

# GO annotation guidelines for the TIGR Prokaryotic Annotation Group

## What is GO?

The Gene Ontology (GO) project began in 1998 as a collaboration between three model organism databases: FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD). The GO consortium has since grown to include many other major model organism databases and sequencing centers (including TIGR). For a complete list of contributors see the GO web page: [www.geneontology.org](http://www.geneontology.org)

The Gene Ontology project was initiated to address the need for consistent descriptions of gene products in different databases and across all species. The GO consists of three structured, controlled vocabularies (ontologies) that can be used to describe gene products in terms of the biological processes in which they are engaged, the cellular components in which they act or live, and the molecular functions which they carry out. The GO was designed to be as species-independent as possible, allowing one system (the GO) to be used for the annotation of all organisms. The controlled vocabularies facilitate querying and retrieval of data from many different sources using a common query structure. There are three separate aspects to this effort: the production and maintenance of the ontologies themselves; the creation of associations (or annotations) between the GO terms and gene products; and the development of tools that facilitate the creation, maintenance and use of the ontologies.

The GO Editorial Office is charged with the responsibility of maintaining the integrity, continuity, and consistency of the ontologies. The bulk of additions and changes that are made to the ontologies are done through that office. This group has as their full-time responsibility the care of the three GO ontologies and therefore, they are the most versed in the lore of GO, its rules, and guidelines. One might wonder whether there is much to change after work has been going on on the ontologies for 7 years now, but in fact there is still quite a bit of work to do. There are around 20,000 terms in GO. There are still errors in content and structure of the terms that date back to the original creation of the term set. These are found and corrected over time as various groups start annotating with the terms in question. In addition, there is a constant need for the addition of new terms. Although the original creators of the GO (in large part Michael Ashburner) did an amazing job at thinking up as many possible terms as they could, it is impossible at the outset to think of every possible term that could be needed for every possible organism. Therefore, as new organism groups have joined the GO, there has arisen a need for terms to represent the areas of biology present in those life forms. A constant stream of requests for new terms and term changes is sent to the GO Editorial Office. The GO has set up a SourceForge site to track these requests and their outcomes and to act as a forum for discussion:

([https://sourceforge.net/tracker/?func=browse&group\\_id=36855&atid=440764](https://sourceforge.net/tracker/?func=browse&group_id=36855&atid=440764)). Within the TIGR prokaryotic group it is usually Michelle who submits requests to the GO SF site. Many of these requests come in to Michelle from the annotation team. Anyone on the team is free to submit requests to SourceForge at any time, however, since writing and shepherding these requests through to completion can be a very time-consuming process (and since Michelle is very familiar with it) most annotators choose to let her submit requests on their behalf.

If a TIGR annotator needs a new term for use immediately, they have the option of creating a **TI term** (TIGR specific term). These terms are entered into our db just as though they are real GO terms and they can be used as real GO terms in annotations. When a TI is made, a corresponding request for the new term is also sent to GO via SF (usually by Michelle). When the new term is created at GO, the TI term is replaced in all annotations with the new GO term and the TI is made a secondary id of the real GO term (more on secondary ids later). If an annotator can afford to wait a few days and if the new terms are fairly straightforward, then it is often best to skip making a TI and just request Michelle send in the SF item since lately the GO editorial office is very fast with straightforward requests.

**Associations (or annotations)** between GO terms and gene products are made not by the GO Editorial staff, but by the annotators and curators of the member databases. This encompasses a wide range of activities. Some groups (for example SGD) do only literature curation of their gene products and perform no sequence analysis of their own. While other groups (for example prokaryotic TIGR) do very little literature curation and a huge amount of sequence based analysis. These variations are largely dependent on the data sets and time frames within which the different groups operate. The yeast research community is very large and very active and has accumulated a wealth of publications on yeast for the SGD curators to draw on. In addition, SGD has a full-time staff of curators dedicated solely to the curation and maintenance of the SGD data set. They have had years (and will continue to have more years) to scrutinize the 6000 genes in the SGD dataset. On the other end of the spectrum is the prokaryotic group at TIGR. Many of the genomes we annotate have a very small publication record, therefore we have no choice but to rely on sequence based methods to determine the putative functions of our gene products. In addition, we work in a much more high-throughput environment where we might annotate 6000 genes in only 3 months with little opportunity to revisit them in the future.

