Annotation Presentation
Week 4

Alternative Start Codons
&
Novel ORFs

You

The Brilliant Student

vs.

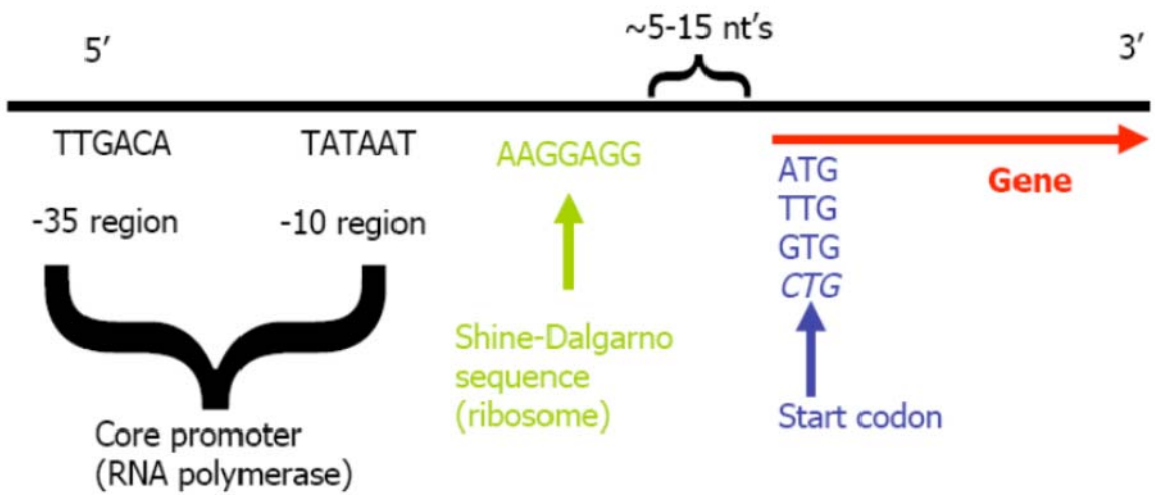
Glimmer

The Automated Gene Caller

The Automated Gene Caller
(i.e., Glimmer)

What does an Automated Gene Caller do?

- Scans all nucleotides in a genome.
- Looks for "punctuation marks" that determine where genes start and stop (i.e., regulatory sequences that define where genes begin and end).



Courtesy of KM Scott

The Brilliant Student (You)

What will you do?

Double check the work of the Gene Caller.

- ✓ Is there evidence supporting <u>proposed</u> start codon?
- ✓ Is there an <u>alternative position</u> possible for the start codon?

Examine all six reading frames.

- ✓ Is there a *novel* open reading frame (ORF) missed by Gene Caller?

Let’s get started with the first task:

Verify the position for the start codon

In the *imgACT* Lab Notebook…

Basic Information Module

- Module Instructions

IMG Gene Object ID

go to the IMG Gene Details page for the proposed gene

enter Gene Object ID (OID)

2501578154

1- Return to Gene Detail page

Quick link to Gene Details page for assigned gene (enter URL)

http://img.jgi.doe.gov/cgi-bin/edu/main.cgi?section=GeneDetail&page=geneDetail&gene_oid=2501578154

DNA Coordinates

go to the IMG Gene Details page for the proposed gene

DNA coordinates

3948986..3951211 (+)

Recall:
You previously entered the DNA coordinates set by the automatic Gene Caller. This information was found on the Gene Detail page.

Gene Detail Page in img/edu

img/edu INTEGRATED MICROBIAL GENOMES EDUCATION SITE

IMG Home | Find Genomes | Find Genes | Find Functions | Compare Genomes | Analysis Carts | MyIMG | Using IMG

Gene Search | BLAST | Phylogenetic Profilers

Gene Detail

Loaded.

Gene Information
Evidence For Function Predictions
Sequence Search
External Sequence Search
IMG Sequence Search
Homolog Display

Keep the Gene Detail page open in separate tab while working

Gene Information

Gene Information	
Gene Object ID	2500607069
Gene Symbol	
Locus Tag	PlimDRAFT_19450
Product Name	Uncharacterized anaerobic dehydrogenase, COG3383
IMG Product Source	COG3383
Genome	Planctomyces limnophilus DSM 3776
DNA Coordinates	2111568..2113793 (-)(2226bp)
Scaffold Source	Planctomyces limnophilus DSM 3776 : PlimDRAFT_4083246_C168 (5423025bp)
IMG ORF Type	
GC Content	0.58
External Links	
Fused Gene	No
Fusion Component	No
Protein Information	
Amino Acid Sequence Length	741aa

Scroll down

Evidence For Function Prediction

Neighborhood

red = Current Gene

green = Positional Cluster Gene in the same KEGG Pathway as the Current Gene

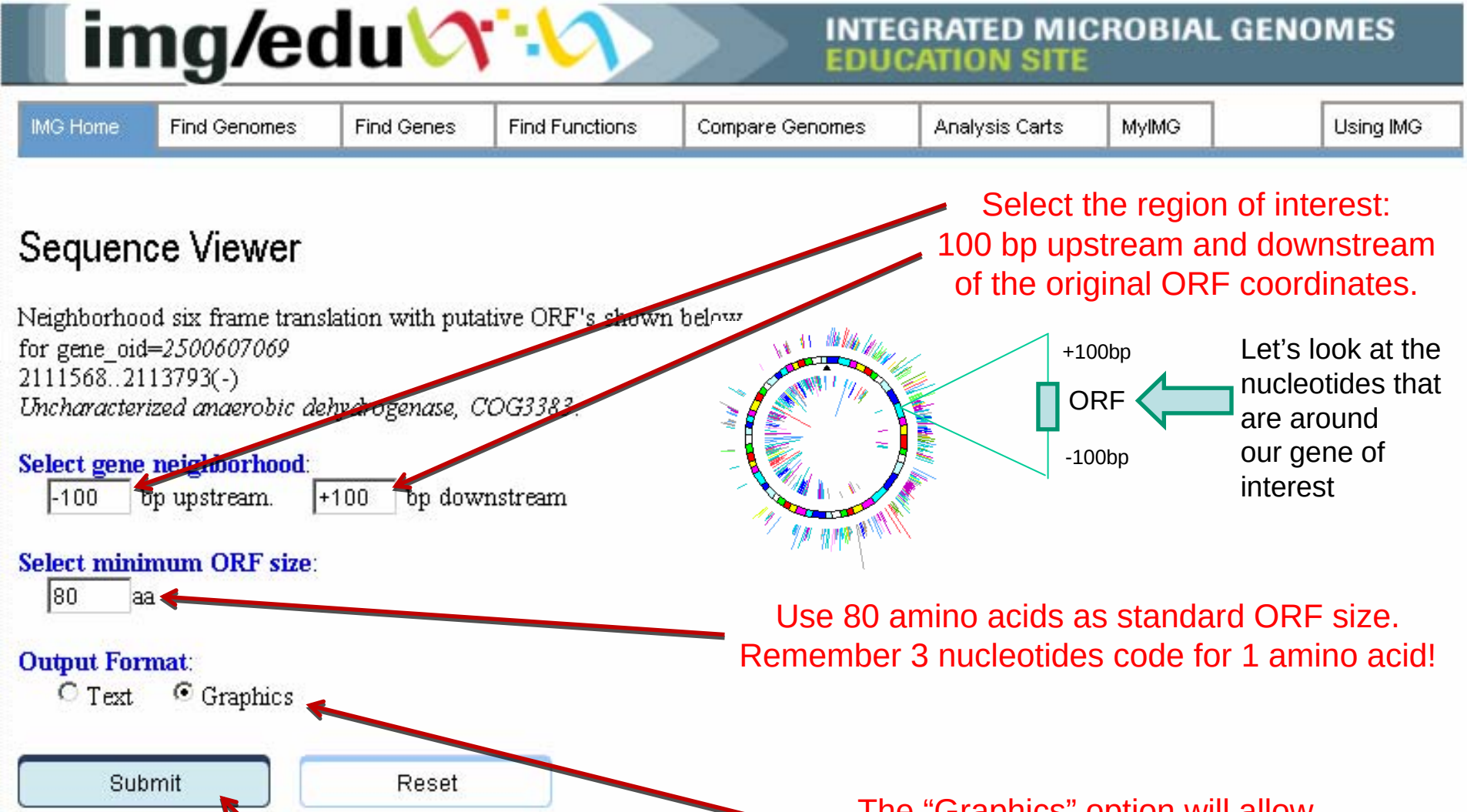
CRISPR array

Sequence Viewer For Alternate ORF Search

Chromosome Viewer colored by COG

Chromosome Viewer colored by GC percentage

"Click" here



Select the region of interest: 100 bp upstream and downstream of the original ORF coordinates.

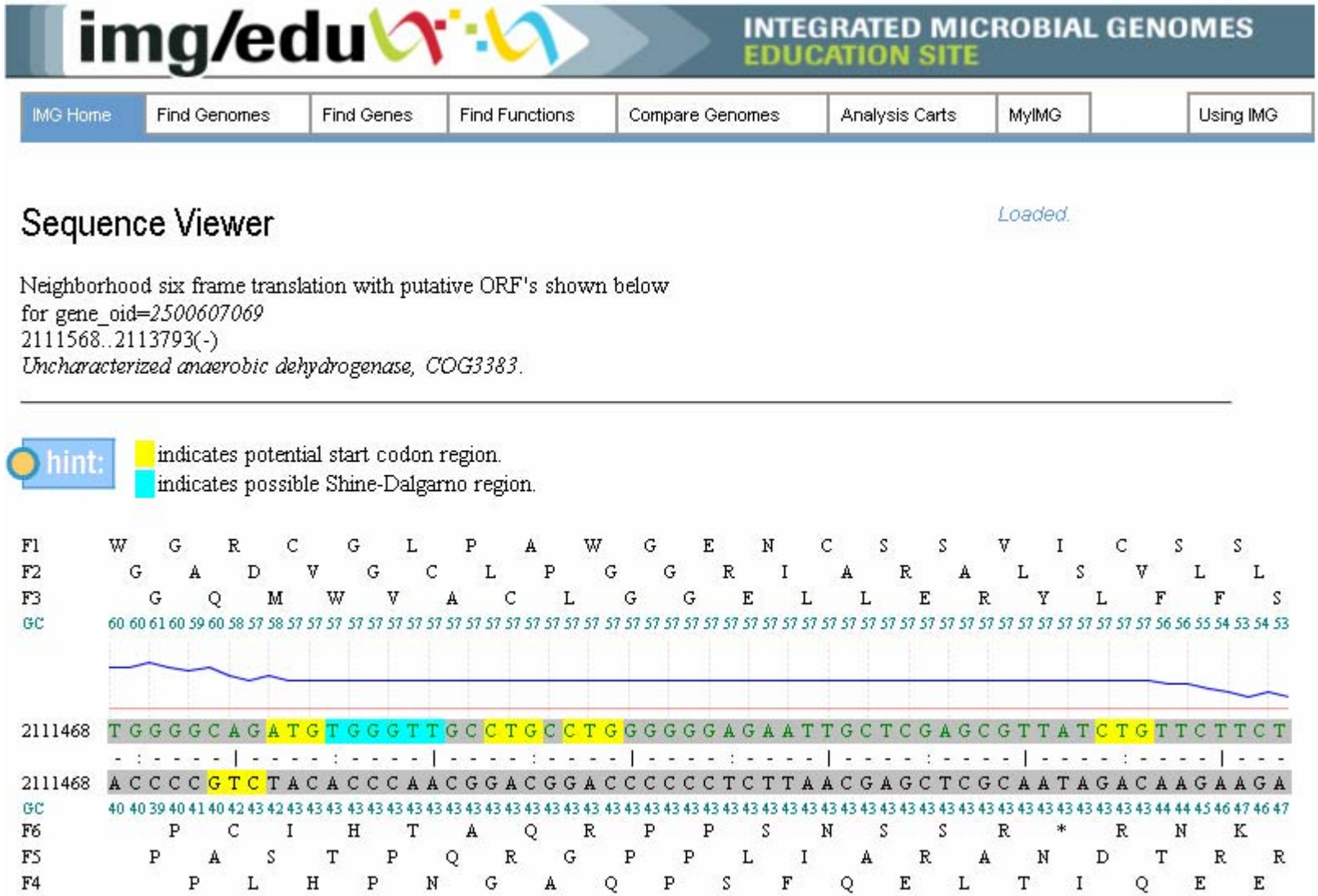
Let's look at the nucleotides that are around our gene of interest

Use 80 amino acids as standard ORF size. Remember 3 nucleotides code for 1 amino acid!

