Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Volume 39, Article 6, 2018

Gamifying Critical Reading through a Genome Annotation Intercollegiate Competition

Ivan Erill[1], Steven Caruso[1], James C. Hu[2]

[1] University of Maryland Baltimore County (UMBC), Department of Biological Sciences, 1000 Hilltop Circle, Baltimore MD 21228, USA
[2] Texas A&M University (TAMU), Department of Biochemistry and Biophysics, College Station TX 77843-2128, USA
(**erill@umbc.edu, scaruso@umbc.edu, jimhu@tamu.edu**)

This workshop explores the use of a web-based inter-collegiate competition to perform Gene Ontology annotation of gene products in sequenced genomes as a tool to motivate and focus students' reading and critical assessment of primary scientific literature while performing a useful task for the scientific community. Students work in teams to generate and peer review gene annotations. The combination of team-based peer competition with a highly structured and publicly-accountable annotation process enhances student involvement and discussion, provides well-defined guidelines for critical reading of primary literature, and engages students in thinking about evidence and source identification in scientific statements.

Keywords: Critical reading, critical thinking, primary literature, Gene Ontology, team-based, online, peer-review, competition

Link to Supplementary Materials: **http://www.ableweb.org/volumes/vol-39/Erill/supplement.htm**

Introduction

Motivation

Motivating students to read and analyze scientific literature remains an outstanding challenge in undergraduate science education. Graduating undergraduates should have the ability to read, interpret and contextualize scientific reports, but conventional approaches, such as paper discussion-based courses for upper-level undergraduates, are too often based on a small set of instructor-selected papers, giving students limited exposure to the literature and no exposure to the process of searching for papers or connecting the literature to computational analyses. At the same time, the life sciences are rapidly shifting towards the use of computational techniques for the analysis of large datasets compiled by emerging and established high-throughput methods. Key to these advances is the availability of high-quality, manually-curated and computer-accessible knowledge, which is stored in international repositories using standardized annotation formats, unique identifiers and ontology-based controlled vocabularies.

This workshop explores the use of a web-based inter-collegiate competition on Gene Ontology annotation to perform functional annotation of gene products in sequenced genomes. It was developed as part of the HHMI SEA-PHAGES program, in which UMBC participates since its inception (Caruso et al. 2009; Jordan et al. 2014). As part of the SEA-PHAGES program, UMBC currently offers two laboratory Phage Hunters courses. The first is dedicated to the isolation and characterization of bacteriophages using microscopy and molecular microbiology methods. The second is devoted to the annotation of the sequenced genomes for some of the isolated phages. The functional annotation of the gene products encoded by bacteriophage genomes is a fundamental component of the genome annotation course. To enhance the quality of student annotations, and to guarantee that the annotation effort becomes visible and useful for the scientific community, in 2015 UMBC teamed up with Texas A&M University, which runs the inter-collegiate annotation competition CACAO (Community Assessment of Community Annotation with Ontologies) as part of the wiki-based Gene Ontology Normal Usage Tracking System (GONUTS) (Renfro et al. 2012). The CACAO-Phage Hunters competition was first piloted as an intramural competition at UMBC in 2015 and extended to a multi-college competition involving several SEA-PHAGES colleges in 2016 and 2017.

Background

The Gene Ontology is a hierarchical, comprehensive systematization of the possible biological roles of a gene product (Ashburner et al. 2000), enabling biocurators to formally describe the involvement of a given gene product in a particular biological process (e.g. response to iron starvation), its specific molecular function (e.g. cholesterol transporter) or its cellular location (e.g. mitochondrial ribosome) (Balakrishnan et al. 2013) (Figure 1). Gene Ontology annotations are made freely available by the Gene Ontology Consortium and have a wide variety of uses (Camon et al. 2004). For instance, Gene Ontology annotations can be used by researchers to uncover functional enrichment patterns in expression data or to compare the makeup of specific biological pathways across different organisms. Given their broad use by the scientific community, the submission of Gene Ontology annotations has to meet formal requirements to guarantee the accuracy of annotations. Importantly, all Gene Ontology annotations must cite a source (typically a peer-reviewed scientific article) containing the evidence on which the annotation is based, and must specify what type of evidence is used in the assertion (e.g. evidence from a mutant phenotype).

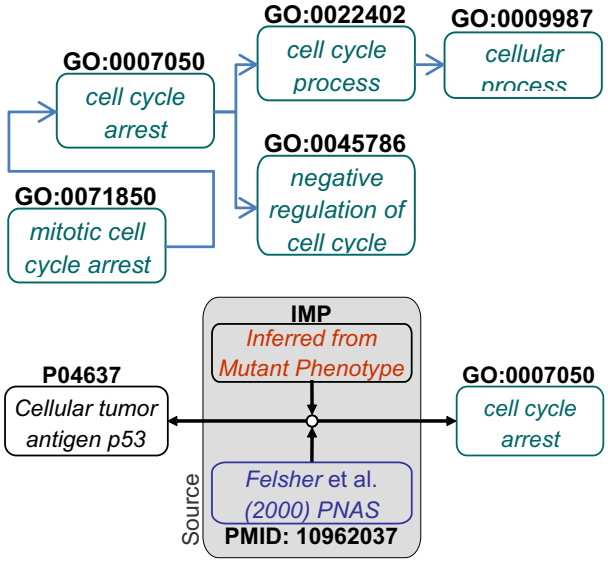


Figure 1. (*Top*) Schematic view of a section of the Gene Ontology covering the ontological neighborhood of the biological process *cell cycle arrest* (GO:0007050). Arrows denote membership relationships between processes (i.e. cell cycle arrest *is_a* cell cycle process). (*Bottom*) Schematic representation of a Gene Ontology annotation. A human *cellular tumor antigen p53 gene* product (UniProt: P04637) is annotated as mapping to biological process *cell cycle arrest* (GO:0007050), based on the results reported by Felsher *et al*. in a 2000 PNAS manuscript: "Overexpression of MYC causes p53-dependent G2 arrest of normal fibroblasts" (PMID: 10962037). The experiment supporting this association in the paper is based on measurements of DNA content (as proxy for cell cycle progression) in wild-type cells and mutants expressing the human papillomavirus E6 oncogene, which facilitates the proteolytic destruction of p53. This is summarized by the evidence code *Inferred from Mutant Phenotype* (IMP).

Implementation

In this lab unit students learn about the structure of the Gene Ontology and its importance for the interpretation of high-throughput biological data. They receive specific instruction on the process of Gene Ontology annotation and, if required, in the use of bioinformatics tools to reliably assess orthology as the means to transfer existing Gene Ontology annotations to genes in the genome they are analyzing. Students work in teams, which compete against other teams from the same and other colleges using the CACAO interface. The CACAO competition is typically organized in alternating, bi-weekly innings dedicated to annotation and/or challenge. To perform annotations, students must read original articles and specifically describe the experiments in those articles supporting their conclusions. Their claims can be assessed and challenged by other teams who have read the article, and students must address those challenges by revisiting the literature source and revising their annotations accordingly. Students and their teams are usually given credit for accurate annotations and challenges, prompting them to carefully read and assess the experiments reported in the articles they use as sources for their annotations. As a result of the peer-competition scheme, students perceive the reading of scientific literature as a competitive challenge, rather than an obligation, and discuss the interpretation of the findings in each article with their team, thereby bolstering the learning experience associated with the reading of primary literature.

Basic Unit Implementation

The lab unit developed at UMBC showcases the application of this methodology to the annotation of bacteriophage genomes, but the approach is generalizable to any publicly available genome. Students may be asked to annotate genes from their favorite organism or genes of interest to the instructor based on a leading topic (e.g. genes involved in metabolism as part of a cell biology lab). The two key ingredients of the lab unit are its setup as a publicly-visible intercollegiate competition, which motivates students to carefully evaluate their assessments based on critical reading of the literature, and the use of the Gene Ontology annotation framework, which provides a principled and targeted way for students to take upon the task of critically reading manuscripts, weighting what constitutes acceptable evidence and extracting relevant information from scientific papers.