In addition to the SourceForge tracker that the GO uses to track ontology changes, there are several other tools GO makes available to users. One is **AmiGO**, a GO ontology browsing and searching tool. Some annotators may wish to use AmiGO, it is located on the GO web page. It has some advantages over the Manatee viewer in speed, but lacks some of the nice search features of the Manatee viewer. Increased development currently going on for AmiGO may make it our choice for ontology viewing in the near future.

## Why is GO annotation useful?

The process of manual annotation, that is the collection and review of evidence to collect information on the on all aspects of a gene product, is a time consuming and therefore expensive process. It therefore behooves us to capture as much of the information (annotation) we get during that process as possible. The GO system allows capture and, very importantly, efficient exchange of annotations. It is important to point out that the manual annotation process as carried out by the prokaryotic team at TIGR is the same, whether we are using GO to annotate or not, however, the difference comes in the ability to capture and disseminate the information you collect during the annotation process. Without GO, one has limited avenues for capturing and communicating what has been learned about a given protein being annotated. The only places to store information about the protein without GO are the common name field, the gene symbol field, the EC# field, and the comment field. However, to do a search of annotation data in these fields one would need to use a text matching tool. This has some limitations. Two examples: the same enzymatic reaction can be known by several very different names (for example "succinate dehydrogenase" and "fumarate dehydrogenase" are actually the same enzyme, but a text matching tool would not know that); the same text string can be used to describe two very different biological concepts (for example "bud formation" in yeast is very different from "bud formation" in a plant but a text matching tool would think these were the same). The use of the GO system alleviates these difficulties and adds value. It allows the capture of more information about the protein: you can also capture the function of the protein, the process in which it is involved, and the cellular location or protein complex in which the protein lives or acts.

Since GO is a controlled vocabulary, all annotators who use GO will, by definition, be using the same terms to describe the same ideas. Since each term has a precise definition everyone will know what other people mean by the assignment of any given term. The same enzymatic function which has 3 alternative names and is described by the name "2-keto-3-deoxy-galactonokinase" at TIGR and "2-oxo-3-deoxygalactonate kinase" at Sequences R Us, will be annotated with the GO term GO:0008671, "2-dehydro-3-deoxygalactonokinase" which is linked to both name variations as synonyms (see more on synonyms below). Different concepts represented by the same text string will be represented by different GO terms that have precise definitions - for example, "periplasmic space (sensu Fungi)", GO:0030287 and "periplasmic space (sensu Gram-negative Bacteria)", GO:0030288. Therefore, problems associated with variations in protein, function, and process names will be alleviated.

In addition, since all users of GO report the data in a consistent format, all GO annotation data can be searched using one tool. Therefore, annotations from any organism across the tree of life from Arabidopsis to Zebrafish and everything in between can all be searched via their GO annotations with one common tool. This is useful in many ways. First, it allows users to see all the different processes that proteins with similar functions are involved in. Second, users can see all the proteins with a similar function regardless of whether they have sequence similarity to each other or not.

## GO ontology structure - parts of a term and relationships between terms

### Each GO term has 3 required parts:

**id number (go\_id):** a unique 7 digit zero-padded id number

**name:** a descriptive text name

**definition:** a text field containing a complete definition of the term.

It is important to note that the GO id number is assigned not to the descriptive name, but to the definition, the meaning of the term. Most of the GO terms now have definitions (this was not always so), with only about 10% still undefined.

There are several other pieces of information that may be associated with a GO term:

**synonyms:** A term may have one or more "synonyms" where the "synonym" is a text string somehow related to the name of the GO term; it may have a narrower meaning, a broader meaning, or be an exact synonym of the name. Synonyms are assigned to GO terms to help the process of finding the GO term you want/need. Many entities are known by more than one name, for example "peptidoglycan" and "murein sacculus" refer to the same structure. Likewise, synonyms can be entered for all of the various alternative names for enzymes.