The "Graphics" option will allow us to look at the actual sequence. This helps us complete the first task.

Click!

After you click submit, you will see something like this:



HOW AM I SUPPOSED TO READ THIS??

Deciphering Sequence Viewer in Graphics Mode

img/edu INTEGRATED MICROBIA EDUCATION SITE

IMG Home | Find Genomes | Find Genes | Find Functions | Compare Genomes | Analysis Carts | MyIMG | Using IMG

Sequence Viewer

Loaded.

Neighborhood six frame translation with putative ORF's shown below
for gene_oid=*2500607069*
2111568..2113793(-)
Uncharacterized anaerobic dehydrogenase, COG3383.

hint: indicates potential start codon region.
indicates possible Shine-Dalgarno region.

```
F1   W  G  R  C  G  L  P  A  W  G  E  N  C  S  S  V  I  C  S  S
F2    G  A  D  V  G  C  L  P  G  G  R  I  A  R  A  L  S  V  L  L
F3     G  Q  M  W  V  A  C  L  G  G  E  L  L  E  R  Y  L  F  F  S
GC   60 60 61 60 59 60 58 57 58 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 56 56 55 54 53 54 53

2111468  T G G G G C A G A T G T G G G T T G C C T G C C T G G G G G G A G A A T T G
         - : - - - - | - - - - : - - - - | - - - - : - - - - | - - - - : - - - - | -
2111468  A C C C C G T C T A C A C C C A A C G G A C G G A C C C C C C T C T T A A C
GC       40 40 39 40 41 40 42 43 42 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43
F6        P  C  I  H  T  A  Q  R  P  P  S  N
F5     P  A  S  T  P  Q  R  G  P  P  L  I
F4        P  L  H  P  N  G  A  Q  P  S  F  Q  E  L  T  I  Q  E  E
```

The nucleotide sequence is highlighted in gray

This tells us about GC content in this part of the genome.

DNA coordinates

forward strand

reverse strand

Orientation depends upon whether your gene is on the + or - strand

WHICH STRAND IS MY GENE ON?

Example of plus strand gene orientation:

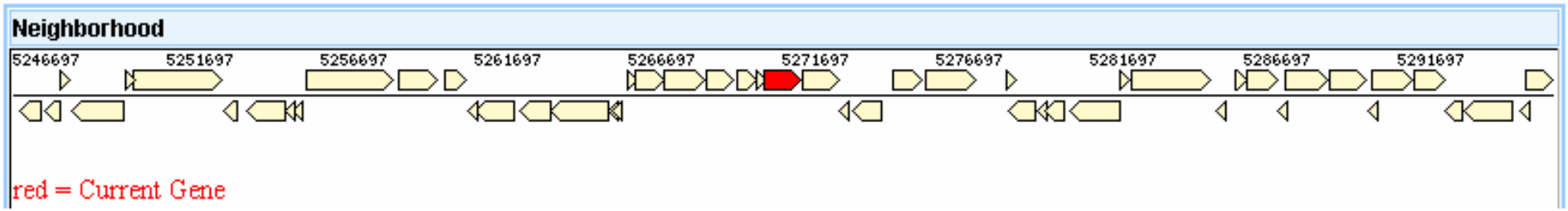
DNA Coordinates

Find this on the Gene Details page for the proposed gene.

enter coordinates

5271121..5272269 (+) (1149bp)

Evidence For Function Prediction



Example of minus strand gene orientation:

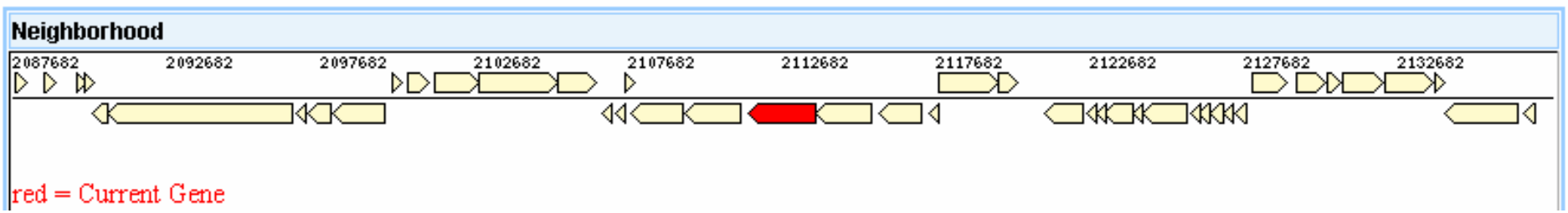
DNA Coordinates

Find this on the Gene Details page for the proposed gene.

enter coordinates

2111568..2113793 (-) (2226bp)

Evidence For Function Prediction



HOW AM I SUPPOSED TO READ THIS??

Deciphering Sequence Viewer in Graphics Mode

Review single letter codes for amino acids!

Sequence Viewer

Neighborhood six frame translation with putative ORF's shown
for gene_oid=*2500607069*
2111568..2113793(-)
Uncharacterized anaerobic dehydrogenase, COG3383.

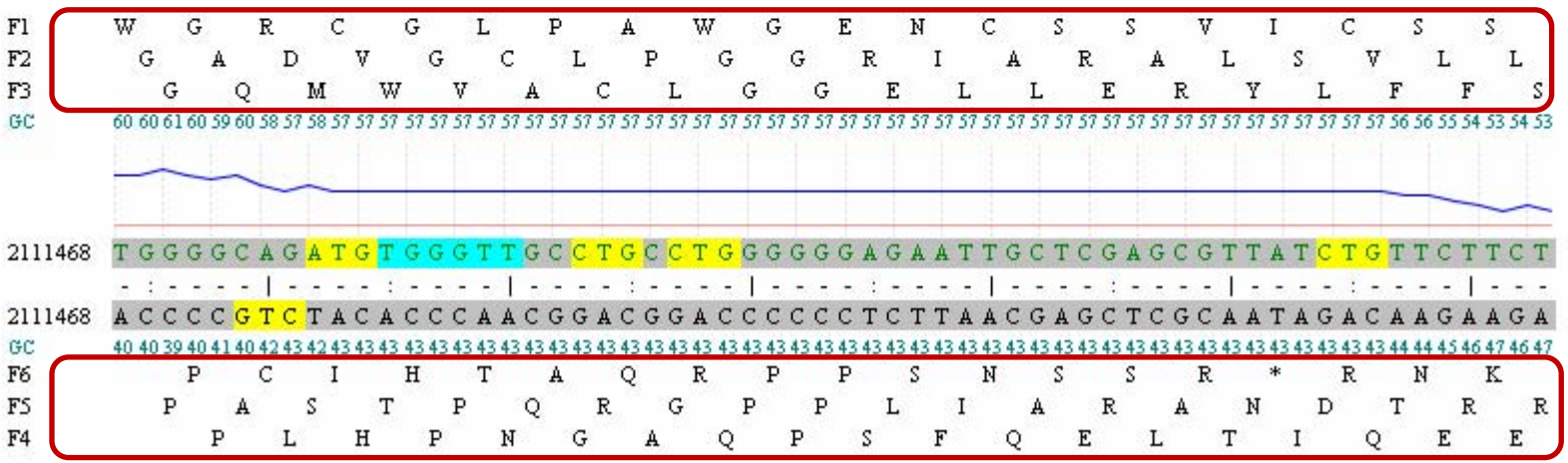
hint: indicates potential start codon region.
indicates possible Shine-Dalgarno region.

Each row above (or below) designated DNA strand represents a different translational reading frame

- amino acids corresponding to codons in nucleotide sequence shifted by one base

AA results from <u>forward</u> DNA strand (F1, F2, & F3)

AA results from <u>reverse</u> DNA strand (F4, F5, & F6)



Helpful tools to keep handy while deciphering!

Insert table of genetic code and table of single letter abbreviations for amino acids

Initiation frequency in *E. coli*: AUG > GUG > UUG > CUG

Stop codons: UAG, UAA, UGA

More on… Deciphering Sequence Viewer in Graphics Mode

Gene expression pathway: DNA → RNA → protein.
We are *interpreting* the protein sequence directly from the DNA, rather than RNA.

START CODON

When you scroll over each nucleotide, DNA coordinates for that position appear above or below the strand.

```
F1      P   L   C   R   T   R   V   C   P   E   E   C   V   H   E   L   V   S   G   *
F2        R   C   V   E   P   A   Y   A   R   R   S   A   F   M   S   W   Y   Q   D   N
F3          A   V   S   N   P   R   M   P   G   G   V   R   S   *   A   G   I   R   I   T
GC      54 56 56 58 58 56 56 56 58 56 56 56 56 54 54 54 56 56 56 56 54 52 50 52 52 54 56 58 60 58 58 60 58 58 58 58 58 58 58 58 60 60 58 60 58 60 62 60 62 64 62 62 60 60 58 58 60 58 58 56
```

2111568

```
2111528 C C G C T G T G T C G A A C C C G C G T A T G C C C G G A G G A G T G C G T T C A T G A G C T G G T A T C A G G A T A A
        - : - - - - | - - - - : - - - - | - - - - : - - - - | - - - - : - - - - | - - - - : - - - - | - - - - : - - - - | - - -
2111528 G G C G A C A C A G C T T G G G C G C A T A C G G G C C T C C T C A C G C A A G T A C T C G A C C A T A G T C C T A T T
GC      46 44 44 42 42 44 44 44 42 44 44 44 44 46 46 46 44 44 44 44 46 48 50 48 48 46 44 42 40 42 42 40 42 42 42 42 42 42 42 40 40 42 40 42 40 38 40 38 36 38 38 40 40 42 42 40 42 42 44
F6      E   A   T   D   F   G   R   I   G   P   P   T   R   E   H   A   P   I   L   I
F5        R   Q   T   S   G   A   Y   A   R   L   L   A   N   M   L   Q   Y   *   S   L
F4        G   S   H   R   V   R   T   H   G   S   S   H   T   *   S   S   T   D   P   Y
```

On forward DNA strand:

GREEN is flanking DNA (100 bp).

