UniProt Hints

http://www.uniprot.org/

Part I: Searching UniProt

(Also look at http://www.uniprot.org/help/text-search)

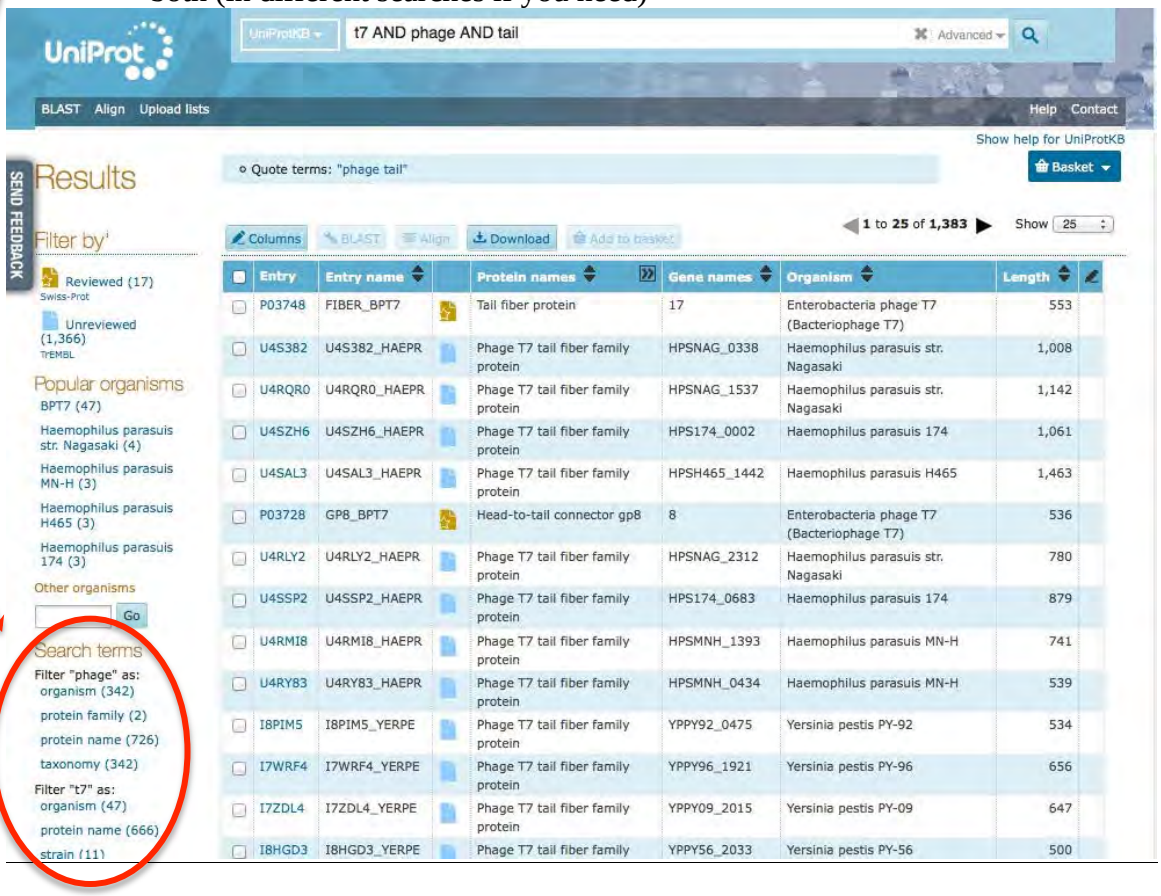
BASIC SEARCHES:

- Need gene or protein name
- Need organism name
  - → coli tolB
  - → tyrosine kinase human

USUALLY SEARCHES WILL NEED MODIFICATIONS:

Use the *Restrictions* at the top of results to refine searches you've done

- If your original search consisted of multiple terms, you will be given the option to apply that term to a specific field
  - Apply "cucumber", "coli", "Bacillus", etc. to the "organism" field
  - Apply your protein/gene name to the appropriate field- if you're not sure, try both (in different searches if you need)



ADVANCED SEARCHES (NOT ESSENTIAL IF YOU MODIFY AS SHOWN ABOVE):

- You can streamline your searches by entering the restrictions directly into the query bar. The commonly used options include protein family (FAMILY:), gene name (GENE:), protein name (NAME:), organism (ORGANISM:), strain (STRAIN:).
- AND/OR/NOT can also be represented by &&, ||, and -/! respectively:
  - → cancer AND gene OR genetic NOT human
  - → cancer && gene || genetic !human
  - → ORGANISM:phage AND NAME:"tail protein"

Part II: How do we select a protein from the search results?

Typically there will be several, to hundreds or thousands, of results. To select the appropriate accession (the "Entry" number in the far left column):

- Make sure the Species/Organism matches what is discussed in the paper
- Make sure the Strain matches what is discussed in the paper
  - "Escherichia coli K12" IS VERY DIFFERENT FROM "Escherichia coli O8"
- Consider the lengths of the potential proteins. Sometimes this info is presented in the paper
- Compare the sequences, if the information is given in the paper
- Sometimes the UniProt, or occasionally a NCBI reference is given in the paper
- If you still have several similar possibilities, **ask for help if you need it**!

Part III: What do the stars (reviewed/unreviewed) mean?

UniProt IDs are assigned to many, many sequences submitted by many, many researchers. Some are actually proteins (which is what they want), but some are fragments, repeats of previously known sequences, etc. The entries that have NOT been professionally reviewed have grey stars- this kind of means "this MIGHT be a protein, but we haven't checked it yet". If you don't have any gold-starred entries, it is perfectly acceptable to use these grey ones- especially with rarer, non-model organisms including many phage.

The gold stars mean the professional reviewers (Swiss-Prot) have verified the entry. If you have a gold-starred option, it will GENERALLY be the correct one; if in doubt, it's usually safe to choose these. If you have multiple gold-starred options that seem to be the same length, same organism & strain, same exact gene, ask for help- it probably won't matter which one you choose though.

Supplementary Material 2

New Term Request (NTR) Walkthrough

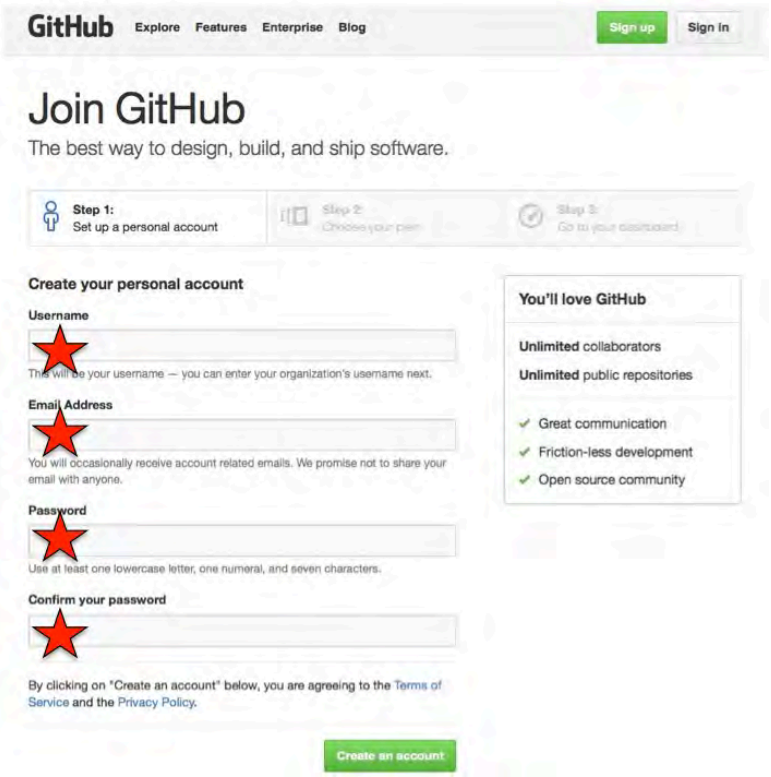
You are not required to make New Term Requests. NTRs do, however, usually count for additional points, so many students are interested in this process. Some terms need modification of their definitions or their placement in the ontology. All requests are sent via message boards to the Gene Ontology staff who review the request and will happily make new terms or modify existing ones if the request is valid. Expect this process to take at least a couple of days to possibly several weeks.
The board is on GitHub.

You can view the entire message board here: https://github.com/geneontology/go-ontology/issues

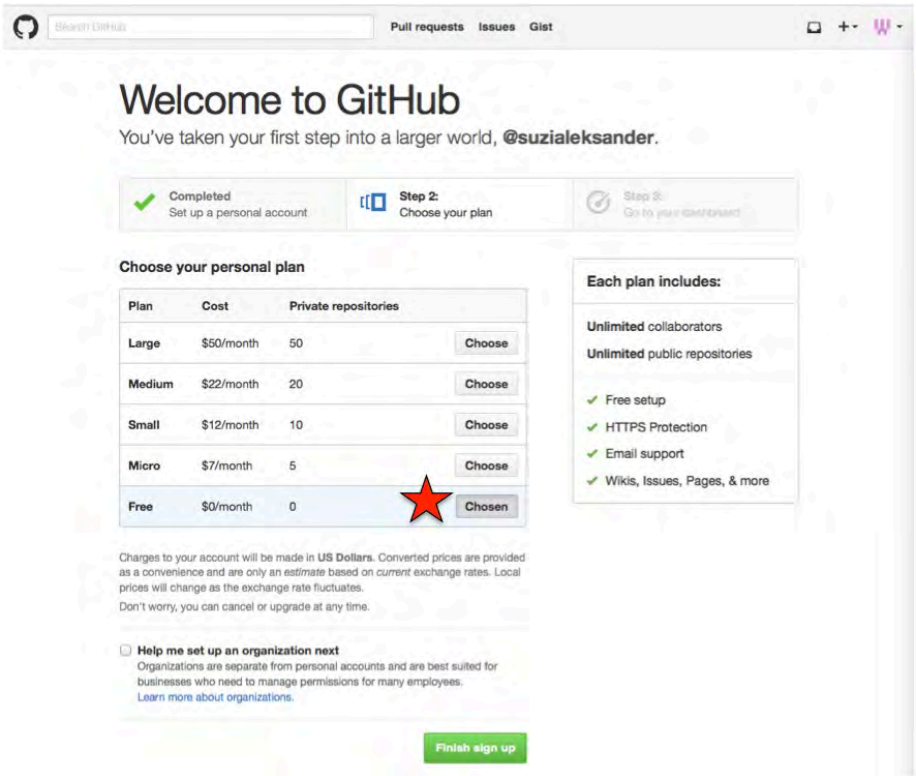
Part I: Logging into GitHub

*Note: TAMU students already have github.tamu.edu accounts. You may have used this already for another class (https://github.tamu.edu/login), but this is **NOT** the same account you will need for this process*

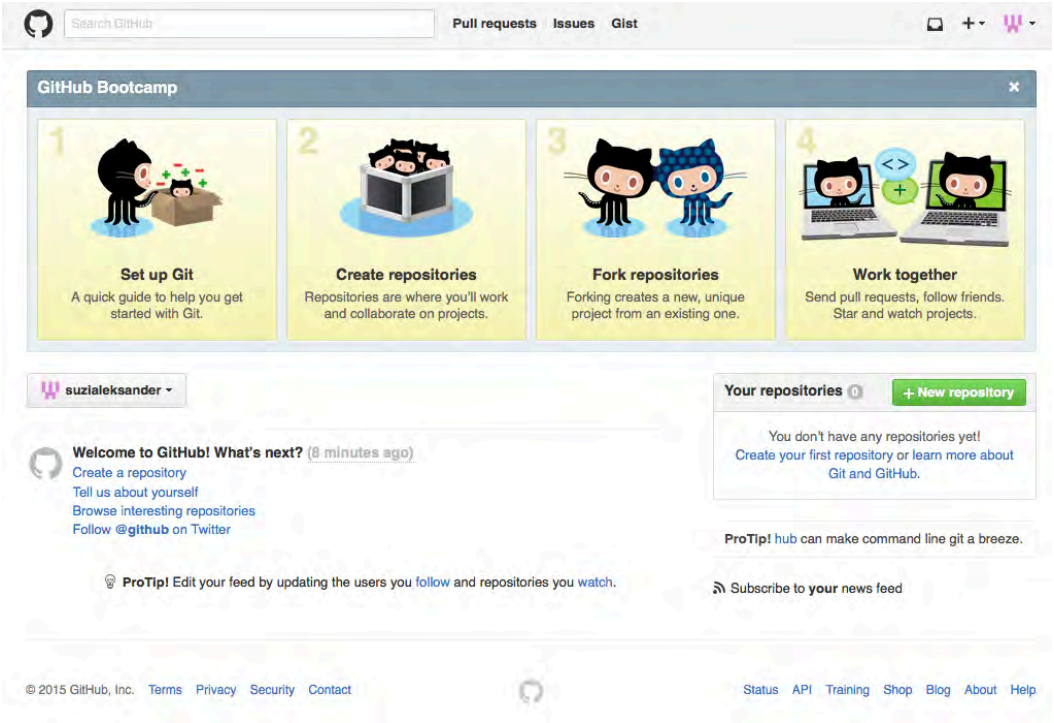
1. Create a GitHub account
    a. Go to https://github.com/join
    All fields appear to be required. Remember to keep your username professional. You CAN use the same username as for your github.tamu.edu, but this is not required.



2. Choose the FREE plan
    a. This will probably already be chosen, but ensure you have selected the $0/month plan.



3. If everything went right, you'll find yourself on some version of a page with a lot of cats. We don't know why, but GitHub seems to like cats. Or they're cats wearing bad octopus costumes, we're not sure. Sorry.



4. Remember to verify your email you used to sign up with.

Part II: Make the Term Request

1. Go back to https://github.com/geneontology/go-ontology/issues and select the green "New Issue" button.

2. Fill out the following information:

Title: "***NTR***: desired GO term name here" (You won't be able to provide a GO ID). If you only need to modify an existing term, use ***OTR*** for Other Term Request.

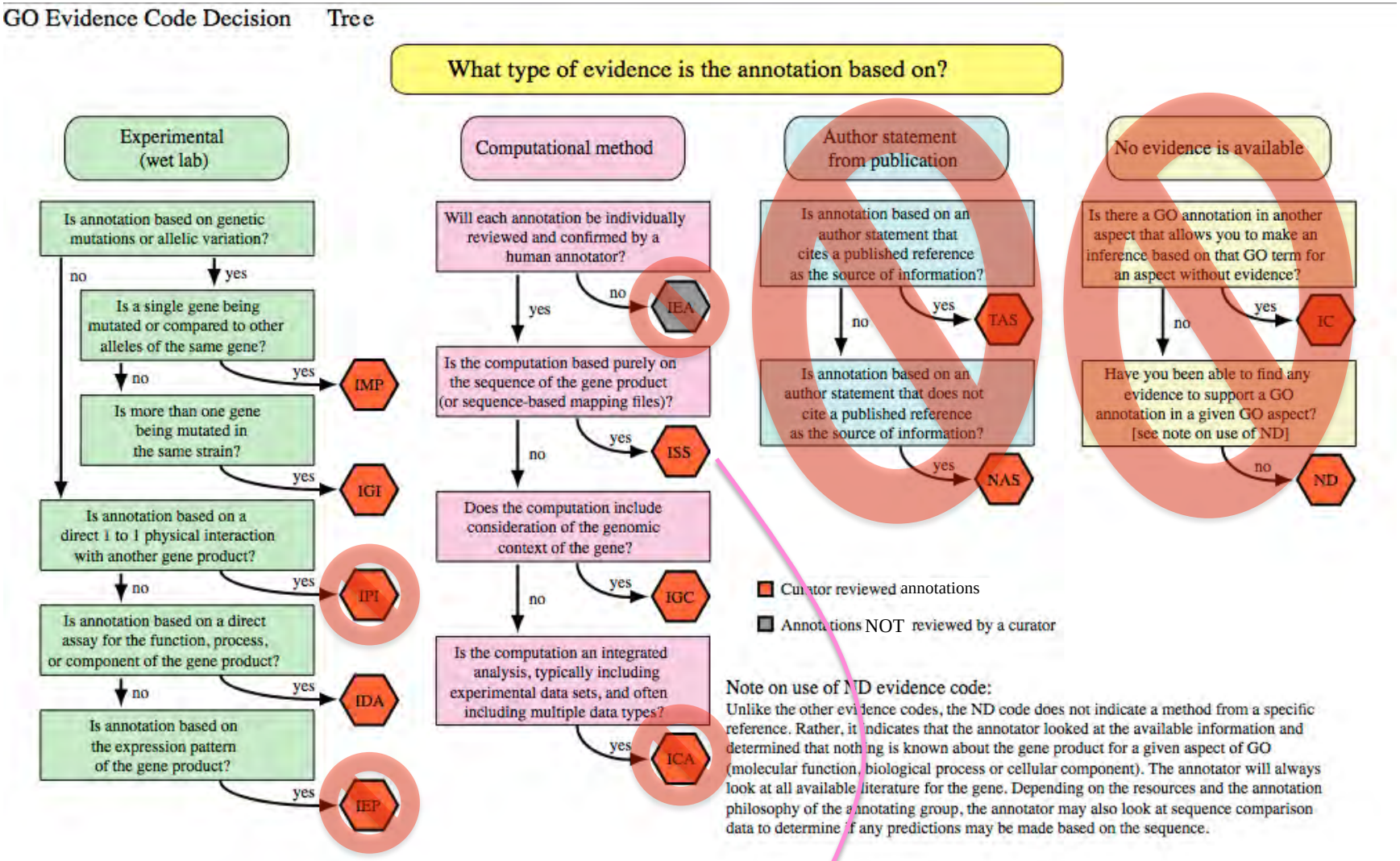
In the description box, include at least the following information, being as specific as possible:

- **Potential GO term name** (again)
- **Definition**- This is essential. You probably identified the need for this term by finding a term that ALMOST worked-feel free to copy & paste that definition, making slight changes as you need
- **Aspect**- Is it a Cellular Component, Biological Process, or Molecular Function? (This will probably be determined by the terms it is related to)
- **Relationships**- Where does this term go in the ontology? Indicate the parent term & parent term GO ID, possibly examples of "sibling" terms, and examples of GO IDs it would be the parent of (not always applicable)
- **References**-Give the PMID of the paper that supports this term. You can also add more info, such as references to specific parts of a paper or multiple PMIDs.
- **Synonyms**- Not required, but helps the Ontology make their decision. Look at the parent term or CHEBI for possible synonyms.

For more examples, look at the message board entries named "NTR" or "OTR"

- https://github.com/geneontology/go-ontology/issues/11541
- https://github.com/geneontology/go-ontology/issues/11280
- https://github.com/geneontology/go-ontology/issues/11966
- https://github.com/geneontology/go-ontology/issues/11346

Supplementary Material 3



ALLOWED CODES FOR ALL CACAO STUDENTS:

1. **IDA: Inferred from Direct Assay**
2. **IMP: Inferred from Mutant Phenotype**
3. **IGI: Inferred from Genetic Interaction** - requires with/from field to be filled in
4. **ISO: Inferred from Sequence Orthology** - requires with/from field to be filled in
5. **ISA: Inferred from Sequence Alignment** - requires with/from field to be filled in
6. **ISM: Inferred from Sequence Model** - requires with/from field to be filled in
7. **IGC: Inferred from Genomic Context**

Use one of these three codes (ISO, ISA, ISM) if the Decision Tree points you to ISS

Supplementary Material 4

Typical Experiments used for Annotating to a Certain Evidence Code

http://www.geneontology.org/GO.evidence.shtml

Inferred from Direct Assay (IDA):

-Enzyme assays using purified components
-*In vitro* reconstitution of a functionally active complex
-Immuno-fluorescence microscopy
-Cell fractionation
-Physical interaction binding assays (when the exact binding partner is unknown)

Inferred from Physical Interaction (IPI)

-2 Hybrid interactions
- Co-purifications
-Co-immunoprecipitations
-Ion/ Metal/Protein Binding Assays (when the exact binding partner is known)

Inferred from Mutant Phenotype (IMP)

-Mutation of a gene that results in a partial or complete impairment of resultant protein
-Any treatment that disturbs the expression or normal functioning of the gene, such as:
    -Overexpression of wild-type or mutant gene
    -RNAi, anti-sense RNAs, antibody depletion
    -Using inhibitors, blockers, modifiers, changes in pH or ionic strength

Inferred from Genetic Interaction (IGI)

-"Traditional" genetic interactions
    -Suppressors
    -Synthetic lethals
-Rescue experiments

Inferred from Expression Pattern (IEP)

-Transcript levels or timing experiments
    -Northern blots
    -Microarray data
-Protein levels
    -Western blots

Evidence Codes (ECs) that require information in the "With/From" field

IPI
IGI
ISS
ISO
ISA

If a protein is being curated to a specific GO term based on a physical or genetic interaction, sequence similarity, or sequence alignment, additional information about the reference protein must be given.

Supplementary Material 5

PUBMED Hints

http://pubmed.gov OR http://pubmed.org

Part I: Searching PubMed

- use the terms "AND/OR/NOT"
- searches can look like: cancer AND gene OR genetic NOT human
  tyrosine kinase AND coli AND characterization

Part II: Good terms to help you find useful papers

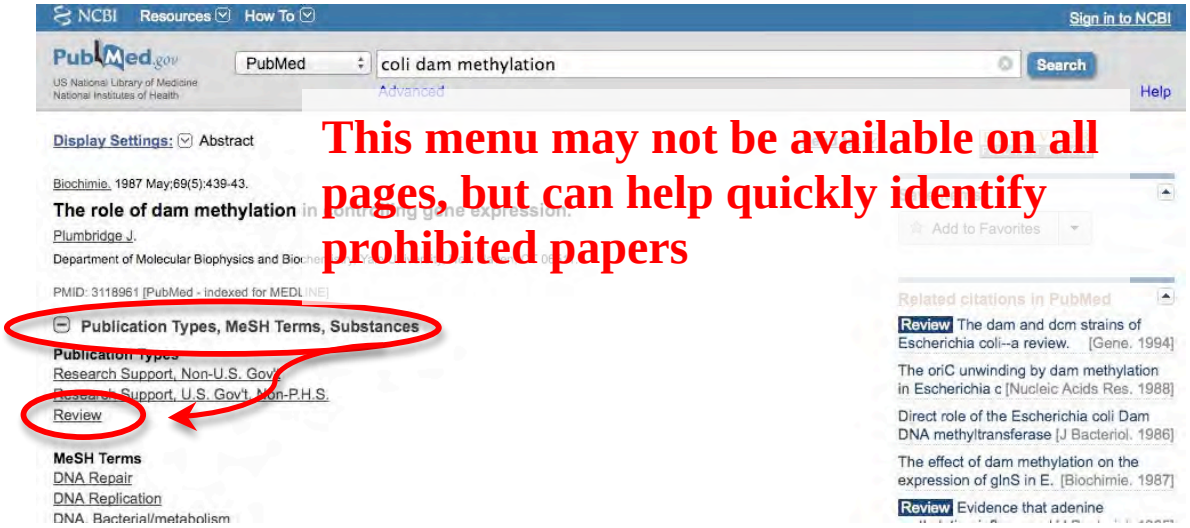
- Specific gene or protein name ("DnaK"), or even just "gene" or "protein"
- Organism name (coli, drosophila, cucumber)
- Words such as:
  - characterization
  - isolation
  - purification
  - localization
- Specific methods: "mass spectrometry"
- GO term from the annotation table on a gene page in GONUTS: "sphingolipid signaling pathway" (don't use the GO ID, eg. GO:0003376)

Part III: How do we select a paper from the search results?

** You will get better at this with practice

- Read the title carefully
- Read the abstract & look for words like "purified" or "mutants" or "enzymatic activity" or " suppressors" or a GO term and so forth
- Download the paper if it looks promising:

- Below the abstract, there will be a couple of expandable links. There may be one that says "Publication Types, MeSH Terms, Substances" or similar- Expand this. If you see the "Publication Type" is a "review", remember you cannot use this for an annotation but you CAN look at the paper's references:



Supplementary Material 6

Introductory & Practice Material

Intro/Practice Material

Training #1 Paper:

- PMID:8227000 **Escherichia coli topoisomerase IV. Purification, characterization, subunit structure, and subunit interactions**

Training #2 Paper:

Practice to identify where in the paper the info came from:

- PMID:21586683 There are several annotations, as seen on TAIR:LQY1, that have been made using this paper. Note that the annotators used C, P, and F terms for the same paper.
- PMID:9288922, see the annots to SGD:AAT2

Training #3 Paper:

Example of a Bad-for-CACAO paper on Disease:

- PMID:22882676

From this paper's introduction, also in the abstract: *Now, in our present study, we have investigated whether ZA induced growth inhibition and apoptosis in PC-3 and DU-145 may be enhanced by the combination with CA or OA, through inhibition of serine/threonine phosphatases in prostate cancer cells*. This investigates the effects of zoledronic acid on specific processes, not the native function of ZA (which is not a protein, but a drug). It would be better to open it and look at the references to find where the serine/threonine phosphatases are ORIGINALLY characterized

Supplementary Material 7

Adding Annotations

*** *Don't forget to* ***log in***. You won't be able to create pages, edit, etc. otherwise. ***

CACAO students have some restrictions that professional curators don't have. There are Evidence Codes, Reference types, and other options that will result in CACAO annotations being rejected even if they're technically correct. Unless you're specifically told otherwise by a judge, the following apply (other restrictions may also be in place for your session):

Don't forget – No annotations using:

- ***binding*** terms (ie. GO:0005515 protein binding, GO:0005524 ATP binding, GO:0008144 drug binding, etc.)
- Forbidden evidence codes:
  - **TAS**: Traceable Author Statement
  - **NAS**:Non-traceable Author Statement
  - **IC**:Inferred by Curator
  - **ND**:No Biological Data Available

To find a UniProt Accession:

1. Go to http://www.uniprot.org
2. In the query box, enter:
   a. a protein name with organism (ie p53 human)
   b. search term (ie cancer)
   c. gene name with organism (**TUBA Prunus dulcis**)
   d. some other relevant term- you can use advanced search options including AND, OR, etc.
3. Click on "*Search*".
4. Find the protein you are looking for & copy the accession. Be careful, *it may not be the first item listed*!

To find a GO term:

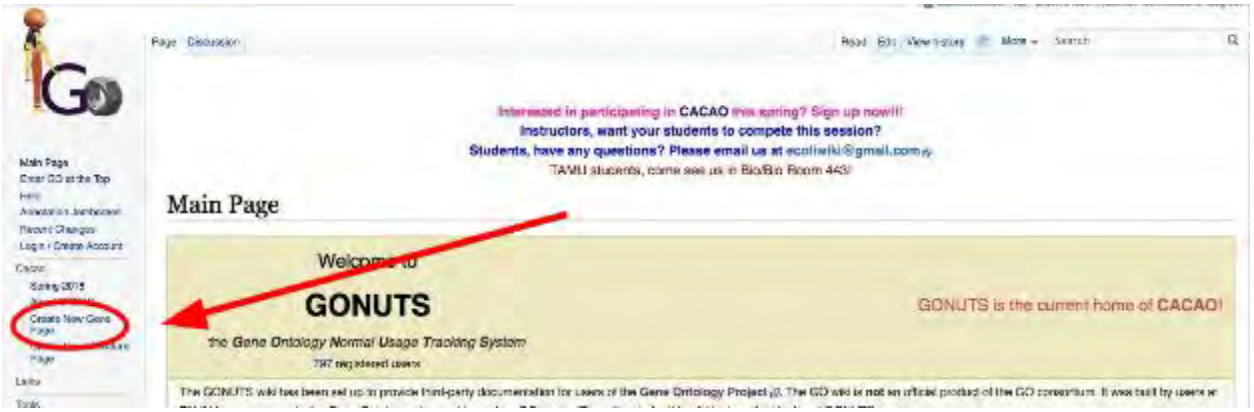
1. Use
    a. QuickGO (http://www.ebi.ac.uk/QuickGO/)
    b. Amigo (http://amigo2.geneontology.org/amigo)
    c. or search on GONUTS (http://gowiki.tamu.edu/wiki/index.php/Main_Page)
2. Search for a possible term
3. Explore the results, including parent and child terms, to ensure the chosen term is applicable and the most specific available term.

