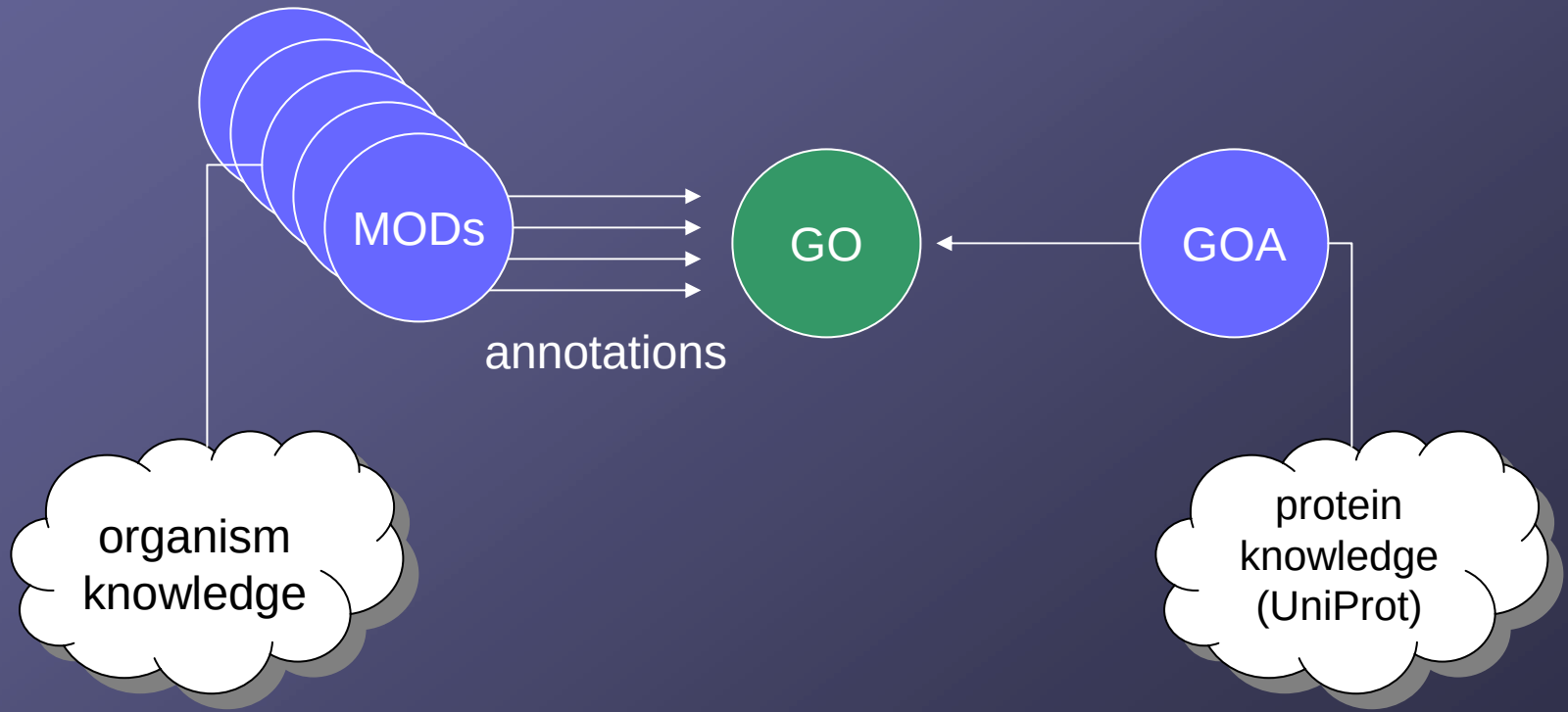
SILS Biomedical Informatics Journal Club

