

A Guide to



Michelle Gwinn Giglio

1

Table of Contents (for the most popular topics)

topic	page #s
Getting started	3-5
“Welcome to Manatee” page and links	6
TIGR role category breakdown	7
Genome calculations	8
“Annotation Tools” page and links	9-18
-Gene List	17
Gene Curation Page	19-45
-BER section	22-31
-HMM section	32-33
-GO section	38-42
Gene Ontology	38-42
Genome Viewer	46-48
TIGR role categories	15-16,43
Annotation Checklist	51

2

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

3

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”

4

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>

Manatee Login

user_name: *NOTE*

password:

database:

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.

5

“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.

Welcome to Manatee

Home | Help | Logout | Logged into [cgspu] as [mgjain](#)
organism: *Shewanella oneidensis* MR-1

This is the main menu page for the Manatee tool. One can access genes directly (with gene's id number or name) or link to additional menus with more options.

ACCESS LISTINGS

- > Annotation Tools
- > Genome Viewer
- > Genome Calculations
- > Role Category Breakdown
- Overlap Summary

ACCESS GENE CURATION PAGE

gene_id:

SEARCH GENES BY PROTEIN NAME

protein name:

CHANGE ORGANISM DATABASE

database:

BLASTN blast nucleotide sequence against nucleotide sequence of predicted genes in this genome
 BLASTP blast protein sequence against amino acid sequence of predicted genes in the genome
 TBLASTN blast protein sequence against the entire genome sequence

Paste nucleotide or protein sequence below:

Run against NCBI databases: **NCBI Blast**

Data file downloads (potentially long download times)

- > GO Dumper (Tab delimited file of GO annotation)
- > Nucleotide Sequence Dumper (Multifasta File)
- > Protein Sequence Dumper (Multifasta File)
- > Coordinate Dumper (Tab delimited file of gene coordinates)
- > Whole Genome Dumper (Nucleotide Fasta File)
- > Annotation Dumper (Tab delimited file of annotation)
- > Genbank Dumper (For use in Artemis, BioPerl, etc.)
- > GFF3 Dumper (For use in GBrowse, JBrowse, etc.)
- > TBL Dumper (For submission to NCBI, along with the nucleotide FASTA)

6

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 Type Distribution

feature name	feature count	feature type	start sites	number	percent
► transcript	4849	transcript	► ATG:	4037	83.3% (2887)
► tRNA	101	tRNA	► GTG:	501	10.3% (323)
			► TTG:	311	6.4% (156)
			► OTHER:	0	0.0% (0.0%)

Numbers in parentheses do not include hypothetical proteins

Information Table

► sequence id:	cgsassembly.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:	cgsassembly.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.

This screenshot shows the 'Welcome to Manatee' page. At the top right, it says 'Logged into [cgspl] as msiglio' and 'organism: Shewanella oneidensis MR-1'. Below this is a navigation bar with links for Home, Help, Logout, and additional menus. The main content area is titled 'ACCESS LISTINGS' and contains a list of tools: Annotation Tools (which is selected and highlighted with a red box), Genome Viewer, Genome Calculations, Role Category Breakdown, and Overlap Summary. Below this are sections for 'ACCESS GENE CURATION PAGE', 'SEARCH GENES BY PROTEIN NAME', and 'CHANGE ORGANISM DATABASE'. There are also options for BLASTN, BLASTP, and TBLASTN, and a text input field for pasting nucleotide or protein sequences. At the bottom, there's a link to 'Run against NCBI databases: NCBI Blast' and a section for 'Data file downloads' with various dumper options. A small number '9' is visible on the right side of the download section.

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.

This screenshot shows the 'Annotation Tools' page. At the top right, it says 'Logged into [cgspl] as msiglio' and 'organism: Shewanella oneidensis MR-1'. Below this is a navigation bar with links for Home, Annotation Tools (selected and highlighted with a blue box), and Genome Summary. The main content area has a section titled 'ACCESS GENE LISTS' with options for molecule type (radio buttons for 'all molecules', 'all genes, ordered by role category', 'main role category' with a dropdown menu set to 'Unclassified', 'single role category' with a dropdown menu set to 'role_id', and 'select coordinate range' with input fields for 'end5:' and 'end3:'). Below this is a large search section titled 'SEARCH GENES BY:' with five input fields: 'gene_id / locus' (highlighted with a green border), 'protein name', 'gene symbol', 'EC number', and 'Comment'. There are also 'submit' and 'reset' buttons at the bottom left of the search section.

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.

The screenshot shows the 'Annotation Tools' page with a blue header bar. The main content area has a light gray background. At the top right, there are links: Home, Help, Logout, Logged into [cgspl], and the organism: Shewanella oneidensis MR-1. Below these are three tabs: Home, Annotation Tools (which is selected), and Genome Summary. A large blue button labeled 'ACCESS GENE LISTS' is at the top left. Underneath it is a dropdown menu for 'molecule' with options: 'All molecules' (selected), 'all genes, ordered by role category', 'main role category' (set to 'Unclassified'), 'single role category' (set to 'role_id'), and 'select coordinate range' (with fields for 'end5:' and 'end3:'). To the left of the search form are two small buttons: 'submit' and 'reset'. The search form itself is titled 'SEARCH GENES BY:' and contains four radio buttons: 'gene_id / locus' (unchecked), 'protein name' (checked and set to 'kinase'), 'gene symbol' (checked and set to 'recA'), and 'EC number' (unchecked). Below the search form is a text input field for 'Comment'.

11

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.

The screenshot shows the 'Annotation Tools' page with a blue header bar. The main content area has a light gray background. At the top right, there are links: Home, Help, Logout, Logged into [cgspl], and the organism: Shewanella oneidensis MR-1. Below these are three tabs: Home, Annotation Tools (selected), and Genome Summary. A large blue button labeled 'ACCESS GENE LISTS' is at the top left. Underneath it is a dropdown menu for 'molecule' with options: 'All molecules' (selected), 'all genes, ordered by role category', 'main role category' (set to 'Unclassified'), 'single role category' (set to 'role_id'), and 'select coordinate range' (with fields for 'end5:' and 'end3:'). To the left of the search form are two small buttons: 'submit' and 'reset'. The search form itself is titled 'SEARCH GENES BY:' and contains four radio buttons: 'gene_id / locus' (unchecked), 'protein name' (unchecked), 'gene symbol' (unchecked), and 'EC number' (checked and set to '1.2.3.4'). Below the search form is a text input field for 'Comment'.