Remember, the same paper/protein can have multiple GO terms, as long as they're not directly related.

To make a gene product page:

1. Once you have the UniProt Accession (ie. P04637), go to GONUTS (http://gowiki.tamu.edu).

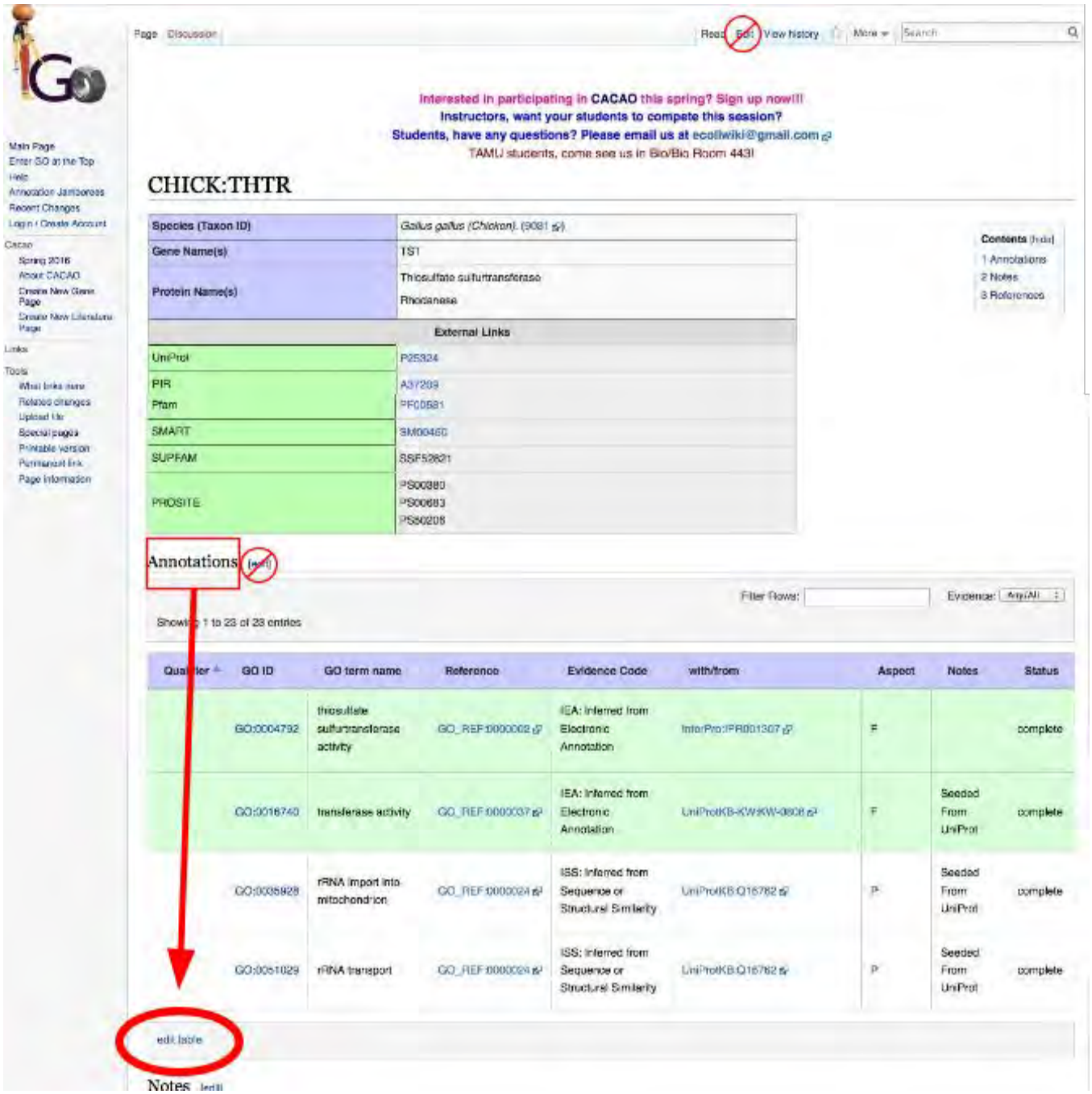
2. Click on the "Create New Gene Page" link in the menu on the left of the page.



3. Paste the UniProt Accession into the empty box on the GoPageMaker page and click "Create". This will make the gene product page in GONUTS automatically, but this may take up to 60 seconds to fetch the info and create the page.

To add an annotation to your gene product:

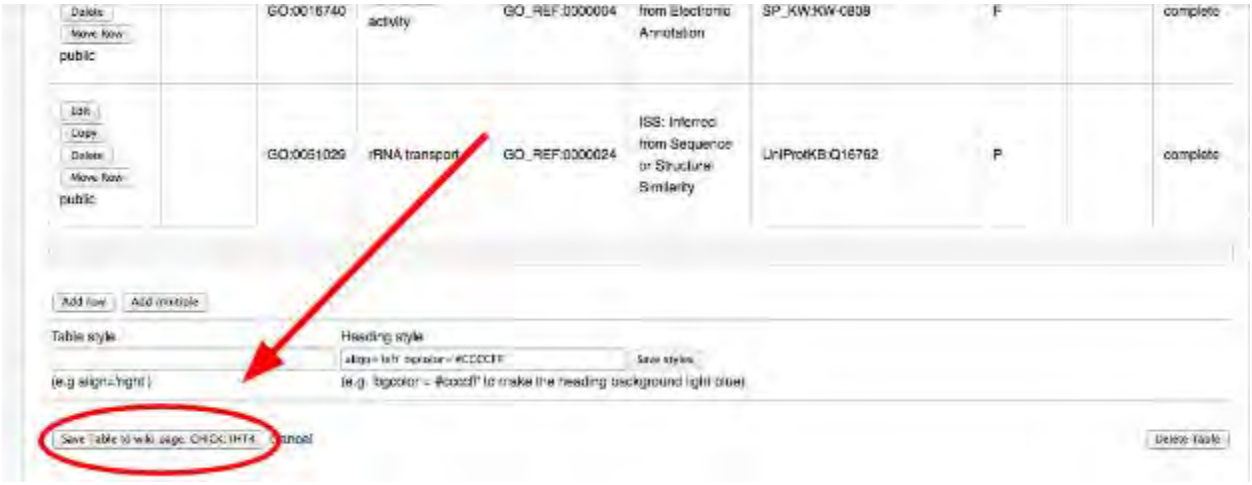
1. Once on the correct *gene page* (**not** the PMID page), scroll down to the bottom of the “Annotations” table. Click on “edit table”. This will make all of the rows available to edit.



2. Scroll to the bottom of the table again and click on “Add row”.

3. You must:
    a. Fill in the form,
    b. ***refresh*** to make sure all the autofill fields are correct, then
    c. ***SAVE*** your annotation row to the table.

4. Save again- this second button saves the table to the page.



IT IS ESSENTIAL TO SAVE TWICE

Failure to save twice will result in the loss of your annotation

**There are *required* fields you must add for every annotation, and some optional fields you might add in rare instances:

Required:

- GO ID (ie GO:0005737)
- Reference (ie. PMID: 1111111)
- Evidence Code (ie. IDA)
- Notes (*specific figure* & method/result in this paper)

Optional:

For some evidence codes, you may have to fill in another field called "with/from".
For some terms, you may add a "Qualifier".

Examples for Optional Fields:

- If you have a paper that shows that a certain phenotype only shows up when you delete 2 genes (gene aaaA and gene zzzZ), you will use the IGI evidence code, which requires the "with/from" field to be filled in. Edit the Annotations table on the gene page for aaaA and add a row so you can put in your annotation. Enter your appropriate GO ID, Reference (PMID:11111111), Evidence Code is IGI (Inferred from Genetic Interaction), which will bring up an additional field. This additional field is the "with/from" field and you must put in an accession for zzzZ (ie UniProt Accession) for this annotation to be considered complete. Note that

you can add the same annotation to zzzZ's annotation table with the accession for aaaA in the "with/from" field.

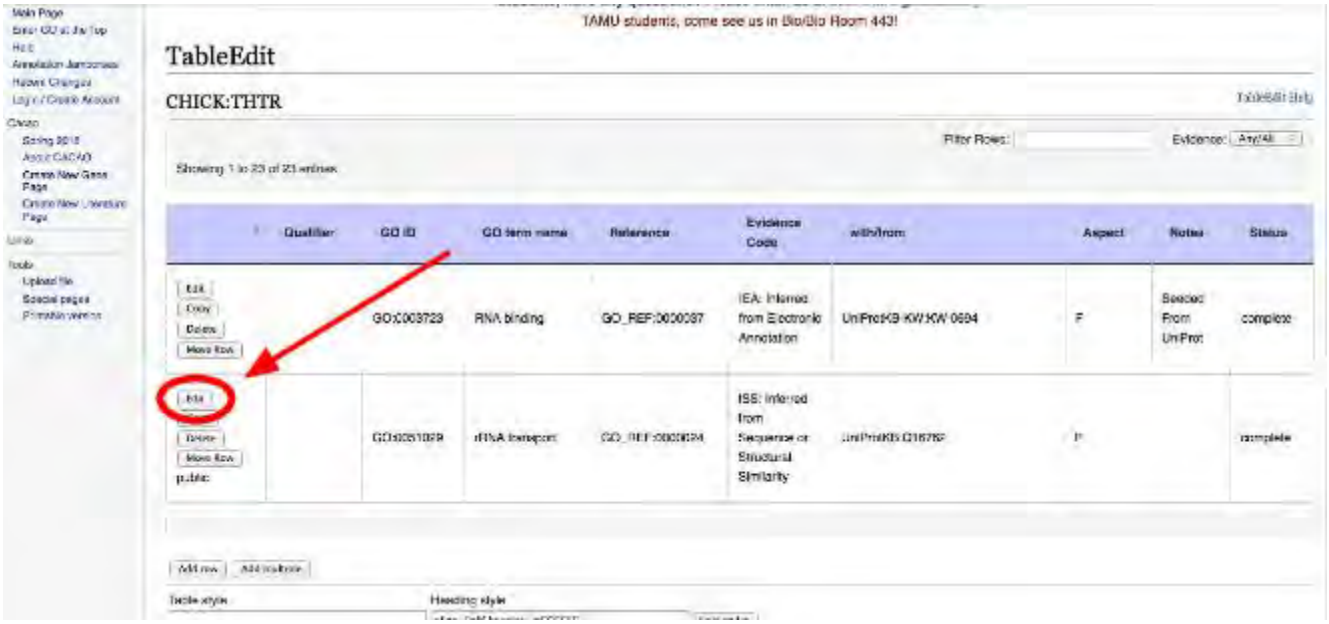
- You have a paper that shows that 2 or more (a complex) of gene products are required to get a certain activity and that the individual gene products alone are not sufficient to show this activity. In this case, you must add a qualifier of "contributes_to". This is ONLY relevant to molecular function terms.

- You have a paper that says "in contrast to previous reports, this protein was not localized to the inner membrane upon subcellular fractionation". In this case, you are allowed to put a "NOT" qualifier in front of an annotation to inner membrane (so long as the experimental result shows this). This qualifier is used when a previously reported result is shown to be incorrect (ie. NOT in inner membrane as shown before), not when there is a negative result reported (ie. Protein A isn't in the inner membrane and has never been reported to be).

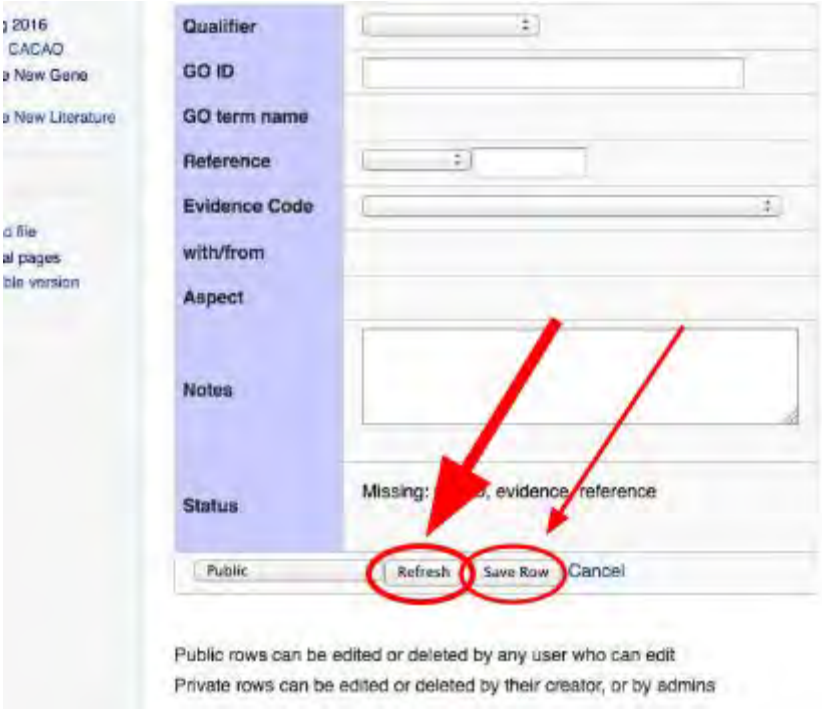
- For more information, see: http://geneontology.org/page/go-annotation-conventions

To edit an annotation you already made:

1. As with making an annotation described above, scroll down to the bottom of the "Annotations" table. Click on "edit table".
2. "Edit" the row you previously made.
   Note: you can only edit your own annotations, and may only do so during *Annotation* or *Open* rounds. All other changes must be made in the form of "Challenges", made during *Challenge* rounds.



3. Make any desired changes.
4. ***Refresh*** **to make sure all the autofill fields update**



5. *Save* the row to the table
6. *Save* the table to the page, exactly as you would if it were a new annotation.

IT IS ESSENTIAL TO SAVE TWICE

Failure to save twice will result in the loss of your changes

Supplementary Material 8

CACAO INSTRUCTOR'S MANUAL

PART I: Setting Up a CACAO Competition
PART II: Making GO annotations
Part III: Challenges
Part IV: Competition Rules
Part V: Assessment
Part VI: Evaluation for Grading
Part VII: The code behind the scoreboard - contained in another GoogleDoc

Table of Contents

Introduction

What is CACAO?

The Community Assessment of Community Annotation with Ontologies (CACAO) is a project to couple annotation to undergraduate education, by having students read papers, and identify inferences about gene function that can be expressed using the controlled vocabulary of Gene Ontology (GO). This is done as an open activity on the web using the GONUTS (http://gowiki.tamu.edu) website. Students not only perform annotation; they also do peer evaluation of annotations from other teams of students either at the same institution or at multiple institutions.

CACAO was created to address one of the basic challenges in genomics: How to efficiently capture the knowledge in the literature about gene function. Associating information from the literature to specific genes and gene products forms a major part of what genome databases do, and a substantial part of the effort is based on manual (i.e. human) curation of database entries based on reading papers. The cost of manual curation has led many genome database projects to create forms of "community annotation" where external experts are recruited to contribute annotations.

The basic problem with community annotation is finding effective incentives to get community members to participate. CACAO was created based on the idea that community members could get teaching credit for having students participate in this important aspect of genomics. The combination of students working in teams, peer review through challenges, and having faculty mentors were all designed to assure that the quality of the resultant annotations would be suitable for submission to genome resources for public use.

CACAO was first run in the Spring and Fall semesters of 2010. The first involved only TAMU undergrads, while the second included two teams of Masters students from University College London. The first two rounds were well received by the students, who enjoyed the competitive aspects of CACAO as well as the sense that they were making real contributions to important biological databases.

Why join CACAO?

We believe that CACAO teams can be a valuable part of your undergraduate programs. Although CACAO resembles other literature-based courses in teaching higher order reasoning skills and critical analysis, two aspects of CACAO are especially appealing to students:

- The fact that information they extract from the literature will become part of major database resources
- The competition with other teams, either within the class or between schools

Both of these are related to CACAO's use of the web.

While the primary benefits of CACAO are based on the quality of the courses that can incorporate it, we believe there are some secondary benefits:

- We provide the web infrastructure, tech support, and materials. This is all currently free of charge, as CACAO is supported by our grants to generate annotations.
- Interaction between CACAO teams at different institutions will help us refine CACAO curriculum resources
- We believe that CACAO could lead to networking opportunities for participating students that could lead to:
    - Opportunities for your students at other institutions
    - Opportunities to recruit students from other institutions

In addition, in our experience, teaching CACAO has been a lot of fun. We hope you'll join us in making CACAO a widespread activity.

GO (Gene Ontology) & Theory of Annotation

What is an annotation? What do we mean by an annotation for this course?

According to dictionary.com, an annotation is an explanatory note added to a text. In terms of genome biology, an annotation can be added at one of three levels: nucleotide, protein or process (PMID:11433356, L. Stein *Nature Reviews Genetics*, 2001). For CACAO, we aim to increase the number of functional annotations for proteins with each annotation including the primary literature evidence used to make the annotation.

What is the GO?

We use a controlled, structured scientific vocabulary from the Gene Ontology (GO), which allows for standardization of annotations and facilitates comparisons across organisms and systems. Additionally, the format of the GO terms aids in computer based reasoning and simplifies data mining. The GO consists of 3 ontologies for annotating gene products: Biological Process, Molecular Function and Cellular Component (Figure 1).

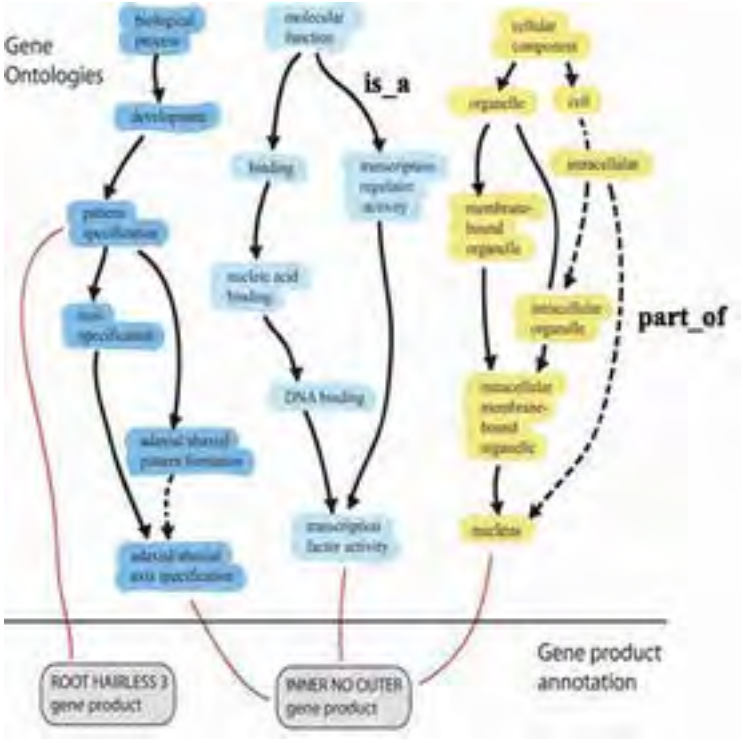
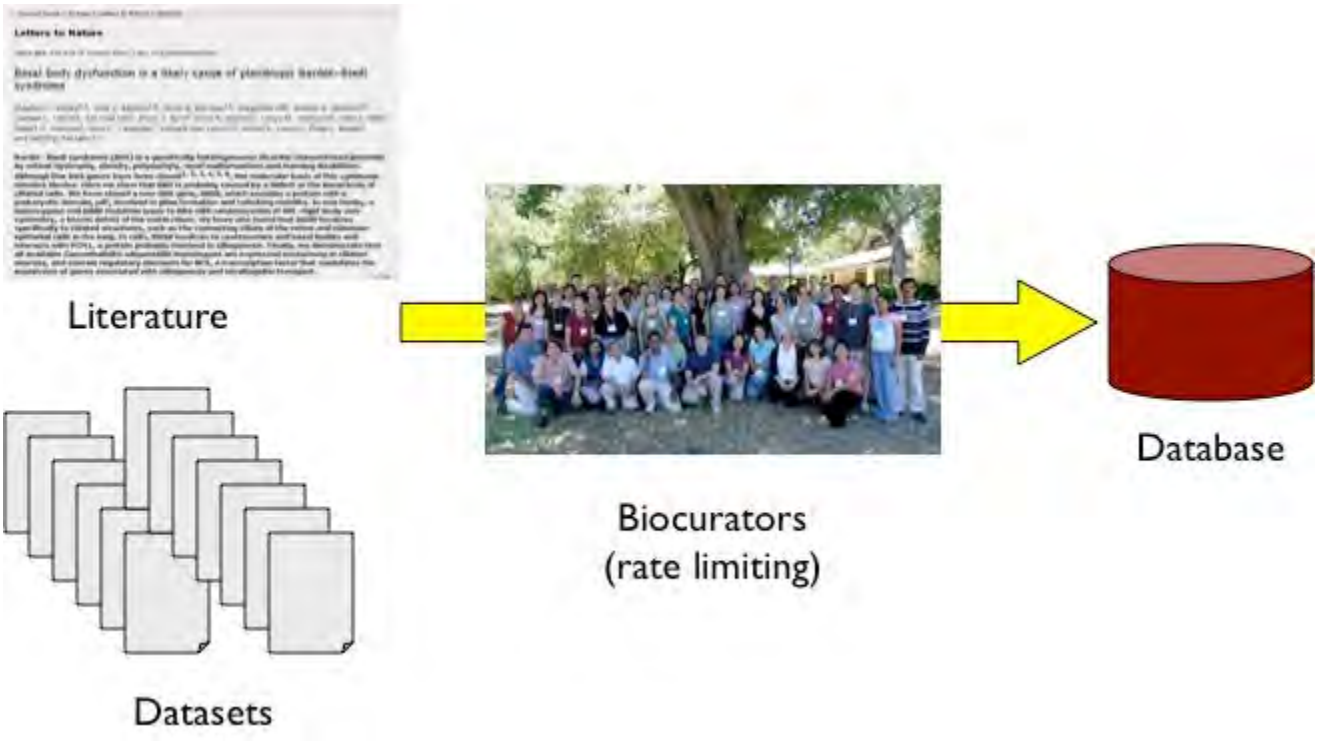


Figure 1: The GO and its ontologies for annotating gene products. Shown are some of the relationships between various terms in GO, including nucleus being a "part_of" intracellular while transcription regulator activity "is_a" molecular function. Taken from the Gene Ontology Consortium.

Who classically makes GO annotations?

Biocurators with PhDs who are usually employed by model organism databases (MODs). They comb relevant literature and datasets, format the information into annotations & enter them into a database.



Why are GO annotations so important?

Functional annotations allow us to:

- infer the function of genes related by common descent, similar expression patterns, phylogenetic profiles, etc.
- understand the capabilities of organisms' genomes
- understand patterns of gene expression in different environments, different tissues, disease states, etc.

Functional annotations done as GO annotations provides a system of assigning function to gene products (proteins, for CACAO) that both humans and computers to compare, contrast, analyze and predict gene product function.

Can you use the same paper to get more than 1 annotation?

YES! You can annotate any paper that hasn't already been annotated. You can annotate multiple BP, MF or CC terms from the same paper.

Can you annotate to less specific terms from the same paper?

NO. You cannot annotate to direct ancestors or descendents of a term. For example, if a paper shows a protein is a kinase and then shows it is more specifically a hexokinase, you can only annotate the most specific term suitable (hexokinase). You may NOT annotate to kinase as well.

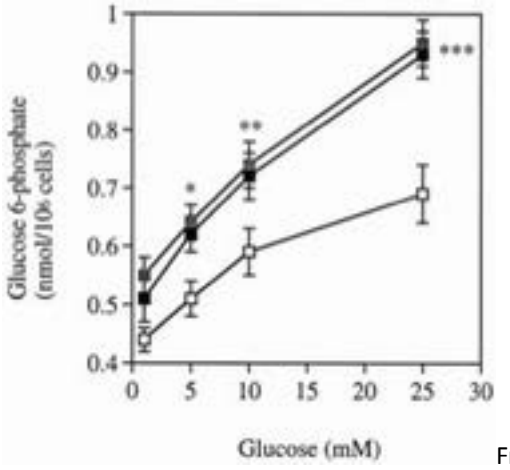
What are the sub-ontologies in GO?

There are 3 sub-ontologies in GO that are filled with terms to describe the attributes of proteins (and RNA products, but we don't use those in CACAO):

1. Molecular Function - activities or jobs of proteins in the cell
2. Biological Process - commonly recognized series of events
3. Cellular Component - where a protein is doing its job in the cell

Molecular Functions

These are the activities or jobs of a protein.



From PMID:9341134

GO:0004347 hexokinase activity

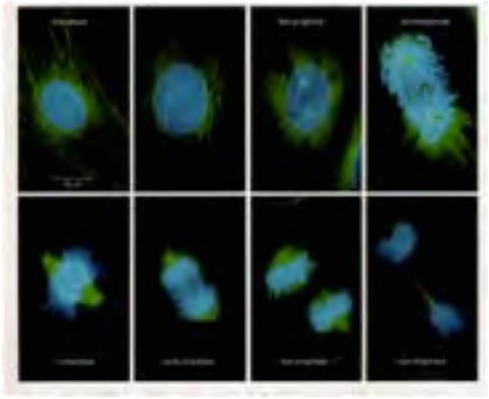
- There are MANY terms, most of which end in "activity".
    - GO:0004535 poly(A)-specific ribonuclease activity
    - GO:0008664 2'-5' RNA ligase activity
    - GO:0032395 MHC class II receptor activity
    - GO:0048018 receptor agonist activity
    - GO:0008080 N-acetyltransferase activity

and many more...

- We do NOT allow CACAO students to annotate to any binding activity terms (i.e. DNA binding, ATP binding, protein binding, etc).

Biological Processes

These are the pathways the protein contributes to or any commonly recognized series of events. These processes are accomplished by coordinated efforts of participating proteins.


From ridge.icu.ac.jp
GO:0051301 cell division

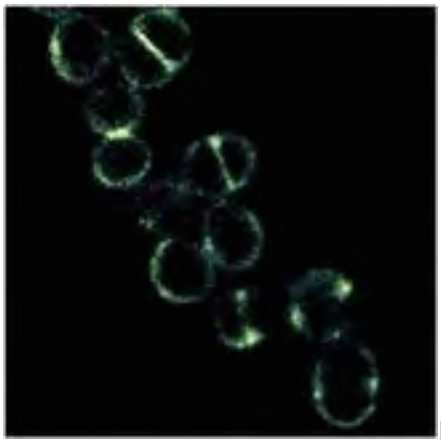
- Terms like:
    - GO:0006351 transcription, DNA dependent
    - GO:0009405 pathogenesis
    - GO:0048481 ovule development
    - GO:0016117 carotenoid biosynthetic process
    - GO:0006508 proteolysis

and so on....

- We do NOT allow CACAO students to annotate to any response to terms (i.e. response to water deprivation, response to heat, response to pH, response to antibiotic, etc)

Cellular Components

These describe where the protein is doing its job in the cell or the complex to which it belongs when doing its job.


From epmm.group.shef.ac.uk
GO:0009274 peptidoglycan-based cell wall

- Terms include:
    - GO:0032807 DNA ligase IV complex
    - GO:0009570 chloroplast stroma
    - GO:0005634 nucleus
    - GO:0005840 ribosome

- GO:0000793 condensed chromosome
- GO:0016020 integral to membrane

Further reading about Gene Ontology:
http://www.geneontology.org/
http://www.geneontology.org/GO.doc.shtml

Learning Objectives

Over-Reaching CACAO Learning Objective:

- students will gain an understanding of where our knowledge of gene products comes from and what techniques scientists used to characterize them.