http://ils.unc.edu/bioinfo/

2005-10-18

Gene Ontology Annotation (GOA) project

- Goal: Annotate proteins in UniProt with GO terms



Problems / Questions

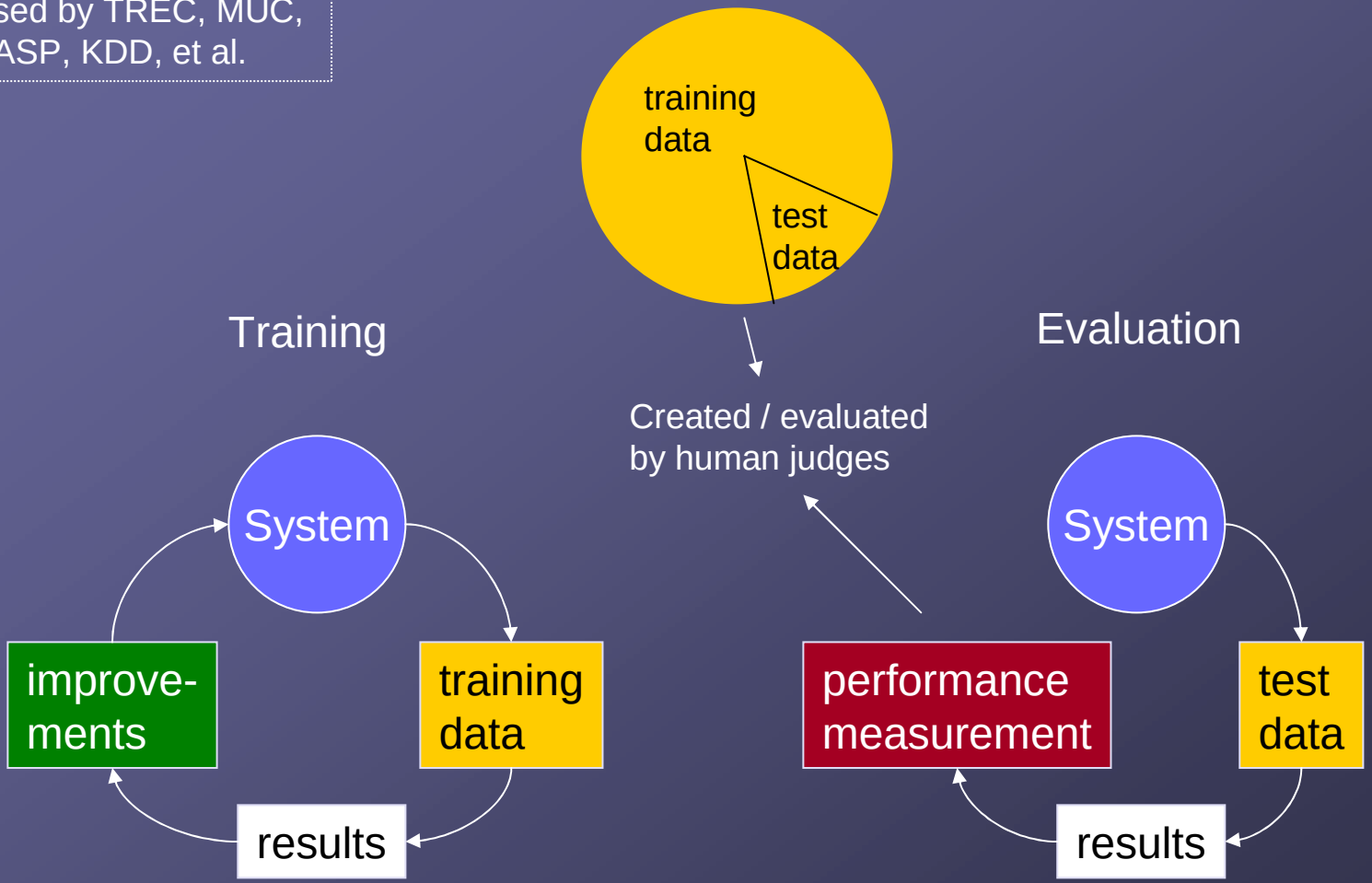
- Protein information continues to grow faster than curators can manually annotate it with knowledge extracted from the literature.
- Automated annotation methods still don't understand natural language well.
- "what do GO curators really need?" [2]
  - A system to find 'relevant' papers and extract "the distinct features of a given protein and species", and then "to locate within the text the experimental evidence to support a GO term assignment." [2]
- RQ: Does "automatically derived classification using information retrieval and extraction" "assist biologists in the annotation of the GO terminology to proteins in UniProt?" [2]