Unit Setup

The activity does not have a substantial upfront formal setup time beyond that invested by the instructors

in familiarizing themselves with the Gene Ontology and the CACAO competition. At the beginning of the semester, the instructor must submit by email the list of students and their associated teams to the CACAO staff (ecoliwiki@gmail.com). Once student accounts have been activated, students can participate in the general CACAO competition (http://gowiki.tamu.edu/wiki/index.php/CACAO), following the predefined annotation and challenge innings.

Unit Organization

Participation in CACAO is entirely and intentionally flexible in both format and allocated time. Instructors may design their CACAO activity to fit the goals of their course, allowing students to participate on all or just a few competition innings, defining the number and quality of expected annotations, constraining or not the subject and/or literature sources and adjusting the grading rubric as desired. Groups of instructors sharing a common topic of interest may request a specific CACAO competition devoted to their topic, and coordinate the specific dates and structure of the competition with CACAO staff.

Time Considerations

Students will need at least one week to familiarize themselves with the concepts behind CACAO and Gene Ontology annotation and with the CACAO interface. Annotations can take anywhere from 20 minutes to a few hours of student work, depending on the clarity of the literature source, the difficulty of the concepts covered and the experience of each student. Time must be allocated by the instructor to assess students' annotations. The CACAO team will provide some assessment support, but the Gene Ontology annotations performed in CACAO are *de facto* exercises in critical reading and reporting, and time should be allocated by the instructor to grade them accordingly.

Target Audience, Difficulty and Required Training

This activity is targeted to undergraduate students in their sophomore or junior year. The activities performed by the students in this unit do not require computing literacy beyond the ability to efficiently navigate web resources. An optional part of the activity (the use of *transfer annotations* if one desires to annotate genes by similarity) does require limited training in the use of widespread bioinformatics techniques for determining orthology, such as BLAST.

Considerations Regarding Critical Reading

The activity focuses on the reading and critical assessment of primary literature and some students may find this challenging. However, the Gene Ontology annotation framework provides a highly structured scaffold to identify and evaluate specific claims made by the authors of a scientific manuscript, facilitating greatly the process for uninitiated students and providing a stepping-stone to the standalone reading of primary literature that students may encounter in upper-level courses. Students may also struggle initially with formal concepts relating to ontologies, which they will likely be unfamiliar with, and with the formalism of Gene Ontology annotations. Time should be allocated to address conceptual issues and to bring formalisms to light through the use of real life examples (e.g. an ontology of cars).

Implementation at Other Levels

This unit could conceivably be implemented at the freshman level, even though we have not directly tested it. In such a setting, critical reading of the literature and comprehension of the ontology formalism will likely become important issues. The instructor should plan for a longer commitment in class time to train on and illustrate both aspects, leveraging the formalism of Gene Ontology annotations to restrict the scope of the critical reading effort.

Notes on Student Handouts and Instructor Notes

As mentioned, CACAO is very flexible in terms of implementation, both content- and format-wise, within a laboratory or lecture course. For this reason, this workshop provides guidance on the implementation and setup of the course, but leaves the specific details of the implementation up to the instructor. This is reflected in the student handouts, which are mainly written for a unit implementation in which the instructor has provided clear guidelines on the scientific literature and gene products to annotate. Background on the main unit concepts is provided in the student handouts, but is applicable to instructor notes. The last sections of the student handouts include troubleshooting instructions for more open-ended implementations of the unit (e.g. broadly targeting bacteriophage genomes for annotation, with no preselected publications), regarding the procedures for identifying genes and manuscripts. The need for such additional instructions is discussed in the instructor notes.

Student Outline

Overview

This lab unit is devoted to genome annotation. In it you will compete in teams with students from your own and other universities to annotate different aspects of genes (their molecular function, location in a cell or their participation in specific cellular processes) using the Gene Ontology as a reference framework. Teams participating in the CACAO competition earn points by submitting correct annotations and challenging inaccurate ones made by other teams.

Objectives

After completing this lab unit you should be able to:

- Explain to a lay audience what ontologies are and what they are used for
- Discuss biocurator as a viable career path in the life sciences
- Summarize how ontologies can be applied to biology
- Describe the Gene Ontology structure and its main sub-ontologies
- Critically review and assess peer-reviewed primary literature in biology
- Generate and critique Gene Ontology annotations based on primary literature
- Utilize the CACAO interface for making GO annotations
- Navigate the QuickGO and UniProt websites
- Differentiate GO terms, evidence codes and their usage
- Explain the differences between different types of GO annotations
- Be familiar with the CACAO interface for making GO annotations
- (Optional) Leverage BLAST and other tools to infer homology

Unit Structure

This lab unit is broadly structured in three different periods: instruction, annotation/challenge and revision. During the instruction period you will receive basic training on the concept of ontology, the overall architecture of the Gene Ontology and the main concepts behind Gene Ontology annotations and their usefulness to the scientific community. You will also be given time to register and familiarize yourselves with the CACAO web interface for Gene Ontology annotation. After completing instruction, you will be able to participate in the annotation and challenge innings defined by the CACAO competition. During annotation innings, you and your team can submit as many Gene Ontology annotations as you like, but you should bear in mind that unsubstantiated or inaccurate annotations will likely be challenged and not will earn you credit. During challenge innings, you can critique other teams' annotations, providing feedback on any errors or inaccuracies present in them. As with annotations, challenges must be substantiated to earn credit. After the last challenge inning is over, you will have the chance to address any outstanding problems raised by challengers or instructor feedback. Once this final revision period is complete, your annotations are considered final and cannot be further modified. If they are accepted, your annotations will be submitted to the Gene Ontology Consortium and incorporated into their growing knowledgebase.

Background

Ontologies and the Gene Ontology

An ontology is a formal representation of a particular real-world domain (Gruber 1993). Ontologies define entities that exist in the real world (e.g. pizzas and their ingredients) and the relationships between them (e.g. toppings are *parts of* pizzas) (**Figure 2**). Ontologies serve two main simultaneous purposes: (1) by providing a unified, controlled vocabulary ontologies eliminate synonyms (e.g. veggie pizza and vegetarian pizza) and disambiguate homonyms (i.e. same word having two different meanings in different contexts); (2) by defining relationships among entities and mappings between entities and their real-world instances ontologies enable computers to reason over the ontology and perform inferences on real-life applications.