CACAO Course Learning Objectives:

1. Gain skills in finding information from online resources for scientific knowledge.
2. Improve critical thinking by analyzing scientific papers to identify experimental evidence that supports inferences about gene function.
3. Improve ability to explain and defend critical analyses.

After completing this course, students will be able to:

- describe different levels of Genome Annotation from gene models to functional annotation to systems annotation
- describe the use of ontologies for annotation
- discuss the nature of gene function
- describe different systems used for classification of genes and gene products
- describe automated and manual approaches to annotation
- compare models for biocuration and challenges for each model
- perform literature-based annotation using Gene Ontology (GO)
- evaluate the quality of literature-based annotations done by others in the competition.

PART I: SETTING UP A COMPETITION

CACAO Competitions

The following section describes how the competition is set up on GONUTS, including making accounts and teams, how to use GONUTS to add an annotation, etc.

Most of this can be done by our team at GONUTS/EcoliWiki if you send Brenley an Excel sheet containing:

1. user names,
2. real names,
3. email addresses,
4. teams.

At current, you must set up the dates of the competition through a Texas A&M Hu Lab member (Brenley or Jim). We are working on establishing a page in GONUTS that you could enter your own dates for a personal competition, but we do not have that available yet.

Duration of Competitions

We can customize CACAO to contain any number of rounds each lasting any length of time for any number of participants. Please contact ecoliwiki@gmail.com, brenleymcintosh@gmail.com for competitions.

Recruiting students & student requirements

CACAO I

Recruiting undergraduate students at Texas A&M for CACAO I

- We used several methods to increase the visibility and enrollment in CACAO because we were usually delayed in getting the course into the TAMU calendar until after registration was already going on.
    1. we posted fliers about the course around TAMU in the science buildings prior to each course.
    2. we contacted the undergraduate advisors for Biochemistry, Genetics and Biology who agreed to send out emails with the flier of the course to their listservs.
    3. several coaches contacted professors instructing biology, microbiology, genetics, or biochem classes and requested 5 minutes during the first week of classes to address their students. A quick summary of the course was given in these classes to recruit participation of students in CACAO.

Prerequisites for undergraduates at Texas A&M for CACAO I

- There are no specific course prerequisites, but each student will need to be an active learner, undaunted by the challenge of digging for information and unafraid to ask questions when he/she gets stuck.

TAMU Recruitment Flyer

CACAO II

We promote our experienced CACAO I students to CACAO II.

Students often wanted to take the course repeatedly, but we decided it was an unfair advantage to have experienced TAMU students competing against students entirely new to the process of annotating. The courses are separate, though offered simultaneously, each for a single credit hour and meeting Tues evenings from 7-9pm CST.

The idea is that these students will assist with mentoring new students, providing feedback on their annotations as well as judging challenges and making annotation assessments. They help us deal with the large number of annotations generated by students in the CACAO competition (i.e. CACAO I students).

Recruiting undergraduate students at Texas A&M for CACAO II

- We ask students who have completed CACAO I if they are interested in registering for CACAO II.

Prerequisites for undergraduates at Texas A&M for CACAO II

- Undergrads must have completed CACAO I to enroll in CACAO II.

Recruiting graduate students at Texas A&M for CACAO II

- Mostly by word of mouth, our recruiting is fairly informal for grad students.

Prerequisites for graduate students at Texas A&M for CACAO II

- No formal prerequisites.
- Students should have a good, solid understanding of Genetics and Molecular Biology.
    - Strongly recommend students to take BICH/GENE 631 and/or BIOL 650.
- Students are expected to do independent work to supplement their background knowledge as needed. In addition, we assume grad students are familiar with the basic operational knowledge of computers and the internet.

Who has participated in CACAO?

Universities across the US and UK. We have had as many as 309 students in a single semester and we have trained more than 700 students as of Fall 2012 (Figure 1).

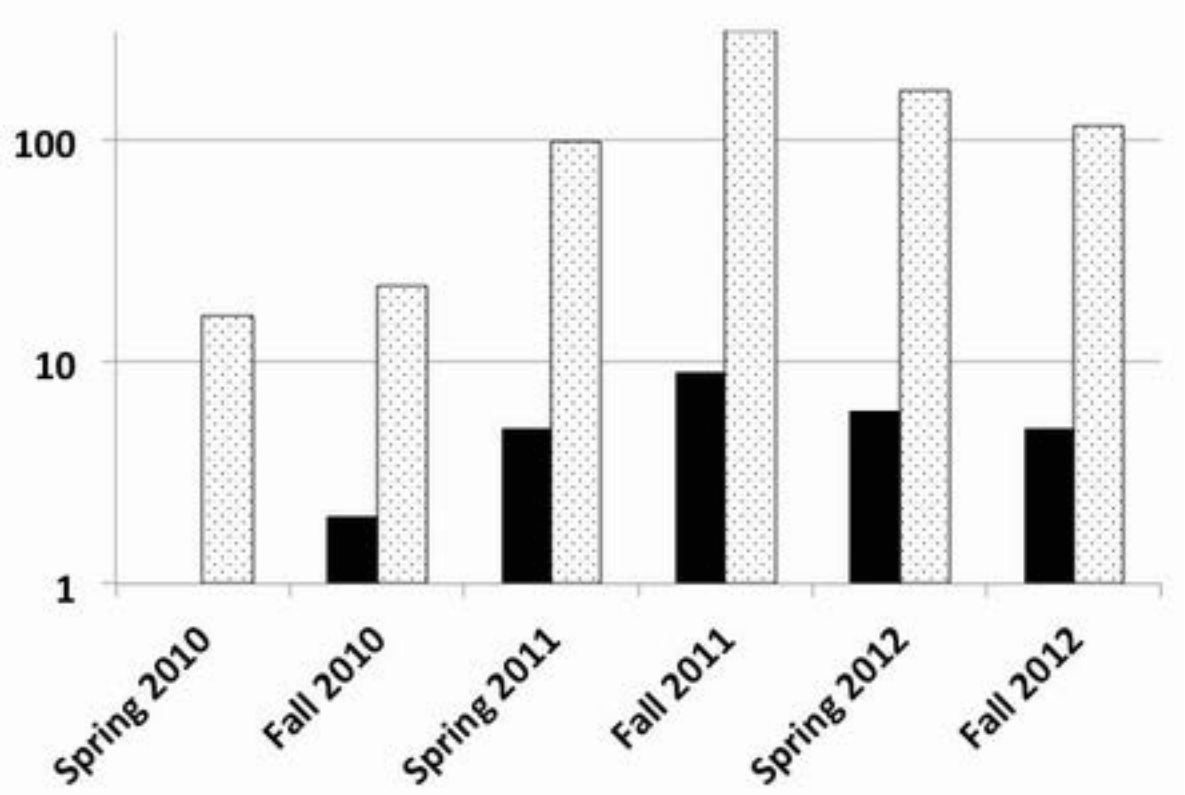


Figure 2: Participation in CACAO by semester. Black bars represent the number of schools and dappled bars represent the number of students.

Competition Set up on GONUTS

All of the set up is Optional & it is preferred that you leave this to us at this time.

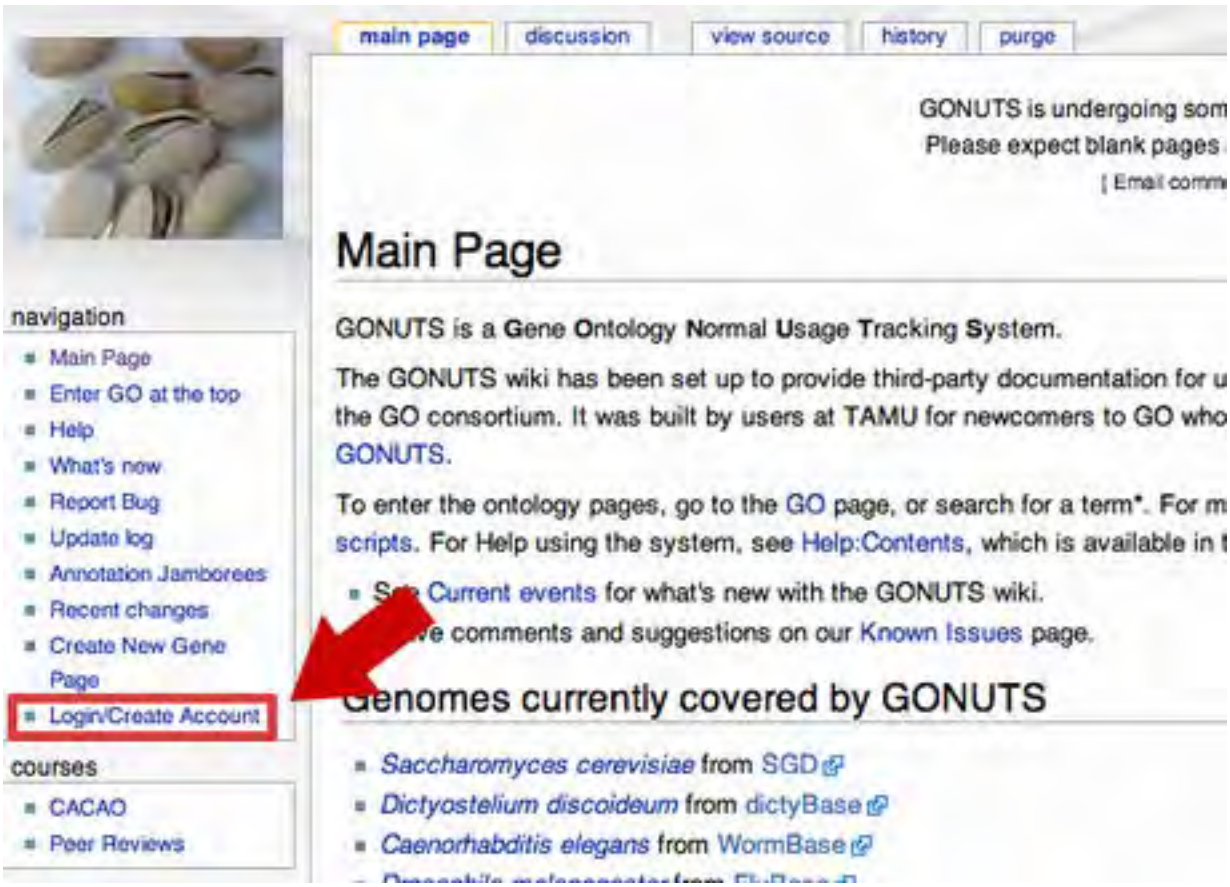
- We can do this step for you if you send us a list of students, their usernames & email addresses - see step #1. We will also need a team name for each group of students.
- Order matters here.
    - Make the students' user accounts
    - Make the competition page
    - Make the team pages & add them to the competition (see below)
    - Add the users to their teams (see below)
- If done in any other order, the competition won't recognize the teams and the teams won't recognize the users (aka students).
- This assumes that the instructor already has a GONUTS account and is logged in (http://gowiki.tamu.edu). If not, please contact us at ecoliwiki@gmail.com or brenleymcintosh@gmail.com.

Making User Accounts for Students (step #1) - optional

Your students must have a login on GONUTS to be able to edit/add/change anything on the site.

1. For each student, you will need:
    a. valid email address
    b. desired username
    c. real name
    d. team name
2. You have two choices:
    a. Send us a list of the usernames/emails and we will make the accounts on your behalf. Skip directly to #3 in this section.
        - We can use a script we have called 'make_groups.php' to do all of this automatically, including making user accounts, setting up user pages, setting up team pages & the competition page!

    b. Create the accounts for your students yourself.
        i. Click on "Login/Create Account" in the menu on the left of every page.



ii. Click on “Create Account” on the next page.

iii. Fill in the form with the username, email and real name. * You do not need to fill in the password fields at this time. *

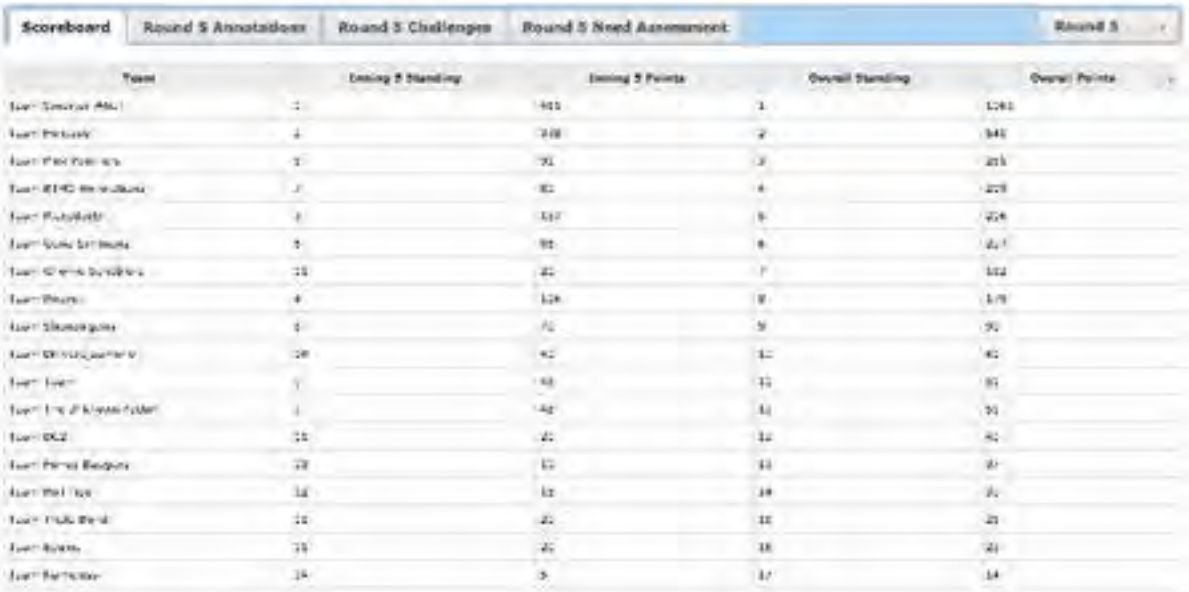
3. Each student will receive an email from GONUTS with a temporary password or a link to click on to confirm their account creation. They can use their username and this password to log in the first time they use GONUTS and can then change their password to whatever they want.

Setting up a Scoreboard (set #2) - optional

1. Click on the CACAO Spring 2011 link at the bottom of the team page

   You can also type in the url:
   http://gowiki.tamu.edu/wiki/index.php/Category:CACAO_Spring_2011

2. When you "create" this page, you will need to enter the following text:
   <scoreboard>
   session = Category:CACAO_Spring_2011
   </scoreboard>
   and then click on "Save page". The session is the current competition (i.e. Category:CACAO_Spring_2011 or Category:CACAO_Sprinog 2013).
3. After saving, the scoreboard will show up on the page.



Teams for CACAO

Team Structure

- Teams - we build our teams with 2 or 3 members in each.
    - We try not to put two freshmen on the same team.
    - Each team will need a name (ie UCL1 or Midnight_Yell). The students can choose their team name, or it can be assigned. It must be a name that has not been used before in any CACAO competition.
    - we have had teams as large as 6 students.

*** If you send us a list of students, usernames, email addresses & their team names, we can do all of this for you.**

Setting Up Team Pages (step #3) - optional

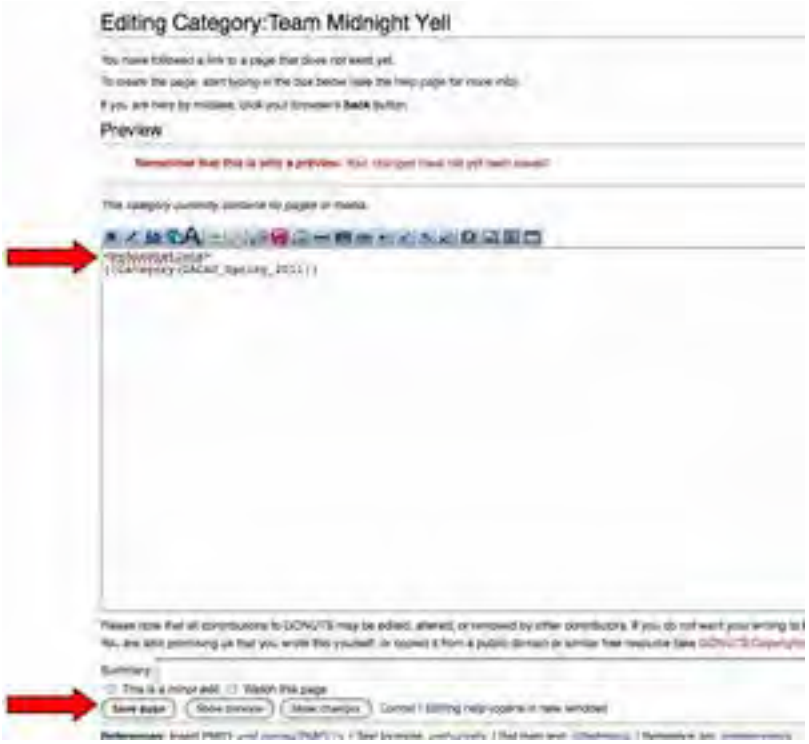
* **Optional** - You do not have to do any this on your own if you send us a list of your students & team names - ecoliwiki@gmail.com, brenleymcintosh@gmail.com.

1. Click on the team name to go to the team page.

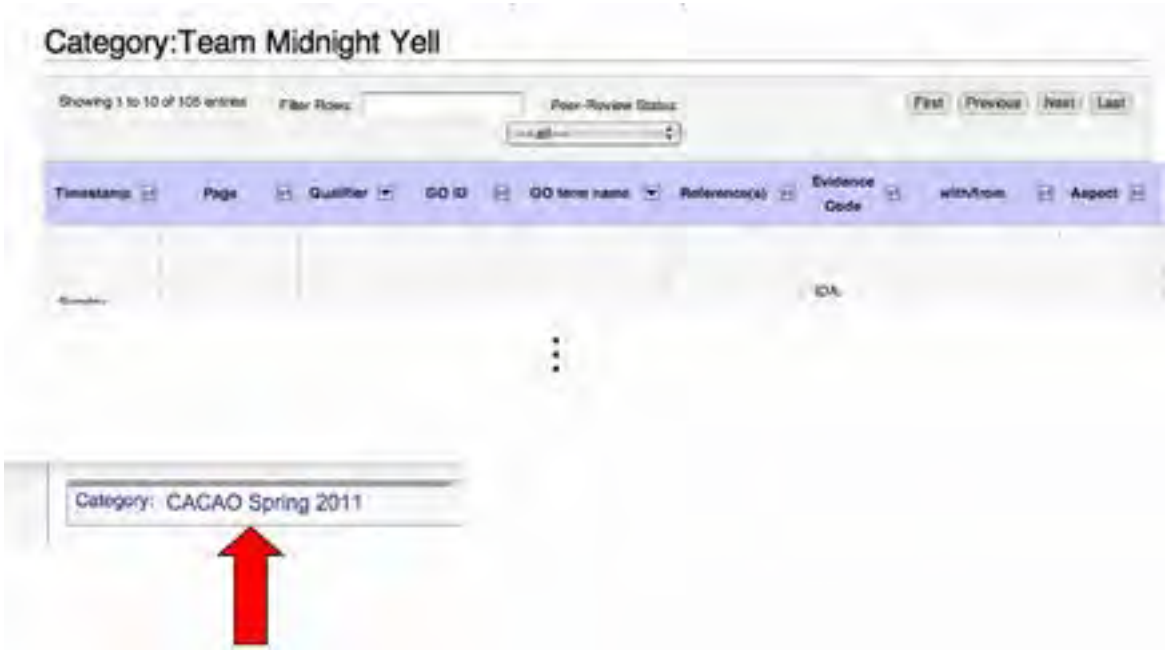
You can also type in the url:
http://gowiki.tamu.edu/wiki/index.php/Category:Team_Midnight_Yell

2. The page will look like the following image. Click on "create".

3. Add the text:
   <cacao>
   session = Category:CACAO_Spring_2011
   </cacao>
   The session is the competition to join (top arrow) to this page and click on "Save page" (bottom arrow).



4. All users that are put on this team (ie you added the group = Category:Team_Midnight_Yell to their user page) will have their annotations show up on this Team page as well. Also, this team page will be a part of the category for the competition (arrow) for the competition they are entered in (i.e. Spring 2011, Fall 2012, Spring 2013, etc).

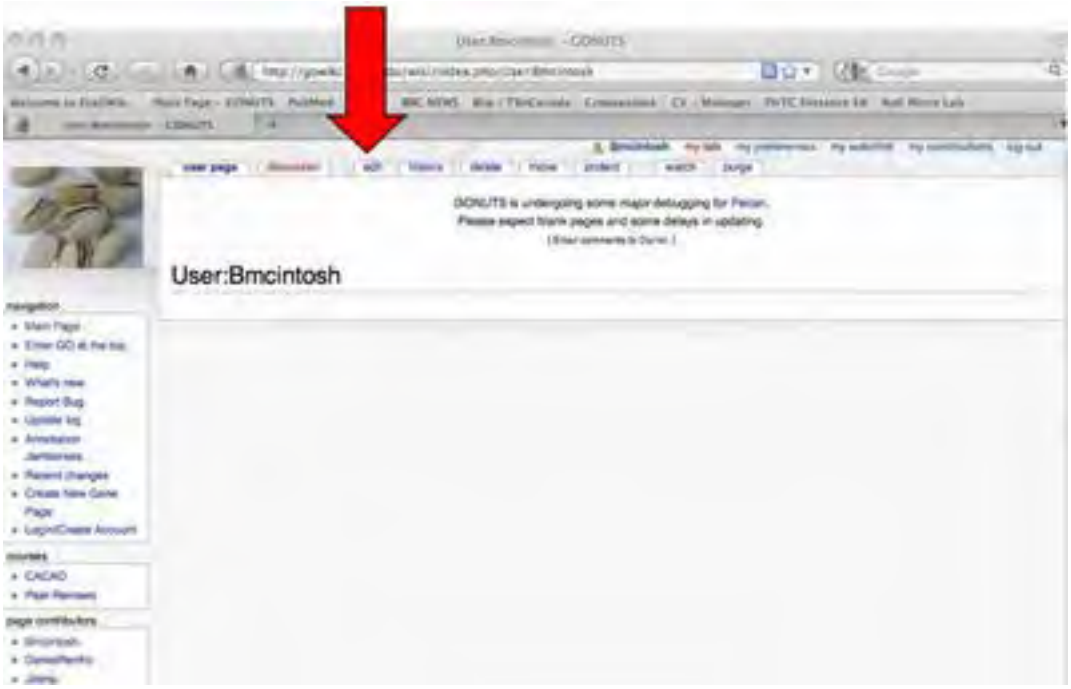

Category:Team Midnight Yell
Showing 1 to 10 of 105 entries
Filter Rows:
Peer-Review Status:
First
Previous
Next
Last
Timestamp
Page
Qualifier
GO ID
GO term name
Reference(s)
Evidence Code
with/from
Aspect
IDA

Category: CACAO Spring 2011

Assigning Teams to Student Users (step #4) - optional

- Once the students have accounts, each will have their own “user page”. For this example, I will use my username (Bmcintosh) and user page and the team name of Midnight_Yell. My user page can be found at:
  http://gowiki.tamu.edu/wiki/index.php/User:Bmcintosh

1. If you click on the user’s name for the first time, GONUTS will automatically send you to a page you can edit (go to #2). This is what my user page looks like (it has already been made and I have edited it before) - if you click on “edit” (arrow), you will be able to add text.



2. The text to add is:
   <cacao>
   group = Category:Team_Midnight_Yell
   session = Category:CACAO_Spring_2011
   </cacao>
   and then save the click on “save page” (bottom arrow).

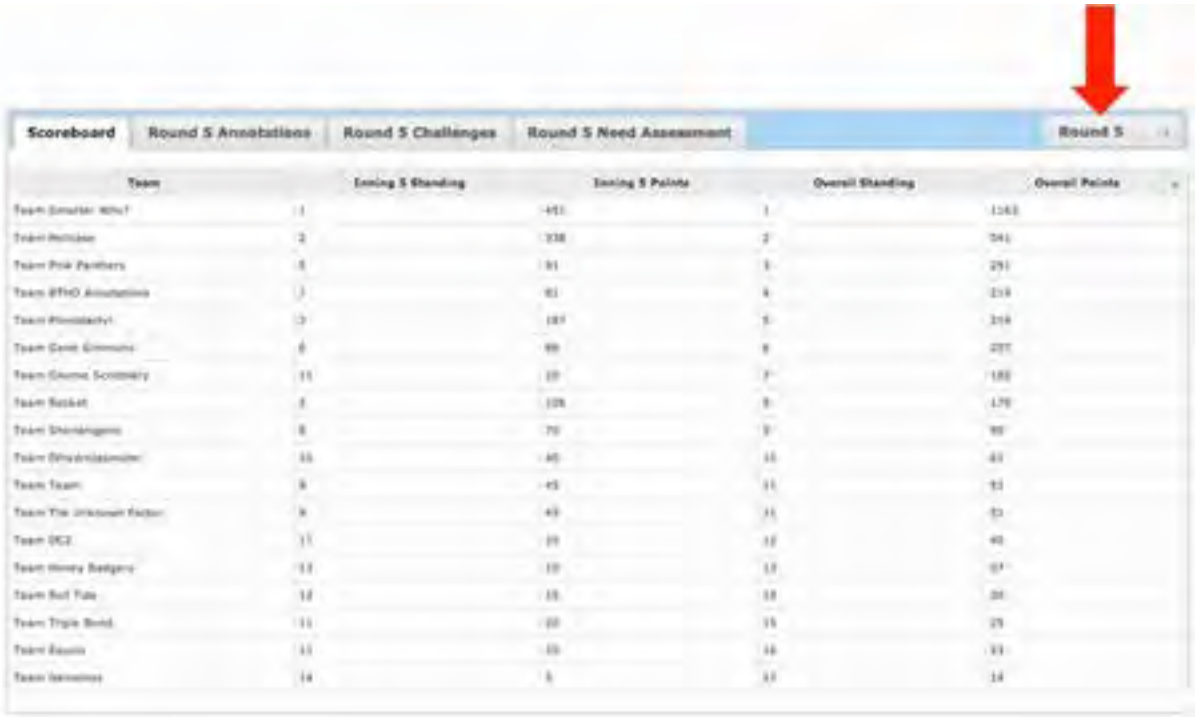
3. This will make a table automatically show up with all of my annotations in it. If the user has no annotations, the table will say so. This puts the user on their team (aka group) and into the current competition (Spring 2011, in this example). The team name shows up as a clickable link at the bottom (see the arrow).


User:Bmcintosh

CACAO Scoreboard

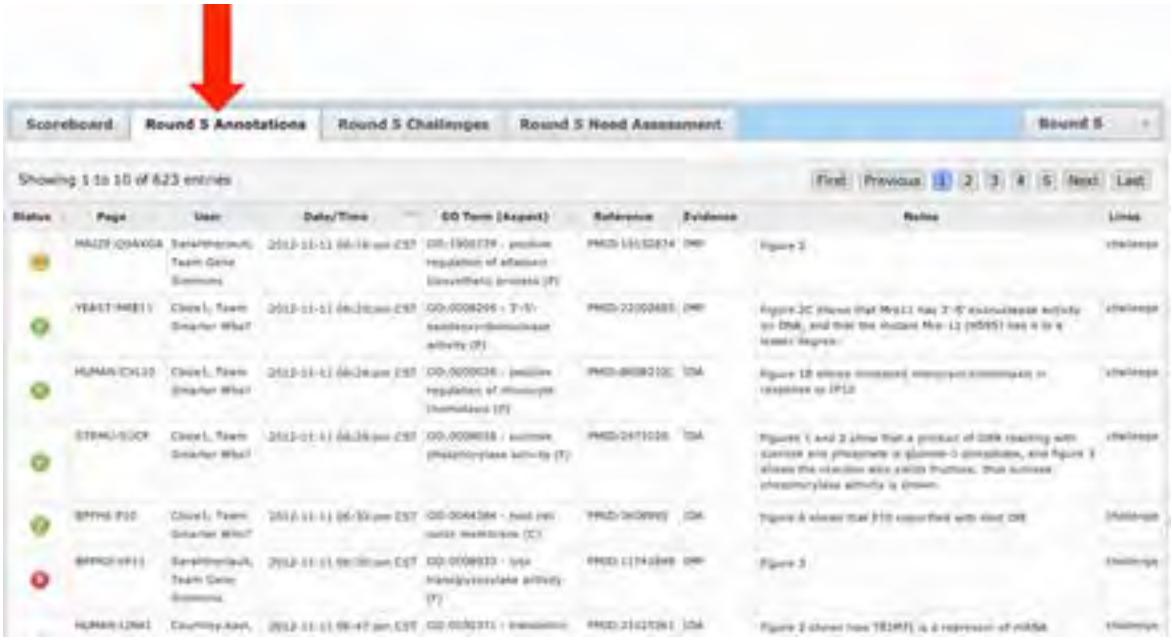
Scores

All of the scores are displayed on the scoreboard by team. You can change the round you are viewing by clicking on the “Round #” dropdown menu at the top right. Each of the team names is a link that you can click on to view the annotations from each team.



