A Guide to

Michelle Gwinn Giglio

Table of Contents (for the most popular topics)

What Manatee Is

- Manatee is a web-based manual annotation tool for accessing and editing annotation data
- Manatee draws information from an underlying database for its displays
- Manatee sends information entered by annotators to the underlying database for storage
- Multiple users can access the same database from different computers when Manatee is run on a server

Getting started with Manatee

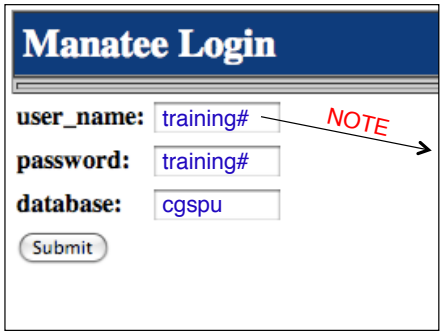
- Start a browser – go to http://manatee.igs.umaryland.edu
- To log into Manatee one must have an account and password.
- When logging into Manatee, one must enter a user account name, a password, and the name of the database on which you wish to work.
- For this class we will be using a training version of the *Shewanella oneidensis* genome database
  - the db name is “cgspu”

Finding Manatee

On the internet:
go to http://manatee.igs.umaryland.edu

There is a bookmark in FireFox on the training laptops.

To download:
go to http://manatee.sourceforge.net/igs/index.shtml



For this workshop, user names and passwords are the same. Use "training#" where number is anything from 1-30.

The database is "cgspu"

Case matters so be sure to use all lower case.

"Welcome to Manatee"

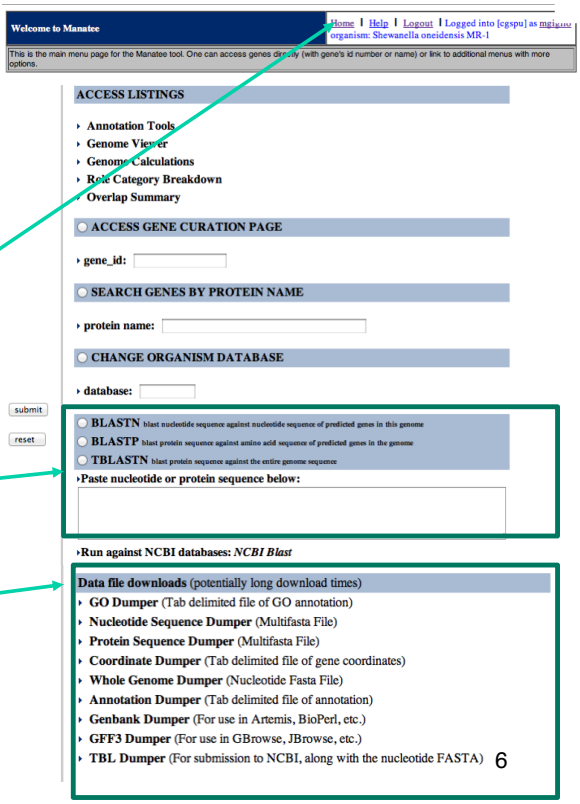
After logging into Manatee, you come to the "Welcome to Manatee" page. Here you will find several menu and search options to choose from.

I will discuss the menu options in more detail in following slides. You can search using a gene id to access a curation page for that gene; you can search by a keyword in a protein name; and if you are working with more than one database you can shift to another database.

In the upper right hand corner of every Manatee page is a navigation bar:
-The "Home" link takes you back to the "Welcome to Manatee" page, from where ever you are within the Manatee tool.
-This area also shows you which database you are logged into, and who is logged in. Clicking on the login name will take you back to the login page.
-The "Help" link should go to page specific documentation. However, these pages are still under development.

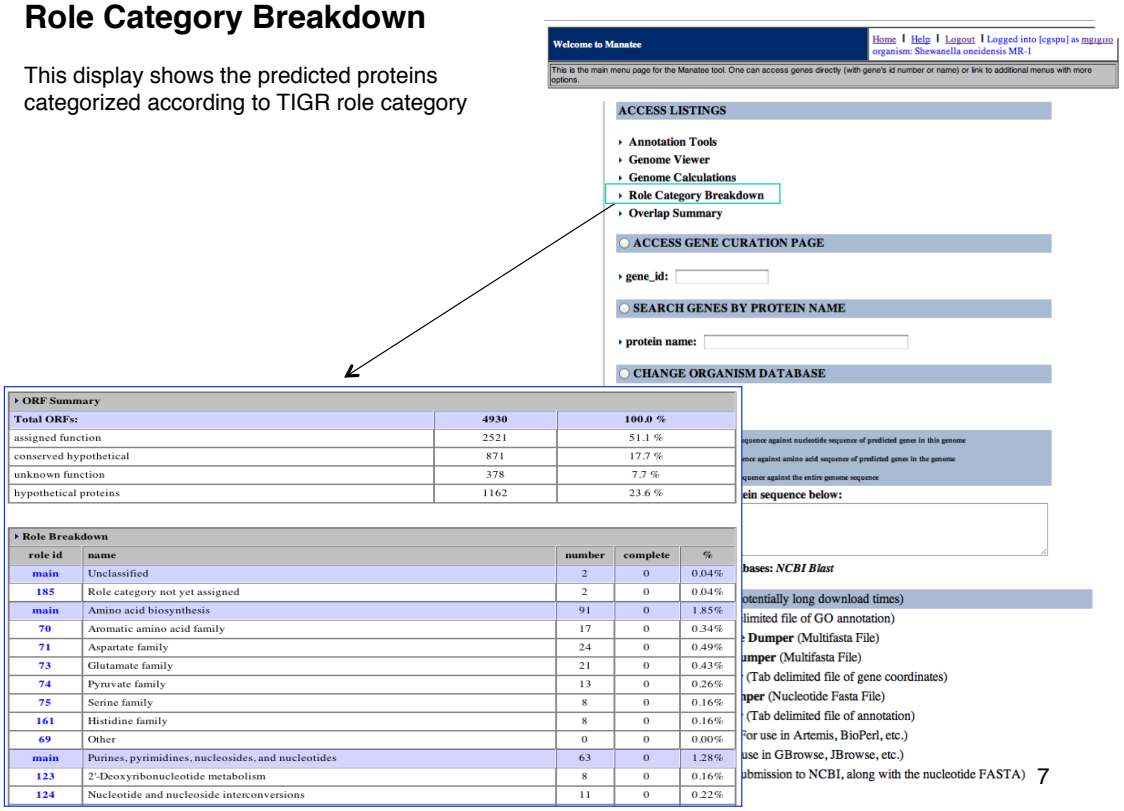
BLAST options
You can BLAST a sequence of interest against the predicted set of genes (nucleotide or protein) or against the entire genome sequence.

Data download options:
At the bottom of the Welcome page are several options for downloading text files containing annotation information. Some of these take a long time to query and load so please be patient. Simply click on the line of interest and the download process will begin.



Role Category Breakdown

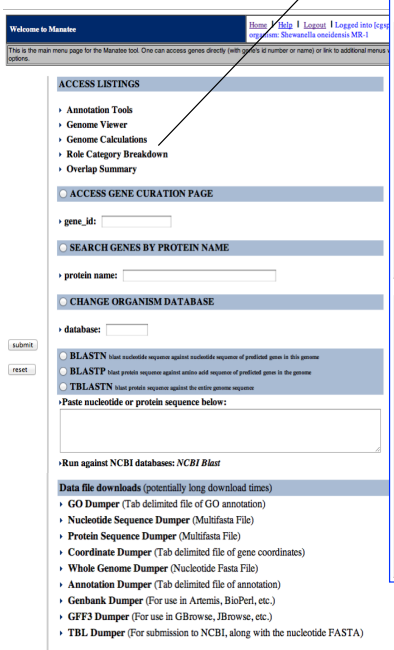
This display shows the predicted proteins categorized according to TIGR role category



ORF Summary		
Total ORFs:	4930	100.0 %
assigned function	2521	51.1 %
conserved hypothetical	871	17.7 %
unknown function	378	7.7 %
hypothetical proteins	1162	23.6 %

Role Breakdown				
role id	name	number	complete	%
main	Unclassified	2	0	0.04%
185	Role category not yet assigned	2	0	0.04%
main	Amino acid biosynthesis	91	0	1.85%
70	Aromatic amino acid family	17	0	0.34%
71	Aspartate family	24	0	0.49%
73	Glutamate family	21	0	0.43%
74	Pyruvate family	13	0	0.26%
75	Serine family	8	0	0.16%
161	Histidine family	8	0	0.16%
69	Other	0	0	0.00%
main	Purines, pyrimidines, nucleosides, and nucleotides	63	0	1.28%
123	2'-Deoxyribonucleotide metabolism	8	0	0.16%
124	Nucleotide and nucleoside interconversions	11	0	0.22%

Genome Calculations



feature name	feature count	feature type
transcript	4849	transcript
tRNA	101	tRNA

start sites	number	percent
ATG:	4037 (2887)	83.3% (85.8%)
GTG:	501 (323)	10.3% (9.6%)
TTG:	311 (156)	6.4% (4.6%)
OTHER:	()	0.0% (0.0%)

Numbers in parentheses do not include hypothetical proteins

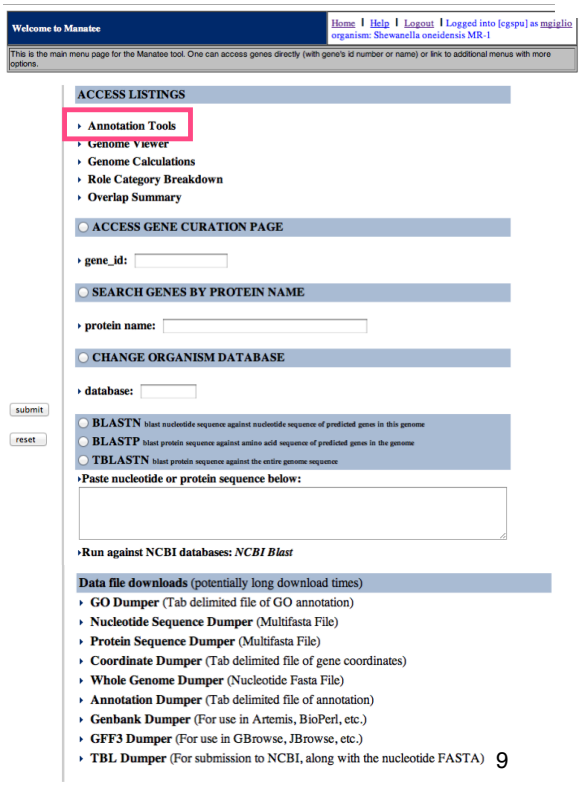
Information Table	
sequence id:	cgsp.assembly.2
type:	chromosome
molecule length:	161613 bp
GC content:	43.7%
base frequencies:	(A) (C) (G) (T) 28.2% 21.4% 22.3% 28.1%
funny characters:	
ORF count:	170
average gene length:	734 nt
percent coding:	77.2%
percent coding OR tRNA, rRNA, or repeat:	77.2%

Information Table	
sequence id:	cgsp.assembly.1
type:	chromosome
molecule length:	4969803 bp
GC content:	46%
base frequencies:	(A) (C) (G) (T) 27.0% 23.0% 23.0% 27.0%
funny characters:	R Y 2 6
ORF count:	4679
average gene length:	904 nt
percent coding:	85.2%
percent coding OR tRNA, rRNA, or repeat:	85.2%

“Annotation Tools”

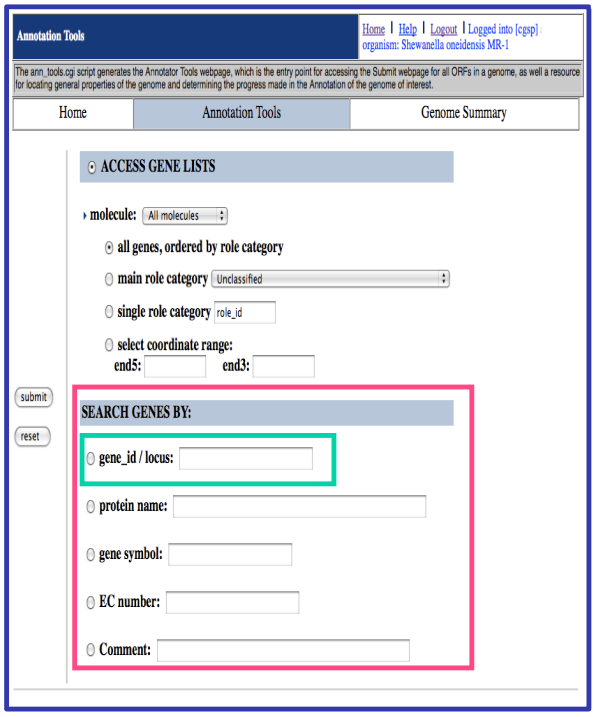
The Annotation Tools section contains most of the tools used during the process of manual annotation.

