Gene Annotation and Naming Guidelines

TABLE OF CONTENTS

C. Homology too weak for functional identification, but still worth recording

3. Specificity uncertain (*ie*, family classification)
    A. Database match is to members of a defined family whose function is known
    B. Database match is to members of a defined family whose function is unknown
4. Limited sequence similarity (*ie*, domains)
    A. The ORF has no characterized matches, but has an above-trusted cutoff match to an equivalog-level DOMAIN HMM
    B. The ORF has no characterized matches, but has an above-trusted cutoff match to a DOMAIN HMM of unknown function
5. Evidence consists only of uncharacterized proteins
    A. ORF only produces full length database matches to conceptual translations in other species; no characterized match or other evidence to indicate true function
    B. As in (A), but the BER alignments are not full-length; instead they show motifs or regions of very good local alignment or regional similarity which warrant mention
    C. ORF is a hypothetical or conserved hypothetical protein with five or more transmembrane regions
    D. ORF is a hypothetical or conserved hypothetical protein with a match to a lipoprotein consensus motif
6. No database match
    A. Lacks significant similarity to any previously published genes from other species, families, or motifs
    B. Has an HMM match but no matches from other species in the BER Skim
7. Final comments and preferences

V. ANNOTATION CONVENTIONS FOR DISRUPTED READING FRAMES

1. Authentic frameshifts/authentic point mutations
2. Degenerate ORFs (Multiple/mixed frameshifts and point mutations)
3. Interruptions and Insertions
4. Truncations
5. Selenocysteine-containing proteins
6. Programmed frameshifts
7. Internal deletions
8. Fragments
9. Fusions

VI. PRACTICAL ANNOTATION

I. INTRODUCTION

This document provides background information and guidelines for microbial annotation. The most accurate annotation comes from experimental work done on the particular protein and genome being annotated. Unfortunately, this level of certainty is almost never available to annotators. Day-to-day annotation largely consists of inferring the function of a sequence from other sequences. It is important to realize that we use an inexact method, that the guidelines described in this document cannot cover every possible combination of evidence, and that when the evidence for function is ambiguous, two annotators may come up with two different but valid annotations for the same sequence. The aim is to provide annotation consistent with the best available evidence. By the same token we want to avoid perpetuating transitive annotation errors which plague the public databases. Annotators are encouraged to consult with each other on difficult cases; and annotation choices, when not obviously supported by the evidence, should be explained in the public and private comments for a sequence.

II. EVIDENCE TYPES

TIGR ‘gene assignment’ is the annotation of Open Reading Frames (ORFs) identified by the gene-finding program Glimmer. Annotation for each ORF is based on the following evidence presented on the ORF’s Gene Curation Page (GCP):

1. BLAST-Extend-Repraze (BER) pairwise alignments

Protein translations of all ORFs are searched against a non-redundant amino acid database using protein BLAST. Matching proteins with significant BLAST scores are preserved and stored in a mini-database. The gene is then extended by 300 nucleotides at both the 5' and 3' ends, and the extended translation is then aligned to the proteins in the mini-database (from the BLAST step) using a modified Smith-Waterman algorithm. The top forty BLAST hits are summarized on the GCP in a table called the ‘BER Skim’, which is linked to the alignments. Especially valuable are alignments to ‘characterized matches’ – sequences for which experimental evidence of function or process has been published. Characterized matches are color-coded in the BER Skim according to evidence type (see item 2 below).

Because of the forty-alignment limit, numerous hits to high-scoring uncharacterized whole genome project sequences (flagged in the BER Skim as ‘wgp=1’) can cause informative matches (such as to *Escherichia col*i sequences) to 'fall off' the bottom of the alignment list. In such cases annotators should run their own BLAST search against SwissProt or other suitable database to identify possible characterized matches. Currently, non-BER Skim accession numbers cannot be entered into the characterized

table as evidence, so annotators should cite the match in the comment area(s) of the GCP instead, and/or use it as GO evidence

2. Characterized matches

TIGR maintains a database of proteins which have been experimentally characterized. Manatee displays this information with color-coded backgrounds and text in the BER Skim. Characterized matches are colored according to the associated tag in the characterized table. These tags are:
DB_PARSE (red): potentially a characterized match, derived from automatic parsing of SwissProt records; requires annotator to check
EXPERIMENTAL> (green): function and process known
EXPERIMENTAL (FRAGMENT): only a fragment has been characterized
EXPERIMENTAL (DOMAIN): only a functional domain of the gene has been characterized
EXPERIMENTAL (PARTIAL) (light blue): either function or process is known, but not both
MISINFORMATION: published function is incorrect
TRUSTED: multiple lines of indirect evidence suggest the gene is characterized, including: a characterized very close homolog; presence in a characterized operon; position effect (conserved gene order); biochemical pathway information. Used only when the lines of evidence are strong.
VOID: a sequence originally thought to be characterized, which turns out not to be

If a high quality match exists to a protein which has itself been experimentally characterized to show a specific function/process, we capture that information as a piece of evidence in the 'characterized match accession' field. Clicking on the accession number in the left hand column of the BER skim automatically puts the accession into the 'add accession 'box. Clicking on an accession already stored in the db and displayed on the GCP will paste it into the 'delete accession' box.

Annotators should always be on the lookout for new characterized matches by checking the literature on candidates that are in the BER Skim (whether flagged by a DB_PARSE, flagged as ‘experimental=1’ or ‘experimental=-1’, or chosen by the annotator). SwissProt accessions in the BER alignment link to a SwissProt page of relevant literature for that protein sequence. Annotators can also scan PubMed for literature related to a particular com_name or gene_sym (*e.g.*, ftsZ) using the following query in NCBI's Entrez PubMed search field:

ftsZ AND pubmed protein [sb]

Characterized matches can be edited via links from the BER alignment view in the GSP. When editing a characterized match, either update the existing entry, or delete it and add a new one.