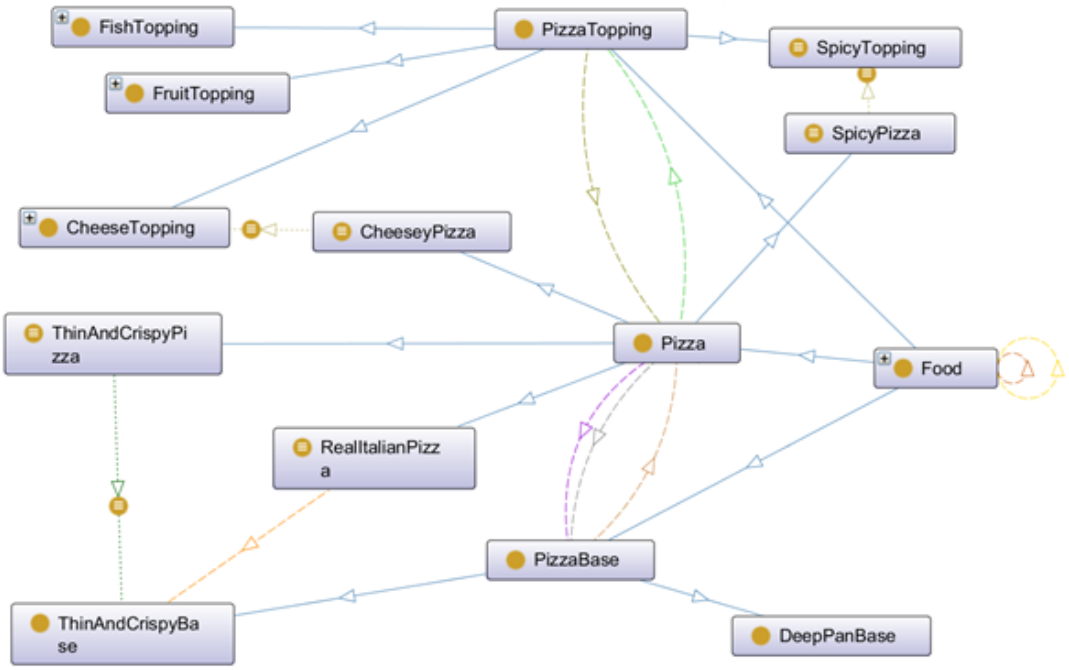


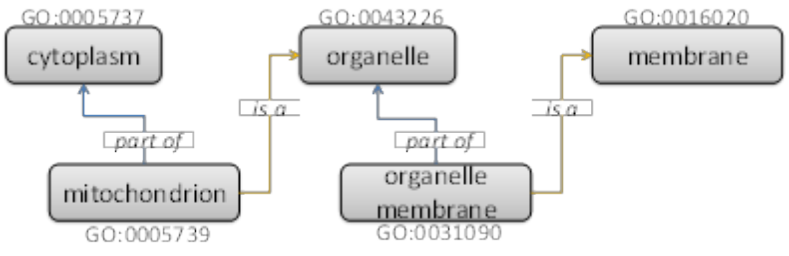
Figure 2. Partial view of the Pizza Ontology developed by ontology researchers at the University of Manchester (Horridge et al. 2004). The figure shows the main entities (Food, Pizza, PizzaTopping and PizzaBase) and the different relationships (e.g. RealItalianPizza *is a* Pizza (and hence Food) that *has part* ThinAndCrispyBase, which *is a* type of PizzaBase). Image was rendered using the OntoGraph Protégé plug-in.

The Gene Ontology (GO) is a specialized ontology that formalizes knowledge on three key aspects of gene products (i.e. proteins, RNAs and derived biomolecules) (**Figure 3**). These three aspects make up the three GO sub-ontologies: molecular function, biological process and cellular component.

- Molecular function refers to activities that occur at the molecular level, such as "catalytic activity" or "binding activity". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform them, and do not specify where, when, or in what context the action takes place.

- Biological process refers to a series of events accomplished by one or more organized assemblies of molecular functions. Examples of broad biological process terms are "cellular physiological process" or "signal transduction". The general rule to assist in distinguishing between a biological process and a molecular function is that a process must have more than one distinct step.

- Cellular component denotes a component of the cell that is part of a larger object, such as an anatomical structure (e.g. rough endoplasmic reticulum) or a gene product group (e.g. a ribosome or a protein dimer)



ID	GO:0043226
Name	organelle
Ontology	Cellular Component
Definition	Organized structure of distinctive morphology and function. Includes the nucleus, mitochondria, plastids, vacuoles, vesicles, ribosomes and the cytoskeleton, and prokaryotic structures such as anammoxosomes and pirellulosomes. Excludes the plasma membrane.

Figure 3. (left) Schematic view of a section of the Gene Ontology, depicting the relationship between different cellular components. The mitochondrion *is a* type of organelle and is also *part of* the cytoplasm, in the same manner that an organelle membrane *is part* of an organelle but *is a* type of membrane too. (right) All terms in the Gene Ontology are defined by a unique identifier and contain the consensus name, synonyms (if any), their primary sub-ontology and a crisp definition.

Gene Ontology Annotations

Beyond an exercise in modeling reality, creating ontologies is not that useful if one cannot map ontology terms to real-world entities. The Gene Ontology provides a highly structured framework to make such mappings, by means of Gene Ontology annotations. Once gene products (e.g. proteins or small regulatory RNAs) in a genome have been mapped to the Gene Ontology one can apply statistical inference and machine learning approaches to interpret data and perform genome-wide comparison. One such example is the use of the Gene Ontology in interpreting data from transcriptome analysis (du Plessis, Škunca, and Dessimoz 2011). If a genome has been mapped to Gene Ontology terms, one can interrogate sets of relevant genes (e.g. genes highly expressed in anoxic conditions) to see if they are enriched in particular subsets of the ontology (e.g. they preferentially map to stress response terms)

A Gene Ontology annotation is therefore a mapping from a given gene product to a specific Gene Ontology term (**Figure 4**). Beyond these two main components, the formalism in Gene Ontology annotations requires that the annotation contain two additional elements: a reference and an evidence code (Balakrishnan et al. 2013). The combination of these two elements is referred to as the *source* for the annotation.

Evidence Codes

Conventional Gene Ontology annotations are typically made by professional biocurators (Howe et al. 2008). Biocurators search the literature for relevant publications containing experimental work that demonstrates the molecular function of a gene product, its involvement in a biological process and/or its location in a particular cellular component. After critically reviewing the results reported in the manuscript, biocurators identify an adequate Gene Ontology term that reflects the findings and determine what type of experimental evidence was used to demonstrate them. For instance, if the authors created a mutant of the human p53 protein and then observed that after irradiation mutant cells, compared to the wild-type, did not advance beyond the G1/S regulation point, a biocurator would use the evidence code Inferred from Mutant Phenotype (IMP) and the GO term "cell cycle arrest" (GO:0007050) to record this observation (**Figure 4**).

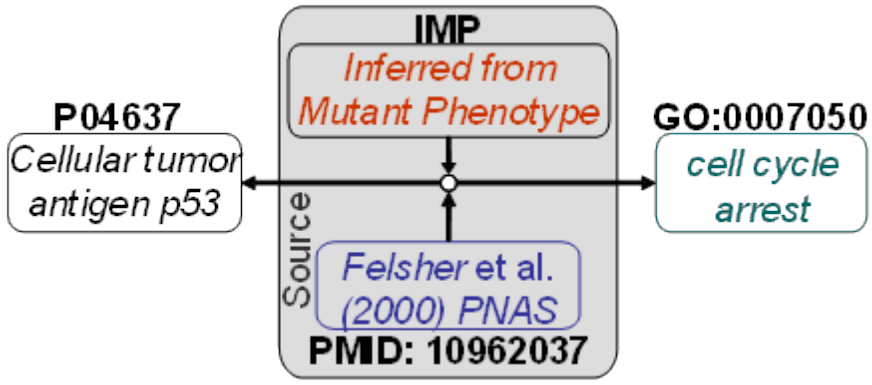


Figure 4. Schematic representation of a Gene Ontology annotation. A human *cellular tumor antigen p53 gene* product (UniProt: P04637) is annotated as mapping to biological process *cell cycle arrest* (GO:0007050), based on the results reported by Felsher *et al.* in a 2000 PNAS manuscript: "Overexpression of MYC causes p53-dependent G2 arrest of normal fibroblasts" (PMID: 10962037). The experiment supporting this association in the paper is based on measurements of DNA content (as proxy for cell cycle progression) in wild-type cells and mutants expressing the human papillomavirus E6 oncogene, which facilitates the proteolytic destruction of p53. This is summarized by the evidence code *Inferred from Mutant Phenotype* (IMP).

In some cases, authors may use computational tools to determine the function of a gene product. For instance, based on sequence analysis a manuscript might report that the mouse protein P02340 is a close homolog of the human p53 protein (P04637) and that it also contains a DNA-binding motif, indicating that P02340 binds DNA in the same way as its human homolog. In such a case, the biocurator might use GO term "DNA binding" (GO:0003677) in conjunction with the evidence code Inferred from Sequence Orthology (ISO) and the identifier for the human p53 protein (P04637) that is used to make such assertion. A full list of evidence codes with usage examples is available at: http://geneontology.org/page/guide-go-evidence-codes. Gene Ontology evidence codes have now been superseded by the Evidence and Conclusion Ontology (ECO), which defines the relationships between different types of evidence (e.g. "loss-of-function mutant phenotype evidence" (ECO:0000016) *is a* type of "mutant phenotype evidence" (ECO:0000015)) (Chibucos et al. 2014). While CACAO still uses native GO evidence codes, it is often convenient to navigate ECO (http://www.evidenceontology.org/) in order to identify the proper GO evidence code to use.