Get there by clicking “Annotation Tools” on the “Welcome to Manatee” page.



Annotation Tools Page: “**Search Genes By: gene_id/ locus**”

This option will take you directly to a page containing gene specific information called the “Gene Curation Page” or “GCP” for short. The GCP displays most of what knowledge we have about a given protein - you will be seeing this page in much more detail later. For now just know that you can reach this page by entering either a gene_id or locus id (e.g. ghi_1234, xyz_23) into this box and then clicking “submit”. The gene_ids displayed in Manatee will be locus ids if those are available, or they will be internal tracking ids that are used prior to locus id assignments. Locus ids (loci) are assigned to proteins sequentially from the origin of replication of the genome (if it can be identified). Loci are unique accessions and are used for public release and display of the proteins.



Annotation Tools Page

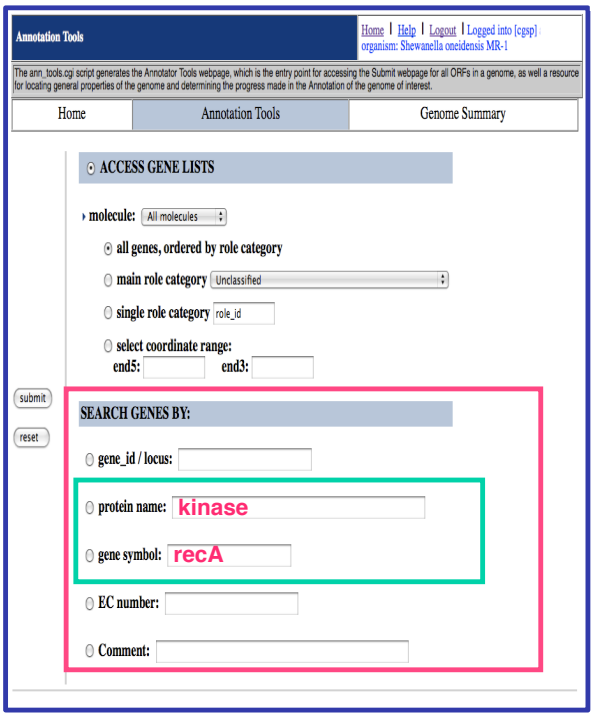
Search genes by: protein name or gene symbol

This is a keyword-based search for the common names and gene symbols that have been given to the genes/proteins
Whatever keyword you enter will be treated as though it has wildcards flanking it. This means that you will get results that include names with your keyword as an individual word and names with words that contain your keyword.
For example, if you search with “kinase”

you could get these:
“adenylate kinase”
“protein kinase”
“sensor histidine kinase”

as well as these:
“glutamate 5-kinase”
“phosphoenolpyruvate carboxykinase”
“ribose-phosphate pyrophosphokinase”

The results will be in the form of a table containing additional information and links to other pages - this table format will be described later.



Annotation Tools Page

Search Gene By: EC number or Comments

EC number

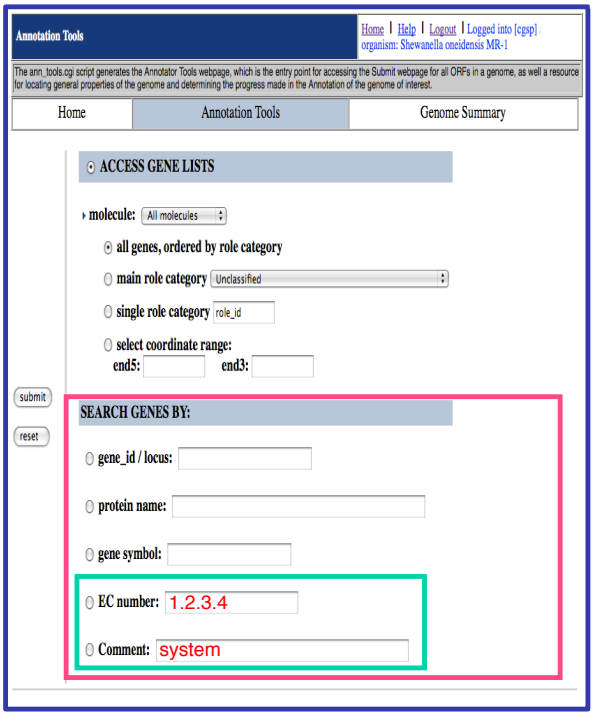
The Enzyme Commission maintains a database of enzymatic reactions which are each assigned an accession number of this format: 1.17.3.2
(this is the id number for xanthine oxidase)

Each position in the number indicates an additional level of specificity, a four position number is the most specific level and identifies a specific enzyme.
For more information go to:
www.chem.qmul.ac.uk/iubmb

For the search, enter an EC number to see a list of all genes in the genome that have been annotated with that particular EC number.

Comment

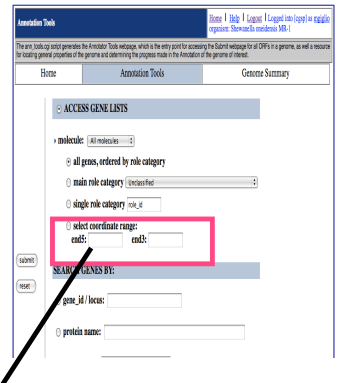
Annotators can add comments containing notes they wish to store for each gene (see Gene Curation Page section of the tutorial). These can be searched by entering text into the box at the right.



Annotation Tools Page
"Access gene lists by coordinate range" search:

Input a coordinate range and you will get a list of genes whose coordinates fall anywhere in that range.

If the genome consists of more than one molecule results from all molecules will be shown



C	seq id	gene_id	locus	end5	end3	gene name	gene symbol	ec	other roles
	CGSP.assembly.1	CGSP_9		9539	10621	ssDNA and dsDNA binding, ATP binding	recF		
	CGSP.assembly.2	CGSP_4807		11286	10078	SOS mutagenesis and repair		2.7.7.7	
	CGSP.assembly.2	CGSP_4808		11739	11335	SOS mutagenesis error-prone repair processed to UmuD' forms complex with UmuC	umuD		
	CGSP.assembly.2	CGSP_4812		16168	16365	orf6 protein	orf6		370
	CGSP.assembly.2	CGSP_4818		19313	19453	plasmid maintenance system antidote protein HigA	higA		
	CGSP.assembly.2	CGSP_4821		21028	20759	recombinase, resolvase family			
	CGSP.assembly.1	CGSP_18		23027	22815	sirA-like family protein			
	CGSP.assembly.2	CGSP_4825		24778	25233	ycfB protein	ycfB		
	CGSP.assembly.2	CGSP_4826		25235	25723	expressed protein			
	CGSP.assembly.1	CGSP_22		28404	29018	elongation factor			
	CGSP.assembly.1	CGSP_24		30558	31082	protoporphyrinogen oxidase	hemG	1.3.3.4	
	CGSP.assembly.2	CGSP_4834		30832	31083	antitoxin module of toxin-antitoxin system, PemI	pemI		
	CGSP.assembly.1	CGSP_26		31656	32210	protoporphyrinogen oxidase	hemG	1.3.3.4	
	CGSP.assembly.2	CGSP_4835		31662	32027	expressed periplasmic protein			
	CGSP.assembly.1	CGSP_36		41557	42123	protein involved in synthesis of threonylcarbamoyladenosine-modified tRNA			

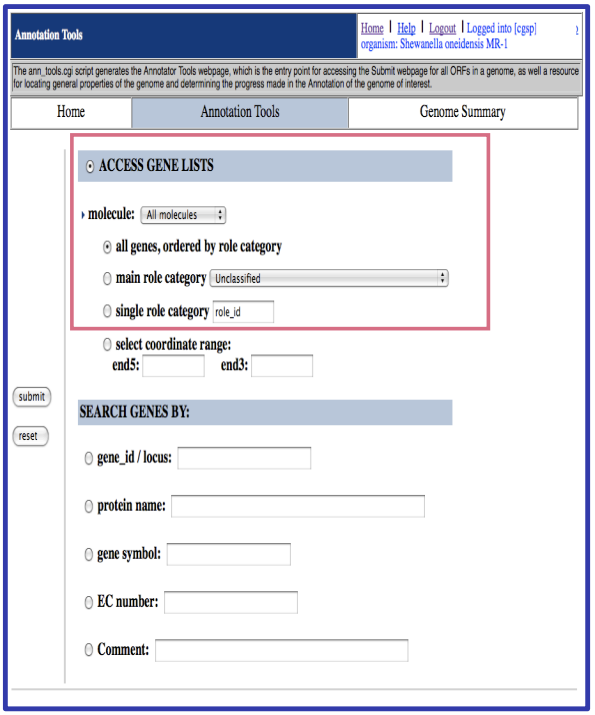
"Annotation Tools": "Access Gene Lists" by role categories

This tool will create a table of genes chosen according to the options in the red box at right. This tool allows one to view the genes organized by TIGR role category.

The first option to select in this section is which molecule you wish to annotate. Some genomes consist of just one chromosome and nothing else, while others can have multiple chromosomes and/or one or more plasmids. If multiple DNA molecules exist for the genome in question, the pull down menu at the top of this section will list them along with their id number. The default selection is "All molecules". To choose just one of the molecules, simply select it from the pull-down menu.

Next, choose one of the 3 options for which role categories you want to see genes from with the toggle buttons: first you can choose all role categories, second you can choose one particular main role category, and third you can choose one particular sub-role category. All of the mainrole categories are listed in the pull-down menu in the main role category selection, to choose one, simply highlight it. In order to select a particular sub-role category you must enter into the box next to "single role category" the id number of the sub-role category. There is a listing of all of the TIGR role categories and their id numbers on the next two pages of this tutorial.

Once you have chosen your desired options, click submit to see a list of the genes that fit your selections.



TIGR Role Categories - Page 1

Unclassified (the automated program was unable to assign a role to these)
185 Role category not yet assigned

Amino acid biosynthesis
70 Aromatic amino acid family
71 Aspartate family
73 Glutamate family
74 Pyruvate family
75 Serine family
161 Histidine family
69 Other

Purines, pyrimidines, nucleosides, and nucleotides
123 2'-Deoxyribonucleotide metabolism
124 Nucleotide and nucleoside interconversions
125 Purine ribonucleotide biosynthesis
126 Pyrimidine ribonucleotide biosynthesis
127 Salvage of nucleosides and nucleotides
128 Sugar-nucleotide biosynthesis and conversions
122 Other

Fatty acid and phospholipid metabolism
176 Biosynthesis
177 Degradation
121 Other

Biosynthesis of cofactors, prosthetic groups, and carriers
77 Biotin
78 Folic acid
79 Heme, porphyrin, and cobalamin
80 Lipoate
81 Menaquinone and ubiquinone
82 Molybdopterin
83 Pantothenate and coenzyme A
84 Pyridoxine
85 Riboflavin, FMN, and FAD
86 Glutathione
162 Thiamine
163 Pyridine nucleotides
191 Chlorophyll
707 Siderophores
76 Other

Central intermediary metabolism
100 Amino sugars
698 One-carbon metabolism
103 Phosphorus compounds
104 Polyamine biosynthesis
106 Sulfur metabolism
179 Nitrogen fixation
160 Nitrogen metabolism
709 Electron carrier regeneration
102 Other

Energy metabolism
108 Aerobic
109 Amino acids and amines
110 Anaerobic
111 ATP-proton motive force interconversion
112 Electron transport
113 Entner-Doudoroff
114 Fermentation
116 Glycolysis/gluconeogenesis
117 Pentose phosphate pathway
118 Pyruvate dehydrogenase
119 Sugars
120 TCA cycle
159 Methanogenesis
105 Biosynthesis and degradation of polysaccharides
164 Photosynthesis
180 Chemoautotrophy
184 Other

Transport and binding proteins
142 Amino acids, peptides and amines
143 Anions
144 Carbohydrates, organic alcohols, and acids
145 Cations and iron carrying compounds
146 Nucleosides, purines and pyrimidines
182 Porins
147 Other
141 Unknown substrate

TIGR Role Categories - Page 2

DNA metabolism
132 DNA replication, recombination, and repair
183 Restriction/modification
131 Degradation of DNA
170 Chromosome-associated proteins
130 Other