12

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		11285	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	16385		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	Central intermediary metabolism
70 Aromatic amino acid family	100 Amino sugars
71 Aspartate family	698 One-carbon metabolism
73 Glutamate family	103 Phosphorus compounds
74 Pyruvate family	104 Polyamine biosynthesis
75 Serine family	106 Sulfur metabolism
161 Histidine family	179 Nitrogen fixation
69 Other	160 Nitrogen metabolism
Purines, pyrimidines, nucleosides, and nucleotides	709 Electron carrier regeneration
123 2'-Deoxyribonucleotide metabolism	102 Other
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	Energy metabolism
176 Biosynthesis	108 Aerobic
177 Degradation	109 Amino acids and amines
121 Other	110 Anaerobic
Biosynthesis of cofactors, prosthetic groups, and carriers	111 ATP-proton motive force interconversion
77 Biotin	112 Electron transport
78 Folic acid	113 Entner-Doudoroff
79 Heme, porphyrin, and cobalamin	114 Fermentation
80 Lipoate	116 Glycolysis/gluconeogenesis
81 Menaquinone and ubiquinone	117 Pentose phosphate pathway
82 Molybdopterin	118 Pyruvate dehydrogenase
83 Pantothenate and coenzyme A	119 Sugars
84 Pyridoxine	120 TCA cycle
85 Riboflavin, FMN, and FAD	159 Methanogenesis
86 Glutathione	105 Biosynthesis and degradation of polysaccharides
162 Thiamine	164 Photosynthesis
163 Pyridine nucleotides	180 Chemoautotrophy
191 Chlorophyll	184 Other
707 Siderophores	
76 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

15

TIGR Role Categories - Page 2

DNA metabolism	Cell envelope
132 DNA replication, recombination, and repair	91 Surface structures
183 Restriction/modification	89 Biosynthesis and degradation of murein sacculus and peptidoglycan
131 Degradation of DNA	90 Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides
170 Chromosome-associated proteins	88 Other
130 Other	
Transcription	Cellular processes
134 Degradation of RNA	93 Cell division
135 DNA-dependent RNA polymerase	188 Chemotaxis and motility
165 Transcription factors	701 Cell adhesion
166 RNA processing	702 Conjugation
133 Other	96 Detoxification
Protein synthesis	98 DNA Transformation
137 tRNA aminoacylation	705 Sporulation and Germination
158 Ribosomal proteins: synthesis and modification	94 Toxin production and resistance
168 tRNA and rRNA base modification	187 Pathogenesis
169 Translation factors	149 Adaptations to atypical conditions
136 Other	706 Biosynthesis of natural products
Protein fate	92 Other
97 Protein and peptide secretion and trafficking	Mobile and extrachromosomal element functions
140 Protein modification and repair	186 Plasmid functions
95 Protein folding and stabilization	152 Prophage functions
138 Degradation of proteins, peptides, and glycopeptides	154 Transposon functions
189 Other	708 Other
Regulatory functions	Unknown
261 DNA interactions	703 Enzymes of unknown specificity
262 RNA interactions	157 General
263 Protein interactions	
264 Small molecule interactions	Hypothetical
129 Other	156 Conserved
Signal transduction	704 Domain
699 Two-component systems	856 General
700 PTS	
710 Other	Disrupted reading frame
	270 NULL

16

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

Home | Help | Logon | Logged into [cgsp] as ...
organism: Shewanella oneidensis MR-1

This List contains ORFs which are currently assigned to TIGR microbial role categories. It is sorted by role category.

All categories > Unclassified (626) > Amino acid biosynthesis (56) > Purines, pyrimidines, nucleosides, and nucleotides (38) > Fatty acid and phospholipid metabolism (26) > Biosynthesis of cofactors, prosthetic groups, and carriers (92) > Central intermediary metabolism (30) > Hypothetical proteins (1750) > Energy metabolism (174) > Transport and binding proteins (282) > DNA metabolism (73) > Transcription (64) > Protein synthesis (131) > Protein fate (180) > Regulatory functions (209) > Signal transduction (36) > Cell envelope (136) > Cellular processes (162) > Mobile and extrachromosomal element functions (188) > Unknown function (743) > Disrupted reading frame (0) > Viral functions (0) > Glimmer rejects (0)

Amino acid biosynthesis

Aromatic amino acid family							Role id: 70			
C	seq id	gene_id	locus	end5	end3	gene name	gene symbol	ec	other roles	start_edit
	cgsassembly.1	cgs_1644		3129854	3131582	anthranilate synthase component I	trE	4.1.3.27		
	cgsassembly.1	cgs_1029		772997	771939	3-deoxy-7-phosphoglyceraldehyde reductase		2.5.1.54		
	cgsassembly.1	cgs_4549		1415682	1416773	3-deoxy-7-phosphoglyceraldehyde reductase		2.5.1.54		
	cgsassembly.1	cgs_2217		3134866	3136056	tryptophan synthase, beta subunit	trB	4.2.1.20		
	cgsassembly.1	cgs_1585		294362	295441	3-dehydroquinate synthase	trB	4.2.3.4		mgiglio
	cgsassembly.1	cgs_196		3559863	3559588	trp operon repressor test3	trR		261	
	cgsassembly.1	cgs_2463		536953	536513	3-dehydroquinate dehydratase, type II	trQ	4.2.1.10		
	cgsassembly.1	cgs_3841		3135062	3136898	tryptophan synthase, alpha subunit	trA	4.2.1.20		
	cgsassembly.1	cgs_3923		3199559	3198774	chorismate synthase	trC	4.2.3.5		mgiglio
	cgsassembly.1	cgs_3689		2509016	2507736	3-phosphoglycerate kinase	trA			

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

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

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

The ORFs can be ordered according to any of the blue headers by clicking on that header. 17

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.

Shewanella oneidensis MR-1 | Role Information For Role_id 77

The role_info.cgi script is executed from the Submit web display page and directs the user to a web display page that contains Single Role Category.

