

Module 5: Alternative Open Reading Frame

Objective

The objectives of this module are:

1. To verify the work of the gene caller by looking for evidence supporting the proposed start codon.
2. To provide evidence for an alternative start codon if the one identified by the gene caller is incorrect.

Materials

To perform this activity you will need:

- Access to the internet on a computer equipped with the most recent version of Firefox (preferred), Chrome or Safari.
- To have completed the sign up for GENI-ACT described in the Signing Up for GENI-ACT section of the manual.

Background

Review the "Prokaryotic Gene Structure, Transcription and Translation" section in the Background document to familiarize yourself with start and stop codons, open reading frames and the process of translation of a bacterial mRNA.

The procedures you will be performing in this module attempt to confirm the start (or initiation) codon for your protein. The gene caller program has automatically analyzed the genome of your bacterium and has already predicted the start and stop codons for your protein, but this is one of the areas where automated gene calling results in errors. The way the gene caller works to predict that a gene exists is to find a start, or initiation codon (ATG, methionine; TTG, leucine or CTG, valine) codon in the genomic sequence followed by an open reading frame of significant length, that ends in a stop codon (TAA, TAG or TGA) in frame with the initiation codon. The gene caller looks at both the top and bottom strands of DNA and in all 6 reading frames to identify long open reading frames.

Another characteristic that supports that a section of genomic DNA encodes a gene is if there is a ribosome binding site consensus sequence immediately upstream (5') of the proposed start codon. In *E. coli* bacteria the consensus (most common) form of the sequence is 5'-GGAGGU-3' and is called the Shine Dalgarno sequence. You can see from the weblogo generated from 149 genes in *E. coli* that there is some variability in the consensus sequence (Figure 5.1). There is the likelihood of efficient translation of an mRNA when a Shine Dalgarno sequence is immediately upstream of a possible start codon. **A shine Dalgarno sequence is not absolutely essential for translation to occur, however, as there are some genes that do not have a Shine Dalgarno sequence upstream of their start codons.**

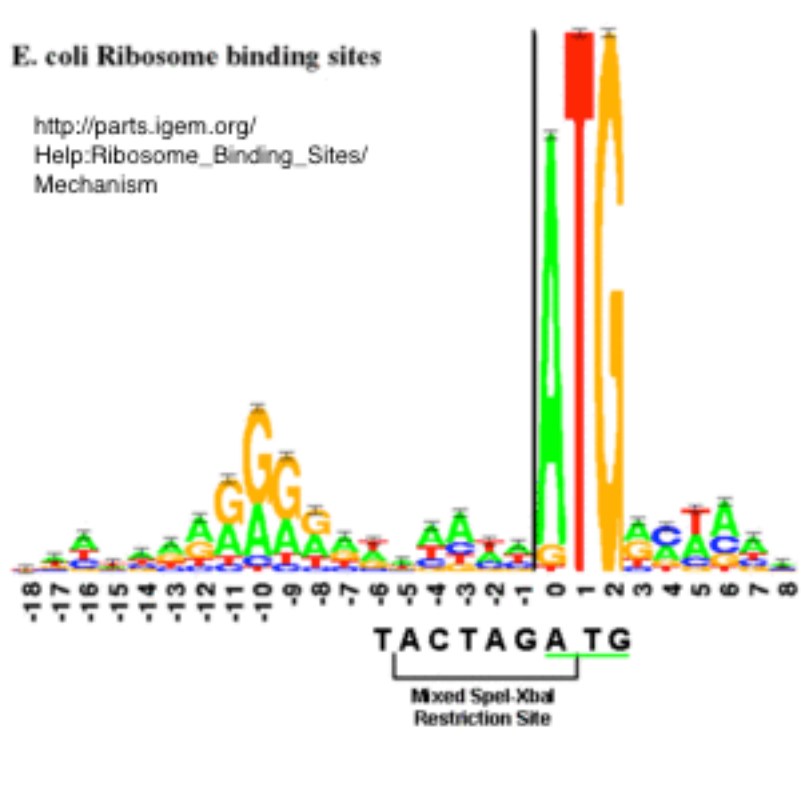


Figure 5.1.
WebLogo of
ribosome
binding site
sequences in
149 *E. coli*
genes.

The “best” open reading frame in a genomic sequence will thus have a well-defined Shine-Dalgarno sequence that is 5-13 nucleotides upstream of the start codon and a significant number of codons that follow prior to a STOP codon being encountered. You will evaluate your gene for this arrangement and make a decision about whether the open reading frame for your gene is correct as called by the gene caller. In the event that you feel the gene caller has made an error, you will provide evidence for the correct start codon and subsequent open reading frame for your gene.

The BLAST and T-Coffee results you obtained in previous modules may offer you potential clues that the proposed open reading frame for your protein might not be correct. Look at that start and stop amino acids in your protein compared to others in your BLAST and T-Coffee alignments. If your protein always starts with amino acid 1, but all the other hits start aligning with a higher number amino acid, it could be that some amino acids are “missing” from the amino terminus of your protein (i.e. a start codon downstream of the true start codon was selected by Gene Caller to translate the DNA sequence). Conversely, if the alignments begin with a higher number amino acid in your gene aligning with lower number amino acids in subjects, it could be that a start codon upstream of the true start codon in your protein was selected by the gene caller to translate the DNA sequence.