Transcription
134 Degradation of RNA
135 DNA-dependent RNA polymerase
165 Transcription factors
166 RNA processing
133 Other

Protein synthesis
137 tRNA aminoacylation
158 Ribosomal proteins: synthesis and modification
168 tRNA and rRNA base modification
169 Translation factors
136 Other

Protein fate
97 Protein and peptide secretion and trafficking
140 Protein modification and repair
95 Protein folding and stabilization
138 Degradation of proteins, peptides, and glycopeptides
189 Other

Regulatory functions
261 DNA interactions
262 RNA interactions
263 Protein interactions
264 Small molecule interactions
129 Other

Signal transduction
699 Two-component systems
700 PTS
710 Other

Cell envelope
91 Surface structures
89 Biosynthesis and degradation of murein sacculus and peptidoglycan
90 Biosynthesis and degradation of surface polysaccarides and lipopolysaccharides
88 Other

Cellular processes
93 Cell division
188 Chemotaxis and motility
701 Cell adhesion
702 Conjugation
96 Detoxification
98 DNA Transformation
705 Sporulation and Germination
94 Toxin production and resistance
187 Pathogenesis
149 Adaptations to atypical conditions
706 Bioosynthesis of natural products
92 Other

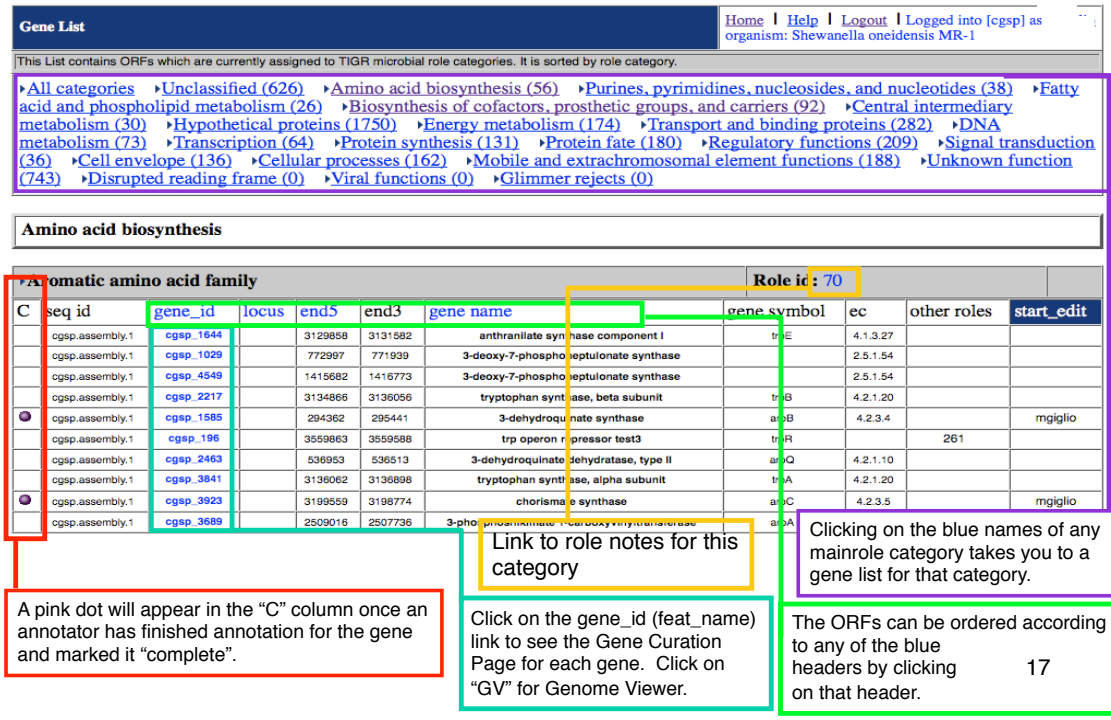
Mobile and extrachromosomal element functions
186 Plasmid functions
152 Prophage functions
154 Transposon functions
708 Other

Unknown
703 Enzymes of unknown specificity
157 General

Hypothetical
156 Conserved
704 Domain
856 General

Disrupted reading frame
270 NULL

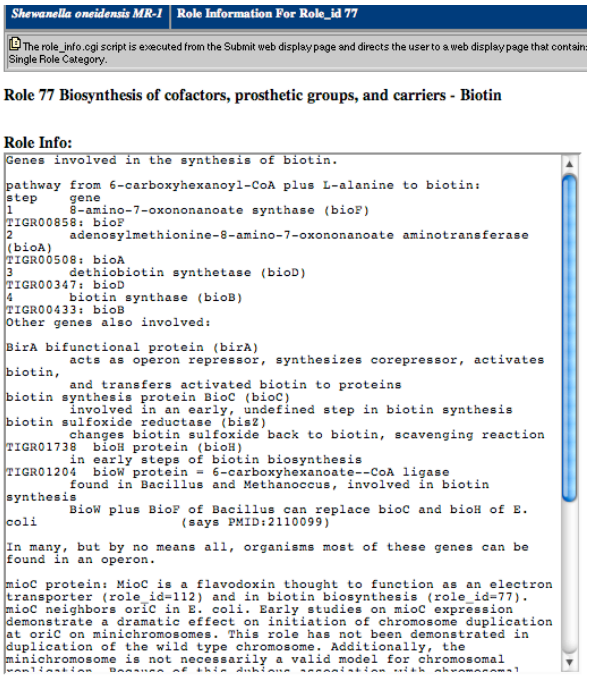
Gene List: The results of your selection from the Access Listings tool are displayed in a gene list containing gene id number, locus (if available), coordinates of the gene (end5, end3), common name of the gene/protein, gene_sym, EC number, and other roles for the protein. Not all of these fields will be populated for every gene. The genes are organized by role category (if your selection included more than one.) There are many features of the gene list, and much information displayed - text describing a feature is boxed in the same color as the feature itself.



Gene list link: Role information page:

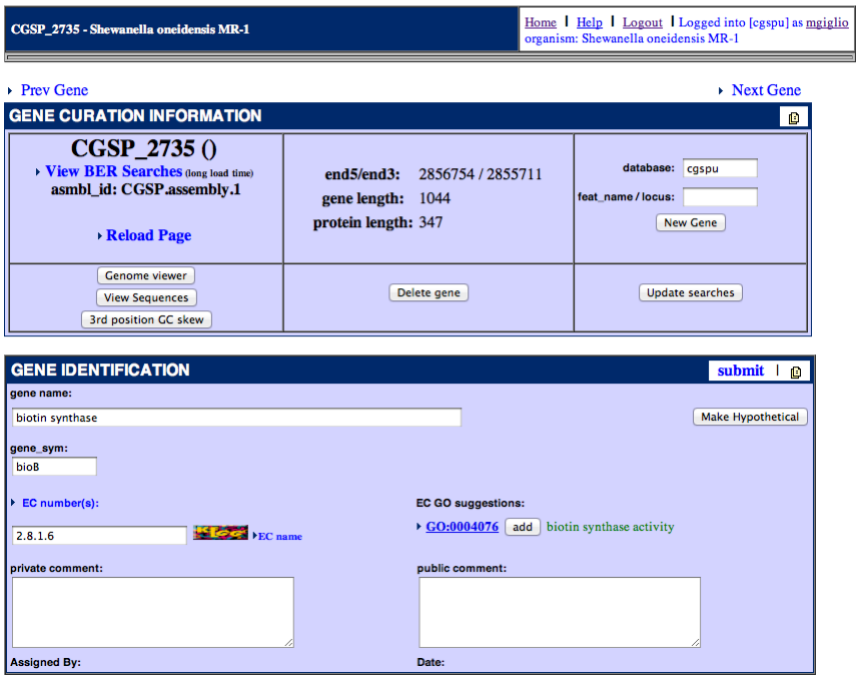
TIGR annotators expert in particular role categories have written "role notes" to aid new annotators and annotators unfamiliar with the category in the annotation process. These notes contain information on what genes belong in the category and what genes don't, on the pathways found in particular categories, and on the TIGR naming conventions for proteins within the category.

The utility of these documents has diminished as metabolic pathway reconstruction tools and the Gene Ontology have become more prominent in the annotation process.



Gene Curation Page

The Gene Curation Page summarizes evidence and annotations for each protein. This page can be accessed within Manatee from many places: any gene list, the "Access Gene Curation Page" option on the Genome Summary/Annotation Tools pages, Genome Viewer, …. and more. The GCP contains a lot of information so we will look at it in sections. I will try to organize the descriptions of each section in roughly the same order that the concepts behind each section were reviewed in the Annotation Overview.



Gene Curation Page

Gene Curation Information

This section contains basic identifying information about the gene and some search and display options.

The **gene_id** of the gene is listed at the top of the page. The gene_id is followed in parentheses by the **locus name** (final loci are assigned to genes at the end of a project, once annotation is complete, but they may get temporary loci during the course of the project).

The **blue link** under these names is a link to a file containing the BER search results for this gene (see later slide). There is another link to this page further down the orf info page (will be seen in a later slide).

To the right of the ORF names is a box containing **coordinates, length, and molecular weight (if available)**. "end5" is the 5' coordinate for the beginning of the coding sequence, "end3" is the 3' coordinate for the end of the coding sequence.

Finally on the extreme right is a box allowing you to move to another ORF info page by typing in the feat_name or locus in the box and clicking "**new gene**". One can also change to an orf in a different genome by **changing the database** in the database box, typing in the new orf number and clicking "new gene".

If you want to reload the GCP, use the "**Reload Page**" link in this section. Do not use the browser's reload button.

Gene Curation Page
Gene Identification

Initial information for this section comes from the automatic annotation pipeline and pFunc. The manual annotation then confirms or changes the information.

gene name: the descriptive name given to the protein
gene sym: the gene symbol for the protein (in this case bioB) (we default to E. coli gene symbols when possible and B. subtilis for Gram + specific things)
EC#: If the protein is an enzyme, we store the Enzyme Commission number. See later slides for info on ECGO term suggestions.
private comment: a field for annotators to note information for later reference by themselves or other annotators. A good place to keep notes.
public comment: comments meant to go out with our public accessions .

Gene Curation Page - BER Skim and Characterized Match

The BTAB SKIM section shows the top hits from the BER search file (see Annotation Overview presentation for more information on BER searches). The first column is the UniRef100 accession number of the match, the second column is percent identity with the match, the third is the length of the matching region (in nucleotides), the fourth is the name of the match protein, the fifth is the most specific taxon common to all the proteins in the UniRef100 cluster, and the sixth column is the p-value from the BLAST search. If the background color in the description column is green, then the match protein is "trusted". Clicking on the **blue accession number** will automatically populate the field in the characterized match section with that accession which can then be used as GO evidence. Clicking on the **blue names of the proteins** in the skim will take you to a page with just the alignment to that protein. The blue **"View BER searches"** link at the top of the skim section will take you to a file of all of the pairwise alignments from the BER search (see later slide). By default the first 25 matches are shown – to see all matches click on the "Show all" button.