3. HMM matches

Protein translations of all ORFs are searched against hidden Markov models (HMMs) built from multiple protein sequence alignments. Each HMM has associated with it a 'noise' cutoff score and a 'trusted' cutoff score. ORFs are considered to be members of the HMM model if they score higher than the trusted cutoff. If an ORF yields a score between the trusted cutoff and the noise cutoff of an HMM, it deserves closer examination before excluding the HMM from consideration; if the HMM is used as the basis of annotation, a score between trusted and noise is grounds for appending ‘putative’ to the name of the ORF.

HMM evidence occupies its own segment of the GCP. Ideally the match should extend across the entire HMM; matches having a >20% length discrepancy compared to the HMM are flagged in red. HMM regions matching the ORF are also graphically displayed in the Evidence Picture on the GCP. The GCP displays hits to HMMs built by TIGR (whose HMM accession numbers start with 'TIGR') and Pfam (whose HMM accession numbers start with 'PF'). TIGR classifies TIGR and Pfam HMMs into fifteen ‘isology’ types, each of which specifies a different level of database match:

EQUIVALOG: a collection of proteins that share function back through their last common ancestor.

EQUIVALOG_DOMAIN: a region with an assignable conserved function that with some regularity shows up in different protein architectures. It can be the sole functional domain in a protein or it can be one of several functional domains in a longer, multifunctional protein.

HYPOTH_EQUIVALOG: a family of uncharacterized proteins hypothesized to be equivalogs.

HYPOTH_EQUIVALOG_DOMAIN: a region with a hypothesized conserved function that, with some regularity, shows up in different uncharacterized proteins architectures. It can be the sole domain in one of these proteins or it can be one of several domains in longer proteins.

PARALOG: a family whose members are all drawn from the same (or very closely related) genome.

PARALOG_DOMAIN: a region shared by members of a family that are all drawn from the same (or very closely related) genome.

SUBFAMILY: formally, a branch of a superfamily. Subfamilies often include fairly closely related proteins with functional heterogeneity. In practical terms, this iso_type is warning against trying to interpret the set of proteins as equivalogs.

SUPERFAMILY: a collection of proteins with the same domain structure, encompassing all homologs, usually including proteins with at least two different functions.

DOMAIN: This is the broadest isology; it connotes a region of similarity shared by proteins homologous over portions of their length, encompassing all homologs, usually including proteins with at least two different functions. The domain itself is not presumed to have the same function in all instances; the HMM describes only a sequence similarity. Contrast to 'equivalog_domain' HMMs, where conserved domain function is presumed.

REPEAT: a region that is found in multiple copies in members of the HMM.

PFAM: a Pfam model not yet even preliminarily classified by isology type at TIGR. These should be approached with skepticism.

PFAM_EQUIVALOG: a Pfam model that appears to find only equivalogs, but cutoffs are probably too lenient for automated annotation. Beware of 'false positive' hits.

PFAM_EQUIVALOG_DOMAIN: a Pfam model that appears to find only equivalog_domains, but again cutoffs are probably too lenient for automated annotation. Beware of 'false positive' hits.

The Pfam HMM ‘gathering’ score is analogous to a TIGRFAM ‘trusted’ score, but is less rigorously assigned. An ORF that scores above gathering but below trusted to a Pfam HMM should be treated with exceptional caution. In fact, annotators should be wary of drawing inferences from any Pfam HMM, including those that have been assigned an isology type, since even they may not have been rigorously reviewed by a TIGR annotator.

Pfam also assigns its HMMs to ‘clans’ (higher-level groupings of protein sequences). As a consequence, multiple Pfam HMMs from the same clan may match the same region of an ORF. If this occurs, annotation should be based on the best-scoring HMM.

4. Genome Properties

A Genome Property comprises a suite of genes known to participate in a metabolic pathway, cellular activity, or cellular structure. Genome Properties also include basic data about prokaryotes such as their Gram staining and genomic GC content. HMM- and context-based rules are used by software to identify potential Genome Property genes during autoannotation. Indicators appear with associated HMM information on the GCP when potential genes of that Property have been found elsewhere in the genome – alerting the annotator that the current ORF may be part of it too. It is often convenient to identify and annotate all the genes in a Property together, by this means.

5. Paralogous families

Paralogous gene families are constructed by searching the translation of all genes of a genome against themselves and clustering genes by sequence similarity. Paralogous families are therefore candidates for bulk annotation, since they may share similar function. They are graphically displayed in the Evidence Picture on the GCP.

6. Transmembrane regions, lipoprotein motifs and other biologically significant patterns

The ORF protein sequence is searched against PROSITE for biologically significant patterns and sites, including the lipoprotein motif. Potential alpha helix transmembrane regions are predicted by TmHMM software (a tool developed by the Center for Biological Sequence Analysis). These are useful in identifying transporters, signal sequences, cell membrane and envelope proteins. Additionally, signal sequences are predicted by the CBSA's SignalP software. The results of all of these searches are graphically displayed in the Evidence Picture.

7. Gene context

Gene context, such as location within a cluster or operon with a common functional theme, can be significant in some assignments - particularly for genes such as ABC transporters or enzymes involved in biosynthetic or metabolic pathways. Genes in the same operon are good candidates for bulk curation, and should have consistent names, role_IDs, gene_syms and GO terms. The gene context of an ORF can be viewed by launching the Genome Viewer from the pull-down menu at the top of the GCP, or in the context of Genome Properties via the Genome Properties information page.