BioCreAtIvE

- Critical Assessment of Information Extraction systems in Biology
- Addresses the problems of comparability and evaluation (multiple text-mining systems using different data and tasks)
- Defines a common task, common data sets and a clearly defined evaluation
- “BioCreAtIvE task 2 was an experiment to test if automatically derived classification using information retrieval and extraction could assist expert biologists in the annotation of the GO vocabulary to the proteins in the UniProt Knowledgebase.” [1]

Standard IR evaluation process

Used by TREC, MUC, CASP, KDD, et al.



Manual annotation process

Protein prioritization

1. Un-annotated
2. Disease relevance
3. Microarray importance

Find relevant papers

- Do existing papers in UniProt entry have GO relevance?
- Supplementary PubMed searches using gene & protein names
- Underlying species is important

Term extraction

- Paper is preferred
- Scan specific sections [Table 1]

Term assignment

- Browse GO for appropriate terms

Automated annotation process

- Identify proteins in narrative text of papers
- Check for presence of functional annotation
- Select GO term and text that provided the evidence [4]

Data

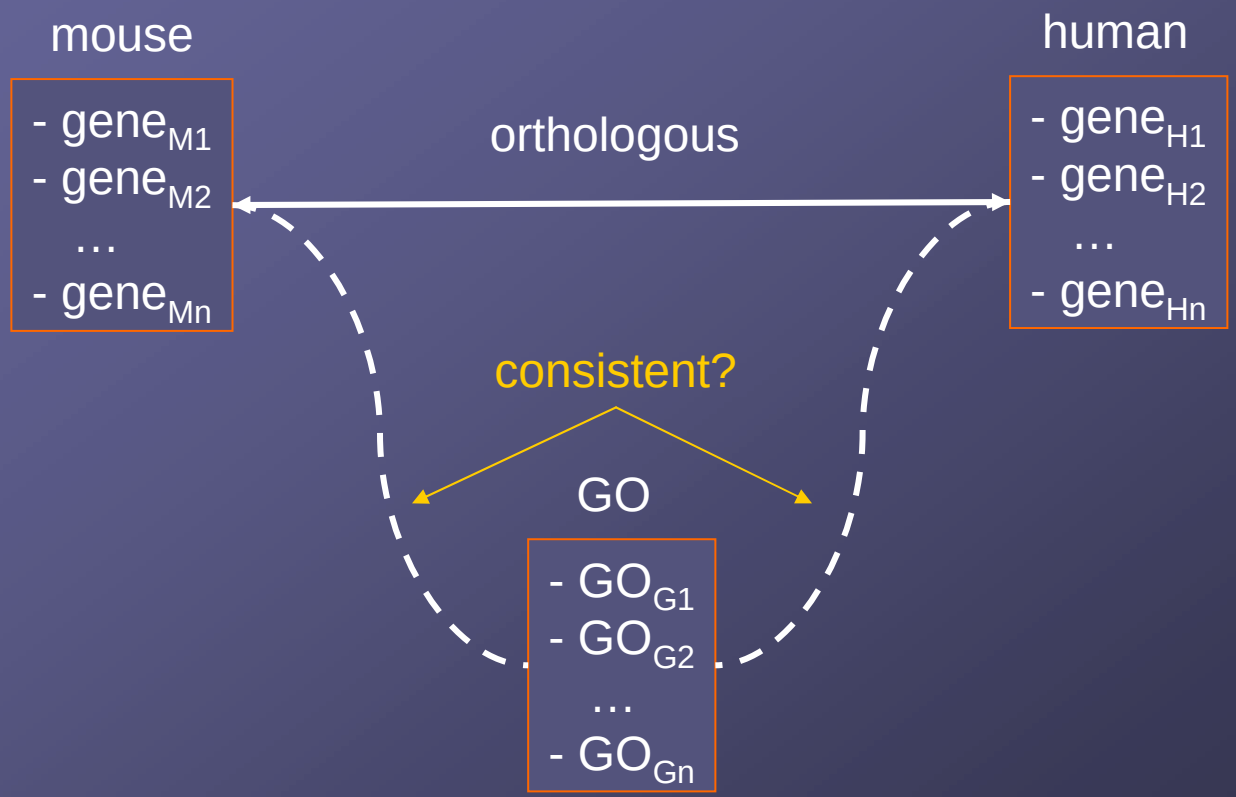
Training set

- ~9,000 existing manually-curated GO annotations in UniProt with PubMed IDs & GO evidence codes
- GO evidence codes ISS, IC, ND ignored
- Some coding problems on older annotations limit the number of usable records [5]

Test set

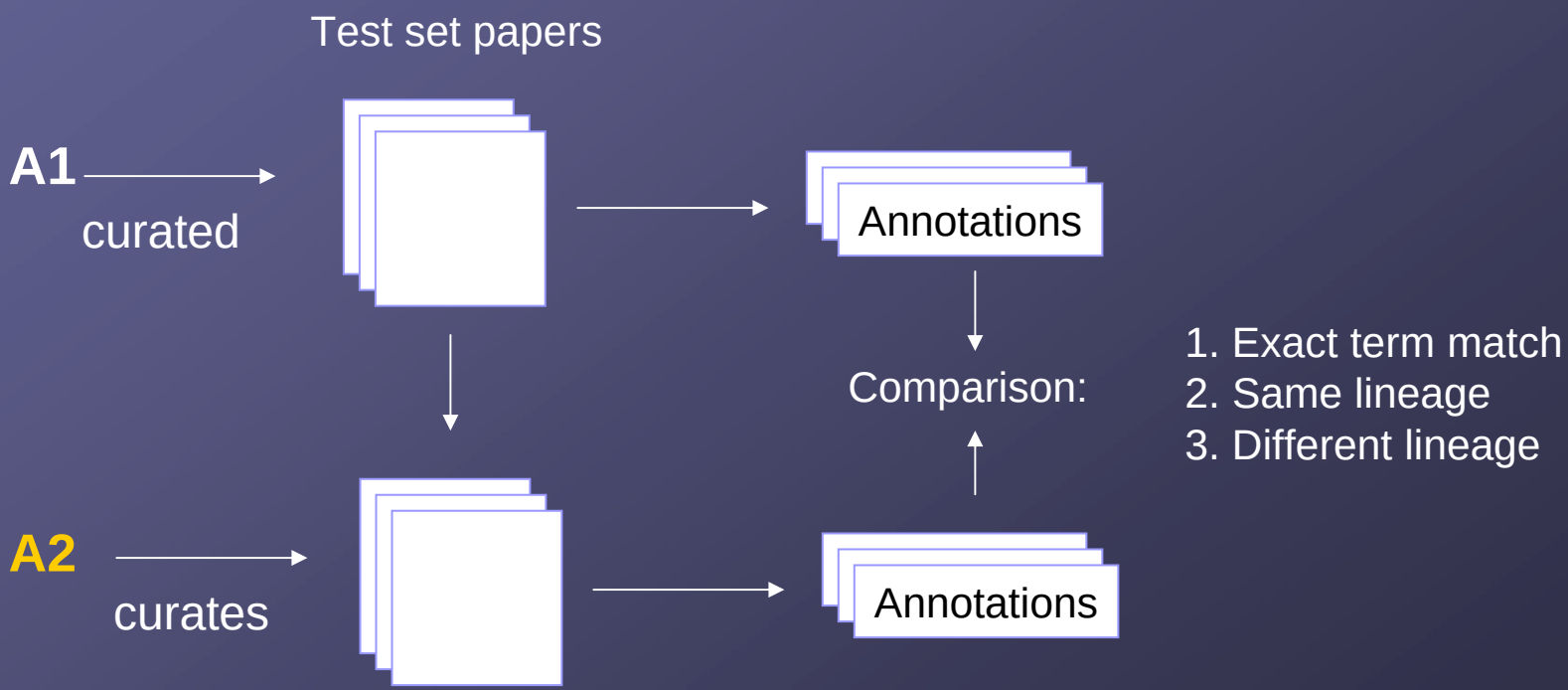
- 200 papers from *JBC* 1998-2002
- Already associated with 286 UniProt entries, but lacking manual GO annotation
- 923 GO terms were manually extracted; avg of 9 terms/protein

