A Guide to

Michelle Gwinn Giglio
2009

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 Mozilla or Firefox on your computer
  - other browsers work fine too, but Manatee is optimized for Firefox.
- To log into Manatee one must have an account and password.
- Each student will have their own account with the format "training#"
- 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 "cgsp"

Finding Manatee

On the internet:
go to http://manatee.igs.umaryland.edu/tigr-scripts/chado_prok_manatee/shared/login.cgi.

To download:
go to http://manatee.sourceforge.net

"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.

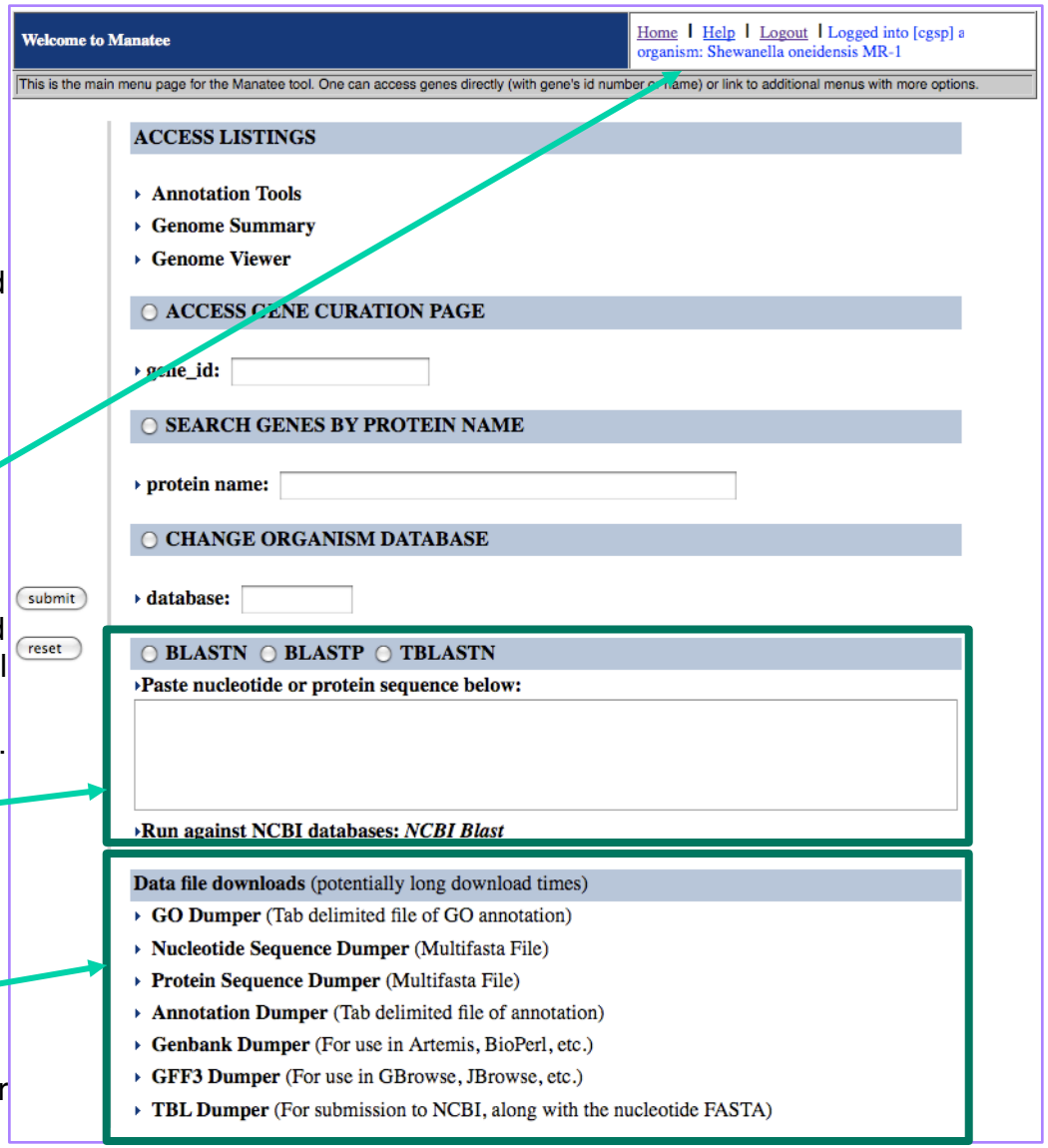
-The "Help" link should go to page specific documentation. However, these pages are still under development.

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:

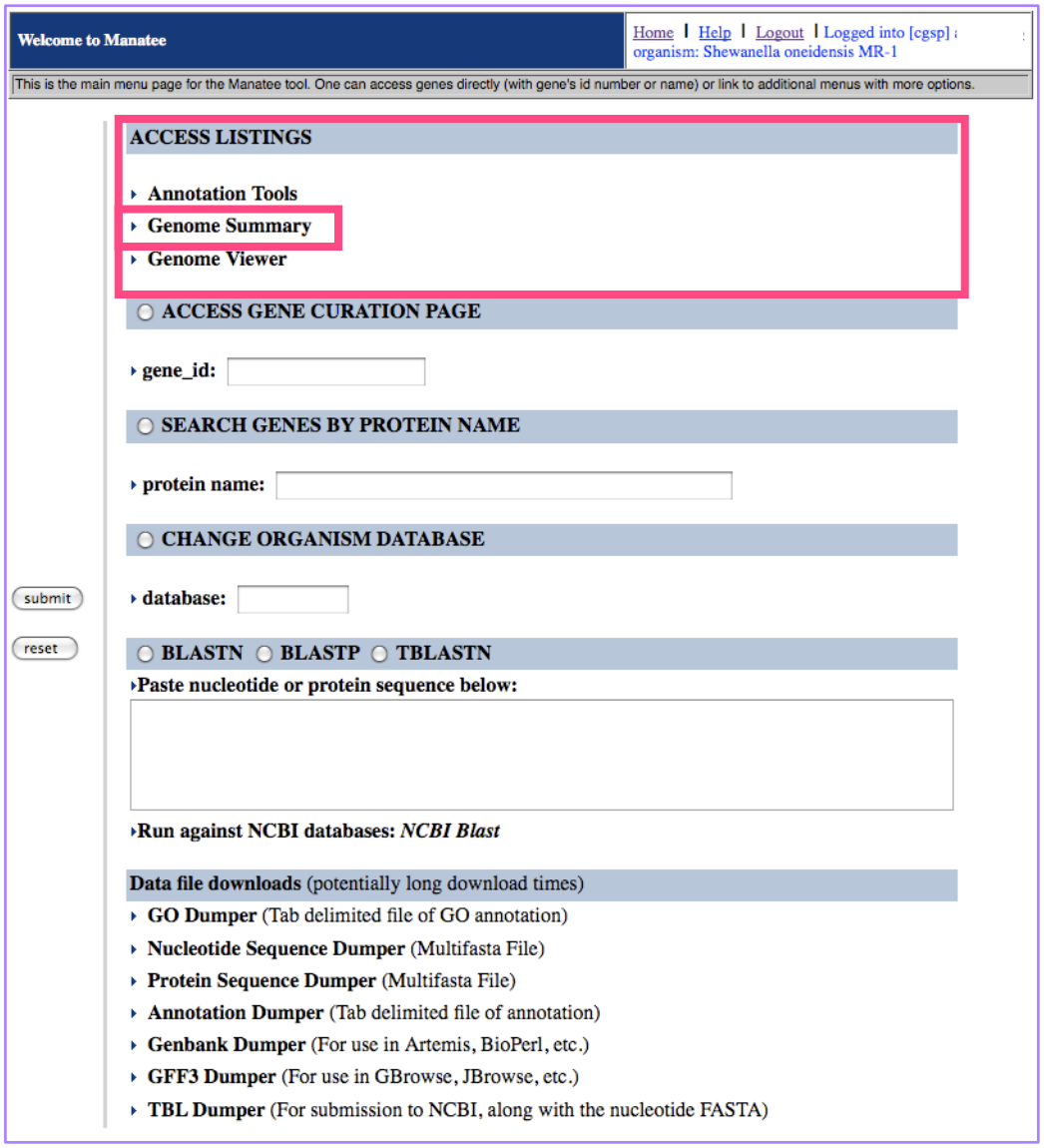
At the bottom of the Welcome page are several options for downloading text files containing annotation information. Some of these take a long time to query and load so please be patient. Simply click on the line of interest and the download process will begin.



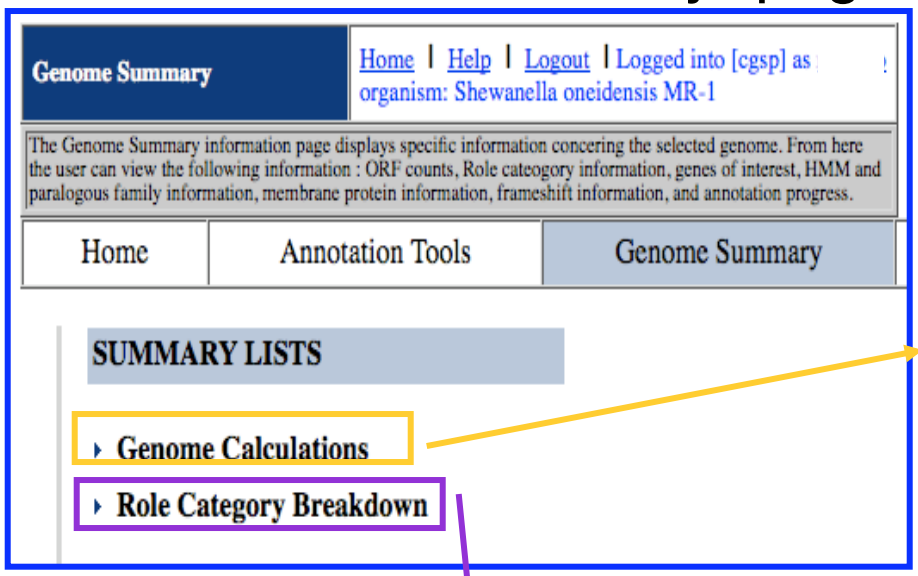
“Genome Summary

The Genome Summary section provides summary information about the annotation content of the entire genome.

Get there by clicking “Genome Summary” on the “Welcome to Manatee” page.



The “Genome Summary” page



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
	chromosome
	4969803 bp
	46%
	(A) (C) (G) (T) 27.0% 23.0% 23.0% 27.0%
	R Y 2 6
	4679
	904 nt
	85.2%
	85.2%

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 %

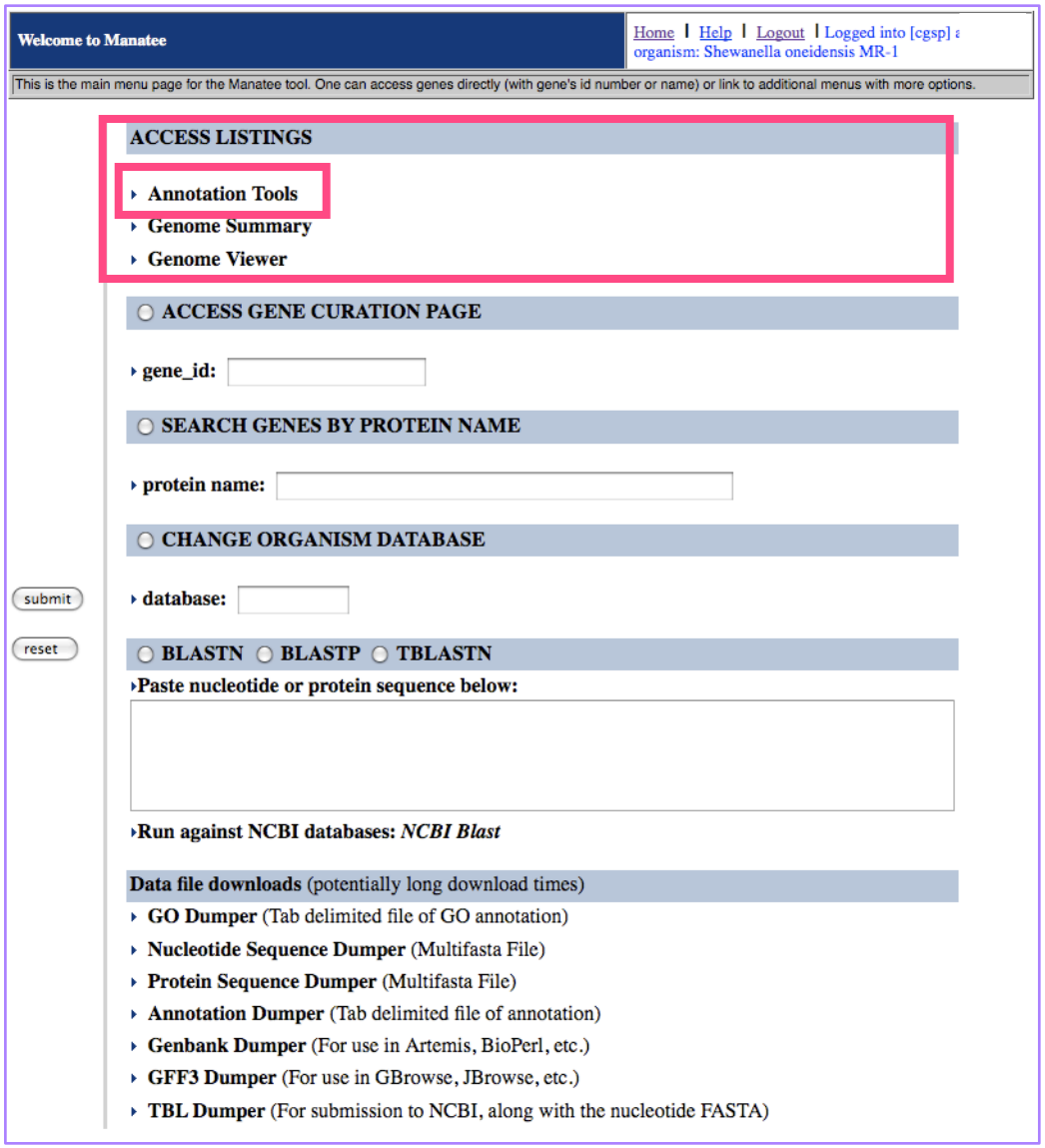
Role Breakdown

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%

“Annotation Tools”