8. Annotator-initiated searches of databases on the Web

Many Web-based bioinformatics databases allow annotators to input a sequence or search term to retrieve evidence of function. The Prokaryotic Annotation group maintains a page of Useful Web Links to these sites, which include BLASTable databases of transporter proteins, enzymes, motifs and metabolic pathways.

III. DESCRIPTORS

We annotate each gene by assigning as many descriptors as are relevant to each gene. In a practical sense, annotation is populating database tables with descriptions of the gene. The annotator has the option of populating the following six fields:

1. Common name (com_name)

The common name of the protein. Nomenclature guidelines are described in detail below - in general this will be the most specific name justified by the evidence. Enzymes should be assigned the standard enzyme commission (IUBMB) name. In most other cases we defer to protein names assigned by SwissProt (specifically *E. coli* annotation). Note that in some cases we incorporate the gene symbol into the common name, *e.g.*, ' cell division protein FtsZ', when the common name alone is not specific. Names of bifunctional genes are separated by a slash as described below in (in 'Types of Database Match).

2. Gene symbol (gene_sym)

The three or four letter gene symbol, *e.g.*, ftsZ (the protein symbol has an initial capital, *e.g.*, FtsZ). We use *E. coli* gene symbols as a standard, and choose them when one is available. If the ORF matches a gene which is not in *E. coli*, then we default to *Bacillus subtilis* as the standard. In cases where there is neither an *E. col*i nor *B. subtilis* gene, we choose a gene symbol from a consensus of those found in the pairwise alignment file. But we try to use standard gene_syms, and avoid ones created or altered by researchers for a specific species.

When the ORF being curated has an analogous function but dissimilar sequence compared to a gene of *E. coli*, we do NOT use the *E. coli* gene_sym.

The form of bifunctional gene_syms depends on whether they share a common prefix,

as described below in 'Types of Database Match'.

For duplicate genes, we do not use hyphenated numbers to distinguish the gene_syms, since this form is classically reserved for alleles. Instead we simply add the number to the gene_sym; however, this is done by the contact BA during final consistency checks, not by the annotator.

Isozymes (isoenzymes) are defined by the IUBMB as multiple forms of enzymes that catalyze essentially the same reaction, and which arise from genetically determined differences in primary structure (*i.e.*, sequence variants or heteromers of two or more polypeptide chains) rather than from modification of the same primary sequence (post-translational modifications). A bacterial example is the three isozymes of acetolactate synthase in *E. coli*. Where a gene_sym exists we number them as we would for other duplicate genes.

Where no gene_sym is available but the annotator thinks a set of ORFs are duplicate genes/isozymes, put the information in the public_comment field (*e.g.*, "1 of 3, other loci are XX#### and XX####").

3. Enzyme Commission number (ec_num)

The Enzyme Commission number is a four part numbering scheme for representing specific enzyme activity; *e.g.*, 1.1.1.1 (alcohol dehydrogenase). This system is curated by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) in consultation with the IUPAC/IUBMB Joint Commission on Biochemical Nomenclature (JCBN). Multiple EC numbers should be separated by spaces.

4. Comments

The GCP contains private and public comment fields for any pertinent comments about the protein. These are useful for documenting additional information you may find from literature searches, constructing multiple sequence alignments, etc. Anything written in the public comment field can be read by the public in both our Genbank accession and on the TIGR CMR. Therefore, exercise caution when writing these comments. Be brief, precise, and scientific; do not use TIGR jargon or any usage that could be considered unprofessional.

5. TIGR roles (role_id)

These describe the biological role of the protein. TIGR originally adapted Monica Riley's role scheme for *E. coli* for use with our prokaryotic projects. We use these categories to both organize the data for easy annotation, and also to analyze genome content. Assignment of a protein to a TIGR role(s) is achieved by assigning the id number of the TIGR role. A full list of these is available by clicking the 'role help' link in the TIGR role section on the GCP. From the GCP annotators can also access text describing the genes that belong in that role, and the specific naming guidelines relevant to that role.

6. GO terms

TIGR has officially adopted the Gene Ontology (GO) classification system for annotating the molecular functions which the genes carry out , the biological processes they are involved in, and the cellular components in which they live and act. These three aspects of a protein are captured in the three controlled vocabularies (ontologies) of the GO system. Every effort should be made to assign at least one GO term from each ontology to each protein. Since some proteins have more than one function, are involved in more than one process, or live in more than one place in the cell, often more than one GO term from each ontology will be required to fully describe a protein. Assign as many GO terms to each protein as are appropriate to fully describe it. GO terms should be assigned at the level of specificity that is supported by the available evidence for the function of the protein. See the GO Annotation Guide for more details.

IV. LEVELS OF NAMING SPECIFICITY

In the course of reviewing data we have developed certain criteria regarding assignments. Many ORFs will fall into one of the categories described below. However, there are many possible combinations of evidence types, so annotators often must weigh the value of various pieces of data, rather than simply apply a 'rule of thumb'.