Role 77 Biosynthesis of cofactors, prosthetic groups, and carriers - Biotin

Role Info:

Genes involved in the synthesis of biotin.

pathway from 6-carboxyhexanoyl-CoA plus L-alanine to biotin:
 step gene
 1 8-amino-7-oxononanoate synthase (bioF)
 TIGR00858: bioF
 2 adenosylmethionine-8-amino-7-oxononanoate aminotransferase (bioA)
 TIGR00508: bioA
 3 dethiobiotin synthetase (bioD)
 TIGR00347: bioD
 4 biotin synthase (bioB)
 TIGR00433: bioB
 Other genes also involved:
 BioA bifunctional protein (birA)
 acts as operon repressor, synthesizes corepressor, activates biotin, and transfers activated biotin to proteins
 Biotin synthase protein BioC (bioC)
 involved in an early, undefined step in biotin synthesis
 Biotin sulfoxide reductase (bioZ)
 changes biotin sulfoxide back to biotin, scavenging reaction
 TIGR01738 bioH protein (bioH)
 in early steps of biotin biosynthesis
 TIGR01204 bioW protein = 6-carboxyhexanoate-CoA ligase
 found in Bacillus and Methanococcus, involved in biotin synthesis
 BioW plus BioP of Bacillus can replace bioC and bioH of E. coli
 (says PMID:2110099)

In many, but by no means all, organisms most of these genes can be found in an operon.

bioC protein: BioC is a flavodoxin thought to function as an electron transporter (role_id=112) and in biotin biosynthesis (role_id=77). bioC neighbors oriC in E. coli. Early studies on bioC expression demonstrate a dramatic effect on initiation of chromosome duplication at oriC on minichromosomes. This role has not been demonstrated in duplication of the wild type chromosome. Additionally, the minichromosome is not necessarily a valid model for chromosomal duplication because of the additional association with the

submit | Update Role Note For 77

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.

The screenshot shows two main sections of the GCP:

GENE CURATION INFORMATION

- CGSP_2735 ()**
- View BER Searches** (long load time)
asmbL_id: CGSP_assembly.1
- Reload Page**
- end5/end3:** 2856754 / 2855711
gene length: 1044
protein length: 347
- database:** cgspu
feat_name/locus:
New Gene
- Genome viewer**
View Sequences
3rd position GC skew
- Delete gene**
- Update searches**

GENE IDENTIFICATION

- gene name:** biotin synthase
- gene_sym:** bioB
- EC number(s):** 2.8.1.6 **EC name:** biotin synthase activity
- private comment:**
- public comment:**
- Assigned By:**
- Date:**
- submit**

19

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.

The screenshot shows two main sections of the GCP:

GENE CURATION INFORMATION

- CGSP_2735 ()**
- View BER Searches** (long load time)
asmbL_id: CGSP_assembly.1
- Reload Page**
- end5/end3:** 2856754 / 2855711
gene length: 1044
protein length: 347
- database:** cgspu
feat_name/locus:
New Gene
- Genome viewer**
View Sequences
3rd position GC skew
- Delete gene**
- Update searches**

GENE IDENTIFICATION

- gene name:** biotin synthase
- gene_sym:** bioB
- EC number(s):** 2.8.1.6 **EC name:** biotin synthase activity
- private comment:**
- public comment:**
- Assigned By:**
- Date:**
- submit**

20

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 .

The screenshot shows two stacked forms. The top form is titled 'GENE CURATION INFORMATION' and contains fields for 'CGSP_2735 (0)', 'View BER Searches (long load time)', 'asmbi_id: CGSP_Assembly.1', 'end5/end3: 2856754 / 2855711', 'database: cgspu', 'feat_name / locus: [empty]', 'gene length: 1044', 'protein length: 347', 'Reload Page', 'Genome viewer', 'View Sequences', 'Delete gene', and 'Update searches'. The bottom form is titled 'GENE IDENTIFICATION' and contains fields for 'Gene name: biotin synthase', 'Make Hypothetical', 'gene_sym: bioB', 'EC number(s): 2.8.1.6', 'EC GO suggestions: GO:0094976 [add] biotin synthase activity', 'private comment: [large text area]', 'public comment: [large text area]', 'Assigned By: [empty]', and 'Date: [empty]'. There are 'submit' and 'cancel' buttons at the top right of each form.

21

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.

The screenshot shows a table titled 'BER SKIM' with a 'View BER Searches (long load time)' header and a 'search date: Mon Feb 21 09:55:54 2011' header. The table has columns: accession, %ID, length, description, taxon, and p-value. The 'description' column contains colored text: green for trusted proteins like 'Biotin synthase' and blue for others like 'Biotin synthase' (which links to a detailed view). The 'taxon' column lists various bacterial genera. The 'p-value' column shows values like 1e-127, 0, 1e-171, etc. A 'Show all' button is at the top right, and a 'submit' button is at the bottom right.

View BER Searches (long load time)				search date: Mon Feb 21 09:55:54 2011	
accession	%ID	length	description	taxon	p-value
UniRef100_A7ZJ14	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_Q08418	86.0	349	Biotin synthase	Shewanella frigidimarina NCIMB 400	1e-167
UniRef100_Q12NN4	85.8	356	Biotin synthase	Shewanella denitrificans OS217	1e-165
UniRef100_A1SS19	83.4	352	Biotin synthase	Shewanella amazonensis SB2B	1e-163
UniRef100_B1KJP7	84.3	353	Biotin synthase	Shewanella woodyi ATCC 51908	1e-162
UniRef100_A8QDN8	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_A8DID7	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	Bermellia 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_Q2SB04	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	Celvibrio japonicus Ueda107	1e-134
UniRef100_UP10000E1102B	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.