**comment:** A text field for the entry of term specific comments, such as, guides to the term's use in annotations. If a term has become obsolete there will be a note here stating why and suggesting alternative terms.

**database cross ref:** This is to store accessions from other databases that have something with the meaning of the GO term, for example there is a one-to-one correspondence between EC numbers and many GO terms, so such terms will have a cross reference to the EC number. This information shows up in the EC number field on the Manatee GO tree page.

**secondary id:** Some terms are secondary ids to other terms. This often occurs when it is discovered that two terms actually mean the same thing. One is merged into the other and becomes secondary to the other. From that point on, the secondary id always points to the primary id. The primary id is the one to use in annotations.

**obsolete terms:** If a term has been made obsolete it will have the tag "is\_obsolete" as true in the GO data file and will likely appear as a child of the "obsolete node" in its tree in Manatee (however this manner of presenting obsoletes will likely change). Terms are made obsolete for many reasons. Often a term is discovered to simply not be appropriate for GO (it may be capturing a gene product rather than a description of a gene product for example), or, if the definition of a term is discovered to be incorrect or too vague the term will be made obsolete and a better new term id with a better definition will be created. As noted above, the GO id number is linked to the definition, not the name. So, if a change is made to the name of a term, such as correcting spelling or slight re-wording, then the id number stays the same. However, if a change is made to the definition (the meaning) of the term then the term is taken out of the active ontologies and is made "obsolete". This is important to do since if the meaning of a term has changed then the annotations that were made between gene products and

that term may no longer be valid under the new definition. Therefore, making the term obsolete alerts users that the term has undergone change serious enough that all annotations to that term must be reviewed.

GO is a **DAG or directed acyclic graph**. What this means to us is that in a DAG a term can have more than one parent. This is different from a hierarchical structure in which terms can have only one parent (like the TIGR roles). It is important for the GO to have the capacity for multiple parentage since GO is trying to capture the highly complex nature of biology in which the same term will sometimes need multiple parents. (Fig. 1)

### Absolute Path in GO Tree: 2 instances detected

```
+Ontology (TI:0000001)[R]
  +Gene_Ontology (GO:0003673)[P]
    +molecular_function (GO:0003674)[P]
      +binding (GO:0005488)[I]
        +nucleic acid binding (GO:0003676)[I]
          +DNA binding (GO:0003677)[I]
            +transcription factor activity (GO:0003700)
          +transcription regulator activity (GO:0030528)[I]
            +transcription factor activity (GO:0003700)[I]
```

Figure 1. An example of a term with 2 parents. This is a screen shot from Manatee. Transcription factors have both "DNA binding" and "transcription regulator" activities and therefore need to be categorized in both places.

There are 2 (soon to be 3) allowed **relationships between terms: "is an instance of" (or "isa") and "part of"**. Soon will be added "regulates", but that has not been added yet. Many of the function terms have "isa" relationships, for example: "ribokinase" isa "kinase" - in this set "ribokinase" is the more specific child of "kinase" (and therefore "kinase" is the more general parent of "ribokinase"). Many of the component terms have "part of" relationships, for example: "cytoplasm" is part of the "cell". It is common to talk of the terms as children and parents of each other, and even as grandchildren or grandparents. In addition, terms which share the same parent are called "siblings". (Fig. 2)

```

+DNA binding (GO:0003677)[I] [add]
  +transcription factor activity (GO:0003700)[I] [add]
    left-handed Z-DNA binding (GO:0003692)[I] [add]
    damaged DNA binding (GO:0003684)[I] [add]
    P-element binding (GO:0003693)[I] [add]
    DNA end binding (GO:0045027)[I] [add]
    methyl-CpG binding (GO:0008327)[I] [add]
    random coil DNA binding (GO:0003695)[I] [add]
    triplex DNA binding (GO:0045142)[I] [add]
  +ribosomal DNA (rDNA) binding (GO:0000182)[I] [add]
    satellite DNA binding (GO:0003696)[I] [add]
    DNA replication origin binding (GO:0003688)[I] [add]
  +single-stranded DNA binding (GO:0003697)[I] [add]
    DNA clamp loader activity (GO:0003689)[I] [add]
    unmethylated CpG binding (GO:0045322)[I] [add]
    centromeric DNA binding (GO:0019237)[I] [add]
    DNA bending activity (GO:0008301)[I] [add]
  +telomeric DNA binding (GO:0042162)[I] [add]
    AT DNA binding (GO:0003680)[I] [add]
    DNA secondary structure binding (GO:0000217)[I] [add]
  +DNA topoisomerase activity (GO:0003916)[I] [add]
    bent DNA binding (GO:0003681)[I] [add]
  +double-stranded DNA binding (GO:0003690)[I] [add]
  +chromatin DNA binding (GO:0031490)[I] [add]