1. Confident assignment

A. Full-length match to an equivalog HMM and/or good database match to protein of known function

The protein should match an EQUIVALOG HMM at a score higher than the trusted cutoff or an EQUIVALOG_DOMAIN HMM, but only if the match covers essentially the full length. Typically, there will be matches to proteins with only one function in the BER search file. The percent identity/similarity should be high along the entire length of the match. In general, consider greater than 35% identity over the entire length of the protein to be significant. (However, this is not a strict cutoff: a 35% identity match between 500 aa-long proteins is much more significant that 35% identity for proteins only 100 aa long, take into account the length of the
proteins. Also, because of the variation in frequency between different amino acids -- *e.g.*, leucine-9% vs. tryptophan-1.4% -- we attribute more significance to matches between 'rare' amino acids). If the gene is an enzyme look for conservation of active sites, substrate, and cofactor binding sites. If the gene is not an enzyme then any PROSITE motifs that are defining characteristics of the protein should be conserved.

An example of this type of match is *Prevotella ruminicola* (gfr) ORFB01133. This ORF matches equivalog HMM TIGR01060: phosphopyruvate hydratase (enolase) with a score well above the 'trusted' cutoff. It also has a characterized match to enolase from Bacillus subtilis (SwissProt accession P37869) with 67% identity and 87% similarity. In

both cases the matches are essentially full-length. The active site H155 and magnesium cofactor binding sites D242, E287 and D314 are conserved (the residue numbering differs by one between the BER alignment and the SwissProt sequence because BER adds an initial Met residue to any matches lacking it). The PROSITE motif is present between residues 346 and 359. The ORF is annotated with com_name (using the official IUBMB name for the enzyme), gene_sym, ec_num, role_ID and GO terms as appropriate. The com_name is lower case:

phosphopyruvate hydratase
eno
role_ID 116 (Energy metabolism: Glycolysis/gluconeogenesis)
>ec_num 4.2.1.11
GO:0004634 (F) phosphopyruvate hydratase activity
GO:0006096 (P) glycolysis

B. Good evidence suggesting two or more different proteins of known function; *i.e.*, a hybrid or multifunctional protein

Note that these guidelines for annotation of hybrid proteins apply to all types of database match, not just to 'confident assignments'.

If different regions of the protein match well to two different equivalog HMMs and/or two different characterized proteins, the gene is probably bifunctional. This can be tricky since a gene that is bifunctional in one species may exist in another species as two separate genes. In this case we want to include all valid names in the com_name starting from the amino terminal end of the protein, separating the different names with a slash (no spaces between slash and adjacent words).

An example of this type of match is *Prevotella ruminicola* (gfr) ORFB00499. This ORF is a high-scoring match to the characterized bifunctional enzyme phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphohydrolase from *E. coli* (P06989). It also matches PFAM_equivalog_domain HMMs for each of the two enzymes (PF01502 and PF01503). Annotate the com_name, gene_sym, role_ID, ec_num, and GO terms as follows:

phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphohydrolase
hisIE
role_ID 161 (Amino acid biosynthesis: Histidine family)
ec_num 3.5.4.19 3.6.1.31
GO:0000105 (P) histidine biosynthesis
GO:0004635 (F) phosphoribosyl-AMP cyclohydrolase activity
GO:0004636 (F) phosphoribosyl-ATP diphosphatase activity

In this case the two gene symbols shared a common three-letter prefix (his). When the two gene_syms do not share a common prefix, use a slash to separate them, *e.g.*, comA/recB. The order of names or symbols should reflect the order of the activities in the sequence.

2. Function uncertain

A. Database match to protein of known function, but minor uncertainty exists

If we think the gene is almost certainly performing this function, but are less than fully confident, precede the com_name with "putative". In this case, the evidence for function is very strong except for one or two missing lines of evidence. The protein may match an EQUIVALOG HMM but the score is between the trusted and noise cutoffs. Within the BER search file, the percent identity/similarity may be lower than in case one (rule of thumb: 30-35% identity). There may be only one or a few examples in the BER search file, or the matches all are from genome projects. Active sites, or substrate/cofactor binding sites may not be as well conserved; PROSITE motif matches may be partial. We also use 'putative if the gene is located in an operon which suggests its function, but the similarity is very low. No gene_sym is assigned to the ORF, but role_ID and GO terms are assigned. A full or partial ec_num may be assigned if appropriate.

An example of this type of match is *Prevotella ruminicola* (gfr) ORFB00166. This ORF matches the equivalog HMM TIGR00091: tRNA (guanine-N(7)-)-methyltransferase, but the score is between noise and trusted cutoffs. There are no characterized matches in the BER Skim; if there is an *E. co*li hit, it has fallen off the table due to the numerous genome project hits. Indeed, independent BLAST search against SwissProt yields a match to the characterized *E. col*i gene for tRNA (guanine-N(7)-)-methyltransferase (P32049). However, the percent identity is <35%. Annotation:

putative tRNA (guanine-N(7)-)-methyltransferase
no gene_sym
role_ID 168 (Protein synthesis: tRNA & rRNA base modification)
no ec_num
GO:0006400 (P) tRNA modification
GO:0008176 (F) tRNA (guanine-N7-)-methyltransferase activity

If the gene name adds clarity or specificity to the assignment preserve it as part of the protein name using the protein naming convention of capitalizing the first letter, *e.g.*, "putative cell division protein FtsZ".

B. Database match to protein of known function (characterized match or equivalog HMM), but function not predicted to be conserved

In this case homology is strong enough to want to record, but unlike a ‘putative’ match, we do NOT believe the query protein has the same function as the match. This might be because critical residues are not conserved (*e.g.*, catalytic residues in an enzyme), or because the function is not predicted to exist in this particular organism (*e.g.*, photosynthetic enzyme matches in a non-photosynthetic organism; eukaryotic proteins in a bacterium), or because sequence homology is strong but other critical evidence is lacking. We name these ‘protein name, homolog’ and assign them to role_ID 157 (Unknown function, general).