[Link to info pages for the match protein UniRef/UniProt.](#)

BER Alignment detail: Boxed Header

66.0/79.7% over 343aa	Enterobacteriaceae
• UniRef100_A7ZJI4 Biotin synthase n=35 Tax=Enterobacteriaceae RepID=BIOB_ECO24	

- 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

```
cgsp.CDS.141942892.1( 7 - 350 of 351 aa)
SP|P12996| IO ECOLI(4 - 346 of 346) iota synthase (EC 2.8.1.6) ( iota synthetase).
%Identity = 66.0 %Similarity = 79.7
Gaps = 1 InDels = 3 Frame Shifts = 0
Primary Frame = 1 [343, 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.

25

BER Alignment detail: alignment of amino acids

cataaaaagtttatctcgccgtacggagggtgccaagtttagcaaccggcgcaggcaactttaaaggtcggattccagtc
 cgaacataggctcatatgaagagaacttcttaatttacagtagaaaaaaatatggttcacgcgcagaagcagcga
 agataaaaagtaggggattttggaaacacatggggtaataacacctcttagtttcaggccccaggcattttttattgggtgc
 -2 9 19 29 39 49 59 69
 PR*NTKIKGCSMSQLQVRHDWKREEIEALFALPMDNLLFKAHSHIREEYDNEVQISRLLS1KTGACPEDCKYCPQSARY
 : | : | : || : : : : : : : : : : : : : : : : : :
 MMADRIHWTVGQAQALFDKPLLELLFEAQTVHRQHFDPRQVQVSTLLS1KTGACPEDCKYCPQSARY
 10 20 30 40 50 60

- 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

26

UniProt- #1

name, EC#
gene_symbol
Taxonomy

Info about
function

UniProt- #2

Sequence info

27

UniProtKB - UniProtKB

Search Blast * Align Retrieve ID Mapping *

Search in Query Protein Knowledgebase (UniProtKB) Search Clear Fields >

Reviewed, UniProtKB/Swiss-Prot P15042 (DNLJ_ECOLI)

Last modified March 23, 2010. Version 110. History...

Clusters with 100%, 90%, 50% identity | Documents (3) | Third-party data | Customize display | Contribute | Send feedback | Read comments (0) or add your own | TAIR XML | BLAST XML | GFF | FASTA

Names and origin - Protein attributes - General annotation (Comments) - Ontologies - Sequence annotation (Features) - Sequences - References - Cross-references - Entry information - Relevant documents

Names and origin

Protein names Recommended name: DNA ligase [EC-5.5.1.2]
Alternative names: Polydeoxyribonucleotide synthase (NAD+)

Gene names Name: ligA
Synonyms: dnlA, lig, lop, pdc
Ordered Locus Names b2411, JW2403

Organism Escherichia coli (strain K12) (Complete proteome) [HAMAP]

Taxonomic identifier 83333 (NCBI)

Taxonomic lineage Bacteria • Proteobacteria • Gammaproteobacteria • Enterobacteriales • Enterobacteriaceae • Escherichia

Protein attributes

Sequence length 671 AA.

Sequence status Complete.

Protein existence Evidence at protein level.

General annotation (Comments)

Function DNA ligase that catalyzes the formation of phosphodiester linkages between 5'-phosphoryl and 3'-hydroxyl groups in double-stranded DNA using NAD as a coenzyme and as the energy source for the reaction. It is essential for DNA replication and repair of damaged DNA. [HAMAP MF_01588]

Catalytic activity NAD⁺ + (deoxyribonucleotide)(n) + (deoxyribonucleotide)(m) = AMP + nicotinamide nucleotide + (deoxyribonucleotide)(n+m). [Ref]

Cofactor Magnesium. [Ref]

Sequence similarities Belongs to the NAD-dependent DNA ligase family. LigA subfamily.
Contains 1 BRCT domain.

Ontologies

Keywords Biological process DNA damage, DNA repair, DNA replication
Ligand Magnesium, Metal binding, NAD, Zinc
Molecular function Ligase, 3D-structure, Complete proteome, Direct protein sequencing
Gene Ontology (GO) Biological process DNA repair, Inferred from electronic annotation, Source: UniProtKB-KW DNA ligase, Inferred from electronic annotation, Source: UniProtKB-KW Cellular component Intracellular, Inferred from electronic annotation, Source: InterPro Molecular function DNA binding, Inferred from electronic annotation, Source: InterPro DNA ligase (NAD+) activity, Inferred from electronic annotation, Source: UniProtKB magnesium ion binding, Inferred from electronic annotation, Source: UniProtKB-KW zinc ion binding, Inferred from electronic annotation, Source: UniProtKB-KW

Complete GO annotation...

Sequence annotation (Features)

Feature key Position(s) Length Description Graphical view Feature identifier

Molecule processing

Chain 1–671 671 DNA ligase [HAMAP MF_01588] PRO_0000161745

Regions

	Start	End	Description	Graphical view
Domain	593–671	79	BRCT	
Nucleotide binding	32–36	5	NAD [PDB:00]	
Nucleotide binding	61–82	2	NAD [HAMAP MF_01588]	
Region	330–334	5	Interaction with target DNA [HAMAP MF_01588]	
Region	453–458	6	Interaction with target DNA [HAMAP MF_01588]	
Region	519–524	6	Interaction with target DNA [HAMAP MF_01588]	
Region	551–556	6	Interaction with target DNA [HAMAP MF_01588]	

Secondary structure

Sequences

Sequence	Length	Mass (Da)	Tools
P15042-1 [UniParc]	671	73,606	Blast go

Sequence

Last modified November 1, 1997. Version 2. Checksum: D2BDC64DBE6526

16 25 39 45 55 60
MESIEQQTE LTTILRHMET LYHVMDAPEI PDAYDRPHF ELRELETKHE ELITPDPSG
79 85 99 105 115 120
RVGAAPLAAT SQRIVSPVME SLNDVYDQES FLAFNKRQVG RLKNNKEVTC CCELXLGLA
146 150 155 165 175 180
VSILVYENVL VSAATRGCGT QEDCITSWV TIRAIPLKHE GENIPARLSTT RGENPLPQG
199 205 210 225 235 240
FEKINEDAAK TGCKYVPAFF NAACASLGEQ DPRITAKRF TFTFCYGVDT EGEGELPFTG
256 265 270 285 295 300
GRILLQPKRNG LVPSDQRTLC ESAEVVLKQ HVVEEDRPF GFDIDQYVTH VNLSLAQQ2C
319 325 330 345 355 360
GPVARAPFNA VAFXKFPAGQI MTFVNDPZFTV VRCMGAIFFV ARLEPVVNAVC VLVSNAZNIN
378 385 395 405 415 420
ADEIEERLGLR IGDKVYVIRRA GDVIPVQTVYV VLSERPECH EVVVPPTPFCF CGSDVERVQG
438 445 455 465 475 480
EAVARICVQG ICGAQGRKSEI KHFVSRBRME VDGMGDXKIIQ QLVKEVYHES PADLFLKLQG
499 505 510 525 535 540
KLTGGLERKF KSAQAVVHAEKAKETTAR FLYALGIREV GRATAAGAAK YFGTLEALEA
556 565 570 585 595 600
ASIKEEIQQVNF DUGVIVAHY HNFPAZEEHR NVISELLAGE VINHPAPIVDE AEEIDSEPPG
618 625 630 645 655 660
KTVVLTGDS QMSRDDAMAR LVELGAKYAZ SVSKXTDLYI AGEAGASKLA KAQEIGIIZI
678 DEAEMLRLRIG S

28

UniProt #3

References Hide | Top

[1] "Nucleotide sequence of the *lg* gene and primary structure of DNA ligase of *Escherichia coli*." Ishino Y, Shinagawa H, Makino K, Tsuchihara S, Salvanesca F, Nakata A. Mol. Gen. Genet. 204:1-7(1986) [PubMed: 3018458] [Abstract]

Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA], PROTEIN SEQUENCE OF 1-13 AND 666-671. Strain: K12 / O600 / ATCC 23724 / DSM 3925 / LMG 3041 / NCIB 10222.