BLACK is DNA sequence of original ORF.

*** = STOP CODON**

```
F1      L   V   Q   G   L   R   R   Q   R   Q   G   R   C   G   V   N   I   D   A   F
F2        S   Y   K   A   C   A   V   N   V   K   G   G   A   E   *   I   L   T   R   F
F3          R   T   R   L   A   P   S   T   S   R   E   V   R   S   E   Y   *   R   V   S
GC      60 60 58 58 58 58 56 56 56 56 56 58 56 56 58 60 62 62 62 62 62 62 60 58 58 58 58 58 58 58 58 60 58 58 58 56 58 60 60 62 60 62 62 64 62 60 60 58 56 54 54 54 54 54 56 58 56 56
```

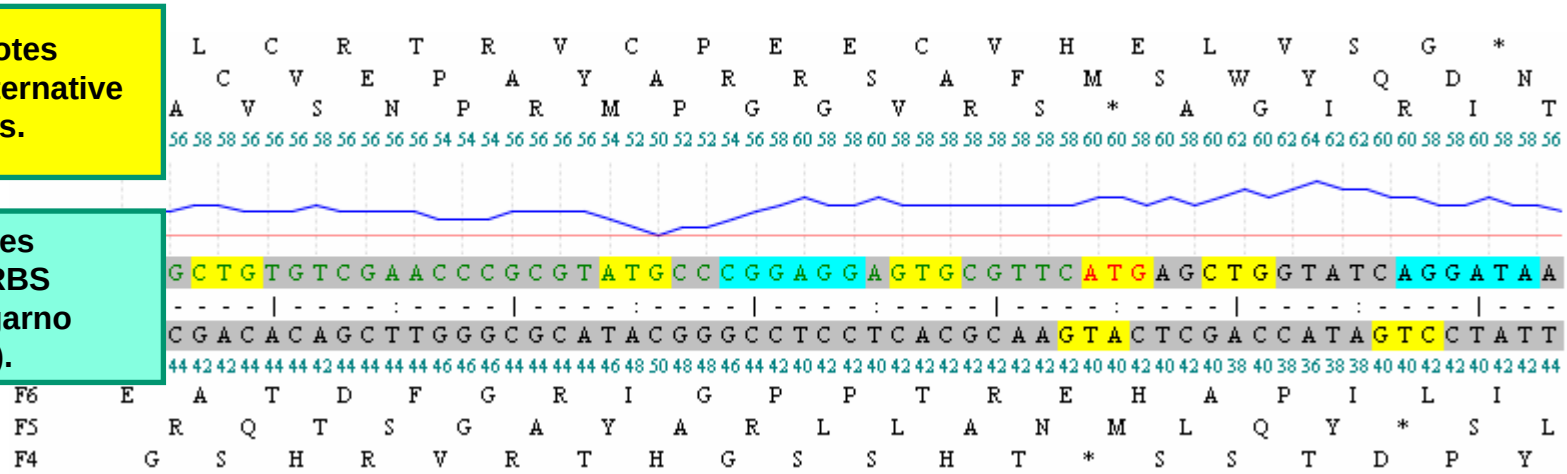
```
2113748 C T C G T A C A A G G C T T G C G C C G T C A A C G T C A A G G G A G G T G C G G A G T G A A T A T T G A C G C G T T T
        - : - - - - | - - - - : - - - - | - - - - : - - - - | - - - - : - - - - | - - - - : - - - - | - - - - : - - - - | - - -
2113748 G A G C A T G T T C C G A A C G C G G C A G T T G C A G T T C C C T C C A C G C C T C A C T T A T A A C T G C G C A A A
GC      40 40 42 42 42 42 44 44 44 44 44 44 44 42 44 44 42 40 38 38 38 38 38 38 40 42 42 42 42 42 42 42 42
F6      G   R   V   L   S   A   G   D   V   D   L
F5        E   Y   L   A   Q   A   T   L   T   L
F4        R   T   C   P   K   R   R   *   R   *   P
```

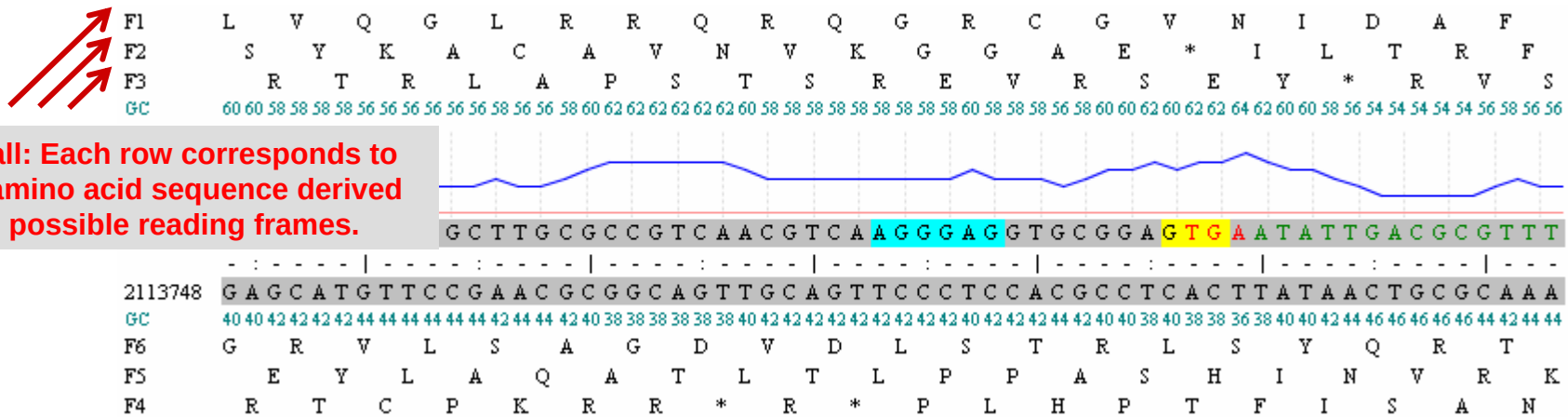
START and STOP CODONS determined by Gene Caller for the your gene are in red font (or located at green/black color font transition in DNA strand). Note they are both on the same DNA strand & polypeptide sequence.

More on… Deciphering Sequence Viewer in Graphics Mode

Yellow denotes possible alternative start codons.

Cyan denotes candidate RBS (Shine-Dalgarno Sequences).





Recall: Each row corresponds to the amino acid sequence derived from possible reading frames.

RBS = Ribosome Binding Site = Shine-Dalgarno Sequence

- <u>Always</u> on SAME DNA STRAND AS THE START CODON

Remember: To initiate protein synthesis (translation), the ribosome interacts with the Shine-Dalgarno sequence in the mRNA immediately upstream of the proper start codon

Insert WebLogo of *E. coli* ribosome binding sites and start codons from http://molbiol-tools.ca/Motifs.htm

Reference: **Figure 1** in Schneider and Stephens (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids. Research* **18**: 6097-6100.

Decision Tree

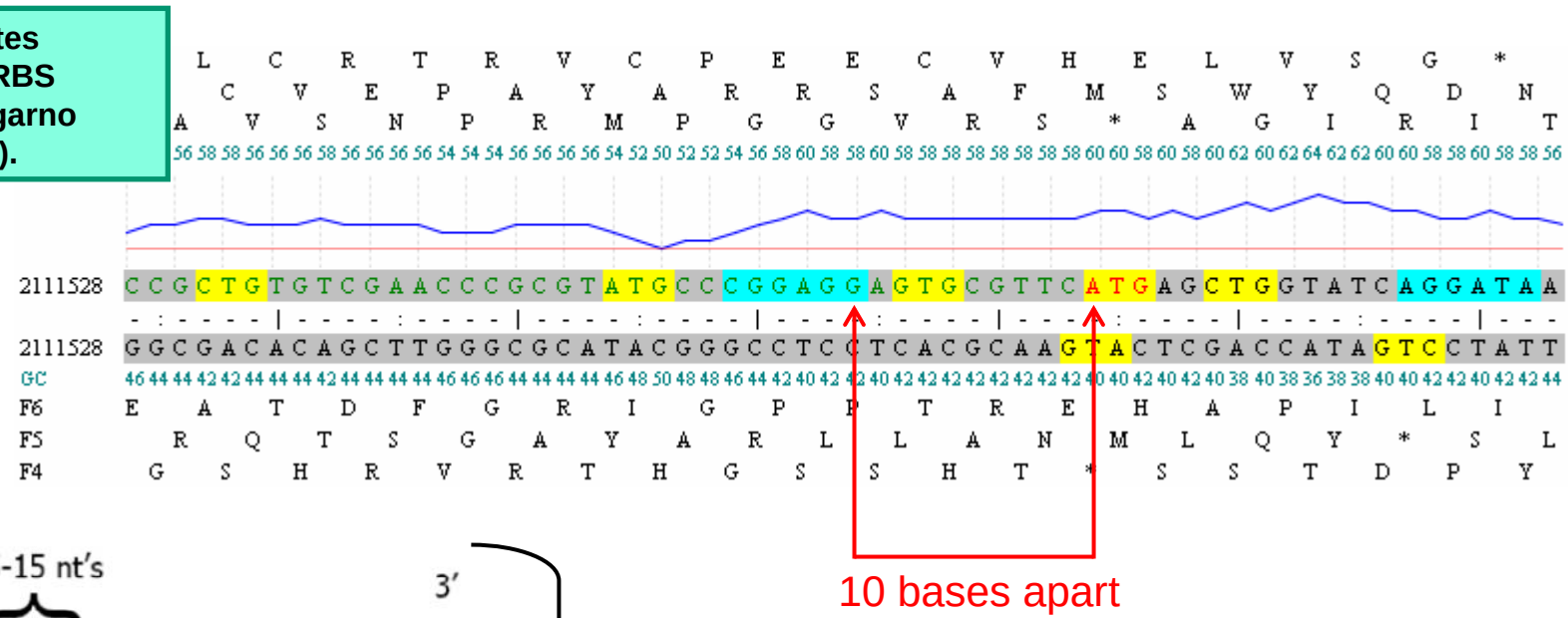
- The following slides contain a Decision Tree to aid in decision making as you proceed through the first task of this module.
- All of the steps provided in the tree will be elaborated upon throughout the presentation.
- The tree should act as an outline to help you map out your plan of attack!

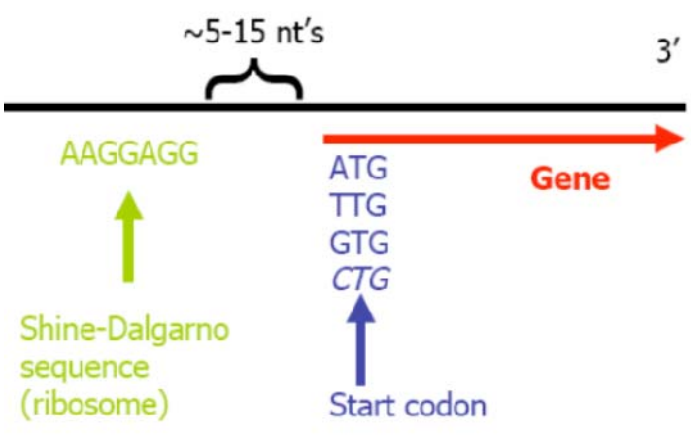
Find your original start codon.