An example of a homolog is *Mycobacterium tuberculosis* (gmt) ORF05771. This ORF matches N5,N10-methylenetetrahydromethanopterin reductase - an enzyme involved in

methanogenesis in methane-producing organisms. Since *M. tuberculosis* does not produce methane, we predict that this ORF cannot have the same function as its chosen pairwise match:

N5,N10-methylenetetrahydromethanopterin reductase homolog
no gene_sym
role_ID 157 (Unknown function: General)
no ec_num
GO:0000004 (P) biological process unknown
GO:0005554 (F) molecular function unknown

C. Homology too weak for functional identification, but still worth recording

In this case there is more uncertainty about function than for the 'putative' class, but the matches may still be informative. No family name is available; no equivalog matches exist; characterized matches, if they exist, have scores considerably below the the 'confident' range. An example is *Myxococcus xanthus* (gmx) ORF06592. The ORF matches two apparently domain-like HMMs for the HipA protein, but both have 'PFAM' isology (*i.e.*, their actual isology type has not been determined); a characterized match to *E. coli* HipA is only 28% identical to the ORF. The SwissProt record for the match does not indicate membership in a protein family, but notes that high levels of HipA are toxic and that the protein associates strongly with HipB -- potentially useful information which would be obscured if the ORF was classed as a 'conserved hypothetical'. In such cases we append 'homolog' to the name of the match:

HipA homolog
no gene_sym
role_ID 157 (Unknown function: General)
no ec_num
GO: 0000004: (P) biological process unknown
GO: 0005554: (P) molecular function unknown

Note that if a family name is available, it should be used rather than 'homolog'.

3. Specificity uncertain (*i.e.*, family classification)

As a general rule when the specific function or name of an ORF cannot be determined from the evidence, if a family designation is available we use that, rather than making an unwarranted stab at precision. Conceptually, a protein *family* (or subfamily) should consist of evolutionarily-related proteins with the same function in different organisms, while evolutionarily related proteins whose functions have diverged form a *superfamily*. In practice these distinctions are not rigidly adhered to in many protein databases, most of whose protein families are defined by sequence relatedness without verification that all family members share the same function. Annotators should examine the documentation for protein families they encounter, to determine if function is likely to be consistent for all members of the family. TIGR defines SUPERFAMILY and SUBFAMILY HMMs as groups of proteins which may not share the same function, while EQUIVALOG isology

is reserved for homologs likely to have conserved function. Proteins which match SUPERFAMILY and SUBFAMILY HMMs above the noise cutoff should be named as "(HMM_com_name) family protein". We do not use ‘subfamily’ or ‘superfamily’ in our com_names. (See the section on HMMs in 'Descriptors', above, for more information on how TIGR defines family HMMs.)

A. Database match is to members of a defined family whose function is known.

In this context a defined family means: a family for which an HMM has been built, or which can be identified through the literature, found in SWISSPROT, PROSITE, or some other similar curated database. Annotators should not coin new family names if there is no such documentation. No gene_sym is assigned to the ORF, though a partial ec_num should be assigned if appropriate.

An example of this type of match is *Mycobacterium tuberculosis* (gmt) ORF05434. This ORF matches several kinases at about the same degree of similarity (*e.g.*, D-arabinitol kinase, xylulose kinase, gluconate kinase) so we can predict that this is a carbohydrate kinase, but cannot accurately predict the substrate specificity. However, these enzymes belong to the FGGY family of carbohydrate kinases, and this ORF also has a match above the trusted cutoff to the HMM for the FGGY family of carbohydrate kinases. So the common name of this ORF is less than completely specific:

carbohydrate kinase, FGGY family
no gene_sym
role_ID 119 (Energy metabolism: Sugars)
no ec_num
GO:0005975 (P) carbohydrate metabolism
GO:0019200 (F) carbohydrate kinase activity

It is also possible in some cases to omit the word 'family', *e.g.*, if the match was to a family of sulfatases whose sulfur esterase function is conserved, but whose substrates differ, the com_name could simply be 'sulfatase'.

If the family name is derived from a protein name, we would use the protein naming convention of capitalizing the first letter, *e.g.*, PfkB family. Also note that term "class" can be used instead of the term "family" if this makes sense or follows pre-established conventions.

B. Database match is to members of a defined family whose function is unknown.

No gene_sym or ec_num is assigned to the ORF; role_ID is 157 (Unknown function: General).

These are similar to (A) except that no reasonable inference about the function of the family can be made. An example of this type of match is *Prevotella ruminicola* (gfr) ORFB00309. This ORF aligns substantially with an uncharacterized family of proteins which has been documented as the "DHH family" in Pfam. This ORF is designated:

DHH family protein
no gene_sym
role_ID 157 (Unknown function: General)

no ec_num
GO:0000004 (P) biological process unknown
GO:0005554 (F) molecular function unknown

It can be difficult to distinguish Cases 2A/2B (putative/homolog functional assignment) from Cases 3A/3B (family assignment). Often the database match to a family of proteins is strong, but the particular best match within this family is difficult to assign. In these cases, construction of a multiple alignment containing the query ORF and entries from the BER file is recommended as this can identify similarity relationships more precisely.

4. Limited sequence similarity (*i.e.*, domains)

There are two broad classes of domain HMMs: those where function is predicted to be conserved (*i.e.*, TIGR and Pfam 'equivalog_domain' HMMs) and those where no such assumption is made (TIGR and Pfam 'domain', 'subfamily domain' and 'paralog_domain' HMMs). In cases where a domain HMM is the main evidence for an annotation, this classification determines whether specific role_IDs and GO terms may be applied. Remember also to consider the length of the domain alignment to the ORF; it is safer to assume that domain and ORF annotation are the same when the domain spans most of the ORF, than when it represents just a small part of a possibly multidomain protein.