[2] O'Connor M.J., Ally A., Ally D., Zhao X., Robichaud M., Backman K. Submitted (APR-1989) to the EMBL/GenBank/DDJB databases
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA].

[3] "Construction of a contiguous 874-kb sequence of the *Escherichia coli*-K12 genome corresponding to 50.0-68.8 min on the linkage map and analysis of its sequence features." Yamamoto Y, Alba H, B Nakamura Y, Nashimoto DNA Res. 4:91-113(1997) [PubMed: 9111111] Cited for: NUCLEOTIDE STRAIN: K12 / W3110 / AT

[4] "The complete genome Blattner F.R., Plunkett G. H.A., Goeden M.A., Rosi Science 277:1453-1474(1997) [PubMed: 9111111] Cited for: NUCLEOTIDE STRAIN: K12 / MG1655 / J

[5] "Highly accurate genome Hayashi K., Morooka N. Mol. Syst. Biol. 2:E1-E5(2004) [PubMed: 15247844] Cited for: NUCLEOTIDE STRAIN: K12 / W3110 / AT

[6] "Direct binding of PtsZ Hale C.A., de Boer P.A.J. Cell 88:175-185(1997) [PubMed: 9111111] Cited for: NUCLEOTIDE STRAIN: PB103.

References Hide | Top

Cross-references

Sequence databases	M30265 Genomic DNA; Translation: AAA24071_1. M24278 Genomic DNA; Translation: AAA24070_1. U00096 Genomic DNA; Translation: AAA24071_1. AP009048 Genomic DNA; Translation: BAA16292_2. U00096 Genomic DNA; Translation: AAB42062_1. LQEC66_B6501B RefSeq AP_003005_1. NP_416906_1.
3D structure databases	PDB 2OWO X-ray PDBsum
ModBase	Search...
Protein-protein interaction databases	
DIP	DIP-10098N.
STRING	P15042.
Proteomic databases	
PRIDE	P15042.
Genome annotation databases	
GeneID	946885.
GenomeReviews	Gene locus JW2405 in contig AP009048_GH. Gene locus JW2411 in contig U00096_GH. ANI: 0.99495
KEGG	

Other databases Hide | Top

Entry Information

Organism-specific:	
EchoBASE	Entry name DNJ_ECOLI
EcoGene	Accession Primary (citable) accession number: P15042
CMR	Secondary accession number(s): P78197, P78198
Entry history	
Integrated into UniProtKB/Swiss-Prot: April 1, 1990	
Last sequence update: November 1, 1997	
Last modified: March 23, 2010	
This is version 110 of the entry and version 2 of the sequence. [Complete history]	
Entry status	Reviewed (UniProtKB/Swiss-Prot)
Annotation project	HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes)

Accession info Hide | Top

29

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

$$\text{Biotinyl-}\text{Biotin} + \text{sulfur} \rightleftharpoons \text{biotin}$$

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_ARATM ; F19206 , BIOB_BACSH ; P53557 , BIOB_BACSU ; P57378 , BIOB_BUCAT ; Q8K9P1 , BIOB_BUCAT ; Q89HK5 , BIOB_BUCBP ; F12997 , BIOB_CITFR ; F46396 , BIOB_CORGL ; F12996 , BIOB_ECOLI ; Q47862 , BIOB_ERWHE ; F44987 , BIOB_HAEIH ; Q92JK3 , BIOB_HELPJ ; Q25956 , BIOB_HELPPY ; Q58692 , BIOB_MEVRB ; F94966 , BIOB_METSK ; F46715 , BIOB_MCYCL ; Q06601 , BIOB_MCYTV ; F12678 , BIOB_SALTY ; Q59778 , BIOB_SCHPO ; P36569 , BIOB_SEMV ; P73538 , BIOB_SYNYS ; P32451 , BIOB_YIRST ;

[View entry in original ENZYME format](#)

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

30

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. Shuan, 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](#) → **Click here to see all the classifications within EC #2 (the transferases)**
[Return to Enzymes home page](#)
[Return to IUBMB Biochemical Nomenclature home page](#)

31

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.

Click to see hits below noise

[submit](#) [all hmms](#)

HMM

TIGR0433: biotin synthase	gene_sym: bioB	ec#: 2.8.1.6	role_id: 77	tc_num: none	Add To Annotation
Isology: equivalog					
Total score: 564.1	Trusted cutoff: 300.00	Gathering cutoff: 300.00	Noise cutoff: 50.00	Total expect: 2e-166	
Trusted cutoff2: 300.00	Gathering cutoff2: 300.00	Noise cutoff2: 50.00			

Coords HMM Coords Score Expect Curation [\[Add To GO Evidence \]](#)

17.313 17.313 / 350

↳ [GO:0004076](#) biotin synthase activity (F)

↳ [GO:0009102](#) biotin biosynthetic process (P)

32

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.

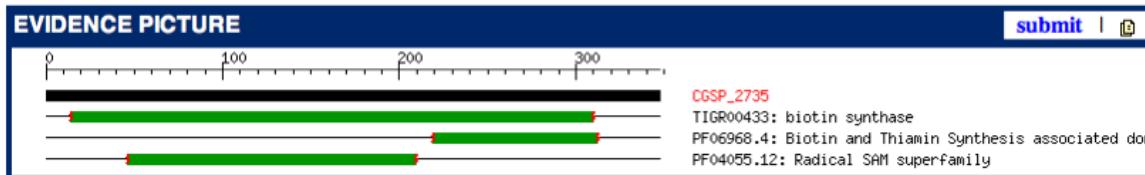
Below is a list of all genes in the organism which hit the HMM and the scores they received. The row with the gold background is the protein of interest. Rows with a green background have scores below the trusted cutoff, rows with a purple background have scores below the noise cutoff.

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_0004078 (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	-160.2	biotin synthase family protein
ORF01034	80	-	76-286	1-350	-178.3	lipoic acid synthetase
ORF03392	162	-	62-370	1-350	-187.3	thiH protein, putative

33

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.



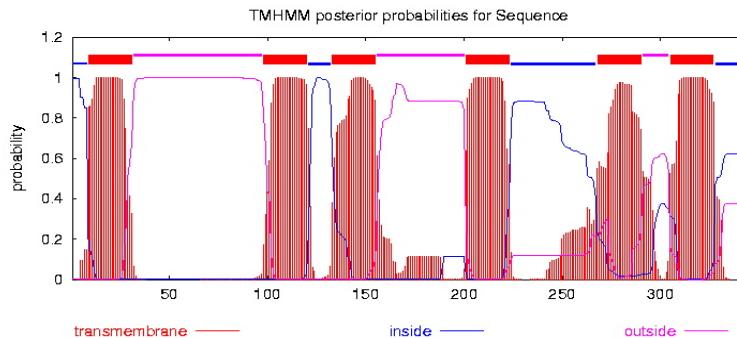
34

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

35

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

[PS01039: Bacterial extracellular solute-binding proteins, family 3 signature.](#)

Match sequence: **GFDIELAKQIAKDA**

Coords	Precision	Recall	Curation
52/65	0.76	0.93	<input checked="" type="checkbox"/> [Add To GO Evidence]

ATTRIBUTES

No Frameshifts Detected.

SIGNAL_P

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.

36

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

Search PROSITE

PROSITE page at ExPASy

NiceSite View of PROSITE: [PDOC00798](#) NOTE: this display is for ORF01166 (documentation)

Bacterial extracellular family 3 signature

PROSITE cross-reference(s)

[PS01039: SBP_BACTERIAL_3](#)

Documentation

Bacterial high affinity transport solutes across the cytoplasmic membrane systems include one or two membrane-associated ATP-b:
<PDOC00185>) and a high affinity are thought to bind the substrate 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 through 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 generally correlated with the

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

This PROSITE entry is copyright by the Swiss Institute of Bioinformatics (SIB). There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.sib-sib.ch/announce/> or email to license@sib-sib.ch).

[View entry in original PROSITE document format](#)

[View entry in raw text format \(no links\)](#)

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.

GENE ONTOLOGY					
delete	gold	assigned	date	evidence	