If your protein has the name “hypothetical” and you did not get a significant number of BLAST hits in even the NR database, it could be that the gene you are working on was called in error by the gene caller. This module will allow you to look at all open reading frames on both strands of DNA, and it might lead you to conclude that the gene that resides in this stretch of DNA actually is on the opposite strand from that

proposed, or that the true product encoded by the gene is in a different reading frame. This is also a reminder to look back at previous results from earlier modules as you move through the annotation of your gene, as data collected in all modules needs to be considered as you formulate hypotheses about the gene you are annotating.

Procedures

1. Log into IMG/M using the following link: <http://img.jgi.doe.gov/cgi-bin/edu/main.cgi>. Click on the Find Genes tab at the top of the page as indicated by the arrow in Figure 5.2 below and select the Gene Search option from the pull down menu.

IMG Content

Datasets	JGI	All
Bacteria	5763	39296
Archaea	269	773
Eukarya	25	220
Plasmids		1192
Viruses		3907
Genome Fragments		1192
Metagenome & Metatranscriptome	4322	5699
Total Datasets		52279

Last Datasets Added On:

Genome	2016-06-26
Metagenome	2016-06-29

[Project Map](#)
[Metagenome Projects Map](#)
[System Requirements](#)

Hands on training available at the [Microbial Genomics & Metagenomics Workshop](#)

The **Integrated Microbial Genomes (IMG)** system serves as a community resource for analysis and annotation of genome and metagenome datasets in a comprehensive comparative context. The **IMG data warehouse** integrates genome and metagenome datasets provided by IMG users with a comprehensive set of publicly available genome and metagenome datasets.

IMG provides users with tools ([IMG UI Map](#)) for analyzing publicly available genome datasets (<http://nar.oxfordjournals.org/content/42/D1/D560>) and metagenome datasets (<http://nar.oxfordjournals.org/content/42/D1/D568>).

IMG Statistics

Metagenome and Metatranscriptome dataset distribution:

Sequenced at:	Engineered		Environmental		Host-associated	
	JGI	All	JGI	All	JGI	All
Metagenome	365	476	2667	2912	420	1419
Metatranscriptome	120	134	649	655	100	102

IMG contains 250 public studies, 5699 public metagenome datasets (5194 unique samples) distributed as follows: (Public Metagenome count / Public Metatranscriptome count)

	Engineered	Environmental	Host-associated
Bioreactor	476 / 134	2912 / 655	1419 / 102
Air	15 / 4	31 / 0	52 / 0
Algae			

Figure 5.2. The IMG/M Home page. The Find Genes tab is indicated by the

2. In the Gene Search window paste the locus tag for your gene in the keyword box, select Locus Tag (list, no MER-FS Metagenome) from the filters pull down menu (Figure 5.3) and click Go.

Quick Genome Search:

My Analysis Carts**: 0 [Genomes](#) | 0 [Scaffolds](#) | 0 [Functions](#) | 0 [Genes](#)

Home Find Genomes Find Genes Find Functions Compare Genomes OMICS My IMG Data Marts Help

Home > Find Genes loaded.

Gene Search

Find genes in selected genomes by keyword. It's required to add selections into "Selected Genomes" unless blocked.
*MER-FS Metagenome supported search filters.

Keyword

Filters

Figure 5.3. The Gene Search window at IMG.

- The page that will appear is called the Gene Details page in the IMG Database (Figure 5.4). It contains a wealth of information known about the gene.

Gene Detail

[Gene Information](#)
[Find Candidate Product Name](#)
[Evidence For Function Predictions](#)
[Sequence Search](#)
[External Sequence Search](#)
[IMG Sequence Search](#)
[Homolog Display](#)

Gene Information

Gene Information	
Gene ID	644990317
Gene Symbol	dnaA
Locus Tag	Ksed_00010
IMG Product Name	chromosomal replication initiator protein DnaA
Original Gene Product Name	chromosomal replication initiator protein DnaA
IMG Product Source	
SwissProt Protein Product	
SEED	[Chromosomal replication initiator protein DnaA] figl478801.4.peg.1954 DNA-replication
IMG Term	
Description	PFAM: Bacterial dnaA protein helix-turn-helix domain; Bacterial dnaA protein; TIGRFAM: chromosomal replication initiator protein DnaA
Genome	Kytococcus sedentarius 541, DSM 20547
DNA Coordinates	209..1729 (+)(1521bp)

Figure 5.4. The IMG Gene Details Page. Only the uppermost portion of the page is shown.

4. In the alternate open reading frame module, the automated gene caller program/algorithm proposes the start and stop codon of the sequence in question, and you will determine if these have been called correctly. To verify the position of the start codon predicted by the Gene Caller, go to the IMG Gene Details page and navigate to the Evidence for Function Prediction section. Click the hyperlink for "Sequence Viewer for Alternate ORF Search" as shown by the arrow in Figure 5.5.