```

Figure 2. A sample GO tree showing siblings. This is another screen shot from Manatee. All of the terms under "DNA binding" are siblings to each other. The "I" at the end of each line indicates an "is a" relationship between the term and its parent. If the relationship was "part of" there would be a "P" there. A "+" at the beginning of the line indicates that the term has children, clicking on the term will refocus the tree on that term and show its children.

## Storing GO annotation data at TIGR

Links between GO terms and proteins are made in each small genome database in the **go\_role\_link** table. In this table "feat\_name" is assigned a "go\_id" and the row in the table is identified with a sequential numeric "id" field, "assigned\_by" and "date" are entered automatically by the db. Links between GO term assignments and the evidence that supports them are made via the **go\_evidence** table. **The go\_role\_link and go\_evidence tables are linked via the "id" field in go\_role\_link and the "role\_link\_id" field in go\_evidence.** The four fields "ev\_code", "evidence", "with\_ev", and "qualifier" are populated for each "role\_link\_id" ("id" from go\_role\_link). "ev\_code" and "evidence" are mandatory fields for each GO term assignment, while "with\_ev", and "qualifier" are only used with certain types of GO annotation (see below). See Figure 3.



Every GO term assigned to a protein must also include supporting evidence for the annotation. There are 4 fields where we store this information (also see Figure 3):

**db field name: ev\_code** (mandatory for all GO annotations)

description: This field holds an evidence code which is an abbreviation for the type of evidence used to make the annotation. Some commonly used ev\_codes are listed below (for a full list and descriptions of each see <http://www.geneontology.org/GO.evidence.shtml>).

ISS - "inferred from sequence similarity" - this should be used any time you are making a prediction of function based on some tool that employs sequence similarity. Included here would be HMM, BER, TMHMM, SignalP, PROSITE, and InterPro evidence. This is the most frequently used ev\_code in prokaryotic annotation.

IMP - "inferred from mutant phenotype" - use this when a researcher has published a paper on the gene in the organism that you are currently annotating, has done some kind of mutant assay, and has found that the protein is involved in some process or function.

IDA - "inferred from direct assay" - use this when a researcher has published a paper on the actual gene in the organism that you are currently annotating in which they have done an actual direct assay of function, for example an enzymatic activity assay.

TAS - "traceable author statement" - use this when you are making an annotation based on something that doesn't have an accession, or can not be easily described with the other ev\_codes, for example: if you use gene cluster (putative operon) evidence to make your annotation, use the TAS ev\_code in conjunction with a public comment describing the nature of the evidence - "Part of the evidence for the annotation of this gene is based on presence of the gene in a cluster of genes with related functions." (more on this below)

ND - "no biological data available" - used when annotating a gene to the "unknown" GO terms (more on this below)

IEA - "inferred from electronic annotation" - this code is used when GO terms are assigned by some automatic process, without human review. At TIGR, there are preliminary assignments which are then reviewed by annotators and changed to an appropriate other ev\_code.