<input type="checkbox"/>	GO:0004076 [add] [edit] (F) biotin synthase activity	migwinn	07/29/04	ISS: PMD:12368813 with TIGR_TIGRFAMS:TIGR00433	Link to GO search tool
<input type="checkbox"/>	GO:0009102 [add] [edit] (P) biotin biosynthesis	migwinn	07/29/04	ISS: PMD:12368813 with TIGR_TIGRFAMS:TIGR00433	Link to GO suggestions

function	process	component
<input type="button" value="▼"/>	<input type="button" value="▼"/>	<input type="button" value="▼"/>

add go id	evcode	reference	with	qualifier
ISS	TIGR_CMRA annotation			
ISS	TIGR_CMRA annotation			
ISS	TIGR_CMRA annotation			
ISS	TIGR_CMRA annotation			
ISS	TIGR_CMRA annotation			

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 ONTOLOGY					submit go sug search ?
delete	go id	assigned	date	evidence	
<input type="checkbox"/>	GO:0004781 add edit	(F)biotin synthase activity	nigurn	07/29/04	ISS: PMID:1238813; with TIGR_CMRFAMS:TIGR04433
<input type="checkbox"/>	GO:0009102 add edit	(P)biotin biosynthesis	nigurn	07/29/04	ISS: PMID:1238813; with TIGR_CMRFAMS:TIGR04433

function	process	component
?	?	?

add go id	ev code	reference	with	qualifier
ISS	TIGR_CMRFAMS	annotation	?	?
ISS	TIGR_CMRFAMS	annotation	?	?
ISS	TIGR_CMRFAMS	annotation	?	?
ISS	TIGR_CMRFAMS	annotation	?	?
ISS	TIGR_CMRFAMS	annotation	?	?

39

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.

40

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

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

The screenshot shows the UniProt GO term editor interface. At the top, there are five rows for adding GO terms, each with dropdown menus for 'add go id' (ISS), 'ev code' (TIGR_CMR:annotation), 'reference' (TIGR_CMR:annotation), 'with' (empty), and 'qualifier' (empty). Below this is a large central panel for an HMM search result for TIGR0433 biotin synthase. The panel includes fields for 'submit', 'all terms', 'Add To Annotation', 'Add To GO Evidence', 'Cords', 'HMM_Cords', 'Score', 'Expect', and 'Curate'. A green box highlights the 'Add To GO Evidence' button. Another green box highlights the 'GO:0004076 biotin synthase activity (function)' entry in the 'EC GO suggestions' list. A red box highlights the 'GO:0004076 biotin synthase activity (F)' entry in the 'GO:0004076 add biotin synthase activity (F)' list. A blue box highlights the 'GO:0009102 add biotin biosynthetic process (P)' entry in the 'GO:0009102 add biotin biosynthetic process (P)' list. A red arrow points from the 'Add To GO Evidence' button to the 'GO:0004076 biotin synthase activity (F)' entry. A green arrow points from the 'Add To GO Evidence' button to the 'GO:0009102 add biotin biosynthetic process (P)' entry. A red arrow points from the 'GO:0004076 biotin synthase activity (F)' entry to the 'GO:0004076 add biotin synthase activity (F)' list. A green arrow points from the 'GO:0009102 add biotin biosynthetic process (P)' entry to the 'GO:0009102 add biotin biosynthetic process (P)' list. A red arrow points from the 'GO:0009102 add biotin biosynthetic process (P)' entry to the 'Add To GO Evidence' button. A red arrow points from the 'GO:0004076 biotin synthase activity (F)' entry to the 'Add To GO Evidence' button.

41

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/>

42

Gene Curation Page - TIGR roles

Click here to view/edit role notes

Click here to enter this role into the "Delete" box

Click on the name of the main role or sub role to take you to a page with the gene list for that main/sub role.

TIGR ROLES

role_id	delete	main role	sub role
77	del	Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin

submit | role help | history

Add role_ids (separate with spaces):

Delete role_ids (click on ids above):

Add or delete role ids with these boxes.

Click here for a list of TIGR roles.

43

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

SUBMIT DATA

Start confidence not calculated.
 Start Site Curated:
 Completed:

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

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.)

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.

Submit Reset

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

44

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.

CGSP_2735 - Shewanella oneidensis MR-1

GENE CURATION INFORMATION

CGSP_2735 ()

View BER Searches (long load time)
asmbL_id: CGSP_assembly.1

end5/end3: 2856754 / 2855711
gene length: 1044
protein length: 347

database: cgspu
feat_name / locus:
New Gene

Reload Page

Genome viewer
View Sequences
3rd position GC skew

Delete gene (Red Box)

Update searches (Purple Box)

GENE IDENTIFICATION

submit |

gene name: biotin synthase

gene sym: bioB

EC number(s): 2.8.1.6 EC name

EC GO suggestions: GO:0004076 add biotin synthase activity

private comment:

public comment:

Assigned By: Date:

45

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.

CGSP_2735 - Shewanella oneidensis MR-1

GENE CURATION INFORMATION

CGSP_2735 ()

View BER Searches (long load time)
asmbL_id: CGSP_assembly.1

Reload Page

Genome viewer (Red Box)

View Sequences
3rd position GC skew

GENE IDENTIFICATION

gene name: biotin synthase

gene sym: bioB

EC number(s): 2.8.1.6 EC name

private comment:
public comment:

Assigned By: Date:

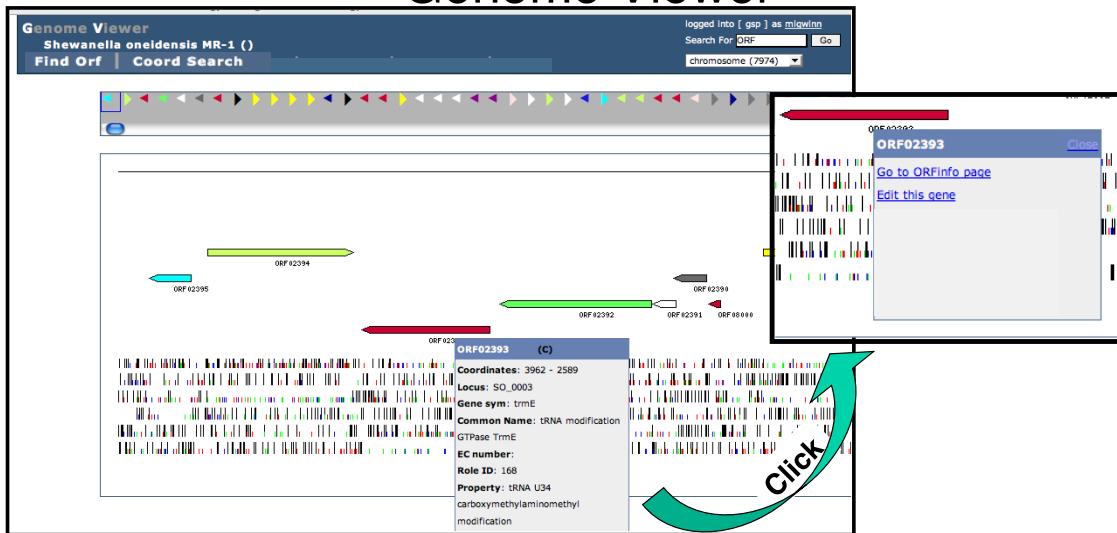
cgsp.transcript.141930837.1

1.261 Mb 1.262 Mb 1.263 Mb 1.264 Mb 1.265 Mb 1.266 Mb 1.267 Mb 1.268 Mb 1.269 Mb 1.270 Mb

cgsP_1637 cgsP_3115 cgsP_4446 cgsP_417 cgsP_300 cgsP_2067 cgsP_209 cgsP_2055

46

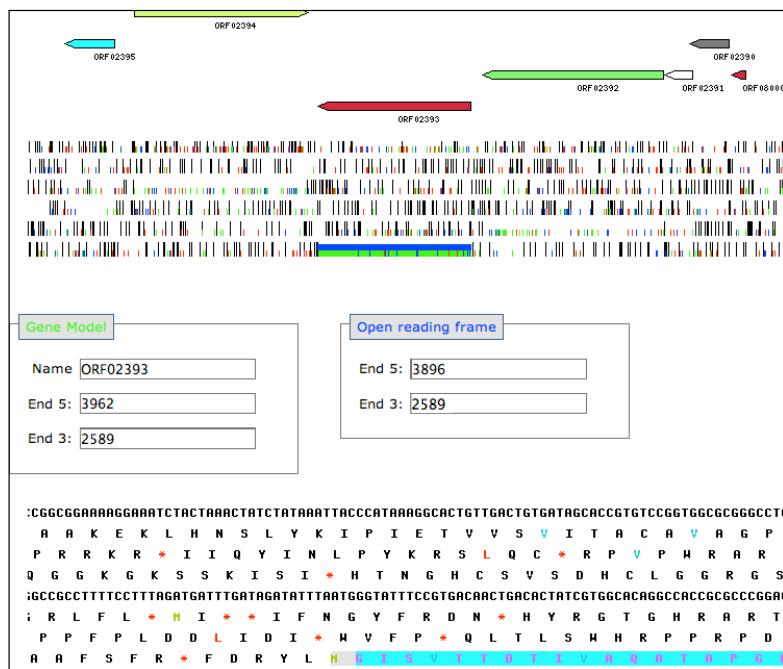
Gene Model Curation in Manatee: Genome Viewer



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.

47

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.

48

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

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

Next page

CGSP_2735 - Shewanella oneidensis MR-1

GENE CURATION INFORMATION

CGSP_2735 ()

- > View BER Searches (long load time)
- asmbi_id: CGSP_assembly.1

end5/end3: 2856754 / 2856754
gene length: 1876
protein length: 347

GENE IDENTIFICATION

gene name: biotin synthase
gene sym: biob
EC number(s): 2.8.1.6
private comment:
public comment:
Assigned By: Date:

CDS