Find Candidate Product Name

Method:

Display Option (for sequence based only):

Find Candidate Product Name

Evidence For Function Prediction

Neighborhood

2785024 5000 10000 15000 20000

red = Current Gene
 ||||| CRISPR array
[Sequence Viewer For Alternate ORF Search](#)

Chromosome Viewer colored by

Conserved Neighborhood

[Show neighborhood regions with the same top COG hit \(via top homolog\)](#)
[Show neighborhood regions with this gene's bidirectional best hits](#)

Chromosomal Cassette Viewer By

COG

Figure 5.5. Link to perform alternate open reading frame (ORF) search on the gene details page.

5. The sequence viewer page gives the option of adding bases upstream and downstream of the original ORF coordinates (Figure 5.6).
 - A. In the Select Gene Neighborhood section, enter the number of base pairs you would like to see upstream (e.g. -99bp) and downstream (e.g. +99bp) of the query gene in the appropriate two boxes [Note: the program works best if you enter the same number of base pairs in both boxes. Also use multiples of 3 so the reading frame that your protein is in for analysis will always be ORF 1].
6. Select the minimum ORF size as 60-80 amino acids (If your amino acid sequence is less than 60 amino acids in size enter the number of amino acids in your protein).

JGI **IMG/M** INTEGRATED MICROBIAL GENOMES & MICROBIOME SAMPLES

Quick Genome

My Analysis Carts**:

Home Find Genomes Find Genes Find Functions Compare Genomes

[Home](#) > Find Genes

Sequence Viewer

Select parameters to view the six frame translation
Gene: [644990317](#) *chromosomal replication initiator protein DnaA*
209..1729(+)

Select gene neighborhood:
 bp upstream bp downstream

Select minimum ORF size:
 aa

Output Format:
☐ Text ☒ Graphics

Figure 5.6.
The sequence
viewer page.

7. You can choose the output format as “Graphics” or “Text” depending on the mode in which you want to view the output, but to begin check the box labeled “Graphics” and Click "Submit" (Figure 5.6).

8. A graphical output is obtained which shows the actual sequence and the six different open reading frames (Figure 5.7).

Sequence Viewer

Neighborhood six frame translation with putative ORF's shown below

Gene: [644990730](#) *glycine/D-amino acid oxidase, deaminating*

427153..428313 (-)