Highlighted Yellow and Red Text.

Does it have a Shine Dalgarno Sequence 5-15bp upstream?

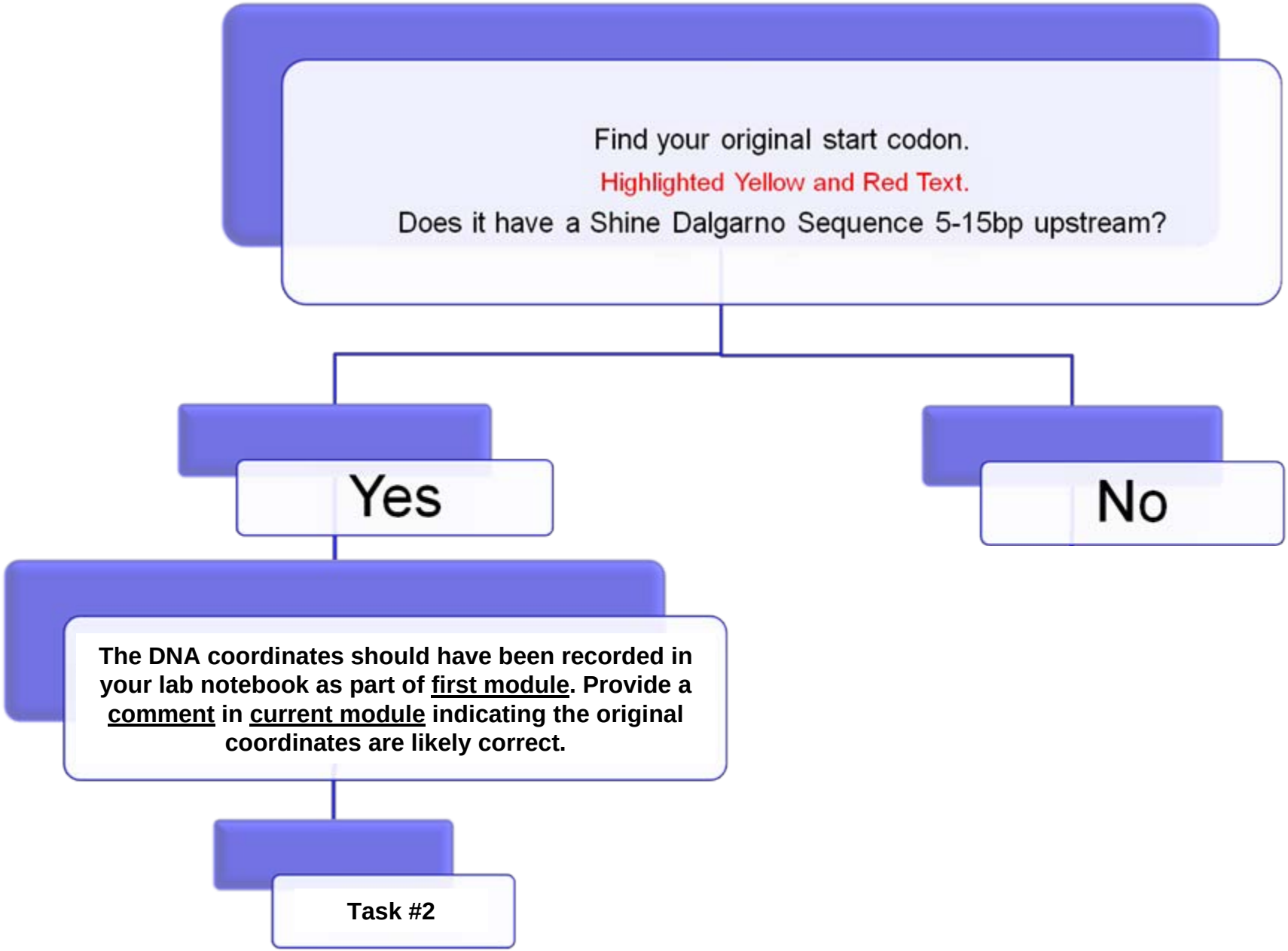
Cyan denotes candidate RBS (Shine-Dalgarno Sequences).





Recall: Start codons should occur about 5-15 bases downstream of the Shine-Dalgarno sequence (RBS)

OID 2500607069


Find your original start codon.
Highlighted Yellow and Red Text.
Does it have a Shine Dalgarno Sequence 5-15bp upstream?
Yes
No
The DNA coordinates should have been recorded in your lab notebook as part of first module. Provide a comment in current module indicating the original coordinates are likely correct.
Task #2

Recording results in your Lab Notebook

Scroll down

Alternative Open Reading Frame Module

- Module Instructions

go to the IMG Gene Details page for the proposed gene

Proposed DNA coordinates

```
enter in lab report
```

Explanation of choice

explanation

Recording results in your Lab Notebook

Alternative Open Reading Frame Module

- Module Instructions

go to the IMG Gene Details page for the proposed gene

Proposed DNA coordinates

```
No change from original DNA coordinates recorded for first module:

3948986..3951211 (+)
```

1- Enter original DNA coordinates with an explanatory comment

Explanation of choice

The original start codon for the ORF identified by automatic Gene Caller has Shine-Dalgarno sequence (CGGAGG) 10 bases upstream of the start codon (AUG).

Find your original start codon.

Highlighted Yellow and Red Text.

Does it have a Shine Dalgarno Sequence 5-15bp upstream?

Yes

The DNA coordinates should have been recorded in your lab notebook as part of <u>first module</u>. Provide a <u>comment</u> in <u>current module</u> indicating the original coordinates are likely correct.

Task #2

No

In the same reading frame as the original start codon, is there an <u>*alternative*</u> start codon upstream of the original proposed start codon?

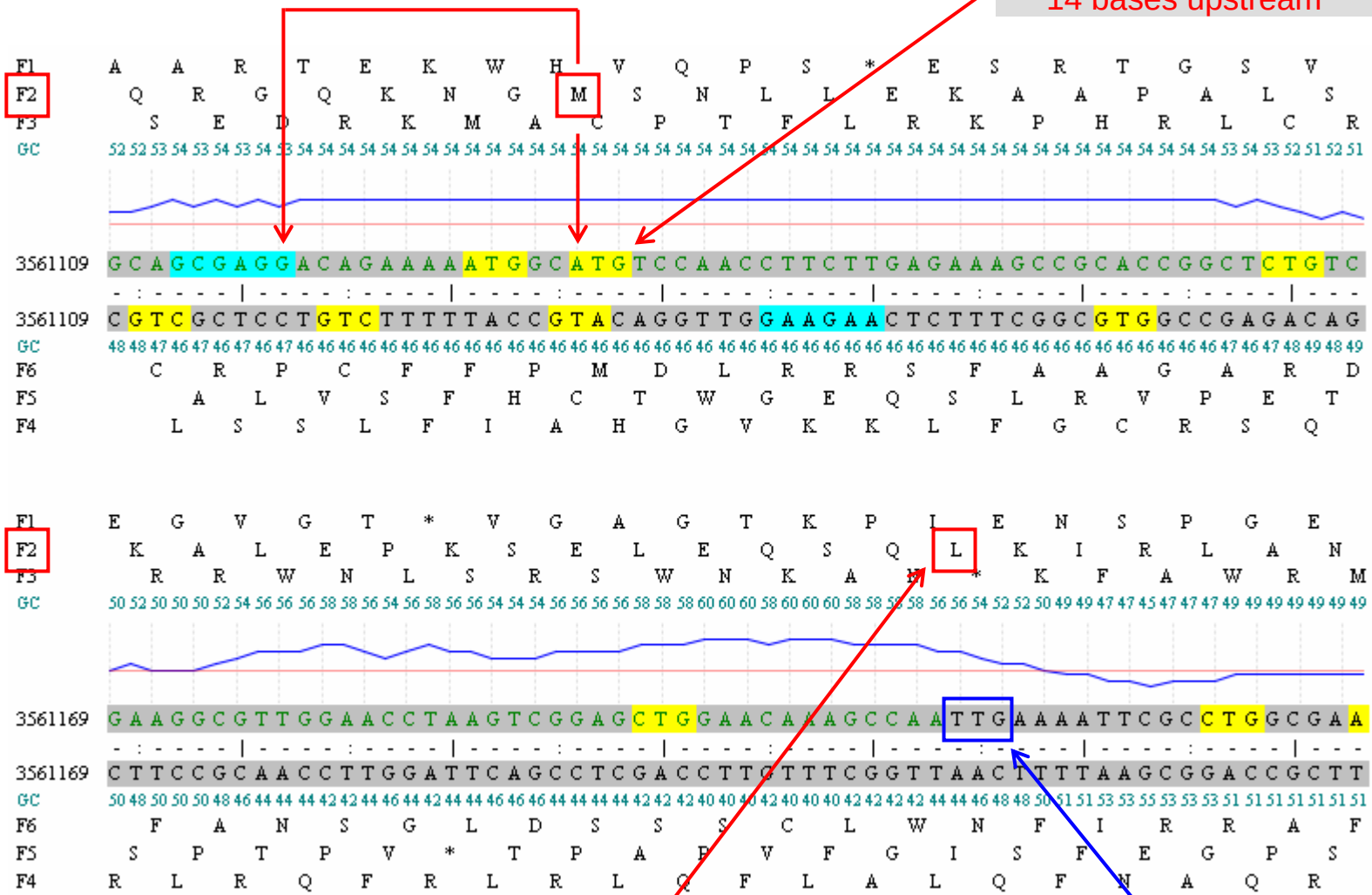
****Only inspect up to 100 bases upstream of the original start codon**

Continue to the next slide

Note there are NO stop codons between the original & proposed alternative start codon in F2

Possible alternative start codon in F2 with Shine-Dalgarno sequence 14 bases upstream