Round by Round Annotations

The annotations from each round can be viewed under the “Round # Annotations” tab.

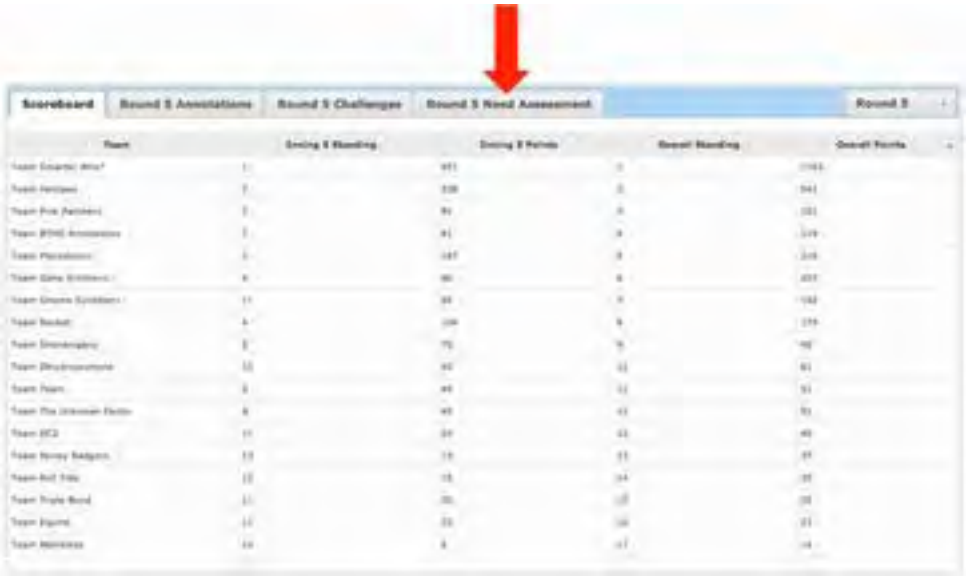


Round by Round Challenges

You can view each round’s challenges under the tab “Round # Challenges”.

Annotations that still need to be assessed

And the annotations still needing assessment are under the “Round # Need Assessment” tab for each round. Please assess any annotations still listed under this tab!



This ends PART I: SETTING UP A COMPETITION.

At this point, you are finished with the technical set up of the competition. Next, you will need to train your students as biocurators.

Each student needs training with:

- Finding proteins & papers to annotate
- Using UniProt and finding accession
- GO (Gene Ontology)
- Formatting an annotation
- Navigating GONUTS (how to add an annotation to a gene page, how to make a gene page using a UniProt accession)
- Challenging & rebuttals

PART II: GO Annotations

Where do we add annotations during CACAO?

Our lab operates and maintains several wikis, including EcoliWiki (http://ecoliwiki.net), SubtilisWiki (http://subtiliswiki.net), OMPwiki (http://microbialphenotypes.net) and GONUTS (http://gowiki.tamu.edu). The scoreboard and challenge system are only on GONUTS. This will be the site to which annotations will be added and stored. It also provides the users with the ability to search the GO terms. The process of making a page and adding an annotation to the page in GONUTS is described later in the section called **Instructions on how to add annotations to GONUTS**.

Why do we use GONUTS?

We use GONUTS because of several significant advantages:

1. GONUTS can be edited by users unlike other closed databases such as UniProt.
    - Users can make a gene page automatically (using a UniProt accession) that has the annotations from UniProt for that protein, but can be added to by the student(s).
2. GONUTS also has all of the GO terms.
    - These terms are searchable and offer information about the term (children and parent terms, a definition, synonyms, usage notes and so forth).
3. We have designed tools on GONUTS that allow students to follow their team scores in real time and a system for peer review/challenging.

Quick Guide to Annotating

We allow students to annotate for 7 days, then challenge for 7 days. We call this two week set a “round”. We generally run 5 rounds per semester or 4 rounds per quarter, with several training weeks beforehand and 1-2 weeks to wrap up.

There is NO single place to start (i.e. protein, paper, GO term, evidence code), but you will ultimately have to do the following things:

1. **Find a suitable article on PubMed about a protein (no reviews, no notes, no Wikipedia articles). Record the PMID number (NOT the PMC number!!)**
2. **Find the SAME protein in UniProt and get the accession/entry number**
3. **Use the UniProt accession to make an editable protein page on GONUTS**
4. **Find a suitable GO term based on figure(s) &/or table(s) characterizing the protein**
5. **Pick a suitable evidence code based on how protein was characterized in those figure(s) &/or table(s)**
6. **Enter (and save) the GO annotation on the protein’s page in GONUTS, complete with notes (indicating the figure(s) &/or table(s) that support it)**
7. **Challenge & refute challenges to team’s annotations**

Choosing something to annotate - PROTEINS

There are a number of ways to pick targets to annotate:

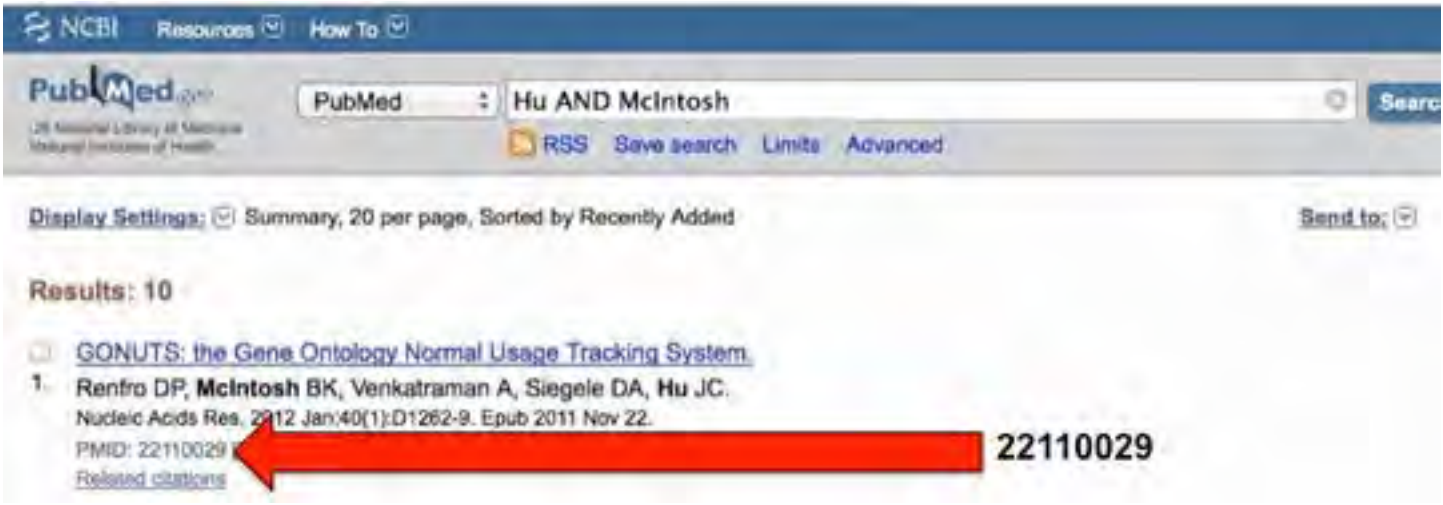
1. randomly
2. topics of interest (ie efflux pump proteins, biofilms)
3. papers you have come across while doing other stuff
4. methods you know or want to learn
5. phenotypes and mutants you are interested in
6. by author
7. by pathway or regulon
8. suggested by another (ie high IEA:manual annotation ratio)
9. current paper mentions another gene product
10. review papers (ie Annual Reviews are excellent sources)

Quick Guide to Annotating

1. **Find a suitable article on PubMed about a protein (no reviews, no notes, no Wikipedia articles). Record the PMID number (NOT the PMC number!!)**
2. Find the SAME protein in UniProt and get the accession/entry number
3. Use the UniProt accession to make an editable protein page on GONUTS
4. Find a suitable GO term based on figure(s) &/or table(s) characterizing the protein
5. Pick a suitable evidence code based on how protein was characterized in those figure(s) &/or table(s)
6. Enter (and save) the GO annotation on the protein's page in GONUTS, complete with notes (indicating the figure(s) &/or table(s) that support it)
7. Challenge & refute challenges to team's annotations

Finding the PMID on PubMed

Students must have the PMID and ONLY the PMID. PMC numbers are NOT acceptable! The PMID number is underneath the Journal name for each record when searching, or under the abstract if on the abstract of a specific paper in PubMed.



PMC numbers can be converted to PMID numbers either on PubMed (ie search the PMCnumber on PubMed) or there are programs that will do this for you (google search convert PMC to PMID).

Quick Guide to Annotating

1. Find a suitable article on PubMed about a protein (no reviews, no notes, no Wikipedia articles). Record the PMID number (NOT the PMC number!!)
2. **Find the SAME protein in UniProt and get the accession/entry number**
3. Use the UniProt accession to make an editable protein page on GONUTS
4. Find a suitable GO term based on figure(s) &/or table(s) characterizing the protein
5. Pick a suitable evidence code based on how protein was characterized in those figure(s) &/or table(s)
6. Enter (and save) the GO annotation on the protein's page in GONUTS, complete with notes (indicating the figure(s) &/or table(s) that support it)
7. Challenge & refute challenges to team's annotations

Finding UniProt Accessions

GONUTS is designed to pull the annotations already on UniProt onto a page and into a table that can be edited. To do this, GONUTS needs the UniProt Accession.

Why do we have to use UniProt accessions on GONUTS?

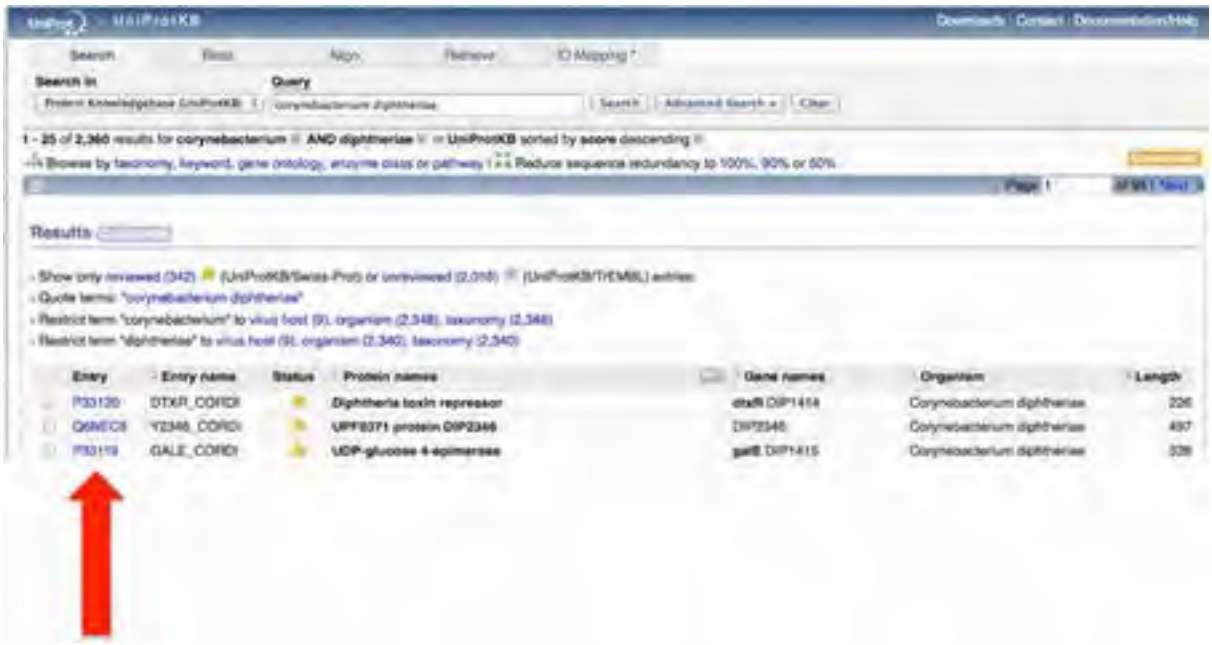
- UniProt records are not editable by the public.
    - They have professional, PhD-level biocurators who add their GO annotations to their protein pages.
- GONUTS is capable of making an editable protein page based on the UniProt entry/ accession number
    - Dr. Hu at Texas A&M operates GONUTS
- We can harvest correct annotations for proteins in GONUTS (that are made using their UniProt accession) & send these back to UniProt and other databases.

Why aren't all of the proteins already in GONUTS from UniProt?

- We make the pages for proteins on a case-by-case basis because there are SO many records in UniProt (hundreds of millions), most of which are never going to be annotated with literature. Most of their records are submitted en masse by authors doing sequencing projects & are studied directly for publication.

To find a UniProt Accession

1. Go to http://www.uniprot.org
2. In the query box, enter a search term (ie cancer), protein name with organism (ie p53 human), gene name with organism, or other relevant term. Click on "Search".
3. Find the protein you are looking for & copy the accession. ***Be careful, it may not be the first item listed!**
4. The accession is 6 letters or numbers and look like **P8A0Y2**.

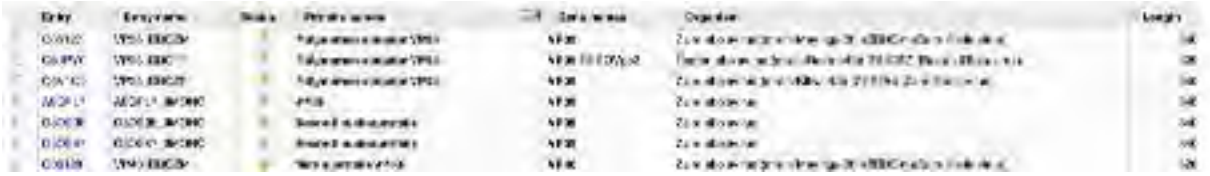


What does that status of the record mean?

Records that have been reviewed by one of the professional biocurators at UniProt are given “gold star” status, but records that are submitted en masse automatically to TrEMBL are given “grey star” status because no human has checked the record yet.

Does the gold star matter?

It absolutely does matter. It is preferable to use the reviewed record, but the accession for a non-reviewed record can be used if there is no reviewed record. If a reviewed record is available, but an annotation is entered using the non-reviewed accession to make the protein page on GONUTS, it will be marked as UNACCEPTABLE and subject to challenges.



How to identify gene products that are suitable:

- If you can identify the gene name (in *E. coli*, the gene names look like ***gyrA*** or ***acs***) and the species (ie E. coli, human, mouse, rat), you should be able to find the accession. Note that not every species has been sequenced and entered into UniProt!!

1. If you start with a subject of interest (ie glycolysis), there are a number of ways to identify the gene. A review article, search on another model organism database (ie EcoCyc, SGD, MGI, etc), search in PubMed/Google Scholar, search Google, look in a textbook or you can look for a gene product name (such as an enzyme name like **phosphofructokinase**). It is very helpful to narrow the search with the species of interest (ie human, mouse, E. coli, etc).
2. If you start with a gene of interest (ie p53 in humans), get the UniProt accession & make the page in GONUTS. It is recommended to not spend a bunch of time reading a paper carefully until you have looked at the GONUTS page to see if the paper has already been annotated by someone else.
3. Once you have a paper and find the UniProt accession, make the page on GONUTS and see if your paper has already been annotated.

Classroom Activity - “UniProtology - the study of the Universal Protein Resource”

- Before this is given, you need to supply the students with background information on the Gene Ontology & also give them the web address of UniProt (http://uniprot.org).
- This activity is meant to be an opportunity for students to navigate UniProt and look at protein pages on UniProt. It can be done individually or in groups and should take roughly 5-10 mins for the students to prepare their answers. Answers can be collected or discussed as a class.

1. How many records are there when you search for *Ebola virus strain Zaire VP35*?
2. How many of those records have been reviewed by professional biocurators are UniProt?
3. What is the difference between the record for Q05127 (http://www.uniprot.org/uniprot/Q05127) and Q6V1Q9 (http://www.uniprot.org/uniprot/Q6V1Q9)?
4. Does this matter?

Answers:

1. 15
2. 12
3. Q05127 is from the Mayinga-76 strain, whereas Q6V1Q9 is from the Kikwit-95 strain.
4. Yes, it matters. You must find the SAME protein in UniProt that is described in the paper, right down to the strain, subspecies and/or any other identifying factor!!

Quick Guide to Annotating

1. Find a suitable article on PubMed about a protein (no reviews, no notes, no Wikipedia articles). Record the PMID number (NOT the PMC number!!)
2. Find the SAME protein in UniProt and get the accession/entry number
3. **Use the UniProt accession to make an editable protein page on GONUTS**
4. Find a suitable GO term based on figure(s) &/or table(s) characterizing the protein
5. Pick a suitable evidence code based on how protein was characterized in those figure(s) &/or table(s)
6. Enter (and save) the GO annotation on the protein's page in GONUTS, complete with notes (indicating the figure(s) &/or table(s) that support it)
7. Challenge & refute challenges to team's annotations

Making Protein Pages on GONUTS Using the UniProt Accession

- Annotations have to be entered on gene pages. You may be able to find existing gene pages where you can add to existing annotations. However, sometimes you will want to annotate a gene that doesn't already have a page in GONUTS.
- Don't forget to log in to GONUTS. You won't be able to create pages, edit, etc otherwise.

1. Once you have the UniProt Accession (ie. P04637), go to GONUTS (http://gowiki.tamu.edu).
2. Click on the "Create New Gene Page" link in the menu on the left of the page.

3. Paste the UniProt Accession into the empty box on the GoPageMaker page and click “Create”. This will make the gene product page in GONUTS automatically, but this may take up to 60 seconds to fetch the info and create the page.

Quick Guide to Annotating

1. Find a suitable article on PubMed about a protein (no reviews, no notes, no Wikipedia articles). Record the PMID number (NOT the PMC number!!)
2. Find the SAME protein in UniProt and get the accession/entry number
3. Use the UniProt accession to make an editable protein page on GONUTS
4. **Find a suitable GO term based on figure(s) &/or table(s) characterizing the protein**
5. Pick a suitable evidence code based on how protein was characterized in those figure(s) &/or table(s)
6. Enter (and save) the GO annotation on the protein's page in GONUTS, complete with notes (indicating the figure(s) &/or table(s) that support it)
7. Challenge & refute challenges to team's annotations

GO terms in GONUTS

- For CACAO, we use GONUTS (http://gowiki.tamu.edu) to both explore GO and to make annotations.
- On GONUTS, you can search for GO terms (and add your annotations to gene pages).
- If you click on "G", the search will suggest corrections if you spell your term incorrectly.

Do we have to use GONUTS to search for GO terms?

No. There are a number of different sites that have the entire GO. The European Bioinformatics Institute (EBI) has a site called QuickGO at http://www.ebi.ac.uk/QuickGO/. The GO Consortium has its own site called AmiGO at http://amigo.geneontology.org/. Any of these sites is fine to use, it just depends on what each user is most comfortable with.

What information is on each GO term page in GONUTS?

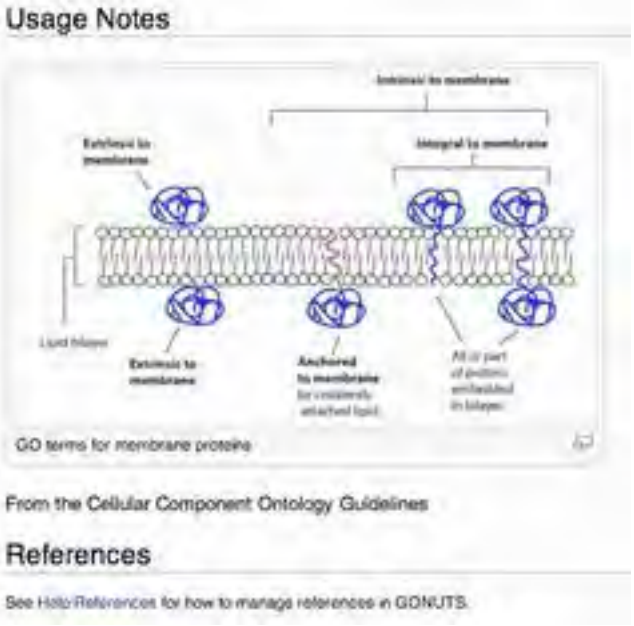
Each term page will divided into several parts.

Part 1. Term Information

There is a box that contains the GO ID and the name (1), which of the three ontologies the term belongs to (2), definition (3), synonyms (4), relationships (5) and a DAG (6) (directed acyclic graph showing relationships).

Part 2 & 3. Usage Notes & References

If edited, these can be helpful in giving examples or further information on how to use the term or how not to! "References" will be added if there are papers described in the Usage Notes.



Part 4. Child Terms

Think of each ontology as a tree with the root being the most general term (ie cellular component). This is the parent to all of the component terms - this makes all of the more

specific terms its children. Thus, each child term like a branch, each child of a child is another branch off that, etc. You can walk the tree to increasingly specific terms (or smaller branches), but the goal of annotating is to find the most appropriate term for the evidence. This might mean that you need to use a less specific (more towards the trunk) or more specific term from where you are currently in the ontology. The Child Terms displayed allows you to see and surf more specific terms.

Part 5. Other proteins annotated to this term already

The last section on every term page can be very helpful when annotating. It consists of a list (if there are any) of gene products from other model organisms (*S. cerevisiae, D. discoideum, C. elegans, D. melanogaster, M. musculus, D. rerio, A. thaliana, S. pombe* and *G. gallus*) that have been annotated to this term. If you click on a link, you will see the GONUTS annotation page for that gene product, which may offer you some insight into other GO terms that might co-occur with the term.

Classroom Activity - “GO figure the subontology”

- This is meant to get the students to think about what each of the subontologies contain in a hands-on activity. You will need to at least have introduced the 3 subontologies of GO (Molecular Function, Biological Process & Cellular Component).
- It can be done individually or in groups and should take less than 5 minutes for the students to prepare their answers. Answers can be discussed as a class using a voting system.

Which subontology would the following terms be in?

1. GO:0003909 DNA ligase activity
2. GO:0071705 Nitrogen compound transport
3. GO:0007124 Pseudohyphal growth
4. GO:0015123 Acetate transmembrane transporter activity
5. GO:0071514 Genetic imprinting
6. GO:0005773 Vacuole
7. GO:0000312 Plastid small ribosomal subunit

Answers:

1. MF
2. BP
3. BP
4. MF
5. BP
6. CC
7. CC

Classroom Activity or Homework Assignment - "Going Nuts on GONUTS"

- Before this is given, you need to supply the students with background information on the Gene Ontology & also give them the web address of GONUTS (http://gowiki.tamu.edu).
- This activity is meant to be an opportunity for students to navigate GONUTS and look at GO terms, IDs, and gene pages. It can be done individually or in groups and should take roughly 5-10 mins for the students to prepare their answers. Answers can be collected or discussed as a class.

1. What is the GO term for GO:0004713?
2. What is the GO identifier for mitosis?
3. How many results (ballpark) do you get when you search for cell division using the Go, Search or G buttons?
4. How many annotations are on page for the mouse protein Bub1a?
5. How many child terms are there for plasma membrane? How many grandchildren?
6. What term is the parent of GO:006825?

Answers:

1. Protein tyrosine kinase activity
2. GO:0007067
3. Go = 1; Search = more than 40!; G = 1
4. 29
5. 2; 229
6. transition metal ion transport

Quick Guide to Annotating

1. Find a suitable article on PubMed about a protein (no reviews, no notes, no Wikipedia articles). Record the PMID number (NOT the PMC number!!)
2. Find the SAME protein in UniProt and get the accession/entry number
3. Use the UniProt accession to make an editable protein page on GONUTS
4. Find a suitable GO term based on figure(s) &/or table(s) characterizing the protein
5. **Pick a suitable evidence code based on how protein was characterized in those figure(s) &/or table(s)**
6. Enter (and save) the GO annotation on the protein's page in GONUTS, complete with notes (indicating the figure(s) &/or table(s) that support it)
7. Challenge & refute challenges to team's annotations

Evidence Codes

Evidence codes describe the type of work or analysis done by the authors. There is an entire ontology dedicated to evidence codes and the GO Consortium uses more than 20 of these. However, for CACAO, students may only use 7. They describe how a protein might be characterized experimentally or computationally.

- IDA:Inferred from Direct Assay
- IMP:Inferred from Mutant Phenotype
- IGI:Inferred from Genetic Interaction
- ISA:Inferred from Sequence Alignment
- ISO:Inferred from Sequence Orthology
- ISM:Inferred from Sequence Model
- IGC:Inferred from Genomic Context

IDA:Inferred from Direct Assay

Describes a direct assay carried out to determine the function, process or component indicated by the GO term.

- enzyme assays using purified components
- *In vitro* reconstitution of a functionally active complex
- Immuno-fluorescence microscopy
- cell fractionation

IMP:Inferred from Mutant Phenotype

Describes cases when function, process or cellular localization is inferred based on the differences between the function, process or cellular localization between two different alleles of the corresponding gene.

- overexpression of wild-type or a mutant gene
- RNAi, anti-sense RNAs, antibody depletion

- using inhibitors, blockers, modifiers, changes in pH or ionic strength to see the effect of absence or significant depletion of the protein.

IGI:Inferred from Genetic Interaction

Describes any combination of alterations in the sequence or expression of more than 1 gene/ protein, including cases when redundant copies of gene must all be mutated to observe the informative phenotype or in which a gene from one organism complements a deletion or other mutation in another species. The with/from field should contain the UniProt accession(s) of the other genes relevant to the experiment.

- traditional genetic interactions -- suppressors, synthetic lethals
- rescue experiments
- functional complementation across species

ISA:Inferred from Sequence Alignment

Describes a PUBLISHED sequence alignment (pairwise or multiple alignments). The with/ from field should contain the other protein(s) relevant to the alignment that have an existing GO annotation based on an experiment to the GO term of interest.

- aligned protein(s) UniProt accession is entered in with/from
- aligned protein(s) MUST HAVE AN ANNOTATION TO THE SAME GO TERM WITH IDA, IMP or IGI EVIDENCE CODE & BASED ON A PUBLICATION.

ISO:Inferred from Sequence Orthology

Describes a PUBLISHED sequence alignment when the proteins are established to be orthologs or a phylogenetic analysis used to define a orthologous groups. The with/from field should contain the other protein(s) relevant to the assignment of orthology that have an existing GO annotation based on an experiment to the GO term of interest.

- orthologous protein(s) UniProt accession is entered in with/from
- orthologous protein(s) MUST HAVE AN ANNOTATION TO THE SAME GO TERM WITH IDA, IMP or IGI EVIDENCE CODE & BASED ON A PUBLICATION.

ISM:Inferred from Sequence Model

Describes a PUBLISHED analysis from some kind of statistical modeling program of a sequence or group of sequences used to make a prediction about the function of a protein. The with/from field should contain the PMID of the publication in which the statistical modeling program was first described.