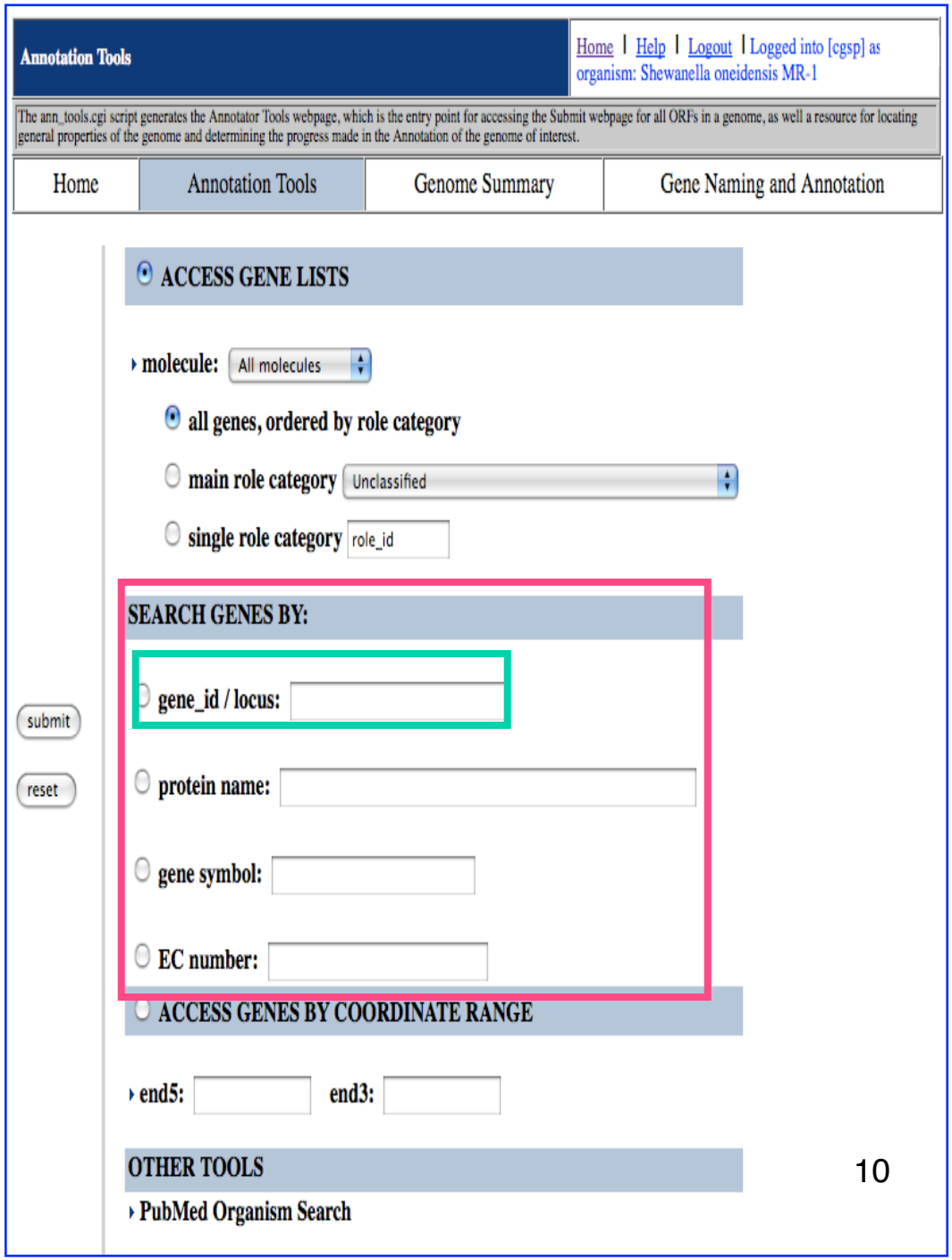
The Annotation Tools section contains most of the tools used during the process of manual annotation.

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

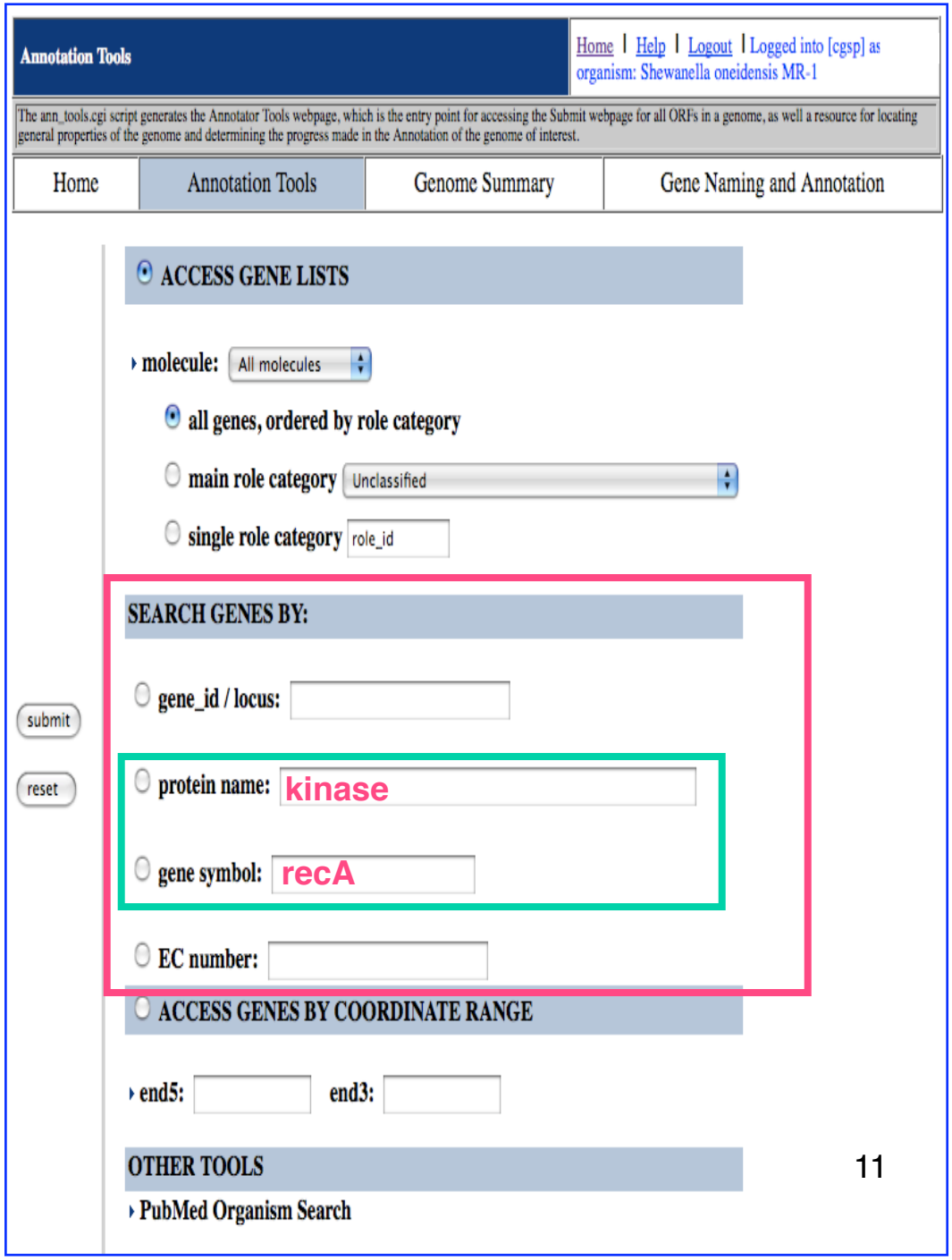
Whatever keyword you enter will be treated as though it has wildcards flanking it. This means that you will get results that include names with your keyword as an individual word and names with words that contain your keyword.

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

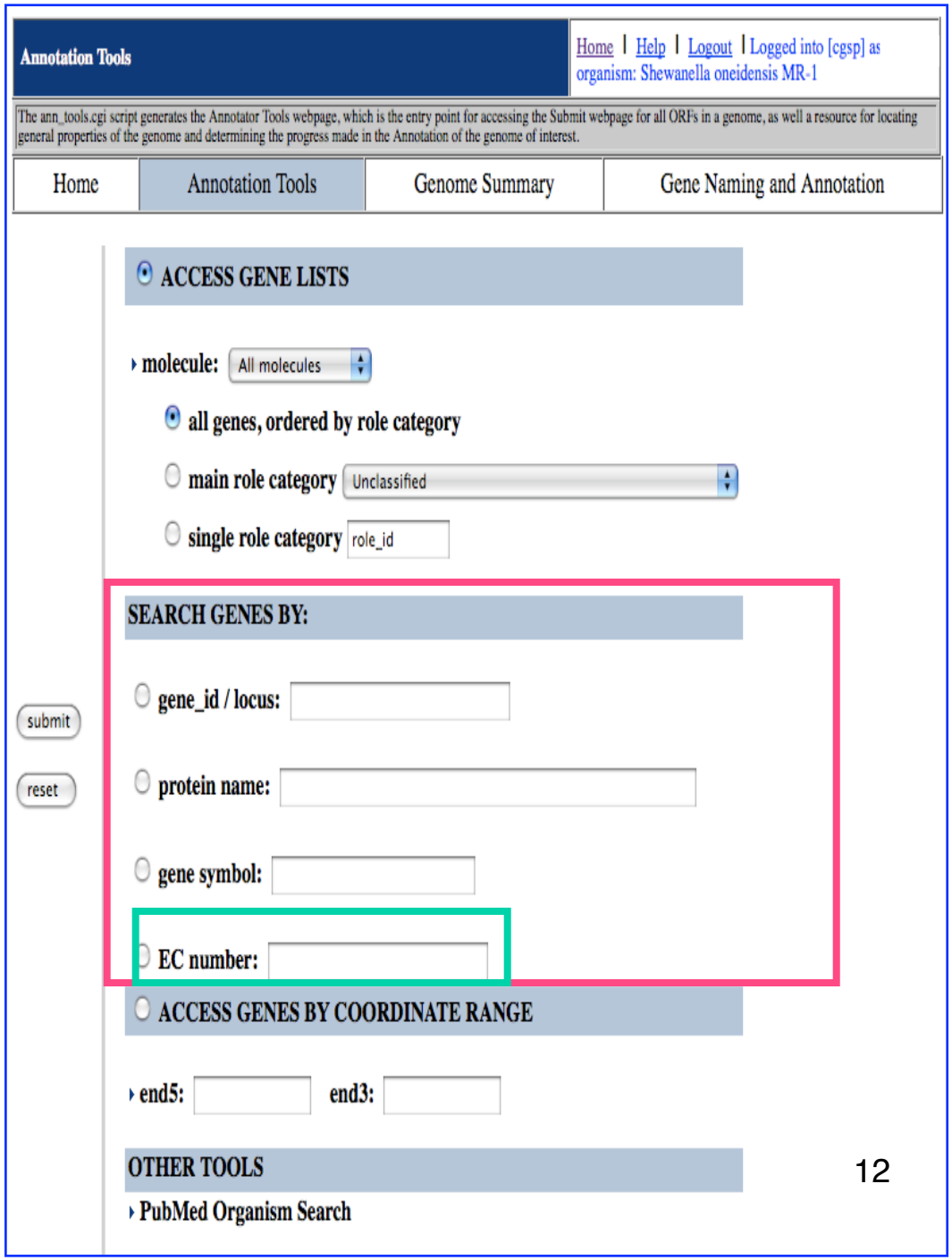
The Enzyme Commission maintains a database of enzymatic reactions which are each assigned an accession number of this format:

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

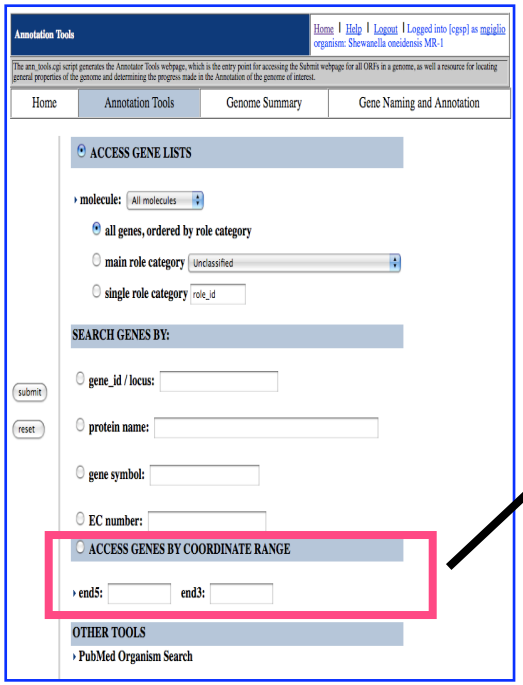
For the search, engter an EC number to see a list of all genes in the genome that have been annotated with that particular EC number.



Annotation Tools Page
“Access genes 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



List of all genes found between **10000** - **20000**

A	C	gene id	locus	end5	end3	role id	gene name	gene symbol	ec
		ORF02375	SO0017	22090	18941	156	conserved hypothetical protein		
		ORF02378	SO0016	18279	18854	132	DNA-3-methyladenine glycosidase I	tag	3.2.2.20
		ORF02379	SO0015	18161	17256	137	glycyl-tRNA synthetase, alpha subunit	glyQ	6.1.1.14
		ORF02381	SO0014	17246	15180	137	glycyl-tRNA synthetase, beta subunit	glyS	6.1.1.14
		ORF02382	SO0013	14311	15111		hypothetical protein		
		ORF02383	SO0012	13791	13129	96 102,	glutathione S-transferase family protein		
		ORF02385	SO0011	10638	13055	132	DNA gyrase, B subunit	gyrB	5.99.1.3
		ORF02386	SO0010	9539	10621	132	DNA replication and repair protein RecF	recF	
		ORFA00005	SOA0024	20332	19523	154	ISSo1, transposase OrfB		
		ORFA00006	SOA0023	19154	19453	94 186,	proteic killer suppressor protein	higA	
		ORFA00007	SOA0022	18774	19079	94 186,	proteic killer active protein	higB	
		ORFA00008	SOA0021	18235	18462	154 270,	ISSo1, transposase OrfB, truncation		
		ORFA00009	SOA0020	17414	18154	154 270,	transposase family protein, truncation		
		ORFA00011	SOA0019	16733	17290	132 154,	TnSon1, resolvase		
		ORFA00012	SOA0018	16362	16739	154 156,	TnSon1, conserved hypothetical protein		
		ORFA00013	SOA0017	16075	16365	703	TnSon1, nucleotidyltransferase domain protein		
		ORFA00014	SOA0016	15911	12945	154	TnSon1, transposase		
		ORFA00015	SOA0015	12878	12732		hypothetical protein		
		ORFA00016	SOA0014	12332	12427		hypothetical protein		
		ORFA00017	SOA0013	11739	11335	132	umuD protein	umuD	3.4.-.-
		ORFA00019	SOA0012	11334	10078	132	umuC protein	umuC	