BER SKIM Show all submit

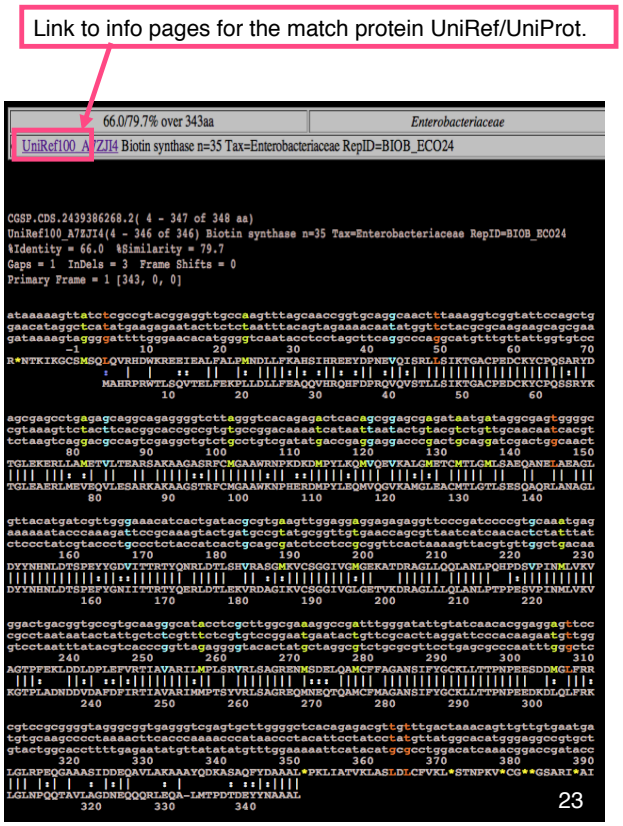
View BER Searches (long load time) | search date: Mon Feb 21 09:55:54 2011

accession	%ID	length	description	taxon	p-value
UniRef100_A7ZJI4	66.0	342	Biotin synthase	Enterobacteriaceae	1e-127
UniRef100_A4Y7Y3	97.1	349	Biotin synthase	Shewanella	0
UniRef100_Q8EDK6	100.0	349	Biotin synthase	Shewanella oneidensis	0
UniRef100_A0KY79	98.0	349	Biotin synthase	Shewanella	0
UniRef100_B8EAJ2	95.7	349	Biotin synthase	Shewanella baltica	0
UniRef100_B0TJN8	88.9	349	Biotin synthase	Shewanella halifaxensis HAW-EB4	1e-171
UniRef100_A8H3I7	88.6	349	Biotin synthase	Shewanella pealeana ATCC 700345	1e-170
UniRef100_B8CQY2	87.7	349	Biotin synthase	Shewanella piezotolerans WP3	1e-170
UniRef100_Q084I8	86.0	349	Biotin synthase	Shewanella frigidimarina NCIMB 400	1e-167
UniRef100_Q12NN4	85.8	356	Biotin synthase	Shewanella denitrificans OS217	1e-165
UniRef100_A1S5I9	83.4	352	Biotin synthase	Shewanella amazonensis SB2B	1e-163
UniRef100_B1KPJ7	84.3	353	Biotin synthase	Shewanella woodyi ATCC 51908	1e-162
UniRef100_A3QDN8	82.9	349	Biotin synthase	Shewanella loihica PV-4	1e-162
UniRef100_A8FX10	82.6	353	Biotin synthase	Shewanella sediminis HAW-EB3	1e-160
UniRef100_D4ZLY1	82.6	353	Biotin synthase	Shewanella violacea DSS12	1e-159
UniRef100_A9DID7	81.8	353	Biotin synthase	Shewanella benthica KT99	1e-158
UniRef100_A0KIC6	70.3	342	Biotin synthase	Aeromonas hydrophila subsp. hydrophila ATCC 7966	1e-140
UniRef100_Q1MZV6	71.1	344	Biotin synthase	Bermanella marisrubri	1e-139
UniRef100_A4SPR7	69.8	342	Biotin synthase	Aeromonas salmonicida subsp. salmonicida A449	1e-138
UniRef100_C4LDC7	70.2	344	Biotin synthase	Tolumonas auensis DSM 9187	1e-137
UniRef100_Q2SBD4	69.3	349	Biotin synthase	Hahella chejuensis KCTC 2396	1e-135
UniRef100_Q5QZ16	68.9	342	Biotin synthase	Idiomarina loihiensis	1e-134
UniRef100_B7S225	69.1	334	Biotin synthase	marine gamma proteobacterium HTCC2148	1e-134
UniRef100_B3PI87	68.5	342	Biotin synthase	Cellvibrio japonicus Ueda107	1e-134
UniRef100_UPI0000E1102B	66.7	349		Glaciecola sp. HTCC2999	1e-133

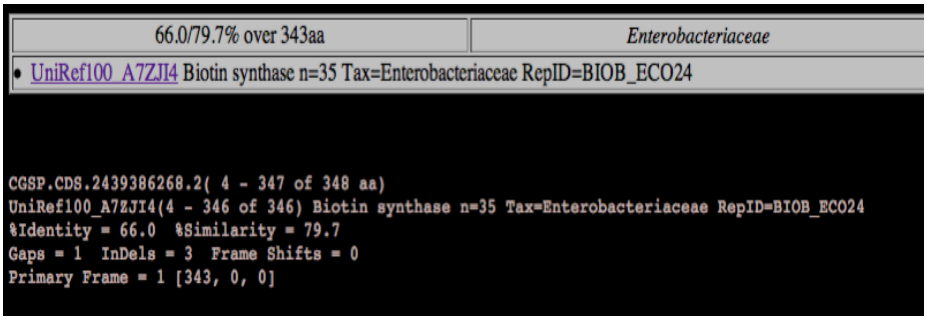
The BER alignment page

This page is accessible by clicking on the “View BER searches” link at the top of the Info page or at the top of the BTAB skim section. Here you will find multiple pairwise alignments of the genome protein to hits found in the BER search. Pages with alignments for one match per page can be accessed by clicking on the match protein name in the Skim. These load much more quickly.

In the header of each alignment will be listed the accession of the UniRef100 match cluster. This is a link that takes you to UniRef/UniProt.



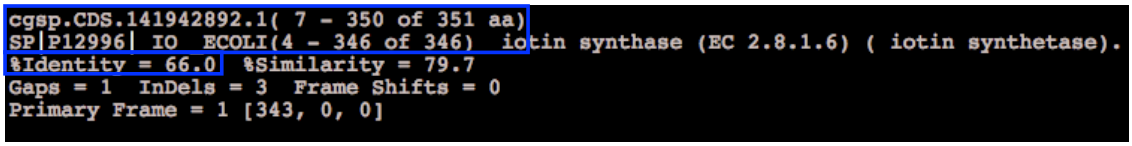
BER Alignment detail: Boxed Header



-The top bar lists the percent identity/similarity and the organism from which the protein comes (if available).

-The bottom section lists the accession number and name for the matching cluster from UniRef100.

BER Alignment detail: alignment header



-It is most important to look at the range over which the alignment stretches and the percent identity

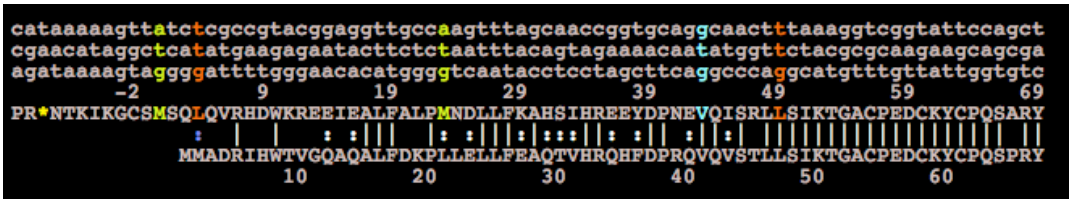
-The top line show the amino acid coordinates over which the match extends for our protein

-The second line shows the amino acid coordinates over which the match extends for the match protein, along with the name and accession of the match protein

-The last line indicates the number of amino acids in the alignment found in each forward frame for the sequence as defined by the coordinates of the gene. The primary frame is the one starting with nucleotide one of the gene. If all is well with the protein, all of the matching amino acids should be in frame 1.

-If there is a frameshift in the alignment (see overview) the phrase “Frame Shifts = #” will flash and indicate how many frameshifts there are.

BER Alignment detail: alignment of amino acids



-In these alignments the codons of the DNA sequence read down in columns with the corresponding amino acid underneath.

-The numbers refer to amino acid position. Position 1 is the first amino acid of the protein. The first nucleotide of the codon coding for amino acid 1 is nucleotide 1 of the coding sequence. Negative amino acid numbers indicate positions upstream of the predicted start of the protein.

-Vertical lines between amino acids of our protein and the match protein (bottom line) indicate exact matches, dotted lines (colons) indicate similar amino acids.

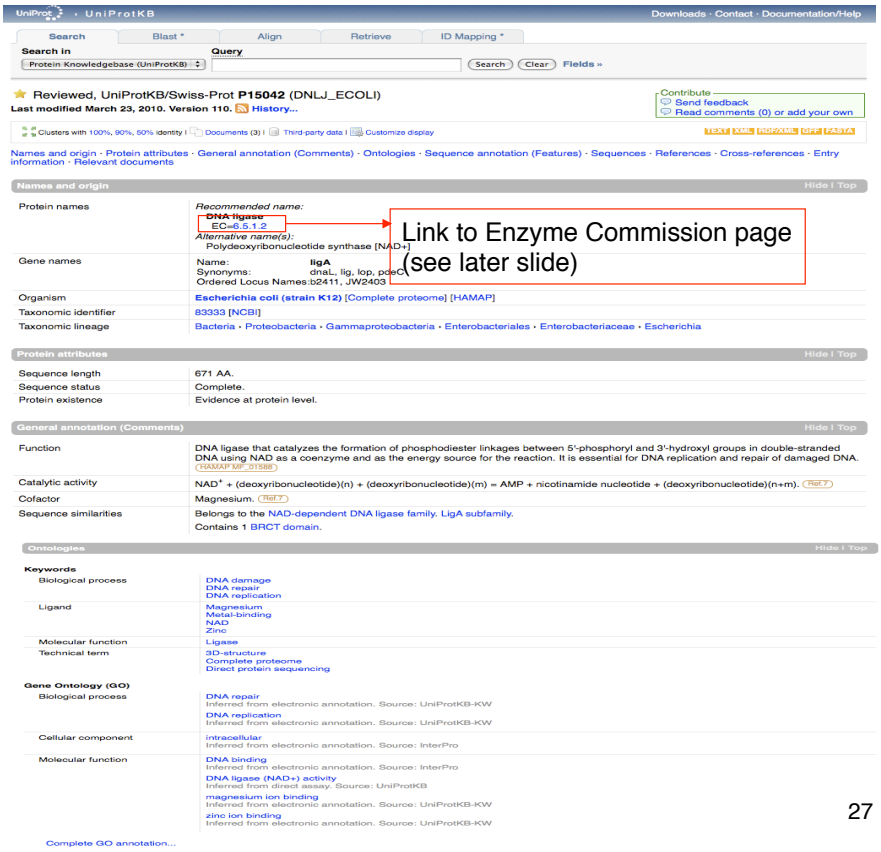
-Start sites are color coded: ATG is green, GTG is blue, TTG is red/orange

-Stop codons are represented as asterisks in the amino acid sequence. An open reading frame goes from an upstream stop codon to the stop at the end of the protein, while the gene starts at the chosen start codon.