**db field name: evidence** (called "reference" in Manatee, mandatory for all GO annotations)

description: A reference describing the nature of the evidence for the annotation to a particular GO term. This can be a paper which describes the experimental characterization of the protein you are annotating, used with the experimental ev\_codes like IMP and IDA (remember if you use the experimental ev\_codes the experiments must have been done on the actual protein you are annotating, not a homolog). Or this can be a paper that describes the process of annotation that we use, for example the genome publications. Or this can be a standard GO reference that has meaning within the GO system, the three of these standard GO references that we use are:

GO\_REF:nd - for use with any of the "unknown" terms

GO\_REF:0000011 - describes the process of using HMM as evidence, for use with HMM evidence

GO\_REF:0000012 - describes the process of using pairwise alignments for evidence, for use with BER evidence

**db field name: with\_ev** (called "with" in Manatee)

description: This is used with ISS evidence to store the accession number of the thing (HMM, BER, etc) your match is with and this field is also used with experimental characterization evidence (for the exact protein you are annotating) with IGI/IPI where you store the accession number of the gene/protein that your gene/protein interacts with.

This field is mandatory for ISS/IPI/IGI annotations, but is not populated for other ev\_codes.

**db field name: qualifier**

description: a set of controlled terms that are used to modify the nature of the GO annotation. Currently, the only qualifier currently used by TIGR prokaryotic annotation is "contributes\_to".

Allowed terms:

1. "contributes\_to" - this is ONLY used with FUNCTION GO terms, it is used when one is annotating the function of an entire protein complex to a subunit of the complex. This qualifier must be used in this case since the individual subunits of a complex do not themselves have the function of the whole complex, rather they contribute to the function of the whole complex. In addition to the function for the whole complex, one can also annotate functions to the individual subunits which they have on their own within the complex. For example, the ATP-binding subunit of an ABC transporter has ATP-binding activity all on its own, but also contributes to the function of the whole transporter. Therefore such a protein would get 2 function GO term annotations:

GO:0005524 "ATP binding"

GO:0042626 "ATPase activity, coupled to transmembrane movement of substances" with the qualifier "contributes\_to"

\*NOTE\* - the term GO:0042626 would be added to generic ABC transporters whose substrate was unknown. If you know the substrate of your ABC transporter, you would assign one of the substrate specific children (or grandchildren) of this term, for example: carbohydrate-transporting ATPase activity (GO:0043211)

polyamine-transporting ATPase activity (GO:0015417)

cation-transporting ATPase activity (GO:0019829)

and many others.....see role notes for more info.

2. "NOT" - used when there is a publication or other evidence that indicates a certain GO annotation, but then there is subsequent information that says that that GO annotation is not appropriate. Used very rarely in the GO community and not at all in TIGR prokaryotic annotation.

3. "colocalizes\_with" - From the GO documentation: "Gene products that are transiently or peripherally associated with an organelle or complex may be annotated to the relevant cellular component term, using the 'colocalizes\_with' qualifier. This qualifier may also be used in cases where the resolution of an assay is not accurate enough to say that the gene product is a bona fide component member."



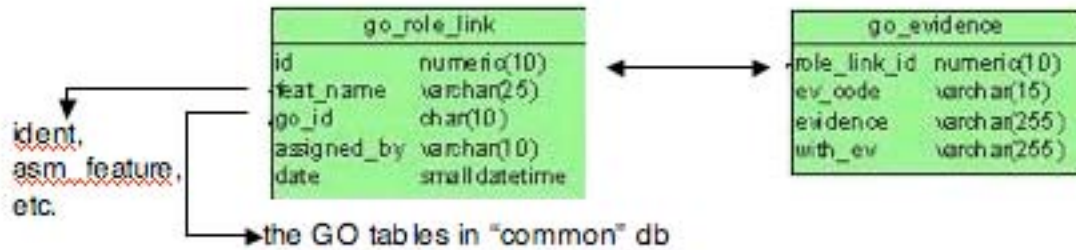


Figure 3. The GO tables in the small genome databases. These are where links between proteins and GO terms are stored.