Note Met is in same frame as original start codon

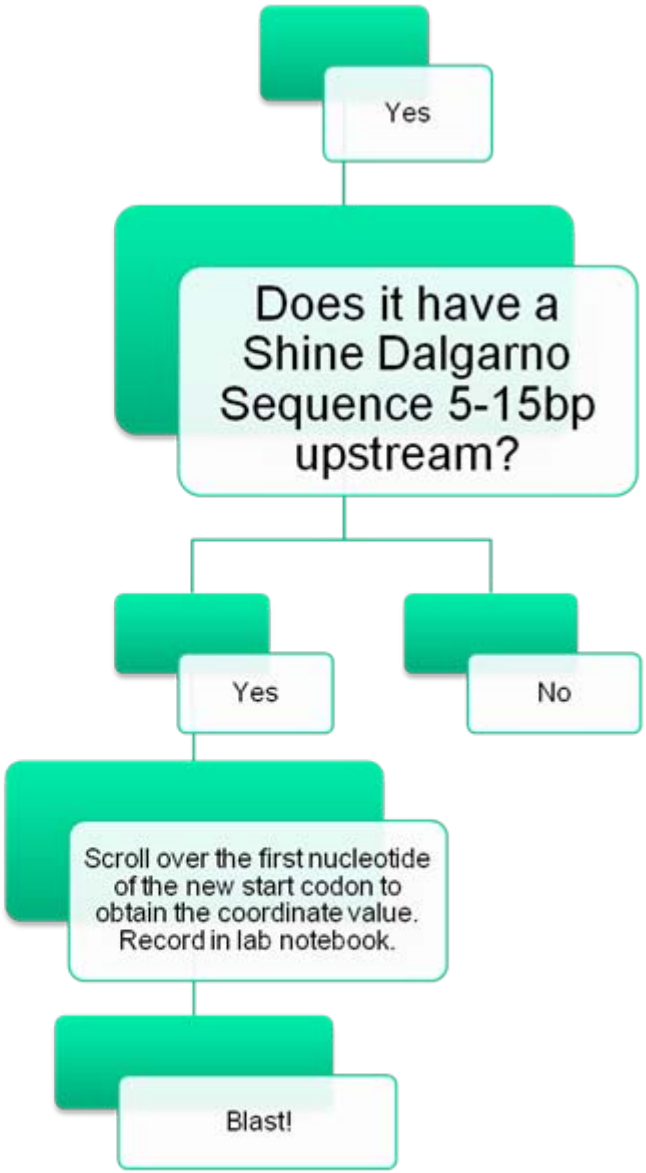


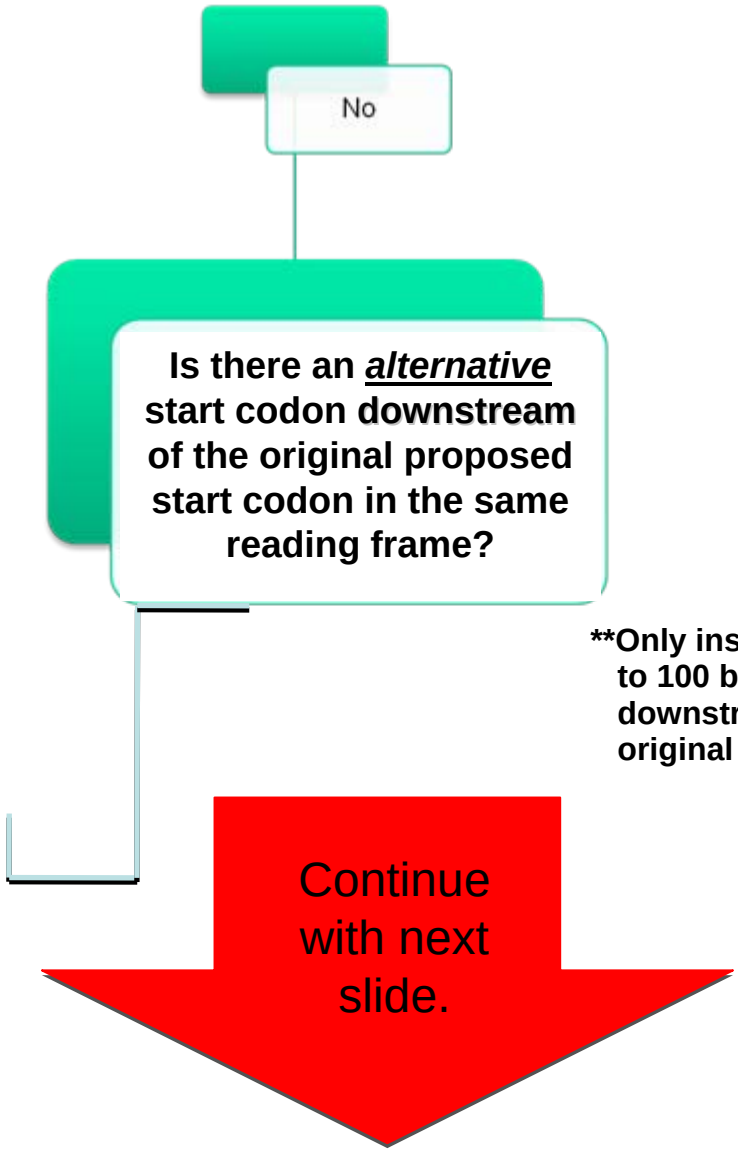
Note Leu is in reading frame 2

Original start codon with NO Shine-Dalgarno sequence upstream

<u>Not</u> highlighted red; Must examine green/black transition in DNA sequence

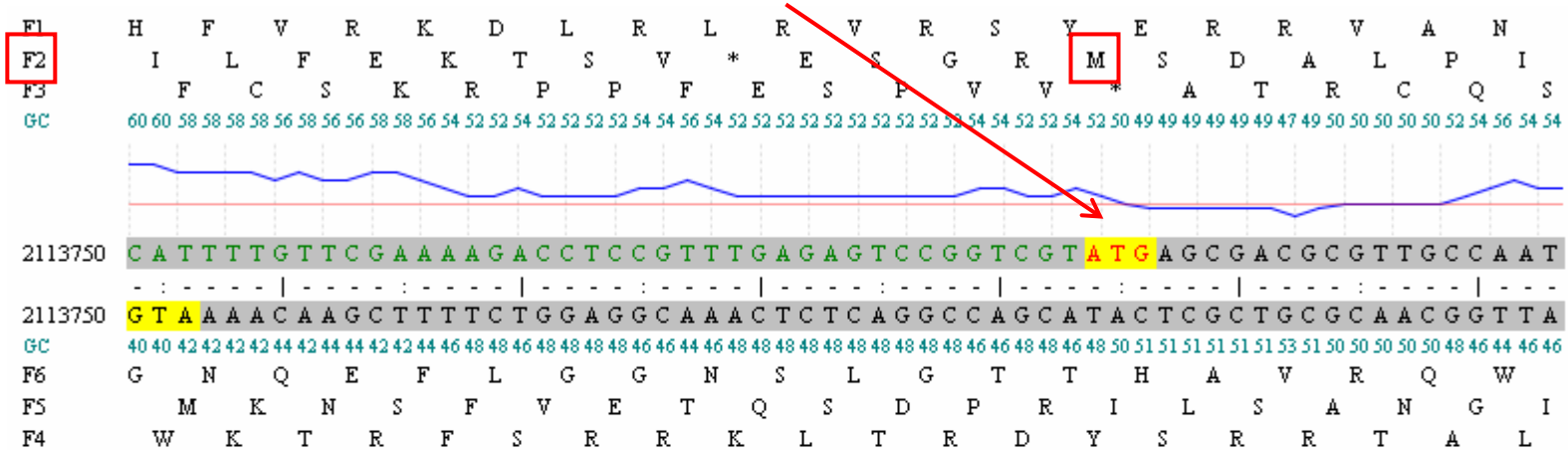
OID 2500608365



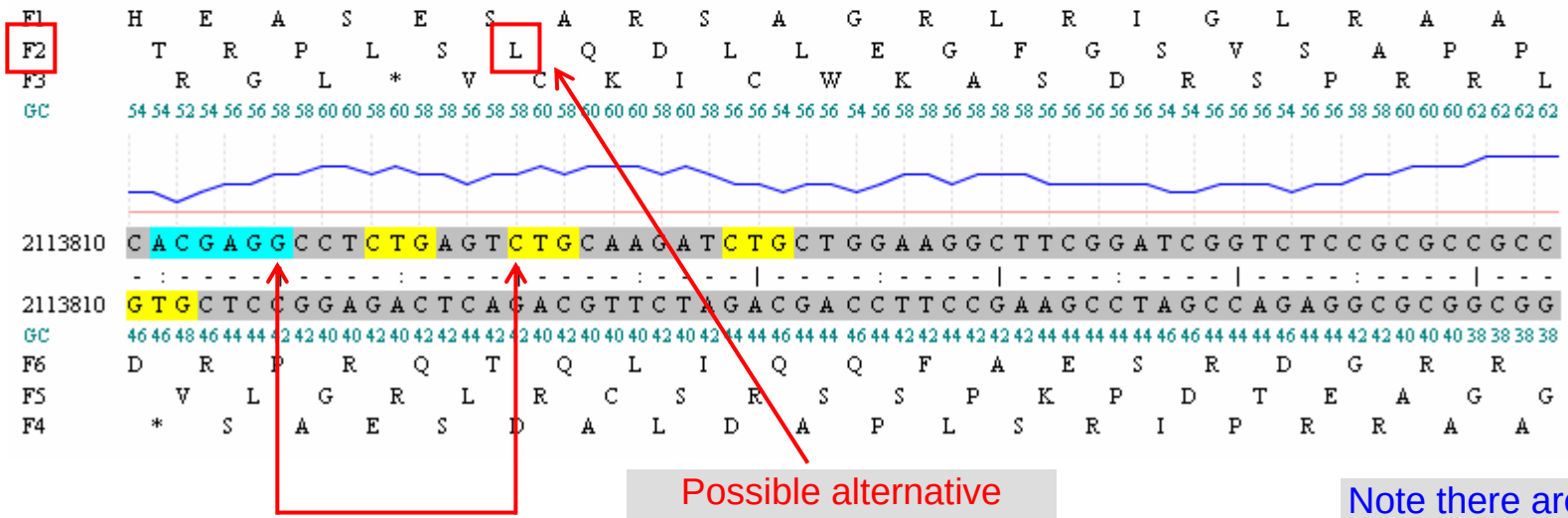


**Only inspect up to 100 bases downstream of the original start codon

Original start codon with NO Shine-Dalgarno sequence upstream



Note Met is in reading frame 2

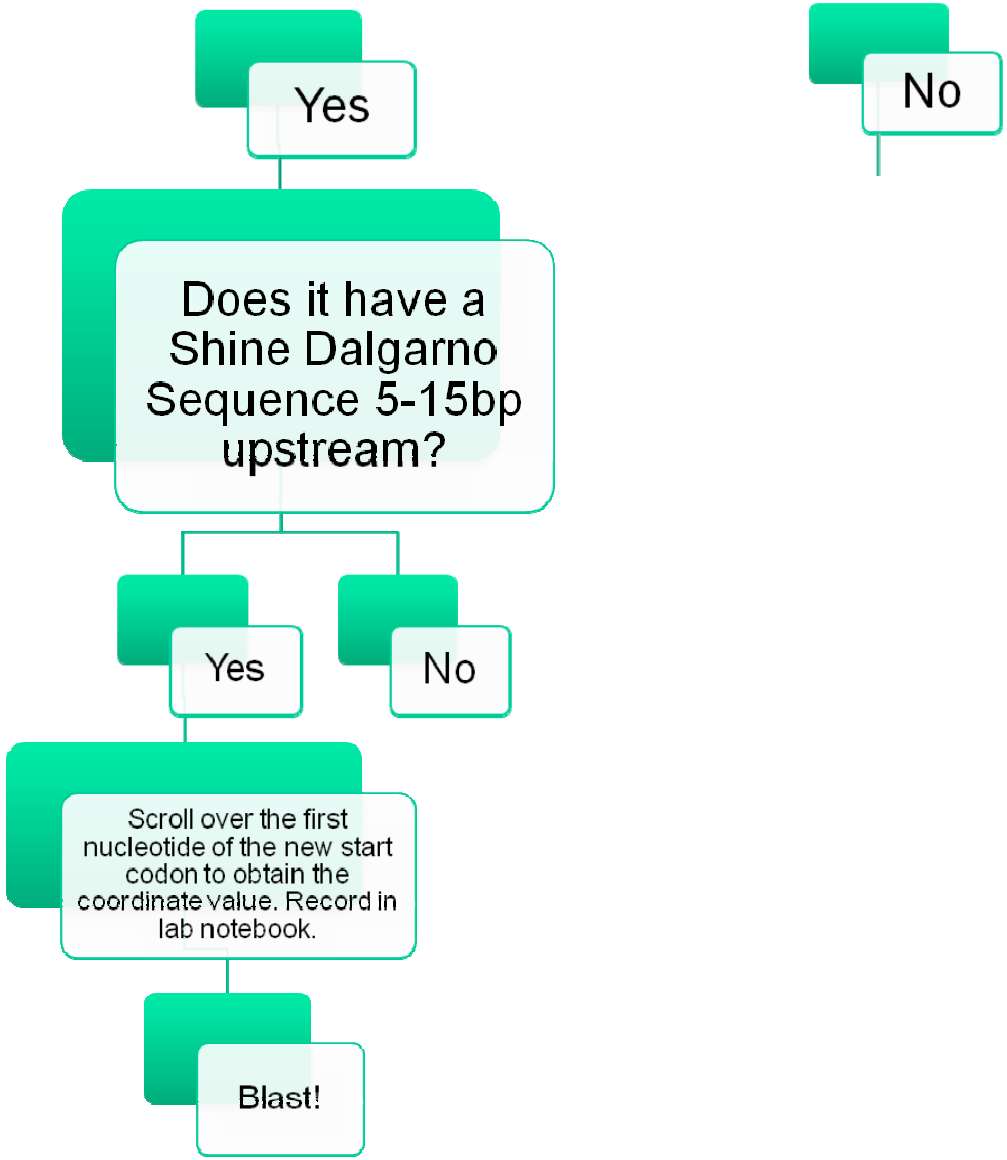


Note Leu is in same frame as original start codon

Possible alternative start codon in F2 with Shine-Dalgarno sequence 10 bases upstream

Note there are NO stop codons between the original & proposed alternative start codon in F2

OID 2500607070


Yes
No
Does it have a Shine Dalgarno Sequence 5-15bp upstream?
Yes
No
Scroll over the first nucleotide of the new start codon to obtain the coordinate value. Record in lab notebook.
Blast!

For genes with possible alternative start codon…It's time to BLAST!

- **BLAST your results:**
  - Construct a "revised" protein sequence in FASTA format (add or subtract amino acid residues in proper reading frame to reflect new start codon position then copy/paste into lab notebook).
  - Submit as query for a BLAST search of NCBI database (Genbank, SwisProt).
- **Your results from BLAST:**
  - Compare results from original blast search with those from new blast search.
    - Determine if statistics have improved.
      - REMEMBER: higher bit score, lower e-value, higher % identity and/or longer alignment length are all good arguments that the alternative start codon is a better choice

Create new headings & boxes for entering amino acid sequence in FASTA format for ORF with alternate start codon

Alternative Open Reading Frame Module

- Module Instructions

1- Add headings and box in lab notebook

go to the IMG Gene Details page for the proposed gene

Protein Sequence for ORF with Alternate Start Codon

Construct this from Sequence Viewer on the Gene Details page for the proposed gene.

enter AA FASTA format

2- Copy/paste modified protein sequence into box

```
>OID 2500608365 Flp pilus assembly protein TadC with alternative start codon
MSNLLEKAAPALSKALEPKSELEQSQLKIRLANAGFHSPQAPMIYLAIKTVCLVVGLVLGGGLGMYRYGTTQAGLT
TLIIAAGAGFYLPEGVLAYLISKRKQAIFLQLPDVLDLLVVCVEAGLGLDAGLRRVAEELKDTAPEICGELAMCNL
QLQMGRNRRDMLHDLGVRTGVDDVKALVAIMIQADKFGSSIAQALRVQSDSMRVKRRQIAEEKAQKTAVQMLFPMV
IFIFPGIFVVLVGPAAIKMMDQLLNKP
```

**NOTE highlighted portion was added to the original amino acid sequence since the alternative start codon (AUG) is located 26 residues upstream of the original start codon (UUG, which encodes Leucine)

OID 2500608365

NCBI BLAST can be accessed from lab notebook link

Sequence-based Similarity Data Module

- Module Instructions

BLAST

go to http://www.ncbi.nlm.nih.gov/blast

BLAST Results for ORF with original start codon (Gene Caller)

```
>ref|NP_866981.1| G secretion system protein TadC [Rhodopirellula baltica SH 1]
 emb|CAD74523.1| G probable secretion system protein TadC [Rhodopirellula baltica
SH 1]
Length=328

 GENE ID: 1791454 tadC | secretion system protein TadC
[Rhodopirellula baltica SH 1] (10 or fewer PubMed links)

 Score =  263 bits (671),  Expect = 1e-68, Method: Compositional matrix a
 Identities = 133/227 (58%), Positives = 168/227 (74%), Gaps = 0/227 (0%)

Query  1    LKIRLANAGFHSPQAPMIYLAIKTVCLVVGLVLGGGLGMYRYGTTQAGLTTLIIAAGAGF