Mouse/Human annotation consistency



Inter-annotator agreement

- Sources of variation:
  - Curator's biological knowledge / experience
  - Curator's standard work practices [should be normalized for the study]
  - Manually-curated annotations could be wrong
  - Curators acting as relevance judges creates bias



Evaluation criteria

Table 3: Evaluation criteria for GO and protein predictions.

Evaluation	Criteria for GO term assignment	Criteria for protein association
High	The GO term assignment was correct or close to what a curator would choose, given the evidence text.	The protein mentioned in the evidence text correctly represented the associated UniProt accession (correct species).
General	The GO term assignment was in the correct lineage, given the evidence text, but was too high level (parent of the correct GO term) e.g. *biological_process* or too specific.	The evidence text did not support annotation to the associated UniProt accession but was generally correct for the protein family or orthologs (non-human species).
Low	The evidence text did not support the GO term assignment. **Note:** The GO term may have been correct for the protein but the evidence text did not support it.	The evidence text did not mention the correct protein (e.g. for Rev7 protein (ligand) incorrect evidence text referred to 'Rev7 receptor') or protein family.

Evaluation criteria

Table 4: Summary of mistakes and curator comments following the task 2 evaluation.

Mistakes	Suggestion/Comment
Predicting obsolete GO terms	Strip obsolete GO terms, i.e. children of *obsolete molecular function* (GO:0008369), *obsolete cellular component* (GO:0008370), *obsolete biological process* (GO:0008370) [25]
Predicting GO terms from Materials and Methods e.g. 'pH' value yielded 'pH domain binding' (GO:0042731), 'CHO cell line' yielded numerous GO terms containing 'acetyl**cho**line'.	Only look in certain sections of the paper for features. See Table 1 for GOA.
Predicting plant GO terms to human proteins e.g. *germination* (GO:0009844)	Look at GO Documentation on *sensu* [24] and strip out unnecessary GO terms.
Highlighting too much text	Set limit on evidence text highlight to be useful for curators. Limit to <5 lines.
Over-predicting GO terms from one line of text	More important to curator to choose a higher level term that is correct than to be too specific and incorrect.
Common GO terms predicted out of context e.g. text 'mapped to chromosome 3q26' yielded GO component term 'chromosome' GO:0005694. Text indicates chromosome number, not where the protein functions. e.g. text '249 amino acid' yielded multiple GO terms i.e. 'amino acid activation' GO:0043038.	Most papers will mention chromosome location and the amino acid length of a sequence. Do not predict GO terms from text if words 'chromosome' or 'amino acid' in evidence text is accompanied by a number.
Choosing first paragraph of paper as supporting text	Although a lot of information can be found in introduction of paper, the task was to choose the highlight which supported the GO term. Whole paragraph highlights do not speed up the curation process. Limit to <5 lines.
Difficulty in interpreting word order e.g. 'RNA binding protein' yielded the incorrect GO prediction 'protein binding'	
Difficulty in predicting correct taxonomic origin of protein.	This can also be difficult for a curator, given lack of evidence in text.
Too many low confidence runs	Only submit data with high confidence level for evaluation. Limit participants to their best run/technique. (little difference between runs, repeat evaluations)

Inter-annotator agreement

Table 5: Inter-annotator agreement.