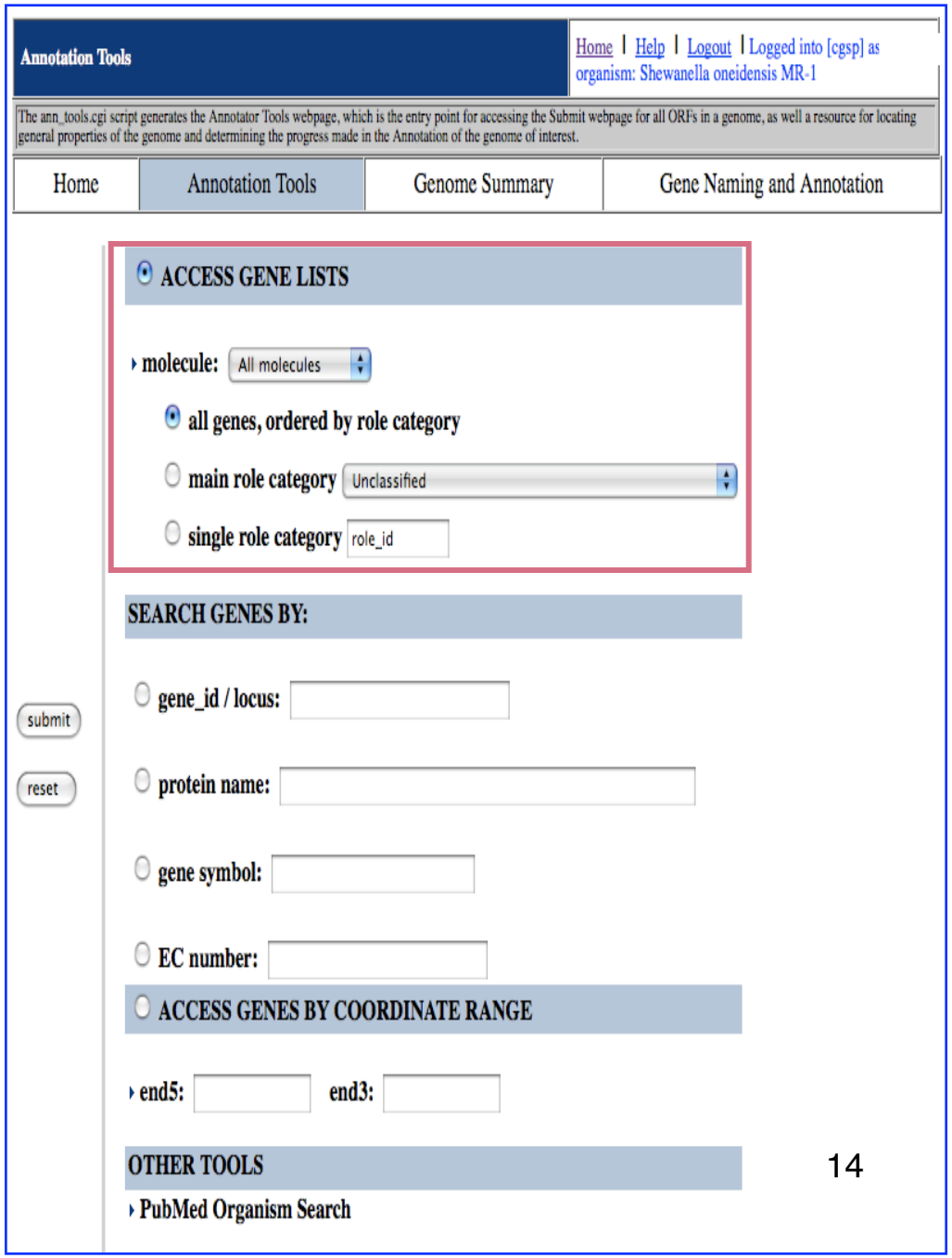
“Annotation Tools”: “Access Gene Lists” section

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
702 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

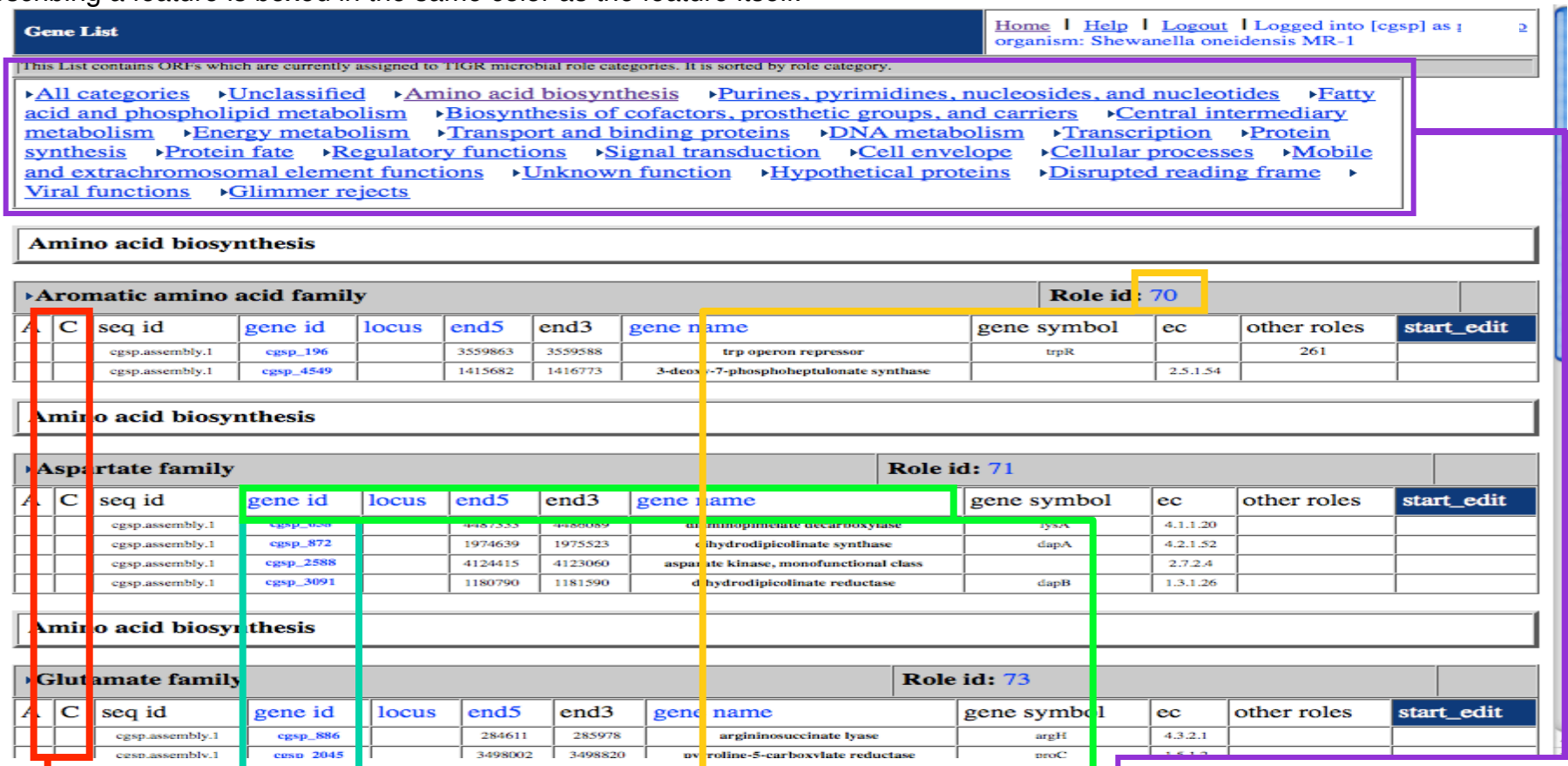
Mobile and extrachromosomal element functions
186 Plasmid functions
152 Prophage functions
154 Transposon functions
708 Other

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.



A pink dot will appear in the "C" column once an annotator has finished annotation for the gene and marked it "complete".

Link to role notes for this category

Clicking on the blue names of any mainrole category takes you to a gene list for that category.

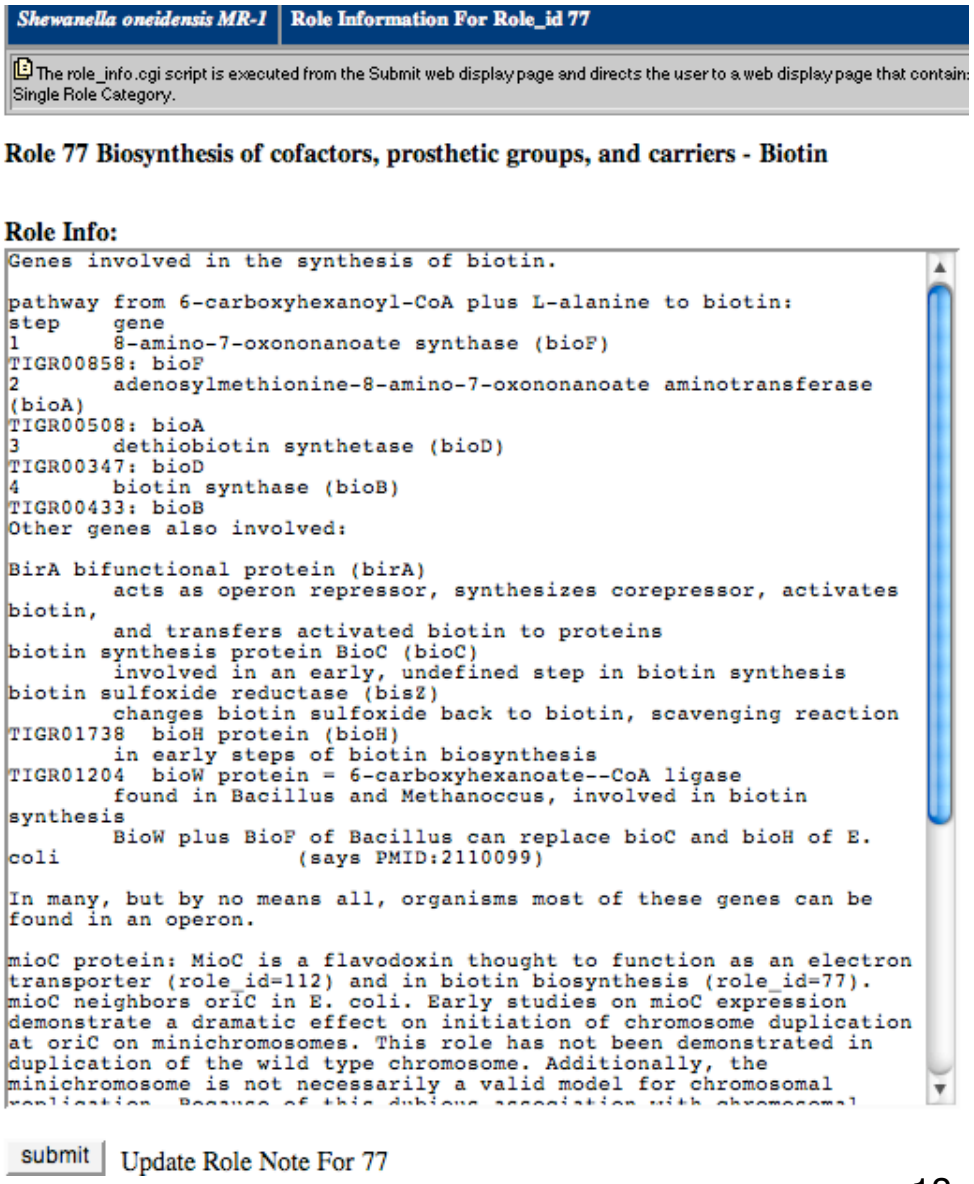
Click on the gene_id (feat_name) link to see the Gene Curation Page for each gene. Click on "GV" for Genome Viewer.

The ORFs can be ordered according to any of the blue headers by clicking on that header.

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 (GCP) is likely the most important page within Manatee, it is certainly the one that annotators spend the bulk of their time looking at and working with.

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 is a very complex page 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 AutoAnnotate. 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 characterized match section is where we enter the accession of a match gene whose function has been characterized in the lab (as opposed to having received its name based on sequence similarity.) This is stored as a piece of annotation evidence. This accession will pop into the go with_ev field in the proper format if you click on "Add to GO Evidence". (more on GO data later)

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 accession of the match protein (from various databases), the second is the percent similarity of the match, the third is the length of the match (in nucleotides), the fourth is the name of the match protein and finally, the P score from the BLAST search.

The color of the background for each entry in the skim indicates whether it is in the characterized table and at what confidence level: **green**=high confidence; **red**=automated process; **sky blue**=partial characterization; **olive**=trusted, used when multiple extremely good lines of evidence exist for function but no experiment has been done; **blue-green**=fragment/domain has been characterized; **fuzzy gray**=void, used to indicate that something that was originally thought to be characterized really is not; **gray**=omnium only

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).

BER SKIM submit

View BER Searches | search date: Wed Oct 23 12:59:20 2002

accession	%sim	length	description	p-value
OMNI:SO2740	100.0	349	biotin synthase {Shewanella oneidensis MR-1}	1.5e-176
SP:P36569	80.7	340	Biotin synthase (EC 2.8.1.6) (Biotin synthetase). {Serratia	2.5e-119
SP:P12996	79.7	342	Biotin synthase (EC 2.8.1.6) (Biotin synthetase). {Escherich	7.2e-120
GP:145425	79.7	342	biotin synthetase {Escherichia coli}	1.5e-119
GP:12620127	79.4	342	biotin synthase BioB {uncultured bacterium pCosHE2}	1.5e-119
OMNI:NTL03EC0855	79.4	342	biotin synthetase {Escherichia coli O157:H7 VT2-Sakai}□GPI13	5.1e-119
OMNI:NTL01YP1094	81.0	340	biotin synthase {Yersinia pestis CO92}□OMNINTL02YP2986 biot	8.3e-119
GP:12620099	79.5	340	BioB-like protein {uncultured bacterium pCosFS1}	9.5e-118
OMNI:NTL02EC0848	79.1	342	biotin synthesis, sulfur insertion? {Escherichia coli O157:H	2.2e-118
SP:Q47862	79.2	339	Biotin synthase (EC 2.8.1.6) (Biotin synthetase). {Erwinia h	3.6e-118
SP:P12678	78.6	344	Biotin synthase (EC 2.8.1.6) (Biotin synthetase). {Salmonell	5.1e-119
OMNI:VC1112	81.8	348	biotin synthase {Vibrio cholerae El Tor N16961}□GPI9655583|g	5.1e-119
OMNI:NTL03ST0726	78.6	344	biotin synthetase {Salmonella enterica serovar Typhi CT18}□G	1.1e-118
OMNI:NTL03PA00501	78.9	348	biotin synthase {Pseudomonas aeruginosa PAO1}□GPI9946364|gbl	7.7e-116
GP:12407614	76.8	339	biotin synthase BioB {uncultured bacterium pCosAS1}	9.1e-113
OMNI:NTL01XC0388	79.2	311	biotin synthase {Xanthomonas campestris pv. campestris ATCC3	2.8e-111
OMNI:NTL01XA0388	78.5	311	biotin synthase {Xanthomonas axonopodis pv. citri 306}□GPI21	6.6e-110
OMNI:NTL02BA0265	77.0	340	biotin synthase {Buchnera aphidicola Sg}□GPI21623185|gbl|AAM6	1.4e-109
OMNI:NTL01XF00065	79.4	309	biotin synthase {Xylella fastidiosa 9a5c}□GPI9104834|gbl|AAF8	8.4e-110
OMNI:NTL01RS0266	79.5	306	PROBABLE BIOTIN SYNTHASE PROTEIN {Ralstonia solanacearum GMI	4.7e-109
SP:P57378	77.3	342	Biotin synthase (EC 2.8.1.6) (Biotin synthetase). [Buchnera	1.1e-107
GP:15419053	79.1	328	biotin synthase {Acinetobacter calcoaceticus}	1.6e-106
OMNI:CC3521	76.2	339	biotin synthase {Caulobacter crescentus CB15}□GPI13425251|gb	3.0e-105
OMNI:NTL01BMA0776	79.8	311	BIOTIN SYNTHASE {Brucella melitensis 16M}□GPI17984969|gbl|AAL	6.3e-105

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 accessions and names for this protein from every database where it is found. These accessions are clickable objects and will take you to the page for the match protein in the database in question.