UniProt- #1

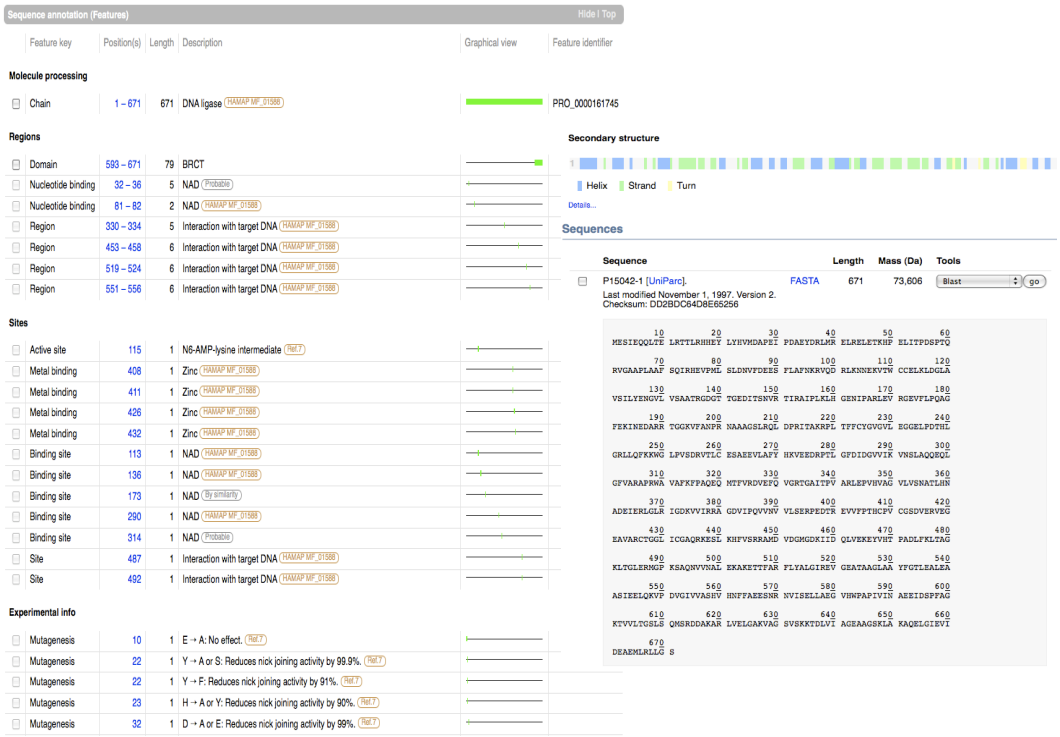
name, EC#
gene_symbol
Taxonomy

Info about function

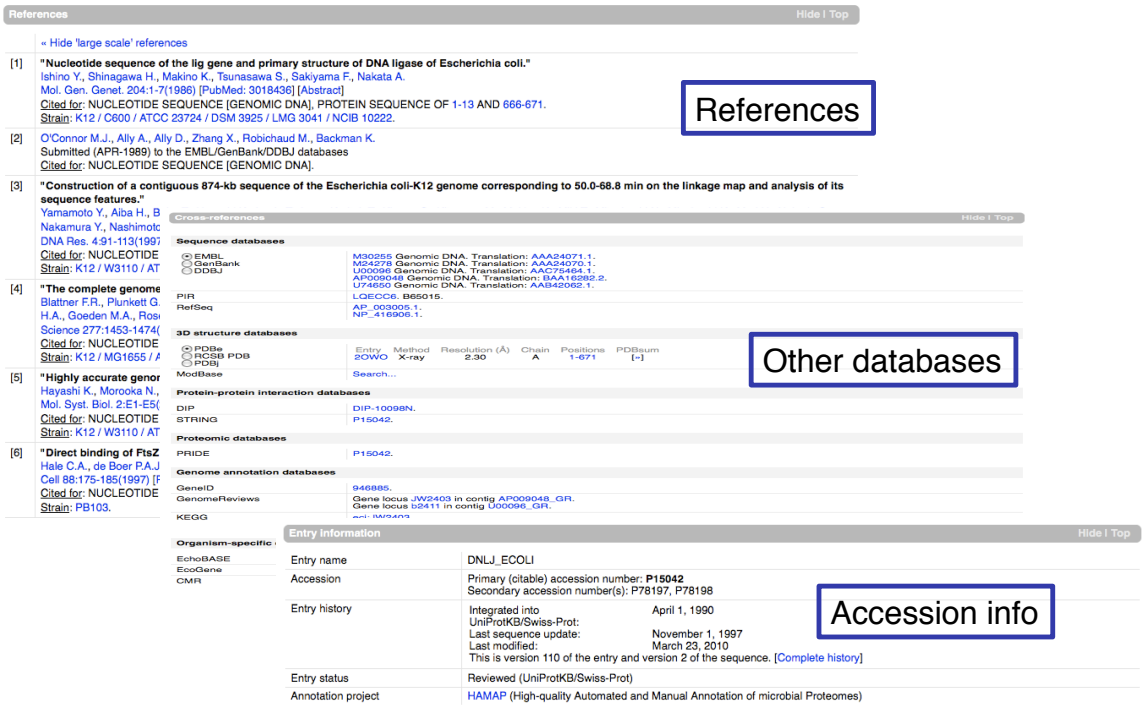


UniProt- #2

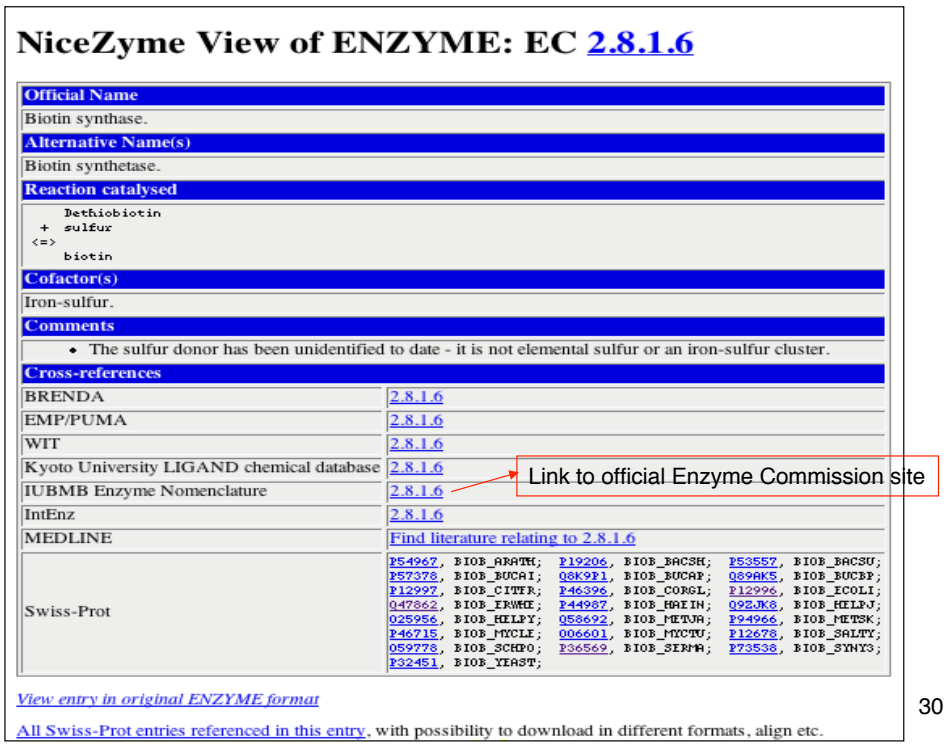
Sequence info



UniProt #3

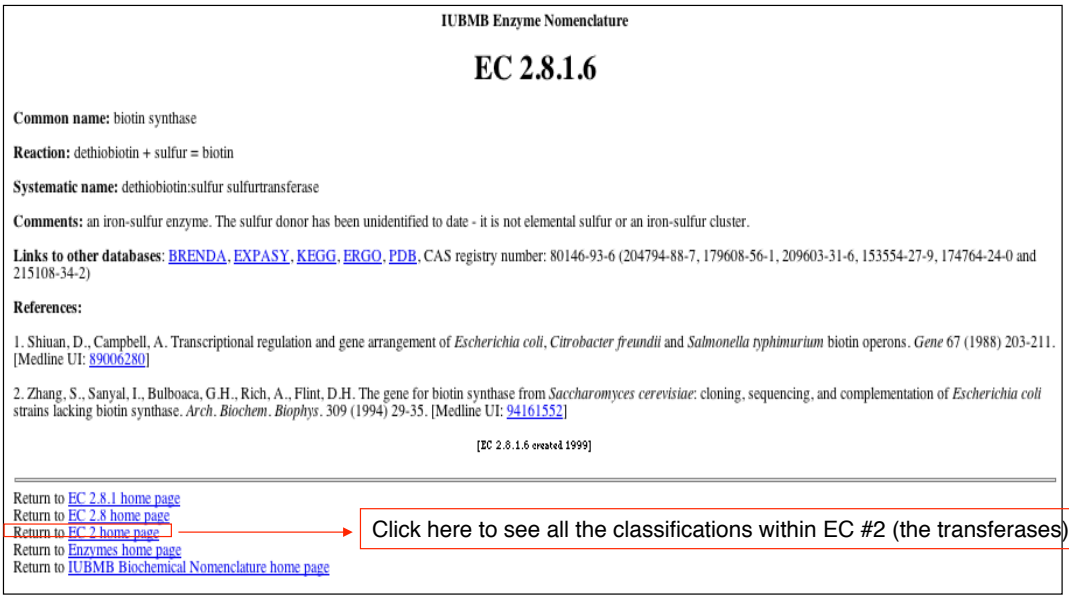


View of EC number info page from Swiss Institute of Bioinformatics site



View of information page for an EC number at IUBMB site

The Enzyme Commission (EC) is part of the IUBMB and is charged with maintaining the database of enzyme classifications. In the EC system, each reaction is assigned a 4 part accession number with each part consisting of an integer, where the numbers are separated by periods. As one moves from the first number to the second to the third to the fourth the nature of the reaction becomes more specific. For example: EC2.-.-.- = "transferase", 2.8.-.- = "transferase, transferring sulfur-containing groups", 2.8.1.- = "sulfurtransferases", and finally 2.8.1.6 = "biotin synthase" (a specific sulfurtransferase, which is a specific class of transferases that transfer sulfur-containing groups). One can see the breakdown of all of the classes within each EC first number (they only go up to 6) by clicking on the home page for each number (see below).



IUBMB Enzyme Nomenclature

EC 2.8.1.6

Common name: biotin synthase

Reaction: dethiobiotin + sulfur = biotin

Systematic name: dethiobiotin:sulfur sulfurtransferase

Comments: an iron-sulfur enzyme. The sulfur donor has been unidentified to date - it is not elemental sulfur or an iron-sulfur cluster.

Links to other databases: BRENDA, EXPASY, KEGG, ERGO, PDB, CAS registry number: 80146-93-6 (204794-88-7, 179608-56-1, 209603-31-6, 153554-27-9, 174764-24-0 and 215108-34-2)

References:

1. Shiuan, D., Campbell, A. Transcriptional regulation and gene arrangement of *Escherichia coli*, *Citrobacter freundii* and *Salmonella typhimurium* biotin operons. *Gene* 67 (1988) 203-211. [Medline UI: 89006280]

2. Zhang, S., Sanyal, I., Bulboaca, G.H., Rich, A., Flint, D.H. The gene for biotin synthase from *Saccharomyces cerevisiae*: cloning, sequencing, and complementation of *Escherichia coli* strains lacking biotin synthase. *Arch. Biochem. Biophys.* 309 (1994) 29-35. [Medline UI: 94161552]

[EC 2.8.1.6 created 1999]

Return to EC 2.8.1 home page
Return to EC 2.8 home page
Return to EC 2 home page
Return to Enzymes home page
Return to IUBMB Biochemical Nomenclature home page

Click here to see all the classifications within EC #2 (the transferases).

Gene Curation page - HMM hits scoring **above noise**

(Text describing the features of the HMM section is boxed in the same color as each feature.)

The blue id numbers for each HMM link to an info page for that HMM.

Key information is the isology type and the "total" and "cutoff" scores.

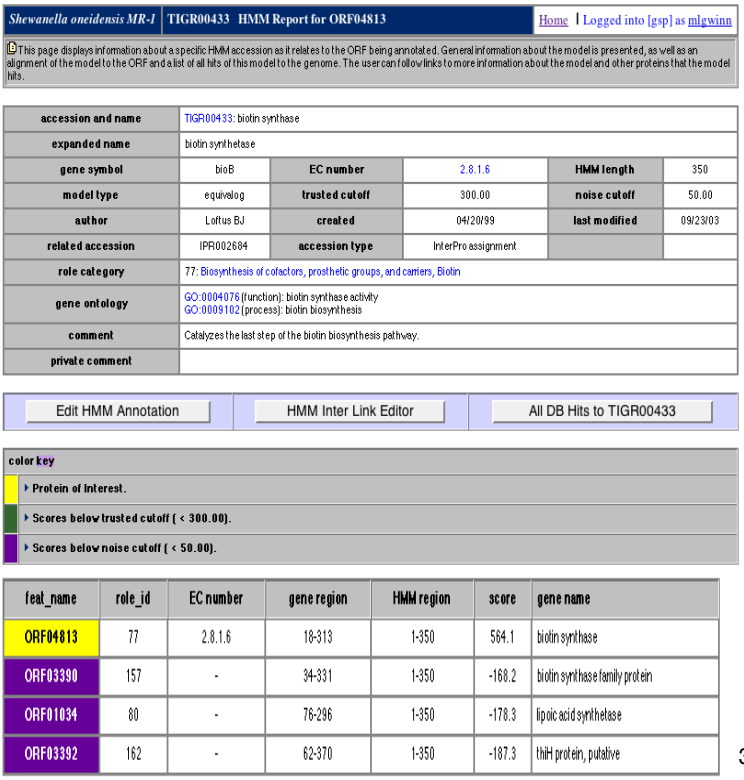
The "Add To GO Evidence" link automatically fills the HMM information into the "with" field in the GO term entry box.

GO terms assigned to each HMM are listed under the HMM (if any). Clicking on the "Add" button here adds not only the GO term id, but also the HMM evidence.

The "Add To Annotation" link will automatically copy the annotation from the HMM to the protein.

HMM report page - to get to this page click on an HMM accession number almost anywhere in Manatee

At the top is information about the HMM including HMM name, associated annotation (gene symbol, EC#, TIGR role, etc.) and comments from the authors.
Below is a list of all genes in the organism which hit the HMM and the scores they received. The row with the gold background is the protein of interest. Rows with a green background have scores below the trusted cutoff, rows with a purple background have scores below the noise cutoff.