-99 upstream +99 downstream

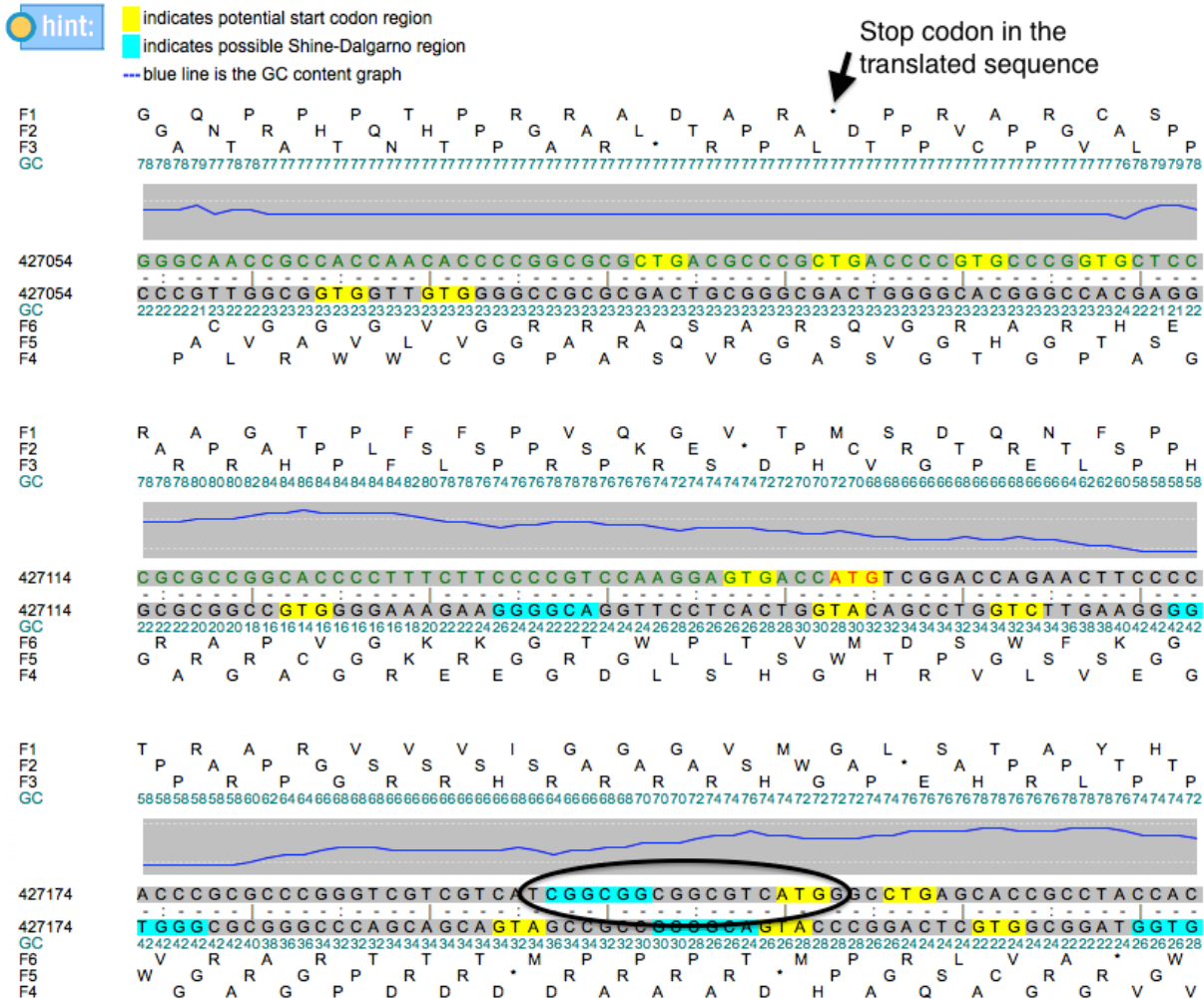


Figure 5.7. The graphical output of adding 99 nucleotides upstream and downstream of the predicted coding sequence of the demonstration gene. The called start codon for the protein is in red and the M in frame 1 above the ATG is the first amino acid in the predicted sequence. The arrow shows a stop codon upstream of the called start codon and the oval indicates a potential start codon downstream of the one called by the gene caller that also has a potential Shine-Dalgarno sequence beginning 7 nucleotides upstream of the ATG (see text for full explanation).

9. The top three rows of letters are amino acids that correlate with the codons or translations of the top strand of DNA (Frames 1, 2 and 3). The next line is the top strand of DNA. The middle line is a place marker to denote nucleotides in intervals of 5 and 10. The sequence of the bottom strand of DNA is then following three rows that are the amino acid translations of the bottom DNA strand (Frames 4, 5 and 6). Note that amino acid sequences are read right to left on frames 4-6 since the complementary strand of DNA is oriented in with its 5' end at the right and 3' end at the left.
10. The bases highlighted in green are the ones that were added upstream or downstream of the actual ORF. The nucleotides of the query gene are black. The bases highlighted in red are the start and stop codons identified by the Gene Caller. Note that the start codon is shown in figure 5.7 and the stop codon is shown in figure 5.8. You will likely need to scroll down the page to find the stop codon for your gene.

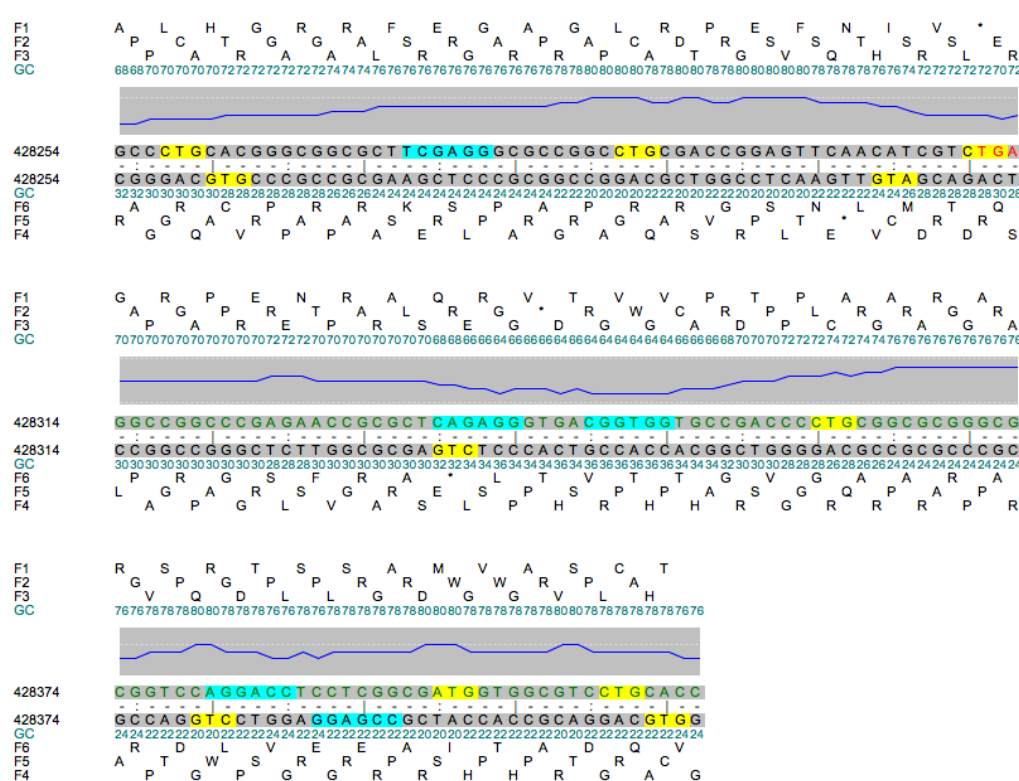


Figure 5.8. The stop codon for the example gene described above (red TGA) in reading frame 1. Holding the cursor over the last nucleotide in the stop codon of the gene you are studying will display the DNA coordinate for the end of the gene.