            L+ +L NAGF    AP+I+  I+ VC  VGL+LGG  G    G TQ  +  L++    GF
Sbjct  102  LREKLINAGFRRESAPVIFKLIQLVCTGVGLMLGGVTGAVLDGLTQGMIIKLLLGMIGGF
```

- ✓ Compare hits from <u>same</u> organism
- ✓ Are statistics improved for proposed ORF?
  - ➢ If <u>yes</u>, enter data into notebook

BLAST Results for ORF with alternative start codon proposed by student!

```
>ref|NP_866981.1| G secretion system protein TadC [Rhodopirellula baltica SH 1]
 emb|CAD74523.1| G probable secretion system protein TadC [Rhodopirellula baltica
SH 1]
Length=328

 GENE ID: 1791454 tadC | secretion system protein TadC
[Rhodopirellula baltica SH 1] (10 or fewer PubMed links)

 Score =  271 bits (694),  Expect = 3e-71, Method: Compositional matrix adjust.
 Identities = 140/253 (55%), Positives = 182/253 (71%), Gaps = 1/253 (0%)

Query  1    MSNLLEKAAPALSKALEPKSELEQSQLKIRLANAGFHSPQAPMIYLAIKTVCLVVGLVLG  60
            ++  LEKAA  ++ ++    E E  +L+ +L NAGF    AP+I+  I+ VC  VGL+LG
Sbjct  77   LTEALEKAATPIASSVSGNEE-EMGKLREKLINAGFRRESAPVIFKLIQLVCTGVGLMLG  135
```

OID 2500608365

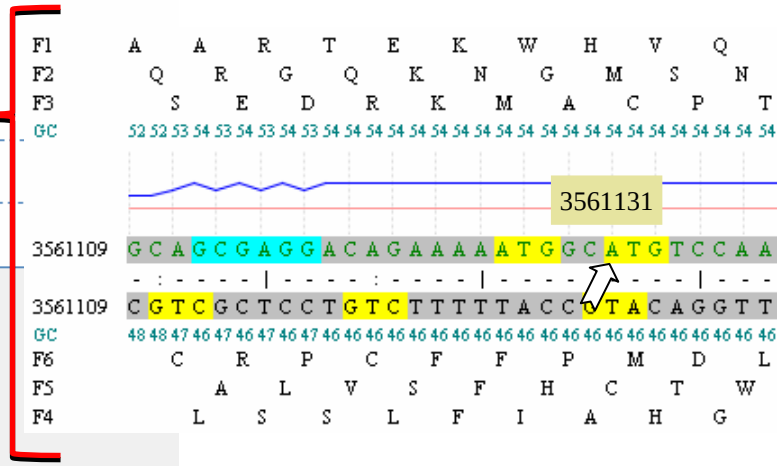
Recording results in your Lab Notebook

Alternate Start Codon for Original Open Reading Frame

Find this on the Gene Details page for the proposed gene.
enter coordinates
enter proposed DNA coordinates

```
3561131..3561898 (+) (765bp)
```

Adjust DNA coordinates & base pairs.



Reasoning

No Shine-Dalgarno sequence was found by automatic Gene Caller immediately upstream of original start codon. However, located upstream of the original start codon there is a Shine-Dalgarno sequence 14 bases upstream of an alternative start codon (ATG). Furthermore, the NCBI protein BLAST statistics for the ORF using the alternative start codon improved in comparison to the search with the ORF having the original start codon. In particular, the score was higher (271 bits compared to 263 bits) and the E-value was lower (3e-71 compared to 1e-68) for the ORF with an alternative start codon. However, the percent identity was lower for the alternative compared to the original (55% compared to 58%). Also of note is that the length of the alignment did not change (328 amino acids for both). Taken together, I believe there is sufficient evidence to suggest that the alternative position for the start codon for this gene is a better choice than the original position. My conclusion is based on three results: (1) presence of Shine-Dalgarno sequence at appropriate distance from the start codon, (2) higher bit score and lower E-value for BLAST search using alternative ORF.

1- Change section heading as shown

BLAST Results

Gene product name (same hit as in Module 2)

```
secretion system protein TadC
```

Organism

```
Rhodopirellula baltica SH 1
```

Length, E-Value, Score, Percent identity, Positives, and Gaps

```
Length=328,  Score =  271 bits (694),  Expect = 3e-71, Method: Compositional matrix adj
 Identities = 140/253 (55%), Positives = 182/253 (71%), Gaps = 1/253 (0%)
```

Alignment of the BLAST hit and the query sequence

```
Query  1    MSNLLEKAAPALSKALEPKSELEQSQLKIRLANAGFHSPQAPMIYLAIKTVCLVVGLVLG  60
            ++  LEKAA  ++ ++    E E  +L+ +L NAGF    AP+I+  I+ VC  VGL+LG
Sbjct  77   LTEALEKAATPIASSVSGNEE-EMGKLREKLINAGFRRESAPVIFKLIQLVCTGVGLMLG  135
```

2- Copy/paste sub-headings & field boxes from Sequence-based Similarity Data module.

3- Fill in with your results from BLAST search using ORF with alternate start codon for gene.

OID 2500608365

Are the BLAST results always better? NO!

For example. . .