The background color of the header will be gold if the protein is believed to be experimentally characterized with the confidence level indicated by the color of the text for the relevant.
(This is seen for the SP accession in this alignment.)

Names in Skim are first entry in header, not necessarily the name you want to use when annotating your protein.

Links to info pages for the match protein in the source db.

BER Alignment detail: Boxed Header

66.0/79.7% over 343aa	*Escherichia coli*
• SP|P12996|BIOB_ECOLI Biotin synthase (EC 2.8.1.6) (Biotin synthetase). (exp=1; wgp=-1; cg=-1; closed=-1; pub=1; rf_status= ;)RF|NP_415296.1|16128743|NC_000913 biotin synthase {Escherichia coli K12;} (exp=0; wgp=1; cg=1; closed=1; pub=1; rf_status=provisional;)RF|AP_001406.1|89107626|AC_000091 biotin synthase {Escherichia coli W3110;} (exp=0; wgp=1; cg=1; closed=1; pub=1; rf_status=provisional;)RF|YP_309738.1|74311319|NC_007384 biotin synthesis	

-The background color of this box will be gold if the protein is in the characterized table and grey if it is not.

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

-The bottom section lists an accession numbers and names for instances of the match protein from the search databases. The accession numbers are links to pages for the match protein in the source databases.

-A particular entry in the list will have colored text (the color corresponding to its characterized status) if that is believed to have experimental evidence - this tells the annotators which link they should follow to find experimental characterization information. Only one accession for the match protein need be characterized for the header to turn gold.

BER Alignment detail: alignment header

```
ORF04813( 7 - 348 of 351 aa)
SP|P36569|BIOB_SERMA(5 - 345 of 346) Biotin synthase (EC 2.8.1.6) (Biotin synthetase).
%Identity = 67.5  %Similarity = 80.7
Gaps = 2  InDels = 9  Frame Shifts = 0
Primary Frame = 1 [340, 0, 0]
```

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

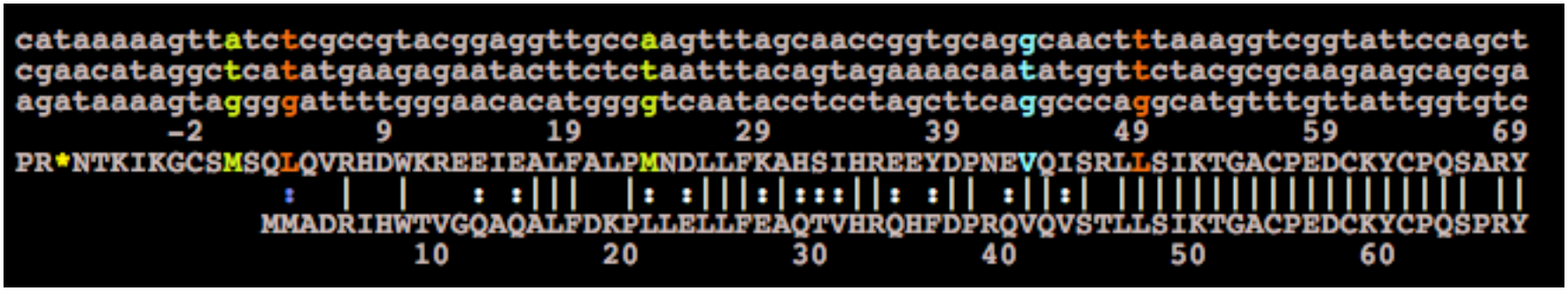
-The top line show the amino acid coordinates over which the match extends for our protein

-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.

Swiss-Prot entry - slide #1 - top of page

SwissProt is an incredibly useful database for manual annotation. All of the genes in SwissProt have been manually annotated by an experienced knowledgeable staff. In addition, along with each protein's annotation is stored additional information on references that describe the protein, cross referened databases in which the protein can be found, motifs which the protein contains, and coordinates of any known features in the protein (and much more.)

accession and version information

name, EC# gene_symbol taxonomy

references with links to abstracts (click on NCBI to see a PubMed abstract of the paper)

NiceProt View of Swiss-Prot: P12996

Printer-friendly view | Submit update | Quick BlastP search

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**BIOB_ECOLI**
Primary accession number	**P12996**
Secondary accession numbers	None
Entered in Swiss-Prot in	Release 13, January 1990
Sequence was last modified in	Release 35, November 1997
Annotations were last modified in	Release 44, July 2004
Name and origin of the protein	
Protein name	**Biotin synthase**
Synonyms	**EC 2.8.1.6** **Biotin synthetase**
Gene name	**Name: bioB** OrderedLocusNames: b0775
From	Escherichia coli [TaxID: 562]
Taxonomy	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.

Link to Enzyme Commission page (see later slide)

References

[1] SEQUENCE FROM NUCLEIC ACID.
MEDLINE=89066784;PubMed=3058702 [NCBI, ExPASy, EBI, Israel, Japan]
Otsuka A.J., Buoncristiani M.R., Howard P.K., Flamm J., Johnson O.;
"The Escherichia coli biotin biosynthetic enzyme sequences predicted from the nucleotide sequence of the bio operon.";
J. Biol. Chem. 263:19577-19585(1988).

[2] SEQUENCE FROM NUCLEIC ACID.
Pearson B.M., McKee R.A.;
"Genetic material for expression of biotin synthetase enzymes.";
Patent number GB2216530, 11-OCT-1989.

[3] SEQUENCE FROM NUCLEIC ACID.
STRAIN=K12 / MG1655;
MEDLINE=97426617;PubMed=9278503 [NCBI, ExPASy, EBI, Israel, Japan]
Blattner F.R., Plunkett G. III, Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B., Shao Y.;
"The complete genome sequence of Escherichia coli K-12.";
Science 277:1453-1474(1997).

[4] CHARACTERIZATION.
PubMed=8142361 [NCBI, ExPASy, EBI, Israel, Japan]
Sanyal I., Cohen G., Flint D.H.;
"Biotin synthase: purification, characterization as a [2Fe-2S] cluster protein, and in vitro activity of the Escherichia coli bioB gene product.";
Biochemistry 33:3625-3631(1994).

[5] MUTAGENESIS OF CYSTEINE RESIDUES.

Swiss-Prot entry - slide #2 - middle of page

useful functional information

links to other dbs where the protein is found or to motif clusters or protein families which this protein is a member of

Comments

- ***CATALYTIC ACTIVITY***: Dethiobiotin + sulfur = biotin.
- ***COFACTOR***: Binds a 4Fe-4S cluster coordinated with 3 cysteines and an exchangeable S-adenosyl-L-methionine, and a 2Fe-2S cluster coordinated with 3 cysteines and 1 arginine.
- ***PATHWAY***: Biotin biosynthesis; last step.
- ***SUBUNIT***: Homodimer.
- ***SIMILARITY***: Belongs to the biotin and lipoic acid synthetases family.

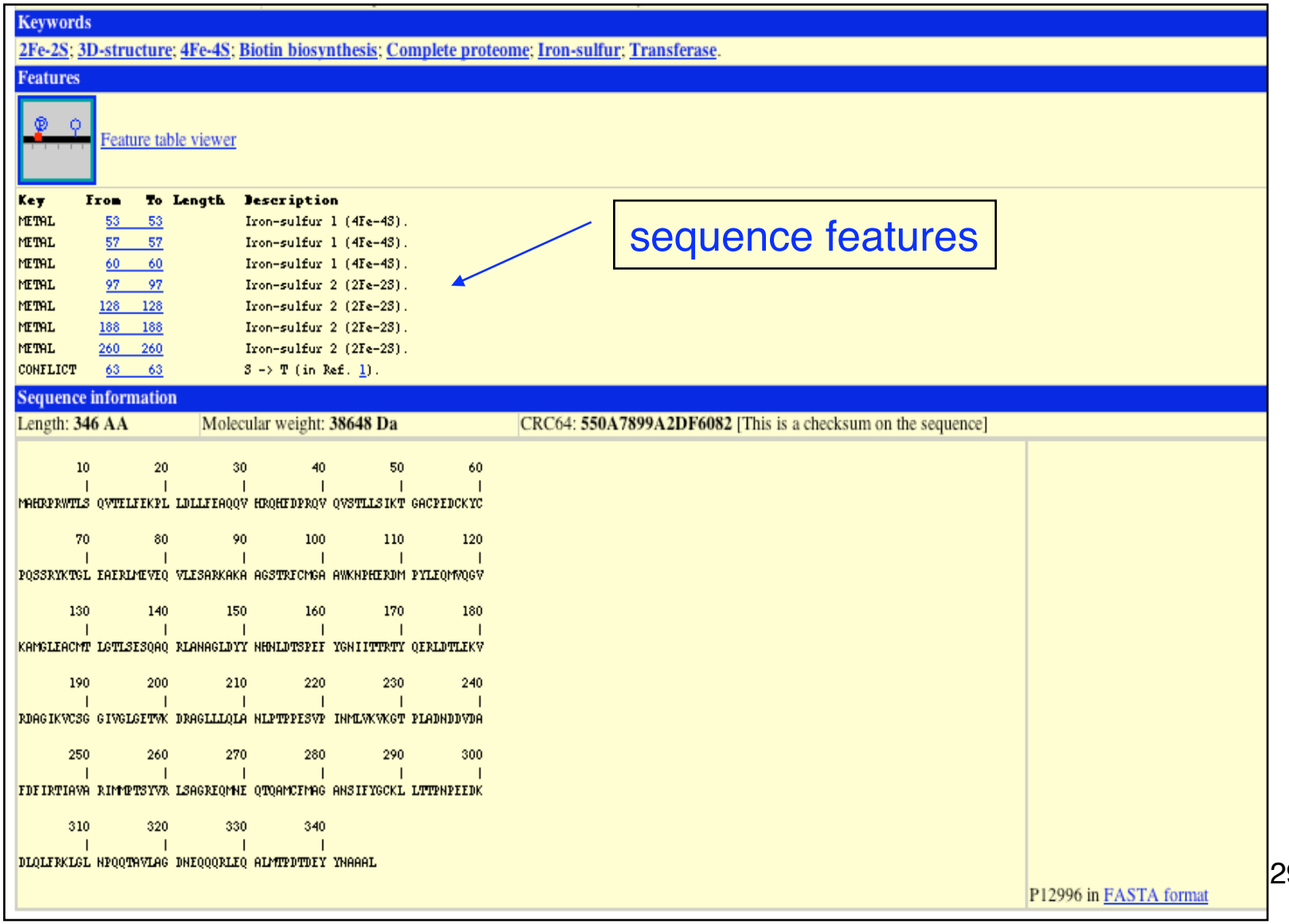
Copyright

Cross-references

EMBL	J04423; AAA23515.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] A11530; CAA00965.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AE000180; AAC73862.1; -.[EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	JC2517; SYECBB.
PDB	1R30; 13-JAN-04.[ExPASy / RCSB / EBI]
ECO2DBASE	E038.6; 6TH EDITION.
EchoBASE	EB0116; -.
EcoGene	EG10118; bioB.
EcoCyc	EG10118; bioB.
CMR	P12996; b0775.
InterPro	IPR010722; BATS. IPR002684; Biotin_synth. IPR006638; Elp3/MiaB/NifB. IPR007197; Radical_SAM. Graphical view of domain structure.
Pfam	PF06968; BATS; 1. PF04055; Radical_SAM; 1. Pfam graphical view of domain structure.
SMART	SM00729; Elp3; 1.
TIGRFAMs	TIGR00433; bioB; 1.
ProDom	[Domain structure / List of seq. sharing at least 1 domain]
HOBACGEN	[Family / Alignment / Tree]
BLOCKS	P12996.
ProtoNet	P12996.
ProtoMap	P12996.
PRESAGE	P12996.
DIP	P12996.
ModBase	P12996.
SMR	P12996; 550A7899A2DF6082.
SWISS-2DPAGE	Get region on 2D PAGE.
UniRef	View cluster of proteins with at least 50% / 90% identity

Swiss-Prot entry - slide #3 - bottom of page

keywords and sequence features with coordinates



Keywords

2Fe-2S; 3D-structure; 4Fe-4S; Biotin biosynthesis; Complete proteome; Iron-sulfur; Transferase.