```
>cgsp_4048
ATGTCGGCACTTCCAAGTTCTCATGATTGAAACCCGGAAAATCGAACGCCATTATTGCC
CTGGCGATGAATGACTATTATTAACCCCAACAGTATCACCGCTGAAGACTACGATCCT
AACGAACTTGCAGATTCACCCGGCTTATTCTCGATCAAAAATCTGGGGCTTGCTCTGAGGATTC
AAATAATTCTCCCGAGACTGGCGCTTACQACACTGGCGCTTGAAGAAAAGACGTCTTACCG
ATGGAAACCGTGCCTCAGCAAGCCGCTACCCGGAAAGCCGGCGGGCTTCCCGCTTCTGT
ATGGCCGGCGCTTGGCTTAACCCCAAAGATAAAGATAATGCTACACCTAACCAATCGTC
CAAGAGGTGAAGCCCTCGGATGAAACCTCTATGACCTTAGGGATGTTAACTGCCAG
CAACGAACTTGCAGATTCACCCGGCTTACQACACTGGCGCTTGAAGAAAAGACGTCTTACCG
CTCTGAATACTACGGCGATCTGATCACCCACCTTACCTTACAAACCCCTTGAAGATACCTT
ACCCGCACTGGCTTATTACACAACTCGCTTAATTACCCACATCCCGAT
ATGGCGATCAATATGTTACTCTAAACAGTACGGGGTACCCCTTGAAGAAAACCTGATGAT
TTAGATCAGATGGCTTCTGGCAACCATGCCGTGGCCCTTATTAAATGCACTOTCG
CGGTGCGCTTATCCGGCGCTTGAAGATAATGACCGATGAACCTCACGGCATCTGTTTC
TTTGGGGCGCGGAACTCGATTTTACGCTGTAAGTTACTGACCCACCCCCAACCCGAA
GAAACTGATGATATGGCTTCTGGCTGGGTTACCCCTGACCAGGGCGCAGCC
GCCCTAATGATGATGAGCAAGCGGTATTAGCTAAAGCTGGCTTATCAAGATAAGCT
TCACCTCAGTTATGATGGGGCGCACTATAA
```

Protein