GO terms	Curator 1 +2	Curator 1+3	Curator 2+ 3	Average
Exact	47	35	35	39
Same Lineage	15	20	19	18
New Lineage	56	39	35	43
Correct	107	91	85	94
Incorrect	11	3	4	18
TOTAL	118	94	89	100
Precision	0.91	0.96	0.96	0.94
Recall	0.70	0.72	0.73	0.72
F-measure	0.79	0.82	0.83	0.82

Where precision is the fraction of manual GO term annotations that are correct (number of correct annotations / (number of correct annotations + number of incorrect annotations). Recall is defined as the fraction of correct GO term annotations that were successfully retrieved during manual annotation (number of correct annotations / number of correct annotations + (number of annotations from new lineage - number of incorrect annotations). New lineage annotations minus incorrect annotations represent total number of the GO terms that the curators should have correctly retrieved from the paper. F-measure = (balanced precision and recall) = 2 × P × R/(P+R).

Table 6: Comparison of BioCreAtIvE test set manual annotations with electronic GO annotation predictions.

	InterPro2GO	SPKW2GO	EC2GO	
Total IEA annotations	635	385	27	
Exact term	151 (0.24)	62 (0.16)	18(0.67)	Correct
Same lineage > granularity	24 (0.04)	10 (0.03)	3 (0.11)	Potentially Incorrect/Correct
Same lineage < granularity	273 (0.43)	170 (0.44)	1 (0.04)	Correct
Total same lineage	297 (0.47)	180 (0.47)	4 (0.15)	Potentially Incorrect/Correct
New lineage	187 (0.29)	143 (0.37)	5 (0.19)	Potentially Incorrect/Correct
Total potential incorrect	211 (0.33)	153 (0.40)	8 (0.30)	
Total minimal correct	424 (0.67)	232 (0.60)	19 (0.70)	
Precision	0.67–1.00	0.60–1.00	0.70–1.00	

Where the GO evidence code IEA is 'Inferred from Electronic Annotation' [27]. 'Same lineage > granularity' means where the electronic mapping (InterPro2GO, EC2GO or SPKW2GO) predicted a GO term that was in the same lineage/branch as the manually curated GO term but represented a more granular/parent term. 'Total potential incorrect' annotations = 'Same lineage >granularity' + 'New lineage'. 'Total minimal correct' annotations = 'Exact term' + 'Same lineage < granularity'. Percentage calculations are represented in parentheses.

Inter-annotator agreement

- Camon's 3 measures of agreement don't allow for:
  - measurement of magnitude of difference, apart from 1 node up or down (parent_of, child_of);
  - cases where similar terms appear in different parts of the tree (polyhierarchy);
  - when new terms must be created for concepts that don't currently exist in GO;
  - measures of annotation quality other than inter-annotator consistency
  - don't adjust for chance or >2 annotators as do statistics such as Cohen's kappa

Table 4. Annotation quality facets, questions, and evaluation methods.