A. The ORF has no characterized matches, but has an above-trusted cutoff match to an equivalog DOMAIN HMM.

Generally the name of the domain HMM can be used in the com_name. If the domain has a defined role_ID, ec_num, or GO terms (*i.e.*, it is an equivalog domain), these may also be applied at the discretion of the annotator.

An example is *Myxococcus xanthus* (gmx) ORF04103. This gene matches TIGR00097: phosphomethylpyrimidine kinase, an equivalog_domain TIGR HMM for the ThiD protein. It's the only HMM evidence that matches the ORF above cutoff. The match is essentially co-terminous (*i.e.*, the alignment spans the entire ORF and HMM). The ORF also has full-length matches to numerous uncharacterized bacterial ThiD proteins, (as well as a partial match to a characterized bifunctional eukaryotic protein). The HMM evidence box also shows that other members of the Genome Property associated with the HMM are present in gmx. The annotation of the HMM was transferred to the ORF:

phosphomethylpyrimidine kinase
thiD
role_ID 162 (Biosynthesis of cofactors etc: Thiamine)
ec_num 2.7.4.7
GO:0008972 (F) phosphomethylpyrimidine kinase activity
GO:0009228 (P) thiamin biosynthesis

B. The ORF has no characterized matches, but has an above-trusted cutoff match to a DOMAIN HMM of unknown function.

If no function can be definitively assigned to the domain, we adapt the domain HMM name to the form '[X] domain protein'. These ORFs belong in role category 157 or 703 (Unknown function: General or Unknown function: Enzyme, depending on the HMM) and receive the most general GO terms (*e.g.*, unknown function or process; catalytic; metabolism, etc.). No gene_sym is assigned to the ORF.

An example is *Myxococcus xanthus* (gmx) ORF05874. This gene matches the Pfam HMM PF03109: ABC1 family, which has a 'domain' isology type. Note that this is a case where the Pfam name and the TIGR isology differ; annotators should defer to the TIGR annotation:

ABC1 domain protein
no gene_sym
role_ID 157 (Unknown function: General)
no ec_num
GO:0000004 (P) biological process unknown
GO:0005554 (F) molecular function unknown

If the HMM name and isology are in conflict, annotators should also notify the HMM team.

5. Evidence consists only of uncharacterized proteins.

A. The ORF only produces full-length BER matches to conceptual translations in other species; there are no characterized matches or other evidence to indicate true function.

The criteria for making this assignment are the same as for an assignment for a protein of known function: a rule of thumb is to look for similarity of at least 35% over 75% of the length. These ORFs are assigned to role_ID 156 (Hypothetical proteins: Conserved). There are numerous examples and they are designated:

conserved hypothetical protein
no gene_sym
role_ID 156 (Hypothetical proteins: Conserved)
no ec_num
GO:0000004 (P) biological process unknown
GO:0005554 (F) molecular function unknown

Note that simply assigning the ORF to role 156 on the GCP automatically fills in the GO terms and other annotation. However, the annotator must manually delete any previous GO terms.

Annotators should promote hypothetical proteins to ‘conserved hypothetical’ status only if hypothetical proteins in the BER table come from species other than the ORF's. A BER table populated only with hypothetical proteins from different strains of the same

species is not sufficient evidence for promotion; these should be demoted to 'hypothetical' using the 'Make This ORF Hypothetical' option in the GCP pull-down menu.

B. As in (A), but the BER alignments are not full-length; instead they show motifs or regions of very good local or regional similarity that warrant mention.

These ORFs are assigned to role_ID 704 (Hypothetical proteins: Domain). There are numerous examples and they are designated:

conserved domain protein
no gene_sym
role_ID 704 (Hypothetical proteins: Domain)
no ec_num
GO:0000004 (P) biological process unknown
GO:0005554 (F) molecular function unknown

In addition, there are two cases of conserved hypothetical proteins in which annotation is based on evidence other than HMMs or pairwise alignments: namely transmembrane region predictions and the lipoprotein consensus motif. In all cases the ORFs are assigned to role_ID 88.

C. The ORF is a hypothetical or conserved hypothetical protein with five or more predicted transmembrane regions.

Transmembrane regions predicted by the TmHMM algorithm are displayed graphically in the evidence picture. If there are five ore more we make the assumption that the protein is membrane-spanning. There are numerous examples of these ORFs; they are assigned role_ID 88 (Cell envelope: Other) and are designated:

putative membrane protein
no gene_sym
role_ID 88 (Cell Envelope: Other)
no ec_num
GO:0000004 (P) biological process unknown
GO:0005554 (F) molecular function unknown
GO:0016021 (C) integral to membrane

D. The ORF is a hypothetical or conserved hypothetical protein with a match to a lipoprotein consensus motif.

Lipoprotein motifs predicted algorithmically are displayed graphically in the evidence picture. If one is found, we assume that the protein is attached to the membrane by a lipoprotein anchor. There are numerous examples of these ORFs; they are assigned role_ID 88 (Cell envelope: Other) and are designated:

putative lipoprotein
no gene_sym
role_ID 88 (Cell Envelope: Other)
no ec_num

GO:0000004 (P) biological process unknown
GO:0005554 (F) molecular function unknown

If the protein has both five or more TmHMM regions *and* a lipoprotein consensus motif, annotators should annotate as a 'putative membrane protein', as above.

Note that SignalP (signal protein motif finder) data alone does not allow us to use names such as ‘secreted protein’ because SignalP returns positive results for periplasmic and membrane-associated proteins as well as for secreted proteins.