Shewanella oneidensis MR-1 | TIGR00433 HMM Report for ORF04813 | Home | Logged into [gsp] as mlgwinn

This page displays information about a specific HMM accession as it relates to the ORF being annotated. General information about the model is presented, as well as an alignment of the model to the ORF and a list of all hits of this model to the genome. The user can follow links to more information about the model and other proteins that the model hits.

accession and name	TIGR00433: biotin synthase				
expanded name	biotin synthetase				
gene symbol	bioB	EC number	2.8.1.6	HMM length	350
model type	equivalog	trusted cutoff	300.00	noise cutoff	50.00
author	Loftus BJ	created	04/20/99	last modified	09/23/03
related accession	IPR002684	accession type	InterPro assignment		
role category	77: Biosynthesis of cofactors, prosthetic groups, and carriers, Biotin				
gene ontology	GO:0004076 (function): biotin synthase activity GO:0009102 (process): biotin biosynthesis				
comment	Catalyzes the last step of the biotin biosynthesis pathway.				
private comment					

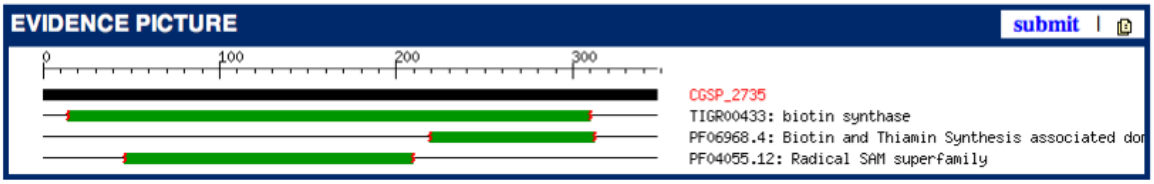
Edit HMM Annotation | HMM Inter Link Editor | All DB Hits to TIGR00433

color key
- Protein of interest.
- Scores below trusted cutoff (< 300.00).
- Scores below noise cutoff (< 50.00).

feat_name	role_id	EC number	gene region	HMM region	score	gene name
ORF04813	77	2.8.1.6	18-313	1-350	564.1	biotin synthase
ORF03390	157	-	34-331	1-350	-168.2	biotin synthase family protein
ORF01034	80	-	76-296	1-350	-178.3	lipoic acid synthetase
ORF03392	162	-	62-370	1-350	-187.3	thiH protein, putative

Gene Curation Page - Evidence Picture - ORF04813

All of the evidence stored for an ORF is displayed in this graphic. The black bar represents the ORF in question. Green bars represent HMMs which hit the ORF above trusted cutoff. Green HMM bars indicate above trusted score, orange indicates above noise but below trusted, red indicates below noise and is generally not shown. Clicking on the colored bars in the graphic opens windows with additional information on that piece of evidence. The evidence picture for ORF04813 does not contain all of the possible evidence types, so later slides will show some evidence pictures from other genes.

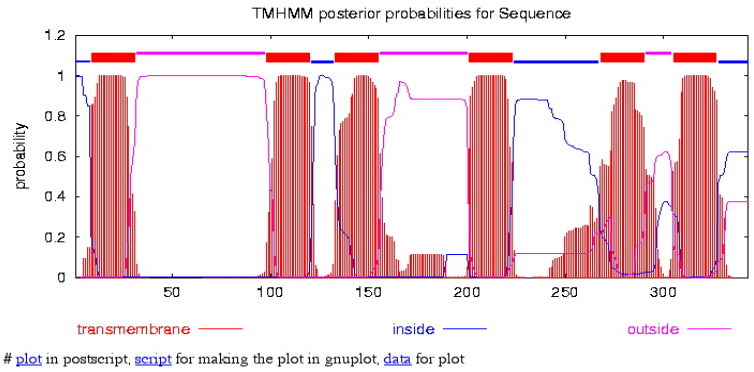


NOTE: this display is for ORF03779

TMHMM result

HELP with output formats

```
# Sequence Length: 343
# Sequence Number of predicted TMHs:  6
# Sequence Exp number of AAs in TMHs: 139.48261
# Sequence Exp number, first 60 AAs:  20.99155
# Sequence Total prob of N-in:        0.99734
# Sequence POSSIBLE N-term signal sequence
Sequence        TMHMM2.0        inside       1     8
Sequence        TMHMM2.0        TMhelix      9    31
Sequence        TMHMM2.0        outside     32    97
Sequence        TMHMM2.0        TMhelix     98   120
Sequence        TMHMM2.0        inside     121   132
Sequence        TMHMM2.0        TMhelix    133   155
Sequence        TMHMM2.0        outside    156   200
Sequence        TMHMM2.0        TMhelix    201   223
Sequence        TMHMM2.0        inside     224   267
Sequence        TMHMM2.0        TMhelix    268   290
Sequence        TMHMM2.0        outside    291   304
Sequence        TMHMM2.0        TMhelix    305   327
Sequence        TMHMM2.0        inside     328   343
```



plot in postscript, script for making the plot in gnuplot, data for plot

Gene Curation Page - PROSITE and Signal P sections on the GCP

NOTE: this display is for a different protein

Click here to see info on PROSITE motif.

PROSITE submit

PS01039: Bacterial extracellular solute-binding proteins, family 3 signature.

Match sequence: GFDIELAKQIAKDA

Coords	Precision	Recall	Curation	
52/65	0.76	0.93	☑	[Add To GO Evidence]

ATTRIBUTES submit

No Frameshifts Detected.

SIGNAL_P submit

SignalP-2.0 Results: [Graphical Display] [Raw output for SP-HMM/NN]

SignalP-2.0 HMM

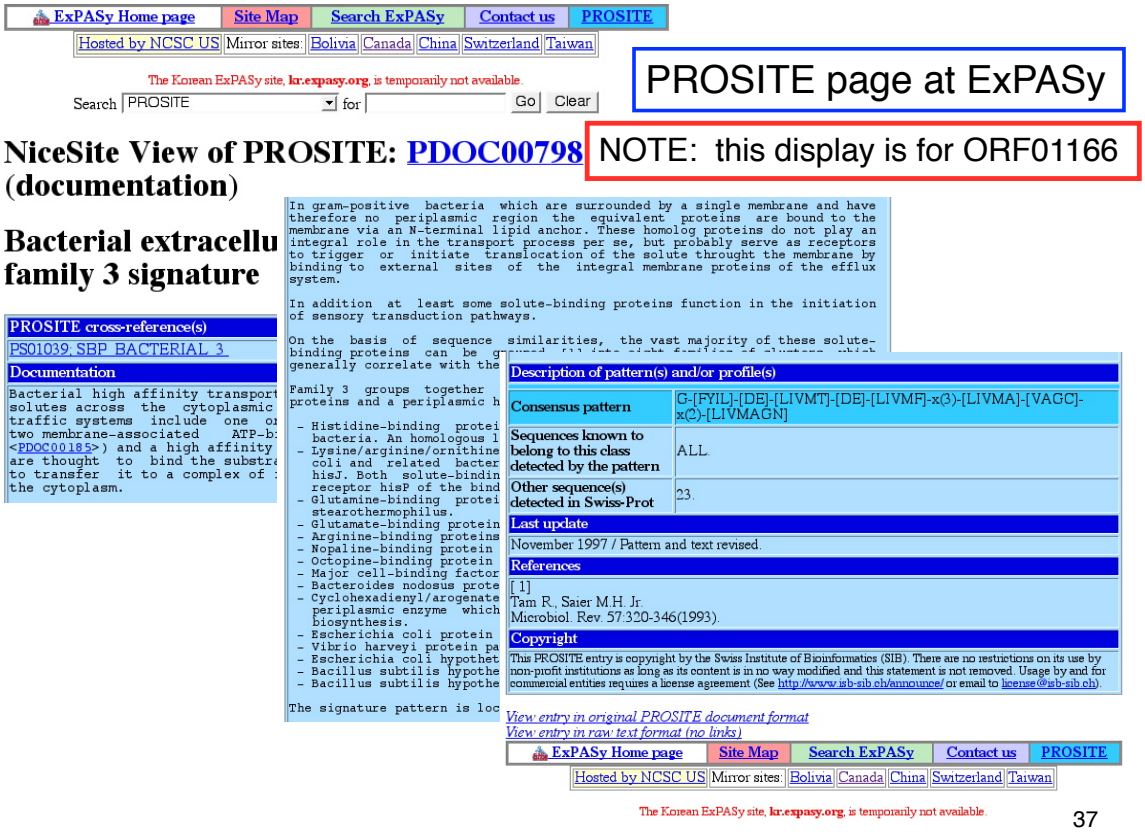
Prediction: No prediction generated ☐ Curated

Signal peptide probability: 0.984

Signal anchor probability:

Max cleavage site probability: 0.340

Click here to see output in graphical form.

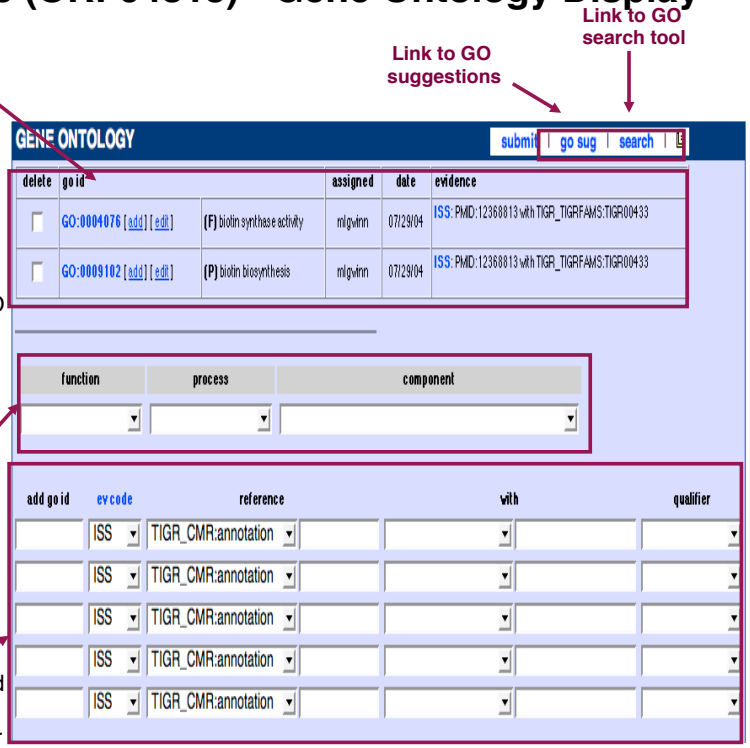


Gene Curation Page (ORF04813) - Gene Ontology Display

Current GO term assignments are listed in table.
-Click id # to see term in tree.
-Click box for GO term to be deleted.
-Click “add” to add additional evidence rows. (or click delete and add to completely redo evidence)
-Click “edit” to edit evidence.
-”Make ISS”(not seen in this example) can be used when the GO term and evidence assigned by AutoAnnotate are correct, clicking this button marks the old association for deletion and automatically puts in the new info for insertion.

These pull downs have commonly used GO terms. If you choose the unknown terms from any pull-down, the evidence will automatically fill in (since it is always the same.)

Fill in the fields in this section to add or change GO term assignments. These columns are detailed on later slides.



GO data entry columns:

The format for all GO data is carefully controlled by the GO. Manatee knows all of the formatting rules and will format the data for you whenever you use the “add” or suggestions buttons. (more on this later)

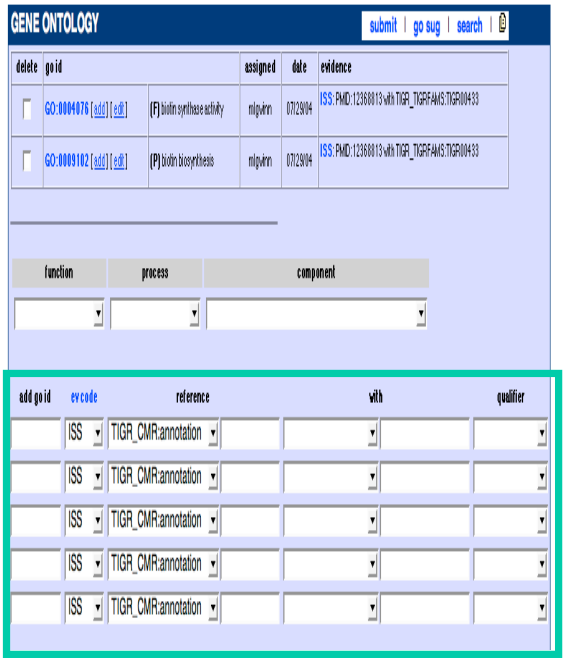
GO id - the format is GO:#######.

ev code - pick an evidence code from the pull down.

reference - identifier for publication or other accessible text that describes experiments, methods, or SOPs as appropriate for the annotation being made. Format is DB:identifier (e.g. PMID:1234567)

with - used with ISS, IPI, IGI, IC, IGC. Format is DB:identifier. (e.g. UniProt:P12345)

qualifier - only used with some annotations. contributes_to is only used when annotating function to a subunit of a complex



Gene Curation Page - GO suggestions and Auto-fill-ins

GO term suggestions and auto-fill-in buttons are located in several places on the Gene Curation Page:

-GO terms assigned to HMMs are listed under HMM hits (if any have been assigned - see the HMM slide for how these look). These are often excellent sources for GO terms. Clicking the “Add” button next to a GO term under an HMM adds both the term id and the evidence to the appropriate fields in the GO entry section. Clicking the “Add to GO evidence” button adds just the HMM accession into the “with” field in the GO entry section.