11. The yellow shaded regions indicate the potential start codon regions and the light blue shaded regions indicate the possible Shine-Dalgarno sequences. The star symbol in the reading frame marks a stop codon.
12. Examine the start codon identified by the Gene Caller for your gene marked in red. Determine if your sequence has a Shine-Dalgarno site 5 to 15 bases upstream of the identified start codon. The Shine-Dalgarno sequence is always on the same DNA strand as the start codon.
13. If the start codon has a proper Shine-Dalgarno sequence, then check if the DNA start coordinate matches with the one you recorded for your gene in the basic information module. Move the cursor

over the first nucleotide of the highlighted in the start codon and a popup box will show up that has the nucleotide number indicated. Make a note of the number. Scroll down the page until you come to the highlighted stop codon in the same reading frame. Hover your cursor over the **LAST** nucleotide in the highlighted stop codon and note the coordinate of that nucleotide. If you have not obtained any other data that suggests an alternative start codon might be present (see below) then you can conclude that the start codon was likely correct. Write a comment in your notebook indicating that the original coordinates are likely correct, and that you have found a Shine-Dalgarno sequence 5-15 bases upstream of the start codon (see figure 5.15 below). It is a good idea to Snip or Grab the upper three rows of the graphical output and save it in your notebook as well. **You do not need to enter any information in the DNA coordinates section of the notebook if you feel the start codon was called correctly.**

14. If you had a significant number of BLAST hits with low e values and high scores in your Sequence Based Similarity module, you can most likely stop at this point and conclude the start and stop codons are correct as called. However, there may be some hints that the start codon is called incorrectly even when a Shine Dalgarno sequence is found 5-15 bases upstream of the start codon called by the gene caller. Such data would include:
 - A. Apparently missing or extra amino acids at the beginning of your protein when compared to its top BLAST hits. This is most easily seen when looking at the multiple sequence alignment of your protein using T-COFFEE in Module 2. **Review your sequence alignment and see if it appears that your protein doesn't start aligning with the majority of sequences in the multiple sequence alignment at the beginning.** This might be indicated if amino acid 1 of your protein didn't match up until, for example, amino acid 20 of the other proteins in the alignment (meaning there may be additional codons upstream of the predicted start codon that should part of your protein), or if amino acid position 1 of the majority of proteins in the alignment match up with, for example amino acid 10 of your protein (meaning the true start codon might be downstream of the one predicted by the gene caller).
15. If you do not find a Shine Dalgarno sequence 5-15 bases upstream of the start codon indicated by the gene caller, as is the case in the example shown in Figure 5.7, or if you have evidence the makes you think that the start codon was called incorrectly, look for an alternative start codon upstream of the original proposed start codon in the same reading frame.
16. If you spot an alternative start codon with Shine Dalgarno sequences 5- 15 bases upstream of it, then scroll over the first nucleotide of the new start codon to obtain the coordinate value. Note that there should be no stop codons between the original and proposed alternative start codon. In the example shown in Figure 5.4, there are **no** start codons upstream of the proposed start codon, and there is a STOP codon present in the upstream translated sequence (Arrow, figure 5.7).
17. If no alternative start codon exists upstream, then look for an alternative start codon **downstream** of the original proposed start codon in the same reading frame. Note that there should be no stop codons between the original and proposed alternative start codon. If you spot an alternative start codon with Shine Dalgarno sequences 5-15 bases upstream of it, then scroll over the first nucleotide of the new start codon to obtain the coordinate value. In the example shown in figure 5.4 there is a start codon downstream of the proposed start codon that has a potential Shine-Dalgarno sequence within the required distance (oval figure 5.7).

18. If you find potential alternative open reading frames either upstream or downstream (or in both directions), you will then need to test them to see if they represent potentially better matches to other proteins in the database. Hit the back button in your browser to return to the sequence viewer start page and then click the “text” option, leaving the number of nucleotides displayed before and after your sequence and the minimum ORF size the same as used in the graphic display above (Figure 5.9).

JGI **IMG/M** INTEGRATED MICROBIAL GENOMES & MICROBIOME SAMPLES

Quick Genome Search

My Analysis Carts**: 0 Genomes

Home Find Genomes Find Genes Find Functions Compare Genomes

Home > Find Genes

Sequence Viewer

Select parameters to view the six frame translation
 Gene: [644990317](#) *chromosomal replication initiator protein DnaA*
 209..1729(+)

Select gene neighborhood:
 bp upstream bp downstream

Select minimum ORF size:
 aa

Output Format:
☒ Text ☐ Graphics

Figure 5.9.
The
sequence
viewer page
with the text
output box
checked to
allow text
versions of
open
reading
frames to be
retrieved.