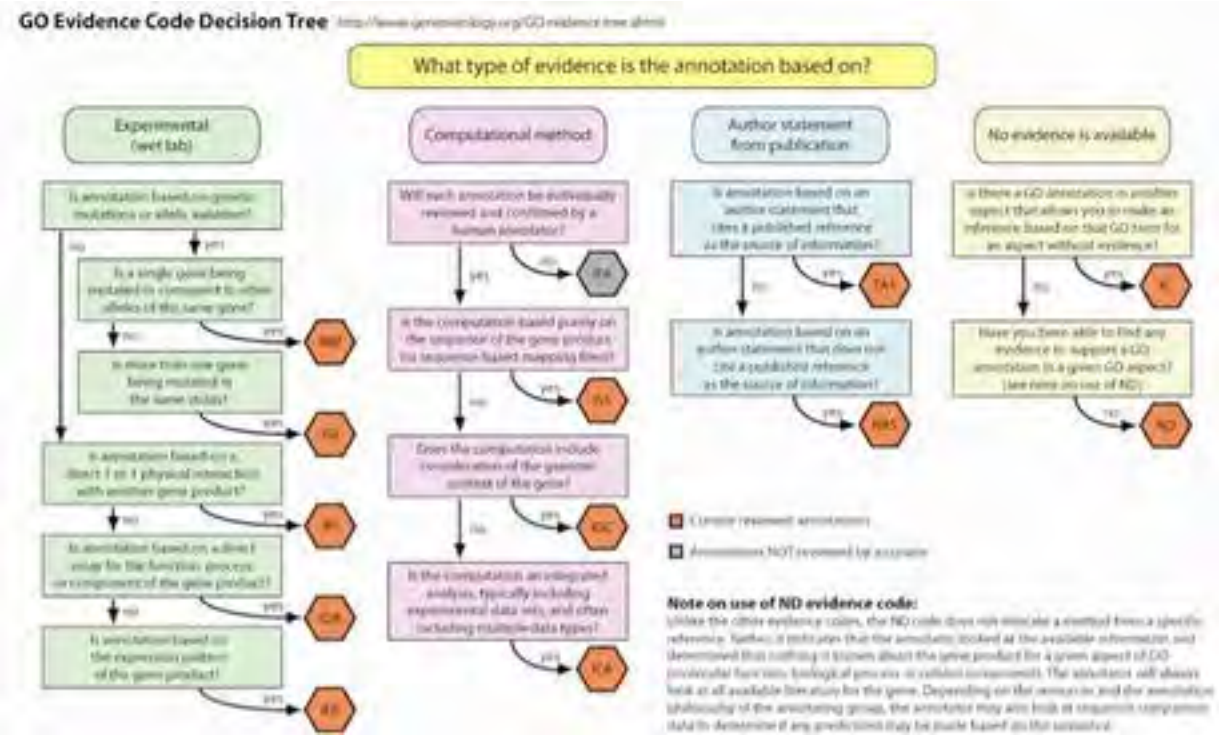
- PMID accession for publication of the modeling program is entered in with/from

IGC:Inferred from Genomic Context

Describes when information about the genomic context of a protein forms part of the evidence for a particular annotation.

- operon structure
- syntenic regions (identity of genes neighboring the gene in question)
- pathway analysis

Any other evidence code is NOT allowed in CACAO. It will be marked as either requires changes or unacceptable and will NOT count towards the required number of annotations in each course.



For more information on evidence codes:

http://geneontology.org/GO.evidence.shtml
http://gowiki.tamu.edu/wiki/index.php/evidence_codes

Classroom Activity - “Evidence de-coded”

- Before this is given, you need to supply the students with background information on the Gene Ontology .
- This activity is meant to be an opportunity for students to work with the evidence codes permitted in CACAO. It can be done individually or in groups and should take roughly 5-10 mins for the students to prepare their answers. Answers can be collected or discussed as a class.

1. You have a paper that shows that a certain phenotype only shows up when you delete 2 genes (gene aaaA and gene zzzZ). What evidence code would you use? What page(s) do you make the annotation on? Does it need something in the with/from field? If it does, what would go in the with/from field?
2. You have a paper that purifies a protein from the mitochondrial membrane. What evidence code would you use? Does it need something in the with/from field? If it does, what would go in the with/from field?
3. You have a paper that shows that a certain phenotype shows up when you delete a gene. What evidence code would you use? What page(s) do you make the annotation on? Does it need something in the with/from field? If it does, what would go in the with/from field?
4. You have a paper that shows that shows a sequence alignment of protein A (the focus of the paper) to protein B. What evidence code would you use? Does it need something in the with/from field? If it does, what would go in the with/from field? What do you need to check before you save this annotation?

Answers:

1. IGI evidence code, which requires the “with/from” field to be filled in. The annotation can go on both aaaA and zzzZ’s protein page. The UniProt accession for the other protein will have to go into the with/from field (i.e. on aaaA’s page I would put the UniProt accession for zzzZ and the opposite on the protein page for zzzZ, which would have the UniProt accession for aaaA in the with/from field).
2. IDA. nothing else is required.
3. IMP. nothing else is required.
4. ISA, which require the “with/from” field to be filled in. This field takes a UniProt accession for Protein B. I need to check to see if Protein B has an experimental annotation (i.e. same GO term as I want to assign to protein A, but with a PMID and IDA, IMP or IGI as the evidence code). If Protein B doesn’t have an experimental annotation to this term, I will have to delete my annotation on protein A’s page.

Quick Guide to Annotating

1. Find a suitable article on PubMed about a protein (no reviews, no notes, no Wikipedia articles). Record the PMID number (NOT the PMC number!!)
2. Find the SAME protein in UniProt and get the accession/entry number
3. Use the UniProt accession to make an editable protein page on GONUTS
4. Find a suitable GO term based on figure(s) &/or table(s) characterizing the protein
5. Pick a suitable evidence code based on how protein was characterized in those figure(s) &/or table(s)
6. **Enter (and save) the GO annotation on the protein's page in GONUTS, complete with notes (indicating the figure(s) &/or table(s) that support it)**
7. Challenge & refute challenges to team's annotations

Editing a Gene Product Page and Adding An Annotation

1. Scroll down to the bottom of the "Annotations" table. Click on "edit table". This will make all of the rows available to edit.

2. Scroll to the bottom of the table again, click on “Add row”.

3. You must fill in the form and save (not update) your annotation.

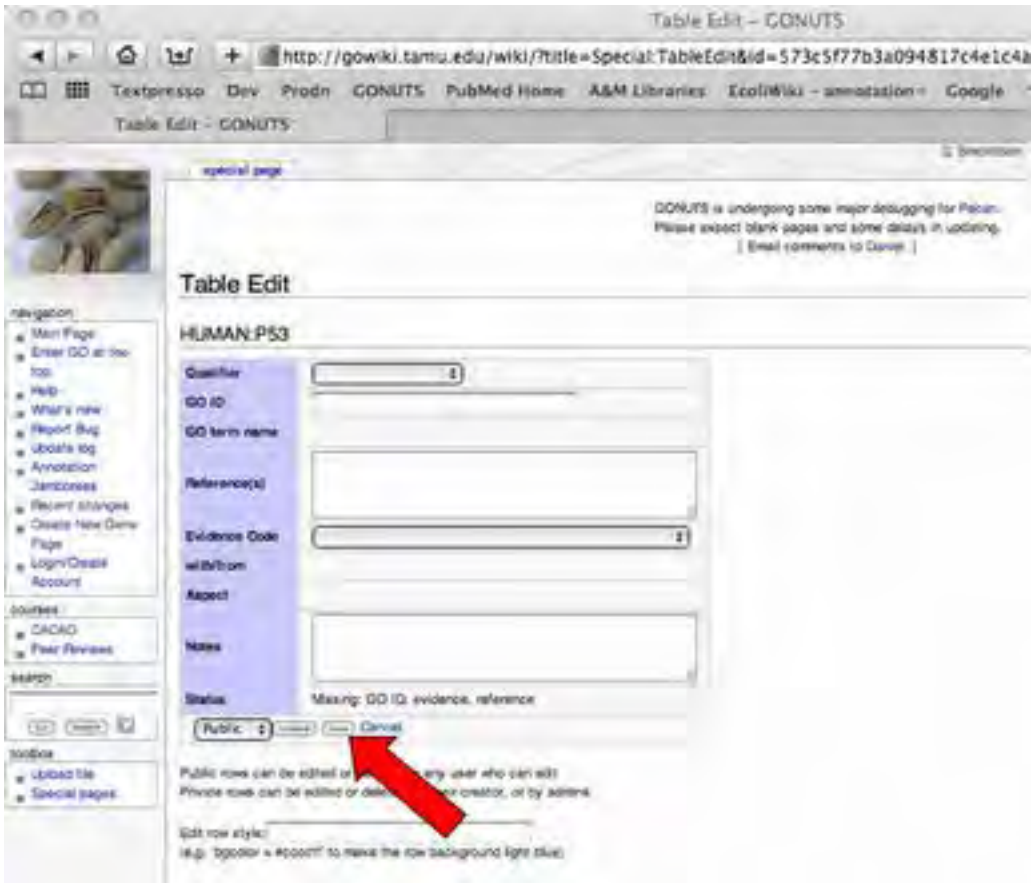
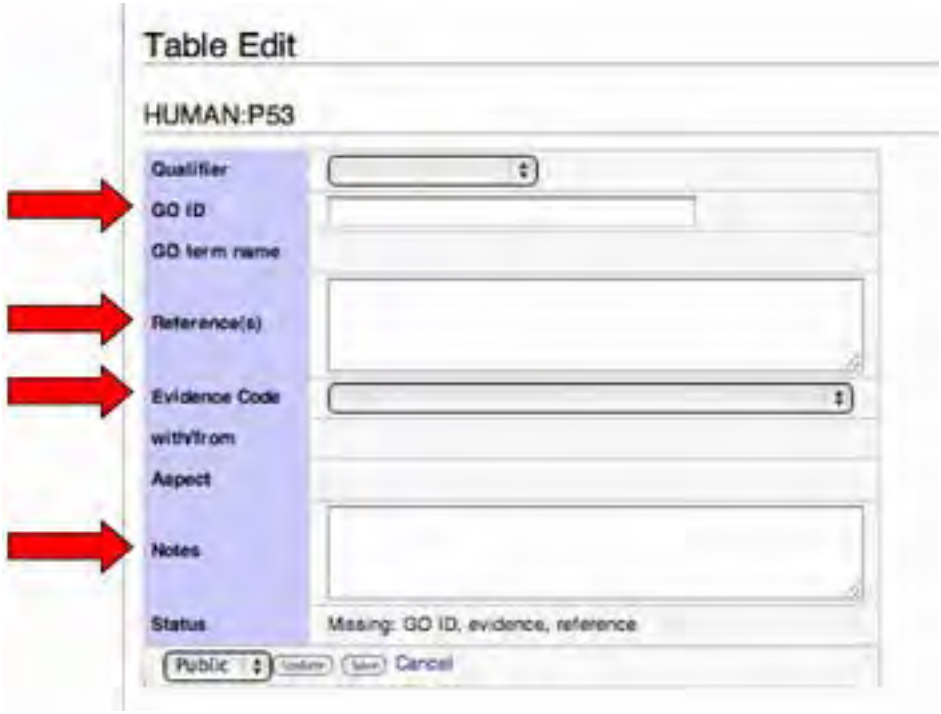


Table Edit – CONUTS
Table Edit - CONUTS
special page
CONUTS is undergoing some major debugging for Pecan.
Please expect blank pages and some delays in updating.
Table Edit
HUMAN:P53
GO ID
GO term name
Reference(s)
Evidence Code
Aspect
Notes
Status
Missing: GO ID, evidence, reference
Public
Cancel
navigation
Main Page
Help
What's new
Report Bug
Upload file
Special pages
search
toolbox

Components of an Annotation

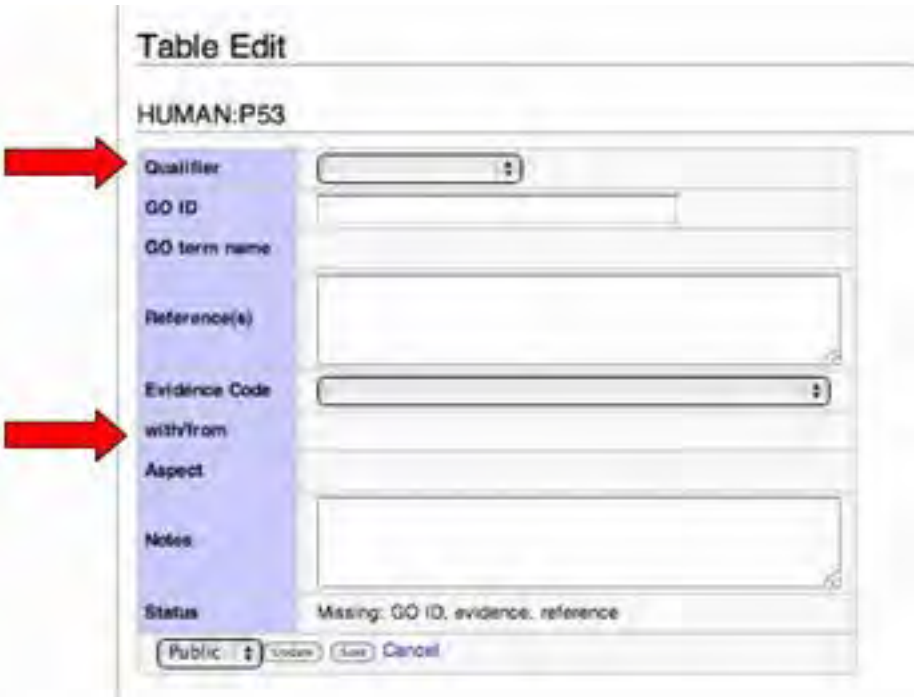
There are required fields you must add for every annotation:

1. **GO ID** - in the format GO:########
2. **Reference(s)** - in the format PMID:######
3. **Evidence Code** - selected from the dropdown menu
4. **Notes** - in the format of a short description of what figure/method/table is the evidence for this annotation



There are also optional fields that may be required for an annotation:

1. For some GO terms, you may add a “Qualifier”. This is VERY rare & the default is NOT TO INCLUDE A QUALIFIER!
2. For some evidence codes, you may have to fill in the “with/from” field.



Using the with/from field:

For IGI:

The with/from field should contain the UniProt accession(s) of the other genes relevant to the experiment.

For ISA:

The with/from field should contain the other protein(s) relevant to the alignment that have an existing GO annotation based on an experiment to the GO term of interest.

- aligned protein(s) UniProt accession is entered in with/from
- aligned protein(s) MUST HAVE AN ANNOTATION TO THE SAME GO TERM WITH IDA, IMP or IGI EVIDENCE CODE & BASED ON A PUBLICATION.

For ISO:

The with/from field should contain the other protein(s) relevant to the assignment of orthology that have an existing GO annotation based on an experiment to the GO term of interest.

- orthologous protein(s) UniProt accession is entered in with/from
- orthologous protein(s) MUST HAVE AN ANNOTATION TO THE SAME GO TERM WITH IDA, IMP or IGI EVIDENCE CODE & BASED ON A PUBLICATION.

For ISM:

The with/from field should contain the PMID of the publication in which the statistical modeling program was first described.

- PMID accession for publication of the modeling program is entered in with/from

QUALIFIERS ARE NOT NORMALLY USED. Don't add a qualifier unless you have checked with an experienced biocurator as it is likely not necessary!!!

contributes_to qualifier:

- If a paper shows that 2 or more (a complex) of gene products are required to get a certain activity and that the individual gene products alone are not sufficient to show this activity.
- This is ONLY relevant to molecular function terms. We assume proteins contribute to their processes.

NOT qualifier:

- If a paper that says "in contrast to previous reports, this protein was not localized to the inner membrane upon subcellular fractionation". Put a "NOT" qualifier in front of an annotation to inner membrane (so long as the experimental result shows this).
- This qualifier is used when a previously reported result is shown to be incorrect (ie. NOT in inner membrane as shown before), not when there is a negative result reported (ie. Protein A isn't in the inner membrane and has never been reported to be).

Where will each annotation appear?

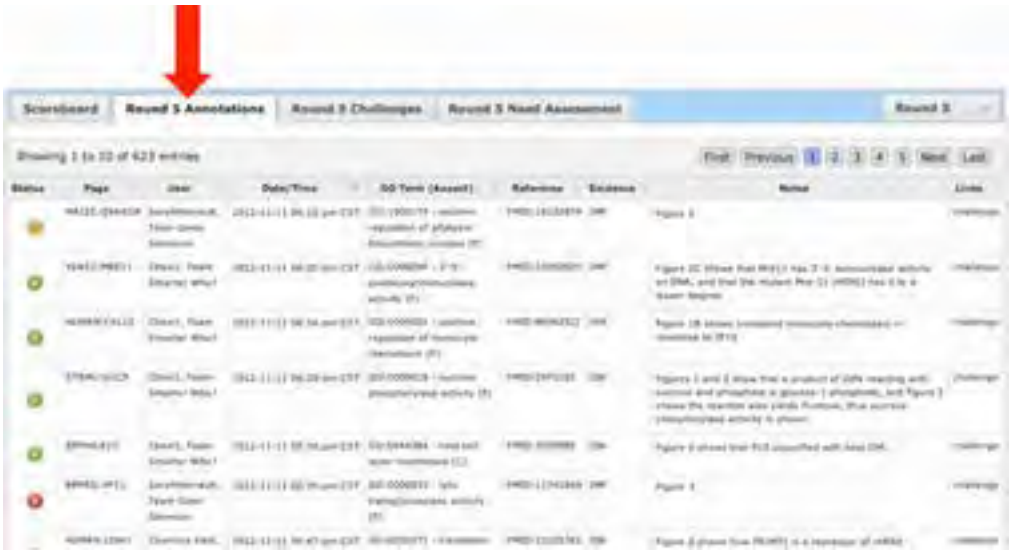
Each annotation made by a team member will show up:

1. In the "Annotation" table on the gene page they added the annotation to.
2. In the table on their user page.
3. In the table on their team page.
4. As 5 points in their team score on the scoreboard page & potentially as a "Submitted Challenge".

What does an annotation look like on a user or team page?

When the annotation (1) is transferred to the user and team page, it will be given a "Timestamp" (2) indicating when the annotation was added to the gene page (3). The user (4) will also be indicated and it will be given a "challenge" (5) link.

Where else is the annotation? The scoreboard under “Round # Annotations” tab.



There are 4 ways to have a correct and complete annotation count towards the total number of annotations required by an instructor:

1. Synthesize a complete and correct annotation.
2. Challenge someone else’s incorrect/incomplete annotation & correct the problems with the annotation entirely.
3. If one of your annotations is challenged, but not fully corrected by the challengers, you can correct the annotation completely for it to count in your total score.
4. Identify term(s) missing in the GO, request the term(s) from the GO Consortium using SourceForge and get the term(s) accepted by the GOC. Each accepted term is equivalent to 1 complete and correct annotation.

Can students make useful GO annotations? Yes.

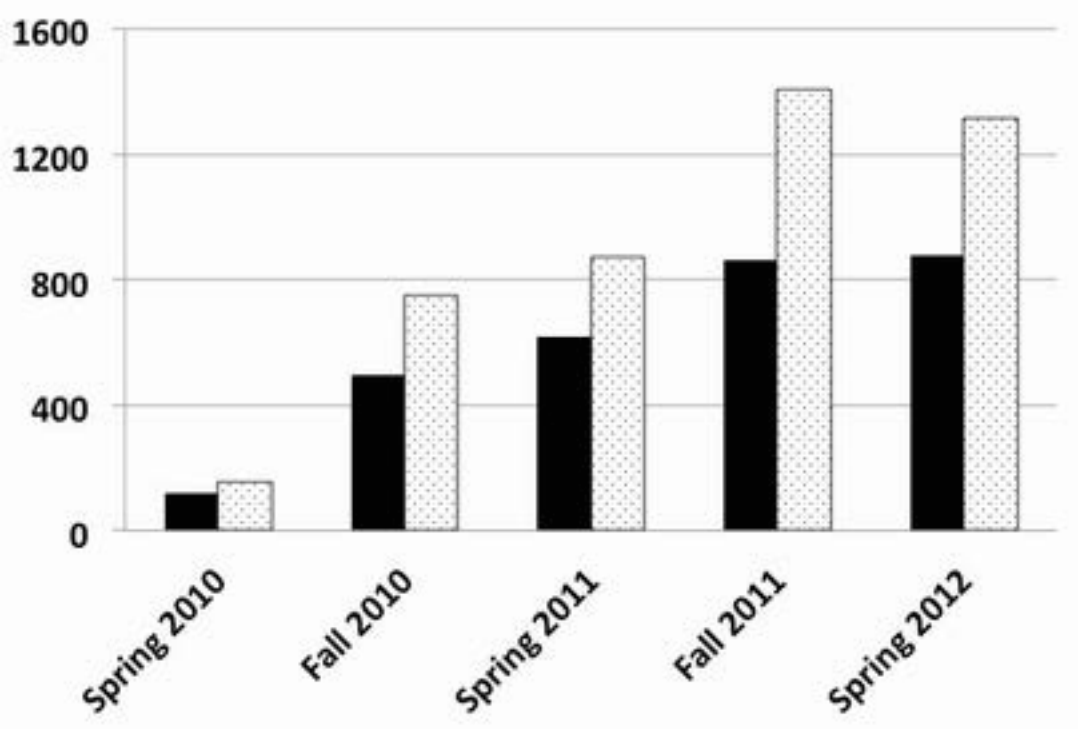


Figure 3. CACAO students contribute many GO annotations per semester. Black bars represent the number of annotations assessed as acceptable or requires changes, while the dappled bars represent the total number of annotations (including those assessed as acceptable, requires changes and unacceptable).

Part III: Challenges

Quick Guide to Annotating

1. Find a suitable article on PubMed about a protein (no reviews, no notes, no Wikipedia articles). Record the PMID number (NOT the PMC number!!)
2. Find the SAME protein in UniProt and get the accession/entry number
3. Use the UniProt accession to make an editable protein page on GONUTS
4. Find a suitable GO term based on figure(s) &/or table(s) characterizing the protein
5. Pick a suitable evidence code based on how protein was characterized in those figure(s) &/or table(s)
6. Enter (and save) the GO annotation on the protein's page in GONUTS, complete with notes (indicating the figure(s) &/or table(s) that support it)
7. **Challenge & refute challenges to team's annotations**

Challenges

Challenging annotations

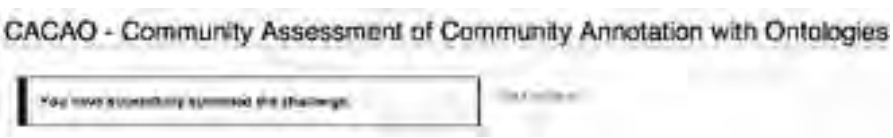
- We allowed students to annotate for 7 days, then challenge for 7 days.
- Each annotation can be challenged multiple times.
- Students cannot challenge their own annotations.

To challenge another competitor's annotation:

1. If a participant decides to challenge another competitor's annotation, click on "challenge" in the appropriate row of either the team's or user's page.

2. Fill in the form for the challenge. A challenger ***must*** fill a reason for the challenge for which a portion of the points for the annotation being awarded to the challenging team if the challenge is accepted by the judges. If the challenger wants additional points, they can include the suitable correction for the error. For the ease of challenging, the annotation they are challenging and so forth is displayed at the top of the page.

3. The challenge will be confirmed when submitted.



Where does the challenge show up once submitted?

Once a challenge has been successfully submitted, it will show up on the same page as the scoreboard in a table called “Round # Challenges”. It is now up to the group that originally authored the annotation to provide a rebuttal to the challenge or any other group that identifies additional problems with the annotation. These can be entered in any round’s challenge period, including the same period that the annotation was first challenged in.

To submit a rebuttal for a challenge:

If an annotation is challenged it will show up under the “Round # Challenges” tab of the scoreboard. It is the responsibility of the teams to monitor their annotations for challenges since we do not have a way to notify the original authors that their annotation has been challenged. Additional challenges or a rebuttal can be entered by clicking on “challenge or judge” under the “Links” heading.

2. They will be offered another challenge form that they can fill in with their rebuttal or challenge.

Part IV: Competition Rules

Competition Rules

1. Completeness of the annotation

a) A fully synthesized annotation includes all fields with required/optional information and a record of the supporting evidence

i) All annotations must at the very least include a GO term, evidence term, reference and a brief description of the evidence in the notes field and may require other components to be considered complete.

ii) A brief description of the evidence means that the annotator must record the figure, table, etc that provides the evidence for the annotation (ie "Figure 2"). It is strongly encouraged that curators add additional information about the experiment/figure/table, in particular for ISS and related evidence codes.

iii) Some annotations may also include a qualifier term as optional information.

- Contributes_to is only used with molecular function terms, never with biological process terms. (If it is annotated as being involved in a process, it is assumed it contributes to it.)
- Colocalizes_with will not be used for CACAO annotations.
- NOT is only appropriate if there is an annotation for this term already on the gene product page that you find evidence to the contrary for or if the gene products is specifically recorded in the literature to have a specific molecular function/biological process/cellular component and you find evidence to the contrary.

iv) The use of certain evidence codes will require the team to enter additional data into the with/from field as optional information.

- IGI, ISA, ISO, ISM must have the with/from field filled in with the database and accession number of the object of similarity/partner. (ie UniProtKB:P01023)
- No annotations will be accepted using the evidence codes EXP:Inferred from Experiment, TAS: Traceable Author Statement, NAS:Non-traceable Author Statement, IC:Inferred by Curator or ND:No Biological Data Available, ISS:Inferred by Sequence Similarity.

vi) All information must be properly formatted for an annotation to be complete and any annotation must be clearly and fully defended in case of a challenge. The team responsible for creating the annotation will be given an opportunity to defend the annotation during the challenge period.

vii) No annotations to binding terms will be accepted (ie. Protein binding, ATP binding, etc)

viii) Annotations from a high throughput method/paper will only be awarded points for the first annotation. Annotators are free to add the other annotations from the high throughput paper, but they will not count towards your team's score (nor will they count for challenge points). Whether a paper is a high throughput paper will be decided by the judges and their decision is final. As a guide, greater than 15 proteins annotated with the same GO term are likely to be considered high throughput data.

ix) Full points (5 points per annotation) will be awarded if the annotation is judged as complete and accurate and is successfully defended if challenged.

b) An incomplete annotation includes any annotation that is missing one or more required/optional fields of information or any information is incorrectly formatted.

i) Missing or incomplete information refers to both required or optional information.

ii) This will primarily be judged by peer review during the competition by the use of challenges. The judges will decide how many points will be awarded to each team (1-5 points for a total of 5 points per annotation).
iii) If unchallenged throughout the competition, an annotation missing a required/optional field will be awarded zero points at the final review.
iv) If unchallenged throughout the competition, an annotation that has a field that is incorrectly formatted will only be awarded 2 points.
v) Judges will have the final decision on completeness.

2. Accuracy of the annotation

a) *All content included in the annotation must be unambiguously accurate.*

i) This includes ensuring the GO term selected is the most appropriate for the evidence cited, the evidence term is appropriate for the methods used, the reference chosen must include the actual data (not a reference to another paper or the abstract), the experiment described provides the evidence for the chosen GO term, the appropriate partner(s) is(are) included in the with/from field (if necessary).
ii) The annotation must be added to the correct gene product page.
iii) If only a part of a figure (ie "Figure 2b") is the evidence for the annotation, this must be correctly identified.
iv) Some annotations may also include a qualifier term. (see above - 1.iii)
v) The use of certain evidence codes will require the team to enter additional data into the with/from field. (see above - 1.iv)
vi) The reference must be a peer-reviewed scientific paper and be in the format PMID:12345678. No annotations will be accepted if reference is a review article, book, news article, conference paper, or blog or other non-reviewed source.
vii) No annotations will be accepted to any binding term (ie ATP binding, protein binding, etc)
viii) If an annotation is challenged on the basis of accuracy, the team responsible for creating the annotation will be given the opportunity to defend the annotation during the challenge period. The accuracy of the annotation must be sufficiently justified upon a challenge.
ix) No annotations will be accepted to any response to terms (ie response to heat, response to drought, response to pH, etc.)

b) *There are one or more errors in the content of the annotation.*

i) This might include selection of a GO term, evidence code or description that is provided in the notes field is not appropriate for the experimental evidence, etc.
ii) If challenged, judges will assign points based on the challenger's reason for challenging, the suggested correction and the rebuttal by the team responsible for the original annotation. (see below - 3. and 4.)
iii) If unchallenged, zero points will be awarded for an inaccurate annotation at the final review.
iv) Expert curators/judges will review the annotations and will have the final decision on accuracy.