6. No database match

A. The ORF lacks significant similarity to any previously published genes from other species, families, or motifs.

These are not assigned to any role category. These ORFs should be demoted to 'hypothetical' using the 'Make This ORF Hypothetical' option in the GCP pulldown menu, or else deleted using the Genome Viewer tool.

Hypothetical ORFs which overlap significantly with ORFs in other categories are good candidates for deletion; however, do not delete hypothetical proteins found within prophage, IS, transposon, or bacteriocin feat-type regions without further evidence, as these may be legitimate.

B. The ORF has an HMM match but no matches from other species in the BER Skim.

Try re-running the BLAST search, as the previous search may be out of date. If there are still no hits to the sequence databases, and the HMM match is to a Pfam HMM, name the ORF '[Pfam name] protein'. These are annotated further as per the rules for family and domain evidence above. If the HMM match is to a TIGRFam, consult the HMM team.

7. Final comments and preferences

Use the term "subunit" instead of "chain", and unless the subunit designation is an Arabic numeral, subunit designation should precede the word "subunit". For example,

DNA polymerase III, beta subunit
DNA polymerase IV, B subunit
sulfate adenylyltransferase, subunit 1

We do not distinguish mature proteins from their precursors, so avoid using the terms "precursor" or "proprotein". Also avoid use of the terms "-like", and "xxx operon protein" unless the protein has already been published with that name.

V. ANNOTATION CONVENTIONS FOR DISRUPTED READING FRAMES

Annotators frequently encounter alignments which contain either frameshifts or stop codons (point mutations), often having been flagged by autoannotation software. These can reflect either sequencing errors or an actual mutational event which has disrupted the gene. If a disruption is suspected that has not been caught by software, the annotator should manually submit a frameshift notification to the lab. This can be done by selecting "Frameshift Report" on the GCP pulldown menu and filling in all necessary fields. Annotators should also add the temporary designation of 'FRAMESHIFT' or 'POINT MUTATION' to the common name of potentially disrupted ORFs (whether found by software or human), so they can be easily identified via the com_name field. The lab will check the sequence and either correct the sequence or identify the frameshift or point mutation as authentic. If a sequencing error is found, it is corrected and FRAMESHIFT or POINT MUTATION is removed from the common name by the annotator. If no sequencing error is found (*i.e.*, the disruption is authentic), the common name is annotated as described below.

If during initial review of the ORF the frameshift/point mutation notification seems unwarranted, and the annotator feels no further checking is needed -- *e.g.*, if a frameshift only occurs in one match out of many in the BER alignment table and/or the disrupted match is to a sequence from a whole genome sequencing project -- check 'No Action' in the frameshift action field at the bottom of the frameshift report page. This removes the frameshift/point mutation flag from the ORF.

We have also established annotation conventions for ORFs which are in some way disrupted. Although authentic frameshifts/point mutations are not unusual, note that interruptions, truncations and degenerate ORFs are uncommon. Following are the types of gene lesions and our naming conventions for them:

1. Authentic frameshifts/authentic point mutations

When an ORF is disrupted by either a single frameshift or a single point mutation and this sequence has been authenticated by the lab, we simply add this information after the common name and add role_ID 270 (Disrupted reading frame) to the GCP. An example is *Geobacter sulfurreducens* (ggs) ORF04572 which is designated:

protein-glutamate methylesterase, authentic frameshift
no gene_sym
role_ID 188 (Cellular processes: Chemotaxis and mobility)
no ec_num
GO:0006935 (P) chemotaxis
GO:0008984 (F) protein-glutamate methylesterase activity

Note that in these cases we do not include a gene name or an EC number, but do include GO terms. This is because the GO Consortium assumes that a single frameshift/PM may be a sequencing artifact. For the same reason, we do not add role_ID 270 (Disrupted reading frame) to annotation of ORFs with single authentic frameshifts or point mutations.

2. Degenerate ORFs (multiple/mixed frameshifts and point mutations)

When an ORF is disrupted by multiple frameshifts and/or point mutations that have been confirmed by the lab, we assume that the ORF is not functionally expressed and we denote this with the term "degenerate" after the com_name. No gene_sym, ec_num, or GO terms are assigned, but the ORF is given two role_IDs, one for the role of the 'nondegenerate' protein and another for 'disrupted reading frame'. An example is *Neisseria meningitidis* (gnm) ORF02090:

sodium- and chloride-dependent transporter, degenerate
no gene_sym
role_ID 141 (Transport and binding proteins: Unknown substrate)
role_ID 270 (Disrupted reading frame)
no ec_num
no GO terms

3. Interruptions and Insertions

Interruptions are cases in which you can find both the amino and carboxyl terminal portions of an ORF separated by some sequence(s), such as a transposon. These should be labeled with "interruption-N" and "interruption-C" after the common name. The remaining annotation is as for degenerate ORFs. An example of this case is *Shewanella oneidensis* (gsp) ORF00706 and ORF00713:

site-specific recombinase, phage integrase family, interruption-N
site-specific recombinase, phage integrase family, interruption-C
no gene_sym
role_ID 132 (DNA metabolism: DNA replication, recombination, and repair)
role_ID 270 (Disrupted reading frame)
no ec_num
no GO terms

However, a small (20-30 amino acid) insertion found in a gene but not in its BER or HMM matches, is treated differently:

- if it is a repeat, or an expansion of a loop region, ignore it
- if it occurs elsewhere in the genome, call it 'insertion' and describe the situation in public_comment;
- if it can be identified as an IS element, call it that
- if none of the above apply, consult the HMM team