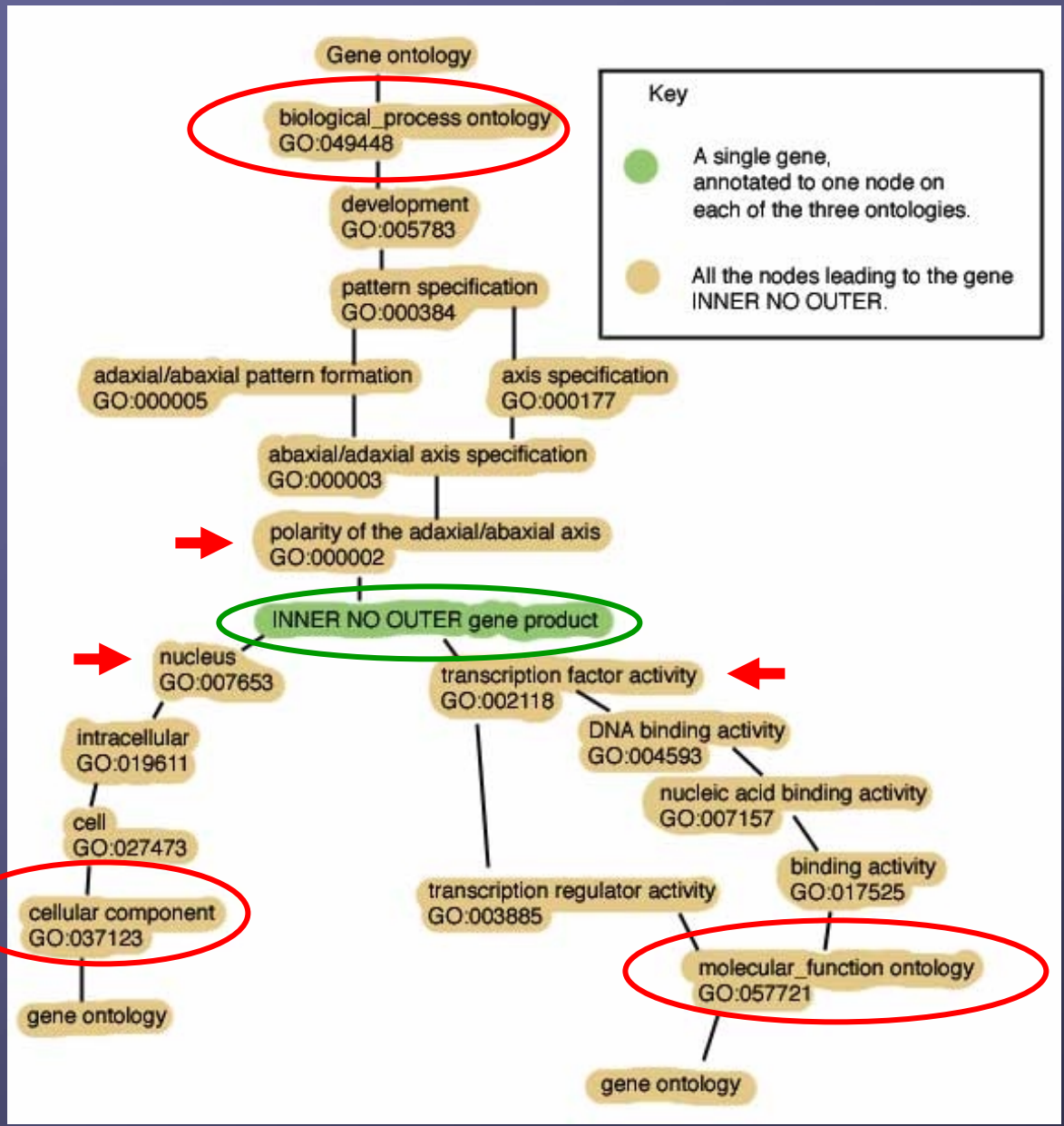
Facet	Research questions	Evaluation methods
Consistency	What is the nature and degree of variance in annotations made by different curators for the same unit of evidence?	Compare variation in similar annotations made by different annotators (inter-annotator consistency).
Specificity	Do the annotations of the same unit of evidence made by different annotators vary in terms of breadth, depth, specificity, etc.?	Compare quantitative and qualitative variation in annotations apart from consistency facets.
Reliability	Does the same curator make the same annotations for the same article at different time points? What factors might contribute to differences in annotation over time?	Compare variation in individual curators' annotations over time (intra-annotator consistency).
Accuracy	How is the accuracy of an annotation evaluated? What are the decision points in the annotation process that influence accuracy?	Define annotation accuracy and how to measure variance. Evaluate the accuracy of selected extant annotations.

from MacMullen, W.J., Identification of strategies for information integration using annotation evidence. NLM F37 proposal (PAR-03-070), 2005-08-04.

Questions

- “Variation is acceptable between curators but inaccuracy is not.” [6]

GO annotation



http://geneontology.org/GO.nodes.shtml

GO multi–organism annotation

http://geneontology.org/GO.annotation.example.shtml