Alternative Methods for Gene Ontology Annotation

Even though large, the amount of available experiments determining different aspects of gene products is vanishingly small when compared to the number of genes present in sequenced organisms. Members of the Gene Ontology Consortium and others have developed tools to automatically annotate gene products in genomes using computational methods to establish homology with annotated genes or to parse manuscripts in order to extract relevant information. The reliability of these methods increases yearly, but computerized approaches are still very far from being as thorough and accurate as human biocurators. For this reason, all computer-generated annotations with no human supervision are tagged with the Inferred from Electronic Annotation (IEA) evidence code.

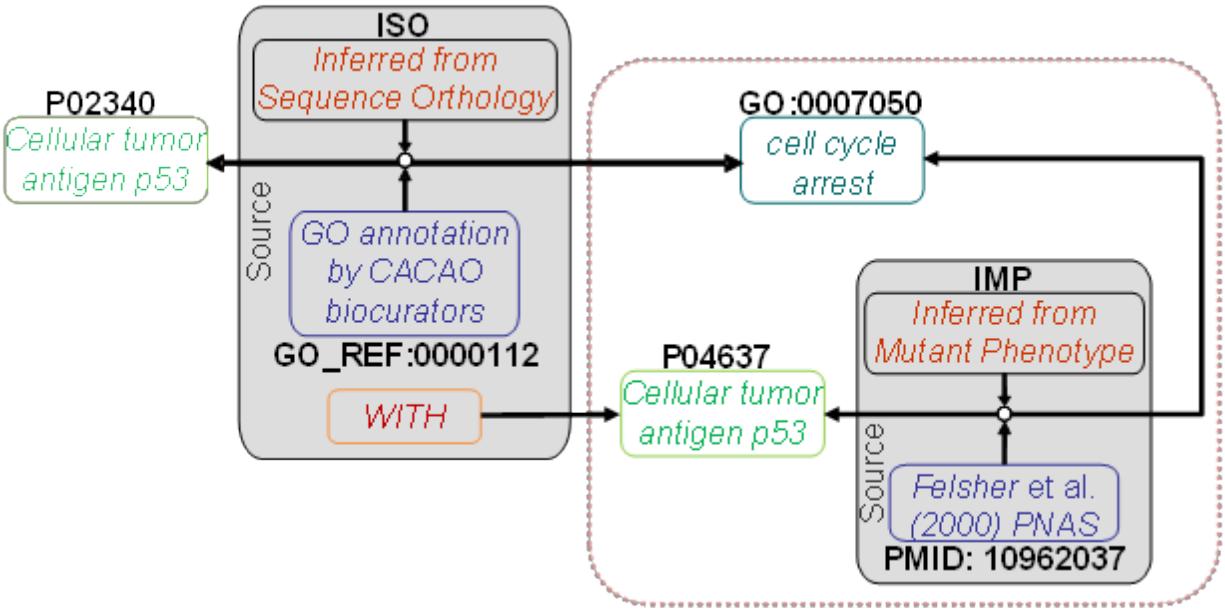


Figure 5. Schematic representation of a "transfer" Gene Ontology annotation. Using the computational tools described in GO_REF:0000112, CACAO biocurators determine that the mouse p53 protein (P02340) is homologous to the human p53 protein (P04637), which has been previously annotated (**Figure 4**) as being involved in *cell cycle arrest* (GO:0007050) based on experimental (IMP) results published by Felsher *et al.* (PMID: 10962037). The assignment of the GO:0007050 term to the mouse P02340 protein is formally defined as deriving from a computational approach (Inferred from Sequence Orthology; ISO) reported in a published reference (GO annotation by CACAO biocurators; GO_REF:0000112) that establishes the homology of the mouse P2340 protein WITH the human p53 protein (P02340), allowing the biocurator to conclude that the mouse P2340 protein also participates in cell cycle arrest (GO:0007050).

Gene Ontology annotations require that a source be referenced in the annotation. Conventionally, the source is a peer-reviewed scientific manuscript reporting experiments, but there are cases in which we may want to capture results following a well-established methodology that are not published in peer-reviewed manuscripts. For instance, biocurators working on the Mouse Genome Informatics (MGI) project at the Jackson Laboratory have developed well-established computational processes to establish homology between rat and mouse genes. MGI biocurators examine, verify and contextualize these computational predictions and use them to assign GO terms to mouse genes based on experimental annotations of rat genes. When they do so, they use a special type of reference (a GO reference; GO_REF:0000008) that describes the methodology they have used in the annotation. As a sutdent participating in CACAO you can make use of a dedicated GO reference (GO_REF:0000112) to annotate gene products for which there is no available experimental literature. As in the case of MGI biocurators, you will do so through the establishment of homology with gene products containing experimental annotations using a variety of computational methods. Instead of referencing a peer-reviewed scientific manuscript, these "transfer" annotations will reference a source composed of a computational evidence code (e.g. ISO), the CACAO GO reference (GO_REF:0000112) and the identifier of the homologous protein containing the experimental annotation (**Figure 5**).

Performing Gene Ontology Annotations

Creating a Gene Ontology annotation entails three separate steps: reading and assessment, mapping and annotating (Figure 4). The first, and most complex step, is the critical reading of a peer-reviewed scientific manuscript and the

assessment of the claims made therein. Mapping refers to the identification in reference databases of the entities detailed in the manuscript (i.e. the gene product accession, the GO term and the evidence code). The last step concerns the use of CACAO to perform the annotation and submit it for review. There are many approaches to reading scientific manuscripts, but for the purposes of Gene Ontology annotations the following procedure is recommended:

- Read the abstract carefully to get a general idea of what the paper is about and what are the main claims made by the authors. Hopefully, one of these claims will involve the function, process or location of gene product.
- Read the introduction and attempt to identify the specific species/strain the authors work on and accurate descriptions (or accession number, if provided) of relevant protein products.
- Use the NCBI RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) and EBI UniProtKB (http://www.uniprot.org/) services to identify the accession numbers of the protein products referenced by the authors (**Supplementary material 1**). If you cannot identify a valid accession number for your gene product, contact your instructor.
- Look at the *Material and Methods* section to familiarize yourself with the main experimental/computational techniques used by the authors.
- Read through the *Results* (or *Results and Discussion*) section. Most annotation-worthy claims in a scientific manuscript will be backed up by figures or tables. Identify the manuscript regions that cite a given figure to understand what the authors seek to accomplish (i.e. demonstrate) with the experiments reported in the figure. A figure reporting an experimental procedure can be the source of one or more annotations.
- Use the QuickGO (http://www.ebi.ac.uk/QuickGO/) or AmiGO (http://amigo.geneontology.org/) web services to see if the aspect the authors seek to validate through their experiments corresponds to a Gene Ontology term. The autocomplete function will suggest GO terms matching your query words. Use the *Ancestor Chart* and *Child Terms* list to navigate the ontology from any given start point. These services also provide guidelines for the annotation of specific topics (e.g. cell death). You should always aim to annotate the most specific GO term possible (i.e. if the manuscript reports the involvement of a gene in apoptosis in hepatocytes you should annotate "hepatocyte apoptotic process" and not its parent term "apoptotic process"). If you cannot find a matching Gene Ontology term, or you believe the existing ones are inadequate (e.g. too general) for the aspect you are trying to annotate, contact your instructor. CACAO has a guide on how to submit new Gene Ontology terms for approval by the Gene Ontology Consortium (**Supplementary material 2**). CACAO students have contributed several GO terms in the past.
- Take your time analyzing the table/figure referenced in the text, and reading the figure/table legend and the text referencing it. Try to identify the type of experimental technique used in the figure (or within a figure panel) and to understand how the use of such technique allows the authors to validate the particular aspect of the gene product they identify in the main text. Ask yourself: does (do the authors claim that) the figure allows us to conclude something regarding the gene product (e.g. does it tell us that it performs a certain molecular function, that is localizes somewhere in the cell or that it participates in a specific biological process)?
- Map the experimental method to one of the Gene Ontology evidence codes. A decision tree and sampler for picking the correct experimental code are available in the CACAO webpage (**Supplementary material 3**, **Supplementary material 4**). The Evidence and Conclusion Ontology (ECO) is also a good resource to navigate experimental techniques and identify the relevant Gene Ontology evidence codes (which map to ECO root terms).
- Note that some evidence codes are not allowed in CACAO. In particular CACAO does not accept IPI (Inferred from Physical Interaction) and IEP (Inferred from Expression Pattern). These codes are not accepted in the competition to avoid the use of manuscripts reporting a high-throughput experiment to perform large numbers of annotations. Evidence codes based on traceable (TAS) or untraceable author statements (NAS), or inferences made by curators (IC) are also not accepted in CACAO. These terms are mostly in disuse and reserved to professional biocurators.
- Note down the GO term and evidence code, the gene product accession number and the manuscript PubMed ID (which you can find through the NCBI Entrez interface; **Supplementary material 5**).
- Write a concise explanation of the deductive process you have followed to determine that the annotation is possible and the terms/codes you have chosen to use. You have examples of such summaries in all previous CACAO annotations.
- Remember that a single manuscript may contain data for several annotations on one or multiple aspects of a single or multiple gene products.