3. Defense of a challenge of an annotation constructed by your team

a) *Successful defense of your annotation.*

i) Your team is able to present and defend the description of the evidence and logic for all components of the annotation and ***NO*** changes are made to the annotation as a result.

ii) Full points remain with your team for this annotation.

b) Lose the defense of your annotation.

i) Your team can retain a portion of the points for an annotation if the challenging team identifies a problematic annotation, but does not suggest the most appropriate correction. The number of points awarded to the challenging team will depend on the severity of the problem(s) with the annotation and what problem(s) are identified by the challengers.

ii) Your team will lose all points for an annotation, which are then awarded to the challenging team, if the judges agree with the challenge ***and*** challenger's suggested correction(s).

iii) Expert curators/judges will have the final decision on challenges.

4. Challenge of an annotation contributed by another team

a) Successful challenge of an opponent's annotation

i) The challenge can be on any annotation contributed during the competition by another team. To be awarded any points, your team must identify a problematic annotation.

ii) If a team is caught arbitrarily entering non-substantive challenges, the team will be penalized. The judges will decide the severity of the infraction and will determine the penalty.

iii) Your team can steal a portion of the points for an annotation made by another team if you correctly identify a part of the annotation that is inaccurate or incomplete (as deemed by the expert curators/judges in cases of dispute). The number of points awarded to your team will depend on the severity of the problem(s) with the annotation, what problem(s) your team has identified, and the accuracy or completion of correction proposed.

iv) To be awarded all of the points for a challenge, your team must present a valid challenge ***and*** propose a complete/accurate correction.

v) Expert curators/judges will review the annotations/challenges and will have the final decision on allotment of points.

b) Lose the challenge of an opponent's annotation

i) There is no penalty for challenging and losing a challenge (where no correction is required to the original annotation).

ii) If a team is caught arbitrarily entering non-substantive challenges, your team will be penalized. The judges will decide the severity of the infraction and will determine the penalty.

iii) If your challenge to an annotation, is inaccurate or incomplete (as deemed by the expert curators/judges in cases of dispute), the other team will lose a portion of their points and your team will be awarded a portion of the points for the annotation. The number of points awarded to your team will depend on the severity of the problem(s) with the annotation and what problem(s) your team has identified.

iv) To be awarded full points for the annotation (and to cause the other team to lose these points from their total score), you must rightly challenge a problematic annotation ***and*** suggest the complete/accurate correction.

v) Expert curators/judges will review the annotations/challenges and will have the final decision on allotment of points.

5. Identification of an ontology development site

a) Successful identification of a term lacking in GO

i) If your team correctly identifies a term that is needed in GO, you will be awarded bonus points that will be added to your team score at the end of the competition.
ii) Expert curators/judges will review the terminology proposed and will have the final decision on allotment of points. For example, 1 point for suggestion of a new 'regulation term', up to 5 points for complex ontology request.

There are 4 ways to have a correct and complete annotation count towards the total number of annotations required by an instructor:

1. Synthesize a complete and correct annotation.
2. Challenge someone else's incorrect/incomplete annotation & correct the problems with the annotation entirely.
3. If one of your annotations is challenged, but not fully corrected by the challengers, you can correct the annotation completely for it to count in your total score.
4. Identify term(s) missing in the GO, request the term(s) from the GO Consortium using SourceForge and get the term(s) accepted by the GOC. Each accepted term is equivalent to 1 complete and correct annotation.

Why are certain annotations not accepted for CACAO?

* No annotations will be accepted using:

1. binding terms (ie. GO:0005515 protein binding, GO:0005524 ATP binding, GO:0008144 drug binding, etc.)
    - the explanation for this is that the binding terms are difficult for experienced curators to use consistently and are still creating much discussion in the GO consortium (the group responsible for developing the GO terms). For example, it is not universally accepted by the GO consortium to annotate an ATPase to ATP binding. It is obvious from its function that it binds ATP, so some groups within the GO consider the annotation to an ATPase sufficient and the annotation to ATP binding thereby redundant and incorrect. Meanwhile, other groups say it is not incorrect to annotate the protein as ATP binding, it is merely unnecessary. Also, ATP would be a substrate, which GO does not want to annotate for each protein (this would be REALLY hard for proteases, for example). However, this is complicated by the fact that there are no terms yet for allosteric interactions, in which case the ATP binding might be a regulatory event.
    - As stated, it is a complicated issue and we don't want to have the students confused by this.

2. response to terms (i.e. GO:0009408 response to heat, GO:0009414 response to water deprivation, GO:0009411 response to UV, GO:0009268 response to pH, etc.)
    - the explanation for this is that the these terms are easy to use to circumvent the point of CACAO. In past semesters, students have looked for all the "drought tolerance" genes in plants. Although these may be interesting, it is similar to the binding terms in searchability and scalability.
    - Also, it is difficult to judge these annotations - is the phenotype sufficient to assign the protein as responding to some stimulus? Is it possible that this protein is not involved in the "response", but rather is upregulated by the protein that is actually doing the responding?
3. Certain evidence codes (EXP:Inferred from Experiment, TAS:Traceable Author Statement, NAS:Non-traceable Author Statement, IC:Inferred by Curator, or ND:No Biological Data Available, IEP:Inferred from Expression Pattern)
    - Our expectation is that students should be able to pick a more specific evidence code than EXP:Inferred from Experiment. This evidence code is a catch-all, but it is more informative to decide what kind of experiment was performed and choose a different experimental evidence code.
    - Although an experimental evidence code, the use of IEP:Inferred from Expression Pattern remains controversial in the GO Consortium. It is widely misused to annotate gene expression (where it really should only be used on

protein or RNA product expression) & it is difficult to explain to students what is and isn't acceptable about this evidence code except on a case-by-case basis, which is time consuming for instructors/assessors.

- The other evidence codes that aren't accepted deal with non-experimental evidence, which isn't the focus of CACAO. We are looking to link gene products to their experimental evidence in published scientific literature.
- more information about the evidence codes can be found at:
  http://geneontology.org/GO.evidence.shtml

Part V: Assessment

ASSESSMENT OF CACAO ANNOTATIONS

4 ways to get a correct and complete annotation:

1. Synthesize a complete and correct annotation.
2. Challenge someone else's incorrect/incomplete annotation & correct the problems with the annotation entirely.
3. If one of your annotations is challenged, but not fully corrected by the challengers, you can correct the annotation completely for it to count in your total score.
4. Identify term(s) missing in the GO, request the term(s) from the GO Consortium using SourceForge and get the term(s) accepted by the GOC. Each accepted term is equivalent to 1 complete and correct annotation.

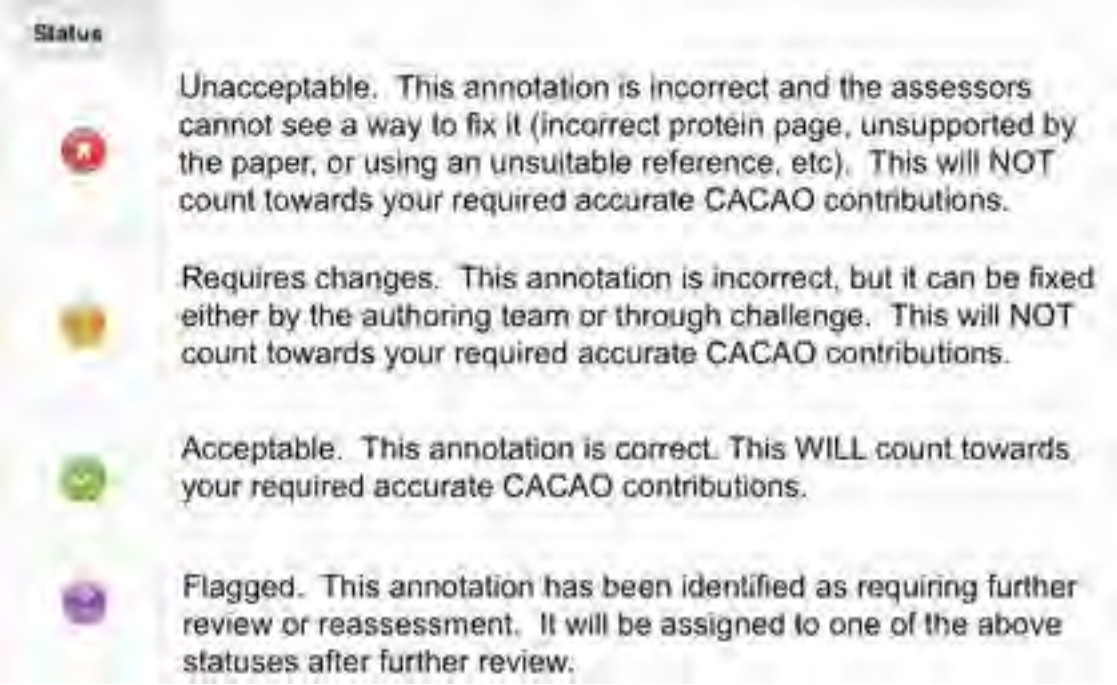
- **Every annotation must be assessed during the competition.**
- **Assessment procedure is described next.**

Assessment of Annotations

- Instructors will be given “judge” status on GONUTS and will be able to judge challenges & enter assessments. Student accounts will not have the ability to judge challenges.
- Assessments are shown in the “Status” column. If there is no icon, it hasn’t been assessed yet. Feel free to assess it!

4 possible assessments for every annotation:

1. Unacceptable
2. Requires changes
3. Acceptable
4. Flagged

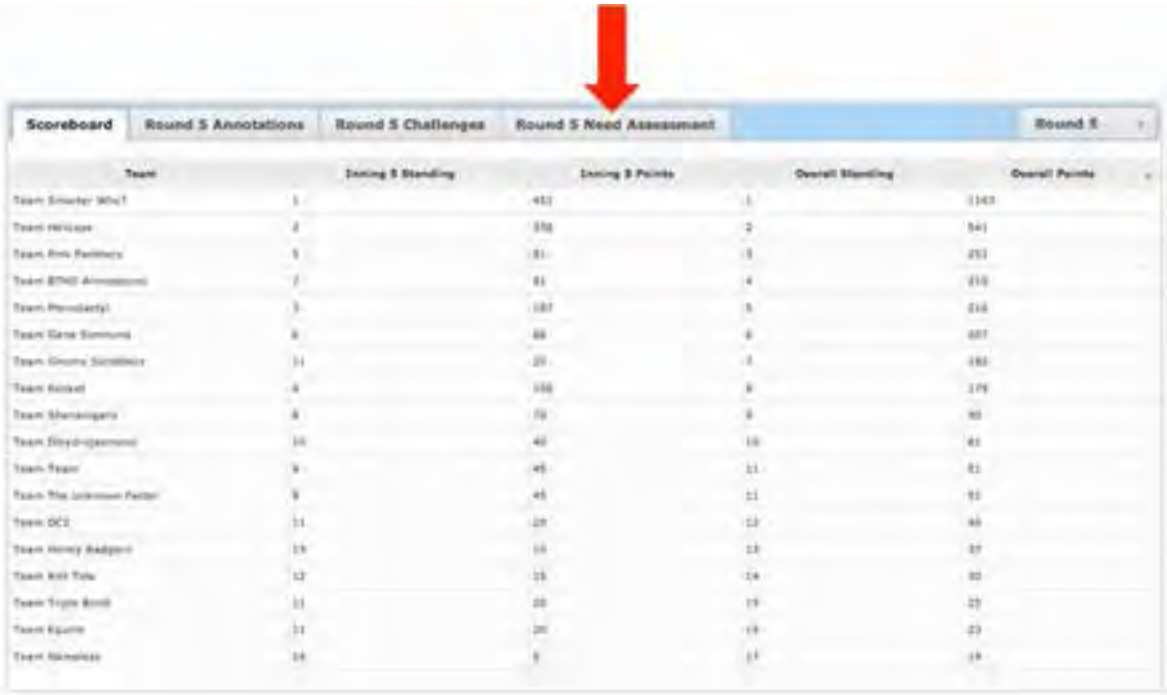


What gets assessed for each annotation?

1. Is the annotation on the right protein's page?
    - identical protein described in paper
2. Is the paper suitable for annotating according to CACAO rules?
    - cannot be a review & it has to have data in it
3. Is the annotation complete?
    - Does it have the 4 required parts?
    - Does it need the optional with/from field?
4. Has the student used information NOT allowed by CACAO rules?
    - illegal evidence code
    - binding terms
    - response to terms
5. Do the notes point to figure(s) &/or table(s) that support the annotation?
    - must point to a figure with sequence alignment, phylogenetic tree, computational model result or experimental data.
    - no crystal structures, models, etc.
6. Is there a more suitable GO term?
    - should it be more or less specific?
7. Does the evidence code fit with the experiment described?
    - IDA, IMP, IGI -- experimental
    - ISA, ISO, ISM or IGC -- computational analysis
8. For IGI, ISO, ISA or ISM, have they entered the correct accession in the with/from field?
    - UniProt accession for IGI, ISO or ISA
    - PMID number for ISM
9. For ISO and ISA, does the protein entered in the with/from have a GO annotation that has experimental evidence for that GO term?
    - does it maintain a DIRECT CHAIN OF EVIDENCE?
10. Is the annotation complete, correct and accurate based on the paper?
    - Would you be willing to submit the annotation to UniProt?

To enter a judgment on an annotation:

1. After the students enter annotations, they will be show up under the "Round # Need Assessment" tab of the scoreboard. If any annotation shows up here, feel free to judge it! Flagged annotations will remain under this tab until resolved.



2. Click on "Challenge" …….....

*** NEEDS A FIGURE***

3. Click on the "New Assessment" tab. It is also possible to enter public comments that the students can see, but the default setting is that the comments will be "Private" that only other instructors with "judge" status can see. This encourages students to continue to refine the annotation without providing the exact answer to the problems.

4. Select an assessment status from the dropdown list and check off the steps that you checked for the annotation (i.e. it is on the right protein page, right GO term, etc).

Judging of the Challenges

- Instructors will be given “judge” status on GONUTS and will be able to judge challenges. Student accounts will not have the ability to judge challenges.

Points for Challenges

- **Points are assigned at the judges’ discretion**
- **Students are not punished for incorrect challenges unless they are challenging for no plausible reason. The primary CACAO instructor reserves the right to take away 1000 points for baseless challenges.**
- Students can get points for only identifying a problem and not suggesting a fix or identifying a problem, but suggesting an unsuitable fix.
- Students can challenge multiple annotations multiple times.
- Students can defend their annotation with their own “challenge to the challenge”, but they cannot challenge their own annotation. If a fix is required, they should have entered it during the annotation period!
- Students responsible for the ***final correction that makes an annotation acceptable*** will be given credit for that annotation.

Identification of Problems

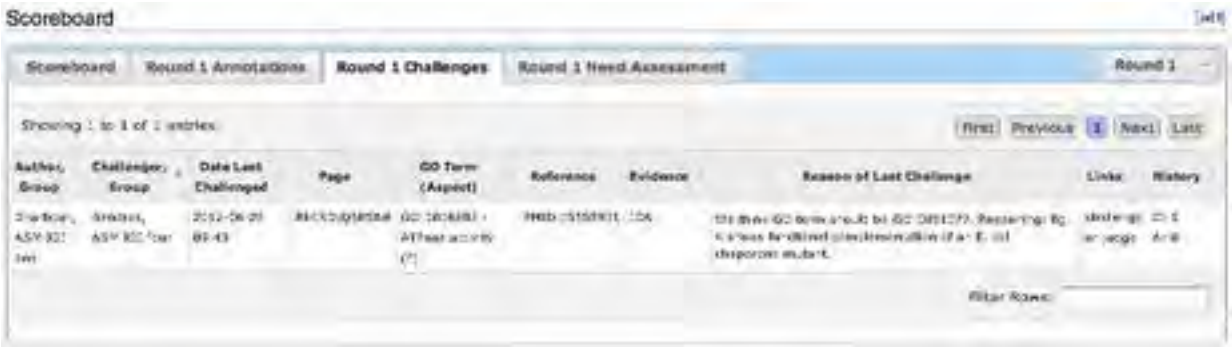
1. Identification of a minor problem = 1 point
    - formatting errors (i.e. entering the PMC number)
    - use of a review article
    - use of an “illegal” evidence code (EXP, IEP, IPI, etc)
    - use of an “illegal” GO term (binding, response to)
    - use of an unnecessary qualifier
2. Identification of multiple problems or one major problem = 2 points
    - evidence codes
    - specificity of GO terms
    - entry in with/from for ISA or ISO annotation doesn’t have an experimental annotation to the same GO term
    - annotation entered on the wrong protein page

Correction of Problems

1. Correction of a minor problem = 1 point
    - how to fix minor error (i.e. converting PMC to PMID, remove qualifier, etc.)
2. Correction of a major problem = 2 to 18 points
    - suggest appropriate fix for any above identified problem (i.e. UniProt id of the correct protein, better GO term, correct evidence code, etc.)

To enter a judgment for a challenge:

1. After the students have challenged and submitted a rebuttal, challenges will be available for judging by clicking on “challenge or judge” under the “Links” heading.



2. First, under the tab “History”, enter the number of points you want to assign for the challenge/rebuttal & click “Save Points” at the bottom right corner of the page.

***** NEEDS A FIGURE*****

3. The challenge will remain in the “Round # Challenges” tab, but will now have an additional number under the “History” heading. “C:1” indicates that the annotation has been challenged once, while the “A:0” indicates the number of assessments by instructors. This will increase by one if you also enter an assessment & the assessment has been saved.

Part VI: Evaluation for Grading

Pre- and Post-Assessment of Understanding

Previously:

- We did pre- and post-assessment of the knowledge and understanding gained by the students using surveys. The surveys were created in a GoogleDoc and emailed to the students on the first and last day of the class. The survey did not count towards their grade, but was meant to give us an idea of what the students knew coming in and what they learned. You will note that the students don't actually have to define or explain a term/topic, but rather have to judge their ability to explain that to a peer.
- The surveys are from our pilot course, which was done with EcoliWiki (http://ecoliwiki.net) not GONUTS (http://gowiki.tamu.edu)
- The surveys are shared as additional GoogleDoc files.

As of Fall 2013:

- We will be doing formal assessments with Dr. Karen Sirum of Bowling Green State University using SALG surveys (salgsite.org).
- Surveys will be pre- and post-completion.

Annotation Requirements (at TAMU)

I require 20 complete, correct annotations for each student by the end of the course. This averages to 5 complete and correct annotations per round. As you see below in the grading section, students get 5 points per annotation and must have at least 100 points from annotating to get an A in CACAO I.

4 ways to get a correct and complete annotation:

1. Synthesize a complete and correct annotation.
2. Challenge someone else's incorrect/incomplete annotation & correct the problems with the annotation ***entirely***.
3. If one of your annotations is challenged, but not fully corrected by the challengers, you can correct the annotation completely for it to count in your total score.
4. Identify term(s) missing in the GO, request the term(s) from the GO Consortium using SourceForge and get the term(s) accepted by the GOC. Each accepted term is equivalent to 1 complete and correct annotation.

Annotation Requirements for Single Rounds

This depends on the instructor preferences, but generally, most instructors require 4-5 complete and correct annotations per round they compete in.

It has been successful to:

- require 4-5 complete & correct annotations
- enter officially in a single round
    - train students live via Skype or using videos on GONUTS
    - give students class time to prepare annotations & challenges
    - experienced CACAO biocurators give students feedback on their annotations
- allow students to compete without supervision in the remaining rounds to fulfill their required number of annotations.

Grading for CACAO I (TAMU)

- We aim for this to be a group activity, so we have emphasized the participation in group work. Our rubrics and peer/self assessments are meant to evaluate this (see below).

Letter graded or pass/fail. Attendance is mandatory.
Grades will be based on:

- Attendance
- Participation in group work
- Annotations
- Challenges

1. The course will be graded on a curve with a median letter grade being somewhere in the B's as indicated in the rubrics below.
2. The synthesis of annotations and challenges is likely to be challenging.
3. There will be no opportunity to earn extra credit by doing extra work.
4. Points distribution:

Group work assessment (self, peer, coaches)	80
Attendance	20*
Annotations and Challenges	unlimited*

Grading Scale

A	200 or more*
B	175-199
C	150-174
D	125-149
F	<124

* Students get 5 points for every successful (complete and correct after assessment by experienced biocurator) annotation or challenge they make, so there is theoretically no upper limit to the high score. The point range for an A is based on what we consider an outstanding effort for the number of credit hours allotted. Because of the competitive nature of CACAO, students often exceed this standard.

Group Work Assessment: 80 Points
Students will be graded on the assessment of his/her own participation as well as by group members and the coaches for his/her participation in group discussions.

Attendance Policy:

Students start with 20 points for attendance. 10 points will be deducted for each unexcused absence. Note that the final attendance score can be a negative number. Attendance policy for this class conforms to student rule 7. See: http://student-rules.tamu.edu/rule07

Annotations & Challenges: 100 Points

Each student will perform in-depth Gene Ontology annotation of genes involved in a biological process of their choice from appropriate organisms. Annotations will be evaluated based on the completeness of the annotations, the appropriate documentation of evidence, contributions to revisions to GO via new term requests or term revision requests, and biological significance. Students may enter refinements and challenges to other students' annotations to clarify, correct or complete the annotation. Challenges will be evaluated based on the logic and difficulty of the challenge.

Rubrics for CACAO I

Rubric #1: Group Work

A - Thorough Understanding	i. Consistently and actively works towards group goals. ii. Is sensitive to the feelings and learning needs of all group members. iii. Willingly accepts and fulfills individual role within the group. iv. Consistently and actively contributes knowledge, opinions and skills. v. Values the knowledge, opinions and skills of all group members and encourages their contribution. vi. Helps group identify necessary changes and encourages group action for change.
B - Good Understanding	i. Works toward group goals without prompting. ii. Accepts and fulfills individual role within the group. iii. Contributes knowledge, opinions and skills without prompting. iv. Shows sensitivity to the feelings of others. v. Willingly participates in needed changes.
C - Satisfactory	i. Works toward group goals with occasional prompting. ii. Contributes to the group with occasional prompting. iii. Shows sensitivity to the feelings of others. iv. Participates in needed changes, with occasional prompting.
D - Needs Improvement	i. Limited understanding. ii. Works toward group goals only when prompted. iii. Contributes to the group only when prompted. iv. Needs occasional reminders to be sensitive to the feelings of others. v. Participates in needed changes when prompted and encouraged.
F - Unacceptable	i. Did not participate in group work.

Rubric #2: Mechanics and Quality of the Annotations & Incorporation of Feedback

A - Excellent	i. Annotations are formatted correctly and are complete. ii. Annotations are made using the first occurrence of evidence. iii. Short descriptions show clear understanding of experimental method/ evidence. iv. GO term selected is the most appropriate for the evidence cited, the evidence term is appropriate for the methods used, the reference chosen includes the actual data (not a reference to another paper or the abstract), the experiment described provides the evidence for the chosen GO term, the optional fields are also filled in correctly (if necessary). v. Upon challenges, annotation is well defended using clear logic and

	appropriate explanations of evidence. vi. Feedback from coaches, judges or instructors is incorporated rapidly and appropriate changes are made to annotations quickly. vii. ***A minimum of 5 annotations per round (for a total of at least 20 over the 5 rounds) are correct and complete***.
B - Very Good	i. Annotations are formatted correctly and are complete. ii. Short descriptions show clear understanding of experimental method/ evidence. iii. GO term selected is not the most appropriate, but is only a single (parent or child) relationship away from the most appropriate term. iv. Upon challenges, an annotation is defended using clear logic, but explanation lacks some detail(s). v. Feedback from coaches, judges or instructors is incorporated rapidly and appropriate changes are made to annotations quickly. vi. ***4 annotations per round (for a total of 15-19 over 5 rounds) are correct and complete.***
C - Satisfactory	i. Annotations are formatted correctly and are complete. ii. Simple explanation of evidence given. iii. GO term selected is not the most appropriate and is more than a single parent/child relationship away from the most appropriate term. iv. Upon challenges, an annotation is defended, but the explanation lacks important details. v. Feedback from coaches, judges or instructors is incorporated and appropriate changes are made to annotations with occasional prompting. vi. ***3 annotations per round (for a total of 10-14 over 5 rounds) are correct and complete.***
D - Needs Improvement	i. Annotations are incorrectly formatted, incomplete or inaccurate. ii. Annotations are made using the wrong evidence code. iii. GO term selected is not relevant to the evidence described. iv. Upon challenges, an annotation is not well defended due to a lack of understanding of evidence or annotation logic. v. Feedback from coaches, judges or instructors is incorporated and appropriate changes are made to annotations when prompted and encouraged. vi. 2***annotations per round (for a total of 5-9 over 5 rounds) are correct and complete.***
F - Unacceptable	i. Annotations were not added to GONUTS. ii. Upon challenges, an annotation is not defended. iii. Did not incorporate feedback from coaches, judges or instructors and appropriate changes are not made to annotations. iv. ***1 or less annotations per round (for a total of 0-4 over 5 rounds) are correct and complete.***

Rubric #3: Knowledge Integration & Application Through Challenges

A - Excellent	i. Challenge was well-organized, comprehensive and persuasive. ii. Demonstrates full knowledge of the annotation and evidence. iii. Presents a logical explanation for the challenge. iv. Presents an easy-to-follow argument that is logical and adequately detailed. v. Presents an excellent (see above rubric) for alternative annotation.
B - Very Good	i. Challenge is well-organized. ii. Presents most of the arguments against the annotation but lacks some details. iii. Presents a very good (see above rubric) for alternative annotation.
C - Satisfactory	i. Challenge is appropriate. ii. Features of the argument lack important detail or does not present sufficient logic. iii. Presents a satisfactory (see above rubric) for alternative annotation.
D - Needs Improvement	i. Challenge is disorganized or illogical. ii. Presents basic background to the annotation, but does not adequately describe the problem to be solved. iii. Presents a needs improvement (see above rubric) for alternative annotation.
F - Unacceptable	i. Challenge is vague, confusing or obviously inappropriate. ii. Argument is poorly contrived or neglects obvious problems with the annotation. iii. Does not present an alternative annotation.

GRADING FOR CACAO II

- Undergrads must have completed at least one semester of CACAO competition.
- Although there are no formal prerequisites for graduate students, students should have a good, solid understanding of Genetics and Molecular Biology. We strongly recommend students to take BICH/GENE631 and/or BIOL650.
- Students will be expected to do independent work to supplement their background knowledge as needed. In addition, we will assume that students are familiar with the basic operational knowledge of computers and the internet.

The course will cover theory and practice of functional annotation of gene products.
After completing this course students will be able to:

- Describe different levels of Genome Annotation from gene models to functional annotation to systems annotation
- Describe the use of ontologies for annotation
- Discuss the nature of gene function
- Describe different systems used for classification of genes and gene products
- Describe automated and manual approaches to annotation
- Compare models for biocuration and the challenges for each model.
- Perform literature-based annotation using Gene Ontology (GO)
- Evaluate the quality of literature-based annotations done by others (peers or students in BICH 460)
- Write a curriculum development section for an NSF CAREER award based on student annotation
- Students who complete this course should be qualified to teach undergraduate annotation courses to their areas of interest, either at TAMU or in their future jobs.