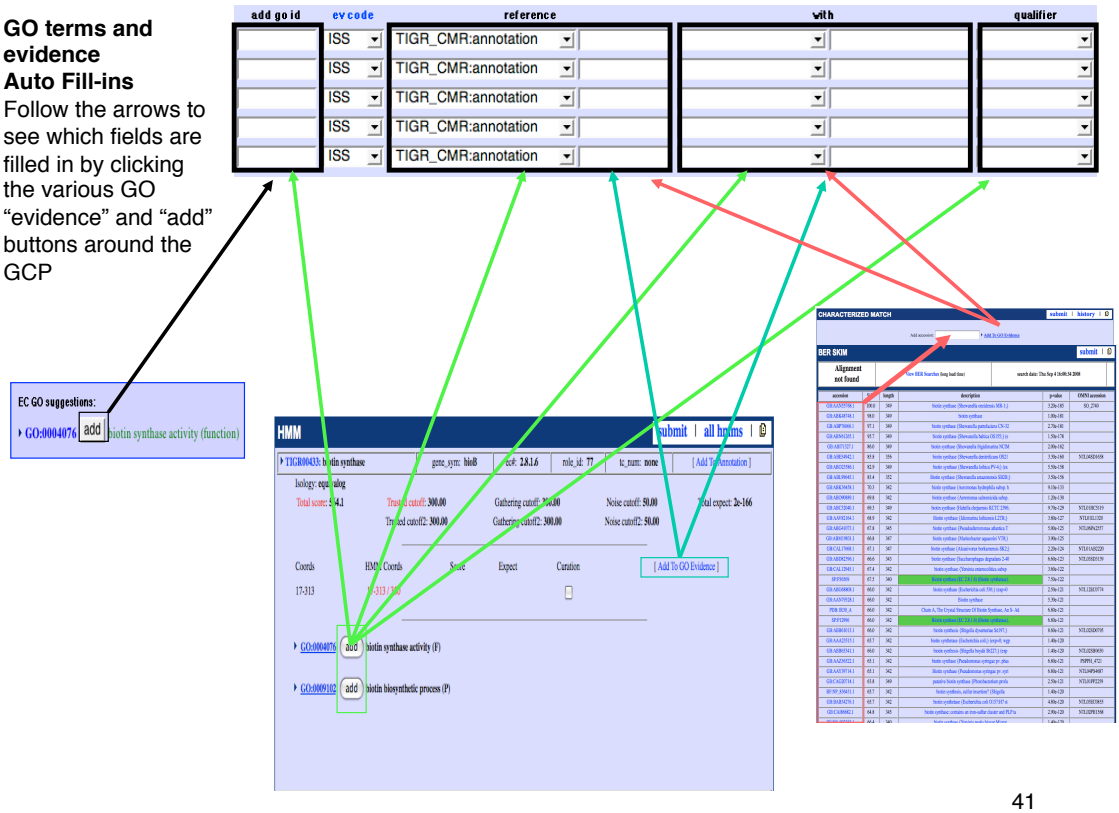
-GO terms corresponding to EC numbers are listed next to the EC box (for enzymes). Clicking the “add” button will put the GO term id into the “add go id” fields in the GO entry section.

-”Add to GO evidence” buttons are also available for Prosite hits, this populates the “with” field with the Prosite accession. Available when a protein has matches to Prosite.

-”Add to GO evidence” is also available for the characterized match accession, this will put the accession of the characterized matching protein into the “with” field entry box.

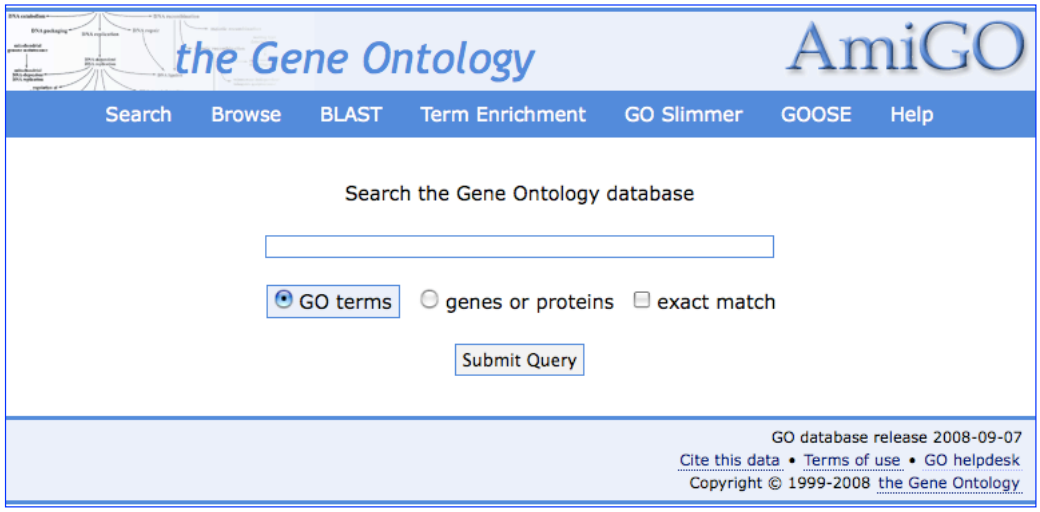
See next page for screen shots.

GO terms and evidence
Auto Fill-ins
Follow the arrows to see which fields are filled in by clicking the various GO "evidence" and "add" buttons around the GCP



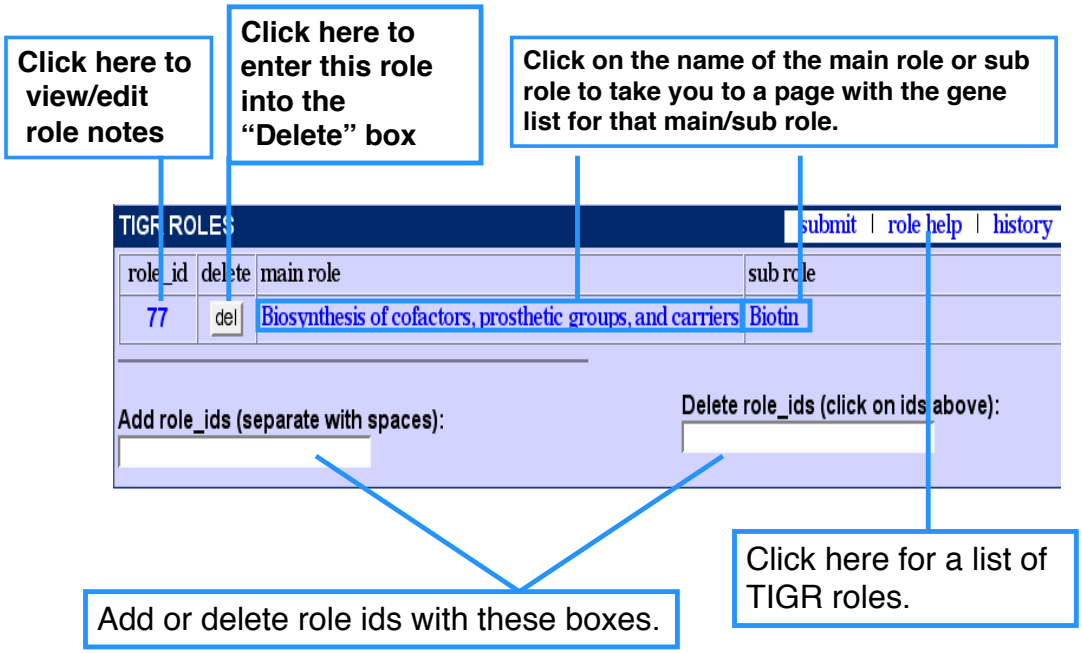
Searching for GO terms: the AmiGO search tool:

In many cases the GCP will not have a suggested GO term that meets an annotators needs. In that situation the annotator can click on "Search GO" in the header of the search section and use AmiGO to find terms.

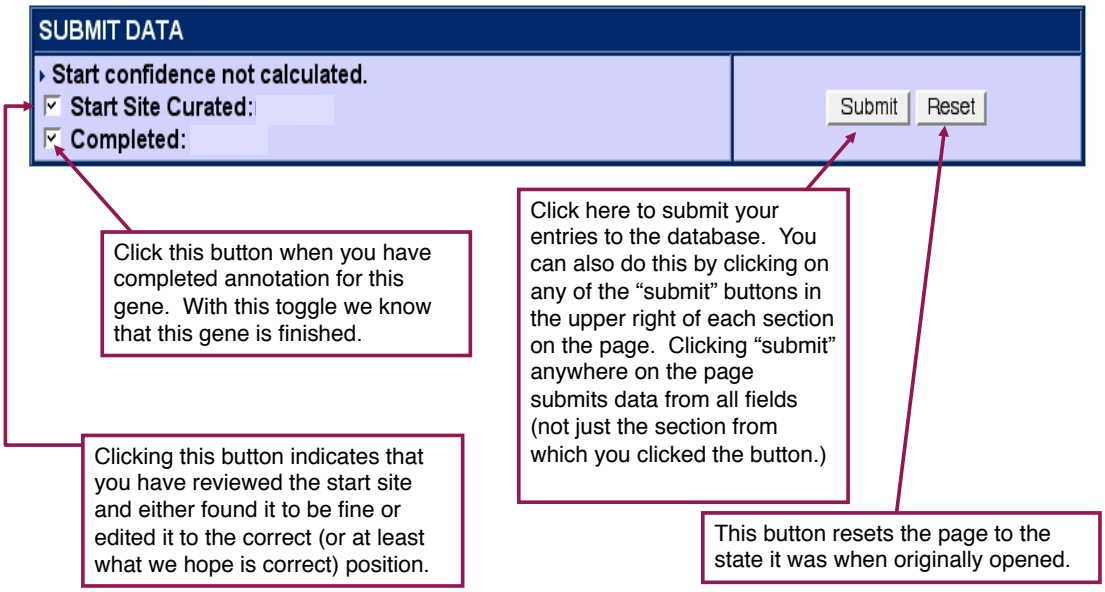


http://amigo.geneontology.org/

Gene Curation Page - TIGR roles



Gene Curation Page - How to get the data into the database: The "Submit" buttons



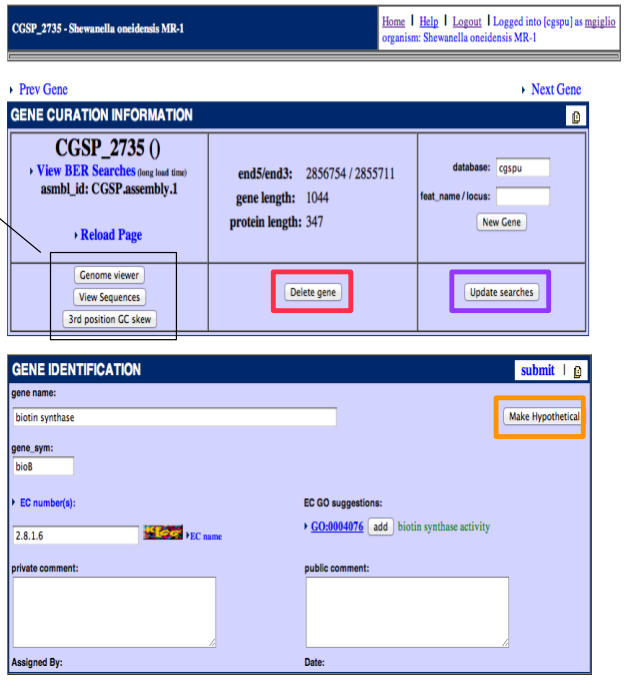
Gene Curation Page - The buttons at the top of the page

The three buttons on the left will be described on later pages.

The "delete gene" button:
This button will delete the gene model from the database, not just the annotation.