```
>cgsp_4048
MSQLQVRIDWKREEIEALFALPMNDLLFKAHISIREEYDPNEVQISRLLSIKTGACPEDC
KYCPoSARYDTGLEKERLAMETVLTTEARSAAKAGASRFCMGAAWRNPKDKDMYPLKQMV
QEVKALGMETCMILGMLSAEQANELAEAGLDYNNHNLDTSPPEYYGDVITTRTYONRLDTL
SHVRASGMKVCSGGIVGMGEKATDRAGLQLQNLANLPQHPDSPVIMNLVKVAGTPFEKLDD
LDPLPEFVRTIAVARILMPLSLRVLRSAGRENMSDLQAMCFAGANSIFYGCKLTTPNPE
ESDDMGFLFRRRLGLRPEQGAASIDDEQAVLAKAAAYQDKASAQFYDAAAL
```

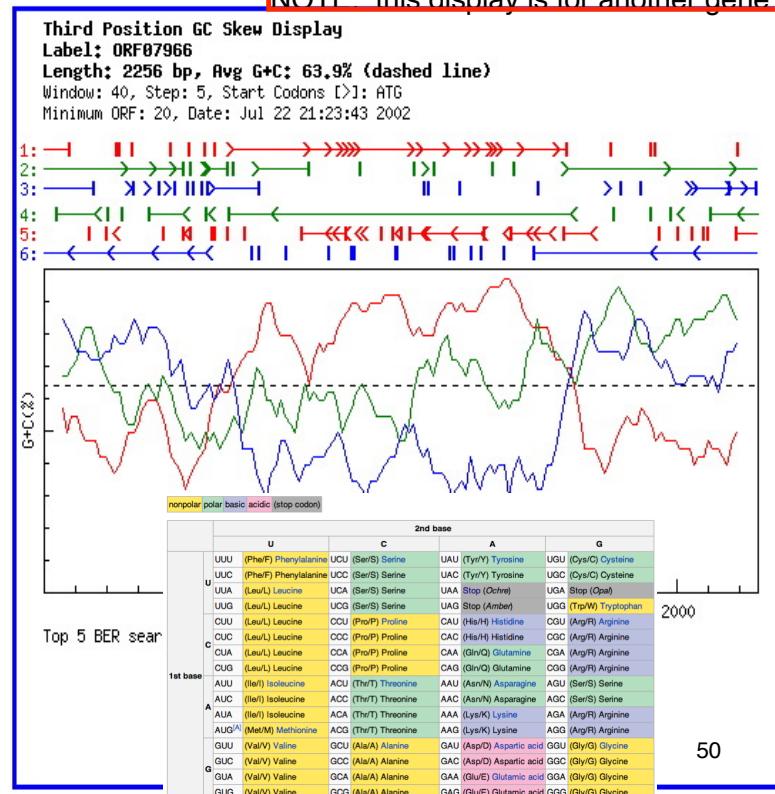
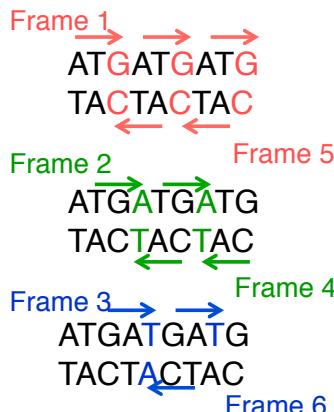
49

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.



50

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!