4. Truncations

These are cases in which some significant segment of the ORF is missing. This should not include ORFs that are just a little shorter than the database match, or ORFs that are simply a separately expressed, functional unit usually seen as a domain in a larger protein. The implication must be that enough of the gene is missing so that it is no longer

functionally expressed. In these cases we add "truncation" after the common name. The remaining annotation is as for degenerate ORFs. An example of this case is *Neisseria meningitidis* (gnm) ORF00002:

DNA helicase, truncation
no gene_sym
role_ED 132 (DNA metabolism: DNA replication, recombination, and repair)
role_ID 270 (Disrupted reading frame)
no ec_num
no GO terms

5. Selenocysteine-containing proteins

These are cases in which an ORF contains the specific in-frame termination codon TGA. In certain organisms an internal TGA is read by a selenocys-tRNA and encodes the amino acid selenocysteine. Other conditions must be met to distinguish such cases from a simple point mutation, *e.g.*, the genome must contain a selenocys-tRNA and the enzyme selenide, water dikinase. Confer with Dan Haft or other annotators before making this conclusion. In these cases we add "selenocysteine-containing" after the com_name. Annotators should also fill out the translation exception form accessible from the drop-down menu on the GCP. Remaining annotation is dictated by the evidence on the GCP. An example of this case is *Desulfovibrio ferrooxidans* (gdv) ORF03102:

formate dehydrogenase, alpha subunit, selenocysteine-containing
fdnG
role_ID 110 Energy metabolism: Anaerobic
role_ID 112 Energy metabolism: Electron transport
1.2.1.2
GO:0006118 (P) electron transport
GO:0008863 (F) formate dehydrogenase activity
GO:0009326 (C) formate dehydrogenase complex

6. Programmed frameshifts

These are cases in which an ORF contains an in-frame termination codon, and a naturally occurring frameshift prior to the termination codon regulates translation of the ORF. Some genes which are known to be regulated by this mechanism are: peptide chain release factor 2 (prfB), DNA polymerase III gamma/tau subunit (dnaZX); phosphoglycerate kinase/triosephosphate isomerase (pgk-tim); adhesin plaA; various phage proteins (lambda tail assembly protein; T7gene 10); transposases from IS1, IS150, IS911 and some from the IS3-family; contingency genes. In these cases we add "programmed frameshift" after the common name. Annotators should also fill out the translation exception form accessible from the drop-down menu on the GCP. Remaining annotation is dictated by the evidence on the GCP. An example of this case is *Dehalococcoides ethenogenes* (gde) ORF01591:

peptide chain release factor 2, programmed frameshift
prfB
role_ID 169 (Protein synthesis: Translation factors)
no ec_num
GO:0003747 (F) translation release factor activity
GO:0006415 (P) translational termination

7. Internal deletions

We define internal deletions as the absence of a region of DNA in the interior of an ORF relative to its orthologs. Internal deletions are shorter than interruptions, but long enough such that we expect the deletion to impair function. In these cases we add 'internal deletion' to the common name. Remaining annotation is as per degenerate ORFs. An example of an internal deletion is *Chlorobium tepidum* (gct) ORF01384:

transposase, internal deletion
no gene_sym
role_ID 154 (Mobile and extrachromosomal element functions: Transposon functions)
role_ID 270 (Disrupted reading frame)
no ec_num
no GO terms

8. Fragments

We can use 'fragment' in the com_name where needed. Fragments are differentiated from truncations in that the latter have a correct beginning or end, while fragments may not. Conceivably these categories can overlap. Include the 'normal' role_ID as well as the role_ID for disrupted reading frames. An example is *Burkholderia mallei* ORF13050:

putative trans-aconitate methyltransferase fragment
no gene_sym
role_ID 71 (Amino acid synthesis: Apartate family)
role_ID 270 (Disrupted reading frame)
no ec_num
no GO terms

9. Fusions

We define fusions as two different protein fragments which have been fused into one reading frame by a deletion event in the genome. There are currently no examples of these.

VI. PRACTICAL ANNOTATION

Based on a survey of annotators, here are the sorts of questions annotators ask themselves when confronted with a new ORF, grouped by evidence type. The evidence types are listed in the order generally reported by annotators when describing which evidence they tend to look at first -- a rough measure of 'importance'. There are many exceptions, and which particular evidence type turns out to be 'key' can vary greatly from ORF to ORF. Nevertheless, answering these questions is the nearest thing to a consensus 'method' for assigning com_names and other descriptors to an ORF:

HMMs

What is the isology type? Is the score above trusted, or in the putative range (between trusted and noise)? Are there clues to function in the HMM documentation? Is there a Genome Property associated with the HMM, and is the property supported for the genome being annotated?

BER Skim, multiple alignments, and trees

Is the ORF already in a public database? (This rarely happens, but when it does expect a 99-100% ID match.) Are the matches all from genome projects? Do any of the records need to be upgraded to characterized matches? Is there a characterized match and if so, what is its percent identity to the ORF? What are the percent identities of the best matches to the ORF? Are the alignments coterminous, or partial? Are functional domains/catalytic sites conserved? Does the clustering of the ORF and the BER Skim sequences in a tree provide clues to function?

Gene context

Do surrounding genes appear to form an operon or a functional cassette? Can they be annotated together?

Motifs, paralogous families

Do these agree with/supplement other evidence? Can paralogs be annotated together? Can TmHMM or lipoprotein evidence be applied?

External database searches

Do BLAST searches against function-specific databases (*e.g.*, for transporters or peptidases) allow assignment to a specific family?