Features

Feature table viewer

Key	From	To	Length	Description
METAL	53	53		Iron-sulfur 1 (4Fe-4S).
METAL	57	57		Iron-sulfur 1 (4Fe-4S).
METAL	60	60		Iron-sulfur 1 (4Fe-4S).
METAL	97	97		Iron-sulfur 2 (2Fe-2S).
METAL	128	128		Iron-sulfur 2 (2Fe-2S).
METAL	188	188		Iron-sulfur 2 (2Fe-2S).
METAL	260	260		Iron-sulfur 2 (2Fe-2S).
CONFLICT	63	63		S -> T (in Ref. 1).

sequence features

Sequence information

Length: **346 AA** | Molecular weight: **38648 Da** | CRC64: **550A7899A2DF6082** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
         |          |          |          |          |          |
MAERPRWTLS QVTELFEKPL LDLLFEAQQV HRQHFDPRQV QVSTLLSIKT GACPEDCKYC

        70         80         90        100        110        120
         |          |          |          |          |          |
PQSSRYKTGL EAERLMEVEQ VLESARKAKA AGSTRFCMGA AWKNPHERDM PYLEQMVQGV

       130        140        150        160        170        180
         |          |          |          |          |          |
KAMGLEACMT LGTLSESQAQ RLANAGLDYY NHNLDTSPEF YGNIITTRTY QERLDTLEKV

       190        200        210        220        230        240
         |          |          |          |          |          |
RDAGIKVCSG GIVGLGETVK DRAGLLLQLA NLPTPPESVP INMLVKVKGT PLADNDDVDA

       250        260        270        280        290        300
         |          |          |          |          |          |
FDFIRTIAVA RIMMPTSYVR LSAGREQMNE QTQAMCFMAG ANSIFYGCKL LTTPNPEEDK

       310        320        330        340
         |          |          |          |
DLQLFRKLGL NPQQTAVLAG DNEQQQRLEQ ALMTPDTDEY YNAAAL
```

P12996 in FASTA format

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

NiceZyme View of ENZYME: EC 2.8.1.6

Official Name

Biotin synthase.

Alternative Name(s)

Biotin synthetase.

Reaction catalysed

```
    Dethiobiotin
  +  sulfur
<=>
    biotin
```

Cofactor(s)

Iron-sulfur.

Comments

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

Cross-references

BRENDA	2.8.1.6
EMP/PUMA	2.8.1.6
WIT	2.8.1.6
Kyoto University LIGAND chemical database	2.8.1.6
IUBMB Enzyme Nomenclature	2.8.1.6
IntEnz	2.8.1.6
MEDLINE	Find literature relating to 2.8.1.6
Swiss-Prot	P54967, BIOB_ARATH; P19206, BIOB_BACSH; P53557, BIOB_BACSU; P57378, BIOB_BUCAI; Q8K9P1, BIOB_BUCAP; Q89AK5, BIOB_BUCBP; P12997, BIOB_CITFR; P46396, BIOB_CORGL; P12996, BIOB_ECOLI; Q47862, BIOB_ERWHE; P44987, BIOB_HAEIN; Q9ZJK8, BIOB_HELPJ; O25956, BIOB_HELPY; Q58692, BIOB_METJA; P94966, BIOB_METSK; P46715, BIOB_MYCLE; O06601, BIOB_MYCTU; P12678, BIOB_SALTY; O59778, BIOB_SCHPO; P36569, BIOB_SERMA; P73538, BIOB_SYNY3; P32451, BIOB_YEAST;

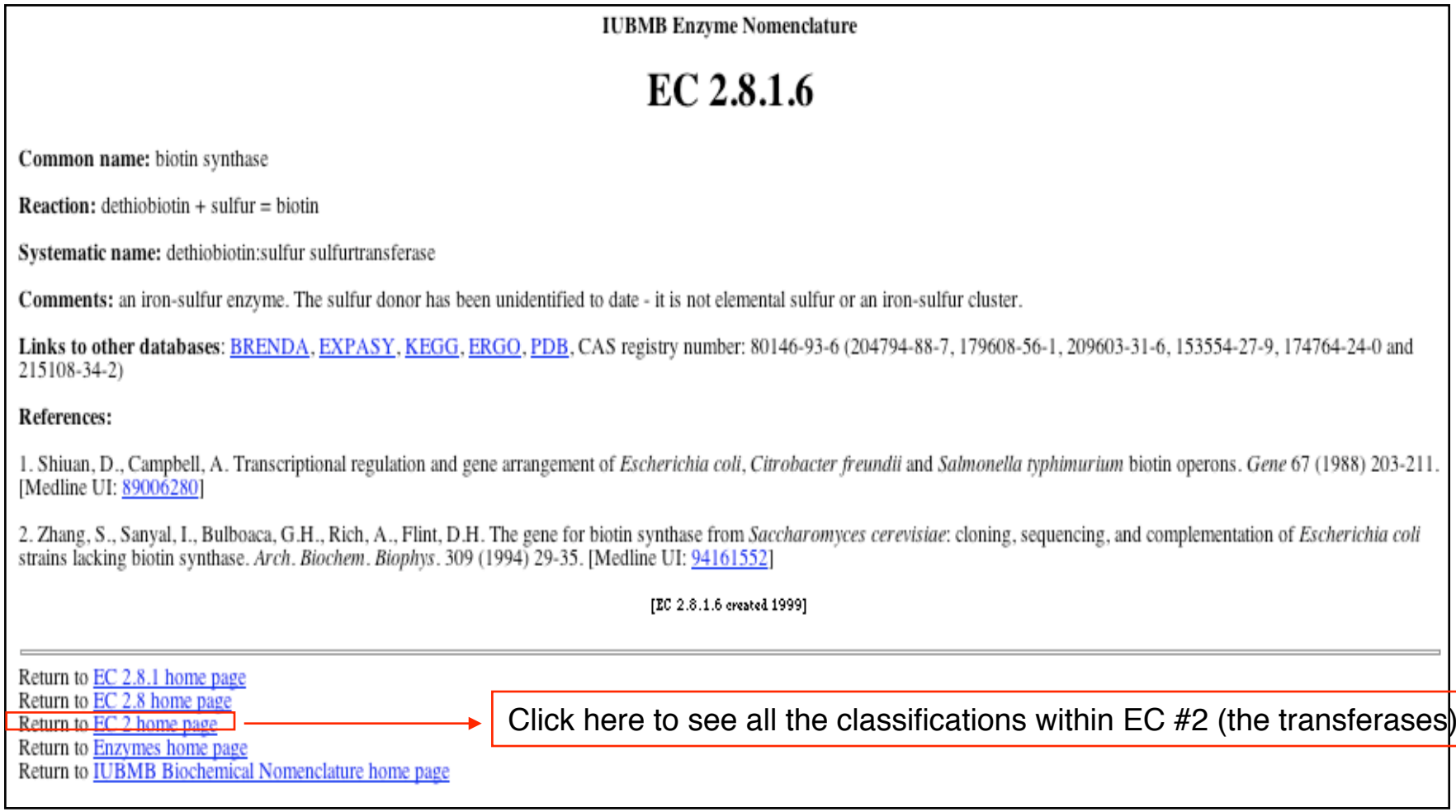
Link to official Enzyme Commission site

View entry in original ENZYME format

All Swiss-Prot entries referenced in this entry, with possibility to download in different formats, align etc.

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.)

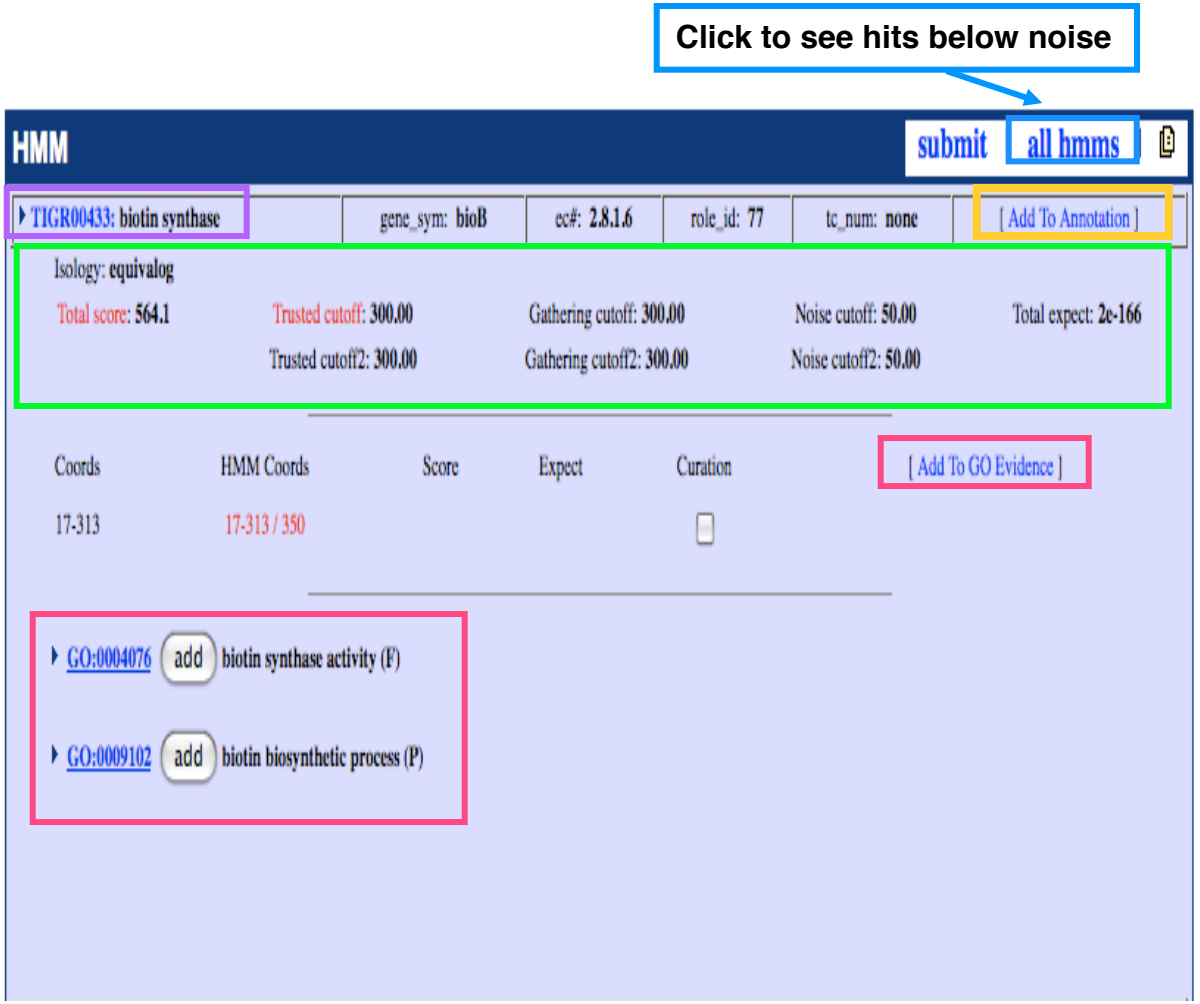
The blue id numbers for each HMM link to an info page for that HMM.

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

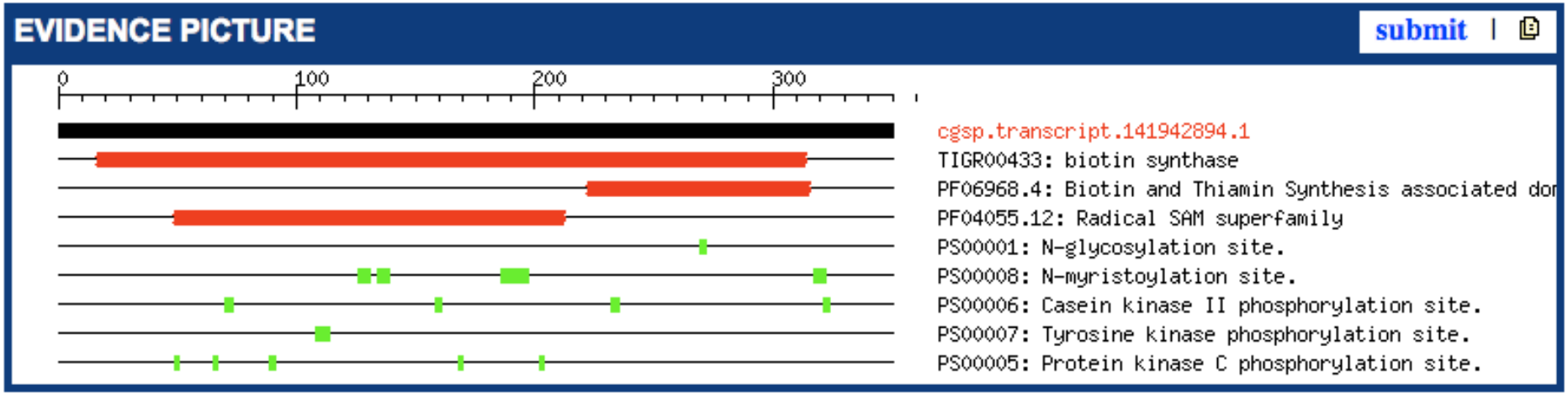
color key

- 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 unless an annotator has decided that the HMM should be included as evidence by toggling the curation box. The pink bar represents the characterized match to this ORF. Characterized matches are shown in different colors that at this time have no meaning. Also shown here is a secondary structure prediction (not run on all genomes). Clicking on the colored bars in the graphic opens windows with additional information on that piece of evidence. To get additional cog info, you must click on the very skinny bar all the way to the left of the cog row. 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.

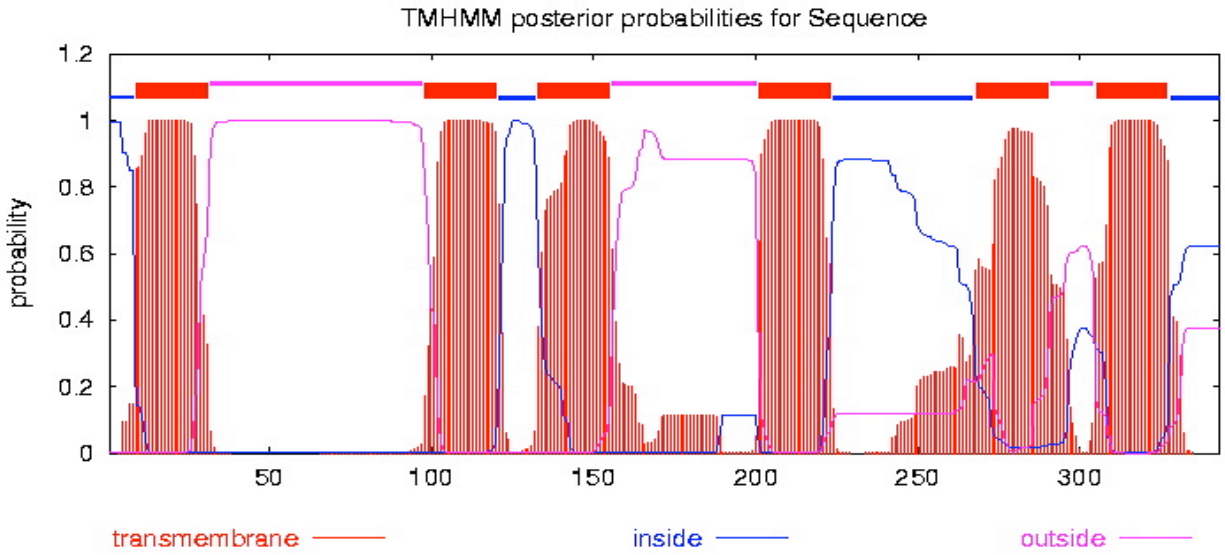


NOTE: this display is for ORF03779

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
```