The CACAO website contains example papers to train on before you perform your first annotation (**Supplementary material 6**).

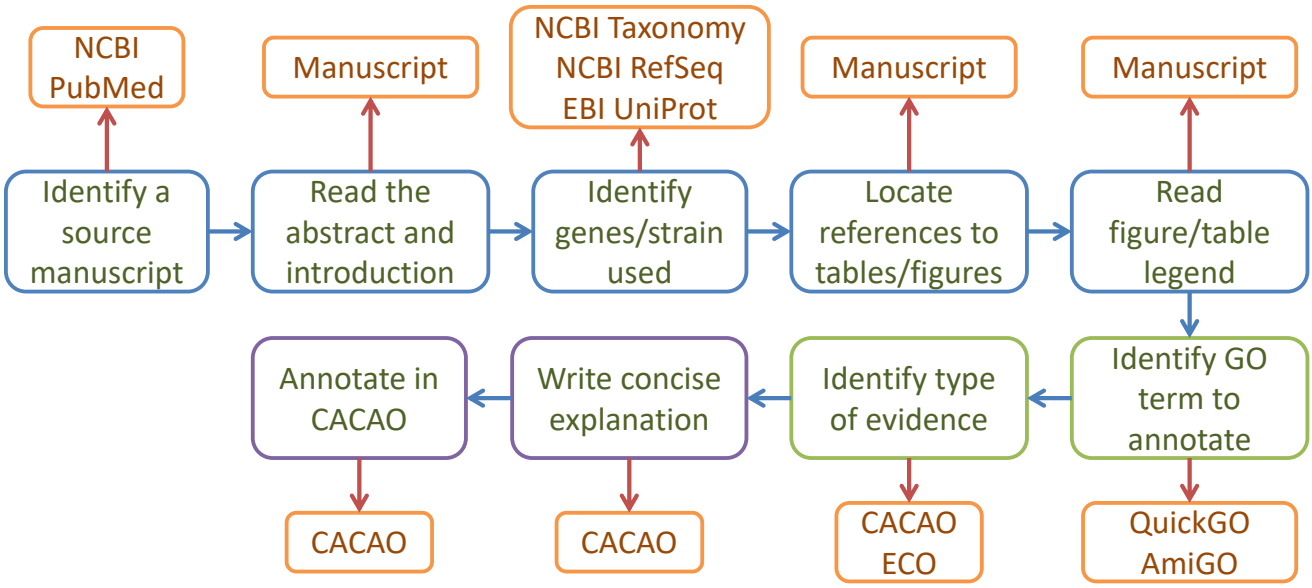


Figure 6. Schematic representation of the steps in a Gene Ontology annotation and the different resources (orange boxes) used in the process. Blue boxes correspond to reading and assessment steps, green to mapping steps and purple to annotation. After completing a successful annotation, students should try to determine if further annotations can be extracted from the manuscript.

Performing Gene Ontology annotations with CACAO

Performing Gene Ontology annotations in CACAO is fairly straightforward once you understand the basic elements of an annotation. CACAO provides a simple, intuitive wiki interface to generate Gene Ontology annotations. Creating a new Gene Ontology annotation in CACAO requires three distinct steps: (1) searching/creating a gene product page, (2) creating the annotation and (3) saving the changes. The following illustrates these three basic steps with the annotation example from **Figure 4**. A more detailed step-by-step annotation example is available on the CACAO website (**Supplementary material 7**).

Searching/Creating a Gene Product Page

The first thing to do is to search CACAO and check whether the gene product already exists in the system. If the gene product is not yet in CACAO, you can create a new gene product page by clicking on *Create New Gene Page* (**Figure 7**). When you do so, CACAO will import all relevant data for the gene, including existing Gene Ontology annotations. You should check whether annotations from the manuscript you desire to annotate from have already been made and verify that the annotation that you intend to perform has not been previously made.

Creating an Annotation

In the gene product page, at the bottom of the list of existing annotations, you will find an *edit table* link (**Figure 7**). Clicking on it will bring you to the annotations table edit page and, at the bottom of the table you will find an *Add row* button that will take you to the data entry page for the annotation (**Figure 7**). On the data entry page, you can enter all the relevant elements of a Gene Ontology annotation: the GO term, the manuscript PubMed ID, the evidence code and your rationale for the annotation.

Saving an Annotation

Once you have entered all the annotation elements, you must save the annotation. In CACAO, which is a wiki, this involves a two-step process. You must first save the row, and then save the table back to the wiki (**Figure 7**).

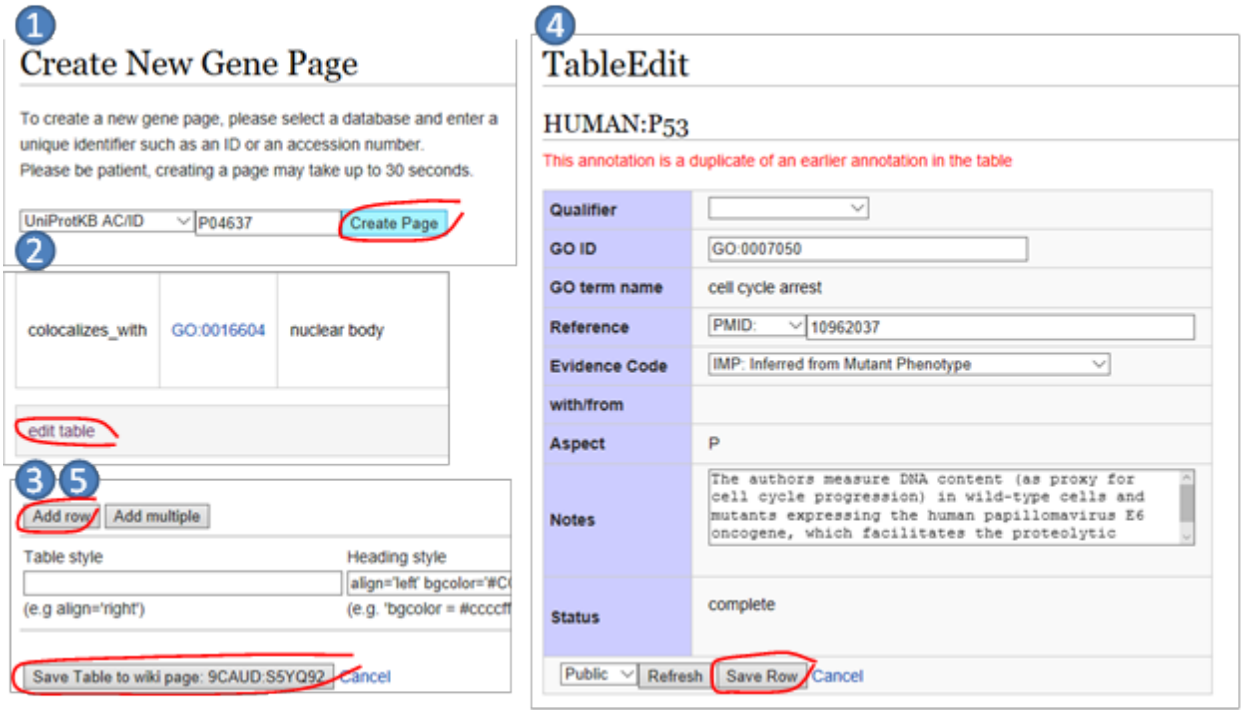


Figure 7. Essential steps of a Gene Ontology annotation in CACAO. (1) If not existent, a gene product page must be created. (2) At the bottom of the annotation list, click *edit table*. Once the edit page for the table loads, click on *Add row* to create a new annotation. (4) Enter the relevant Gene Ontology annotation information, including a detailed note explaining your rationale for the annotation. Click refresh to populate GO term name and aspect and hit *Save Row* before leaving the page. (5) Once you return to the edit table page, you must also *Save the table to wiki page* for the added row (annotation) to be saved.