Letter graded.
Grades will be based on:

- Preparation for each lecture and participation in the discussions
- Evaluation of peer and student evaluations
- Grad only: Usage notes for ontology terms
- Grad only: Ontology term requests and Annotation requests to GO Consortium

Points Distribution	

Undergrad		Grad	
Mentoring	100	Mentoring	100
Annotation Evaluation	100	Annotation Evaluation	100
		Extra Activities	100
Total	**200**	**Total**	**300**

Grading Scale			
A	150+	A	250+
B	125-149	B	200-249
C	100-124	C	175-200
D	75-99	D	150-174
F	<75	F	<150

Mentoring: 100 Points
Each student is required to provide written feedback on annotations made by students in the Community Assessment of Community Annotation with Ontologies (CACAO) competition during the first round of participation by any student. Students may join CACAO during any point in the semester and BICH 489/689 students are expected to provide written evaluations for these students throughout the semester. Each student is expected to lead group discussions with students enrolled in CACAO at TAMU.

Attendance Policy:
Students start with 20 points for attendance, which is included in the points for mentoring. 10 points will be deducted for each unexcused absence. Note that the final attendance score can be a negative number. Attendance policy for this class conforms to student rule 7. See: http://student-rules.tamu.edu/rule07

Annotation evaluations: 100 Points
Each student will review annotations of other students in the course, as well as annotations being done in parallel by undergraduates doing GO annotation as part of the Community Assessment of Community Annotation with Ontologies.

Extra Activities (Grad only): 100 Points
Students will be graded on extra activities that reflect deeper understanding and use of the Gene Ontology. This will include usage notes for Gene Ontology added to the GONUTS wiki, term requests to the GO Consortium on the GO Sourceforge tracker, and Annotation requests on the GO Sourceforge tracker. Undergrads are welcome to do these activities, but they are not required.

Part VII: Documentation of the code behind the scoreboard

New features Requested for Spring 2013

Most of these are located here (http://hexamer.tamu.edu/team/wiki/index.php/Cacao_improvements_brainstorming), but for those without an account on hexamer, below is a list.

1. Auto-assess an annotation as ***flagged*** if an annotation is challenged again since being assessed by an experienced biocurator.
2. Check boxes on the assessment form to indicate components that have been verified as either correct or incorrect.
3. Student tracking page with all of the statistics for each student, including the number of annotations with their values (assessments), the number of challenges entered with points assigned & the TOTAL number of annotations each student has credit for.
4. A check box to indicate if a student has fixed the annotation as it appears on the protein's page.
5. The ability to block CACAO students from choosing illegal evidence codes.
6. The ability to check the content entered certain fields before a student saves an annotation to a protein page. i.e. PMID field should ONLY contain numbers.
7. Once an annotation has been challenged, it cannot be changed on the protein page (row is locked from editing).
8. Students should not be able to delete each others' annotations.
9. A way to give credit to a student for developing the GO & have that show up as an annotation on the tracking page.
10. A way to give credit to a student for being the one who COMPLETELY corrected another's annotation & have that show up as an annotation on the tracking page.
11. A "like" or "I support" button for students who spend a bunch of time checking an annotation that they then decide is ok.
12. We still need a system to assign credit for an annotation to a student who completely corrects an annotation through a challenge.

1. Re-flagging

When an annotation has been assessed, it generally gets forgotten. It has been assigned a value (acceptable, requires changes or unacceptable). However, we want to encourage students to repeatedly challenge and refine each others' annotations. This means we need to be able to identify annotations that have been assessed, but then get challenged (or re-challenged).

Daniel Renfro has written code that will automatically check to see if an assessment was the last entry for an annotation before another student entered a challenge and will re-assess the annotation as "FLAGGED" so that we can look at the challenge.

2. Check boxes for assessment

What gets assessed for each annotation?

1. Is the annotation on the right protein's page?
    - identical protein described in paper
2. Is the paper suitable for annotating according to CACAO rules?
    - cannot be a review & it has to have data in it
3. Is the annotation complete?
    - Does it have the 4 required parts?
    - Does it need the optional with/from field?
4. Has the student used information NOT allowed by CACAO rules?
    - illegal evidence code
    - binding terms
    - response to terms
5. Do the notes point to figure(s) &/or table(s) that support the annotation?
    - must point to a figure with sequence alignment, phylogenetic tree, computational model result or experimental data.
    - no crystal structures, models, etc.
6. Is there a more suitable GO term?
    - should it be more or less specific?
7. Does the evidence code fit with the experiment described?
    - IDA, IMP, IGI -- experimental
    - ISA, ISO, ISM or IGC -- computational analysis
8. For IGI, ISO, ISA or ISM, have they entered the correct accession in the with/from field?
    - UniProt accession for IGI, ISO or ISA
    - PMID number for ISM
9. For ISO and ISA, does the protein entered in the with/from have a GO annotation that has experimental evidence for that GO term?
    - does it maintain a DIRECT CHAIN OF EVIDENCE?
10. Is the annotation complete, correct and accurate based on the paper?
    - Would you be willing to submit the annotation to UniProt?

Check Boxes

We will use check boxes to track these when each annotation is assessed by an experienced biocurator.

3. Statistics page

A single page in the wiki that shows all of the statistics for every student in CACAO I.

- total # of *de novo* annotations
    - # acceptable
    - # requires changes
    - # unacceptable
    - # flagged
- total # of challenges
- total # annotations

There would also be a page to track the statistics of CACAO II students.

- # of annotations assessed
- # of challenges judged

4. Check box to track changes on the protein page

Last semester (Fall 2012), we ran into trouble with keeping track of annotations that the students had very nicely changed for us on the protein pages. Some students wanted to incorporate the changes they needed, even though they knew they wouldn't get credit ultimately for the annotation. It is good b/c the students are keen to make their changes, but it is difficult to keep track of annotations that have been changed on the protein page.

5. Block students from choosing an illegal evidence code

This has to do with better column rules & is what Suzi is working on. If students are CACAO users, then they shouldn't get the choice of any other evidence codes other than those permitted.

They should ever only be able to choose:

- IDA
- IMP
- IGI
- ISA
- ISO
- ISM
- IGC

6. Check validity of certain fields before saving

This has to do with better column rules as well.

- PMID field should only ever accept numbers.
    - would NOT accept
        - PMC1111 (PubMed Central #)
        - P1A2Z1 (UniProt ID)
- If student picks IGI, ISA, ISO or ISM -- MUST have the with/from filled in

If a field does not follow certain column rules, GONUTS should not save the annotation, but rather give a warning and return the student to the annotation form they were working on.

7. Row locking upon challenge

We have discovered in past semesters that students will occasionally change the annotation shown on the protein page AFTER the original annotation was challenged. Technically, we want to freeze the challenged state of the annotation so that refinements can be entered via challenges, not by editing the annotation on the protein page.

8. No deletion of other students' annotations

This would be a sneaky way to game the system. We have only heard of this happening to a single annotation over the past 6 semesters, but it is a concern we know about.

9. Credit for GO development equivalent to an annotation

We currently haven't had that many GO terms requested of and accepted by the GO Consortium, so keeping track of giving credit for these hasn't been a problem previously. Nevertheless, students are able to use SourceForge requests to develop the GO & I tell them I will count

10. Credit for final correction equivalent to an annotation

There are 4 ways to have a correct and complete annotation count towards the total number of annotations required by an instructor:

1. Synthesize a complete and correct annotation.
2. **Challenge someone else's incorrect/incomplete annotation & correct the problems with the annotation entirely.**
3. **If one of your annotations is challenged, but not fully corrected by the challengers, you can correct the annotation completely for it to count in your total score.**
4. Identify term(s) missing in the GO, request the term(s) from the GO Consortium using SourceForge and get the term(s) accepted by the GOC. Each accepted term is equivalent to 1 complete and correct annotation.

Somehow, we need a way to keep track of any student who completely corrects an annotation through a challenge.

11. Like button

We would love to have a way for students to “Like” an annotation (similar to a Facebook like).

I have regularly heard about how students have spent hours looking at an annotation, only to decide it is perfectly fine. They would like to be able to show their effort and support of another student’s annotation by a “Like”.

The Meat of the CACAO code follows...

General Definitions

User (aka Participant)

- Each participant in the cacao class can either be a *student* or an *instructor*. A participant is defined as anyone who has the `<cacao>` parser-tag on their User page, and/or has the `cacao_instructor` permission as defined by `$wgCacaoInstructor`

Instructor

- The difference between a participant and an instructor is simply the assignment of the `$wgCacaoInstructor` permission. This can be set in [Special:Userrights] page in the appropriate wiki. See also: mediawikiwiki:Manual:User_rights.

Group (aka Team)

- A group is a set of participants organized into a Category using Mediawiki's category system. Typically a group represents a *team* of users and has a parser-tag on the page to show the team's annotations.
- A team can have any number of participants, but we usually have 2-3 students per team.

Session (aka Competition)

- A *global session* is the highest level of organization of an instance of CACAO. It contains any number of groups, which can challenge each other.
- Participants and groups cannot challenge groups in other sessions.

Inning/Rounds

- An *inning* is a temporal designation for a part of CACAO. Typically a CACAO class has multiple innings, letting the students complete iteratively in rounds rather than in one large competition throughout the semester. This also facilitates universities that only want to participate/compete during a subset of the semester.
- All innings must consist of at least one *period.*
- Innings are defined in LocalSettings using the `$wgCacaoGlobalSessions` variable.

Period

- A *period* is a subset of an inning. Innings usually have a number of periods which define what types of actions participants (mostly non-instructors) can do. Examples include:
    1. annotation
    2. challenge
    3. closed
- Periods are defined in LocalSettings using the `$wgCacaoGlobalSessions` variable. See the **Scheduler** below.

Row

- A *row* is typically an instance of class `wikiBoxRow` from TableEdit.

- In GONUTS it almost always refers to a row in the `Annotation_Headings` table -- the annotation table found on gene-pages where GO annotations are made. Row objects are an important part of the cacao infrastructure.

Annotation

- An *annotation* is an instance of class `CacaoModelAnnotation` and is the unit of work in CACAO.
- *Annotations* are not specifically the same as *rows*; annotations contain a row object along with other information such as associated challenges, timestamps, *etc.*
- Points for annotations are based on a basal amount defined by the global variable `$wgCacaoPointsForAnnotation`.

Challenge

- A challenge is a declaration or an objection to the precision or the legitamacy of an annotation. The challenging team must provide a logical reason for the challenge, and instructors can assign any amount of points to a challenge.
- Participants are not allowed to challenge their own annotations or annotations of their group, unless a previous challenge has been submitted by another team. In this way *challenges* are more like iterative *refinements* of an annotation and less like single objections.
- An annotation can have any number of challenges by any/all teams.

Open Challenge

- An *open challenge* is terminology reminiscent of early CACAO semesters. It is defined here as any annotation for which there is a challenge which does not have points assigned. If an annotation has "open challenges" then it is necessary for an instructor to adjudicate those challenges.

Assessment

- An assessment is an appraisal of an annotation. No points are assigned for assessments. Each assessment consists of a description and the actual *assessment*, which is a defined list containing terms like "acceptable", "unacceptable", "requires changes", *etc.*
- Assessments are necessary for determining if an annotation is good enough to go into the `gene_association` file after CACAO is done.

Judgement

- The act of assigning points to challenges (as well as typically entering as assessment.) At this point an annotation can be reviewed for inclusion into the `gene_association` page.

Scoreboard

The scoreboard is an ajax-driven parser-tag that is usually found on session pages. The scoreboard evolved from a simple algorithm to determine the points of each team to a set

of tabs that shows annotations/challenges/assessments for each round. The scoreboard has a number of classes and a significant amount of code associated with it.

Code-Specific Definitions

Controller

- The controller in the CACAO MVC is simply a SpecialPage. After the request is handed off from Mediawiki to class `CacaoSpecialMain`, it creates an appropriate model, updates it, creates an appropriate view, and then tells the view to render.

Status

- A small class (`CacaoStatus`) that helps identify problems and exceptions in the software. Each model returns a status object after running its `update()` method.

View

- A piece of code that creates a user interface (in HTML.)
- In the `cacao` code there are two levels of views: regular views and sub-views (explained below.)

SubView

- A *subview* is a small class that creates HTML for a specific purpose. Many tables and forms are implemented as subviews to compartmentalize the code and keep things orderly.
- This means that you will need to create an instance of a subview and give it data before calling its `render()` method.

Model

- The *model*s in the cacao software represent concepts such as: annotations (`CacaoModelAnnotation`), challenges (`CacaoModelChallenge`), and points assigned to a challenge (`CacaoModelPoints`.)

Module

- The term *module* in the code most always refers to the Resource-Loader module. This *module* is defined early as a multidimensional array and contains information (paths, dependencies, *etc.*) used by the ResourceLoader to load JavaScript and CSS for `cacao`.
- The entire module can be added to a page easily using the `$wgOut->addModules()` method.

Scheduler

- The *scheduler* is a singleton of the class `CacaoScheduler`. It contains all the temporal-based criteria for actions in CACAO.
- For example: Is a user without instructor permissions allowed to annotate during a 'challenge' period? (answer: no)

Dates in the competition

Datetimes

- All dates and timestamps are represented as Datetime objects in PHP. In the MySQL database they are represented as VARCHAR(14) like this: YYYYMMDDHHIIMMSS.
- See also SQL Timestamps.
- You can get a formatted representation from a Datetime object by using it's format() function like so:

```
$format = 'YmdHis';
$datetime->format( $format );
```

Variables

These variables are placed within LocalSettings.php. It does not matter whether they are defined before or after the CACAO code is included.

wgCacaoBagOStuff

- This is an instance of HashBagOStuff. It is a temporary (gets removed at clean-up/exit) storage place for values that won't change during this request. For example -- names and groups of users. Expensive (time-consuming) calculations can be done once, cached temporarily, and then looked-up when needed.

wgCacaoDebug

- When set to true (or any expression that evaluates true,) the scheduling code is turned off - meaning that anyone can preform any action at any time. This is typically for debugging when we don't want to check the $wgCacaoGlobalSessions array.

wgCacaoGlobalSessions

- A large array defining the global sessions and their innings. Each session can have multiple innings, and each inning can have multiple (mutually exclusive) periods. Each period must contain at least a start and an end. See below for an example:

```
/**
 * Global session assignment. Session-names need to be in the following format: "Category:_______".
 *  It's best to just copy-and-paste the appropriate part of the URL from your browser into this
 *  array of settings.
 */
$wgCacaoGlobalSessions = array(

    // Older cacao sessions; defined here so that their <nowiki><scoreboard /></nowiki>s still show/render
    'World_Series_2011_CACAO'          => array(),
    'Category:CACAO_Fall_2010'         => array(),
    'Category:CACAO Spring 2011'   => array(),
    'Category:Penn State CACAO'        => array(),

    // Current cacao sessions
    'Category:CACAO Fall 2011' => array(
        array(
            'annotation'   => array( 'start' => '2011:09:19 00:00:00', 'end' => '2011:09:25 23:59:59' ),
            'challenge'    => array( 'start' => '2011:09:26 00:00:00', 'end' => '2011:10:02 23:59:59' ),
        ),
        array(
            'annotation'   => array( 'start' => '2011:10:03 00:00:00', 'end' => '2011:10:09 23:59:59' ),
            'challenge'    => array( 'start' => '2011:10:10 00:00:00', 'end' => '2011:10:16 23:59:59' ),
        ),
        array(
            'annotation'   => array( 'start' => '2011:10:17 00:00:00', 'end' => '2011:10:23 23:59:59' ),
            'challenge'    => array( 'start' => '2011:10:24 00:00:00', 'end' => '2011:10:30 23:59:59' ),
        ),
        array(
            'annotation'   => array( 'start' => '2011:10:31 00:00:00', 'end' => '2011:11:06 23:59:59' ),
            'challenge'    => array( 'start' => '2011:11:07 00:00:00', 'end' => '2011:11:13 23:59:59' ),
```

```
            ),
            array(
                'annotation'    => array( 'start' => '2011:11:14 00:00:00', 'end' => '2011:11:20 23:59:59' ),
                'challenge'     => array( 'start' => '2011:11:21 00:00:00', 'end' => '2011:11:27 23:59:59' ),
            ),
            array(
                'closed'              => array( 'start' => '2011:11:28 00:00:00' )
            )
    )
);
```

wgCacaoInstructor

- This variable defines the *name* of the role of the CACAO instructor. Using this information, CACAO sets up permissions for annotation, challenging, and assessing. An example:

```
$wgCacaoInstructor = 'cacao_instructor';
$wgGroupPermissions[$wgCacaoInstructor]['*'] = true;
```

- This variable has a default setting of 'cacao_instructor'. It's suggest that the value for this variable *not contain any space.*

wgCacaoModules

- Used by the ResourceLoader and the efCacao_RegisterModules function to register what JS and CSS are needed.
- Looks something like this:

```
$wgCacaoModules = array(
    CACAO_MODULE => array(
        'scripts' =>  array(
            'js/ext.cacao.js',
            'js/jquery-ui-1.8.16.custom.min.js',
      'js/jquery.expander.min.js',
      'js/jquery.ui.selectmenu.js'
        ),
        'styles' => array(
            'css/ext.cacao.css',
            'css/jquery-ui-1.8.16.custom.css',
      'css/ui.expandable.css',
      'css/jquery.ui.selectmenu.css'
        ),
    'dependencies' => array(
      'jquery.tipsy',
            'ext.datatables'
        ),
    'messages' => array( ),
    'localBasePath' => dirname( __FILE__ ),
    'remoteExtPath' => 'Cacao',
    'group' => 'Cacao'
    )
);
```

wgCacaoPointsForAnnotation

- The number of points a team is awarded for making a single annotation - the value should be an integer. The default is 5 points.

wgCacaoScheduler

- A global object that handles all the time-based constraints for Cacao. It reads the wgCacaoGlobalSessions array and has some methods for determining current inning/round, and also the*permissions* for different types of users at different times.

wgCacaoTableTemplate

- This variable defines the TableEdit template for the annotation table.

wgCacaoScoreboardAlreadyRendered

- Used only once to make sure that we only render one scoreboard per page -- otherwise the JavaScript won't work and the wiki will probably die (from too many heavy-requests.)

wgCacaoTagAlreadyRendered

- Similar to above. Used only once to make sure that we only render one (type of) parser-tag per page -- otherwise the JavaScript won't work.

Parser Tags

The CACAO software requires the use of "parser tags" in some pages. These tags are similar to HTML or XML tags, and in general look like: *>tag_name /<*. These tags are placeholders for more complicated content that usually needs to be created dynamically. The <myAnnotations /> tag was used in previous versions of CACAO to show a user or group's annotations. This tag has been deprecated in favor of the <cacao /> tag. It will still work, however, and will behave exactly as the <cacao /> tag does.

For example, putting the <cacao /> tag within a user's Userpage will show his/her annotations for CACAO to show up once the page is saved.

Below is a list of the available tags for CACAO.

cacao (replaced the older code called myAnnotations)

- The <cacao /> tag will show the annotations for a particular user or group, depending on which type of page it is on. If it is on a page in the User namespace, it will show all annotations for that user, regardless of which group that user is currently in. If it is on a page in the Category namespace, it will show all annotations that were made for that group -- even if those users are not currently in that category.
- It is important that the members of the group are included in the Category of their group so that their annotations show up using the <cacao /&gt tag. For example, if a user specifies his/her group as "Foo", but does not have the associated wikitext [[Category:Foo]] in their page (which adds them to the group), then their annotations will not show up on the Category:Foo page. The <cacao /&gt tag will show a message saying that they are not currently in the group, but won't fix the problem.
- Usage:

```
<cacao>
group   = Category:Foo
session = Category:Bar
</cacao>
```

group (aka Team)

- The group that this user is in. It can be specified with or without the "Category:" prefix (as all groups are categories.)

session

- The global session that this user is in. It can be specified with or without the "Category:" prefix (as all session are categories containing multiple groups.)

scoreboard

- The `<scoreboard />` tag shows the current standing of all groups participating in a global session of CACAO. This tag is typically found on the Category page that contains all the groups for a class. Given no parameters, the tag assumes that it is on a global session page and tries to display a scoreboard for that session. Alternatively, you can specify a session using the `session=`parameter like so:

```
<scoreboard>
session = Category:Foobar
</scoreboard>
```

- In this case it will display a scoreboard for the `Foobar` CACAO session.
- The scoreboard is a parser-tag, meaning that the parser will interpret the <scoreboard> ... </scoreboard> tags and render it (the parser will generate some HTML and insert it into the page.) This is the result of the `execute()` function.
- Once the initial HTML is created and the page is sent to the user, their browser will combine this with the JavaScript and CSS from the ResourceLoader. The JavaScript is written to make an ajax call to the wiki (actually to it's API) -- which is handled by the CacaoScoreboardAPI class. That class is essentially a wrapper for this class, which also contains the code to answer the ajax call and create the HTML for the scoreboard table itself. (The API class contains a lot of other code, too.)
- The renderScoreboard() method is what actually does most of the work. Although most of the the code for creating the scoreboard is procedural, it has been split into separate methods for easy/easier reading.

SpecialPage(s)

Cacao (Main)

- Class CacaoSpecialMain (defined in includes/CacaoSpecial/CacaoSpecialMain.php) is the controller for the entire Cacao extension. It is quite short -- the main use being to delegate responsibility to other parts of the software.

Admin

- The admin special page is for keeping track of the CACAO 2 students. It shows the number of atomic assessments and the number of annotations assessed per round for users with the wgCacaoInstructor permission.

Scripts

The scripts live in the `scripts/` subdirectory and can be invoked using the PHP interpreter like any other PHP script. Most of them require a `--wiki/-w` flag. Run the script without any arguments (or with the `--help` flag) for usage.

Cacao.php

- All configuration happens in the `cacao.php` file. There is a lot going on there; things like:
    1. checking to see if Mediawiki is loaded
    2. checking to see if our (Hu Lab) code library is loaded
        - this abstracts a lot of common functionality we use for extensions
    3. set up global default variables
    4. define global constants
    5. add credits to Mediawiki
    6. add things to the $wgAutloadClasses global
        - this is built-in to Mediawiki and loads classes on-demand
    7. register special page
    8. initialize the parser-tags
    9. hook into Mediawiki
    10. register language and alias files (see [http://www.mediawiki.org/wiki/Manual:Special_pages#The_Messages.2FInternationalization_File this page on MW)
    11. define global level functions

changed_annotations.php

- This script finds annotations that have been changed in some way (edited or challenged) since they were last assessed. It will print some details about the annotations. Most useful when piped to a file and read later.

detect_duplicates.php

- This script looks for duplicates in all the annotations for a given session. It will find annotations in the CACAO database with the same row-data, regardless of what protein-page their are on.

global_stats.php

- This script gives some generic statistics about a global-session/semester.

```
% php ./global_stats.php --wiki ~daniel/Sites/gonuts --session "CACAO_Spring_2012"
Total number of...
  annotations                1176
    perfect                  320     27.21%
    acceptable               18      1.53%
    unacceptable             393     33.42%
  challenges                 489
```

assessments	338
users/students	167
teams	73

make_groups.php

- This script will create the appropriate accounts and pages on GONUTS for a new CACAO session, and it will do so *correctly.* This is the **preferred way to create a new session**.
- Create a tab-delimited file containing the following fields and set it on the command-line using the `--file/-f` flag.
    1. username
    2. last name
    3. first name
    4. email
    5. group

remove_unacceptable_annotations.php

Removes rows in the "*Annotations_Headings*" table on protein pages for annotations that were marked as "unacceptable." This is **irreversible**.

retrofit.php

Retroactively assigns annotations to a user based on times. This script is a front-end for doing things in the database that are hard to do by hand (because there are lot of foreign keys.) Give the script a username, group, and session to assign annotations to, along with a start and end time.

```
--wiki      -w      Path to the Mediawiki installation.
--username  -u      The user to use.
--group     -g      The group to use when assigning annotations.
--session   -s      The session to use when assigning annotations.
--start     -S      Find annotations after this timestamp - formatted as YYYYMMDDHHMMSS.
--end       -e      Find anontations before this timestamp - formatted as YYYMMDDHHMMSS.
```

SQL

The SQL file for Cacao is contains 6 CREATE TABLE definitions and 2 CREATE VIEW statements. The older data-store for the Cacao extension was just a single table. This proved inadequate for having a more-complicated class -- such as multiple challenges per annotation. The SQL file (see below) has many comments which try and explain the use of each particular column/field.

Conventions Used

I've skipped using the MW convention of adding /*_*/ and /*i*/ prefixes before table names and indexes, because this SQL isn't handled directly from MW. Many of the field names have backticks around them because many of them are reserved words in MySQL. Indexes were added after the CREATE TABLE directives to make things more clear. See the Mediawiki conventions at:http://www.mediawiki.org/wiki/Manual:Coding_conventions#MySQL

Page Titles

- When storing a page title, use Title Object's getFullText() method, this gives both the namespace prefix as well as the page's title. Most of the things here reference the page_id anyway.

Timestamps

- The MySQL table backend for MediaWiki currently uses 14-character VARCHAR fields to store timestamps. The format is YYYYMMDDHHMMSS, which is derived from the text format of MySQL's TIMESTAMP fields (but MUCH easier to read.)

Foreign Keys

- FOREIGN KEY constraints have been forgotten about because I tried and tried to get them to work, but it was to no avail. I kept getting something like "Can't create table 'GO_dev_wikidb.#sql-22d_13a7b7' (errno: 150)" and it was really frustrating. Also, see http://verysimple.com/2006/10/22/mysql-error-number-1005-cant-create-table-mydbsql-328_45frm-errno-150/ about this error.

Tables

cacao_annotation

- The table that holds information about the annotation. Data is entered into this table by the function efCacao_AddAnnotation, defined in Cacao.php.
- The function efCacao_DeleteAnnotation in the same file is used to remove annotations from the table when a row gets deleted. The model for this table is class CacaoModelAnnotation.
- **Foreign Keys:**

```
FOREIGN KEY ( row_id ) REFERENCES [[TableEdit|ext_TableEdit_row]]( row_id )
FOREIGN KEY ( annotation_user ) REFERENCES [[#cacao_user|cacao_user]]( id )
```

cacao_challenge

- This is the table that holds all the challenge information.
- **Foreign Keys:**

```
FOREIGN KEY ( challenge_annotation ) REFERENCES [[#cacao_annotation|cacao_annotation]] ( annotation_id )
FOREIGN KEY ( challenge_user ) REFERENCES [[#cacao_user|cacao_user]] ( user_id )
FOREIGN KEY ( challenge_text ) REFERENCES [[#cacao_text|cacao_text]]( text_id )
```

cacao_assessment

- All the assessment data.
- **Foreign Keys:**

```
FOREIGN KEY ( assessment_annotation ) REFERENCES [[#cacao_annotation|cacao_annotation]](annotation_id)
FOREIGN KEY ( assessment_text ) REFERENCES [[#cacao_text|cacao_text]]( text_id )
FOREIGN KEY ( assessment_client ) REFERENCES [[#cacao_user|cacao_user]]( user_id )
```

cacao_points

- **Foreign Keys:**

```
FOREIGN KEY ( points_challenge ) REFERENCES [[#cacao_challenge|cacao_challenge]]( challenge_id )
FOREIGN KEY ( points_user ) REFERENCES [[#cacao_user|cacao_user]]( user_id )
```

cacao_user

- The user table -- keeps information about the user, what team they are currently on, and what session they are in. A major resource-hog for the software has been calculating this type of data; this table is an effort to keep the costs low.
- Whenever a user saves their user-page, the efCacao_ArticleSave function gets run. This function checks for a valid cacao parser tag on the user-page and updates the data in this database table if necessary. This cuts down on the number of times we have to calculate group/session info.
- **Foreign Keys:**

```
FOREIGN KEY ( user_id ) REFERENCES USER( user_id )
FOREIGN KEY ( team_id ) REFERENCES page( page_id )
FOREIGN KEY ( SESSION ) REFERENCES page( page_id )
```

cacao_text

- A simple table consisting of just an id and a text field (free-text.) This is used to store all the text data for the challenges and assessments. Both of those tables reference this one.