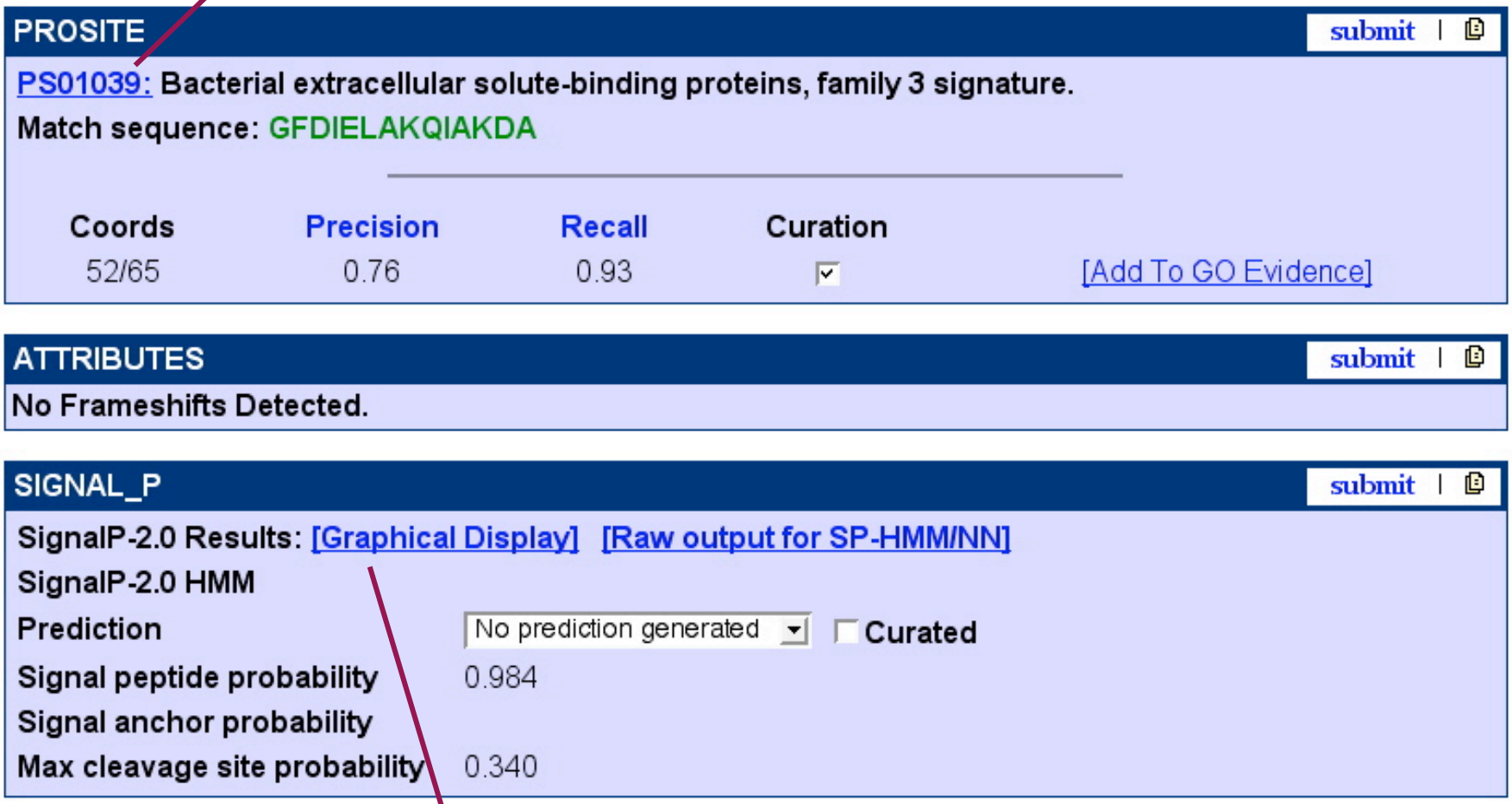


plot in postscript, script for making the plot in gnuplot, data for plot

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.



Click here to see output in graphical form.

ExPASy Home page | Site Map | Search ExPASy | Contact us | PROSITE

Hosted by NCSC US | Mirror sites: Bolivia | Canada | China | Switzerland | Taiwan

The Korean ExPASy site, **kr.expasy.org**, is temporarily not available.

Search PROSITE for Go Clear

PROSITE page at ExPASy

NOTE: this display is for ORF01166

NiceSite View of PROSITE: PDOC00798 (documentation)

Bacterial extracellu family 3 signature

PROSITE cross-reference(s)

PS01039; SBP_BACTERIAL_3

Documentation

Bacterial high affinity transport solutes across the cytoplasmic traffic systems include one o two membrane-associated ATP-b <PDOC00185>) and a high affinity are thought to bind the substra to transfer it to a complex of the cytoplasm.

In gram-positive bacteria which are surrounded by a single membrane and have therefore no periplasmic region the equivalent proteins are bound to the membrane via an N-terminal lipid anchor. These homolog proteins do not play an integral role in the transport process per se, but probably serve as receptors to trigger or initiate translocation of the solute throught the membrane by binding to external sites of the integral membrane proteins of the efflux system.

In addition at least some solute-binding proteins function in the initiation of sensory transduction pathways.

On the basis of sequence similarities, the vast majority of these solute-binding proteins can be g
generally correlate with the

Family 3 groups together
proteins and a periplasmic h

- Histidine-binding protei bacteria. An homologous l
- Lysine/arginine/ornithine coli and related bacter hisJ. Both solute-bindin receptor hisP of the bind
- Glutamine-binding protei stearothermophilus.
- Glutamate-binding protein
- Arginine-binding proteins
- Nopaline-binding protein
- Octopine-binding protein
- Major cell-binding factor
- Bacteroides nodosus prote
- Cyclohexadienyl/arogenate periplasmic enzyme which biosynthesis.
- Escherichia coli protein
- Vibrio harveyi protein pa
- Escherichia coli hypothet
- Bacillus subtilis hypothe
- Bacillus subtilis hypothe

The signature pattern is loc

Description of pattern(s) and/or profile(s)

Consensus pattern	G-[FYIL]-[DE]-[LIVMT]-[DE]-[LIVMF]-x(3)-[LIVMA]-[VAGC]-x(2)-[LIVMAGN]
Sequences known to belong to this class detected by the pattern	ALL.
Other sequence(s) detected in Swiss-Prot	23.

Last update

November 1997 / Pattern and text revised.

References

[1]
Tam R., Saier M.H. Jr.
Microbiol. Rev. 57:320-346(1993).

Copyright

View entry in original PROSITE document format
View entry in raw text format (no links)

ExPASy Home page | Site Map | Search ExPASy | Contact us | PROSITE

Hosted by NCSC US | Mirror sites: Bolivia | Canada | China | Switzerland | Taiwan

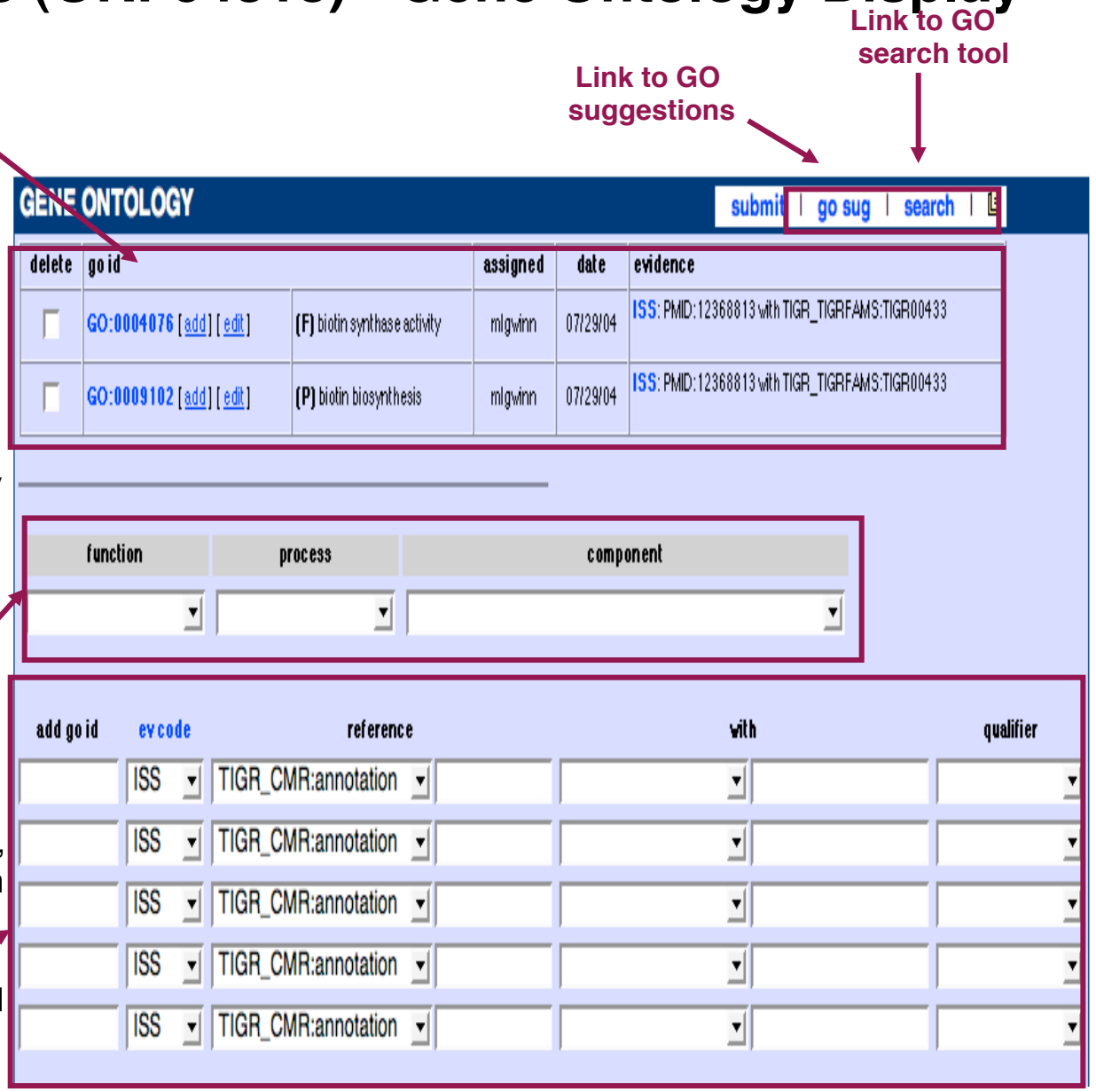
The Korean ExPASy site, **kr.expasy.org**, is temporarily not available.

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)

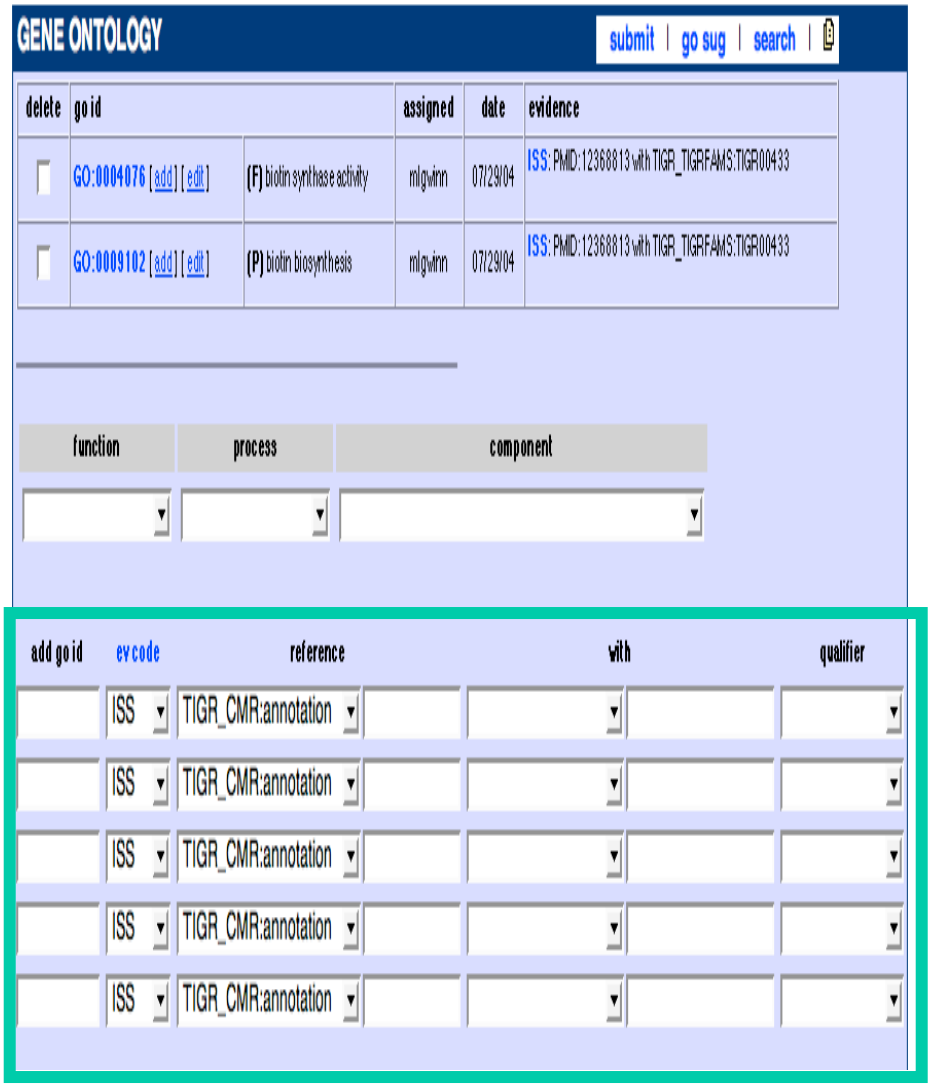
GO id - the format is GO:#######.

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.