Alternative Open Reading Frame Module

- Module Instructions

go to the IMG Gene Details page for the proposed gene

Protein Sequence for ORF with Alternate Start Codon

Construct this from Sequence Viewer on the Gene Details page for the proposed gene.

enter AA FASTA format

```
>OID 2500607070 Fe-S-cluster-containing hydrogenase components_alternative start codon
LQDLLEGFGSVSAPPSKGEQMVHRLLQKQRELSAVELFAKEHAGGTINGS
KLPDQARFYASLLPATPVGPGQQYGFEVDLDRCSGCKACVTACHSLNGLD
DSETWRDVGLLIGGTETLPVMQHVTAACHHCLEPACMTACPVNAYEKDAF
TGIVRHLDDQCFGCQYCTLACPYNVPKYHAAKGIVRKCDMCSNRLKNGEA
PACVQACPHEAISIRIVDVSRVTENAEADHFLPAAPEPYITLPTTTYRTT
RVFPRNMLPADYYSVSPQHPHWPLIVMLVLTQLSVGAFAAGSFLEEALDT
ELATAFRPIHATGALVLGLLALGASTLHLGRPLYAFRGILGFRHSWLSRE
IVAFGLFAGLAVPFAGLCWGLPLLEVSGSPWGKLAGELLPTLSPSVACVG
VIGVFCSVMIYVFTRRELWSLERTLIRFSLTTILLGVATIWLMMMLAVGF
LSDDEWHQLAQNLTRPLARSVIILTTLKLLYDISLLRHLATFRNSPLKRS
ALLVVGPLRGFSIGRLVLGVVGGVVIPAAFAAVPIHDPIQFTTTSAVFVG
QMWVACLGGELLERYLFFSAVSNPRMPGGVRS
```

**Recall that this gene had residues deleted from the original amino acid sequence since the alternative start codon (CUG, which encodes Leucine) is located 12 residues downstream of the original start codon (AUG)

OID 2500607070

BLAST Results for ORF with original start codon (Gene Caller)

```
>ref|ZP_01093385.1| molybdopterin oxidoreductase, iron sulfur subunit [Blastopirellula
marina DSM 3645]
 gb|EAQ77915.1| molybdopterin oxidoreductase, iron sulfur subunit [Blastopirellula
marina DSM 3645]
Length=536

 Score =  374 bits (960), Expect = 2e-101, Method: Compositional matrix
 Identities = 235/574 (40%), Positives = 304/574 (52%), Gaps = 61/574 (1

Query  28   SKGEQMVHRLLQKQRELSAVELFAKEHAGGTINGSKLPDQARFYASLLPATPVGPGQQYG
            + G  ++  LL +Q++L+AVE F++ HA  T      P  A  Y  LLPA     G+QY
Sbjct  13   TDGFDLIQSLLAEQQDLTAVERFSQRHADATT-----PLMAPLYRELLPAAAPSAGEQYA
```

- ✓ Compare hits from <u>same</u> organism
- ✓ Are statistics improved for proposed ORF?
  - ➢ If <u>no</u> or <u>not really</u>, enter data into notebook anyway with a comment

BLAST Results for ORF with alternative start codon proposed by student!

```
>ref|ZP_01093385.1| molybdopterin oxidoreductase, iron sulfur subunit [Blastopirellula
marina DSM 3645]
 gb|EAQ77915.1| molybdopterin oxidoreductase, iron sulfur subunit [Blastopirellula
marina DSM 3645]
Length=536

 Score =  374 bits (960), Expect = 1e-101, Method: Compositional matrix adjust.
 Identities = 235/574 (40%), Positives = 304/574 (52%), Gaps = 61/574 (10%)

Query  16   SKGEQMVHRLLQKQRELSAVELFAKEHAGGTINGSKLPDQARFYASLLPATPVGPGQQYG  75
            + G  ++  LL +Q++L+AVE F++ HA  T      P  A  Y  LLPA     G+QY
Sbjct  13   TDGFDLIQSLLAEQQDLTAVERFSQRHADATT-----PLMAPLYRELLPAAAPSAGEQYA  67
```

OID 2500607070

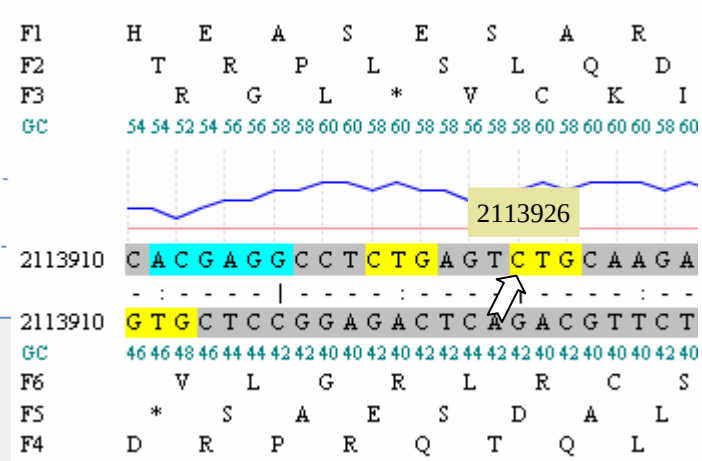
Recording results in your Lab Notebook

Alternate Start Codon for Original Open Reading Frame

Find this on the Gene Details page for the proposed gene.
enter coordinates
enter proposed DNA coordinates

```
2113926...2115574 (-)(1746bp)
```

Adjust DNA coordinates & base pairs.



1- Change section heading as shown

Reasoning

No Shine-Dalgarno sequence was found by automatic Gene Caller immediately upstream of original start codon. However, located downstream of the original start codon there is a Shine-Dalgarno sequence 10 bases upstream of an alternative start codon (CTG). Although the NCBI protein BLAST statistics do not markedly improve for the ORF using the alternative start codon, they also are not significantly different (or worse). Thus, I cannot rule out the possibility that the proposed alternative start codon is a better choice than the original start codon.

BLAST Results

Gene product name (same hit as in Module 2)

```
molybdopterin oxidoreductase, iron sulfur subunit
```

Organism

```
Blastopirellula marina DSM 3645
```

Length, E-Value, Score, Percent identity, Positives, and Gaps

```
Length=536,  Score =  374 bits (960),  Expect = 2e-101, Method: Compositional matrix adjust.
 Identities = 235/574 (40%), Positives = 304/574 (52%), Gaps = 61/574 (10%)
```

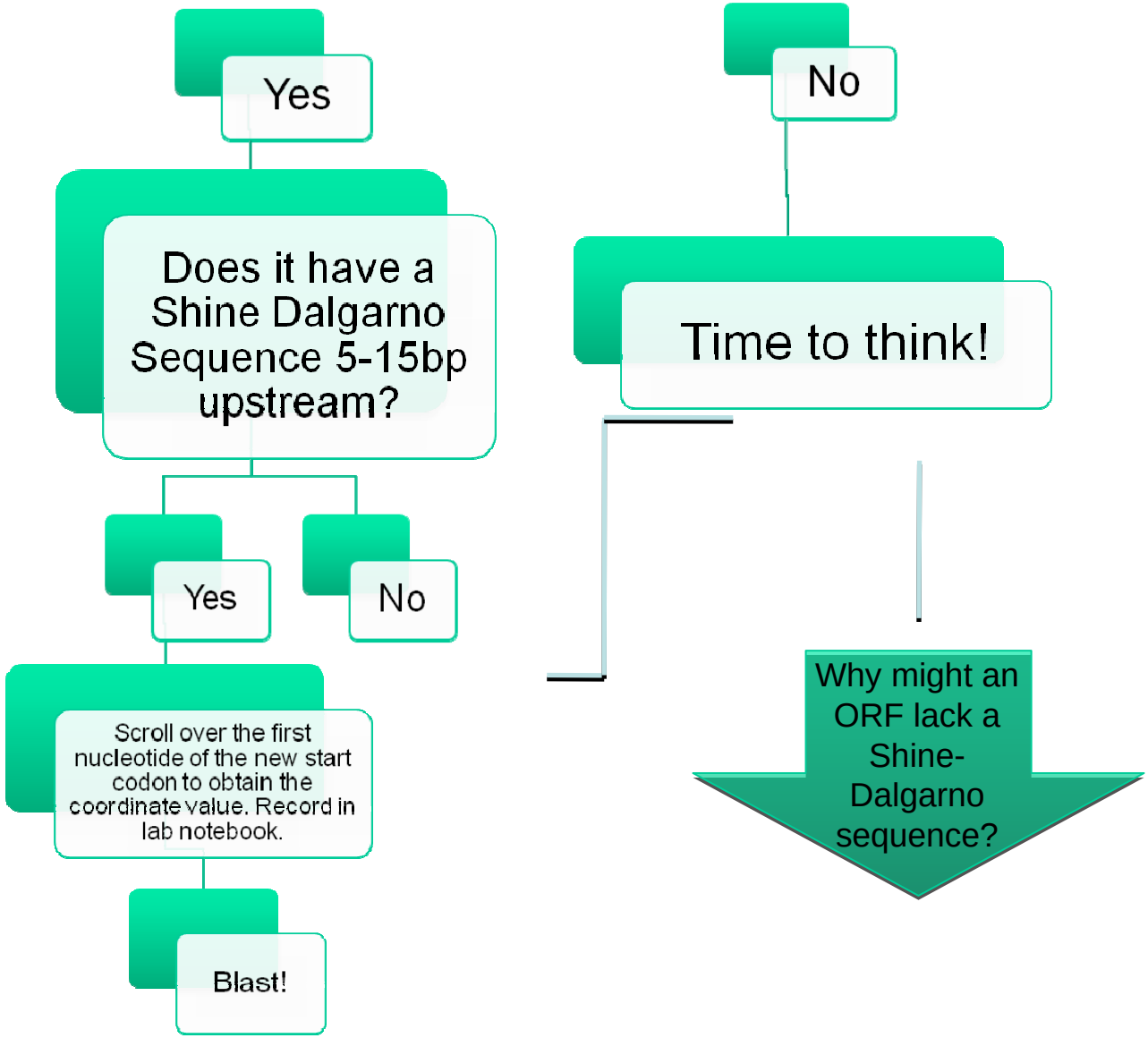
Alignment of the BLAST hit and the query sequence

```
Query  16   SKGEQMVHRLLQKQRELSAVELFAKEHAGGTINGSKLPDQARFYASLLPATPVGPGQQYG  75
            + G  ++  LL +Q++L+AVE F++ HA  T      P  A  Y  LLPA     G+QY
Sbjct  13   TDGFDLIQSLLAEQQDLTAVERFSQRHADATT-----PLMAPLYRELLPAAAPSAGEQYA  67
```

2- Copy/paste sub-headings & field boxes from Sequence-based Similarity Data module.

3- Fill in with your results from BLAST search using ORF with alternate start codon for gene.

OID 2500607070


Yes
No
Does it have a Shine Dalgarno Sequence 5-15bp upstream?
Time to think!
Yes
No
Scroll over the first nucleotide of the new start codon to obtain the coordinate value. Record in lab notebook.
Blast!
Why might an ORF lack a Shine-Dalgarno sequence?

WHY? WHY? WHY NOT?!

Interpreting Your Negative Results

(i.e., no Shine-Dalgarno, no alternative start codon, etc.)

- Maybe there is flexibility in the amount of sequence conservation needed in the Shine-Dalgarno (S-D) that allows ribosome binding

Insert **Figure 33.9** from Garret & Grisham *Biochemistry* (2nd Ed.)
Alignment of various Shine-Dalgarno sequences recognized by
E. coli ribosomes.

WHY? WHY? WHY NOT?!

Interpreting Your Negative Results

(i.e., no Shine-Dalgarno, no alternative start codon, etc.)

- Consider possible mutations in DNA sequence
  - Remember “draft genome” problems?