Identifying Manuscripts and Gene Products

Identifying manuscripts with reliable Gene Ontology annotations is not trivial, and in many ways it is more art than science. For starters, many manuscripts simply do not contain relevant annotations for gene products. Some articles are reviews, which may well cite original research articles with relevant annotations but which, by themselves, cannot be used for annotation (since experiments are not carried out in the article). Many other articles, by their nature and topic, just do not contain research material for gene product annotation. For instance, an epidemiological article is unlikely to demonstrate the cellular component, molecular function or biological gene process a gene product locates, performs or participates in.

Finding Manuscripts for Annotation

Finding manuscripts for annotation should not be too difficult (NCBI PubMed currently contains more than 27 million citations for biomedical literature), but can get a bit tricky depending on your specific assignment. NCBI PubMed (https://www.ncbi.nlm.nih.gov/pubmed) is by far the best resource for this purpose, and it has the added bonus that, once you locate the manuscript, you will have a PubMed identifier (PMID) for it (CACAO works primarily with PubMed identifiers, even though other manuscript identifiers are accepted under special circumstances).

Searching NCBI PubMed

The NCBI PubMed (and other NCBI databases) is accessed through a comprehensive search interface that predates Google by almost two decades. You can search with simple terms <*Escherichia coli*>, or enforcing the combination <(*Escherichia* AND *coli*)>. You can specify that you want to see the words in the title or abstract (*Escherichia*[Title/Abstract]) AND (*coli*[Title/Abstract]). You can also set up personalized Filters to see specific types of records (like those linking to a protein record. Full instructions on how to use PubMed search can be found at https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.PubMed_Quick_Start.

Searching via Other Services

NCBI PubMed is a powerful and convenient resource, but by no means the only one. Google Scholar (https://scholar.google.com) can do a fair job at locating manuscripts that might not show up easily on PubMed. PubMedCentral (https://www.ncbi.nlm.nih.gov/pmc/) and EuropePMC (https://europepmc.org/) provide different types of

search features to retrieve open-access manuscripts (which will also have a PMID and which do not depend for access on the particular journal subscriptions of your school).

Linking Manuscripts to Gene Products

In theory, an article reporting experimental work on a gene product should be an obvious source of Gene Ontology annotations. However, this is not necessarily the case. Given that performing a Gene Ontology annotation is quite time consuming, you should try to first triage any candidate manuscript before investing too much time on it. The next sections provide a few clues on what can go wrong and how to identify it (and address it if possible).

UniProt Identifiers

Annotations in CACAO need a unique identifier for the gene product. CACAO restricts annotations to a specific type of gene product (proteins) and uses a single source for protein identifiers: the UniProtKB database (http://www.uniprot.org/uniprot/). This means that in order to perform a Gene Ontology annotation in CACAO you will need a UniProtKB identifier. And therein lies the problem, because not all the species and strains are represented in UniProtKB. In the last few years, there has been an unprecedented surge in the number of (mostly bacterial) genomes sequenced, leading to thousands of identical protein records predicted from the genome sequences (*Escherichia coli* alone has almost 4,000 complete genomes available, most of the with identical translated protein sequences). Faced with this surge, UniProt decided to implement a redundancy reduction strategy (http://www.uniprot.org/help/proteome_redundancy) by designating some strains as reference proteomes in UniProt, and relegating other strains to the UniParc archive (with no UniProtKB identifiers). If you cannot find a match in UniProt for the gene product reported in the manuscript, check with your instructor and/or CACAO staff (ecoliwiki@gmail.com). It is possible to annotate the gene product reported in the manuscript using the reference UniProt protein, but you should make this explicit in the annotation notes. Specifically, you should be able to locate and report in the notes the accession number of the proteome of the particular strain your organism is in and of the reference proteome you will be using (through http://www.uniprot.org/proteomes/), and detail in the notes how you have established that the protein you are annotating is a homologue of the one in the reference proteome, following the guidelines in http://gowiki.tamu.edu/wiki/index.php/Category:CACAO_GO_REF.

Undefined Species/Strain

Believe it or not, many scientific manuscripts reporting experimental results do not clearly identify the species/strain the work has been carried out on. Or, if they do so, they identify them in a substantially oblique manner. For instance, some manuscripts identify the strain they work on with the name of the derivative strain (e.g. an *E. coli* K-12 MG1655 strain in which a specific gene has been knocked out). The specific strain used should be named in the Abstract, the Introduction or the Materials and Methods section. In many cases, a Table with the strains used will be listed in the Materials and Methods section. If the authors use a derivative strain, they may mention at some point where it derives from, or a quick Google search with the derivative strain name may do the job. If both venues provide infructuous, NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) or Genomes Online (GOLD; https://gold.jgi.doe.gov/organisms) may do the trick. If you cannot easily find the parent of a derivative strain with these resources or if the authors simply do not state the strain's name, discard the manuscript and look for another one.

Undefined GENE PRODUCT

Gene names for which a likely annotation is possible will typically be mentioned in the abstract or the introduction (and obviously more in detail in the Results section), so scanning these two initial segments of the manuscript for a gene mention in some kind of assertive statement (e.g. "we show that") will allow us to quickly gauge whether a gene product may be annotated. As with strains, authors are sometimes not very precise about what gene or genes they are working on. This is particularly problematic in model organism (fly, worm, mouse…) and human literature, where gene names have a long history, typically multiple original naming conventions with their adherents and detractors, and where the model organism context tends to imply that the reader will know about the gene through offhand references. In many cases, a search on NCBI RefSeq or EBI UniProt with the synonym used in the manuscript will quickly resolve the issue, but in some others this may not prove easy. In such cases, as with undefined strains, it is better to discard the manuscript and move onto another.

Cited References

Balakrishnan R, Harris MA, Huntley R, Van Auken K, Cherry JM. 2013. A guide to best practices for Gene Ontology (GO) manual annotation. Database J. Biol. Databases Curation 2013:bat054. doi:10.1093/database/bat054.

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 J. Biol. Databases Curation 2014. doi:10.1093/database/bau075.

Gruber TR. 1993. A Translation Approach to Portable Ontology Specifications. Knowl. Acquis 5:199–220. doi:10.1006/knac.1993.1008.

Horridge M, Knublauch H, Rector A, Stevens R, Wroe C. 2004. A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0. The University Of Manchester.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger, S, White, O, and Yon Rhee, S. 2008. Big data: The future of biocuration. Nature 455:47–50. doi:10.1038/455047a.

du Plessis L, Škunca N, Dessimoz C. 2011. The what, where, how and why of gene ontology—a primer for bioinformaticians. Brief. Bioinform. 12:723–735. doi:10.1093/bib/bbr002.

Materials

This lab unit requires that each student have access to a computing device with Internet connection. Even though the CACAO website is accessible using tablets and smartphones, it has not been optimized for display or interaction on these units and hence desktop or laptop computers are recommended. Extensive textual and audiovisual training materials for students and instructors, including videos illustrating the annotation process, are available on the CACAO website (see http://gowiki.tamu.edu/wiki/index.php/CACAO and the supplementary material).