Views

cacao_open_challenges_view

- Define a view for all the "open" challenges.
- That is, any annotation for which there is one or more challenges that do not have points associated with them. This assumes a 1:1 ratio between the challenges and the points, meaning that each challenge can only be scored once.

- The session_id is included here so we can use it in a WHERE clause like so: SELECT * FROM cacao_open_challenges WHERE session_id = 7700066

```
CREATE VIEW cacao_open_challenges_view AS
 SELECT annotation_id,
   session_id,
   COUNT(challenge_id) AS num_challenge_ids,
   COUNT(points_id) AS num_point_ids
 FROM cacao_annotation
   INNER JOIN cacao_challenge ON (challenge_annotation = annotation_id)
   LEFT JOIN cacao_points ON (points_challenge = challenge_id)
   INNER JOIN cacao_user ON (cacao_user.id = annotation_user)
 GROUP BY annotation_id;
```

cacao_user_membership_view

- This view is just for us admins to look at when trying to debug users/groups. Many times I want to see who is/was in what group and not have to do the joins...this is that view.

```
CREATE VIEW cacao_user_membership_view AS
 SELECT user_id,
   user_real_name,
   user_name,
   user_email,
   team.page_title AS team,
   session.page_title AS session,
   user_timestamp
 FROM cacao_user
   INNER JOIN page AS team ON (team_id = team.page_id)
   INNER JOIN page AS session ON (session_id = session.page_id )
   INNER JOIN user USING (user_id);
```

Setup

Requirements

Because the CACAO system was developed in-house for the Hu Laboratory at Texas A&M University, it requires a number of MediaWiki extensions and code libraries developed by our lab.

extension

1. TableEdit
2. DataTables

libraries

1. The Hu-Lab code/ library, an extension of the MediaWiki framework for biological wikis.

LocalSettings.php

- We'll need to make sure that the appropriate things are setup in LocalSettings.php:

```
// global extensions path
  $wgExtensionsPath = 'extensions/';

// Hu Laboratory Code
  require_once( '/Volumes/pentamer/shared/code/trunk/library/Setup.php' );
```

- see also Code Library
- Then add the cacao code:

```
// Community Assessement of Community Annotation with Ontologies
  require_once( $wgExtensionsPath . 'cacao/cacao.php' );
```

Supplementary Material 9

Step-by-step transfer annotations in CACAO

A brief guide to GO annotation using the CACAO interface

This is meant as a brief introduction to GO annotations. For a fully featured, well-written introduction that will address most of the doubts you may have after reading this, please see Balakrishnan et al. (2013) *Databases* "A guide to best practices for Gene Ontology (GO) manual annotation" [PMID: 23842463].

GO annotation basics

In this CACAO lab unit we will be making Gene Ontology (GO) annotations of gene products in a genome of interest. A GO annotation consists in establishing a link between a gene product (e.g. the Bacillus phage Troll "Tail assembly chaperone"; UniProt accession S5YQ92) and a GO term describing a specific aspect of its biology. In GO, we distinguish among three major biological components for a gene product: molecular function, biological process and cellular location. Hence, a GO annotation links a gene accession number to a GO term in any of these categories. GO is an ontology, meaning that GO terms are linked by familial relationships (e.g. "sequence specific DNA binding" GO:0043565 being a *child* of "DNA binding" GO:0003677).

Here is a brief summary from the GO Consortium site (http://geneontology.org/) on what the three biological components are meant to indicate:

Cellular Component
These terms describe a component of a cell that is part of a larger object, such as an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

Biological Process
A biological process term describes 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". Examples of more specific terms are "pyrimidine metabolic process" or "alpha-glucoside transport". 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 steps. A biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

Molecular Function
Molecular function terms describes 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 the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are "catalytic activity" and "transporter activity"; examples of narrower functional terms are "adenylate cyclase activity" or "Toll receptor binding". It is easy to confuse a gene product name with its molecular function; for that reason GO molecular functions are often appended with the word "activity".

Regular and "transfer" GO annotations

In a "regular" GO annotation, biocurators identify a peer-reviewed article where experimental evidence for the molecular function, biological process and/or cellular component of one or several genes is provided. Reading the paper, biocurators then make assertions on, say, gene `X` having molecular function `Y`, where `X` is an accession number for the gene and `Y` is a GO term.

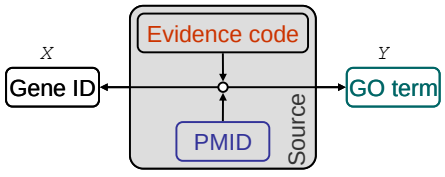


Figure 1 – Schematic diagram for "regular" GO annotations. A gene (`X`) is annotated as having GO term (`Y`), which specifies a well-defined function/process/component, using an evidence code and the PubMed ID (PMID) of a scientific article as the *source* for the annotation.

In doing so, biocurators identify the *source* of the annotation in the following way: (1) they cite the original paper with the evidence (providing its PubMed ID number) as the *reference* for the annotation, and (2) identify an appropriate *evidence code* term to summarize the type of evidence that is provided in the paper to warrant such assertion. For instance, if a study shows that protein X is part of the ribosome through immunoflueorescence techniques, a curator can use the GO evidence code *Inferred from Direct Assay (IDA)* to annotate protein X to the GO term GO:0005840. The following page provides a list of all the possible GO evidence codes you can use in a GO annotation: http://geneontology.org/page/guide-go-evidence-codes. See here for the evidence codes that you are authorized to use in CACAO annotations (http://gowiki.tamu.edu/wiki/index.php/evidence_codes)[1].

Transfer annotations

In this lab unit we may deal with an organism the sequence of which has just been recently sequenced. This means that no experimental work has been done on our organism of interest and, therefore, we cannot perform "regular" GO annotations (there are not scientific manuscripts to annotate from). Instead, what we seek to do is to *transfer* annotations from another organism in which there is experimental evidence for the annotation. The way this is most frequently done is through homology. Remember that two genes are homologous if they share similarity due to shared ancestry. Using sequence and structure search methods (such as BLAST or HHPred) we can establish that two sequences are similar. Using appropriate thresholds (listed here) and our own judgment, we can use the observed similarity to postulate homology. Once we postulate that two genes are homologous, we can make use of our knowledge of the underlying biology to decide if functional annotations made on one gene should transfer to the other or not[2].

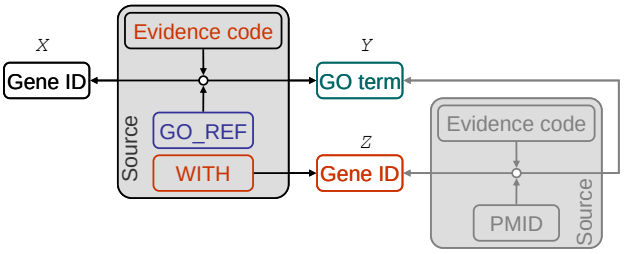


Figure 2 – Schematic diagram for "transfer" GO annotations. A gene (X) is annotated as having GO term (Y), by establishing that gene X is homologous to gene Z, where the annotated function/process/component (Y) has been established through experimental means. The source for the annotation is now the general protocol (*GO_REF*) and the specific method (evidence code) used for establishing homology between X and Z, as well as the identifier for Z in the *WITH* field.

[1] Certain evidence codes (and some types of annotations) are disabled in CACAO. CACAO is normally run as a competition where the number (and quality) of annotations determines the winning team. To avoid contestants submitting many weak annotations based on papers using high-throughput methods (e.g. protein-protein interaction yeast-to-hybrid assays) to score points, evidence codes such as Inferred from Physical Interaction (IPI) or Inferred from Expression Pattern (IEP) have been disabled. If you find that you need to use such codes, please contact your instructor.

[2] In some cases, the transfer makes complete sense, in other cases, no sense at all. For instance, a yeast protein can be annotated as being localized to the nucleus (cellullar component), but that annotation makes no sense on a bacterial homolog of the protein. Whenever you are transferring annotations between markedly different species you should attempt to explain the possible role of the protein in the recipient's biology.

Our annotation process is now, therefore, slightly different from the regular case. In the case of "transfer" annotations, we will be stating that gene X has function/process/component Y, based on its similarity with another gene (Z), for which that function/process/component has been annotated using experimental evidence.

The *source* for our annotation is therefore different than that of "regular" annotations and the evidence codes we will use are also different. To make these assertions, we use evidence codes such as Inferred from Sequence Orthology (ISO) or Inferred from Genomic Context (IGC). Because it is us, and not the authors of a paper, who are making the claim that the annotation of gene Z should be transferred to X, the source does not cite a scientific article, but rather a specialized *GO reference* (GO_REF) that describes the general procedure used by the biocurator to determine the correspondence. We will use the CACAO GO_REF (GO_REF:0000112; Gene Ontology annotation by CACAO biocurators), the description of which you can find here (https://gowiki.tamu.edu/wiki/index.php/CACAO_GO_REF). Crucially, the *source* of the "transfer" annotation includes also another element, defined by the *WITH* field of GO annotations. This is the identifier for the gene (Z) that we are transferring the annotation from.

The "transfer" annotation process

In "regular" GO annotations, biocurators typically start with a paper and then look for experimental evidence of function/process/component for one or more genes in the paper, then proceed to annotate these. The situation in transfer annotations is fundamentally reversed. Here we start with the genes of our genome of interest, for which we seek to make annotations via homology. Our workflow is therefore as follows:

1) We use search methods to identify putative homologs
2) We scan GO annotation databases and PubMed to see if *any* of the putative homologs has
   a) existent GO annotations
   b) a paper with experimental evidence of function/process/component that we can use to make GO annotations
3) We use homologs with existent GO annotations (or make *regular* GO annotations on homologs) to make our *transfer* annotations (i.e. transferring the homolog annotation to the gene in our genome of interest)

Note that, in many cases, putative homologs will not have existing GO annotations. That means that in order for you to annotate a gene in the genome of interest you will have to perform *two* annotations: a first "regular" annotation on the homolog and a second "transfer" annotation on the target genome.

Alternative workflow

Notice that it is possible to use an alternative workflow. This alternative workflow depends on first identifying one or more model organisms that are evolutionary "close" to our organism of interest. These model organisms will have abundant experimental literature that we can annotate on (or maybe even already-made annotations). We just need to check before we start annotating that the gene in the model organism will transfer to our organism of interest according to the criteria outlined in the CACAO GO_REF and made explicit here: https://gowiki.tamu.edu/wiki/index.php/Category:CACAO_GO_REF.

Annotations using the CACAO interface

To facilitate and standardize the annotation process in genomes of interest, we use the CACAO/GONUTS interface, developed by Jim Hu at Texas A&M University. CACAO is an intercampus

annotation competition. For reference on CACAO and GONUTS, see all the great instructional material already available here: http://gowiki.tamu.edu/wiki/index.php/Category:CACAO (see *Help for Students* section)

Getting gene products for our genomes of interest

In order for our annotations to be incorporated into the Gene Ontology (GO), we must annotate gene products with an assigned accession. Most gene products are proteins, and hence we will use UniProt identifiers. Proteins get UniProt accessions after the genome sequence has been successfully submitted to the NCBI GenBank or EBI ENA databases. Once you have identified a gene to annotate in the genome of interest, get its protein accession from the NCBI RefSeq database (e.g. YP_008051028.1). Go to the UniProt website and search with this accession. You should get as a result an item with a UniProt accession (R4TBI6).

Checking for annotations

The first thing to do once you have a candidate gene for annotation is to check that it has no previous annotations. Go to QuickGO and search with your UniProt accession.

The fact that your gene is not on QuickGO does not mean that it has not been annotated; it might have been picked up by another student and may be already annotated in CACAO. Check this by searching with the code `9CAUD:` followed by the UniProt accession (R4TBI6; that is: 9CAUD:R4TBI6) in the CACAO interface.

Creating a gene page

Chances are that your gene will not even be in CACAO to start with. To add your gene to CACAO, click on the *Create New Gene Page* link on the left panel and paste your UniProt accession into it, then hit *Create Page*.

Figure 3 – Gene page creation in CACAO.

Making an annotation

Once you have created a gene page (or using an existing one), you can make annotations using the *edit table* link and then clicking on the *Add row* button. This will bring up a form to create the annotation (Figure 5). The annotation fields are as follows:

- **Qualifier**: this allows you to modify the annotation to indicate, for instance, that the GO term used is *NOT* applicable to your gene.
- **GO ID**: this is the identifier of the GO term. You can use QuickGO and AmiGO to browse GO terms and find the one that is most adequate for describing the function/process/component you want to annotate.

- **Reference**: this will either be the CACAO GO_REF or the PubMed ID of the article you are annotating from.
- **Evidence code**: this is the evidence code term that best captures the method used to make the annotation.
- **with/from**: if you are making a "transfer" annotation using the GO_REF, you will use *WITH* and enter here the UniProt ID for the identified homolog you are annotating from. Note that you can use multiple homologs to make your annotation.
- **Aspect**: this just refers to the type (function/process/component) of annotation our term belongs to
- **Notes**: here you should summarize the process used (e.g. HHpred with probability *X* and coverage *Y* identifies gene *Z* as a putative homolog for this gene). You can see examples in any annotation already on CACAO, such as this one. Take also a look at the instructions on the GO_REF (Figure 4). You should also note here the rationale for transferring the function/process/component from the homolog to your gene, especially when the homolog is not a gene from a similar organism (that is, why you think the same function/process/component applies to your gene).

```
go_ref_id: GO_REF:0000112
title: Gene Ontology annotation by CACAO biocurators
authors: Ivan Erill, James Hu, Community Assessment of Community Annotation with Ontologies
year: 2017
abstract: This GO reference describes the criteria used by biocurators participating in the Community
Assessment of Community Annotation with Ontologies (CACAO) to annotate gene products from genomes of
interest through the use of computational methods to establish and manually validate function or homology
to gene products. In particular, this GO reference describes the criteria used to make annotations based
on evidence codes ISS, ISA, ISO, ISM and IGC. To perform ISS-, ISA-, and ISO-based annotations on a gene
product, CACAO biocurators use sequence- and structure-based search algorithms (e.g. BLASTP, HHPred) to
establish homology, conservation of sequence and structure functional determinants between the target gene
product and gene products from other organisms with published GO annotations supported by experimental
codes and lacking NOT qualifiers. These gene products are referenced in the WITH field of the annotation
using their xref database accession. ISM-based annotations make use of published computational methods
(e.g. TMHMM, SignalP) to predict gene product structure, localization or function. IGC-based annotations
are made on the basis of suggestive evidence for function based on synteny. Parameters and criteria for
use of all computational methods (e.g. e-value) are listed and versioned in the publicly available CACAO
documentation (http://gowiki.tamu.edu/). Annotations made by CACAO biocurators are reviewed by CACAO team
instructors before their release.
```

Figure 4 – The Gene Ontology annotation by CACAO biocurators GO_REF:0000112.

Figure 5 – Making an annotation in CACAO. Adding a row to the annotation table (left) and making the annotation (right).

Once you have filled up the required fields, click *Save Row* and then, on the *TableEdit* page don't forget to click on the *Save Table to wiki page* button (Figure 5). Otherwise your annotation will NOT be saved.

Making a GO "transfer" annotation for a gene in our genome of interest

Making "transfer" annotations on a genome of interest is not easy. First, and foremost, you must not rely on the assigned function (if any) in the genome annotation. These annotations are likely automatic and do not intend to be a permanent and validated functional association for the gene. The following example describes the process of making an annotation for the YP_008430876.1 - TROLL_93 "tail assembly chaperone" gene product in *Bacillus phage Troll*, a recently sequenced bacteriophage genome. It is intended to be an illustration for the process, not a direct template you should follow in your annotations, and the main steps and concepts introduced apply to any other genome of interest.

BLASTP and HHpred

The first thing to do is to run a search with the main programs we use in this unit: BLAST and HHpred. In this case, we modify the BLASTP parameters to ask for 5,000 targets. This is a good trick, because it allows you to detect similarity with more distant things that the lot of closely related genomes (usually poorly annotated in bacteriophages) that populate the first ~100 rows. Another convenient trick is to exclude Eukarya to speed up and focus the search (since we will rarely be able to make use of hits with eukaryotic organisms to faithfully annotate a bacteriophage gene).

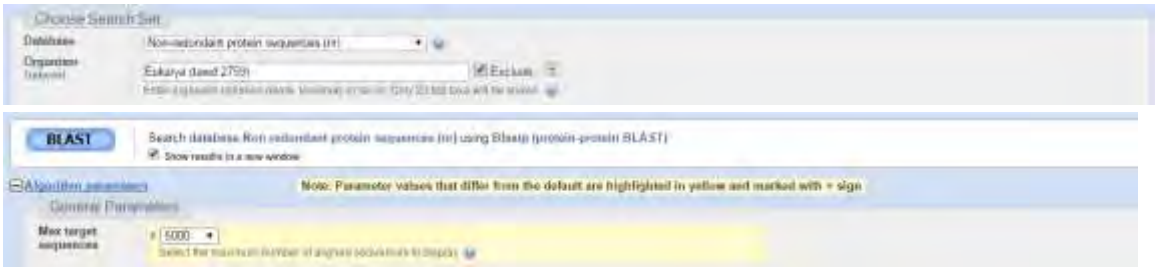


Figure 6 – Setting up the BLAST search.

When trying to annotate with these search tools, the first thing to do is to look for hits on relevant model organisms. Most experimental work and serious annotation on bacteriophages has been done on a handful of them. These include Enterobacteria phage lambda, *Enterobacteria phage T4* and *Enterobacteria phage T7*. Closer to *Bacillus phage Troll*, several *Mycobacteriophages* (L5, L1, TM4 and D29) have been carefully annotated, and the same is true for *Bacillus phage SPO1*, *Listeria* phages A511, PSA and A118, and *Staphylococcus aureus* phages G1, phiMR11 or SA4, *Bacillus cereus* bacteriophages BCP78 and B4, and *Bacillus phage vB_BceM-Bc431v3* (this is by no means an exclusive list).

The BLASTP and HHpred results in this case are rather disappointing. HHpred returns only high-quality hits to generic domains, which we cannot use. BLASTP does not return any slam-dunk hits (such as a hit to an *Enterobacteria phage lambda* tail chaperone, which would come in handy). In fact, only a few entries in the BLASTP result list hit genes annotated as "tail chaperones"[3]. The first one comes from

[3] BLASTP will not provide you with an extensive list of results. Identical proteins, for instance, will be masked and only one representative will be reported. If you find experimental evidence for what looks (from the given name/function) like a homolog of your gene, it is good practice to perform a BLASTP limiting your search to that specific *Organism*.

Bacillus phage Moonbeam. Following the protein accession [AIW03469.1], we can see on the right *Related information* tab that we are lucky, since there seems to be at least one article in PubMed citing this gene. This is a recent *Genome Announcement* paper on the "Complete Genome Sequence of Bacillus megaterium Myophage Moonbeam" by Cadungog *et al.* [PMID:25593264]. The protein is also accessible at UniProt, with accession number A0A0A0RPE2. Reading the paper, we find the following statement:

> Several functional proteins were identified using BLASTp and InterProScan analyses (6, 7). Genes encoding structural proteins include a capsid protein, portal, prohead protease, tail proteins, tail chaperones, tape measure protein, tail proteins, and multiple components of the baseplate. The tail chaperone had an unusual +1 frameshift to its secondary product, where most Caudovirales use a −1 frameshift to encode their secondary tail chaperone (8).

This leads us to reference 8: Xu *et al.* "Conserved translational frameshift in dsDNA bacteriophage tail assembly genes." [PMID:15469818]. Here we find this:

> The gene encoding the tape measure protein is easily recognizable in the genome because it is very long (usually more than 2 kb) and the encoded protein is predicted to be largely α-helical. Furthermore, the order of the tail genes is highly conserved. Notably, the major tail protein gene is always upstream of the tape measure protein gene, and, as we show here, between these two genes there are typically two overlapping open reading frames (ORFs) that are related by a programmed translational frameshift.
>
> In bacteriophage λ, between the major tail protein gene *V* and tape measure protein gene *H*, two proteins−gpG and gpGT−are encoded, the second by a −1 translational frameshift (Figure 1A) (Levin et al., 1993). Both proteins are required for tail assembly even though neither is part of the mature tail structure. Near the end of gene *G*, a "slippery sequence" in the mRNA, 5′-GGGAAAG-3′, causes about 3.5% of the ribosomes to slip back one nucleotide, with the shifted ribosomes then continuing to read in the −1 reading frame to make a larger fusion protein, gpGT.

and this:

> comparative genomic studies show that dsDNA-tailed phages with very similar virion morphology often have structural genes with very different primary sequences (Brussow and Hendrix, 2002). Yet the head and tail genes of these phages normally have the same or similar functional gene order despite the frequent lack of demonstrable homology (Casjens et al., 1992).

Enterobacteria phage lambda gene *G* (lambdap14; NP_040593.1) has a UniProt accession (P03734), and QuickGO shows three associated annotations using evidence code IEA (Inferred from Electronic Annotation).

Figure 7 – Annotations for Enterobacteriophage lambda gene *G* (P03734).

Making a "regular" annotation

IEA codes are not really what we are aiming for. IEA annotations are tags automatically assigned to genes using rudimentary mappings to function (e.g. the presence of an InterProt domain or of keywords indicating function in its product definition). IEA annotations are deleted one year after their generation[4]. You can consider them as a to-do list of sorts for UniProt biocurators. Hence, the idea is to properly annotate the *Enterobacteria phage lambda* gene *G* ourselves, so that we can then transfer the manual annotation. The first place to look is the cited reference (Levin et al. 1993), but this paper is not freely accessible. A PubMed search for it, however, will return also related papers, which we can check.

[4] You may find in QuickPro that a gene you are about to annotate already has several IEA annotations. These can give you pointers as to what you can expect to annotate for that gene, and the WITH record may provide you with some clues as to where the annotation comes from. Remember, however, that these are low quality, automatically generated annotations. You can not use them (i.e. enter them in CACAO and count them as annotations if they are not present there), and the sources they are based on (conserved domains or keywords) are unlikely to be of any use for performing a regular or transfer annotation. In other words, for intents and purposes of this work you should proceed as if there were not IEA annotations.

Figure 8 – Following reference (Levin et al. 1993).

On first inspection, one of them seems to provide enough information for an annotation (Xu *et al.* "A Balanced Ratio of Proteins from Gene G and Frameshift-Extended Gene GT Is Required for Phage Lambda Tail Assembly" [PMID:23851014]). The paper clearly demonstrates that the G protein (and its frameshited GT version) is required for tail assembly. Tail assembly (as QuickGO will tell us) is a well-defined biological process with GO term "GO:0098003 - viral tail assembly". This term has two children (fiber and baseplate assembly) but the paper does not specify the role of the *G* gene to this level of detail.

Hence, we'll create an annotation in CACAO for *Enterobacteria phage lambda* gene *G*. This annotation will use the IMP (Inferred from Mutant Phenotype) as the evidence code, and use Figures 3 and 4 of the paper as the main source of evidence[5]. The annotation note will read something like:

"The authors use a plasmid construct with all essential tail genes to analyze the effect of mutations in plate formation. In particular, they introduce mutations that remove the slippery sequence resulting in the G-T frameshift. These mutations lead to the direct production of gpGT fusion protein (and no G protein at all; Fig. 3). The authors show that such mutants do not generate active tails (Fig. 4)"

The page for *Enterobacteria* phage lambda gene *G* already exists in CACAO/GONUTS (LAMBD:VMTG) and contains the three IEA annotations we saw in QuickGO, so we will just add ours to the table.

[5] An annotation on a published article cannot be made based on the information provided in the abstract. You should read the manuscript and identify (and name in your Notes) the specific figures/tables/paragraphs of the paper that provide the experimental evidence that you are using for determining the GO term and the evidence code of your annotation.

Figure 9 – Making a "regular" annotation in CACAO.

Making a transfer annotation

Now that we have a nice and tidy (one hopes) "regular" annotation for lambda phage gene *G*, we must figure out how to "transfer" the annotation to our *Troll* gene (TROLL_93, YP_008430876.1, S5YQ92). Ideally, we would rely on a BLASTP or HHpred hit. The paper by Cadungog *et al.* [PMID:25593264] asserts that there is homology between Bacillus phage Moonbeam CPT_Moonbeam72 (A0A0A0RPE2) tail assembly chaperone and those studied by Xu *et al.* [PMID:15469818]. Using targeted BLASTP against the genomes of several phages listed by Xu *et al.* in their supplementary information (e.g. Bacteriophage PSA), our Troll protein consistently matches the tail assembly chaperone in several of them, but never below the threshold e-value listed in the CACAO GO_REF instructions. And BLASTing directly against the Enterobacteria phage lambda genome does not generate any valid hits.

Hence, a nice and tidy homology-based annotation is not possible. That, however, does not mean that a transfer annotation is impossible.

Browsing the literature (starting with Xu *et al.* [PMID:15469818] and following references and cross-listed papers in PubMed), we can identify several instances that remark on the conservation of the *Enterobacteria phage lambda* G-T-H gene arrangement. For instance, Schuch* and Fischetti (2006) remark on their "Detailed Genomic Analysis of the Wβ and γ Phages Infecting Bacillus anthracis: Implications for Evolution of Environmental Fitness and Antibiotic Resistance" [PMID:16585764]:

> Recently, a highly conserved programmed translational −1 frameshift was found to be common among the tail assembly genes of most double-stranded DNA phage (70) ... Analysis of the γ and Wβ sequences did identify two putative orthologs of *G* and *T*, *orf11* and *orf12*, which are of the appropriate size and, like *G* an *T*, are encoded between a major tail protein (*orf10*) and a tape measure protein (*orf13*). Unlike, *G* and *T*, however, the *orf11* and *orf12* loci do not overlap and appear to lack a conventional slippery sequence. Despite this, a nonconventional slippery sequence, providing either a −2 or a +1 frameshift, could fuse the *orf11* and *orf12* products and is worthy of further investigation.

By combining the weak but consistent similarity of TROLL_93 with many other reported tail assembly chaperones and the multiple statements about synteny of the frame-shifting tail chaperones [TROLL_93-TROLL_92] preceding the tapemeasure gene (which is reasonably well-annotated in Troll: TROLL_94), we have solid grounds to postulate that TROLL_93 is a distant homolog of *Enterobacteria phage lambda* gene *G* (with TROLL_92 playing the part of *T*), and hence transfer the involvement of this putative tail chaperone in tail assembly from *Enterobacteria phage lambda* to *Bacillus phage Troll*.

We will do this by means of an *IGC – Inferred by Genomic Context* evidence code, using the *Enterobacteria* phage lambda gene *G* and several other homologs with known synteny conservation in the *WITH* field. Our reference will be the CACAO GO_REF. The note will read as follows:

"BLASTP shows that the protein coded by TROLL_93 is a homolog of the "tail assembly chaperone" AIW03469 (coded by CPT_Moonbeam72), which is known to be homologous to several phage tail assembly chaperones displaying a conserved frameshift and preserved gene organization consisting of the two frameshifted chaperones (TROLL_93 and TROLL_92) upstream of the tapemeasure gene (TROLL_94) [PMID:25593264, PMID:15469818, PMID:16585764]. Examples of these include the chaperones coded by *Listeria* Bacteriophage PSA *ORF11* (CAC85567), and *Enterobacteria* phage lambda gene *G* (AAA96546) or *Streptococcus thermophilus* bacteriophage Sfi19 *orf117* (AAC39294). Given the strong synteny conservation and the specific nature of the genes involved, the TROLL_93

gene product can be assumed to have a similar chaperone role in tail assembly to its Enterobacteria phage lambda counterpart."

Figure 10 – Making a "transfer" annotation in CACAO.

And hence, by means of a "regular" and a "transfer" GO annotation we have managed to annotate the product of gene TROLL_93 to a specific biological process ("GO:0098003 - viral tail assembly) as experimentally established in *Enterobacteria phage lambda*. So on to the next gene/annotation…