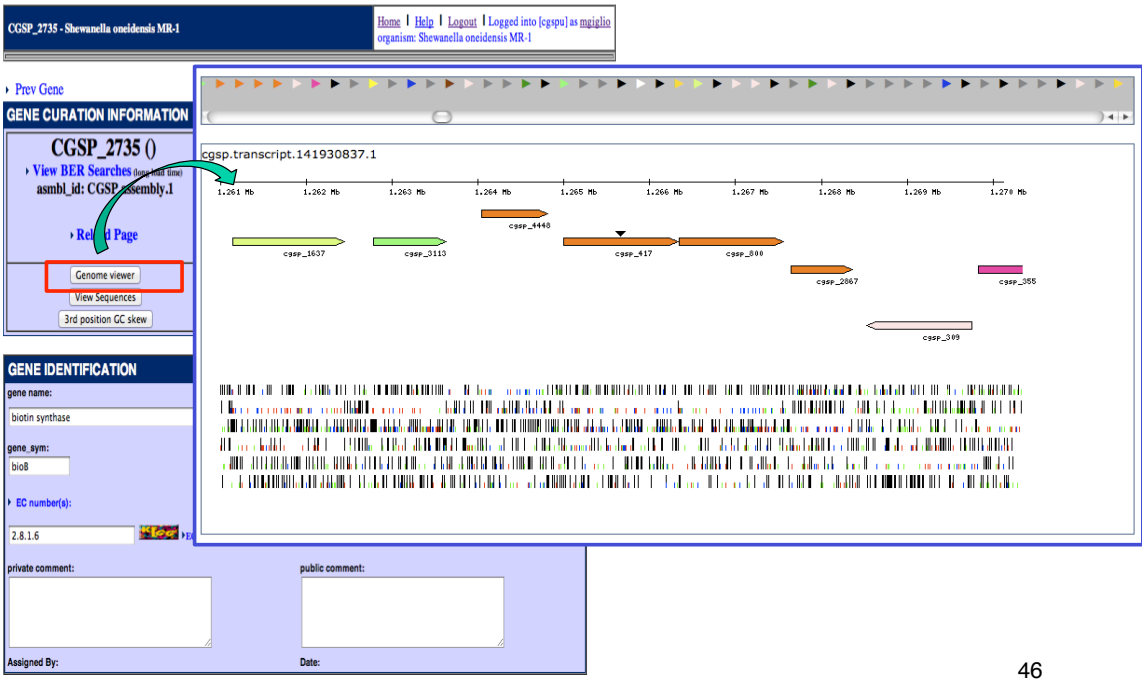
The "make hypothetical" button will delete the annotation for the gene, but leave the gene model intact.

The "update searches" button will launch a new round of evidence searches for this particular protein. It will take 10-15 minutes for this search to finish.

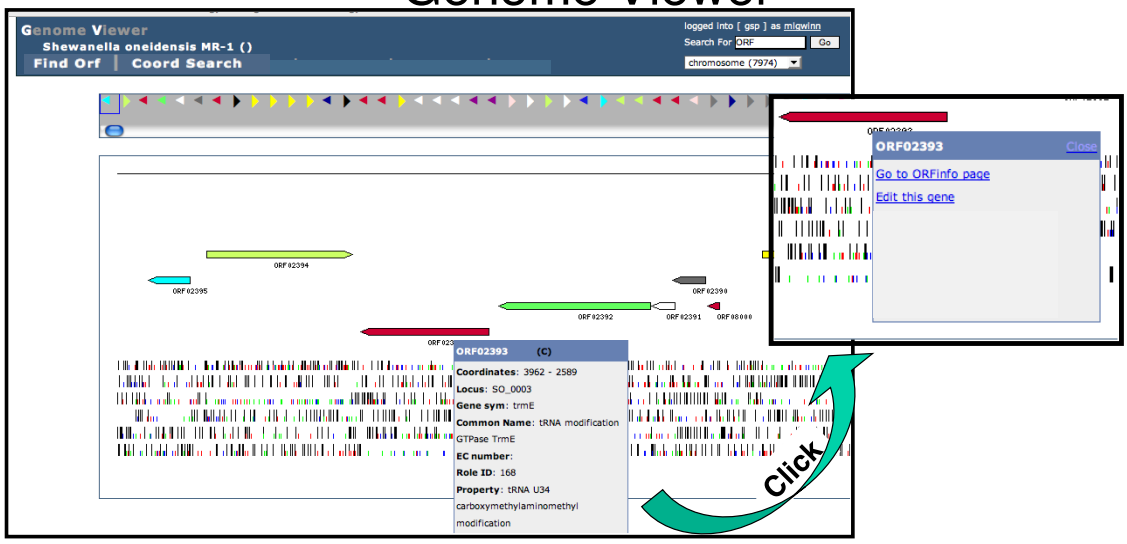


Genome Viewer

Access Genome Viewer from the Welcome to Manatee page or the pull down on the Gene Curation page. Genome Viewer provides a linear view of the coding genes and other features along the DNA molecule and provides a tool for gene model curation.



Gene Model Curation in Manatee: Genome Viewer



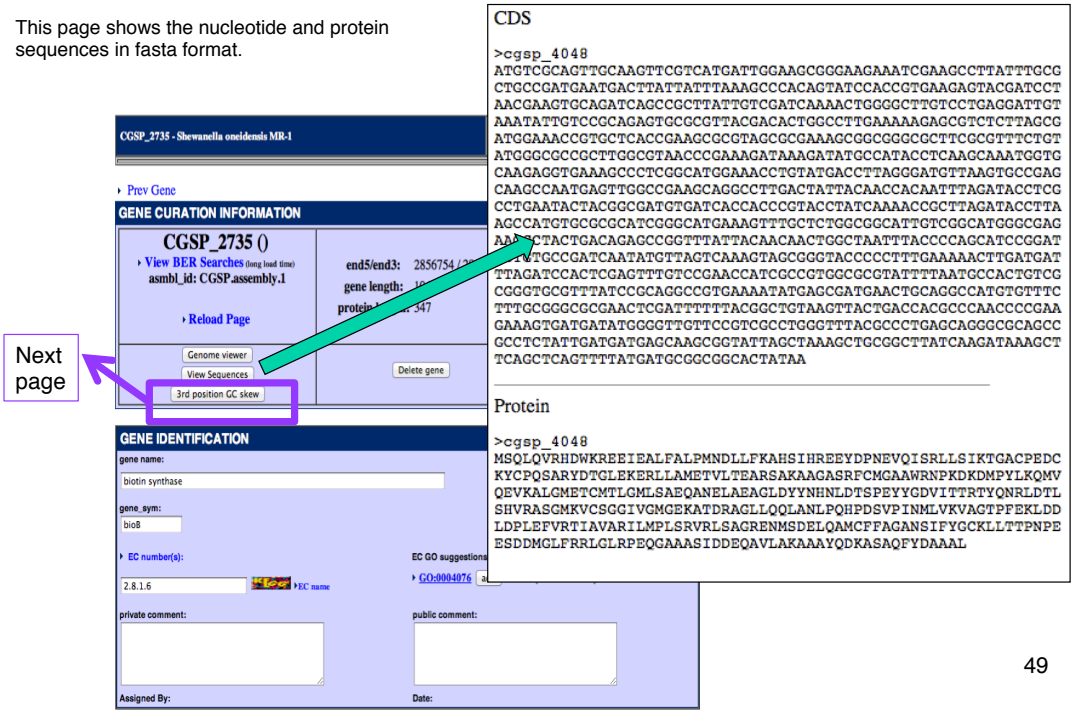
The arrows represent the predicted gene set. They are color-coded according to TIGR role id. The small arrows on the top represent the entire molecule along one scroll bar. The larger arrows depict a zoomed in view of one area of the genome. Mousing over the arrows brings up a box with info on the protein. Clicking on a small arrow will focus the zoom view onto that gene. Clicking on the info box in the zoomed view pops up a new box with links to other tools/pages. Underneath the zoomed view of the predicted genes is a graphical representation of a 6-frame translation of that region of DNA.

Genome Viewer - Gene Edit Page

Choosing to edit a gene brings up this view. Two boxes with coordinates for the predicted gene and for the ORF in which it resides are displayed. At the bottom is a text version of a six frame translation of the sequence in the area. Predicted genes are highlighted. Start sites are color-coded. Clicking on a "start" in the sequence will bring up a box asking you to confirm the change.

Buttons at the top of the Gene Curation Page - View sequence

This page shows the nucleotide and protein sequences in fasta format.



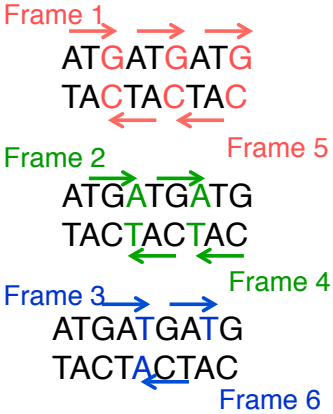
Next page

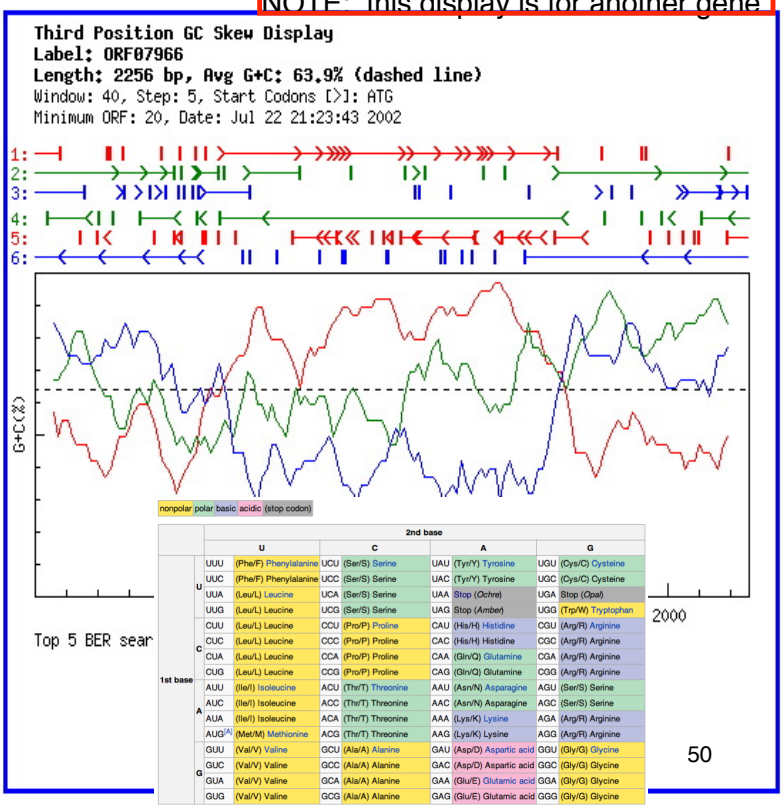
Links from the Gene Curation Page - Third position GC skew

NOTE: this display is for another gene

In organisms whose DNA has a high GC content it can sometimes be helpful to look at third position GC skew to help resolve overlaps.

Due to the nature of the genetic code, the third position is the least constrained of a codon and therefore will be able to reflect the higher GC content of the overall genome. Therefore one should see a markedly higher GC content in the third position of the correct frame.





Manual Annotation Checklist

- Look for HMM hits
  - evaluate what the HMMs are telling you - exact function? family membership? domain?
- Look at BER results
  - looking for proteins in the skim which are characterized (colored backgrounds)
  - many proteins are characterized but not marked so in our tables - may need to check proteins with white backgrounds to see if they are characterized
  - color coding does not indicate quality of match only that the match protein has been experimentally characterized
  - evaluate the alignment - what percent ID over what length? active sites? binding sites?
  - fill in characterized match accession number (by clicking on the accession in left column)
- Check Genome Viewer to view neighboring genes - annotate all genes in an operon together
- Look at TMHMM, SignalP, Prosite, region, etc.
- Decide what you think the protein should be named
- Fill in appropriate fields for common name, gene symbol, EC#, comment as needed.
- Decide what GO terms you need
  - find them on the Gene Curation Page (HMMs, EC number) or with the GO search tool AmiGO
  - change/remove any IEA GO annotations
  - add GO evidence from HMMs, BER, Prosite, etc.
- Review TIGR role and change as needed
- Check start site
  - Look at several BER matches, here you want to look at the best hits regardless of whether they are experimentally characterized
  - adjust if necessary – using Genome Viewer
  - check start site box when finished curation
- Check “complete”, click “submit” and your done!