## Annotating gene products to GO terms

The GO is designed for the annotation of gene products so it can be used to annotate both protein and RNA products. Currently, at TIGR (prokaryotic) we have only used GO to annotate proteins. As mentioned earlier, the process of evidence review and functional assignment is independent of GO annotation. As an annotator one should carefully examine all of the available evidence for a given protein and decide what (function/process/component) you think the gene (has/is involved in/is located in) and then find the correct GO terms that correspond to that information. Manatee has several handy tools built into it to help in finding GO terms. First there are several places on the Gene Curation Page which suggest terms that might be of use. These include: GO terms assigned to HMMs, GO/EC mappings, GO terms assigned to Genome Properties, GO terms mapped to InterPro hits, GO terms assigned to matching proteins from *V. cholerae* and *B. anthracis*. These suggestions are listed in the corresponding sections on the GCP or in the "GO suggestions" section at the bottom of the GCP (or linkable from the upper right corner of the GO section). The mappings to EC numbers should generally be exactly correct (assuming the EC number assigned to the gene is correct), however, if at first the GO term appears wrong, it could be that the GO term is using an alternate name for the enzyme than what we are using in our name, so before panicking check the EC site for alternate names. GO uses the official EC names as the GO term name for a particular enzymatic function and will add the alternate names as synonyms. TIGR now has a policy to use the official EC name for our protein common names as well. However, in the past, TIGR used the Swiss-Prot names which were often not the official EC names, so you may see inconsistencies in the data. If autoannotate or another annotator has assigned an EC number to a protein, then the EC GO suggestion should show up on the GCP. If this does not happen, then it likely means that a mapping between a GO term and that EC number has not been made either here at TIGR, at GO, or both. In that instance, email Michelle who will update the data and contact GO if necessary. If an EC number has not been assigned already to a protein, then the EC GO suggestion will be empty, however, you can find the EC number by using the EC search tool on the GO search page, accessible from the GO section.

Manatee has helpful "ADD" buttons next to all of the GO term suggestions which when clicked automatically fill the GO term into the "Add" column in the GO data entry

section. In addition to automatic GO term entry there are several places on the GCP where one can use one click to enter evidence into the "with" field. These include two places under HMMs: the "Add to GO evidence link" adds the HMM accession into the first available "with" field, and the "ADD" buttons next to GO terms assigned to the HMM add both the GO term and corresponding HMM evidence. Also, there is the "Add to GO evidence" button in the characterized match section for adding the characterized match accession to the "with" field. There is the "GO" link under "add GO evidence" in the Genome Properties section. Finally, if there is TMHMM evidence for a protein, there will be a link in the GCP section to add that evidence. All of these buttons are there to help make adding GO terms and evidence as easy as possible. Since Manatee knows all of the evidence storage rules, it puts the information into the fields in the correct format so that annotators need not remember the formats. In addition, it greatly cuts down on copy/paste/typing errors.

There are also handy pull-down menus for each ontology containing some of the more frequently used GO terms for each.

**For more information on all of these Manatee features, please see the Manatee tutorial.**

If none of the suggestions on the GCP are useful, there is a link to Manatee's GO search page from the upper right hand corner of the GO section on the GCP. Here one can search the ontologies with a GO id or GO term name keyword. Also can search the EC to GO mappings with an EC number. And finally, one can search: the GO annotations for correlations between GO terms (input a GO id and see if there is another GO id that is often assigned in conjunction with the input id, helpful for finding process and function terms when you already know one but not the other), for keywords in the names of proteins that have been annotated to GO, and with GO ids to see lists of proteins that have been assigned that GO term.

**For more information on all of these Manatee features, please see the Manatee tutorial.**

Any time one clicks on a GO id anywhere in Manatee one gets a tree showing that term in the context of its parents, children, and siblings (Fig. 1 and 2). If one gets to one of these trees from a GCP, one can use the "Add" buttons in the trees to add GO terms to the GCP. (Fig. 2)

**For more information on all of these Manatee features, please see the Manatee tutorial.**

When assigning GO terms to a gene it is best if you can assign at least one GO term from each ontology, but at the very least assign function and process. Since many proteins have more than one function and/or are involved in more than one process and/or live in more than one place in the cell, it is often appropriate to assign more than one GO term from each ontology. One should assign as many GO terms as needed until the aspects of the protein are completely described.