19. After clicking Submit a list of amino acid FASTA-format sequences of all possible ORFs in the sequence range you selected will be displayed (Figure 5.10). This is an unfiltered list of ORFs, meaning that these sequences will **NOT** necessarily begin with bona fide start codons. You can see in figure 5.10 that there are two open reading frames in frame 1 (the frame in which the example amino acid sequence is found (see arrows and explanation in figure 5.10) The first ORF is an example of ORFs that is too short and the second is an example that does not begin with one of the three potential microbial start codons (ATG, methionine, M; TTG, leucine, L or CTG, valine, V).

img/edu INTEGRATED MICROBIAL GENOME EDUCATION SITE

IMG Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Cart My IMG OMICS Companion Systems

Home > Find Genes Loaded.

Sequence Viewer

Neighborhood six frame translation with putative ORF's shown below
 Gene: [644990730](#) *glycine/D-amino acid oxidase, deaminating*
 427153..428313 (-)
 -99 upstream +99 downstream

hint: To test ORF translation, copy and paste the sequence to BLAST and InterPro scan.

>644990730_1_ORF1 Translation of 644990730 in frame 1, ORF 1, threshold 1, 13aa
 GQPPPTFRADAR

>644990730_1_ORF2 Translation of 644990730 in frame 1, ORF 2, threshold 1, 405aa
 PRARCSRAGTPFFPVQGVMSDQNFPTRARVVVIGGGVMGLSTAYHLAKQGVQDVVLVER
 GELGAGSTCKAAGGVRAQFSDAVNIELGMRSLVFRNFPPELFDQDIDLDECYLFLLERE
 EDLRTFERNVELQRSMGLESRTSVEEAKELSPLISTEGLIAGVWSPEAGHCTPESVVQG
 YARAARALGVRIIRHCEVTDVVREGDTITSVETAQGSIAITDTVVCCAGAWSRALGDMVGV
 DLPVDPVRRELLVTEPMPDLPANVPFTIDFSTTMYFHREGPGLLVGMSNQDEEPPGFSLEH
 TDEWLEQVVEAAGRRVPVLEEVMASRWAGLYEVTDPHNALIGEAGVSRFLYATGFSGH
 GFLQGPVAGQVMAELYLGQTPSVDTVLTALHGRRFEGAGLRPEFNIV

>644990730_1_ORF3 Translation of 644990730 in frame 1, ORF 3, threshold 1, 33aa
 GRPENRAQRVTVVPTPAARARSRTSSAMVASCT

>644990730_2_ORF1 Translation of 644990730 in frame 2, ORF 1, threshold 1, 31aa
 GNRHQHPGALTADPVPGAPAPAPLSSPSKE

>644990730_2_ORF2 Translation of 644990730 in frame 2, ORF 2, threshold 1, 22aa
 PCRTRTSPAPGSSSSAAASWA

>644990730_2_ORF3 Translation of 644990730 in frame 2, ORF 3, threshold 1, 41aa
 APPTTWPSRVCRSTCSWSAASSAPAPPAPPAACAPSSPMR

>644990730_2_ORF4 Translation of 644990730 in frame 2, ORF 4, threshold 1, 70aa
 TSSWACAARSSGSTRSCTRTSTWTSATCFWSCARRTCALSSATSSCSARWGWRRAASP
 ASRRPRSSPR

Annotations:
 First open reading frame in reading frame 1. From first added nucleotide (-99) to the first stop codon that appears in the added sequence.
 Second open reading frame in reading frame 1 (ORF2). From first nucleotide after the stop codon in ORF 1 until the next stop codon. Note that the ORF does not begin with one of the 3 possible start codons. It is just an open reading frame, not a predicted protein sequence. This frame does, however, include the amino acid sequence predicted to be encoded by the gene used as an example, which begins at the first M in line 1 of the sequence.

Figure 5.10. The text output of the open reading frame search. The gene under investigation will be in Frame 1 because a multiple of 3 (+/-99) nucleotides were added to the beginning and end of the gene sequence. Note the two ORFs indicated in frame 1, the first being the result of the stop codon that occurred within 99 nucleotides of the upstream added sequence.

20. Figure 5.11 is a copy of the ORF 2 in reading frame 1 in which the amino acids that were added upstream of the start methionine (M) are colored red.

>644990730_1_ORF2 Translation of 644990730 in frame 1, ORF 2, threshold 1, 405aa

PRARCSRAGTPFFPVQGVTMSDQNFPTRARVVVIGGGVMGLSTAYHLAKQGVQDVVLVER
 R GELGAGSTCKAAGGVRAQFSDAVNIELGMRSLVFRNFPPELFDQDIDLDECYLFLLERE
 EDLRTFERNVELQRSMGLESRTSVEEAKELSPLISTEGLIAGVWSPEAGHCTPESVVQG
 YARAARALGVRIIRHCEVTDVVREGDTITSVETAQGSIAITDTVVCCAGAWSRALGDMVGV
 DLPVDPVRRELLVTEPMPDLPANVPFTIDFSTTMYFHREGPGLLVGMSNQDEEPPGFSLEH
 TDEWLEQVVEAAGRRVPVLEEVMASRWAGLYEVTDPHNALIGEAGVSRFLYATGFSG
 H GFLQGPVAGQVMAELYLGQTPSVDTVLTALHGRRFEGAGLRPEFNIV

Figure 5.11. The 2nd ORF in frame 1 from figure 5.7. See the text for a full explanation.