Notes for the Instructor

As with any organized activity, logistic issues may arise during this lab unit. Most can be readily addressed through simple interventions and adequate planning of the activity. The sections below detail some of the issues encountered in the past by participating instructors and proposed remedial actions, as well as some general notes on lab unit implementation. The CACAO website contains a detailed instructor manual covering all aspects of the CACAO competition and with step-by-step instructions on how to perform different tasks (**Supplementary material 8**).

Registering Students

You should contact the CACAO staff before the start of the semester, and send in your team rosters as soon as possible after classes start. The registration system will send an email to students with a hyperlink they need to access to activate their accounts. Depending on your institution's email filter configuration, such emails may be sent to a spam folder. Remember to instruct your students to check their inbox and spam folders in the days following your team registration request. See the CACAO webpage Help:CACAO for additional instructions on setting up teams and competitions.

CACAO Innings

CACAO is organized in innings, which can be devoted to annotation, challenge or both (open). During challenge innings students cannot enter new annotations or revise old ones. Make sure students are aware of the planned timeline for each inning and that they plan their annotations accordingly.

Student Training and Instructor Feedback

Most, if not all students participating in the lab unit will be completely unfamiliar with the concept of ontologies and annotation, and will have limited, if any, experience in reading primary literature. Even though the Gene Ontology structures the critical reading process, instructors should expect some confusion and errors during the first round of annotations. Training material is available to walk students through the annotation process, but past experience shows that a live walk through an example annotation by the instructor, requesting input from the students on the different steps (e.g. how to interpret a figure, how to choose the appropriate GO term or evidence code, etc.) is very effective in minimizing misconceptions about the GO annotation process. Providing instructor feedback right after, or within, the first annotation period is also a very effective way to make sure students understand what is expected in an annotation and why some of the entered parameters may be incorrect. A detailed guide on how to enter student feedback, assess and grade annotations, and generally interact as an instructor with the CACAO system is available on the CACAO website (**Supplementary material 8**). It is also important to emphasize that CACAO annotations require that students enter a note detailing the deductive process in their annotation. Student notes are a primary element in assessing whether an annotation is correct or not, both for CACAO instructors and for the ultimate end users of the annotation.

Scientific Scope

The scientific scope of a CACAO-based lab unit is dictated entirely by the instructor. It can be restricted to a particular family of gene products (e.g. transcription factors), to a specific genome (e.g. kangaroo), a taxonomical clade (e.g. Diprotodontia), a biological process (e.g. cell migration) or any other arbitrary subdivision. While it is tempting to not impose restrictions and allow students to annotate any gene of their interest, this has some practical drawbacks. On the one hand, the difficulty of annotations can vary substantially depending on the topical area. This may be due to different factors (e.g. in some fields gene/strain nomenclature is very casual, complicating their mapping to univocal gene product identifiers), but can impact the ability of some students to generate robust annotations. On the other hand, instructors that decide to remove restrictions on annotation must be comfortable with the assessment of methods and the interpretation of results in a very wide range of biology domains, or must be willing to allocate the time to familiarize themselves with a substantial number of unforeseen topics.

Teaching Assistants

Like undergraduate students, graduate students in the biological sciences serving as teaching assistants will likely not be familiar with ontological concepts and the Gene Ontology. Hence, the instructor should be prepared to provide some advance training to teaching assistants on the essentials of ontologies and the Gene Ontology, and

have them practice with the CACAO interface beforehand. As with instructors, care should be taken to make sure that teaching assistants are familiar with the range of topics selected for annotation, and with the use of the necessary resources to locate and select manuscripts and protein identifiers. Our experience reveals that graduate students with no bioinformatics background may face a steep learning curve in this unit, but that proper guidance and the use of the extensive training resources available in CACAO should allow them to perform their work after one week of training.

Competition Scores

CACAO automatically registers annotations and challenges. Once they are assessed and approved, valid annotations and challenges are used to compute the final scores of each team. The scoreboard will display scores during the entire competition, but students should be made aware that only accepted annotations will be used in the final tally.

Grading

Grading of the lab unit is typically linked with CACAO team and/or individual scores, but grading particulars are entirely up to the discretion of the instructor. Different institutions have used different grading criteria in the past, with variable outcomes. An approach often considered by instructors is to offer points for posted annotations and issued challenges, but this typically results in students focusing on the quantity, rather than the quality, of annotations and challenges. An important goal of CACAO is to generate high-quality annotations that can be submitted to the Gene Ontology and used broadly by the scientific community, and quality is hence encouraged over quantity. Instructors should bear in mind that they are ultimately responsible for the grading of their school's annotations and hence focusing on quantity will increase the grading effort. It will also likely diminish the chances of the school winning the competition and decrease the impact of both peer review and instructor feedback. As a consequence, most schools currently participating in CACAO use a grading system based on the number of approved annotations and challenges.

CACAO as Unit for Deployment in a Lab Course

CACAO is designed to be implemented in a very flexible manner, and emphasizes participation over victory, since the ultimate goal of CACAO is to motivate and guide students through a critical learning activity. Even though participating in (and winning) the full competition may be an attractive goal for students and instructors alike, implementing a full CACAO competition as a unit within a pre-established lab course requires a significant restructuring effort to free up the necessary time for a productive student experience. Different schools have tried different models and have had different experiences implementing CACAO within their courses, but most have found it useful to engage in CACAO in a gradual manner, with instructors learning the ropes with their students in the first year and adjusting both timelines and grading schemas to their purposes. The following are some implementation examples from previous CACAO participants:

Single Shot CACAO

If you are not using a group-based learning approach in your class, CACAO might provide an ideal opportunity to test the waters. You can assign groups in class, have them participate in CACAO and then evaluate students individually, or add in an extra-credit factor for the team score. Aim for a short participation period (e.g. 2 weeks), low numerical expectations for annotations and challenges (e.g. 1) and allow for at least one previous week of training.

Vanilla CACAO

If you are using team-based learning in your class (or any other group-based teaching methodology), just use the same teams for CACAO and then integrate the CACAO score into your team grading scheme. Aim for a moderate CACAO participation (e.g. 2-3 weeks), allowing for at least one previous week of training, define low numeric expectations for annotations and challenges (e.g. 2-3) and define grading based on *accepted* annotations/challenges, as a percentage component of the course team grade.

Extra Sugar CACAO

If you are not sure if CACAO will work for your course, you can use it as an opt-in (grade enhancer) solo activity. Register your whole class as a single team and track, at the end of the semester, which students have participated and what they have accomplished to give partial or total extra credit to each student. The only downside of this approach is that you still have to provide some background and training and do so in a more personalized manner, but part of the extra credit may involve independent student work on the CACAO training materials.

Pure Aroma CACAO

If you decide that Gene Ontology annotation is your thing, you can set up a 1-2 credit course dedicated exclusively to Gene Ontology annotation, or use CACAO to complement a broader hands-on course on biological databases and ontologies which can include the creation of Wikipedia pages, contributions to developing ontologies and other units.

Unavailable Evidence Codes and GO Terms

Unavailable Evidence Codes

CACAO does not allow annotations using some evidence codes. A list of accepted evidence codes is available on the CACAO website. Among experimental codes, CACAO does not accept IPI (Inferred from Physical Interaction) and IEP (Inferred from Expression Pattern). These codes are not accepted to prevent students from submitting many annotations resulting from a single manuscript reporting a high-throughput experiment (e.g. a genome-wide transcriptome analysis for IEP). If you and other schools are participating in a topic-focused CACAO and the literature in that field makes extensive usage of low-throughput IEP or IPI techniques, you can contact the CACAO staff to remove this restriction. CACAO does not accept evidence codes based on traceable (TAS) or untraceable author statements (NAS), nor inferences made by curators (IC) based on their knowledge. These evidence codes are nowadays rarely used in the Gene Ontology and they are reserved to professional biocurators.

Unavailable GO Terms

When parsing primary literature, students are likely to encounter experiments reporting functions, processes or cellular/extracellular locations not currently defined (or not specialized enough) in the Gene Ontology. Students are welcome to submit new term requests (NTRs) to the Gene Ontology. Accepted NTRs will earn students points in CACAO and expose them to the generative process of the Gene Ontology (which uses GitHub) and the community behind it. Details on the NTR process are available on the CACAO website (**Supplementary material 2**).