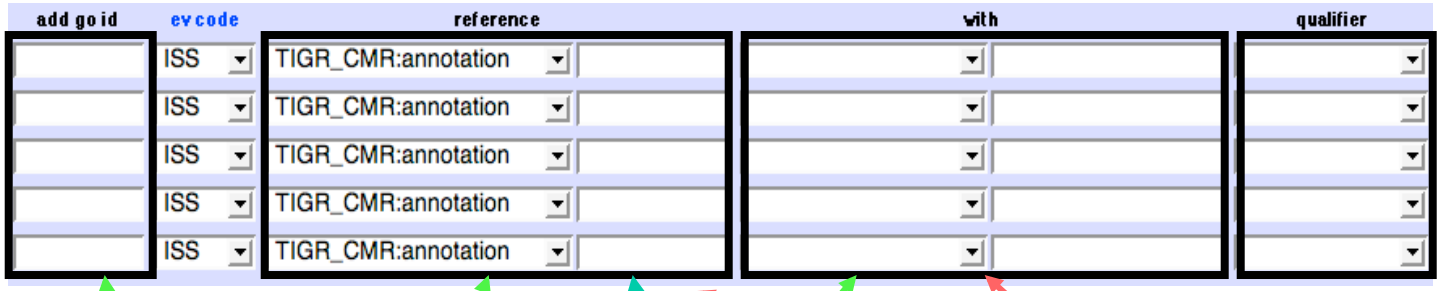
-”Add to GO evidence” buttons are also available for Prosite hits, this populates the “with” field with the Prosite accession. Available when a protein has matches to Prosite.

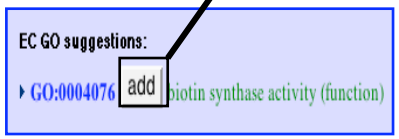
-”Add to GO evidence” is also available for the characterized match accession, this will put the accession of the characterized matching protein into the “with” field entry box.

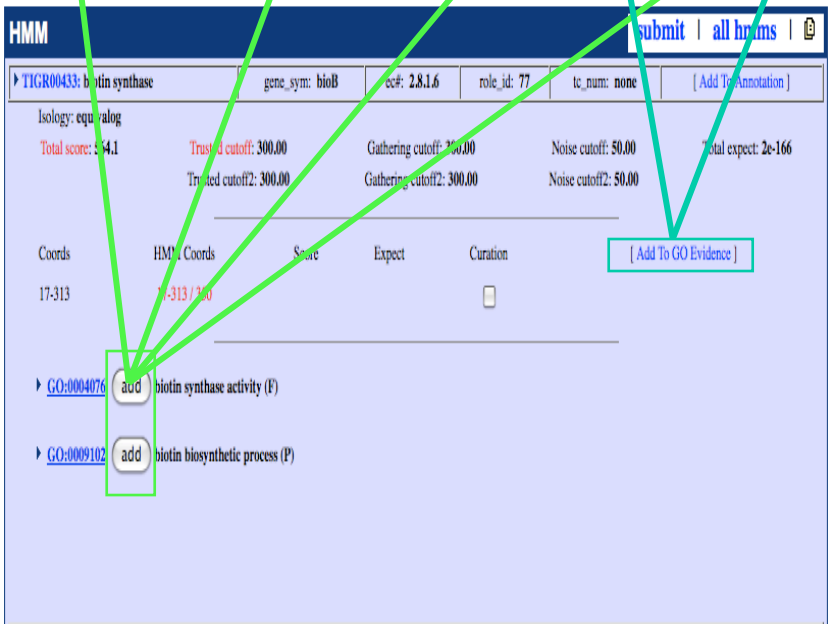
See next page for screen shots.

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

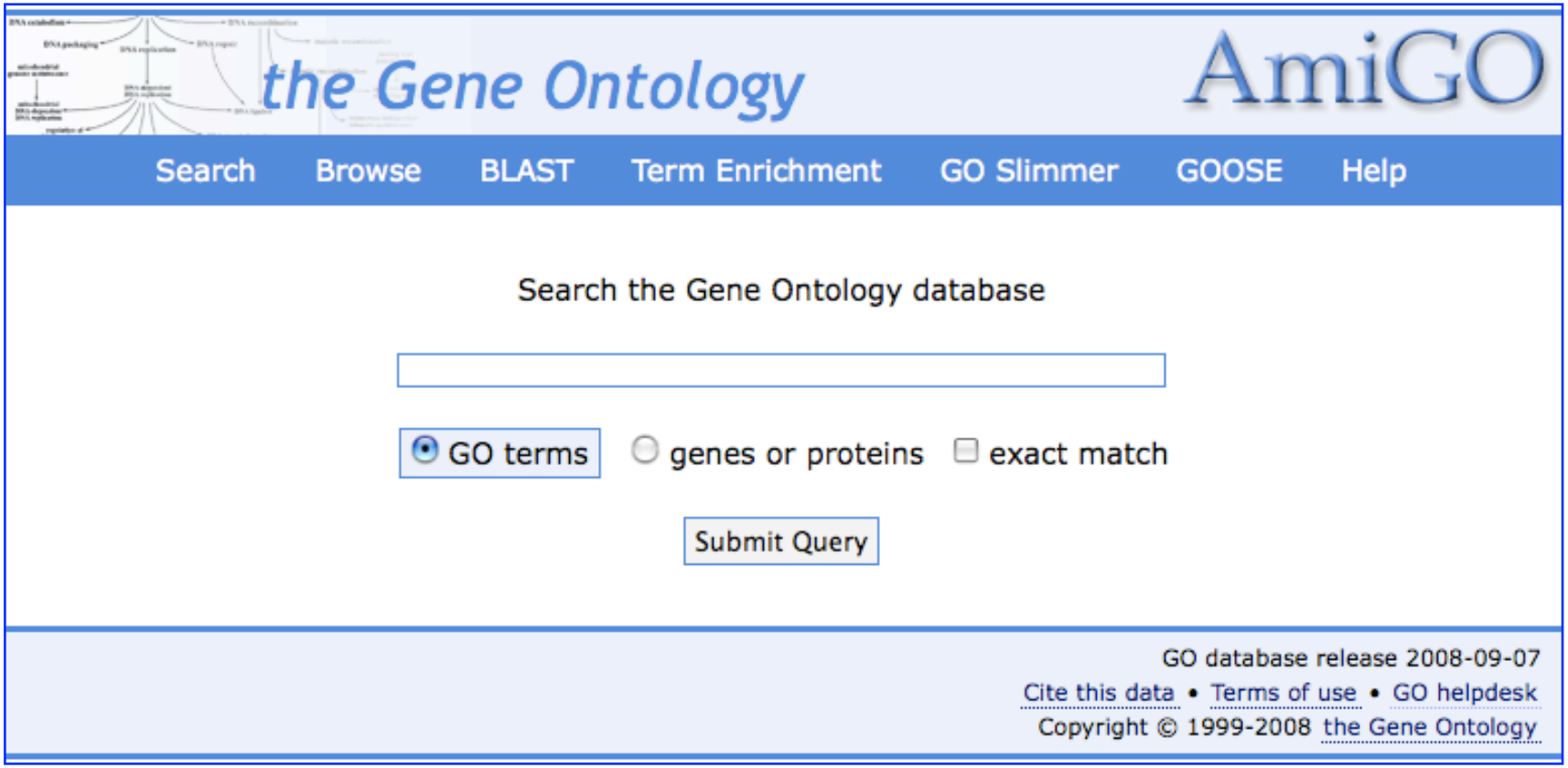






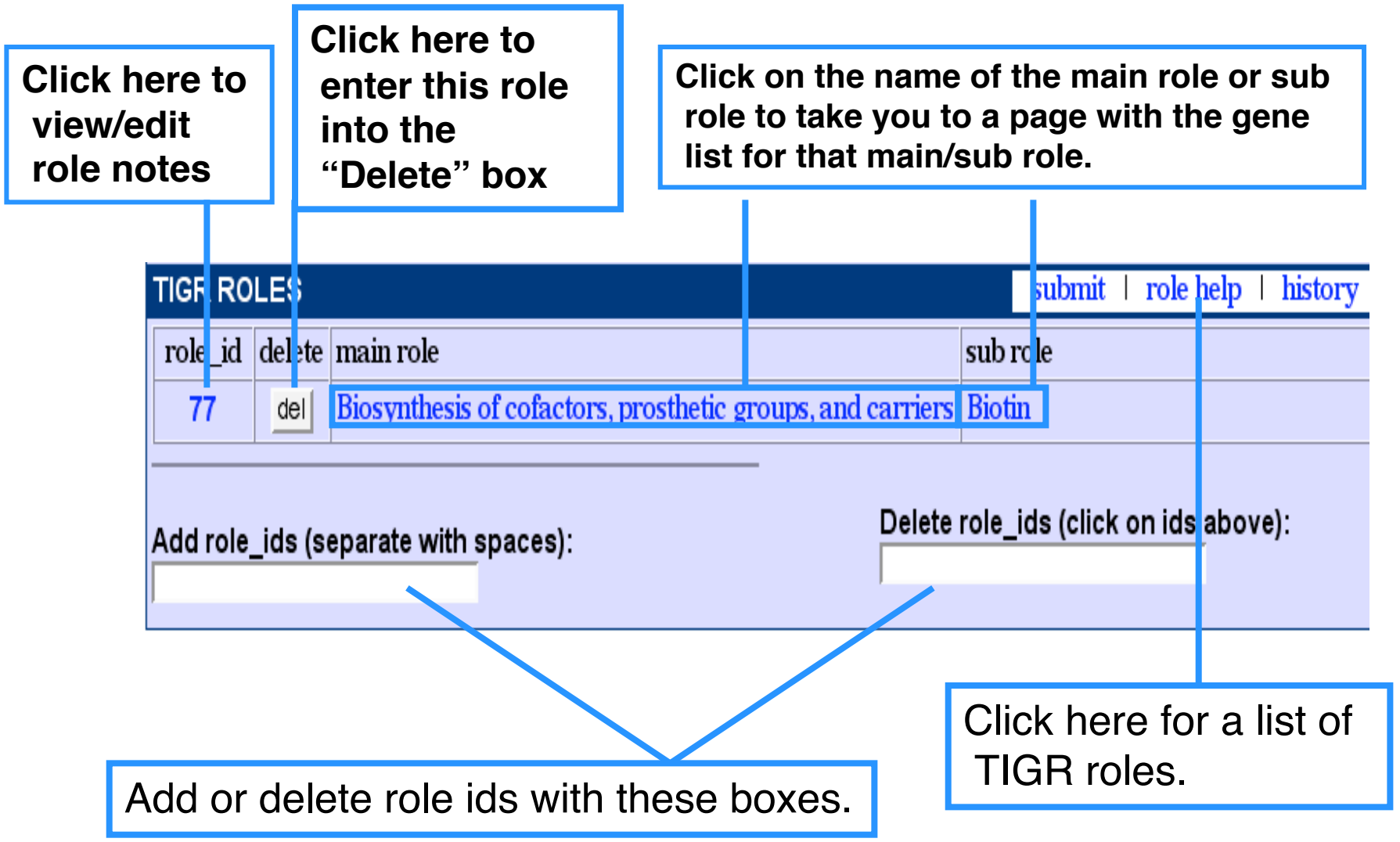
Searching for GO terms: the AmiGO search tool:

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

Click this button when you have completed annotation for this gene. With this toggle we know that this gene is finished.

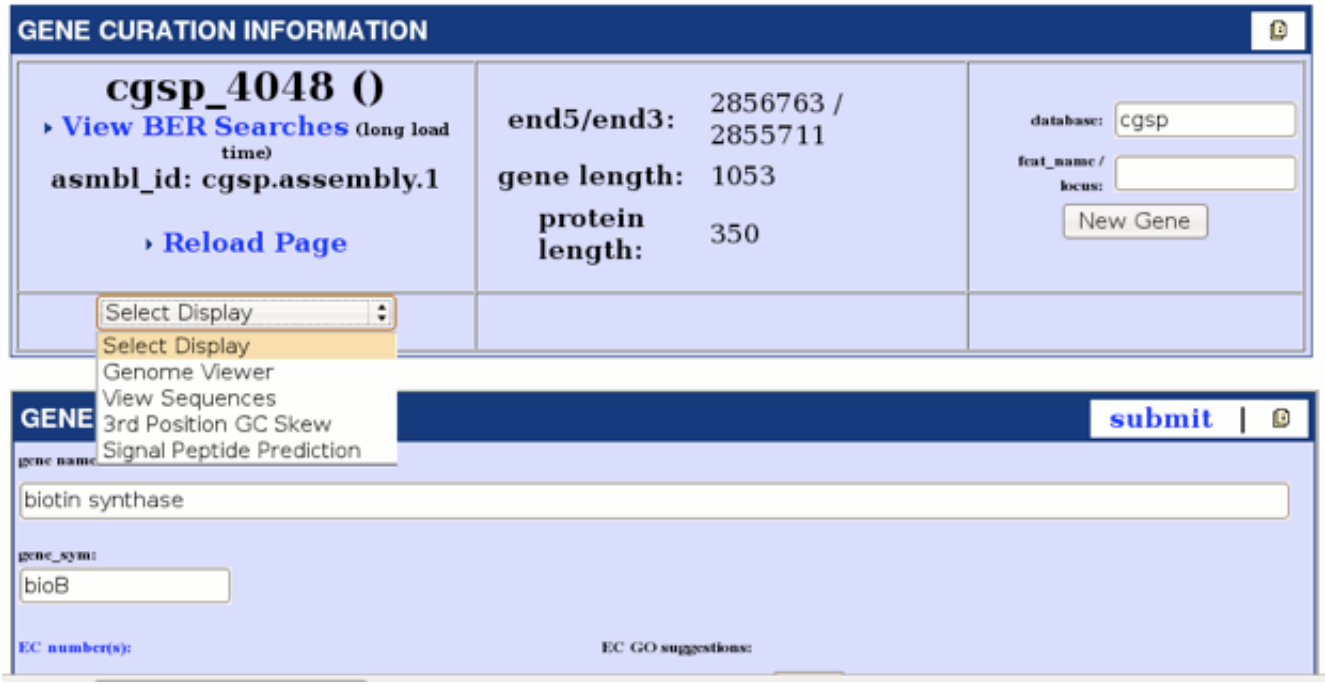
Clicking this button indicates that you have reviewed the start site and either found it to be fine or edited it to the correct (or at least what we hope is correct) position.

Click here to submit your entries to the database. You can also do this by clicking on any of the “submit” buttons in the upper right of each section on the page. Clicking “submit” anywhere on the page submits data from all fields (not just the section from which you clicked the button.)

This button resets the page to the state it was when originally opened.