21. Figure 5.12 is the same sequence that also has the sequence downstream of the called start codon colored blue up the potential open reading frame identified in the previous section (oval, figure 5.7) that had a start codon and a potential Shine-Dalgarno sequence immediately upstream.

>644990730_1_ORF2 Translation of 644990730 in frame 1, ORF 2, threshold 1, 405aa

PRARCSRAGT**PFFPVQGV**TMSDQNFPT**RRARVVVIGGGV**MGLSTAYHLAKQGVQDVVLVE
 R GELGAGSTCKAAGGVRAQFSDAVNIELGMRSLEVFRNFPPELFDQDIDLDECYLFLLERE
 EDLRTFERNVELQRSMGLESRITSVEEAKELSPLISTEGLIAGVWSPEAGHCTPESVVQG
 YARAARALGVRIIRHCEVTDVVREGDTITSVETAQGSIA~~TD~~TVVCCAGAWSRALGDMVGV
 DLPVDPVRRELLVTEPMPDLPANVPFTIDFSTTMYFHREGPGLLVGMSNQDEEPPGFSLEH
 TDEWLEQVV~~EA~~AGRRVPVLEE~~V~~GMASRWAGLYE~~VT~~PDH~~N~~ALIGEAE~~G~~VS~~R~~FLYATGFSG
 H GFLQGPVAVGQVMAELYLGQTPSVDVTALHGRRFEGAGLRPEFNIV

Figure 5.12. ORF from figure 5.8 showing amino acids that would be deleted (blue) if the downstream start codon was correct. A BLAST search would be done for the sequence in black to compare the e value and score with that obtained for the original amino acid sequence as described in the text.

22. Identify the portions of open reading frames that correspond to the alternative start codons you identified above (Figure 5.11 and 5.12). Select the amino acid sequences encoded by the alternate ORFs that start with a potential start codon and then BLAST them at a time the as done before in the Sequence-Based Similarity Module. Construct a “revised” FASTA header that explains the difference from the proposed ORF to make it easier to keep track of the different alternative ORFs you work with. For example, in Figure 5.11 the sequence to BLAST would begin with the M that immediately follows the red highlighted amino acids (i.e. the original start codon). The BLAST search data for the example protein with the original start codon is shown in Figure 5.13. The comparison BLAST search, using the open reading frame downstream from the original that has a potential Shine-Dalgarno sequence, would be done using the sequence the begins with the M just after the red and blue highlighted amino acids shown in Figure 5.12 (i.e. the black text in figure 5.12). Copy and paste the sequence into the BLAST search window and then add a FASTA header (remember that a FASTA header must be by itself on the first line of the sequence submitted for searching and that it must start with the > symbol) that would read something similar to:

>Alternative open reading frame downstream of called start codon for ksed_00010

- A. Submit each as a query for a BLAST search (<http://www.ncbi.nlm.nih.gov/blast>) in the NCBI website using the same database(s) you used in your original BLAST search. The results of the BLAST searches for the example sequence are shown in Figure 5.13 (original sequence) and 5.14 (downstream open reading frame in black from figure 5.12).

- B. Compare results from original blast search with those from new blast search. Make sure you are comparing hits from the same organism (compare accession numbers or click on highlighted matches in the BLAST search to identify from what organism a BLAST hit results).

Descriptions

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

[Alignments](#) [Download](#) [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	RecName: Full=Sarcosine dehydrogenase, mitochondrial; Short=SarDH; Flags: Precursor [Rattus n...	189	189	96%	1e-51	34%	Q64380.2
<input type="checkbox"/>	RecName: Full=Sarcosine dehydrogenase, mitochondrial; Short=SarDH; Flags: Precursor [Mus mus...	188	188	96%	2e-51	34%	Q99LB7.1
<input type="checkbox"/>	RecName: Full=Dimethylglycine dehydrogenase, mitochondrial; AltName: Full=ME2GLYDH; Flags: I...	186	186	96%	6e-51	32%	Q63342.1
<input type="checkbox"/>	RecName: Full=Sarcosine dehydrogenase, mitochondrial; Short=SarDH; AltName: Full=BPR-2; Flag...	183	183	95%	1e-49	33%	Q9UL12.1
<input type="checkbox"/>	RecName: Full=Dimethylglycine dehydrogenase, mitochondrial; AltName: Full=ME2GLYDH; Flags: I...	182	182	96%	2e-49	31%	Q9DBT9.1

Figure 5.13. BLAST results for the sequence with the start codon as called by the gene caller.