If you can not find any evidence for an aspect of the protein, then you should assign the "unknown" term for that aspect - each of the three aspects of the GO have one:

"biological process unknown", "molecular function unknown", "cellular component unknown". The ev\_code for the unknown terms is always "ND" which stands for "no data". The reference that should be used is "GO\_REF:nd" or "GO\_REF:0000015". Manatee knows the rules and will always fill in the correct info for you if you use the pull-down menus to fill in the "unknown" terms.

## Information and notes on specific topics:

### 1. When **operon/gene cluster information** is evidence for annotation:

In this case use the TAS ev\_code, nothing in with\_ev, leave TIGR\_CMR:annotation as reference, and write something in both the internal and the pub\_com fields that says what evidence you used for the annotation. For example: "Part of the evidence for the annotation of this gene is based on presence of the gene in a cluster of genes with related functions."

### 2. When to use **TAS**:

Use TAS whenever there is evidence that you use for annotation that is not of a sequence similarity nature and therefore does not have an accession number that you can put into the with field, for example: gene cluster/operon evidence. Leave with\_ev blank, leave TIGR\_CMR:annotation as the reference, add text to internal and public comment fields that explains what the evidence is, for example: "Part of the evidence for the annotation of this gene is based on presence of the gene in a cluster of genes with related functions."

### 3. The "**unknown**" terms:

Each of the 3 GO ontologies have "unknown" terms, they are: "biological process unknown", "molecular function unknown", "cellular component unknown". These are to be assigned to a gene after you have looked to find a function/process/component annotation but can not determine what that information for a particular protein. Annotation to these terms indicate that an annotator has actually looked for a function/process/component but could not find one. These annotations allow one to differentiate genes that have not yet been annotated (those without GO terms at all) and those that have been annotated but for which process/function/component could not be assigned.

When using these terms assign ev\_code "ND", reference "GO\_REF:nd", leave with and qualifier blank.

### 4. What **specificity** do you assign?

One of the useful things about the GO system is that there are terms at all levels of specificity (or "granularity" in GO speak) so that whatever level of functional specificity you have confidence in for the protein in question can be reflected in GO annotation.

## Available evidence for 3 genes

#1

-HMM for "ribokinase"  
-characterized match to ribokinase

#2

-HMM for "kinase"  
-characterized matches to a "glucokinase", AND a "fructokinase"

#3

-HMM for "kinase"

## Function

catalytic activity

kinase activity

carbohydrate kinase activity

ribokinase activity

glucokinase activity

fructokinase activity

## Process

metabolism

carbohydrate metabolism

monosaccharide metabolism

hexose metabolism

glucose metabolism

fructose metabolism

pentose metabolism

ribose metabolism

Above is an example illustrating the varying levels of specificity and is explained below. Assign GO terms only with as much specificity as the evidence supports, just as you assign common names only with as much specificity as the evidence supports. Assign proteins that have only family level evidence to more general GO terms. For example "kinase", "oxidoreductase", "transporter", etc. If you know only that the protein is an enzyme use "catalytic activity" (function) and "metabolism" (process).

Quality and specificity of evidence dictates GO term specificity. In the above figure, we see a brief view of the function and process ontologies and a list of the evidence available for 3 different genes. Gene #1 has the most specific evidence of the three and can therefore be assigned the most specific GO terms: "ribokinase activity" and "ribose metabolism". Gene #2 has intermediate evidence, we know the gene is a kinase from the HMM, but we have matches to two different characterized proteins. In this case we can say that the protein is a carbohydrate kinase, but not specifically which hexose the

kinase acts on so the GO terms are: "carbohydrate kinase activity" and "hexose metabolism". Gene #3 has the least specific evidence and therefore is assigned the least specific GO terms. We only know that it is a kinase, but have no idea what kind of substrate it acts on. The appropriate GO terms are: "kinase activity" and "metabolism". For more specific examples of GO terms assigned to specific TIGR genes, please see the "Gene naming and annotation guidelines" document.