Gene Curation Page - The pull down menu

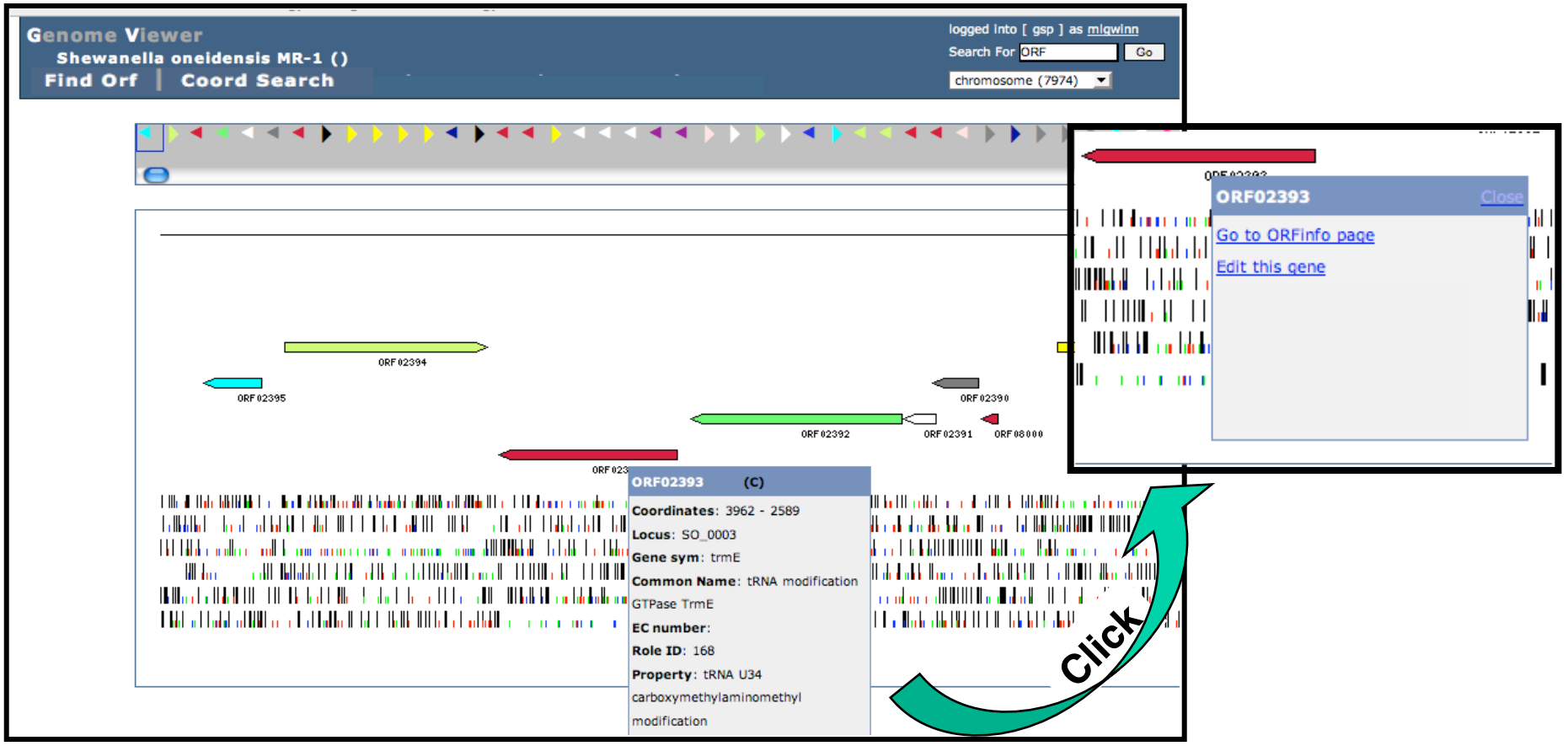
If you click on the pull down menu you will get a selection of options. Each of these when selected will generate a new page with the desired information. (Later slides show examples of some of these.)



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



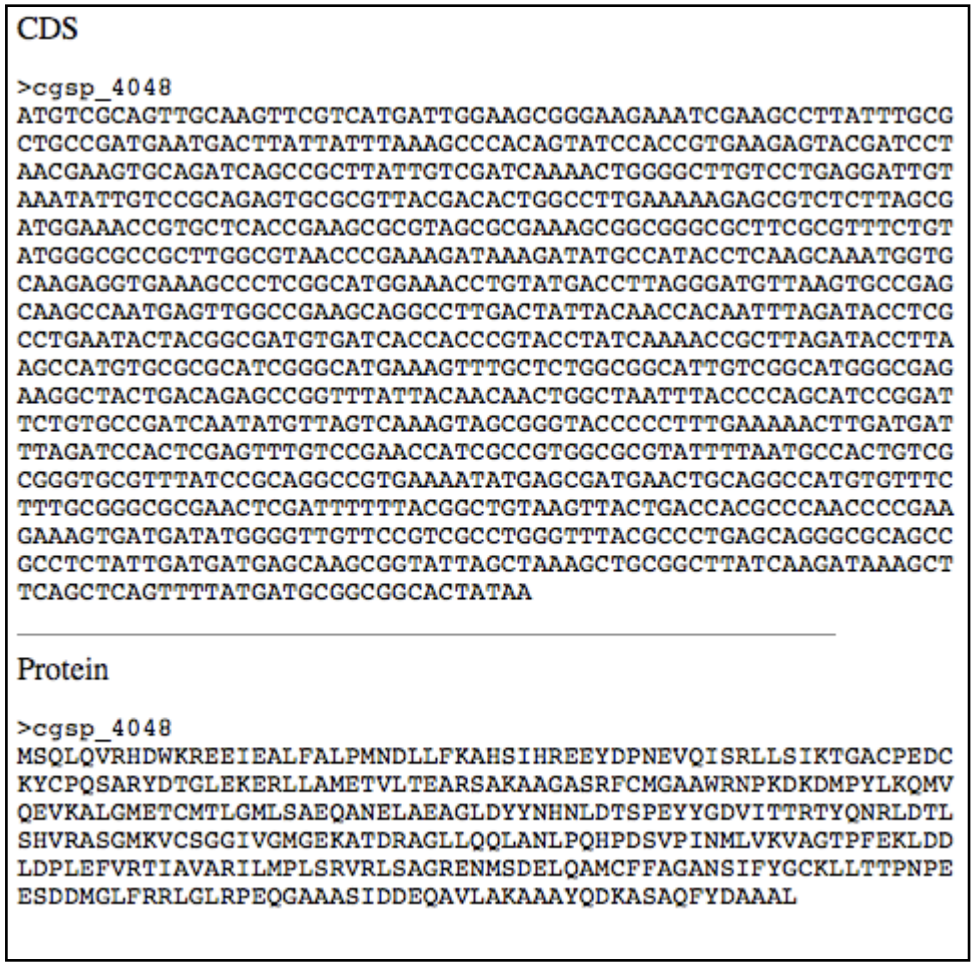
The arrows represent the predicted gene set. They are color-coded according to TIGR role id. The small arrows on the top represent the entire molecule along one scroll bar. The larger arrows depict a zoomed in view of one area of the genome. Mousing over the arrows brings up a box with info on the protein. 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.

Links from the Gene Curation Page - View sequence

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



CDS

>cgsp_4048
ATGTCGCAGTTGCAAGTTCGTCATGATTGGAAGCGGGAAGAAATCGAAGCCTTATTTGCG
CTGCCGATGAATGACTTATTATTTAAAGCCCACAGTATCCACCGTGAAGAGTACGATCCT
AACGAAGTGCAGATCAGCCGCTTATTGTCGATCAAAACTGGGGCTTGTCCTGAGGATTGT
AAATATTGTCCGCAGAGTGCGCGTTACGACACTGGCCTTGAAAAAGAGCGTCTCTTAGCG
ATGGAAACCGTGCTCACCGAAGCGCGTAGCGCGAAAGCGGCGGGCGCTTCGCGTTTCTGT
ATGGGCGCCGCTTGGCGTAACCCGAAAGATAAAGATATGCCATACCTCAAGCAAATGGTG
CAAGAGGTGAAAGCCCTCGGCATGGAAACCTGTATGACCTTAGGGATGTTAAGTGCCGAG
CAAGCCAATGAGTTGGCCGAAGCAGGCCTTGACTATTACAACCACAATTTAGATACCTCG
CCTGAATACTACGGCGATGTGATCACCACCCGTACCTATCAAAACCGCTTAGATACCTTA
AGCCATGTGCGCGCATCGGGCATGAAAGTTTGCTCTGGCGGCATTGTCGGCATGGGCGAG
AAGGCTACTGACAGAGCCGGTTTATTACAACAACTGGCTAATTTACCCCAGCATCCGGAT
TCTGTGCCGATCAATATGTTAGTCAAAGTAGCGGGTACCCCCTTTGAAAAACTTGATGAT
TTAGATCCACTCGAGTTTGTCCGAACCATCGCCGTGGCGCGTATTTTAATGCCACTGTCG
CGGGTGCGTTTATCCGCAGGCCGTGAAAATATGAGCGATGAACTGCAGGCCATGTGTTTC
TTTGCGGGCGCGAACTCGATTTTTTACGGCTGTAAGTTACTGACCACGCCCAACCCCGAA
GAAAGTGATGATATGGGGTTGTTCCGTCGCCTGGGTTTACGCCCTGAGCAGGGCGCAGCC
GCCTCTATTGATGATGAGCAAGCGGTATTAGCTAAAGCTGCGGCTTATCAAGATAAAGCT
TCAGCTCAGTTTTATGATGCGGCGGCACTATAA

Protein

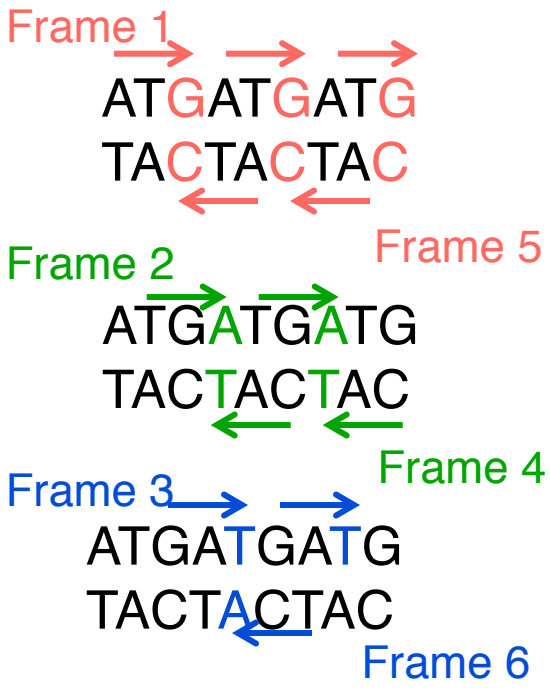
>cgsp_4048
MSQLQVRHDWKREEIEALFALPMNDLLFKAHSIHREEYDPNEVQISRLLSIKTGACPEDC
KYCPQSARYDTGLEKERLLAMETVLTEARSAKAAGASRFCMGAAWRNPKDKDMPYLKQMV
QEVKALGMETCMTLGMLSAEQANELAEAGLDYYNHNLDTSPEYYGDVITTRTYQNRLDTL
SHVRASGMKVCSGGIVGMGEKATDRAGLLQQLANLPQHPDSVPINMLVKVAGTPFEKLDD
LDPLEFVRTIAVARILMPLSRVRLSAGRENMSDELQAMCFFAGANSIFYGCKLLTTPNPE
ESDDMGLFRRLGLRPEQGAAASIDDEQAVLAKAAAYQDKASAQFYDAAAL

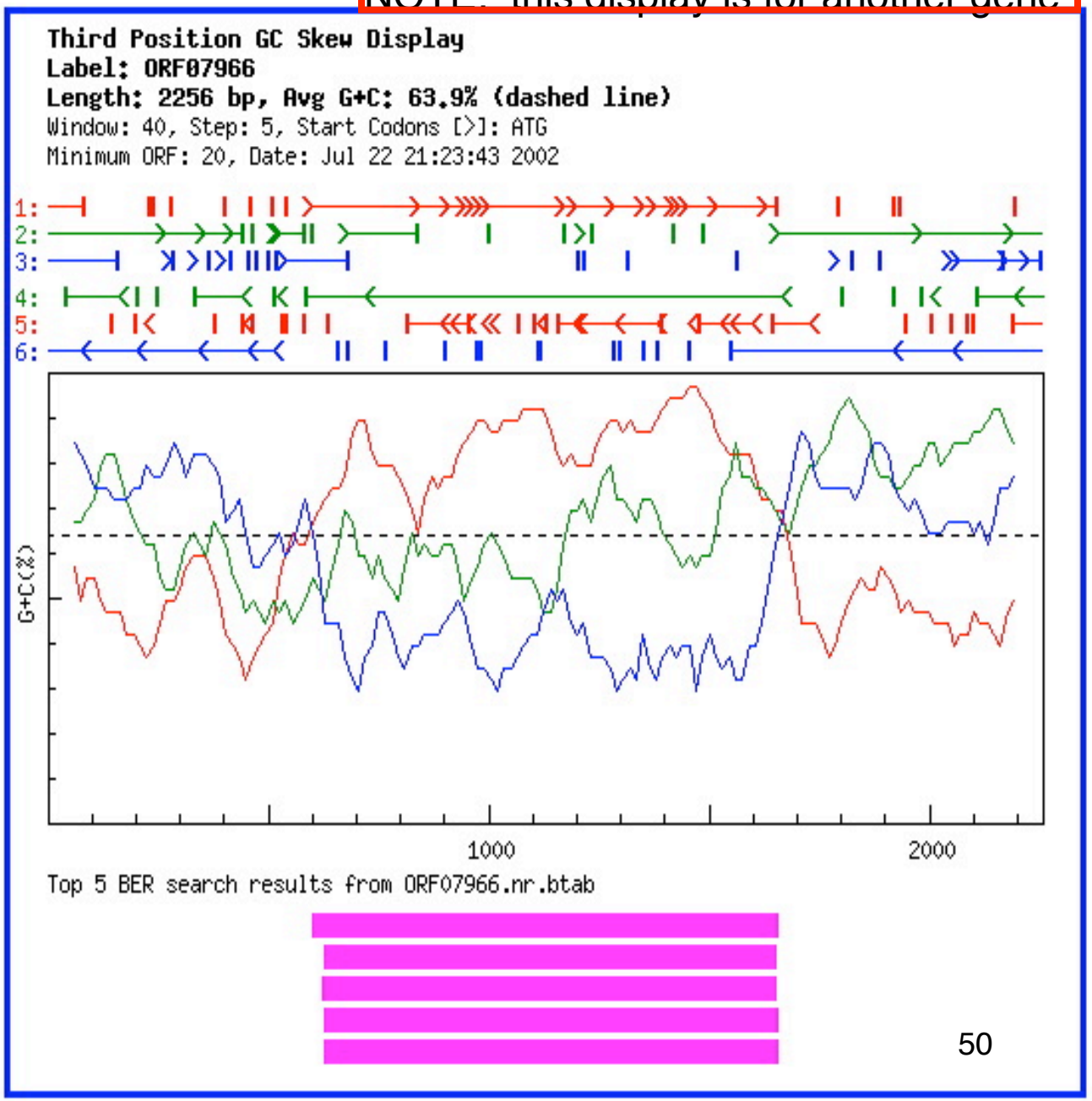
Links from the Gene Curation Page - Third position GC skew

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!