Transfer Annotations

Transfer annotations were not a part of the original CACAO and were introduced in the context of SEA-PHAGES annotations in 2015. Transfer annotations can be used to annotate gene products in genomes with scant (if any) published experimental work, by leveraging the experimental work done in related organisms to infer different aspects of a gene product via computational means. This may be appealing to both instructors and students, but requires additional training and can substantially complicate the annotation process. Transfer annotations can be approached in two different ways. Students may start with the genome of a species of interest and use BLAST and other tools to identify homologous genes in other species, then check whether such homologs contain annotations or have associated publications reporting experimental work that can be leveraged for annotation. Given the large number of results returned by BLAST and other computational tools, this is often time-consuming and can be frustrating, as it involves checking manually many search results. If the instructor has prior knowledge on relevant model organisms that are evolutionary close to the species they are targeting in CACAO (e.g. humans for chimpanzees), they can instruct students to search for publications likely to contain experimental data in the model organism, use BLAST or similar tools to determine homology in the target organism and, if it exists, transfer existing annotations (or make new ones) from the model organism into the target organism. A lengthier explanation of what transfer annotations are and how they are done, together with a walkthrough a regular and a transfer annotation is available in the CACAO website (**Supplementary material 9**).

Troubleshooting the Unit

Given its broad scope and diverse implementation modes, the first implementation of this unit in a laboratory course is likely to have some glitches. Most of implementation problems arise from inadequate comprehension of the annotation problem and excessive degrees of freedom in the implementation. To address the former, it is recommended that the instructor walk the students through an annotation, exemplifying the approach to critical reading in the context of Gene Ontology annotation. Several examples of articles and their annotations are available on the CACAO website (http://gowiki.tamu.edu/wiki/index.php/helpful_handouts_for_students), but instructors willing to target a specific topic/organism should work on a manuscript from that subfield and use it to illustrate at least one annotation to students. To address the latter, it is recommended that the first iteration of CACAO be performed using a set of articles preselected by the instructor, and that the scope and format of the unit be opened up only after the instructor feels comfortable in guiding the students through the issues posed by articles lacking valid annotations, using non-allowed evidence or strains not mapping to UniProt protein identifiers. The last section of the student handout details these cases and their solutions. In practice, one should consider that the goal of a CACAO unit is to develop critical reading skills while performing a valuable service to the scientific community. Having students annotate a restricted set of preselected articles on a given topic/organism will not negatively impact any of these goals, but opening up the scope of the task without providing appropriate guidance to students certainly may.

Cited References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25:25–29. doi:10.1038/75556.

Balakrishnan R, Harris MA, Huntley R, Van Auken K, Cherry JM. 2013. A guide to best practices for Gene Ontology (GO) manual annotation. Database J. Biol. Databases Curation 2013:bat054. doi:10.1093/database/bat054.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. 32:D262–D266. doi:10.1093/nar/gkh021.

Caruso SM, Sandoz J, Kelsey J. 2009. Non-STEM undergraduates become enthusiastic phage-hunters. CBE Life Sci. Educ. 8:278–282. doi:10.1187/cbe.09-07-0052.

Jordan TC, Burnett SH, Carson S, Caruso SM, Clase K, DeJong RJ, Dennehy JJ, Denver DR, Dunbar D, Elgin SCR, Findley, AM, Gissendanner, CR, Golebiewska, UP, Guild, N, Hartzog, GA, Grillo, WH, Hollowell, GP, Hughes, LE, Johnson, A, King, RA, Lewis, LO, Li, W, Rosenzweig, F, Rubin, MR, Saha, MS, Sandoz, J, Shaffer, CD, Taylor, B, Temple, L, Vazquez, E, Ware, VC, Barker, LP, Bradley, KW, Jacobs-Sera, D, Pope, WH, Russell, DA, Cresawn, SG, Lopatto, D, Bailey, CP and Hatfull, GF. 2014. A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. mBio 5:e01051-01013. doi:10.1128/mBio.01051-13.

Renfro DP, McIntosh BK, Venkatraman A, Siegele DA, Hu JC. 2012. GONUTS: the Gene Ontology Normal Usage Tracking System. Nucleic Acids Res. 40:D1262-1269. doi:10.1093/nar/gkr907.

Acknowledgments

The authors wish to thank Suzanne Aleksander and previous members of the CACAO team for their excellent technical support and dedication to the project. The authors would also like to express their acknowledgement to the SEA-PHAGES Program for their support, and to Allison A. Johnson, from Virginia Commonwealth University, and other SEA-PHAGES CACAO participants for their support in implementing this initiative. The authors would also like to thank the GO Consortium for help in prioritizing student ontology term requests and the EBI for help with importing CACAO annotations to the GOA corpus, as well as the National Institutes of Health for past support for development of CACAO and GONUTS. Finally, the authors would like to thank the participants in the ABLE 2017 workshop for many insightful comments and feedback to improve both the unit and this manuscript.

About the Authors

Ivan Erill received his B.S. and PhD in Computer Science from the Universitat Autònoma de Barcelona. He is currently an Associate Professor at the University of Maryland Baltimore County, where he runs a research laboratory on bacterial comparative genomics and applies active learning paradigms to the teaching of bioinformatics and computational biology.

Steven Caruso completed a B.S. in Biological Sciences and Psychology and a Ph.D. in Biological Sciences from the University of Maryland Baltimore County, where he currently teaches genetics, microbial, and molecular biology as a Senior Lecturer.

Jim Hu received his BS in Biology from Stanford University and an MS and PhD in Molecular Biology from the University of Wisconsin-Madison. After postdoctoral work at MIT, he joined the faculty at Texas A&M University where he is currently a Professor of Biochemistry and Biophysics. His research group currently works on annotation and ontology development for microbial phenotypes and integration of biocuration with innovative teaching. He teaches courses in genomics, critical reading with GO annotation, and effective research presentations.

List of Supplementary Material

Supplementary material 1 – Hints on how to identify a UniProt accession number for a protein [https://gowiki.tamu.edu/wiki/images/1/1e/Uniprot_hints-2.pdf].

Supplementary material 2 – Instructions for submitting a new term request to the Gene Ontology [https://gowiki.tamu.edu/wiki/images/1/10/NewTermRequest.pdf].

Supplementary material 3 – Evidence code decision tree [https://gowiki.tamu.edu/wiki/images/3/32/CACAO_deciontree.pdf].

Supplementary material 4 – Tips on experiments and their associated evidence codes [https://gowiki.tamu.edu/wiki/images/e/ee/Experiments_and_their_evidence_codes.pdf].

Supplementary material 5 – Hints on how the identify the PubMed ID for a scientific manuscript [https://gowiki.tamu.edu/wiki/images/4/48/PUBMED_hints-2.pdf].

Supplementary material 6 – Practice manuscripts for GO annotations [https://gowiki.tamu.edu/wiki/index.php/helpful_handouts_for_students].

Supplementary material 7 – Step-by-step walkthrough of an annotation in CACAO [https://gowiki.tamu.edu/wiki/images/2/2a/Annotation_path.pdf].

Supplementary material 8 – CACAO instructor's manual [https://gowiki.tamu.edu/wiki/index.php/File:CACAOInstructorsManual.pdf].

Supplementary material 9 – Step-by-step walkthrough a regular and a "transfer" Gene Ontology annotation in the context of CACAO [https://gowiki.tamu.edu/wiki/index.php/File:Step_by_step_transfer_annotations_in_CACAO.pdf].

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/.**

Papers published in *Tested Studies for Laboratory Teaching: Peer-Reviewed Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

Erill I, Caruso SM, Hu JC. 2018. Gamifying Critical Reading through a Genome Annotation Intercollegiate Competition. Article 6 In: McMahon K, editor. Tested studies for laboratory teaching. Volume 39. Proceedings of the 39th Conference of the Association for Biology Laboratory Education (ABLE). **http://www.ableweb.org/volumes/vol-39/?art=6**