#### 5. GO terms and **"putative" genes:**

Do not add specific GO terms to a gene you have named putative if you are lacking in confidence about the function. Very "strong" putatives can get more specific GO terms, but in general, putatives should have more general GO terms.

#### 6. GO terms and **"homolog" genes:**

There are two classes of "homolog" genes: those with high quality evidence for a particular function but where that function is not expected in the given organism and those with poor quality evidence but where there is enough similarity to warrant mention. For the first case, it may be possible that the function could exist in the bug in question, but be involved in a different process, this will need to be determined on a case-by-case basis. For the second case, assign the "unknown" GO terms as the evidence does not warrant making any claims about function/process/component.

#### 7. GO terms and **hypothetical proteins:**

Both the plain hypothetical proteins and the conserved hypothetical proteins should be assigned the three unknown terms. (This is a policy change from the past in that GO has now agreed that the plain hyps should get the unknown terms as well, when before they said they should get no GO annotation.)

#### 8. **Data consistency:**

It is important to maintain data consistency, not just with GO, but with all of the annotation we assign to proteins. That means that genes in the same operon, or with the same function, or with the same general function should all have consistent names using the same nomenclature standard, consistent GO terms at consistent levels of specificity, consistent gene syms (using the same symbol format), and consistent TIGR roles. In particular for operons, make a decision about the whole operon after looking at all the genes in it, is the function (or pathway or complex) there or not?, then annotate all the genes in the operon accordingly. When annotating a particular role, make sure the format of the names and assignment of GO terms is consistent from protein to protein when the type of function is the same, this kind of situation arises a lot in transport.

#### 9. GO terms for **genes with translation problems.**

All genes designated authentic frameshift or authentic point mutation can be assigned GO terms with the same specificity as if when the FS or PM were missing. We do this because, it is not known whether or not the FS or in-frame stop exists the population of the species as a hole or whether it exists only in the DNA sample used for the sequencing project. In addition, there may be some kind of read-through process that

allows translation of the product of which we are unaware or the protein may be active in some kind of truncated form. However, genes that are in any of the categories that fall into TIGR role 270 "Disrupted reading frame" (including degenerate (multiple FS and in-frame stops), truncations, interruptions, fragments, etc.) do not receive GO terms as it is highly unlikely that these seriously disrupted ORFs are in any way functional.

## How are the GO ontologies themselves stored at TIGR?

Every night the latest copy of the GO obo file containing the newest version of the three ontologies is downloaded and the information is updated in our database. At TIGR we store a version of the GO ontologies in the database called "common".

There are 3 tables that store the actual ontology information: **go\_term, go\_link, and go\_synonym.**

The go\_term table stores the actual information associated with each GO term including id number (go\_id), name, which ontology the term comes from (type), definition, and comment. The go\_link table stores the relationships between the GO terms where each row in the table stores the relationship for a pair of GO terms. "parent\_id" is the go\_id of the term that is the parent of the other term in the pair, known as the "child\_id". The "link\_type" field tells us what kind of relationship exists between the two GO terms. Possible types of relationships are: "isa" which says that the child term "is an instance of" its parent (ribokinase is a kinase); "partof" which says that the child term is "a part of" its parent (cytoplasm is part of a cell), and "supercedes" which says that the parent term "supercedes" the child term (used for secondary ids where the secondary id becomes a child of the primary id with the supercedes relationship, secondary id terms always point to the primary id, the primary id should be the one used in annotations). The go\_synonym table stores the synonyms to the names of the GO terms. Manatee uses these three tables for the GO term text search, GO id search, and all GO trees.

We also store other GO information in the common database. In the "go\_map" table we store links between GO terms and other data sets. Stored here are links between GO terms and EC numbers, InterPro, etc. Manatee uses this table for the EC number search and suggestions. We store annotations to GO, both those from TIGR and elsewhere, in the go\_gene\_association and association\_evidence tables. Manatee searches these tables for the protein common name search, the GO associations keyword search, and the GO correlations search.

See attached schema.