Descriptions

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

[Alignments](#) [Download](#) [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	RecName: Full=Dimethylglycine dehydrogenase, mitochondrial; AltName: Full=ME2GLYDH; Flags: I...	172	172	96%	4e-46	31%	Q63342.1
<input type="checkbox"/>	RecName: Full=Sarcosine dehydrogenase, mitochondrial; Short=SarDH; Flags: Precursor [Rattus n...	172	172	97%	1e-45	33%	Q64380.2
<input type="checkbox"/>	RecName: Full=Sarcosine dehydrogenase, mitochondrial; Short=SarDH; Flags: Precursor [Mus mus...	171	171	97%	3e-45	32%	Q99LB7.1
<input type="checkbox"/>	RecName: Full=Dimethylglycine dehydrogenase, mitochondrial; AltName: Full=ME2GLYDH; Flags: I...	168	168	96%	2e-44	30%	Q9DBT9.1
<input type="checkbox"/>	RecName: Full=Sarcosine dehydrogenase, mitochondrial; Short=SarDH; AltName: Full=BPR-2; Flag...	167	167	97%	4e-44	32%	Q9UL12.1

Figure 5.14. BLAST results for the sequence with the alternative start codon. Note the score is less and the e value greater for the top hit as compared to the same values in figure 5.13

- C. If the BLAST statistics have significantly improved for the proposed alternative ORF (E value a significantly smaller number and the score a significantly higher number) then you have evidence that the alternative ORF is a better choice for the translation of the gene. If, on the other hand, the E value increased and or the score decreased, the alternative ORF is not as good a match as the original for other proteins in the database. You would then conclude that you have no evidence to support an alternative start codon for your protein. This is the case for the example protein in which the downstream alternative ORF has a higher e value and lower score compared to the sequence as called by the gene caller (See Figures 5.13 and 5.14).


- D.** Enter a summary of your findings into notebook by way of a comment (Figure 5.15). You can Snip (PC) or Capture (Mac) images of the BLAST scores of the original and alternative open reading frame protein sequences (i.e., as shown in figures 5.13 and 5.14 above and add upload them to the notebook as well to document your findings. **If there is no evidence for a better start codon, snip the first three rows of the graphical alternative open reading frame results, upload the image to the notebook and comment about why your findings support the proposed start codon as being correct. DO NOT enter any information in the proposed DNA coordinates box of the notebook if your conclusion is that the start codon has been called correctly.**

[.] Alternative Open Reading Frame


[Module Instructions](#)

DNA Coordinates

go to the IMG Gene Detail page (click on the IMG Gene OID at the top of the Notebook)

Proposed DNA coordinates (if different from those predicted by IMG) 

..

Explanation of choice 

The coordinates appear to be correct as called. A shorter alternative open reading frame in ORF1 with a potential Shine-Dalgarno sequence was investigated, but the BLAST E value for the shorter ORF was increased and the score was decreased.

Figure 5.15. The lab notebook for the Alternative Open Reading Frame Module. Always write a conclusion about whether an alternative ORF exists for your protein. State that the coordinates appear correct as called for your gene if you feel there is no alternative ORF and your rationale for saying so in the “explanation of choice” box. If you feel that you have identified a better ORF for your gene, write the new start and stop coordinates for the gene in the box and provide a detailed explanation of why you feel the alternative ORF is correct for your protein, including the results of BLAST as described in the text.

- 23.** There are several reasons why you might not find a Shine-Dalgarno sequence upstream of an ORF:
- A.** Mutations or sequencing errors in the DNA sequence that mask a true Shine-Dalgarno Sequence could have occurred.
 - B.** The gene of interest is a part of an operon (determined in a later module). In that case, a ribosome “skid” could happen, allowing translation to take place even without a Shine-Dalgarno sequence.
 - C.** There may be some flexibility in the amount of sequence conservation needed in the Shine-Dalgarno sequence that allows ribosome binding that are unique to your bacterium, and thus not all possible Shine-Dalgarno regions may be found by IMG.

24. Additional work in this module should be done if you are working on a “hypothetical” protein that had few, if any, significant BLAST hits in even the NR database when you did your BLAST searches in the Sequence Based Similarity Module. Such findings may indicate the Gene Caller was wrong to call your gene in the first place, and that a completely different ORF exists. You will determine if this is the case by looking for novel (new) open reading frames.
- A. Use the TEXT option in the sequence viewer. A NOVEL ORF could be in a different reading frame on the same strand as the original ORF (Frames 2 and 3) or be on the opposite strand of DNA (frames 4,5 and 6).
 - B. Select each open reading frame that meets the criterion of being significant in size (greater than 60-80 amino acids) and BLAST them one at a time.
 - C. Compare the BLAST scores obtained from the alternative ORFs to the scores and numbers of hits you obtained from your sequence. If there are more hits with better BLAST scores, you have evidenced that the gene caller made a mistake in determining that your original sequence was a “real” gene.
 - D. Record details of what you did and the amino acid sequence of any translations found with significant hits and scores in your notebook and state that you feel the sequence that you were working on was called in error.
 - E. You would then go back and repeat the earlier modules with the new amino acid sequence and add those results to your notebook.