- Consider a ribosome “skid” if your gene is part of an operon
  - Maybe your gene is in the middle or at the end of an operon. And perhaps only the first gene in the operon has a S-D upstream of the start codon and has a stop codon within 5-15 nt of the start codon for the next gene in the operon. In this case, the genes may be close enough that the ribosome does not need to completely dissociate from the RNA transcript, instead it “skids” along and begins translation of the next gene in the operon without needing to bind a second S-D.
    - Look at the ortholog neighborhood map!
      (Review Gene Context in HGT module)

Recording results in your Lab Notebook

Alternate Start Codon for Original Open Reading Frame

1- Change the heading

Find this on the Gene Details page for the proposed gene.
enter coordinates
enter proposed DNA coordinates

```
No change from original DNA coordinates recorded for first module:

2111568..2113793 (-)(2226bp)
```

2- Enter original DNA coordinates with an explanation

Reasoning

Although the ORF identified by automatic Gene Caller does not have an obvious Shine-Dalgarno sequence upstream of the start codon, an alternative position for the start codon also was not apparent. It is possible that the S-D sequence for this gene is not well conserved or that there are point mutations in the DNA sequence that made the S-D unrecognizable (e.g., draft genome problem). Notably, upon inspection of the ortholog neighborhood, it is possible that this gene is part of an operon. As shown below, the gene (light green) preceding my gene (red) may be close enough to be transcribed together as an operon. Consistent with this hypothesis, the green gene (which encodes a Fe-S-cluster-containing hydrogenase component as ascertained by the automated Gene Caller) does indeed have a start codon preceded by a Shine-Dalgarno sequence (provided the alternative position is correct). The two genes also may have a functional relationship (e.g., hydrogenase vs. dehydrogenase) as would be expected if part of an operon. If my gene is indeed part of an operon, then it's possible that the initiation of translation occurs from a ribosome "skid" after completion of translation of the preceding gene.

Planctomyces limnophilus DSM 3776 : PlimDRAFT_4083246_C168

NOW for the second task:

Hunting for Potential
Novel Open Reading Frames (ORFs)

Return to the Gene Detail Page

img/edu INTEGRATED MICROBIAL GENOMES EDUCATION SITE

IMG Home | Find Genomes | Find Genes | Find Functions | Compare Genomes | Analysis Carts | MyIMG | Using IMG

Gene Search | BLAST | Phylogenetic Profilers

Gene Detail

Loaded.

Gene Information
Evidence For Function Predictions
Sequence Search
External Sequence Search
IMG Sequence Search
Homolog Display

Gene Information

Gene Information	
Gene Object ID	2500607069
Gene Symbol	
Locus Tag	PlimDRAFT_19450
Product Name	Uncharacterized anaerobic dehydrogenase, COG3383
IMG Product Source	COG3383
Genome	Planctomyces limnophilus DSM 3776
DNA Coordinates	2111568..2113793 (-)(2226bp)
Scaffold Source	Planctomyces limnophilus DSM 3776 : PlimDRAFT_4083246_C168 (5423025bp)
IMG ORF Type	
GC Content	0.58
External Links	
Fused Gene	No
Fusion Component	No
Protein Information	
Amino Acid Sequence Length	741aa

Scroll down

Evidence For Function Prediction

Neighborhood

red = Current Gene

green = Positional Cluster Gene in the same KEGG Pathway as the Current Gene

CRISPR array

Sequence Viewer For Alternate ORF Search

Chromosome Viewer colored by COG

Chromosome Viewer colored by GC percentage

"Click" here again

Change output from “Graphic” to “Text”

Sequence Viewer

Neighborhood six frame translation with putative ORF's shown below
for gene_oid=*2500607069*
2111568..2113793(-)
Uncharacterized anaerobic dehydrogenase, COG3383.

Select gene neighborhood:
-100 bp upstream. +100 bp downstream

Select minimum ORF size:
80 aa

Output Format:
Text Graphics

Submit Reset

Keep parameters the same as first task.

USE TEXT option to look for NOVEL open reading frames

Click submit!

Sequence Viewer

Loaded.

Neighborhood six frame translation with putative ORF's shown below
for gene_oid=*2500607069*
2111568..2113793(-)
Uncharacterized anaerobic dehydrogenase, COG3383.

NOTE: Usually the result with the longest amino acid sequence length is your original ORF.

hint: *Copy and paste sequence to BLAST and InterPro scan to test ORF translation.*

```
>2500607069_1_ORF2  Translation of 2500607069 in frame 1, ORF 2, threshold 80, 102aa
ERKAGTCRLGCSAETFHGPLSTDSGKARAPLSGVHQHRPDAHRRDGIARRGGQVRHGPQA
RRRQYAAVHGDGSGRLQRGVRFRCASLHLRRLRRIGRHGVRG

>2500607069_3_ORF3  Translation of 2500607069 in frame 3, ORF 3, threshold 80, 140aa
AWMRVSSPPRTVGRTTKSWREFSKRKSKDCGSSVPTRPTRGSTRTWPVKCFRGSISWLCR
TCTTIRKRPRWLTSSFQPLAGARKRGRLLIPSGALVVSRRSGVRRGKRWPTSTSFNWLLS
IGAAESNSGAGLRPKPSLKF

>2500607069_4_ORF3  Translation of 2500607069 in frame 4, ORF 3, threshold 80, 123aa
KMWKSASAFPGARLTFLIRPMRRSELINVPSFSPQPAAGRTRSAIWAVSVLWYMSCTTRK
SSRESISRARFWLIHEWAVLVQMIHSPLIFSLRIPSMISWYDQLFSVGMTLSSMLRMPAI
FAR

>2500607069_5_ORF2  Translation of 2500607069 in frame 5, ORF 2, threshold 80, 426aa
LVRPLVMHRNEDLSLSHSRSHERHRPHAASSRLDGDPGVMGDLECSCISRIDFDKRSLGI
QLPEDRRLRSAGLRMPLRSASPSSQQNERIRLVGQFRQRAGGLKEEFCTTVGMIKTTILE
ETPLLYGGGHPFRAGPLHPPLFFDPAVVLNAREVAGHTARTLLQNFKDGFGRRPAPELLS
AAPILSNQLKDVEVGQRFPRRTPDLLDTTNAPLGINKRPLFLAPASGWKDEVSHLGRFRI
VVHVLHNQEIEPRKHFTGQVLVDPRVGRVGTDDPQSFDFLFENSLHDFVVRPTVLGGDDT
LIHAQNARDLRAVIGVLEIVSAQKVGRVAEEPRTHGVALARDRIRSGSGFANVPRHQCQI
DDRLRGTHPLVALIDPHRPPERDSLFLVDRAGESLNLVRLKTRFGCSAGQAIVVHEAGEF
LESRRR
```

HINT: A NOVEL ORF could be in a different reading frame from the original ORF

We are interested in all ORFs with sequence length of at least 80 aa

Sequence Viewer

Loaded.

Neighborhood six frame translation with putative ORF's shown below
for gene_oid=*2500607069*
2111568..2113793(-)
Uncharacterized anaerobic dehydrogenase, COG3383.

---

hint: *Copy and paste sequence to BLAST and InterPro scan to test ORF translation.*

```
>2500607069_1_ORF2  Translation of 2500607069 in frame 1, ORF 2, threshold 80, 102aa
ERKAGTCRLGCSAETFHGPLSTDSGKARAPLSGVHQHRPDAHRRDGIARRGGQVRHGPQA
RRRQYAAVHGDGSGRLQRGVRFRCASLHLRRLRRIGRHGVRG

>2500607069_3_ORF3  Translation of 2500607069 in frame 3, ORF 3, threshold 80, 140aa
AWMRVSSPPRTVGRTTKSWREFSKRKSKDCGSSVPTRPTRGSTRTWPVKCFRGSISWLCR
TCTTIRKRPRWLTSSFQPLAGARKRGRLLIPSGALVVSRRSGVRRGKRWPTSTSFNWLLS
IGAAESNSGAGLRPKPSLKF

>2500607069_4_ORF3  Translation of 2500607069 in frame 4, ORF 3, threshold 80, 123aa
KMWKSASAFPGARLTFLIRPMRRSELINVPSFSPQPAAGRTRSAIWAVSVLWYMSCTTRK
SSRESISRARFWLIHEWAVLVQMIHSPLIFSLRIPSMISWYDQLFSVGMTLSSMLRMPAI
FAR
>2500607069_5_ORF2  Translation of 2500607069 in frame 5, ORF 2, threshold 80, 426aa
LVRPLVMHRNEDLSLSHSRSHERHRPHAASSRLDGDPGVMGDLECSCISRIDFDKRSLGI
QLPEDRRLRSAGLRMPLRSASPSSQQNERIRLVGQFRQRAGGLKEEFCTTVGMIKTTILE
ETPLLYGGGHPFRAGPLHPPLFFDPAVVLNAREVAGHTARTLLQNFKDGFGRRPAPELLS
AAPILSNQLKDVEVGQRFPRRTPDLLDTTNAPLGINKRPLFLAPASGWKDEVSHLGRFRI
VVHVLHNQEIEPRKHFTGQVLVDPRVGRVGTDDPQSFDFLFENSLHDFVVRPTVLGGDDT
LIHAQNARDLRAVIGVLEIVSAQKVGRVAEEPRTHGVALARDRIRSGSGFANVPRHQCQI
DDRLRGTHPLVALIDPHRPPERDSLFLVDRAGESLNLVRLKTRFGCSAGQAIVVHEAGEF
LESRRR
```

Copy/paste the complete FASTA sequence into a BLAST query box

Most translations will not produce significant hits

E-value too high!

Sequences producing significant alignments:		Score (Bits)	E Value	
emb|CAI59731.1|	homeodomain transcription factor bW2 [Sporiso...	39.3	0.70	
ref|YP_002179979.1|	HTH-type transcriptional regulator araB [...	37.7	2.2	G
ref|YP_002200969.1|	LysR-family transcriptional regulator [St...	36.2	6.8	G
ref|YP_002191839.1|	lysR-family transcriptional regulator [St...	35.8	7.8	G
ref|YP_001768978.1|	multi-sensor hybrid histidine kinase [Met...	35.4	9.5	G

. . . Or will give no hits at all

What do I enter in my lab notebook?

- ✓ Create headings and boxes as indicated
- ✓ Enter “No significant hits”

Novel ORFs

Translation of genomic region obtained from TEXT view of Sequence Viewer

BLAST Results

```
No significant hits.
```

If your BLAST search does produce a significant hit…

A potentially valid NOVEL ORF will give you a high % identity, low e-value, & high bit score.

- ✓ Create headings and boxes as indicated in your notebook
- ✓ Fill in boxes with result from BLAST search using sequence for NOVEL ORF

Novel ORFs

Translation of genomic region obtained from TEXT view of Sequence Viewer

BLAST Results for significant hit(s):

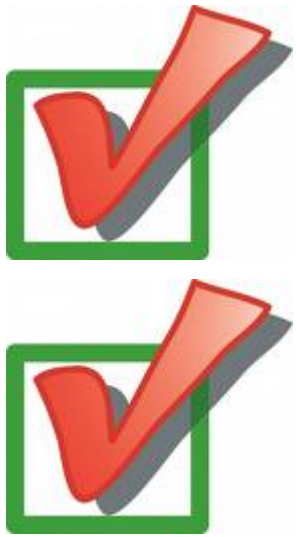
Gene product name (top hit)

Organism

Length of alignment, Score, E-Value, Identities, Positives, Gaps

Pair-wise alignment of the database hit and the query sequence

Module tasks complete



Are you keeping up with your annotations?

The 1st of 3 *imgACT* notebook checks
will occur at the end of this week.